Contents

Chapter 1

Bioinformatics & Functional Genomics

1.1 Sequencing Technologies

Reference: [Yunlong Liu's lectures at YouTube], [Metzker, Sequencing Technologies - the next generation, NRG, 2010], [Illumina NGS introduction]

Sanger sequencing [Brooke, Genetics, Figure 18-19]

- Sequencing by synthesis principle: similar to analysis of mixture (e.g. SDS-PAGE, HPLC), use synthesis to create a mixture of smaller molecules, then identify the compoenents of mixtures.
- Need primer and DNA polymers.
- Terminator (ddNTP) and normal nucleotide mixture: during synthesis, random stop and creates a mixture of oligonucleotide with different lengths. This mixture can be separated because of length difference and each oligonucleotide can be identified with the labeled ddNTP.
- Remark: let p be the fraction of ddNTP, then to get a length L segment, need to incorporate a normal NT at each of the $(L-1)$ step. The probability is $(1-p)^{(L-1)}$, which limits the length L. Also, at large L, it is harder to separate different segments which differ by only one NT.

Strategy of next generation sequencing (NGS) and the challenges:

- Strategy: break DNA into pieces, and read the multiple pieces simultaneously in spatially separate wells/beads.
- Amplification: for any single piece, if only one nucleotide is incorporated, the fluescence is too weak for the imaging system, so most systems require amplification of templates.
- DNA synthesis: to read one bp a time, need the reactions to proceed in stepwise fashion.

Illumina platform: [Liu lecture; Illumina NGS introduction, Figure 3]

- Challenge: physically separate millions of DNA synthesis reactions? How to start them: need primer and amplification.
- Idea: use adapters for DNA (server as primers), and put adapters in the surface (physical separation). Need clusters of identical fragments to generate enough signal during the sequencing step.

- Step 1: Library preparation. Fragmentation of DNA sample, and ligation of adapter molecules in both ends of fragments. Adapters allow hybridization of fragments to flow cells in step 2 and contain PCR primers. Take less than 90 mins.
- Step 2: Cluster amplification. The library is loaded to a flow cell, and the fragments hybridize to the surface. Then bridge amplification: repeated synthesis and denaturation add DNA sequences in the surface (clusters).
- Step 3: Sequencing by synthesis. Now we have clusters of DNA fragments bound to the surface. To sequence each cluster, at each cycle, at four NTs (reversible terminators) with different colors: one cluster would have one specific color per cycle, which can be read.
- Library multiplexing (Illumina, Figure 5): Index for each library: a short k-mer (unique to each sample) ligated to DNA fragments. Pooled sequencing, then demultiplex (removing index sequences from the reads).
- HiSeq: 1B short-reads (35-100bp) in one run. Single-end or paired-end. 30X coverage of human genome. Error rate: about 1-2% per reads (i.e. about 1bp error in a 75bp read), and the error rate is much higher in the 3' end.
- Paired-end sequencing: sequencing from both ends of a fragment. The insert size is known (range). Can control for insert size (from a few hundred to 10k) during library preparation.

SOLiD platform:

- Idea: sequence each bp twice to reduce the error rate.
- Amplification: emusion PCR. In liquid (water-oil) bubbles, some contain one bead (attached with 3K linked primers) and one DNA molecule. In these bubbles, do PCR, eventually, one bead would have 30K identical DNA linked to the bead.
- Sequencing by ligation: many probes, each probe is eight-mer, but only the first two bases are informative, with 4 possible different colors (16 diNT and 4 colors - 2 base encoding or 2BE). Repeated ligation with probes so that every dinucl. is read. The primer starts at different positions, so that each base is sequenced twice.
- Properties of 2BE: each change of nucl. (e.g. SNP) will create two color changes. So if there is only a single color change, it is measurement error.

PacBio:

- Single molecule real-time (SMRT) sequencing. Many holes, within each hold, a DNA polymerase. Whenever a labeled NT enters the hole (used by polymerase), the fluorescence is detected.
- Advantage: very long reads ($>$ 1K, could be 5K).

Nanopore sequencing:

- Reference: `https://nanoporetech.com/community/specifications`
- MiNION: 512 pores per flow cell. Run time: 48hrs. Speed: 70bps or 500bps (fast). Read length: 5-10kb. Error rate: 5-30%. Number of reads in a single run: 70*48*3600/10k = 600K. Yield: 600K * 10K = 6Gb. Regent cost per run: \$99. Flow cell cost: \$500-\$900. Flow cell life time: 72hrs.
- Applications of nanopore sequencing: Rapid Pathogen identification, Human repeat regions, transcriptome reconstruction.

Concepts of NGS:

- Single-end vs. paired-end: In DNA fragments of length 200-500bp, could sequence from one-end of length about 100 bp (Illumina) - single-end; or sequence from both ends of length about 100bp - paired end. For SOLiD, the lengths are slighly different. Paired-end sequencing is particulary good for RNA-seq (improve alignment).
- Mate-pairs: similar to paired end, but much longer fragments, and good for structural variation and de novo assembly.

Comparing different sequencing methods: on number of reads and read length, sample needed, amplification requirement, speed and error rate

- Illumina, SOLID: read length is limited, may be due to bridge amplification (the bridge must be short), or higher error rates at the end of growing primer.
- Single molecule sequencing: need smaller amount of sample; no amplification, so better for quantitative measurements (e.g. RNA-seq)
- PacBio: longer reads, faster but higher error rate (very short interphase time).

Biases and other adverse factors that may affect NGS data accuracy [Wang, NGS book, Chapter 4]

- Biases: may affect the representation of DNA fragments.
- Biases in DNA fragmentation and size selection: sonication can introduce bias, e.g DNA strand breaks after C are more often than expected. Size selection may favor fragments with high melting temperature (GC rich).
- Biases in adapter ligation: usually end repair, add 3-dA tail in DNA fragments and 5'-dT overhang in adapter. But this is biased against DNA fragments starting with a T.
- Biases in PCR: biased against extreme GC or AT-rich sequences.
- Biases and errors in sequencing: similar to PCR, DNA polymerase in sequencing has the same bias. Also equipment operation error/biases, e.g. air-bubbles, dust.

Early-stage NGS data analysis: common steps [Wang, NGS book, Chapter 5]

- Main steps: (1) Base calling: image processing. (2) Data QC and preprocessing. (3) Read mapping.
- Base calling: The result is FASTQ format. Ex. Illumina file, one read per row, with sequence and quality score (per base). About 250GB from HiSeq 2000 (could be 200M reads).
- Genome alignment: The result is SAM/BAM file. about 1TB from HiSeq 2000.

Quality metrics in NGS data analysis:

- Base quality: for each nucleotide, quality of base calling, reported by sequencer. Phred scores: $-10\log_{10}(p)$ where p is the error probability (when we call a base, the chance that it is wrong). So Phread score of 20 is generally required, corresponding to an accuracy of 99% (error rate 0.01).
- Mapping (alignment) quality: for each read, reported by the aligner. Ex. how many mismatches, indels.
- Consensus quality (variant call quality): for each genomic locus, reported by the variant caller.

Raw sequence data: FASTQ format. Similar to FASTA, four lines per read: Seq_ID, sequence, + (optional comment), quality of each base.

- Base quality: Phred score Q. The Q value is encoded using ASCII code: i.e. the integer number of Q is mapped to a character by ASCII. Sometimes, a shift is applied to the integer value of Q.

- Variant of sequence data format: .pseq from Illumina. One read per line: machine number, run number, lane number, single end or paired-end, etc.

NGS data QC and preprocessing [Wang, NGS book, Chapter 5]

- Basic goal is to examine the quality of data, and filter low-quality data, at the base/position or read level. Two strategies: (1) Examine metrics for quality: e.g. Q-scores; (2) Based on possible sources of errors, identify signatures of errors: e.g. duplicate reads from PCR.
- Q-scores: most Q-scores should be about 30. Even in late phase positions, most should be above 20. Could also use %N (missing bases): should be low.
- Percent of each base across base positions: should be constant across positions.
- Signatures of errors: existence of artificial sequences including adapters and PCR primers, duplicate reads.
- Filtering: remove low-quality reads; low-quality base calls and artificial sequences should be trimmed. Remove duplicate reads (better done after mapping, see read alignment section).
- Software for NGS QC: FastQC, NGS QC Toolkit, FASTX-Toolkit.

Computing needs for NGS data management and analysis [Wang, NGS book, chapter 6]

- Challenges of NGS data: (1) Storage: a single run can generate 10-100 GB data. (2) Sharing and transfer: centralized repository or cloud.
- Software tools for NGS data: (1) Galaxy: bridge system to run command-line tools, OS-independent, run on a web browser. (2) Bioconductor.

1.1.1 Sequence Alignment

Alignment file format: SAM/BAM format (BAM is the binary version of SAM). Headers and in the main text, one alignment (for one read) per line: many columns. [Yunlong Liu's lectures; Wang, NGS book, 5.3]

- Reference: Li et al, The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009
- The headers: @HD: file information, the sorting order of alignment. @SQ: reference sequence information, the species, assembly, etc. @RG: the read group information, the sequencing center, etc. @PG: alignment program.
- Alignment fields see [Wang, NGS book, Table 5.1].
- Col 2: FLAG, pairing, strand, alignment information, etc. Ex. FLAG = 163, in binary format, 11000101000, each bit encodes some information. The first bit - whether it has other fragments (1 for paired-end). The second bit - whether the read is aligned.
- Col. 4: the leftmost position of the alignment, often 1-based (including). Sometimes (e.g. in SAM), 0-based (C++ style).
- Col 5: Map quality, in Phred-scale.
- Col 6: CIGAR string, the alignment information, e.g. 8M2I4M1D3M: 8 matched positions, one insertion of size 2, 4 matched positions, 1 deletion of size 1, then 3 matched positions.
- Col 9: inferred fragment size, for paired-end reads (0 for single-end reads).
- Col 10: sequence, col 11: the quality scores

- Optional fields: alignment score (AS), CM (edit distance with the template), H1/H2: number of indels of size 1,2, and so on.

Overview of read mapping: [Yunlong Liu's lectures]

- Background: principle of indexing. In general, if we need to search repeatedly of units in a large collection, then it is better to build an index of the units on the collection (where each unit occurs).
  - Example 1: k-mer (read) search in a large genome. Build an index of k-mers on the genome.
  - Example 2: word search in a large document collection. Build an index of words on all documents.
- Consideration of short-read aligners: in addition to speed and memory, alignment quality (whether allow gaps), base quality scores.
- Hash-based strategy for aligning short-reads: seed-extension paradigm
  - Seeds: suppose length is k, we build an index of all k-mers (size 4^k), and implement with a hash.
  - Seed-extension: for any query, first map the seed, then extend: check which of the matches also extends.
  - Existing methods: may hash reads (Eland, MAQ) or hash the genome (BFAST, SOAP).
- Allowing mismatches: given a query of length m (or seed length), find all genomic locations that match the query in all but k positions (usu. $k = 1$).
  - Strategy: suppose we have the query, ACTT, define many possible "patterns" (similar to regular expression), e.g. ACT*, AC*T, A*TT, or *CTT. Then we build indices of all these patterns. To search for query: first find all patterns that match the query (with the mismatch requirement), then extract genomic locations of these pattern using the indices.
  - Remark: this is similar to the problem of regular expression search - need to build indices of regular expressions.
  - For sequencing problem, the pattern is defined by the location of wild cards (masks). There are many possible designs of the patterns based on the number and positions of masks: suppose the seed length is $m = 8$ and $k = 1$, we could define patterns as: 11110000 and 00001111, where $w = 4$ (w is the number of counted bases), or 10101010 and 01010101 where $w = 4$; and so on.
  - Optimal design: exists for any given m, k and w.
- Suffix tree: the hash-based method does not work well for alignment to multiple identical copies of a substring in the reference.
  - The algorithmic problem is: given a string S, design a data structure of the string s.t. searching of any substring of S is efficient.
  - The suffix tree is a tree representation of the string: (1) each edge represents a substring; (2) each path from a leaf to the root represents one suffix of the string.
  - Why substring search is efficient? Suppose we search Q in S, if Q is a substring of S, then Q must be the prefix of some suffix of S, say F. Then Q must be in the path corresponding to F, starting with the root.

Examining and operations on mapping files [Wang, NGS book, 5.3]

- Under ideal conditions, most aligners map about 70-75% reads. This is due to: repetive regions, short reads, sequencing error, polymorphism, algorithmic limitations.
- Multi-reads: usually should be removed. Some programs will probabilistically allocate the reads.

- Duplicate reads: what may cause them? (1) DNA fragmentation process unlikely to create exactly identical fragments. (2) Adapter ligation and PCR, fragments generally have multiple copies. (3) Sequencing: random sampling process, usually one fragment is sampled only once. However, if some fragments are heavily amplified (PCR overamplification), we will have duplicate reads. When DNA amount is very small (low library complexity), we may have significant duplicate reads.
- What to do with duplicate reads? Because they come from PCR, they do not represent the true proportion in the original DNA pool. In general, we should remove duplicate reads.
- Detection of duplicate reads: because of sequencing errors, the duplicate reads are not necessarily 100% identical. Better to detect duplicate reads after mapping.
- Tools: SAMtools and Picard. Many possible operations with SAM/BAM: e.g merging files, sort reads, read filtering (e.g. duplicate), indexing reads for fast access.
- Visualization can be important: IGV.

Burrows-Wheeler transform (BWT) [Wiki]

- Motivation: if we cn put the same substrings together, then we can compress. Ex. text with many "the"s. Consider all rotations, then we have many strings like "he ... t". If we sort them and take the last column, there will be many ts that we can compress.
- Definition of BWT: given a string X, we have a table of all its rotations, then sort the strings, and the last column is the BWT of X.
- How inverse BWT works? After BWT, we take last column (BWT), then imagine we append it to the first column, we have the pairs of consecutive characters. The first column is easy to know just form the BWT string because it is sorted. Thus we can reconstruct all length-2 substrings of the original string. Repeat this process of: appending BWT string, then sort. To see that this is true:
- Lemma: for any string in the table, the last and first column are always consecutive in the original string. More generally, the last and the first K-columns are consecutive.

Sense from sequence reads: methods for alignment and assembly [Flicek & Birney, Nature Methods, 2009]

- Comparing with database searches (BLAST): millions of read to be aligned from each run, so computational requirement is much higher; scoring of alignment quality is not based on evolutionary data, but polymorphism.
- Strategy: heuristically finding the putative genomic regions, followed by fine-scaled alignment in putative hits. Note that the alignment score should take into account the sequence quality (low quality sequence allows more mismatches).
- Hash-based alignment: Examples: MAQ, SOAP.
  - Indexing: either the short reads or the reference sequence could be indexed, and the other one is aligned to the index. Difference is memory requirement: e.g. indexing genome would require constant memory.
  - Creating hash table: choose index of the form, e.g. 110011 (1: exact match, 0: mismatch). Ex. 4-mer index would be 00AC, 00AT, ..., TC00, TG00. For each read, a region used for seed selection, check if the region matches to any index.
  - For each short read, generate multiple entries in the hash corresponding to possible mismatches. Then do a refined alignment.
- Burrows-Wheeler transform methods: 10-30 times faster than hash-based method.

- First step: BWT of the reference genome.
- Build FM index that enables fast search: storing suffix arrays.

Fast and accurate short read alignment with Burrows Wheeler transform [Li & Durbin, Bioinfo, 2009]

- Background: prefix trie. Given a string X, we can efficiently store all its prefix strings, in a way that supports efficiently query (of substring). The trie should compress the common suffix of prefix strings: e.g. GOOGOL has prefixes: G, GO, GOO, GOOG, GOOGO, GOOGOL. Two prefix GO and GOOGO has the same suffix. We would "share" the common suffix in the trie.
- **Observation**: any substring of a string S must be the suffix of some prefix of S, and as such, is a path from some node to the root of trie of S.
- Searching a substring using prefix trie: this could be achieved by DFS, we match the branch with the query string at each step until we find the match.
- Suffix array and suffix array interval: Figure 2, the suffix array is the order of the rotated strings in the BWT of X. For any query string W, its position at X must be an interval of the suffix array.
- Exact matching by backward search: the problem of searching is reduced to finding the boundaries of the suffix array interval. One can show that there is an optimal substructure of the two boundaries. The aglorithm is $O(|W|)$ for query string W. This is basically prefix trie search without explicitly putting the trie in memory.
- Inexact matching: find all matches with at most z differences. The idea is to do DFS on the prefix trie. However, given W, we can obtain the array $D(\cdot)$: which basically says at each position i, how many mismatches we expect in the prefix $W[0, i]$. This would allow us to reduce the search space.

1.1.2 Genome Assembly

Fragement assembly problem (whole-genome short-gun sequencing) [Jones, Introduction to Bioinformatics Algorithms, 8.3, 8.4, 8.9]

- Fragment assembly problem: reads from Sanger sequencing - 500-700bp. How to assemble them to form the whole genome?
- Statistical consideration: criteria of a good assembly? Example, suppose we have reads, A, B, C, D and E, what is the underlying sequence (S)?
  - Error consideration: the reads are generated from the sequence S. Because of low error rates, the actual sequence of the reads should match S well. A related consideration is whether the coverage is relatively uniform (as expected).
  - Sequence length: even the reads are error-free, we have certain preference for S. For example, the sequence could simply by the sum of all reads, but this sequence is quite unlikely because we expects reads to overlap, or the sequence should be short.
  - In summary, the sequence should be relatively short (a parsimonious explanation of the reads), and the errors should be low. On the other hand, parsimony may not be biologically correct, since real sequences often contain many repeats.
- Algorithmic formulation: Shortest Superstring problem. In the absence of any sequence errors, the problem can be formulated as: given n strings $S_1, \cdots, S_n$, find a string S that contains all S_i's as substrings with minimum length.
  - Example: given two strings, the problem is simple: find the overlap between S_1 and S_2 (suffix of S_1 and prefix of S_2). So the general strategy is to identify all overlaps (tail-to-head), and connect them.

  - Intuition: all reads can be ordered by their positions in the genome. Thus the problem is to find a path from the start to end reads. The challenge is that one reads may overlap with multiple reads.
  - Formulation: represent each string as a node, and the edge of the two nodes represents two overlapped reads with the weight corresponding to the overlap size. Given any path, it can be easily show that the total lenth of the resulting sequence would be the sum of read lengths minus the sum of read overlaps. So to minimize the sequence length is equivalent to maximize the total overlap. If we set the weight of the edge as $-overlap(s_i, s_j)$, then the problem is to find a path with minimum weight, and can be reduced to TSP (ordered graph).

- Challenges: sequencing errors (1-3%), double strands, repeats are especially challenging.
  - Alu (300bp) repeats more than 1M times (with only 5-15% variation). And 200,000 LINE repeats (about 1,000 bp)
  - Large scale repeats: human T-cell receptor locus contains five closely located repeats of the trypsinogen gene, which is 4 kb long and varies only by 3% to 5% between copies.
  - Dealing with repeats: BAC (about 150K bp) - repeats are far less likely within a BAC. Mate-pair reads with fixed gap.: assemble the reads of length 500-700 bps.

- The problems caused by repeats:
  - Error vs. repeats: e.g. we have a cluster of reads from a region, A, B, C, D and E. But we may have a read B' that is very similar to B from a different region in the genome. The decision is between: (1) B' is from the same region but with error; (2) B' does not belong to this region: this could be modeled by a path $A - B - C - D - E - B'$ where $E - B'$ has zero weight.
  - Order ambiguity: suppose the true sequence is $R - X - R - Y - R$, where R is a repeat. The problem is that it is indistinguishable from $R - Y - R - X - R$ since the boundary reads are identical in both cases ($R - X$, $X - R$, $R - Y$, $Y - R$).

- Strategy: overlap-layout-consensus approach
  - Overlap: Finding potentially overlapping reads, ie. find the best match between the suffix of one read and the prefix of another. Because of sequencing errors, the step involves a variation of dynamic programming for alignment.
  - Layout: Finding the order of reads along DNA. The hardest step, NP-complete. In practice, heuristics are used, e.g. greedy approach, at each step, merge the two reads with highest overlap.
  - Consensus: Deriving the DNA sequence from the layout. At any position, from the reads aligned at that position, find the consensus.

Sequencing by hybridization (SBH) [Jones, 8.5-8.8]

- Universal DNA arrays: contain all possible probes of length l. The idea is that by hybridization, the information of all k-mers in the sequence will be provided (no positional information yet).

- A different perspective on the sequencing problem: we could view the problem from the perpestive of basic units/k-mers of the sequences, instead of reads. Ex. in the overlap graph approach, maximum the overlap among reads is (roughly) equivalent to maximizing the coverage of each position. Or if k-mers are taken as unit, this is roughly maximizing the support of all k-mers.

- Idea: if we know all k-mers of the sequence S, then we may be able to reconstruct S. Intuitively, this is simply a path through all $L - k + 1$ k-mer's of S, where L is the length of S. Obviously, each pair of k-mers overlap at a $k - 1$-mer, so the k-mer information can be represented as a graph: its nodes are the $k - 1$-mers, and two nodes are connected by an edge if it is supported by a k-mer in the reads. The problem is then to find a path through all the edges.

- Euler path problem: given a sequence S, define the spectrum of S as the set of all k-mers of S. The problem is: given the spectrum, what is the sequence S? We first construct the de Brunjn graph from the spectrum, then finding a sequence containing all k-mers from the spectrum corresponds to finding a path visiting all edges of the graph. This is the problem of finding an Eulerian path.
- Algorithm for finding Euler cycles:
  - Theorem: A connected graph is Eulerian if and only if each of its vertices is balanced (equal in and out degrees). Proof idea: start with a node v, and take the path using untraveled edges, until it cannot proceed. It must end with v because of the balancing condition. Repeat this process until all edges are traveled (possible because of balancing). Then show that all paths can be merged.
  - For Euler path: the start and end nodes are semi-balanced. We can simply add an edge from the end to the start. And the Euler path problem can be easily converted to the Euler cycle problem.
- Problems/challenges:
  - The de Brunjn graph may not be Eulerian because of (1) errors (a dead-end branch from some node); (2) gaps from insufficient coverage (the graph may be disconnected) or the nature of data (e.g. RNA-seq).
  - Repeats: create circuits in the graph, e.g. given the sequence R-X-R-Y-R, it will have two circuits, one from R to X to R, another from R to Y to R. The Eulerian path will not be unique.
  - Coverage information not used.
  - Loss of information in the reads.

Sense from sequence reads: methods for alignment and assembly [Flicek & Birney, Nature Methods, 2009]

- Comparison of De bruijn graph approach with overlap graph approach (read-centered), (1) A read is split into multiple nodes; (2) Repeats represented differently, e.g. a repeat ABACD, in the graph, it will be A, then a loop to B and back to A, then to C to D. (3) Computational time: scales linearly with the number of reads (rather than quadratic). Intuition is that the k-mers heavily compress the data.
- Procedure of De bruijn graph approach (Figure 3)
  - Create the graph: hash all k-mers. Each k-mer a node, and two nodes are connected if they overlap by (k-1) mers.
  - Graph simplification: linear stretches.
  - Error correction: remove low-coverage tips and bubbles (due to sequence errors).
  - Read the sequence.

Genome assembly reborn: recent computational challenges [Pop, BiB, 2009]

- Computing overlap graph: could take $O(n^2)$ time, as one needs to compute overlap for each pair of reads. In practice, through simple indexing strategies (e.g. exact k-mer), close to linear time.
- Greedy algorithm: an unassembled read is chosen to start a contig, which is then repeatedly extended by identifying reads that overlap the contig on its 3' end until no more extensions are possible. The process is repeated in the 5' direction using the reverse complement of the contig sequence. The assembly continues in an iterative fashion by scanning through the unassembled reads.
- Why greedy algorithm may fail? Suppose we have a true sequence A-B-C-D, but A and C are repeats, then we may have A-D have the highest overlap (effectively from C-D region), and this leads to two contigs A-D and B-C. See Figure 1.

- Practical strategies of greedy algorithm: Reads are considered (perhaps in both intitiation and extension) in decreasing order of quality defined by a combination of (1) depth of coverage (other reads confirming sections of read); or (2) quality values and (3) the presence of at least one perfect overlap with another read. To avoid misassemblies the extension process is terminated once conflicting information is found, i.e. two more reads could extend a contig. Figure 1D.
- OLC approach: overlap, the layout step, then consensus. Layout step:
  - Construction of unitig: unambiguous path, no forks. Error removal to reduce simple forks due to sequencing errors.
  - To reconstruct longer fragments, use other information, e.g. mate-pairs. Algorithms such as network-flow analysis.

De novo assembly of short sequence reads [Paszkiewicz & Studholme, BiFG, 2010], Assembly algorithms for next-generation sequencing data [Miller & Sutton, Genomics, 2010],

- Challenge of de novo assembly using short reads:
  - The number of reads is three to four order of magnitude larger than Sanger sequencing, thus OLC method does not scale well (require evaluation of pairwise overlaps, thus quadratic to the number of reads)
  - More sensitive to errors
  - Repeats: could be longer than reads.
- Review of genome assembly in the past:
  - Typical Sanger draft sequences have so-called contig N50s of 20-200 kilobases (half of all bases are in contigs of this length). The reads are at about 1000bp, and the error rate is 0.1% (vs. $> 1\%$ of NGS). Cost using Sanger technolgoy remains at tens of millions of dollars per genome.
  - Most NGS-based genome assembly efforts use 454 or a hybrid of short and long reads.
  - Panda genome assembly [Nature News and Views on Panda genome]: 73-fold total coverage of the panda genome with 50- and 75-base-long reads (about eight times the average coverage of a typical Sanger genome project). Contig N50 of 40 kilobases, and these contigs were joined to yield scaffolds with an N50 of 1.3 megabases. Total of 3,805 scaffolds, compared with less than 100 in the dog. Cost about 1M in raw data production.
- Feasiblity of NGS de novo assembly by simulation: ex. E. coli, with 30-base reads, 75% of the genome could be assembled into contigs of longer than 10kb and 96% of genes were covered by a single contig. C elegans, 51% of the genome is covered by contigs of at least 10kb. However, in the real data, the errors are not uniformly distributed, and this reduces the quality of assembly.
  - EULER, ALLPATHS deal with errors by correcting errors before assembly
  - Velvet: remove the low-coverage contigs, by setting a parameter "coverage cut-off".
- Assembly algorithms differ in how to deal with errors, sequence variations (heterozygosity), resolve repeats, reverse complement, etc.
- Iternative exension approach:
  - SCAKE: k-mer extension. For the starting read u, take the 3' k-mer of u and check if there is a perfect of the k-mer in another read. If found, extend the read u; if not, reduce the value of k (until it reaches a defined threshold).

  - VCAKE: similar to SCAKE. The difference is: consider all reads overlapping with the seed sequence and extends the seed sequence one base at a time using the most commonly represented base from these matching reads, provided that the set of reads passes certain conditions.

- Comparison of Eulerian path strategy and OLC:
  - Pairwise overlaps between reads are never explicitly computed, hence no expensive overlap step is necessary, rather overlaps are implicitly represented in the deBruijn graph
  - Efficient algorithms exist for finding a Eulerian path in a graph, in contrast to OLC.
  - However, in general, an exponential number of distinct Eulerian paths can be found in a graph, corresponding to the many different ways a genome can be rearranged around its repeats.
  - Chopping up the reads into a set of k-mers results in a loss of long-range connectivity information implied by each read.
  - Errors lead to the creation of 'new' k-mers and dramatically increase the size and complexity of the resulting deBruijn graph

- Scaffolding: further assemble contigs into longer assemblages known as scaffolds. This step exploits additional information in paired sequence reads and/or conservation of gene-order in related biological species.
  - Use read pairs: the contigs are modelled as nodes and matching read-pairs are modelled as edges connecting the pair of contigs. Again the algorithm consists of finding an optimal path through the graph. Because of errors, some read-pair information may be false. So in practice, multiple read-pairs are needed to establish the order and orientation of contigs: e.g. 5 in ABySS and 3 in SOAPdenovo.

- Metrics of sequence assembly:
  - Continguity: N50 length.
  - Accuracy: (1) base accuracy: frequency of calling the correct nucleotide at a given position. (2) Mis-assembly rate refers to the frequency of rearrangements, significant insertions, deletions and inversions.

- Transcriptome assembly challenges: fewer repeats but unique challenges, alternative transcription and alternative splicing, extreme sampling bias (a few transcripts dominate), contamination of genomic DNA, etc.

- Sample preparation approach to overcome the limitation of short-reads: e.g. break down the genome into multiple libraries of different size-classes. Or first label the 550-bp fragments from the genome and then prepare "a library of fragments in which every short fragment is linked to a label that indicates which 550-nucleotide it came from".

- Filtering sequence reads before de novo assembly: reads containing 'N's, any homopolymer reads (e.g. strings of As), trimming of reads using quality scores. This may improve assembly and reduce memory usage. However, with velvet, score-based trimming and filtering does not yield improvements in assembly quality.

An Eulerian path approach to DNA fragment assembly [Pevzner & Waterman, PNAS, 2001]

- Idea: (1) correcting errors in the reads before constructing de Brunjn graph; (2) some of the read information is lost in converting to the de Brunjn graph, restore it by using "read paths".

- Error correction: the general idea is that even though S is unknown, the set of all k-mers of S (its spectrum) is known approximately.

  - Spectrum Aligment Problem: suppose we first obtain the spectrum T of the sequence, e.g. using all solid k-mers (the k-mers with enough support). We use T to correct all reads. For each read s, find the minimum number of mutations in s so that s is consistent with T.
  - Error Correction Problem: a better model is given all reads, and a specified number Δ (the upper bound of the number of errors in reads), introduce at most Δ corrections in each read s.t. the total length of S_k (the specturm of S) is minimized.

- Eulerian Superpath Problem: Every read corresponds to a path in the de Bruijn graph called a read-path, and the fragment assembly problem corresponds to finding an Eulerian path that is consistent with all read-paths.

Velvet: Algorithms for de novo short read assembly using de Bruijn graphs [Zerbino & Birney, GR, 2008]

- Challenges of assembly using short reads: the overlap graph is extremely large. With repeats, more ambiguity.

- Why De Bruijn graph is good for high-coverage, short reads? Reduce the redundancy. Example, given any short segment, there are many reads covering this segment, thus in the overlap graph, many nodes with about the same information. If we use k-mers as elements, then only need a single set of nodes to represent all the information.

- Graph structure and construction:
  - Graph structure: a node is a series of overlapping k-mers, or "string graph", where the basic nodes are strings, and two nodes are connected if the superstring consisting of the two is supported by a read. The idea is that in a de Bruijn graph, many consective nodes can be merged into a single one to reduce graph complexity. Also each node has the reverse complete of each of its k-mers (this is called a block).
  - Constructing graph: reads are first hashed according to k-mers, $k = 21, 25$. For each k-mer, the hash table has the reads of the k-mers. This step represents each read as a set of k-mers. Another hash table of reads, recording information of which of its k-mers are overlapped by another read.
  - Graph simplification: merging the nodes that have only one out-degree with the nodes with only one in-degree.

- Error correction: remove errors using graph topology
  - Read errors in the edge: the error will create a set of additional nodes, and they manifest as a "tip" in the graph. Because the errors are in the end, the tip are typically short. Velvet removes tips of length less than $2k$.
  - Interal read errors: will create bubbles in the graph. Similar to the first case, an error create a branch, but because it is internal, the branch will later by joined by the main path (after passing the error). Velvet uses a Tour Bus algorithm to remove bubbles.
  - Other kinds of errors, e.g. an indel creates errorenous connections: they are typically low-coverage reads. So remove low coverage nodes in the graph.

- Resolving repeats: velvet uses read-pair information to resolve repeats.

- The selection of k:
  - Large k-mers: the sequence graph will be broken in many places and it will be difficult to determine if a dead-end branch arises from a read error or from a lack of k-mer coverage.
  - Smaller k-mer: increase the connectivity of the graph by simultaneously increasing the chance of observing an overlap between two reads and the number of ambiguous repeats in the graph.

  - In practice, given the limited number of possible values for k, it is common to try out various values in parallel then choose the one that produces the highest N50 contig length.

- Parameters for The Tour Bus algorithm: to decide whether two paths should be merged based on (1) both paths must contain less than 200 nodes; (2) their respective sequences must be shorter than 100 bp; (3) the sequences must be at least 80

- Computational considerations: Velvet has four stages: hashing the reads into k-mers, constructing the graph, correcting errors, and resolving repeats. The main bottleneck, in terms of time and memory, is the graph construction. The initial graph of the Streptococcus reads needs 2.0 Gb of RAM.

Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler [Zerbino & Birney, PLoS ONE, 2009]

- Pebble: use read-pair information to resolve repeats.

- Rock Band: exploit sparse long read datasets within a short-read assembly to resolve repeats and extend contigs.

- Identifying unique nodes (assembly contigs): the first step for both algorithms. For any contig, we compute average contig coverage depth, and if the contig is not unique (ie. repeats), then the coverage would be significantly higher than the expected. To obtain the expected coverage, compute the median coverage (length-weighted) over all contigs (ρ). The log-odds ratio of the contig is computed, testing if the node being unique over that of it being twice in the genome (Poisson distributions). The unqiue nodes are those with uniqueness cutoff $F \geq 5$.

ABySS: A parallel assembler for short read sequence data [Simpson & Birol, GR, 2009]

- Distributed de Bruijn graph: each computer node stores part of the graph, defined by a set of k-mers.
  - Hash table of k-mers: for each k-mers, store the information where (which computer node) the k-mer is located. Design the hash function s.t. the set of all possible k-mers are evenly distributed over available nodes.
  - Adjacency information: for each k-mer, also store whether one of eight possible edges (four possible extensions in either side) exist in the graph.
  - Implementation: uses the MPI (Message Passing Interface) protocol for communication between nodes.

- Assembly algorithm: read error detection by removing dead-ends and bubbles. After these steps, do vertex merging and contig merging using read-pair information.

- Assembly evaluation: only contigs greater than 100 bp in length. Contigs aligning to the reference genome with fewer than five consecutive base mismatches at the termini and at least 95% identity are considered to be correct, except in the case where an alignment contains a gap greater than 50 kb.

SOAPdenovo: De novo assembly of human genomes with massively parallel short read sequencing [Li & Wang, GR, 2010]

- Preassembly sequencing error correction: less important for small datasets since the errors can be easily removed from the graph during assembly, but essential for large datasets, as this greatly reduces memory usage.
  - Method: the correct K-mers appear multiple times in the reads set, while random sequencing error-containing K-mers have low frequency. Build a 17-mer hash table, and for each read, find the 17-mer that might be wrong (frequency less than 3), and change it to the highest 17-mer allele (if its frequency is above 3).

  - With the human data, the total number of distinct 25-mers was reduced from 14.6 billion to 5.0 billion.

- Computational requirements: use 25-mer for de Bruijn graph.
  - The preassembly error correction of the raw reads was the most time consuming step, costing more than 20h. Remark: dynamic programming for each read, so slow.
  - The de Bruijn graph construction step had the highest peak memory usage (140 Gb). A graph of 5B nodes, thus memory costly.

A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies [Zhang & Shen, PLoS ONE, 2011]

- Background: quality-value file for the high throughput short reads is usually highly memory-intensive, only a few assemblers, for example, SHARCGS, and ALLPATHS-LG adopt it in the assembly process.
- Overview of de novo methods: two types of data structure for efficient access of reads: string-based or graph-based. String-based assemblers, implemented with Greedy-extension algorithm, are mainly reported for the assembly of small genomes, while the latter ones are designed aiming at handling complex genomes. Two types of graph-based strategies: De Bruijn graph and overlap-layout-consensus (OLC).
  - Graph-based: Edena (OLC), Velvet and SOAPdenovo
- Difficulty with repeats: some methods take advantage of paired-end (PE) sequencing, including SOPAdenovo, ABYSS, Velvet.
- Computational and RAM: graph-based methods seem better in these aspects. Compare with 36-mer short reads assembly, only OLC, De Bruijn and hybrid assemblers can be applied for 75-mer short reads assembly.
- Accuracy and genome coverage:
  - Edena has very strong performance on all settings: close to or more than 90% in both metrics.
  - Velvet and SOAPdenovo: lower accurancy but similar coverage. However, when handling large datasets such as short reads from C.elegans genome, SOAPdenovo had similar performance as Edena.
  - String-based tools: not able to run on 75 read data.
- Size of contigs: For dataset of very small size, string-based assemblers produced fewer but longer reads than De Bruijn graph-based tools. However, it became reverse when the size of dataset increases. For De Bruijn assemblers, Velvet produced better assembly result than SOAPdenovo when assembly of 75-mer short reads datasets, because of the wider range of K value to be chosen in Velvet.

Comparative studies of de novo assembly tools for next-generation sequencing technologies [Lin & Deng, Bioinfo, 2011]

- Motivation: (1) the assembly performance of SOAPdenovo (v1.05) has dramatically improved for long read assembly. (2) Impact of factors such as varying depths of coverage, sequencing errors, read lengths and extent of GC content of the sequence reads.
- Metric of contig quality:
  - N50 length is the longest length such that at least 50% of all base pairs are contained in contigs of this length or larger (short contigs could recover most of bps, and at length increases, it is getting more difficult to cover sequences).

  - Assembly error rates (AER), we aligned the output contigs to the benchmark sequence, and calculated the number of mismatched bases from alignment results

- Evaluation: Sequence coverage and assembly error rates were analyzed by blastz.
- Impact of depth of coverage on N50 length: usually reach plateau as depth increases (denote this as DCAP). For single-end, DCAP for SSAKE and Edena is about 50; and for VCAKE, Velvet, ABySS and SOAPdenovo is about 30-40.
- Sequence coverage and error rate:
  - AER: SSAKE, VCAKE, SOAPdenovo and Euler-sr generated higher assembly error rates than Edena, Velvet and ABySS.
  - Sequence coverage: all tools are close, with SCAKE, VCAKE and SOAPdenovo slightly higher.
- Running time and memory: In general, SOAPdenovo and ABySS were more efficient than other tools in terms of runtime and memory usage. Velvet becomes abnormal for large paired-end datasets.
- Summary: Velvet, ABySS and SOAPdenovo has relatively low DCAP, low AER (SOAPdenovo is somewhat higher), and computationally efficient. Edena is somewhere between SCAKE/VCAKE and ABySS.

Genetic variation and the de novo assembly of human genomes [Chaisson & Eichler, NRG, 2015]

- Challenges of WGSA: gaps, repeats (Figure 1)
- De Bruijn graph: repeat representation as two forks, Figure 3B.
- Single molecule sequencing (SMS) assembly: not De Bruijn graph. Pairwise alignment.
- De novo assembly strategies: combination of reads with different insert size. A combination of high coverage reads from a short-insert library (sequences 200500 bp long), a lower coverage of a medium-insert library (sequences 2 kb, 5 kb and 10 kb in length) and a sparse coverage of long-insert (40 kb in length).
- Quality of mammalian genome assemblies: Few modern genome assemblies exceed an N50 contig of 100 kb (average = 41 kb). Tens to hundreds of thousands of sequence gaps, most of which correspond to various classes of repeat and gaps (due to GC content).

1.1.3 Detecting Sequence Variations

Genotyping and Genomic variation discovery [Wang, NGS book, Chapter 9]

- Data preprocessing:
  - Local realignment: indels create problems for read mapping. So (1) identify potential regions where realignment is needed (e.g. GATK RealignerTargetCreator). This could use known indels, eg. from 1000 Genome. (2) Realignment: e.g. GATK IndelRealigner.
  - Base quality recalibration.
- SNV calling: GATK UnifiedGenotyper for genotype calls, and GATK HaplotypeCaller for joint SNV and indel calling. This involves local de novo assembly of haplotypes, then read mapping and make variant calls.
- Somatic and de novo mutations: e.g. JointSNVmix, Mutect (most popular). Model idea, let G_N, G_T be genotype of normal and tumor tissues respectively, the key is the prior $P(G_N, G_T)$. We can write it as $P(G_N)P(G_T|G_N)$, the conditional probability depends on mutation rates.

- Indel calling: basic idea is to build a new haplotype, and count the number of reads supporting the indel in the alternative haplotype and make calls. Could do de novo haplotype assembly.
- Calling variants from RNA-seq: popular tools eSNV-Detect, SNPiR and SNVMix.
- Evaluating VCF: metrics such as HWE, call quality difference between major and minor alleles, Ti/Tv ratio, strand bias.

Overivew of SNV detection [Li & Wang, SNP detection for massively parallel whole-genome resequencing, GR, 2009]:

- Sources of errors in reads:
  - Error bias: in Illumina platforms, $A \leftrightarrow C$ and $G \leftrightarrow T$ errors are more common because of the overlap in the fluroescence spectrum of the dyes.
  - Cycle (coordinate) of the base in the reads: the bases at the 3' end have much higher error rates.
  - Duplicate reads: during PCR, the same errors may be amplified, and are difficult to distingush from true variants.
  - Misalignment: especially a problem when the true genotype in the sample contains indels (relative to the reference genome).
  - Question: one error in amplication should appear in only one cluster, thus generates only a single read?
- SNP detection strategy: infer the (diploid) genotype at each individual sample (model the error at each base); the population level information (SNP rate) can be used for prior, or a multi-sample strategy can be used. Several issues to address:
  - Alignment errors: if the sample contains indels, one may not align the reads correctly, so infer the haplotypes of samples, and realign the reads.
  - Quality score recalibration: the Phred quality score does not capture a number of possible biases/errors, so need to correct for it to obtain the true error probabilities at each base.

Genotype and SNP calling from next-generation sequencing data [Nielsen & Song, NRG, 2011]

- Sequencing depth: typically, low coverage sequencing if coverage < 5, and high coverage if > 20. In general, the power to detect low-frequency variants, or association mapping is maximized by sequencing many individuals at low depth rather than sequencing fewer individuals at a high depth.
- Overview of steps: base calling and alignment; read mapping; realign, remove duplicate reads and recalibrate quality scores; SNP calling (single or multi-sample); filtering and SNP or genotype quality score recalibration.
- Base calling and alignment:
  - Illumnia errors mainly come from the synthesis process becoming desynchronized between different copies of DNA templates in the same cluster. Base calling becomes less accurate in later cycles as the extent of asynchrony is exacerbated with each sequencing cycle.
  - Base calling quality: Phred score of 20 corresponds to a 1% error rate in base calling.
  - Alignment criteria: the optimal choice of the tolerable number of mismatches may differ between different organisms. Ex. more polymorphism in fruit fly than in human, thus different requirement for read mapping.
  - Alignment algorithms: hash-based (generally more accurate) or BWA (faster, memory efficient).

  - Regions with higher levels of diversity: e.g. MHC region, generally difficult for read mapping. De novo assembly may be a good idea.

- Recalibrate base-calling quality scores:
  - Why recalibration? The base calling Phred scores do not take into accout the base position, the cycle, and other features.
  - Recalibration algorithm: training data: sites with no known SNPs. Any base is classified first into different categories based on the features including the raw quality score from base calling, the position of the base in the read, the dinucleotide context and the read group. For each category, the algorithm estimates the empirical quality score by using the number of mismatches with respect to the reference genome.

- Genotype and SNP calling: calling genotypes of individuals or detecting the presence of variants in a sample.
  - Early methods: define a Q (Phred score) cutoff and call genotype of each individual. Typically, using Q20 as cutoff, and call a heterozygous genotype if the proportion of the non-reference allele is between 20% and 80%. Often used in high-depth sequencing.

- Probabilistic methods for SNP and genotype calling:
  - General idea: let X be all reads at a site and G be the genotype at the site, we determine G by $P(G|X)$. This depends on the prior $P(G)$, which can incorporate information such as MAF, and the probability of the i-th read given G (assuming independence of reads), $P(X_i|G)$, which is determined by the recalibrated quality score. The independence assumption can be violated e.g. by alignment errors.
  - Assigning prior $P(G)$: one could use known polymorphisms e.g. HapMap or MAF estimated from the entire sample (assuming HWE).
  - Incorporating LD information: a straightforward adaptation of imputation algorithms can be used for NGS data. The use of LD leads a significant improvement of genotype calling in 1000GP.
  - Comparison of methods: using GATK single-sample, GATK multi-sample and GATK-Beagle (GATK multi-sample followed by Beagle, which uses LD information). Accuracy gains from multiple samples (80 to 87%) and LD (87 to 96%). Note that LD is mostly useful for high and moderate frequency variants.

- Filtering: based on deviations from the HWE, systematic differences in quality scores for major and minor alleles, aberrant LD patterns, extreme read depths, strand bias, and so on.

- Association mapping taking genotype uncertainty into account:
  - If the error structure is the same in cases and controls, tests that are robust to violations from the HWE will not suffer from an excess of false positives, however, the power is reduced.
  - Method: the use of genotype posteriors effectively sums over all possible genotypes. Similar to GWAS using imputed data (genotype uncertainty).
  - Allele frequency estimation (SFS) can also be affected by genotype uncertainty.

- Implication for estimation of Site Frequency Spectrum (SFS): using highest genotype calls (GC) or GC > 0.95 for calling genotypes, the SFS is highly skewed towards singletons. Explanation: suppose a variant occurs in two samples, but the power of detecting this variant in both samples is low, so often we observe it only once. Idea: account for the lower GC probability from many other individuals the summations should approach true frequency.

Variant calling: filters and QC [personal notes; ASC projects]:

- Rate of variants in the data can be used as an important QC metric. Ex. cases vs. controls; T vs. NT. in family data; de novo mutations. It can also be used to compare across different samples, different genders, and so on.
  - Remark: need to understand the expected rates under null hypothesis, e.g. how it depends on the sample size.
- Other kinds of patterns/summary statistics in the data that can serve as QC: ex. frequency of singletons or SFS more generally, frequency of multi-allelic sites. Population genetics can often tell the null (neutral) expectation.
- Sample filters: e.g. the rates of variants in the samples - whether it's much higher than other samples.
- Genotype filters: depth of coverage (DP), genotype quality (GQ), etc. Ex. in ASC, require DP ≥ 10 and GQ ≥ 30.
- Variant filters: often use MAF filter, also multi-allelic variants, etc.
  - Remark: Note that how MAF is defined matters, e.g. in case-control data, ideally cases should not be used. To avoid inflation error, use the external control is better.
- Homologous and repeat regions: may have higher sequencing and mapping errors. Ex. XY homologous regions when mapping reads to X chr.

Sequencing studies in human genetics: design and interpretation [Goldstein, NRG, 2013]

- Considerations of sequencing studies:
  - Adjust analysis criteria (e.g. specificity-sensitivity trade-off for variant calling) for the purpose. Some analysis: e.g. undiagosed disease requires maximizing sensitivity, while case-control studies of complex diseases requires a balance.
  - Possible bias/variations of detecting variants: sample preparation step, sequencing reactions (variation across lanes, flow cells and machines).
  - Coverage: generally recommend 40x. When variants are expected to occur in multiple samples, lower coverage may be OK. Detecting de novo mutations and structural variations generally require higher coverage. In de novo studies, one of parental alleles may be missed with low coverage, and falsely assigned as de novo.
- Variant detection:
  - Ideas for improving the baseline caller: read re-alignment for orphan read pairs, clusters of read pairs with aberrant insert size, de novo assembly of anomalously aligned reads.
  - Multi-sample variant calling: can introduct batch effects. Best to call each individually or call all samples (including cases and controls) together.
  - Popular methods: GATK, Platypus, SAMtools.
- A common problem in variant calling: the difference between a putative causal variant that is truly not present in controls, or merely not called owing to missing or poor-quality data. Need data of metrics for each base: e.g. coverage, read mapping quality, genotype likelihoods, and they can be stored as Genome VCF (gVCF) file.
- Recommendation: relational database of variant calls, annotations, coverage data, quality metrics and sample relationship. SVA is one implementation, but not easy to add new samples. PLINK/Seq: not comptatible with gVCF.

The role of replicates for error mitigation in next-generation sequencing, [Robasky et al, Nature Reviews Genetics, 2014]

- Sources of experimental errors in NGS: sample preparation, library preparation, sequencing. Ex. somatic mosaicism, chimeric reads (error in ligation during library prep).
- Filtering strategies: read depth, quality scores (base, alignment, calling), strand bias (normally reads should be balanced in two strands), allelic imbalance, sequence context. Generally quality scores are recommended.
- Optimizing parameters using replicates: concordant and discordant calls across replicates.

Analysis of accuracy and power of variant calling using binomial model [personal notes]

- Problem: suppose we consider homozygous genotypes only. Let G be the true genotype (either A or a), and X be the number of reads with alternative allele a. We call a variant when X is greater than a threshold d. Suppose sequencing error is ϵ, defined as the probability of a in a read when the underlying genotype is A, and vice versa. Our goal is to estimate false positive rate and power, as a function of sequencing depth n.
- Analysis: the type I error rate is given by $P(x \geq d|G = A)$. The distribution of x is $\text{Bin}(n, \epsilon)$, so the error rate:
$$p = 1 - \text{pbinom}(d - 1, n, \epsilon) \tag{1.1}$$
where pbinom$(\cdot)$ is the R function for binomial distribution. The power is given by $P(x \geq d|G = a)$, and the distribution of x is $\text{Bin}(n, 1 - \epsilon)$, so it is:
$$\text{Power} = 1 - \text{pbinom}(d - 1, n, 1 - \epsilon) \tag{1.2}$$
- Examples: at $n = 5$, $\epsilon = .03$ and $d = 3$, we have, $p = 0.008$ and Power $= 0.999996$. For homozogous variants, need very low coverage.
- Remark: in real data, more complex error model, and we may need to call SNVs across the population, so probably need simulation to do this estimation. See [Quantifying single nucleotide variant detection sensitivity in exome sequencing, BMC Bioinfo, 2012]

mPileup (SamTools): A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data [Li, Bioinfo, 2011]

- Notation: let d_i be the sequence data of the i-th sample, and g_i be the genotype. Let m be the ploidy (2).
- Genotype likelihood: suppose the genotype is g, the genotype likelihood is $P(d|g)$. Let ϵ_j be the error rate of read j. When $d_j = g$, $P(d_j|g) = g(1 - \epsilon_j) + (m - g)\epsilon_j$; and when $d_j \neq g$, $P(d_j|g) = (m - g)(1 - \epsilon_j) + g\epsilon_j$. Write $L_i(g)$ as the genotype likelihood of sample i.
- Estimation of genotype frequencies: our goal is to estimate the true genotype frequencies across multiple samples, e.g. for testing HWE. Let ξ_g be the frequency of genotype g, with $\sum_g \xi_g = 1$. We assume $0 \leq g_i \leq m_i$ for subject i, where m_i is the ploidy. The likelihood:
$$P(d|\xi_1, \cdots, \xi_m) = \prod_i \sum_g L_i(g)\xi_g \tag{1.3}$$
This can be solved with EM. The update rule is given by Equation (7) of the paper. The testing of HWE can be done via LRT.

- Estimation of the number of non-reference alleles: this is required to call a variant (if the number is equal to 0 or not). Let $G = (g_1, \cdots, g_n)$ be the genotype configurations. The likelihood of allele count $X = k$ is:

$$L(k) = P(d|X = k) = \sum_G P(G|X = k) \prod_i P(d_i|g_i) \tag{1.4}$$

The computation involves summing over G, and this can be done recursively using Dyanmic Programming. We first show that, assuming HWE:

$$P(G|X = k) = \frac{\prod_{i=1}^n \binom{m_i}{g_i}}{\binom{M}{k}} \tag{1.5}$$

when $k = \sum_i g_i$, where $M = \sum_i m_i$ is the total number of chromosomes. To see this, we have k alleles to allocate to M chromosomes, and there are $\binom{M}{k}$ ways. For a given G, the number of ways of allocating alleles becomes product of $\binom{m_i}{g_i}$. Next, we write the likelihood as:

$$L(k) = \frac{1}{\binom{M}{k}} \sum_{g_1,\cdots,g_n} \delta_{k,G} \prod_{i=1}^n L_i(g_i) \binom{m_i}{g_i} \tag{1.6}$$

where $\delta_{k,G}$ indicates if G is consistent with k alleles. We define recurrence variable $z_{j,l}$ as the likelihood of data up to sample j subject to the constraint that the total number of alleles is l:

$$z_{j,l} = \sum_{g_1,\cdots,g_j} \delta_{l,G[1..j]} \prod_{i=1}^j L_i(g_i) \binom{m_i}{g_i} \tag{1.7}$$

We have the recurrence equation, considering the genotype of sample j:

$$z_{j,l} = \sum_{g_j=0}^{m_j} z_{j-1,l-g_j} L_j(g_j) \binom{m_j}{g_j} \tag{1.8}$$

- Variant calling: we need to determine the posterior $P(X = k|d)$. The prior $\phi_k = P(X = k)$ can be Wright-Fisher prior, or Allele Frequency Spectrum determined empirically (e.g. EB). The result can be expressed as Phred score: $Q = -\log_{10} P(X = M|d, \phi)$, where d is sequence data and ϕ_k is the prior of allele count (AFS).

GATK: A framework for variation discovery and genotyping using next-generation DNA sequencing data [DePristo & Daly, NG, 2011]:

- Intuition of the model: instead of calling each sample independently, take information from (1) AF in the population: in general, variants are rare in human population, and this should be taken into account. (2) Other samples: the idea is that if other samples already contain variants, the chance that the current sample would contain more variants is reduced (because high-frequency alleles are rare).

- Local realignment: given all the reads across all individuals, first find all possible haplotypes - indels from HapMap, from all reads. Then compute the likelihood of each haplotype and select alternative haplotypes by LRT. Intuition: if the genome contains an indel, there may exist some reads that cover the genome with only a few bp difference, thus still mappable.

- Base quality score recalibration: the true error probabilities at a base depends on the Phred score, R, the coordinate (cycle) of the base in the read, C, and the dinucleotide context, D. In a gold standard dataset (e.g. all positions with no SNPs in the population), we simply tabulate R, C, D and errors (mismatches to the reference). The recalibrated quality score:

$$Q_{\text{empirical}}(R, C, D) = \frac{\text{mismatch}(R, C, D)}{\text{base}(R, C, D)} \tag{1.9}$$

- Single sample SNP calling: suppose GT_i is the diploid genotype of the i-th sample, D_{ij} is the j-th read of the i-th sample. Then D_{ij} may come from either of the haploid of GT_i. For a particular haploid, say B, we have:

$$P(D_{ij}|B) = \begin{cases} 1-\epsilon_{ij} & D_{ij} = B \\ \epsilon_{ij} P(B \text{ is true}|D_{ij} \text{ is miscalled}) & \text{otherwise} \end{cases} \tag{1.10}$$

  where ϵ_{ij} is the (recalibrated) quality score of D_{ij}. The term $P(B \text{ is true}|D_{ij} \text{ is miscalled})$ is: given that this is a miscalling (probability ϵ_{ij}), what is the chance that it will be called as B (D_{ij}?). If we want to infer the individual genotype, we could use population-based prior of genotypes and use Bayes Thereom to infer the posterior probability of GT_i [Li & Wang, 2009].

- Multi-sample SNP calling: we are only interested in if there is a SNP at this position across all samples. Suppose q_i is the number of alternative alleles (B) of the i-th individual, then $q = \sum_i q_i$ is the total number of B alleles in all samples. We are interested in whether $q = 0$ given data. The prior distribution of q is given by the infinite site model (dependent on the homozygotisity). So $P(q = 1) = \theta$, where $\theta = 4N\mu$ is approximately 6E-4 (for one bp). This small prior ensures that variants are called rarely. The likelihood:

$$P(D|q) = \sum_{GT \in \Gamma} \prod_i P(D_i|GT_i) \tag{1.11}$$

  where Γ is the set of all genotype assignments for the N individuals that contain exactly q alleles. The summation can be done via expectation-maximization or exhuasitive summation (exponential number). The probability of a variant segregating at the site at some frequency is given by a quality score:

$$\text{QUAL} = -10 \log_{10} P(q = 0|D) \tag{1.12}$$

- Variant quality score recalibration: still need the filtering step as there are errors not captured by the above model, e.g. strand bias (errors are not distributed randomly). The variant quality score above also depends on some SNP error covariates. Use Gaussian Mixture Model to estimate the probability of each variant call being true, capturing the intuition that variants with similar characteristics as previously known variants are likely to be real, whereas those with unusual characteristics are more likely to be machine or data processing artifacts.

- Remark: the model is conditioned on the total number of alleles in all samples, and involves exponential number of enumerations (the total number of alleles creates dependency between samples). Can we do better here without introducing sample dependency?

A Likelihood-Based Framework for Variant Calling and De Novo Mutation Detection in Families [Li & Abecasis, PLG, 2013]

- Idea: in family data, we can use the pedigree to specify the distribution of genotypes (Mendelian transmission). This is similar to the standard segregation analysis on family data where the phenotypes (comparing with reads in NGS data) depend on genotypes.

- Genotype likelihood: consider one genotype G, and all reads from this genotype R. A genotype can take 10 different values (diploid). Let b_j be the j-th base, then $P(b_j|G)$ depends on the error rate at j, e_j. When G is homozygous, say AA, this prob. is $1 - e_j$ if the base is A; and $e_j/3$ if this is an error. When G is heterozygous, say AC, this is:

$$P(b_j|AC) = \frac{1}{2}[P(b_j|A) + P(b_j|C)] \tag{1.13}$$

  The genotype likelihood (GL) is the product of $P(b_j|G)$.

- Model of reads: consider any site, let G be the set of all genotypes of a family (unobserved), and R be the set of all reads. Then we have this likelihood:

$$P(R) = \sum_G P(R|G)P(G) = \sum_G \prod_i P(R_i|G_i)P(G_i|G_{fi}, G_{mi}) \tag{1.14}$$

  where G_{fi} and G_{mi} are the genotypes of the parent of the i-th individual, and when i is a founder, these are AF in the population. The summation can be done via Elston-Steward peeling algorithm. To call genotype of i-th subject, we compute $P(G_i|R)$ summing over all other unknowns.

- Integrating de novo mutations: the only change is that we replace transmission probability (based on Mendelian law) to mutation rate. So we have the mutation rate matrix in the equation. Finally to call a de novo mutation, we compute the likelihood under two scenariors, and the prob. of de novo is: $L_{\text{denovo}}/(L_{\text{denovo}} + L_{\text{Mendel}})$.

Monovar: single-nucleotide variant detection in single cells [Zafar and Chen, NM, 2016]

- Process of generating reads in SCS: amplification (WGA), then sequencing. During amplification, allelic dropout (ADO) may happen, and amplification may also introduce errors. The model should account for three sources of errors: ADO, amplification and sequencing. Note that ADO happens at the allele level (missing all reads from that allele), while amplification and sequencing errors occur at the read level.

- Genotype likelihood capturing amplification and sequencing error: we consider genotype 0 or 2 first because ADO has no effect. Suppose we have n reads from a sample (cell), let d_i be the read i. Let β_i be the genotype of the fragment after amplification, then if there is no amplification error, our probability of getting d_i is $1 - e_i$, where e_i is the sequencing error rate. If there is amplification error, our probability of getting d_i is $e_i/3$. Let p^g_β be the probability of getting allele β from genotype g after amplification, our genotype likelihood is:

$$P(d|g = 0) = \prod_i [e_i/3(1 - p^0_{d_i}) + (1 - e_i)p^0_{d_i}] \tag{1.15}$$

- Genotype likelihood capturing ADO: ADO only affects the genotype likelihood when the true genotype is 1. In this case, ADO makes the actual genotype g' that is either A or a. This leads to different distribution of sequencing reads. Let p_{ad} be the prob. of ADO. When there is no ADO, the genotype likelihood is similar to before. When there is ADO, we have:

$$p(d|g = 1, ADO = True) = \frac{1}{2}[p(d|g = 0) + p(d|g = 2)] \tag{1.16}$$

- Variant calling: similar to mPileup. The prior of variant allele count comes from population genetics.

- Genotyping single cells: note that even when we genotype only one cell, it's better to use all cells. Intuitively, it is difficult to call a new variant because of prior penalty. When we do the multi-sample calling, our prior of genotype for a single cell is effectively the posterior from all other cells, which may allow us to overcome prior penalty.

- Application to single cell WES data: 50 cells, 60x. Explorative analysis with MDS and hierarachical clustering reveal several subclones/cell populations.

- **Lesson**: even with single cell data, clustering analysis can reveal subpopulations without phylogeny analysis.

1.1.4 Detecting Structural Variations

Genome structural variation discovery and genotyping [Alkan & Eichler, NRG, 2011]

- Types of SVs (Figure 1): operationally we consider SVs of size > 50 bp. Historically > 1 kb. Deletion, novel insertion, tandem and interspersed duplication, inversion and transloation.
- Array CGH: similar to microarry detection of gene expression. Use probes (oligonucleotides, typically 50-75mers) in the array, and hybridze with test and reference samples (labeled), then determine the signal ratio. The copy number in a region directly correlates with the signal intensity (log ratio).
  - Common arrays: 2.1M and 1M per array. Detection of a CNV typically requires a signal from at least 3-10 conseuctive probes.
  - Ultra-high resolution arrays: 24M to 42M probes. Can find CNVs down to 500 bp.
  - Caveat: effect of reference sample: a loss in test sample is indistinguishable from a gain in reference sampel. So the well-characterized reference sample is key.
- SNP arrays: usual arrays used in GWAS, allelic-specific oligonucletodie (ASO) probes. The total signal intensity (regardless of which allele it comes from) reflects the copy number, but in addition, the allele ratios (BAF) in a region provide additional information. Ex.
  - Normal diploid region: normal intensity, BAF is 0 (AA), 1(BB) or 0.5 (AB).
  - Deletion: low intensity, BAF is either 0 (AA) o 1 (BB).
  - Duplication: one more chromosome copy. High intensity, BAF is either 0 (AAA), 1 (BBB), 1/3 (AAB) or 2/3 (ABB).
  - Uniparental disomy (UPD): the two copies are IBD (from father or mother). Normal intensity, BAF is either 0 or 1.

  and so on.
- Comparison of array CGH and SNP arrays: per probe, CGH offers more information, but SNP arrays can detect copy number neutral events such as UPD. Limitations of array based approaches:
  - It is easier to detect deletions in both array platforms.
  - Both platforms lose sensitivity at CNVs of lower than 10kb.
  - Break point resolution: limited.
- NGS based methods:
  - Read pair: information from "discordant pairs". If the pairs are mapped too far, likely deletions; too close, insertions. Can also detect inversions: two discordant RPs with opposite strandeness. Discordancy is defiend based on both the span size and strandness. Orientation inconsistencies can delineate inversions and a specific class of tandem duplications. Most popular, but resolving ambiguous read mapping is an issue.
  - Read depth: copy number gain or loss leads to change of read depth. Break point resolution poor.
  - Split read: split means alignment to the genome is broken. Specific sequence signatures from deletions, insertions, inversions and duplications. Good break-point detection.
  - Assembly: ultimate approach, but difficult, especially in repeat regions.
- Evaluation of methods: for both array and NGS approaches, the overlap of different platforms or approaches (one of the four strategies) is low. Read pair: 90% deletions. Read depth: poor breakpoint resolution, but dosage information. Split read: only reliable in unique regions of genomes (because only one read can be mapped).

- Overcome challenge of lower coverage: pool reads from multiple individuals to detect common CNVs. MoGUL, VariationHunter.

Computational tools for copy number variation(CNV) detection using next-generation sequencing data: features and perspectives [Zhao & Zhao, BMC Bioinfo, 2013]

- Advantages of NGS approaches vs. array: higher sensitivities and ability to detect novel CNVs (multiple signatures), estimation of copy numbers, higher resolution and precise breakpoint.
- Overview: each strategy utilizes one type of signature, and often detects different kinds of CNVs. Ex. large insertions cannot be detected using paired-end reads.
- Paired-end mapping (PEM): clusters of discordant paired-reads, whose distances are significantly different from expected insert size. The key is to identify such discordant clusters: some use predefined threshold, some use proabilistic test. Clustering could be hard clustering (ignore multiple-mapped reads) or soft clustering.
  - Limitation: cannot detect insertions whose sizes are larger than insert size (dependent on library preparation).
- Split read (SR): for a read pair, one can be uniqued mapped, while the other not because of break points introduced by CNVs/SVs. Allow break point detection. This would be a signature of CNV/SV (the unique mapped read reduces false positives).
  - Representative: Pindel. For partially mapped reads, two cases: (1) if deletion, split read into two fragments; (2) if insertion, split read into three fragment where the middle is the inserted part. This means that only small insertion is allowed (1-20 bp).
  - Limitation: rely on uniquely mapped read, thus applicable only to unique regions.
- Read depth (RD): general ides is to divide the genome into disjoint windows, estimated normalized read depth in each window, then merge adjacent windows into regions.
  - Why normalization of read depth? RD can be affected by a number of things including PCR, sequencing, mappability.
  - Main factors affecting RD: GC content, mappability.
  - Methods for normalization: (1) Single sample: comparison with regional read depth, e.g. using Z-score; or normalization by GC content. (2) Paired sample: comparison with control. (3) Multiple sample: comparison with population mean.
  - Methods for merging: (1) Recusrively localize breakpoint: until the adjacant windows have significantly differ RD. (2) MSB: Merging until the new window has significantly different RD with the merged window. (3) CNVeM: allow ambiguous reads using EM. (4) HMM: CNAseg, JointSLM: using single or multiple samples.
  - Advantages: large insertions and exact copy numbers.
  - Limitation: cannot detect copy neutral events.
- RED approach using WES data: distinuous reads. So cannot use existing segmentation (merging) methods. Cannot detect small scale events. Main methods: CoNIFER and XHMM.
- De novo assembly (AS): only in high coverage regions.
  - Cortex assembler: de Bruijn graph, then find bifurcation structures in which the sequence of the sample differ from the refernce genome.
- Combination approach: e.g. use both PEM and RD. Enables detection of CNVs with exact breakpoints, and with various spanning lengths (esp. larger inserts).

- Comparison of four main approaches:
  - PEM: all types of SVs. Cannot estimate the copy number, and insertions larger than the insert size.
  - RD: copy number, good performance for large CNVs. Not applicable for detection of breakpoints, copy neutral events such as inversion or translocations, or small CNVs (e.g. ¡1kb).
  - AS: novel mutations. Perform poorly on repeat and duplicate regions. Computationally expensive.
  - SR: precise breakpoints, sensitive on deletions and small insertions. But low sensitivity in low complexity regions.

A shifting level model algorithm that identifies aberrations in array-CGH data [Magi & Torricelli, Biostatistics, 2010]

- Shifting level model: let x_i be the data at a window i. The idea is that x_i is determined by underlying rate m_i, which correlates in space. Intuition: m_i of adjacant windows will generally be close, but occasionally large changes may happen (shifting level). The model is:

$$x_i = m_i + \epsilon_i \tag{1.17}$$

  where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. For m_i, there are two models: (1) With probability $1-\eta$, m_i will be the same as m_{i-1}; with probability η, it will differ from m_{i-1} by a normal RV:

$$m_i = m_{i-1} + z_{i-1}\delta_i \tag{1.18}$$

  where z_{i-1} is Bernoulli RV with parameter η, and $\delta_i \sim N(0, \sigma_\mu^2)$. (2) With prob. $1-\eta$, m_i will be the same, but with prob. η, it will change and we will resample m_i from $N(\mu, \sigma_\mu^2)$. The paper uses version (2) SLM.

- HMM: write $\omega = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\epsilon^2)$. The likelihood can be decomposed as:

$$p(x, m, z|\theta) = p(x|m, \sigma^2, \omega), p(m|z, \mu, \sigma^2, \omega)p(z|\eta) \tag{1.19}$$

  We define state $q_i = (m_i, z_i)$, then the model can be represented as HMM. We need to discretize the mean so the number of states is finite, say K. The transition matrix is basically: with probability $1-\eta$, it will stay at the current state k; with probability η, it will jump to another state. The jumping probability from j to k (could be equal) is given by g_{jk}, which takes into account that the probability of moving to a very high or very low state is low - we do not need one parameter for each transition, instead, the transition rates are derived from the underlying distribution from SLM.

- **Remark**: the key idea is that we do not need to have many transition probabilities even if we have many states (levels of discretization) - these are all based on underlying normal distributions (SLM).

Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth (XHMM) [Fromer & Purcell, AJHG, 2013]

- Challenges: read depth in WES data affected in each step in the experiment, including target capture, PCR amplication, sequencing and read mapping. As a result, read depth can vary by order of magnitude even in diploid genome. So we cannot directly use the read depth to infer copy number change events.

- Existing methods: normalization by known confounders such as GC content, background depth. For cancer data, use paired case-control.

- Normalization by PCA: the idea is to find out the main factors influencing read depth through PCA; then adjust for them. First center the target read depth about mean (targets: exons). Construct the individual-by target read-depth matrix, and do PCA. Found that top 100 PCs correlate with a number of factors: sample batch, mean depth, target GC and combinations. However, some strong PCs do not correlate with any of these factors. For each sample, we have these latent factors, we can adjust the read depth of every target. The adjusted read depth would then be comparable across samples and targets.
- Remark: how would PCA adjust for between-target variation within a sample? Intuitively, suppose high GC content is a main factor for read depth, then across many samples, the high-GC targets tend to have high read depth. We will learn a latent variable that represent GC content (even though it does not vary much across samples). Adjusting this latent variable thus correct for the influence of GC on read depth. So in XHMM, our interest is mainly between target variation within a sample; but we use individual-by-target data to learn about these hidden variables, and then adjust for them within a sample.
- HMM: treat each target (exon) as one unit. Three state HMM, Deletion, Duplication or Diploid. The mean read depth of Diploid state (normalized, after PCA) is 0, and the mean RD of Deletion state and Duplication state are $-M$ and M respectively. The transition prob:
  - Deletion/Duplication to Diploid: p, intepretation is exome-wide CNV rate. Choose $p = 10^{-8}$.
  - Diploid to Deletion/Duplication: q, $T = 1/q$ intepreted as the average size of CNV (number of targets). Choose $T = 6$.
- HMM taking into account of distance between targets: when the two targets are very far, there is not much information from the first state on the second state. So the transition probabilities from the first to the second state, when the first is not Diploid, should be the same as transition probabilities from Diploid. Let $f = e^{-d/D}$ where d is the distance and D is the expected distance between two targets. Then the transition probabilities from Deletion state to other states are linear combination of the probabilities in the base model and of the Diploid states, weighted by $1 - f$ and f. See Table 2.
- Computational pipeline:
  - Filtering extreme targets and samples: (1) targets: extreme GC content, repeat regions, etc. (2) Samples: extremely low/high coverage.
  - PCA normalization. The data is now the residual terms, after controlling for top PCs.
  - Filtering targets with extreme variability in normalized depth.
  - HMM decoding of CNVs. The results are expressed as a set of probabilities, or Phred scores.
  - Calling CNVs in family data.

1.1.5 ChIP-Seq

Overview: [Yunlong Liu's lecture]

- Pattern of ChIP-seq signals:
  - Class I: punctate signal, two close peaks in positive and negative strands respectively. Sequence-specific TFs
  - Class II: mixture of punctate and borader signals. Ex. RNA Pol II (diffused along the transcript)
  - Class III: range from nucleosome-size domains to very broad enriched regions. Epigenetic markers

- Strand-specific structure of tag distribution: many fragements, some starting from the left-end of the TFBS, some starting from the right-end. However, short reads can only cover the 5'-end of the two types of fragments. As a result, two peaks in either side of the peak. (Imagine a very large protected region by TF binding, the center of the region will not get any sequence reads.)
- Peak identification: a set of candidate peaks.
  - To deal with the strand-specific distribution: use tag shift (move all reads half of the average fragement length) or use extension (add the read counts to the center of the peak, imagining the reads are long)
  - Sliding window to smooth the signal (tag density)
- Statistical signficance:
  - No negative control: fit the (uniform) background distribution with Poisson or NegBin. Choose the threshold according to the background.
  - With negative control: estimate the background distribution from the negative control or simply subtracting the reads from the background.
- Peak calling: based on minimum quality criteria, e.g. minimum absolute signal threshold or minimum enrichment ratio. Additional filter may be applied, e.g. tag directionality - fraction of reads in + strand vs. in - strand.

QC of ChIP-seq data [personal notes]

- Read mapping: the set of nonredundant reads that can be mapped to the genome should be high; and the rate of repeat reads should be small.
- Biological replicates: consistency of peaks, or correlation of signal intensities across replicates.
- Biological annotations/expectations: the set of known peaks should be recaptulated; evolutionary conservation; distribution of peaks (TSS, intergenic vs. promoter, etc.).
- Motif finding in the peak regions.

The Analysis of ChIP-Seq Data [Wenxiu Ma and Wing Hung Wong, Methods in Enz. 2012], Computational methodology for ChIP-seq analysis [Shin & XS Liu, Quant. Bio, 2012]

- Sequencing statistics: usu. 25-30 bp reads. Mappable reads: map to a unique location (with up to 2 mismatches allowed). In current platform, approximately half of the reads can be uniquely aligned back to the reference genome. For mouse, typically 10M reads is sufficient; for a more compact genome, one can attain higher signal intensity at the same sequencing depth.
- Saturation: The saturation point of the sequencing depth is defined as the minimum number of reads which would enable the detection of all true protein-DNA binding loci. To test if the data has reached saturation point. Randomly sample half of the reads, and compare the results with the full set of reads in terms of motif enrichment.
- Negative controls: typically just DNA input or mock IP. Important to indentify the genomic regions expected to have more reads for reasons unrelated to the TF binding, e.g. GC-rich sequences (Illumnia), TSS (open chromatin).
- Biological replicates: important for additional variability. Check the consistency of replicates (if there are multiple ones, could remove one that is very different from the rest). To combine the replicates, one way is to take the intersection. Another approach: pool all reads from the replicates, and call peaks from pooled data.

- Read mapping: Generally consider only uniquely mapped reads. For multi-mapped reads, a simple solution is random assignment. Often the first 25bp of 30bp reads for quality reason. In human genome, it is estimated that about 75% of the genome are non-redundant. The mapping results can be assessed by: (1) the percent of mappable reads, typically 1/3 to 1/2. (2) The percent of nonredundant reads (two reads that are not identical) among all mappable reads: generally should be above 50%.
- Tools for read mapping: Bowtie: very efficient. Maq: take the sequencing quality stat. into account. SeqMap: consider indels. One consideration is the sequencing platform, e.g. SOLID requires support of colorspace.
- Bimodal peaks: (1) tag-shifting: simple, but the distance may not be uniform since peak width can vary greatly. (2) Call Waston and Crick peaks separately, and define the boundary of peaks (CisGenome).
- Peak calling with no negative control: need to estimate the background rate of reads. Ex. divide the genome into windows, count the number of reads falled into each window, and fit the distribution. Negative Binomial is found to be a better fit than Poisson.
- Peak calling with negative control: simplest is to do fold change - often use 5-fold. Enough number of peaks (e.g. greater than 50%) of 20-fold difference often suggest good quality. However, the ratio approach does not take the variable background rates into account.
- The CisGenome approach: two-sample comparison
  - For each window, the number of ChIP reads k_{1i} and the number of controls reads k_{2i}. The distribution $k_{1i}|n_i \sim \text{Bin}(n_i, p_0)$ where $n_i = k_{1i} + k_{2i}$ is the total number of reads in this window. The parameter p_0 is the probability of observing a ChIP-read in this window by chance.
  - Estimating p_0: the simplest approach is: among all windows with only one read, estimate the fraction of windows with only one ChIP read (probability due to chance). However, such windows may be few. A more principled approach: let r_0 = the number of background reads (not mapped to peaks) in ChIP / the number of background reads in control, then $p_0 = r_0/(1 + r_0)$. Estimate r_0 iteratively: first use all reads, then remove peaks, reestimate r_0 and call peaks; and so on.
- FDR estimation: once the peaks have been identified, expected number of windows passing the threshold divided by the observed number. The expected number can be computed from permutation (e.g. switch ChIP and control status) or random sampling.
- De novo motif analysis: in the top regions, e.g. FDR 0.01 or top 1000 to avoid noises from weaker sites. The results can be assessed using motif enrichment in peak regions: (1) the number of motif matches in peak regions vs. the number of matches in random control regions. (2) The distance of motif matches to the peak centers.
- Use of other prior information for quality assessment of ChIP-seq data: (1) The conservation scores of motif matches (or peaks) vs. flanking regions. (2) Genomic distribution of peaks: e.g. distance to TSS.
- Differential peak calling in different conditions: (1) Simple methods: run peak calling in separate conditions followed by intersection analysis to identify unique peaks. However, it may miss peaks barely above and below the peak calling cut-off. (2) More specialized algorithms: using HMM, MACS2, etc.
- Tools for ChIP-seq analysis: Integrated software: Cistrome, CisGenome. A few popular peak callers: MACS, Sissr, SPP, USeq, PeakSeq

Model-based Analysis of ChIP-Seq (MACS) [Zhang & Liu, GB, 2008]; Identifying ChIP-seq enrichment using MACS [Feng & Liu, Nature Protocol, 2012]

- Shift-size determination: heuristically define peak regions (e.g. k-fold enrichment in sliding windows), then identify the distance between Watson and Crick peaks. Compute the average over 1000 such regions: shift size. Ex. for FOXA1 data, shift size is estimated to be 126bp. MACS will shift the tag by $d/2$.

- Peak detection: using Poisson distribution on a test region (of size $2d$). MACS compares the read count with the expected rate λ_{local}. The main idea is that: because of local chromatin structure, CNV, etc. the read counts could be biased, so we should use the local rate to test the enrichment instead of the global rate. When the control is available, the rate in the region in the control data should be used as well.

  - Without control: $\lambda_{\text{local}} = \max(\lambda_{BG}, \lambda_{5k}, \lambda_{10k})$, where the rates are the genome-wide rate, the rates in the nearby 5k and 10 regions, respectively.
  - With control: $\lambda_{\text{local}} = \max(\lambda_{BG}, \lambda_{\text{region}}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$, where λ_{region} is the rate in the control data of the test region. For other rates (1k, 5k, 10k), the rates are also calculated using the control data.

  Accounting for read depth difference between ChIP and control: linear scaling down the number of reads from the larger sample.

- FDR control: MACS calls peaks at a given p-value threshold, e.g. 10^{-5}. To obtain FDR, do permutation of ChIP-control labels. The empirical FDR is the number of reads passing the threshold in control divided by the number in ChIP.

PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls [Rozowsky & Gerstein, NBT, 2009]

- Confounding factors: the density of mappable bases in a region (not all regions are equally mappable), the underlying chromatin structure (open chromatin tends to generate more reads) Mappability of mammalian genome: even though only 47.5% of the genome is nonrepetitive, 79.6% of the genome is uniquely mappable using 30 bp tags. The fraction increases to 89.3% as the length of the sequenced tag is increased to 70 nt.

- Mappability map: for each region, compute the fraction of sequences that can be uniquely mapped. Slightly enriched near TSS (i.e. these regions tend to be more mappable). Analysis of control ChIP-seq data: not random. Peaks in the input-DNA signal (control) exhibits distinctive enrichments of signal proximal to TSSs. And this effect is stronger than the mappability issue.

- PeakSeq step 1: creating signal map. Use only uniquely mappable reads, tag shifting by 200bp. The signal map is then the integer count of the number of overlapping DNA fragments at each nucleotide position.

- First pass: identify regions or peaks in the ChIP-seq data that are substantially enriched compared to a simple null background model. The analysis is based on the 1M segments (not really global). Within each segment we use the mappability map to correct for the variation in mappability between segments.

- Normalizing the control data to ChIP data: use linear regression of the tag counts in windows.

- Second pass: use binomial distribution to determine the p-values of potential BSs from the first pass. Then do FDR correction using BH procedure.

Evaluation of algorithm performance in ChIP-seq peak detection. [Wilbanks et al. PLoS ONE, 2010]

- Background: with paired-end reads, the fragment length can actually be measured directly allowing more precise determination of binding sites. However, few software supports this option (?).

- Difference in different software:
  - Dealing with bimodality in tag density: either by shifting tags in the 3' direction, or by extending tags to the estimated length of the original fragments.
  - Peaks: some minimum height criteria at which enrichment is considered significant, and some minimum spacing at which adjoining windows, clusters or local maxima (G-KDE) are merged into a single peak region.
  - Several methods leverage the bimodal pattern in the strand-specific tag densities to identify protein binding sites, either as their main scoring method ("directional scoring methods") or in an optional post-processing filtering step.
- Evaluation strategies: 3 ChIP-seq datasets, all TFs have well-defined motifs. (1) Sensitivity: how often the known binding sites are recovered. (2) Specificity: how often the predicated peaks contain the canonical motifs.
- Conclusion: remarkably similar performance with regards to sensitivity and specificity. The programs differed most significantly in the spatial resolution of their estimates for the precise binding region. The best estimates of precise binding location were provided by Spp (directional scoring), followed by MACS.
- Implementation comparison: spp is run as a package from within the statistical program R. CisGenome: GUI (Windows only), an integrated platform for ChIP-chip and ChIP-seq analysis, combined with downstream motif finding and an integrated genome browser.

GREAT improves functional interpretation of cis-regulatory regions [NBT, 2010]

- Problem: given a set of regions, how to test its GO enrichment? The region to gene mapping is ambiguous.
- Idea: suppose among N regions, 100 of them are close to genes involved in a process G, then we estimate the expected number by: randomly choosing 100 regions, how many of them will be close to genes in G?
- Method:
  - For each gene, define its regulatory domain. The domain could overlap for different genes.
  - For a given GO category G, obtain the fraction of genome that are annotated with G (i.e. the union of all regulatory domains of all G genes). Call it π.
  - Binomial test: a total of N regions, K of them are close to the genes of G, then binom.test(K, N, π).

Architecture of the human regulatory network derived from ENCODE data [Gerstein et al, Nature, 2012]

- Peak calling: use two independent callers, SPP and PeakSeq. For most high-quality datasets, there was a high degree of overlap ($> 80\%$) between the peak sets from the two peak callers.
- For biological replicates, use a measure of consistency of peak calling results between replicates - irreproducible discovery rate (IDR). The IDR score of a peak represents the expected probability that the peak belongs to the noise component, and is based on its ranks in the two replicates.

DBChIP: Detecting differential binding of transcription factors with ChIP-seq [Liang & Keles, Bioinfo, 2012]

- Detecting consensus sites from multiple replicates: use the peaks called by other programs.

- Normalization by library size for ChIP samples: use the library size itself is problematic, because the true binding sites can contribute substantially to the total library size. Use median ratio strategy: the ratio between any two samples is the median of the ratios between binding site counts and is robust to the fluctuations of the high affinity differential sites.

- Normalization of control samples (in comparison of ChIP sample): for the i-th sample, and site j, calculate $y_{ij} = x_{ij} - f_i z_{ij}$, where f_i is the normalization factor between the i-th ChIP sample and its matching control sample and z_{ij} is the corresponding read count in control. The normalization factor is computed by NCIS (Liang and Keles, BMC Bioinfo, 2012).

- Detecting differential binding: use Negative Binomial regression to test difference. The dispersion parameter is estimated using edgeR: the default is to use a common dispersion parameter, but can also estimate dispersion parameter for each site.

diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates [Shen & Nestler, PLoS ONE, 2013]

- Proprocessing: remove regions with very low read counts (e.g. repeats) and very high counts (PCR problem); otherwise, the estimation of background rate may be distorted.

- Normalization: calculating a numeric factor for each sample so that each raw read count can be linearly scaled using its corresponding factor. Two steps: the factor for each window of each sample (comparison with median of this window across all samples); then the median of all factors over all windows of a sample.

- Test of differential binding: exact NB test of DESeq.

- Peaking calling and FDR: call significant window using p-value, then merge overlapping windows. FDR control is done by BH.

- Questions:

  - p-value of the peaks, which may contain overlapping windows?

Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm [Zhang & Wang, GR, 2013]

- Motivation: MeDIP uses IP of methylated DNA (CpG), so the signal is proportional to the level of methylated CpG; MRE uses restriction enzyme sensitive of methylated CpG, and the signal is proportional to the level of unmethylated CpG. The two technologies provide complementary measurements of methylation and can be used together to improve detection of differential methylation.

- Model overview:let x_1 and x_2 be the reads mapped to a region in two samples using MeDIP, and y_1 and y_2 be those using MRE (normalized counts - assuming they are integers). Let μ_1, μ_2 be the fraction of methylated CpG in two samples, and m the number of CpGs in the region. Then we have x_1 follows Poisson distribution with rate:

$$\mathrm{E}(x_1) = L_1 m \mu_1 / \sum_i m_i \mu_{1i} \tag{1.20}$$

  where L_1 is the library size, and the denominator is the total number of methylated CpGs in the genome. Similarly we have $\mathrm{E}(x_2)$. For MRE data:

$$\mathrm{E}(y_1) = L_1 m (1-\mu_1) / \sum_i m_i (1-\mu_{1i}) \tag{1.21}$$

  Our test is $H_0 : \mu_1 = \mu_2$.

- Reduced statistical problem: this is similar to a two-sample Poisson test. Suppose we observe x_1 and x_2 events in interval T_1 and T_2 respectively, with the rates:

$$\mathrm{E}(x_1) = \mu_1 T_1 \qquad \mathrm{E}(x_2) = \mu_2 T_2 \tag{1.22}$$

  And for a different type of event, we observe y_1 and y_2 events:

$$\mathrm{E}(y_1) = (1-\mu_1) T_1 \qquad \mathrm{E}(y_2) = (1-\mu_2) T_2 \tag{1.23}$$

  To form a test statistic, the idea is if $\mu_1 > \mu_2$, then $x_1 > x_2$ (scaled to interval size) and $y_2 > y_1$. So we have $x_1 y_2 > x_2 y_1$, and the differnce reflects the difference of μ_1 and μ_2. To "standardrized" the distribution of $x_1 y_2 - x_2 y_1$, we consider the conditional distribution:

$$P(x_1 y_2 - x_2 y_1 | x_1 + x_2 + y_1 + y_2 = n) \tag{1.24}$$

  Conditioned on the sum of events, x_1, x_2, y_1, y_2 follow multinomial distribution and this allows us to derive the distribution above.

PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data [Zhang & Sartor, Bioinfo, 2014]

- Background: alternative strategies for dealing with replicates in ChIP-seq data
  - Separate analysis (SA): on each pair and then stipulate rules to combine the peak-finding results, such as requiring the peaks to be found in all pairwise comparisons. Problem: may not have natural pairing (the results may be sensitive to pairing); may miss moderate peaks (not occur in each pairwise comparison).
  - Combining replicates (CR): combine the replicates (CR) in each group and run one-ChIP-versus-one-control analysis to identify all possible peaks. Problem: false positives may occur where binding is present in only one or a subset of the samples.
  - Irreproducible discovery rate (IDR)
  - edgeR or DEseq: no necessary processing steps for ChIP-seq data and not take advantage of local chromatin information.
  - diffReps: exact NB test.
  - R packages: DiffBind and DBChIP, rely on other peak callers to generate peak sets for each individual sample first and conduct analysis on the candidate regions
- Motivation: analysis of either a group of ChIP-Seq samples together with controls or compare two groups of ChIP-Seq samples, with or without controls. Favor regions with consistent read counts across replicates.
- PePr preprocessing:
  - Remove duplicate reads
  - Fragment size (shift) estimation
- Normalization:
  - Control samples: divide the genome into 1000bp bins. The mean of all ChIP libraries is used as the reference sample, towards which every sample will be normalized. The normalization factor is the ratio of the total number of reads in the background bins in reference over in an input sample. The number of reads in each window for the input sample is multiplied by its normalization factor.
  - ChIP samples: normalize for different immunoprecipitation efficiencies.

- Testing differential binding: use NB model for normalized counts.
  - Estimate overdispersion parameter: use the log-likelihood for a window (similar to edgeR), but also borrow information for nearby W windows using the triangular weight.
  - Statistic test: Wald test with log-transformation of γ (the enrichment).
  - FDR correction: BH procedure.
- Defining peaks: The significant windows that are localized in the same genomic area are merged.

1.2 Functional Genomics

1.2.1 Genome Editing

CRISPR-Cas systems for editing, regulating and targeting genomes [Sander & Joung, NBT, 2014]

- Double-strand break (DSB) of DNA: usually repaired in cells by either non-homologous end joining (NHEJ) or homology-directed repair (HDR). NHEJ will introduce indels and HDR can introduce point mutations or indels based on DNA templates.
- Naturally occurring CRISPR-Cas9 system:
  - Bacterial has CRISPR array in the genome. When virus randomly insert DNA, the invading DNA may be incorporated into the CRISPR array by chance. The protospacer regions contain these foreign DNA.
  - The CRISPR array gets transcribed, creating crRNA (CRISPR RNA). The second RNA, trans-activating CRISPR RNA (trancRNA), forms hybrid with crRNA.
  - Cas9 protein associates with crRNA:trancRNA complex, which can then recognize foreign DNA, adjacent to PAM sequence. PAM is commonly NGG, but alternative PAMs exist. PAMs typically not found in bacterial.

  This is bacterial immune system: the role of crRNA:trancRNA is similar to antibody.
- Engineered CRISPR-Cas9 system: fused crRNA:trancRNA complex, called guide RNA (gRNA). The length of target sites: 20nt.

CRISPR: gene editing is just the beginning [Nature, 2015]

- Defective cas9 (dCas9): tethered to another protein, which turn on or off genes: dCas9 could block RNA polymerase binding (CRISPRi), or linked to an activator protein (CRISPRa).
- CRISPR epigenetics: tether dCas9 with histone modification enzymes to modify epigenomic marks.
- CRISPR for short sequences (e.g. ncRNA): The Cas9 enzyme will cut where the guide RNA tells it to, but only if a specific but common DNA sequence is present near the cut site. Cpf1 may have more sequence options.
- Combining CRISPR with a chemical or light-triggered switch.
- Targeting RNA [Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell, 2016]: gRNA to target RNA molecule, fused with GFP. This can track RNA movement in real time in live cells.

Beyond CRISPR: A guide to the many other ways to edit a genome [Nature, 2016]

- Use of CRISPR-cas9 for treating genetic disease: the challenge is that the system may be too large to fit in a virus. Use mini-cas9.

- Expand the rich: need a specific neighboring sequence. Cpf1 has different sequences and more specific.
- True editors: "But burning a page of the book is not editing the book" - Church. The effiency of editing is low. New system: disabled Cas9 and tethered to it an enzyme that converts one DNA letter to another.
- NgAgo: DNA-based, avoid the use of neighboring sequence.

CRISPRi experiments

- Introducing DNA: need to introduce two main pieces, sgRNA and dCas9 (or dCas9 fusion protein). Co-transfection, or transduction (may need sorting).
- Testing the effects on main target: cells expressing reporter construct (promoter and GFP), then use sgRNA to target the test sequence or GFP.
- Testing off-target effect: typically use RNA-seq to determine expression of all the genes.

Biology of CRISPRi: what determines the efficiency

- sgRNA length
- sgRNA target sequence: the distance to TSS, the strand, TF binding sites (study of SV40 promoter), and local chromatin structure.
- Effector domain

Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. [Qi & Lim, Cell. 2013]

- Repression of transcription by dCas9:sgRNA complex (coexpression of dCas9 and sgRNA) in E. coli:
  - sgRNAs targeting the nontemplate DNA strand: 10- to 300-fold of repression. Those targeting the template strand showed little effect.
  - Targeting of the sgRNA to the 35 box: 100 fold repression.
  - Targeting sequences about 100 bp upstream of the promoter showed no effects.
- Mechanism: dCas9 function as an RNA-guided DNA-binding complex that could block RNA polymerase (RNAP) binding during transcription elongation.
- Factors that Determine CRISPRi Silencing Efficiency:
  - Repression was inversely correlated with the target distance from TSS
  - sgRNA: requires 20bp or so. Less than 12 bp, no effect.
- Effect of dCas9:sgRNA on HEK293 cell line: the effect is modest and more dependent on the target locus, factors such as the distance from TSS and the local chromatin state may be critical parameters in determining repression efficiency.

CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes [Gilbert & Qi, Cell, 2013]

- Motivation: efficiency of CRISPRi in human cells is low (2-fold), test linking dCas9 with effector domains to repress gene expression.
- CRISPRi with effector domain to silence or active gene expression: use HEK293 cells expressing reporter (GFP or Gal4). The outcome is GFP level, measured by flow cytometry. Try two strategies of introducing CRISPRi:

  - Co-transfection of sgRNA targeting GFP, and dCas9-KRAB fusion protein (KRAB has the highest repression efficiency in the experiment).
  - Lentiviral construct containing dCas9-KRAB; then flow cytometry to sort the subpopulation of cells expressing the lentiviral construct. Then transfection or use another lentiviral construct to introduce sgRNA. Results: 6 out of 9 sgRNAs targeting GFP knocked down GFP expression by at least 75% with a 15-fold repression for the best sgRNA.

- CRISPRi-mediated gene knockdown is highly specific to the target gene: using RNA-seq to show that no gene other than GFP changes by more than 1.5 fold.

- CRISPRi as a tool for mapping and perturbing regulatory elements:
  - sgRNA targeting promoters: using SV40 early promoter of our GFP reporter. dCas9 has no effect on transcription. dCas9-KRAB efficiently repressed GFP expression. The effect depends on the position of sgRNA: sgRNA blocking AP-1 site has a stronger effect than those blocking SP-1 site.
  - sgRNA to block TF binding: if we design sgRNA to target the direct binding site of transcription activator, strongly abolish gene expression.

- Lessons:
  - Directly targeting coding sequences: does not always repress gene expression, 6 out of 9 sgRNA works in the lentivirus experiment on GFP-expressed HEK293 cells.
  - Targeting promoter sequences: the effect depends on the location of sgRNA. It seems that directly targeting activator binding site has a strong effect, even without KRAB.
  - sgRNA is a limiting factor for CRISPRi function: the efficiency may be affected by sgRNA stability, loading into dCas9, or binding to DNA. Alternately, dCas9 binding may be strongly dictated by the local chromatin structure of the target sequence.

Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation [Gilbert & Weissman, Cell, 2014]

- Background/motivations:
  - RNAi has pervasive problems with off-target effects.
  - CRISPR/Cas9: focus on LoF studies involving frameshift disruptions, limiting utility for the study of essential genes. Additionally, ds-DNA break can be toxic. Also indels formed from error-prone DNA repair are often short and in-frame.

- Goal: establish CRISPRi/CRISPRa as a tool for whole-genome screen, e.g. for cellular growth or susceptibility to chemicals.

- Challenge and idea: suppose we are targetting many DNA sequences at a time (including multiple sgRNA for one gene) in a native cell, and we want to determine the effects of many these perturbations in a mix of cells. The challenge is that we cannot barcode the genes (which gene is targeted by which sgRNA). So the idea is to measure cell numbers, and use sgRNA as barcode for cells.

- High Throughput screens using CRISPRi:
  - Library of sgRNAs targetting 49 ricin-resistance genes (1,000 sgRNA per gene, in 10kb window near TSS). The ricin-resistance phenotype is roughly linear to the expression level of these genes. Thus measure sgRNA frequency in the cell population to indirectly measure transcriptional repression.
  - Experiment: lentivrial sgRNA library, infection of cells expressing dCas9 fusion protein, then do growth analysis to determine cell numbers expressing each sgRNA.

  - Results: strongest repression for most genes at -50 to +300 bp around TSS. Neither DNA strand nor GC content matters. The effect is significantly stronger than shRNA.
  - High specificity: the results are very sensitive to mismatch between sgRNA and target DNA.

- Define phenotypic scores of a sgRNA: suppose the number of cells (sgRNA frequency) at $t = 0$ is C_0, and the number of cells at t is C, then we have this relation: $C = C_0 2^{\gamma t}$, where γ is the rate. So we have: $\gamma = \log_2 C/C_0 = \log_2 e$. In the experiment, we also normalize γ, by obtaining Z scores: subtracting γ_0 be the rate of the control (wild type or untreated), and divide by standard deviation obtained from using multiple control sgRNA.

- CRISPRa: all genes show phenotypes in the CRISPRi screen, but only a subset show phenotype in CRISPRa screen.

- Genome-scale CRISPRi screen platform: library of 10 sgRNA/gene, targeting all the protein-coding genes in human genome. Phenotype is cellular growth. Individual sgRNA can have profound effect on cellular growth (up to 256 fold depletion). The top genes are involved in essential cellular functions, such as transcription, translation and DNA replication.

- CRISPRi is reversible and inducible: dCas9-KRAB under the control of an inducible promoter. dCas9-KRAB does not create a permanently repressive chromatin state.

Genome-scale CRISPR-Cas9 knockout screening in human cells [Shalem & Zhang, Science, 2014]

- CRISPR library: construct containing sgRNA, Cas9, delivered by lentiviral vector. On average 3-4 sgRNAs per gene.

- Efficiency of CRISPR vs. shRNA: use GFP to show that CRISPR is more efficient than RNAi (low fluroescence level).

- Application to drug resistance study: the top genes are enriched with known targets, and better scoring consistency among sgRNAs.

- RIGER algorithm: designed for RNAi. A gene may be targeted by multiple RNAi. We have a ranked list of all RNAi's, and use GSEA analysis on the RNAi of the same gene to score that gene.

More specific CRISPR editing [NM, 2016]

- Determine off-target binding of Cas9: use disabled Cas9, then ChIP-seq to determine the locations of binding. For most gRNAs, dCas9 bound at several (tens to hundreds) of off-target sites. Catalytically active Cas9 cleaved some, although not all, of the off-target binding sites

- Improving the specificity: the requirement for two gRNAs, for precise orientation of the target sites, and for correct spacing between half-sites. Dimeric Fok1-dCas9 fusions.

Mapping regulatory elements [NBT, 2016]: CRISPR screening for CRE activities

- p53 and ER sensitive CREs: changes of cell proliferation (growth screening: both positive and negative).

- Test effect of CRE mutations on its activity by using GFP as reporter. Library construction: avoid using plasmids, instead first put dummy gRNA to prepositioned sites in the genome and then replacing the dummy gRNA with a pooled library of gRNAs through CRISPR-Cas9based homologous recombination.

Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens [Dixit and Regev, Cell, 2016]

- Key idea: introduce a mRNA that contains the guide RNA barcode, then this mRNA will be added a cell barcode. By sequencing this mRNA, we can know which gRNA targets which cell.
- Application: BMDCs with LPS. Target TFs and cell cycle regulators (about 50 sgRNAs).
- Linear model: log-read count vs. design matrix (knockout) and covariates.
- Technical Covariates: number of transcripts in a cell. Probabiility that perturbation was successful (similar to EM) - on average 66% success rate. Biological covaraites: cell type, cell states, e.g. cell cycle.
- PVE: guide RNA explains very little. Largest factor: cell states.
- Learn other effects: fitness effects, cell size effects, cell states effects, MOI sensitivity. Do simple Wilcoxin-rank sum test.
- Validation: correlation of beta of guides. Much higher of guides within genes.
- Results: Cell state effects of TFs.
- Genetic interactions: analysis of cells with multiple guides. Types of interactions: buffering, synergistic, dominant, additive.
- Lessons: (1) modeling uncertainty of CRISPR. (2) Should also correct for covariates. Ex. a guide may change fitness, and hence cell states.

A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response [Adamson and Weissman, Cell, 2016]

- Experiment design: first choose candidates from the genomewide CRISPRi screen. Then perturb-seq on these genes: deletions and scRNA-seq, both single and combinatorial.
- UPR epistatisis experiment: deletion (single, double or triple) of three UPR sensor genes, with three different treatment (inducing each branch). Then profile expression. Find the transcriptional program influenced by each branch: (1) PERK/ATF4 had the largest regulon with many targets uniquely under its control. (2) ATF6 and IRE1a showed more overlap, consistent with a more common transcriptional regulatory mechanism. ATF6 has stronger activating effects on its targets. (3) Many genes showed some sensitivity to all branches, particularly a group of very high-abundance stress-response genes.
- CRISPRi screens of ER homeostatis: output is UPRE-GFP. Hit include known/expected ER genes. Unexpected genes (not seen in yeast screen): transcription, translation, mitochondrial. Some of the hits may affect reporter system, rather than UPR.
- UPR Perturb-seq experiment: 91 sgRNAs targeting 82 strongest hits from CRIPSRi screen. Many hits repress all three branches. Use the gene expression profile across perturbations to cluster functionally related genes.

Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells (Mosaic-seq) [Xie and Hou, Mol Cell, 2017]

- Background: Drop-seq: beads containing polyT and, cell barcode, UMI and PCR handle.
- Method (Figure 1C): in the lentiviral construct, a random 12bp linked to gRNA as barcode. This random barcode is expressed in Drop-seq with cell barcode.
- Validation experiment: HBG1/2 (human beta-globin) expression, targeting enhancers. Show that targeting SH2 enhancers significantly reduce expression of genes.

- Analysis methods: (1) normalization of read counts of a gene: read count in a cell divided by median read count (or total) across all cells. (2) DE analysis: at a given normalized expression threshold, number of cells below the threshold in gRNA vs. control by hypergeometric test.
- Data: sgRNA library targeting 71 enhancers.
- PIM1 expression: only 1 enhancer has effect out of 12 in the same TAD. Similar in HBG1/2. Generally one major enhancer per gene?
- Active enhancers: enriched with p300 and RNAP2, less enriched with K4me1 and K27ac.
- Estimate penetrance (%cells) and repression (0 to 100%) of each enhancer. For every gene, obtain the distribution of the two metrics for all enhancers.
- Combinatorial enhancers: pairwise combinations, show many examples of significant reduction, e.g. PIM1.
- Remark: limits of the method: KRAB a universal repressor?
- Lesson: only a small percent of enhancers may actually be transcriptionally active.

Pooled CRISPR screening with single-cell transcriptome readout (CROP-seq) [Datlinger and Bock, NM, 2017]

- Method: normally gRNA expression is driven by RNA Pol 3, and does not have polyA. Create a lentiviral construct with gRNA s.t. it is expressed by RNA Pol 2. Also has antibiotic resistance gene for selection of successful transfection.
- Evaluation of the method: (1) Efficiency of genome editing: $> 90\%$. (2) Assigning gRNAs to cells: dominant gRNA more than 3 times frequent than the other gRNAs combined. The majority of cells can be assigned to unique gRNAs. Only 2.7% cells have more than one gRNA (the difference is less than 3).
- Application to TCR induction: 20 nontargeting gRNAs (negative control), 9 for essential genes (positive control) and 23 TFs. 80 cells per gene, and 1300 cells with non-targeting gRNAs as control.
- Quality control: (1) Depletion of cells with gRNAs of essential genes (Figure 2B): positive control. (2) Similarity of gRNAs targeting the same genes: transcriptome similarity. Overall, transcriptome targeting TFs and negative controls are distinguishable.
- Effect of gRNAs: (1) Define T-cell induction gene signature: PCA on scRNA-seq data, find genes that separate naive and induced T cells. (2) Heat map of gRNAs vs. expression of signature genes reflecting TCR activities ((Figure 2DE).

On the design of CRISPR-based single cell molecular screens [Hill and Trapnell, NM, 2018]

- Background: gRNA expression requires Pol 3, and it does not have poly-A tail.
- Perturb-seq type of design (Fig. 1a): vector has U6 promoter (POL 3) and sgRNA; and Pol 2 promoter followed by trans-gene, UTR and barcode (mutation BC). The BC and gRNA is 2.4kb away.
- CROP-seq design (Fig. 1b): overlapping transcripts, both pol2 and pol3, so no need of mutation BC, since gRNA is expressed.
- Template switching in Perturb-seq design: increases with large distance between gRNAs and BC (2.4kb), about 50%. Show that in arrayed pertubration (one gRNA per pool) vs. pooled perturbation.
- CROP-seq limitation: may not detect gRNA in scRNA-seq. Increase from 50% to 94% if add target amplification step.

- Experiment: target 32 Tumor suppressor genes and 6 non-targeting controls in breast cancer cell line. Treatment of DNA damaging agent and control. About 6000 cells.
- TP53 effects: (1) Cluster distribution is different: clear from t-SNE plot, a small cluster consisting mostly of TP53-gRNA cells. (2) DEGs of TP53 vs. NTC in two conditions: about 4K and 2K genes respectively.

Exploring genetic interaction manifolds constructed from rich single-cell phenotypes [Science, 2019]

- 112 genes: known to have growth phenotypes in K562 cells. Do combinatorial perturbation, and use growth as readout. Find pairs that have interactions in fitness effects.
- A subset of interacting pairs: do Perturb-seq. 287 perturbations measured across 110,000 single cells. Results shown in manifold: e.g. interactions have the same phenotypic outcomes, may have different transcriptomic outcomes.
- Remark: genetic heterogeneity, same phenotypic changes can be achieved by different transcriptome level changes.

Model-based understanding of single-cell CRISPR screening (MUSIC) [Duan and Qi Liu, NC, 2019]

- Pre-processing: imputation of expression using SAVER. Filtering by cells (overall expression of cells are similar to controls) and by perturbation (min. number of cells, use 30). Note: this works for strong perturbations, but not for weaker ones.
- Running LDA: normalization of expression, then rounding to integers. LDA R package.
- Association of perturbation (PE) with topics across cells: normalization of topic proportion in each cell, and then do t-test.
- Quantify overall effect of PEs, and correlation of PE effects. Association test: use t-test between PE cells vs. control cells.
- Data: 14 published datasets from multiple technologies.
- Application to CROP-seq data of cancer cells: Fig. 3, 4 topics, genes with similar functions perturb the same topics. Also high correlation among similar PEs.
- Evaluation of pre-processing: filtering of 6-10% of cells by QC. Filtering of about 40% by sgRNA efficiency. Filtering of zero-expressed genes do not significantly change the results. Note: estimate that 20-30% cells with detected gRNAs have no phenotypic effects, so add a filter by whether DEGs are similar in the cells vs. control cells - filtered if more similar to control cells.
- Choice of negative control: negative vs. blank (no gRNAs detected) control, results are similar.

High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals convergent mechanisms altering neuronal differentiation [Lalli and Mitra, GR, 2020]

- LUHMES NPC line: derived from mid-brain, cultured for 5 years without chrom. abnormality. Characterizing differentiation: in vitro differentiation into dopaminergic neurons. Comparison with BrainSpan fetal cortex: Pearsons r = 0.69.
- Remark: different neuron subtypes probably have reasonably high correlation, so r = 0.69 does not mean that the system captures other neuronal subtypes.
- Candidate genes: from SFARI, choose 14, highly expressed during LUHMES differentiation.
- Validation of gRNAs and CRISPRi vectors: 3 gRNAs per gene. Show that 11/12 genes repress gene expression by >50% using RT-PCR.

- CROP-seq experiment: 14 target genes, 5 non-targeting control gRNAs, low MOI. Differentiation for 7 days, then do scRNA-seq. Total of 14K cells. 11K cells express gRNA, and filter for 8K cells with only a single gRNA. Number cells per target: 900 for NTC, and 200-1000 for target genes.Confirm the repression of all target genes.
- Psuedo-time (PT) analysis of all single cells: follow continuous trajectory (Fig. 3A). Confirm expression of marker genes: proliferation genes go down quickly; other neuronal marker genes steadily increase expression. Note: PT analysis can capture more subtle changes vs. clustering.
- Effects of perturbation on trajectories: 6/14 genes, cell compositions (among different time points) have changed (Fig. 3D). 4 genes show delayed differenitation and 2 show accelerated differentiation (Fig. 3E).
- DEG analysis: total of 800 genes show DEG across all genes (Table S1). DEG analysis stratified by stages: early or late (Table S2, S3). Note: S2 and S3 have 3000 and 2000 DEGs across all genes, but its not clear what FDR cutoff is used (likely 100-200 DEGs per perturbation after cutoff).
- Early stage: cluster cells by perturbations, found that several genes, when perturbed, show similar transcriptome profiles. Also the DEGs have significant overlap (around 20%?). Recurrently dysregulated genes: enriched in cell cycle.
- Late stage: also see recurrently dsyregulated genes, enriched in neuron maturation processes, e.g. neuron projection.
- Lesson: (1) The effects of genetic perturbation can be studied by change of composition of cell types/states/PT. (2) DEG can be made more powerful by controlling for PT/stages. Note: this is helpful even if perturbation itself does not change cellular stages. Since cellular states (and cell cycle) are main determinants of gene expression, so regressing them out would improve the power, if the effects are not genetic.

1.2.2 Optogenetics

Optogenetics: the age of light [Hausser, NM, 2014]

- Different uses of optogenetic probes: (1) Readout side: imaging synaptic release, intracellular calcium (a proxy for neural activity) and membrane voltage. (2) Manipulation side: both activators and inhibitors. Optogenetic inhibitor is particularly significant, as it creates loss-of-function experiment in neural circuits.
- Key advances in optogenetic probes: molecular tinkering of probes, e.g. inhibitory probes deeper in the brain, probes record better the calcium release.
- Challenges: (1) the level of stimulation risks driving neuronal responses outside the physiological range. (2) light stimulation and optogene expression are not uniform across the target neuron population. (3) Cannot selectively activate subtypes within that population,

1.3 Predicting Protein and Variant Function

Consequence of coding mutations: many mechanisms that could lead to change of protein function and activity [personal notes]

- Enzymatic activities: mutations in active sites.
- Protein-DNA interactions: for TFs and chromatin genes.

- Protein cellular localization: often important (e.g. for membrane proteins), and disruption of signal peptide can lead to change of localization.
- Protein stability.
- Protein-protein interactions.

Protein function prediction: towards integration of similarity metrics [COSB, 2011]

- Network-based method for function prediction: most importantly, guilt-by-association.
- Constructing networks: the edges indicate functional similarity or associations between genes. The data sources: yeast-two-hybrid; co-expression; conserved genomic neighborhood; phylogenetic co-occurrence and literature co-occurrence.
- Local network methods: consider nearest neighbors and the functions of a node are predicted from its annotated direct neighbors.
- Non-local methods: detection of modules in the network. However, not all functionally coherent groups of proteins can be represented through modules. For example, transmembrane receptors bind to many extra- and intra-cellular molecular partners but they much less frequently form complexes with other membrane proteins.
- Global methods: optimize annotations by optimize some cost function, which reflects the topology of the graph and yields a distribution of numerical labels (discrete or continuous, positive and negative) indicating functional memberships.

Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes [Petrovski & Goldstein, PLG, 2013]

- RVIS: let X be the number of variants in a gene, including syn. ones (measure of mutational burden) and Y be the number of common NS. variants (tolerance), do regression of $Y \sim X$, and residua be the RVIS. Lower values mean constrained genes.
- Comparison with cross-species comparison: very weak correlation with dN/dS from human-chimp, $r = 0.11$. Possible explanations: selection across species is about fitness consequence, eg. speed of running, sexual attractiveness.
- Enrichment of DNMs in ID and EE, but little in ASD missense mutations.
- 2D plot of gene intolerance and variant deleteriousness: most important variants are those in the corner.
- Remark:
  - RVIS is not normalized by length, as for larger genes, the variance of the residual should be higher.
  - RVIS vs. Samocha: negative selection affects both number and frequency of variants (more senstive). Samocha does not use information on frequency, so RVIS might have some advantage here.

1.3.1 Annotation of Coding Variants

Inferring causality and functional significance of human coding DNA variants [Sunyaev, Human molecular genetics, 2012]

- Statistical analysis of DAF:

  - Presence in healthy controls at appreciable frequency: support that allelic variant is likely a benign polymorphism segregating in the population (the variant is probably not involved in the disease phenotype with high penetrance - useful for Mendelian diseases).
  - Differences in global and even local ancestry may complicate conclusions because many rare variants are specific to individual human populations.

PolyPhen-2: A method and server for predicting damaging missense mutations [Adzhubei & Sunyaev, Nature Methods, 2010]

- Goal: given a protein sequence (wildtype), and a mutation, predict how likely the mutation leads to damaging effect on the protein/organism.
- Data: (1) HumDiv data: collect 3,155 mutations from the UniProtKB database causing Mendelian diseases. The set of non-damaging mutations were collected from close homologous sequences in chimp. (2) HumVar data: 13,032 human disease-causing mutations from UniProt and 8,946 human (nsSNPs) without annotated involvement in disease. Note: in website, said, HumDiv good for complex diseases and HumVar for Mendelian.
- Method: use the Navie Bayes method (other classification methods were tested, but did not outperform). The continouse features are analyzed with entropy-based discretization. The main features are:
  - Eight sequence-based features: PISC score - whether the AA matches the profile using the AA substitution matrix (BLOSUM), the sequence identity to the closest homolog, CpG context, etc.
  - Three structure-based features: the change of hydrophobicity surface, etc.
- PolyPhen-2 outperforms both PolyPhen and SIFT in both datasets. Between the two datasets, lower accuracy on HumVar data, probably because the nsSNPs assumed to be nondamaging in the HumVar dataset included a sizable fraction of mildly deleterious alleles.
- Choice of classifer model: for Medelian diseases, recommend to use HumVar trained model because one need to distinguish highly damaging mutations to mildly damaging ones. For complex phenotypes, recommend to use HumDiv trained model (distinguish damaging to non-damaging variants).
- Categories of mutations: probably damaging if the probabilistic score is above 0.85, corresponding to FP rate of 10% on HumDiv data and 19% on HumVar data. Possibly damaging if the score is above 0.15, corresponding to FP rate of 18% on HumDiv data and 40% on HumVar data.

A Combined Functional Annotation Score for Non-Synonymous Variants [Lopes & Zeggini, Hum Hered, 2012]

- PolyPhen-2: report naive Bayes posterior probability and the categories based on cutoffs.
- SIFT scores: at any position, the frequency of the variant vs. the frequency of consensus/wild-type. The variant is considered deleterious if the ratio (SIFT score) is below 0.05. The SIFT score can be scaled (complement) to [0,1]: scores closer to 1 indicate that the amino acid substitution is deleterious.
- Other combinations: PANTHER - profile HMM approach. - measure the purifying selection. The combination of PolyPhen-2 and SIFT is the most robust.
- CAROL: weighted Z scores to combine PolyPhen-2 and SIFT. The weight is determined by P_k (the complement of SIFT or PolyPhen-2) - higher P_k, higher weight.
- Evaluation datasets:
  - Positive data: (1) dbSNP variants: annotated by LSDBs for disease-causing mutations. (2) HGMD-PUBLIC. Also not found in 1000 Genome Projects.

  - Negative data: 1000 Genomes Project frequency higher than 10% in all populations.
  - In total, 2,939 positive and 14,536 negative control variants. Use about 1/3 for training and the rest for testing.

- Score distributions between positive and negative sets: PolyPhen-2, about 80% have scores close to 1 in the positive set, but only about 23% in the negative set.

- Results: AUC on the combined data: PolyPhen-2 - 0.834, SIFT - 0.819, CAROL - 0.849. The main advantage of CAROL is the lower rate of missing data: about 12% for PolyPhen-2, but less than 1% for CAROL.

- Remark: (1) data N.A. (2) Negative data: common variants.

Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies [Li & Sham, PLG, 2013]

- CONDEL is based on a Weighted Average of the normalized Scores (WAS) for combining scores from different algorithms and is available in Ensembl's Variant Effect Predictor.

- Benchmark datasets: ExoVar data for classifying of rare nsSNVs. Positive data: 5,340 alleles with known effects on Mendelian diseases from the UniProt database; 4,752 rare nsSNVs with at least one homozygous genotype for the alternative/derived allele in the 1000 Genomes Project.

- Benchmark datasets: HumVar data used by PolyPhen-2. Positive: 22,196 human disease-causing or loss of activity/function mutations in UniProtKB. Negative: 21,151 common (MAF$>$ 1%) nsSNVs with no reported disease association.

- Results on ExoVar data using AUC (10-fold cross validation): among five individual methods, PolyPhen-2: 0.826, MutationTaster: 0.853, are the best. CONDEL (combine multiple ones): 0.844 and logit (logistic regression): 0.876.

- Similar analysis on HumVar data: similar trend. The performance is somewhat better: e.g. PolyPhen-2 AUC is 0.87.

- Prior, prediction score and posterior: on average, around 5% of rare nsSNVs are pathogenic. At this prior, the posteriors range from 6.8E-4 to 0.20 when the prediction scores are increased from 0 to 1.

Annotating pathogenic non-coding variants in genic regions, NC, 2017 (TraP) [Gelfman and Goldstein, NC, 2017]

- Creating a classifier: 20 features chosen in random forest. (1) locations in the variants: in all transcripts of a gene or only some. (2) GERP. (3) the effect on canonical splice site. (4) change of splicing: e.g. new cryptic splice site vs. original splice site. Splice site defined as: 3 exonic and 6-20 intronic, using PSSM to define the strength. (5) Creation or disruption of binding sites of splicing factors.

- Training data: known pathogenic syn. and intronic mutations vs. DNMs in healthy subjects. Pathogenic variants TraP scores above 0.4, and most of benign ¡ 0.1.

- Correlation with MAF in ExAC syn. variants and WGS intronic variants: no such trend for CADD. Similar, ClinVar evaluation: GERP and TraP good classification, but not CADD.

- Application to Epilepsy DN syn. mutations: significant burden in TraP vs. control DNMs, but no such trend in GERP or CADD. Found 7 genes with large TraP scores in DN syn. and enriched with Epi. genes.

1.3.2 Annotation of Noncoding Variants

Review of current methods for functional annotation of variants:

- Supervised-learning methods: CADD, GWAVA
  - Very crude model of selection, either divergence or polymophism data (CADD does not use polymorphism data)
  - Feature combinations: generally additive. Ex. if both A and B (TFs) are bound to an element, it is much more likely to be a true CRE.
- Evolutionary methods: GERP, phyloP, PhastCons, fitCons
  - GERP, phyloP, PhastCons: divergence only
  - fitCons: both divergence and polymorphism. But has two main problems (1) all elements in a class have the same constraint; (2) prediction not at the nucleotide level (most of evaluation are at the level of elements).
- Main opportunities for improving annotation of non-coding variants:
  - Target specific cell types/tissues or phenotypes
  - Incorporate phenotypic information of variants: eQTL, disease association, and so on.
  - Nucleotide, allele-specific prediction

Computational approaches to interpreting genomic sequence variation [Ritchie & Flicek, Genome Med, 2014]

- Biological rule-based annotation:
  - Ensemble VEP (variant effector predictor) and similar tools: using rules of coding sequences.
  - Non-coding sequences: overlap with CREs and other regulatory information, including motifs. Tools: HaploReg, RegulomeDB. Rules, e.g. eQTL and TF binding/motif > overlap with open chromatin.
  - Tissue context may be important: e.g. LoF prediction and expression in a phenotype related tissue.
- Constraint based annotation:
  - GERP: observed number of substituations vs. expected number.
  - Other toos: PhastCons, PhyloP
  - SIFT (coding): construct PSSM, then the probability of observing an allele at a given position (smaller means deleterious). Usually deleterious if SIFT score $< .01$.
  - FATHMM (coding): build HMM from MSA, and incorporate pathogenecity information.
- Supervised prediction tools:
  - Coding: PolyPhen, MutationTaster, Condel, CAROL
  - Genomewide: GWAVA uses many regulatory annotations and training with HGMD vs 1GP. CADD: training using fixed vs. simulated variants.
- Lessons:
  - Constraint is a less important metric for predicting function for non-coding variants. Likely one major opportunity to improve the non-coding variant annotation.

- Tissue context is important for predicting functions of variants: for both coding and non-coding sequences.

Identification of altered cis-regulatory elements in human disease [Mathelier, & Wasserman, TiG, 2015]

- Experimental determination and prediction of CREs and TFBSs
  - Chromatin states: DNase-seq, ATAC-seq, FAIRE-seq
  - Enhancer activity assay: STARR-seq
  - eRNA measurement by CAGE
  - Histone marks and segmentation
  - TFBS prediction: predicting TFBS in ChIP-seq data (not all have direct binding); integrative approach such as CENTIPEDE
- Regulatory impact of DNA variations: mechanisms
  - Change of TF binding by DNA variation can be caused by: change of TF binding motif; change of co-factor binding motif; change of chromatin state (e.g. one allele is in open chromatin, the other in closed one). TF motif change explains only a small fraction of allele-specific binding (ASB). [Karczewski & Synder, PNAS, 2011] explore TF co-factors through ASB.
  - The impact depends on sequence context: e.g. TFBS redundancy, combinatorial/cooperative TF binding.
  - TFBS conservation: may be helpful to assess the context, however, many TFBSs are not conserved.
- Experimental study of regulatory impacts:
  - Allele-specific binding (ASB): can be inferred from ChIP-seq data.
  - Massively parallel reporter assay: [Patwardhan & Shendure, NBT, 2012] synthesizing 100K mutant haplotypes of three mouse liver enhancer, diverging by 2-3% from wild type. Most of the variants have a modest effect on the enhancer function, and only 22% significantly affected expression.
- Computational prediction of regulatory variants:
  - TFBS motif change: may also combine with phylogenetic footprinting
  - GWAVA, CADD, DANN. DANN is similar to CADD except that it uses deep neuron network, instead of SVM (lack of non-linearity).
- Outsiding questions:
  - Associate distal enhancers with genes: study show that the information of some tissues can be used to infer relation in new tissues.
  - Regulatory variant prediction: some may introduce new TFBSs.
- Lessons:
  - Regulatory variation: TF motif change may explain only a small part. Other explanations: co-factor motif change, chromatin state change.
  - To properly understand the consequence of regulatory variation: CRE context, redundancy, combinatoial or cooperative binding are all important.
  - Evolutionary studies (both cross-species and intra-species) can be used to understand/infer the rules of transcriptional regulation and the consequence of variation.

FunSeq: Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics [Khurana & Gerstein, Science, 2013]

- Background: limited studies of non-coding mutations in cancer. Recently, identified noncoding driver mutations in the TERT promoter in multiple tumor types.
- Pattern of purifying selection on different categories of genes and elements:
  - Define purifying selection: (1) Coding: enrichment of rare nonsynonymous SNPs from 1GP data (DAF $< 0.5\%$). (2) Non-coding: use fraction of rare variants (FRV). Genome-wide average is 0.6.
  - Purifying selection on genes: Average fraction of rare variants is 67%, missense 0.71, synonymous 0.61. LoF tolerant genes < disease genes (GWAS, Mendelian) < essential genes < cancer genes.
  - Selection on non-coding sequences: missense (0.71) > UTR (0.62) > TFBS (0.61) > enhancer and DHS (0.61). Constraints on enhancers and DHS are only slightly above genomewide average.
  - Selection on non-coding sequences in different tissues: stronger on spinal cord, brain, kidney; liver is about average; below average: ESC, breast, skin. Core-motif regions bound in a "ubiquitous manner" (i.e, across many tissues) are under stronger selection than those bound by TFs in a single cell line.
  - Selection on TFBS motifs: stronger on some families of TFs, eg. HMG and Forkhead.
- Defining sensitive regions:
  - Multiple categories of non-coding sequences: ncRNAs, UTRs, transcription factor (TF) peaks from ENCODE (majority of them, divided into distal and proximal), and DHS (about 30, tissue-specific ones). In total, 677 categories.
  - Categories with highest level of selection, include binding sites of some chromatin and general TFs (e.g., BRF1 and FAM48A) and core motifs of some important TF families (e.g., JUN, HMG, Forkhead, and GATA).
  - Ultrasensitive and sensitive regions: top categories measured by FRV. Ultran-sensitive: about 0.02% of human genome, FRV 0.66. Sensitive regions: 0.4% of human genome. The total size is about 0.13Mb, similar to all missense sequences (0.15Mb).
  - Highly enriched with disease-casuing variants: 40 and 400 fold enrichment in ultran-sensitive and sensitive regions respectively.
- Pattern of positive selection:
  - Focus on selective sweep model: continental populations show extreme differences in DAF (HighD sites). Alternative scenarios not examined: positive selection on standing variation.
  - Similar enrichment of positive selection in coding, DHS/enhancers.
- Pattern of somatic mutations:
  - Somatic variants tend to be enriched in missense (5x), LoF (14x), sensitive (1.2x) and ultrasensitive (2x) sequences. TF-motif-breaking/conserving ratios of TFBSs: 3 vs 1.4 in germline.
  - Some elements are recurrent in cancer samples: e.g. the promoter of RP1 is mutated in two out of seven prostate cancer samples.
- FunSeq for predicting non-coding variants that are likely driver mutations in cancer: apply to one cancer type (could be multiple samples). For noncoding variants: first common variants or not (filter), then add a score 1 to a variant for each of the condition met: ENCODE region, sensitive regions, ultrasensitive, motif breaking, target gene known and target gene is a hub. So the score ranges from 0-6. Additional score from recurrent somatic mutations.

- Questions:
  - Use fraction of rare SNPs as a measure of negative selection: is this sound? Control for mutation rate difference across regions?
- Lessons:
  - DHS/enhancers overall are under weak selection, close to genome-wide average. Using additional features: TFBS motifs (and breaking) can enrich the signal of selection.

A general framework for estimating the relative pathogenicity of human genetic variants (CADD) [Kircher & Shendure, NG, 2014]

- Motivation: annotation of variants that are general, allele-specific, measuring deleterious (in the evolutionary sense) and cover both SNVs and short indels.
  - Ex. conservation metrics cannot distinguish nonsense and misesene mutations.
- Training data: fixed mutations from human-chimp comparison, and random variants (sampling by de novo mutation rates). Fixed mutations are generally devoid of deleterious mutations. A total of 14.9 million SNVs and 1.7 million indels.
- Annotations: 63 distinct ones, including
  - Conservation: GERP, PhastCons, PhyloP
  - Regulatory information: DHS and TF binding
  - Transcript information such as distance to exon-intron boundaries or expression levels in commonly studied cell lines
  - Coding sequences: PolyPhen, SIFT, Grantham
- Examnation of individual annotations: distinguish fixed and simulated variants
  - LoF and missense mutations are depleted (especially near the TSS). The best-performing annotations were protein-level metrics such as PolyPhen and SIFT (AUC around 0.86).
  - Conservation metrics were the strongest individual genome-wide annotations (AUC around 0.55-0.6)
  - Complete table of individual annotations (Table S3): regulatory annotations not informative, e.g. K27ac 0.52, TFBS peaks 0.51, DHS 0.51.
  - Not much interaction found between these annotations.
- Combined AnnotationDependent Depletion (CADD): SVM with a linear kernel, a limited number of interactions. Results are represented by C-scores.
- Distribution of C-scores:
  - Highest in LoF mutations, next highest for missense and canonical splice-site variants. 76% of potential SNVs with C score of ≥ 20 were noncoding.
  - Also distinguish different genes: e.g. LoF in olfactory receptors have lower scores than essential genes.
  - C-scores negatively correlate with DAFs.
- C scores in distinguish functional or pathogenic variants from benign ones
  - KMT2D: C-scores distinguish pathogenic and benign variants (from ESP).

  - ClinVar: for genome-wide variants, C-scores better than GERP, PhyloP, PhastCons, AUC 0.91 vs. 0.83-0.85.
  - ClinVar missense variants: C-score AUC 0.93, better than PolyPhen (0.88). If limit to all variants where PolyPhen, SIFT and Grantham are defined: C-score AUC 0.93 vs. PolyPhen (0.91), SIFT (0.89).

- Application of C-scores to diseases: SNPs found from GWAS have higher C-scores than control SNPs (matched by AF).

- Discussion:
  - Training data: not a perfect representation of selected variants (influenced by local mutation rates, biased GC, etc.)
  - C-scores are not calibrated with pathogenicity.

- Remark/Questions:
  - Regulatory information are almost completely un-informative. Why? This is contradictory to many findings showing the relevance of enhancers for human diseases. Also contradictory to the finding that enhancers tend to be more constrained.
  - Training data: none of the two sets are highly enriched with deleterious variants, thus weak training set.

Functional annotation of noncoding sequence variants (GWAVA) [Ritchie & Flicek, NM, 2014]

- Training data: HGMD non-coding mutations and three different types of controls: random, matching distance to TSS, neighboring sequences (1000GP variants near 1kb) of each variant in the HGMD set.

- Annotations:
  - Open chromatin: DHS
  - TF binding: ENCODE ChIP-seq peaks + motif matches
  - Histone marks: ENCODE ChIP-seq calls; segmentation
  - RNA Pol 2 binding
  - CpG islands; G+C content calculated over the 100 bp surrounding each variant.
  - GERP scores: both at the specific variant and averaging over 100 bp surrounding each variant
  - Human variation: DAF around 1kb window of each variant
  - Genic context: distance to TSS, and splice sites.

- Informative features: for the region-matched control set, the most informative features (defined by Gini coefficient) are distance to TSS, average heterozygosity and DAF (measure constraint in population genetic data), GERP scores, GC content. The next ones are DHS and H3K27me3 (significantly lower Gini coefficient, not sure if significant).

- Classification algorithm: by Random Forest.

- Validation: recurrent somatic mutations in COSMIC have higher GWAVA scores than nonrecurrent ones.

- Remark: comparing with CADD, the additional features that may be beneficial include selective constraint in human population, GC content.

Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence: INSIGHT [Gronau & Siepel, MBE, 2013]; A method for calculating probabilities of fitness consequences for point mutations across the human genome [Gulko & Siepel, NG, 2015]

- Idea: merge similar sequences (e.g. defined by epigenomic features) as one group and infer selection on this group, from both polymorphism and divergence data. Instead of using population genetic model based on demography, use comparison of tested sites and neutral sites to estimate fraction of selected sites.
  - The collection of elements is assumed to be reasonably homogeneous and coherent
- Pattern of polymorphism and divergence under selection:
  - Strong negative (SN) and positive selection will generally cause mutations to reach fixation or be lost rapidly. So will not contribute to polymorphism.
  - Weak negative (SN) selection: allow polymorphisms to persist for longer periods of time, but will tend to hold derived alleles at low frequencies.
  - Positive selection (in addition to neutral sites) contribute to divergence.
- Probabilistic model: assume we have many blocks, $b \in B$ is one block. For each block, we have an element of interest E_b and flanking element F_b. Within each block, the data are O_i the divergence data for the i-th nucleotide position, and X_i the polymorphism data (derived allele frequencies). Latent variables: S_i - the selection status, Z_i - the ancestral allele at the MRCA of the target population and the closest outgroup, A_i - Ancestral allele at the MRCA of samples from the target population. The key parameters are ρ: fraction of selected sites, η: Ratio of divergence rate at selected sites to local neutral divergence rate, γ: Ratio of polymorphism rate at selected sites to local neutral polymorphism rate. The model consists of several (conditional) distributions:
  - $P(S_i|\rho)$: binomial model.
  - $P(Z_i|O_i)$: standard phylogenetic model, used to infer Z_i.
  - $P(A_i|Z_i, S_i, \eta)$: model of divergence, stronger selection (η) would imply reduced divergence.
  - $P(X_i|A_i, S_i, \gamma)$: model of polymorphism. Under neutral model, we have low-, intermediate-, and high-frequency derived allele (probabilities $\beta_1, \beta_2, \beta_3$); but under the selection model, only low-frequency derived alleles are permitted.

  The site model is applied to all elements, E_b and F_b, except that in F_b, all sites evolve neutrally. The parameters are obtained by multiplying over all elements.
- Inference: MLE of ρ, η, γ (for all blocks of a group), but also need to deal with the neutral parameters (block-specific), including branch length, local mutation rates, and β's. The parameters would indicate mode of selection: $\rho > 0$ selection, positive selection $\eta > 0$, WN $\gamma > 0$. LRT is used to test these parameters (for a class of elements). To test each element, sum over site-wise posterior probabilities associated with selection.
- FitCons method:
  - Cluster of DNA elements by epigenomic features: each epigenomic feature has 2-4 different levels (discretized). Apply clustering on each cell type/sample separately: three normal cells, and 165,000 to 224,000 sites per class.
  - Use INSIGHT to learn the parameter ρ, the fraction of sites being selected, for each class. This serves as FitCons score for the class.
  - Project ρ for each position in the class into the genome.

- Advantages of FitCons method over linear function of annotations: capture non-linear relationship between annotations. Ex. expression level and DHS: whether DHS increases selection depends on if the expression level. Possible explanation: 5' UTR (high expression) more open but less conserved than 3' UTR.
- Evaluation: in TFBS from ChIP-seq data of the same cell types; or in eQTL in the same cells. Methods to compare: (1) conservation based: PhyloP, PhastCons, GERP; (2) integrated: CADD (which uses only divergence), RegulomeDB. FitCons performs significantly better.
  - In eQTL data, the conservation based methods perform poorly, probably due to the fact that eQTL are common variants.
  - Comparison of CADD: perform no better than PhyloP, GERP. In the CADD paper, evaluation set is enriched with coding variants.
- Estimation of the fraction of human genome under selection: 7.5%.
- Utility of polymorphism data: define FitConsD scores, which is identical except that it does not use polymorphism data. The scores, fitCons and fitConsD, are highly correlated, $R^2 = .88$.
- Combine scores from multiple cell types: a simple heuristic procedure of combining the scores from multiple cell types perform nearly as good as the cell type-specific scores in the target cell types but much better on elements from mismatched or pooled cell types.
- Remark:
  - Does the model relate strong and weak selection? Treated as separate parameters, thus no sharing of information - intuitively, an element under selection between species is also likely under some selection in the population. Likely answer: use information of ρ, thus combine both strong and weak selection.
  - The model estimates a single level of selection for all blocks of a group, but we expect heterogeneity among the blocks. We should be able to estimate selection at the level of individual elements, e.g. from divergence alone.
  - The score of individual mutations: solely determined by the element containing the mutation (in fact all elements of the same epigenomic features). However, some mutations are obvioulsy more damaging even within a single element: motif-disrupting mutation is an obvious example; alternatively, it may be obvious that a site is neutral or under relaxed selection from divergence data.

Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes [Petrovski & Goldstein, PLG, 2015]

- Define non-coding intolerance scores: use only 5' and 3' UTR and 250bp upstream of TSS. Three schemes: ncGERP, ncCADD and ncRVIS. GERP score: rejection of substitution (RS), and it is OK to take average over a region: total rejection of substitution. ncRVIS scores: defined using 690 WGS samples.
- Correlation of non-coding intolerance scores with gene dosage. Use HI and lack of LoF variants to predict gene dosage sensitivity. ncGERP performs the best. Ex. OMIM HI genes, ncGERP has AUC of 0.78.

Predicting the Effects of Noncoding Variants with Deep learning-based Sequence Model [Zhou and Troyanskaya, Nature Methods, 2015]

- Training data: 919 chromatin features (TF binding, ATAC-seq, etc.), 200 bp bins. For training set, take 1000bp sequences around a positive bin.

- CNN: 3 convolution layers, at each layer, 300-1000 kernels. Also pooling layers. The last layer use logistic function. 919 output.
- Validation: cross-validation. AS DHS. Priortization of HGMD regulatory variants. , better than CADD, GWAVA and FunSeq2.
- Remark: a major challenge is to understand how it works. Idea: compare cell types to learn about weights.
- Remark: Is all regulatory information encoded by sequence? What is the role of epigenetic memory? Idea: run the program genome-wide, and predict active sequences, and see how often they are actually real, and how they is modified by heterochromatin or epigenetic features.

A spectral approach integrating functional genomic annotations for coding and noncoding variants (EIGEN) [Ionita-Laza, NG, 2016]

- Idea: suppose we have two classes (pathogenic or not), we can fit a mixture model. The challenge is that the parametric form of the distributions are unknown. If all features are conditionally independent on classes, then two features are correlated only because their scores tend to be higher or lower (at the same time) in the same class. This allows us to infer the scores of the two classes of any feature from covariance of features.
- Conditional independent model: let μ_j and μ_{j0} be the average scores of feature j in functional and non-functional classes respectively. We assume all features are standardized s.t. the mean of each feature is 0, then we only have one free parameter μ_{j0} for each feature. Let Σ_1 and Σ_0 be the covarance of features in two classes. Using Law of Total Covariance, it is easy to show that the covariance of features:
$$Q = \pi\Sigma_1 + (1-\pi)\Sigma_0 + R \tag{1.25}$$
where $R = (1-\pi)/\pi \cdot \mu_0^T \mu_0$. Plug in $\Sigma_1 = \Sigma_0 = 0$, we need to solve the equations: $Q = R$. It has k choose 2 equations of k unknowns (k is the number of features). We can solve this by least square. It is also easy to show that μ_0 is the leading eigenvector of Q. Using this result:
$$A = xx^T \Rightarrow Ax = xx^T x = \|x\|^2 x \tag{1.26}$$
for any vector x.
- Blockwise CI model: The idea is that we can use block CI to derive the parameters. FOr the system to be solvable, need at least three CI blocks.
- Prediction of variant class: the intuition is that μ_0 can be used to rank features: larger values mean the features have distinct scores in functional vs. nonfunctional classes. For a variant i, let Z_i be its score, and e the eigvenvector of Q, then the score of i is $Z_i e$.
- Application to coding variants: generally PPH, SIFT and MA best, slightly better than Eigen.
- Application to noncoding variants: For ClinVar, evolutionary conservation more important, so GERP performs best. For GWAS/eQTL, conservation less important, probably due to weak selection of common variants, and regulatory features more important, so EigenPC is the best.
- Remark:
  - Eigen does not take into accout the correlation of features in scoring a variant, thus a block with many features may dominate the results.
  - For predicting functional variants: the results depend on what type of functional variants are being evaluated. Ex. common variants and very rare variants may be associated with different features, espeically conservation features.

Tissue-specific functional effect prediction of genetic variation and applications to complex trait genetics [Backenroth & Iuliana, 2016]

- Background: limitations of evolutionary (not tissue-specific), biochemical (not functional) approaches.
- Method overview: four tissue-specific marks H3K4me3, H3K4me1, H3K27ac, H3K9ac in 127 cell types. Mixture model of two components: functional and non-functional, applied on each tissue separately.
- Comparison of histone marks across 127 cell types: MDS plots, show that the cell types are clustered as expected, e.g. by developmental origins (ectoderm, or endoderm).
- Sharing of functional marks across tissues: more sharing in promoters than enhancers. Replicate cis-eQTL sharing: skin and fat.
- Model: m variants, and k annotations. Let Z_{ik} be the k-th annotation of variant i. Indicator C_i for the status of variant i (functional or not). Assume Z is from a two component mixture model. The PDF of each component is modeled non-parametrically, allowing the correlation of histone annotations (block-wise independent). Parameteric model: mixture model where the component distribution is multivariate Bernoulli.
- Gene-based test: (1) selecting SNPs of a gene: functional ones from prediction (only 7.7% SNPs remained), then assign to genes using ENCODE correlation approach; (2) compute the BF of a SNP, similar to Sherlock. (3) BF of a gene is the average of BF of all SNPs (Note: the paper uses average of log-BF?)
- Results: about 4% of genome are functional. Found association of genes with 21 traits.
- Remark: comparison of parametric vs. non-parametric models? Similar performance (Discussion).
- Remark: possible improvements:
  - The method basically finds promoters, enhancers, insulators, etc. and predict all variants in them functional. This ignores the general annotations such as conservation, that can be predictive of functional variants.
  - No information sharing across tissues.

FUN-LDA: A LATENT DIRICHLET ALLOCATION MODEL FOR PREDICTING TISSUE-SPECIFIC FUNCTIONAL EFFECTS OF NONCODING arVARIATION [Backenroth & Iuliana, 2016]

- Model idea: variants (sequences) in each tissue is modeled as a mixture of different functional classes (active promoters, enhancers, etc.). Different tissues share the same component distributions (how annotations are determined by the classes), but have different proportions of different classes.
- Notation: for each variant i, it has k annotations (union of all tissues). l: number of tissues, m_j: the variants in tissue j, with $\sum_j m_j = m$. For variant i, let t_i be its corresponding tissue. Note: a variant may belong to multiple tissues, but we "duplicate" the tissues. So we treat each variant as belonging to one tissue.
- Model: For each tissue, decide the fraction of causal variants, π_j for tissue j. Next for variant i, sample C_i (variant i functional or not) from Bernoulli π_j. Next sample Z_i (data) given C_i: $Z_i|C_i = 0$ follows F_0, and $Z_i|C_i = 1 \sim F_1$. The two distributions are non-parameteric (multivariate kernel density). This is LDA model: we have F_0, F_1 shared across all tissues (topics), and the data of each tissue is a mixture of these distributions (documents, a variant is like a word).
- Inference: we cannot use the standard EM because of shared F_0 and F_1. Using variational Bayes.
- Annotations used: 7 histone marks and DHS from Roadmap data (127 tissues).

- Fun-V-LDA: valley in the histone marks are predictive of TF binding (similar to DNase footprint). Fit the LDA model to valley scores.
- Fitting LDA model: 9 classes. In some analysis, require two classes only: functional vs. not. So combine active promoters and enhancers into one "functional" class.
- Causal tissues of complex traits: LD score regression. Weight SNPs by their functional scores. Results generally consistent with expectation.
- Validation of predicted functional variants: MPRA data in LCL, use 800 variants with AS activities as positive data and sequences with no effects as negative data (20K variants). AUC: 0.66 (ChromHMM) vs. 0.69 (Fun-LDA).

Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data (LINSIGHT) [Huang and Siepel, NG, 2017]

- Model: similar to FitCons, no η (selection parameter from divergence). Selection of site i is measured by ρ_i, probability of selection and γ_i, the strength of selection in polymorphism data. Both parameters depend on annotations at site i: via two separate linear models. The parameters of the model are estimated by ML using genomewide data. For each site ρ_i is the LINSIGHT score.
- Training: the model inference is entirely based on human variation data (see Figure 1). Estimate parameters from across the genome. 48 genomic features: some conservation, predicted TFBSs and chromatin marks, RNA-seq.
- Validation: comparison with CADD, DeepSEA, Eigen, GWAVA in HGMD and ClinVar.
- Evolutionary constraints at enhancers: depend on context. Enhancers expressed in multiple tissue types are more constrained. Enhancers in immune, male reproductive, sensory tissues are less constrained.
- Q: Does the method require some neutral sequences for training?
- Remark: the constraint of an enhancer is entirely determined by its annotations - no random effects. Does not use mutation rates, rather, use only the change of SFS (three categories only) relative to neutral.

The human noncoding genome defined by genetic diversity [Di Lulio and Telenti, NG, 2018]

- Data: 11K WGS data.
- Method: defining mutability of a base using 7-mers. Align all 7-mers in the genome, and at the middle position, count (1) Fraction of 7-mers that have at least a SNV; (2) the fraction of SNVs with AF $> 0.01\%$ (i.e. 3 or more copies). The metric (1) is effectively the relative mutation rate (conditional probability).
- Method: defining constraint/intolerance (CDTS) of a region of 550bp. Compute the number of SNVs with AF $> 0.01\%$ (i.e. relatively common variants) in the region, and compare with the expectation from 7-mer model (multiply (1) and (2) above).
- Top CDTS percentile regions: highly enriched with protein coding sequences, promoters and enhancers, and depleted with H3K9me3 and H3K27me3. Comparison with top GERP regions: overlap mainly in protein-coding sequences.
- Non-coding constraint: can be explained partially by the constraint at protein coding sequences. Promoters or matched enhancers (using pcHi-C) of essential genes tend to be more constrained.
- Top CDTS regions are 10-20 fold enriched with pathogenic variants.

- Q: What explains the difference of constrained NCEs between GERP and intra-species variation? Possible that there are some artifacts: the top ones are enriched with regions with low mutability (not explained by 7-mer model). With GERP/Phastcons: more substitutions, better estimate of local neutral rates.

Whole-genome deep learning analysis identifies contribution of noncoding mutations to autism risk [Zhou and Troyanskaya, NG, 2019]

- DeepSEA annotations: (1) Creating post-transcriptional annotations: train on 230 CLIP-seq datasets (80 unique RBPs) using genic regions. (2) Creating transcriptional annotations: within 100KB of TSS. Expand to 2000 features. Better performance than previous version (Fig. S2).
- Creating disease impact scores (DIS): using HGMD training data, fit a L2-regularized logistic regression, on transcriptional and post-transcriptional effects separately.
- DNMs in SSC probands vs. siblings (Figure 1b): the DIS-scores of DNMs are significantly higher in probands and siblings, for both DNA and RNA effects. The effect size gets bigger with more strigent genes: LoF intolerant and FMRP targets (for RNA) or ASD relevant (for DNA).
- Analysis on 14 gene sets and 10 genomic regions (e.g. distance cutoff to TSS for transcription, and to exon for PTR): Figure 1c, mutation burden is defined as the difference of DIS in probands vs. siblings. For all variants, the effect size is about 0.02-0.03. Using LoF intolerant or ASD risk genes, RNA scores show largest effects of 0.1 to 0.2, higher than DNA scores (effect robust to removing protein-coding regions). Full results in Table S2.
- Cell type specificity: (1) Obtain tissue-specific genes from GTEx 52 tissues. (2) Mutation effects to tissue-specific genes in probands vs. siblings. Strongest tissues are brain (Figure 2a).
- Finding processes targeted by high impact DNMs: genes with stronger DIS in probands vs. siblings, the neighborhood in networks are enriched with relevant processes (Figure 2b).
- Experimental study of DNMs: choose 59 high DIS transcriptional mutations, do allele-specific reporter assay in a cell line - most show effects (Figure 3)
- Overall contribution of noncoding DNMs to ASD: 4% vs. 5% (LoF) and 3% (missense). Method: excess of DNMs divided by number of probands. E.g. for LoF, 331 minus 221.8 (expected, based on sibling data, adjusting for overall proband/sibling ratio) divided by 2,508 probands, leading to an estimated contribution of 5.4%.
- Remark: not clear if mutation burden is observed for distal regions, and gene-based analysis probably uses DNMs close to TSS.
- Remark: strongest effects among various gene sets are seen in RBP binding sequences.

Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk (Expecto) [Zhou and Troyanskaya, NG, 2018]

- Model: for a given gene, obtain +/-20kb of TSS, divided into 200 bins (200bp). For each bin, evaluate 2002 features from DeepSEA. Then do spatial transformation for each of the 2002 features: compress 200 bins into 10 features, measuring overall strength of a feature in both upstream and downstream in a spatially decayed fashion (5 spatial scales x 2 = 10). Linear model to predict expression.
- Top features: in almost all tissues, top 10 features are TF and histone features, but not DHS.
- Performance in prediction of gene expression: R = 0.7 - 0.8 in held out gene expression data.
- Evaluation in eQTL: for strongest eQTLs, predict the signs of effects with 92% accuracy.

- Application in GWAS: several examples in immune traits, top SNPs by ExPector have effects in reporter experiments, while GWAS top SNPs not.
- Evaluation of HGMD variants: most variants are predicted to reduce expression. Exception: TERT.
- Lesson: open chromatin may not be a good predictor of gene expression level, as they can both increase or decrease expression.
- Remark: how do we evaluate the impact of variants outside promoters?
- Remark: can we use predicted eQTL effects to do TWAS?

1.4 Genetic Genomics

Next-generation genomics: an integrative approach [Hawkins & Ren, NRG, 2010]

- Large collaborative projects: ENCODE/modENCODE/mouse ENCODE, TCGA, Roadmap Epigenomics Project.
- Types of high-throughput data:
  - Sequence variations: genotyping and sequencing.
  - Transcriptome: microarray and RNA-seq. The latter has advantages in identifying alternative splicing variants, non-coding RNAs (ncRNAs), gene fusion, etc.
  - Epigenomic data: DNA methylation and histone modification, DNAase I hypersensitivity data (DHS-chip, DHS-seq).
  - Interactome: protein-DNA interaction (ChIP-chip, ChIP-seq), protein-RNA interaction (CLIP-seq), protein-protein interaction (Y2H, IP followed by mass spectrometry), long-range interactions in genome (3C, 4C, 5C).
  - Genetic interactions: E-MAP in yeast (being applied to cells of higher organisms).
- Applications:
  - Mapping pathways related to metastasis, apoptosis and senescence: through RNAi screens (reference 48-53).
  - Understanding mutations in yeast: how they are related to key pathways in yeast through E-MAP data (reference 56).

1.4.1 Genotype-phenotype Map

A global view of pleiotropy and phenotypically derived gene function in yeast [Dudley & Church, MSB, 2005]:

- Data: 4710 yeast mutants under 21 experimental conditions (chosen to be relatively different/independent from each other). Under each condition, assign a value: no growth, slow growth, or full growth. Normalize the growth values of each strain under an experimental condition by its value under the YPD control condition.
- 551 mutants with growth defects in only one or two conditions. The remaining 216 highly pleiotropic genes with growth defects in 3-14 conditions. An example of pleiotropic cluster: three large, multiprotein complexes, SAGA, Swi/Snf, and Ino80, involved in cadmium, cycloheximide, hydroxyurea, and glycerol.
- Protein complexes vs. phenotypes:

  - Distinction within a module: groups of proteins within a complex may belong to different phenotypic classes, for example, the Cti6-Sap30-Ume1 and Dep1-Pho23 groups. Addtional support: the cdc73-leo1 and cdc73-rtf1 synthetic lethal interactions (all in a complex)
  - Consistency: several members of a protein complex, Paf1 and Cdc73, have common phenotypic profiles, across a large number of conditions.
  - Phenotype similarity of protein complexes: two-thirds of the complexes scoring in the range of greater phenotype similarity (score greater than 0.5).

- Pleiotropic genes: e.g. snf1 is assigned to two distinct biclusters, representing different biological functions, one related to chromatin modification, the other related to vacuole transport.

- Pleiotropicity distribution: Most genes (70%) that display growth defects under some conditions have a relatively low degree of pleiotropy (affecting only a small number of pheotypes). Still significantly higher than random expectation.

- Lessons:

  - Bicluster structure: groups of genes with similar phenotypes. This implies that if one learns the structure from a given set of phenotypes, one could use this information in a new phenotype.
  - Protein complexes: at least partially explain the bicluster structure. However, even within protein complexes, considerably heterogeneity.

Integrated genome-scale prediction of detrimental mutations in transcription networks. [Francesconi & Lehner, PLG, 2011]:

- Motivation: what are the phenotypic consequences of regulatory pertubation by change of TFBSs?

- Strategy: the regulatory changes would affect the expression of target genes, thus we could analyze the effect of TFBS perturbation through: (1) the effect of TFBS change on the promoter/enhancer, which depends on promoter architecture, etc.; (2) the effect of gene expression change on the phenotype, which depends on the role of gene in the network. The phenotypic effects are quantified by conservation data. Analyzing the effects of (1) and (2) on the conservation of TFBS.

- Network role of the gene/regulator and TFBS conservation:

  - Binding site conservation relates more to the importance of the regulator than the target gene. Binding sites are also more conserved in the promoters of genes that are harmful when overexpressed.
  - Compensation of TFs: binding sites are less constrained in the promoters of genes targeted by multiple different TFs. Controlling for possible confounders such as the number of different binding sites for each TF upholds this conclusion. When found in the same promoters, binding sites for TF pairs linked by a negative epistatic interaction are less conserved between species than other TF pairs.
  - TFBS are more conserved in the promoters of regulatory genes (TFs or signaling proteins). We find that binding sites for TFs in the top of the hierarchy are more conserved.

- Promoter context and TFBS conservation:

  - Stornger binding sites are more conserved within and between species.
  - Factors that may affect conservation: distance to a transcription initiation site, location in a nucleosome-free region, overlap with another site, and location in a divergently transcribed promoter.
  - More copies of a particular TF binding site in a promoter usually associate with reduced conservation. This result is stronger for promoters that are only targeted by a few different TFs.

Predicting phenotypic variation in yeast from individual genome sequences. [Jelier & Lehner, NG, 2011]:

- Method:
  - Estimate the likelihood that the function of a protein is affected by sequence variations, using SNP annotations (missense,nonsense) and conservation.
  - For a phenotypic trait, we know which genes may be important from deletion screens, defined a gene set of this trait. Compute a score that represents the total perturbation of the relevant gene set.
  - For each strain, wrt. a phenotypic trait, rank the strains according to the score.
- Prediction of effect on genes:
  - SNPs: use SIFT. However, SIFT depends on high quality multiple alignment, thus coverage only 73
  - Premature stop codons and indels: For indels, we compared the occurrence rates in essential and nonessential genes, assuming the rate in essential genes to mostly reflect functionally neutral or falsely reported variations.
- Prediction of effect on phenotypes:
  - Data: A total of 177 gene sets for 115 distinct phenotypes were retrieved from the Saccharomyces Genome Database (SGD)
  - To calculate a prediction score S for a strain h and a condition i, we combined the estimated change-of-function probabilities per gene, correcting for the overall sequence divergence of each strain by normalizing to the expected score per gene.
- Network-guided pruning: genes without predicted functional connections to the other genes in a set may be considered less likely to represent genuine contributors to a phenotype. Network from YeastNet. Removing genes that were unconnected (or weakly connected) within the network of each gene set substantially reduced the size of each gene set without affecting the overall performance of our prediction method.
- Phenotypes studied: mainly resistance to stress conditions (heat, UV, oxidative, etc) and chemicals.

The Majority of Animal Genes Are Required for Wild-Type Fitness, [Ramani & Fraser, Cell, 2012]:

- Background: in yeast, though the majority of genes have no detectable fitness defect under any given condition, almost every gene is individually required for normal growth in at least one environmental or genetic condition.
- Population-level phenotyping (multigeneration population assays) shows most C. elegans genes are needed for normal growth.
- Comparison with yeast: in yeast, in which the assay sensitivity is far higher, only about 40% of genes are required for wild-type growth, whereas in the worm, the number is greater than 70% (and rises to almost 100% if we consider only genes with no paralogs).
- Explanation: Most cells in any multicellular animal have different genetic networks and are effectively in as different conditions as two yeast cells exposed to different environmental conditions. A worm gene that might not be required for neuronal function might play a key role in the germline, for example.

1.4.2 Genetic Interactions

Conceptual background of genetic interactions:

- Reference: [Dixon & Boone, Systematic mapping of genetic interaction networks, Annu Rev Genet, 2009]
- Defining genetic interactions (GI): two genes show GI if the double KO has a phenotype different from what is expected when the two are independent. A widely adopted model assumes that the effects of mutations in independent genes combine in a multiplicative manner.
- Negative interactions: double mutants exhibiting a more severe phenotype than expected, such as synthetic sickness or synthetic lethality. Two interpretations:
  - Reflect the function of two genes operating in parallel biological pathways.
  - Two genesin the same essential pathway may share a synthetic lethal interaction if each mutation contributes to decreased flux through the pathway.
- Positive interactions: double mutants exhibiting a less severe phenotype than expected from the multiplicative model. Also referred to as alleviating interactions. Several cases:
  - Symmetric positive interactions: once the function of a complex is disrupted by the removal of one component, the phenotype cannot be made worse by the removal of additional components.
  - Masking interactions (assymetric): growth is better than the expected double mutant fitness and resembles the fitness of the sickest single mutant.
  - Genetic suppression (assymetric): a double mutant with increased fitness relative to the sickest single mutant.

Models of negative genetic interactions: (personal thoughts)

- Replication model: when one of the element duplicates the function of another, then removing both would have a large effect while removing one would have a smaller effect. Example: a person has two kidneys, with almost identical functions.
- Compensation model: this happens when the two elements are involved in a bigger process, when one of them is disrupted, the process is disturbed, and it produces feedbacks to other parts of the process (e.g. up-regulation) to compensate for the loss. Ex. a linear pathway with end-product inhibition, when one enzyme has a lower activity, the end product is reduced, which feedbacks to the other enzymes.
  - Remark: in the pathway model, the compensation is distributed across multiple enzymes.
- Dependency model: when one element is removed, other parts that depend on this element may be affected, creating GIs. Example: in the kidney case, if one kidney is removed, overall the system is weaker, and many parts may be affected.
- Remark: both replication and compensation models are also invoked to explain the robustness of system, while perturbing one element has a small effect on the system.

Technologies of genetic interaction mapping:

- Reference: [Dixon09]
- Yeast deletion library: 1000 essential genes that are maintained as heterozygous diploids, 4800 strains that are viable as haploids or homozygous diploids under regular growth conditions.
- Synthetic genetic array (SGA) analysis: large-scale mating and meiotic recombination, high-density arrays of yeast colonies on a solid agar surface, ultimately resulting in the isolation of haploid double mutants.

- The dSLAM (diploid synthetic lethality analysis with microarrays) approach: map synthetic lethal interactions by measuring the relative abundance of double mutants in a mixed population
- Dosage Lethality and Dosage Suppression Genetic Networks: combinations of overexpressed genes in specific gene deletion mutant backgrounds

Analytic methods for genetic interactions:

- Reference: [Dixon09]
- Functional similarity analysis: Clustering of negative genetic interaction profiles: infer the composition of biochemical complexes based on shared patterns of interactions. E.g. the early secretory pathway, chromatin modifying complexes, the homologous recombination pathway and the 26S proteasome.
- Protein complex/module mapping: positive genetic interactions between members of the same biochemical complex.
- Pathway interactions: [Kelley05] enrichment of GIs between genes of two pathways would suggest the interactions between patwhways.
- Epistasis and pathway order mapping: especially for regulatory relationship.
- Network topology analysis: the degree distribution of genes, what are the determinants of degree distributions, etc. The results provide overall picture of the organization of the network.

Main biological findings of GI network analysis: [Dixon09]

- Frequency of GI: the synthetic lethal network is a sparse network. The frequency of true synthetic lethal interactions is less than 1%.
- Essential genes: exhibit, on average, five times more GIs than nonessential genes.
- GIs between related genes: the frequency of synthetic lethal interaction between functionally related genes ranges from 18% to 25%
- Relation to pathway structure: SL (negative) GIs are more frequent between genes lying in different pathways, whereas physical interactions are more frequent among genes within the same pathway (Figure 4c). However, when a pathway or complex contains at least one essential gene, it is often enriched for within-pathway SL interactions (Figure 4d).

New frontiers in systematic GI networks:

- Expanding the spectrum of phenotypes: ex. growth phenotype: filamentous growth, morphological phenotype and protein localization
- Pathway-specific reporters: Ex. [High-dimensional and large-scale phenotyping of yeast mutants, PNAS, 2005], [Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science, 2009]
- GI Networks between species:
  - Less than 5% of synthetic lethal GIs in S. cerevisiae were conserved in C. elegans. On the other hand, a study that quantified worm mitotic spindle morphology as a phenotype detected moderate but significant (29%) conservation of GIs between S. cerevisiae and C. elegans.
  - Scer vs. Spom: some (30%) synthetic lethal GIs are conserved between these two divergent species. Differences in genetic network buffering capacity, for example due to gene duplications, could account for some of the 70% of genetic interactions that appear to be species-specific.

Ordering gene function: the interpretation of epistasis in regulatory hierarchies [Avery & Wasserman, TIG, 1992]:

- Model 1: we consider a signal-response phenotype, where S is signal and T is trait (response), X and Y are two genes we want to study that regulate this process. In the first model,

$$S \to X \to Y \to T \tag{1.27}$$

  Note that X may affect other phenotypes as well, so ΔX and ΔY have different phenotypes (which allows us to infer the gene order). In particular, when S is ON, we have: $w(\Delta X \Delta Y) = w(\Delta X)$, the phenotype of Y is masked.

- Model 2: X supresses the activity of Y:

$$S \to X \dashv Y \to T \tag{1.28}$$

  Then ΔX and ΔY have different phenotypes under different signal states, and $w(\Delta X \Delta Y) = w(\Delta Y)$, the phenotype of X is masked.

Genetic interactions in yeast [Tong & Boone, Science, 2004]:

- Methods:
  - SGA (synthetic genetic array) screens: from a query gene, double mutants with other genes, then test lethality or reduced fitness (in a single medium).
  - 132 SGA screens: query genes are actin-based cell polarity, cell wall biosynthesis, microtubule-based chromosome segregation, DNA repair.
- Results:
  - About 4000 genetic interactions, involving 1000 genes.
  - Relationship among biological processes: construct networks of GO terms, two GO terms are linked if they can be bridged by a genetic pair.
  - Characterizing unknown functions: cluster genes by their genetic interaction profiles, then the function of a gene can be predicted by other genes in the same cluster. Also identify additional components of a biological process.
  - Genetic and protein-protein interaction: enriched with PPIs, but cannot predict PPIs (many are indirect). However, gene pairs with the same neighbors often encode PPI.

Systematic interpretation of genetic interactions using protein networks [Kelley & Ideker, NBT, 2005]:

- Idea: many genetic interactions are from the dependence within pathways (sub-processes), or between pathways. Use the PPI networks to infer/define pathways, and interpret GIs in terms of within-pathway and between-pathway interactions.
- Methods:
  - Physical network: PPI, metabolic, and protein-DNA (ChIP-chip).
  - Scoring within-pathway GI: similar to finding dense components, using both genetic and physic interactions.
  - Scoring between-pathway GI: links between two dense components are more than expected by chance.
  - Prediction of new GIs: if every members in one pathway are linked to every other memebers in another pathway, except one link, then that link is likely a GI (limited to pairs instead of whole pathways).

- Results:
  - About 40% of 4,800 GIs can be explained by within or between pathways, while between-pathway GI dominates.
  - Examples of between-pathway GI: prefolding complex (microtubule formation) and dynactin complex (involved in cargo transport along microtubules). A global picture of between-pathway interactions (Figure 2b).

From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions [Ulitsky & Shamir, MSB, 2008]:

- Background: most aggravating interactions occur between pathways (about 40% from [Kelley & Ideker]). [Modular epistasis in yeast metabolism, NG, 2005] similar idea of module-level epistasis, applied to metabolic genes.
- Aim: a set of modules, and the module pairs that exhibit significant complementarity (negative GIs). Such an approach can have applications: (1) to suggest with high confidence novel functions for individual genes, (2) to identify novel functions of complete modules and to highlight interplay between modules.
- Idea: divide into modules s.t. (1) within module, positive or negative GI; (2) between modules, negative GIs. Also prefer the genes within module to have PPI.
- Model:
  - Noation: Siblings (within the same module), cousins (within two interacting modules), strangers (within two unrelated modules), CMP - complementary module pair.
  - GI model: define the latent structure of gene-to-module assignment. Siblings: neutral or positive GIs; cousins: neutral or negative GIs; strangers: neutral or positive or negative GIs. For each pair, model the GI (S-score) as a mixture of two or three components, and infer the latent structure by maximizing the likelihood.
  - Incorporating PPI: CorrelatedConnected, genes within modules must be a connected component in PPI network.
- Data: ChromBio E-MAP, focused on chromatin biology.
- Importance of PPI: the 'CorrelatedConnected' formulation outperforms other alternatives. We therefore used the results of the CorrelatedConnected formulation in all subsequent analysis.
- Module map: 62 modules, most modules are relatively small protein complexes (2-25). CMPs are often between related modules. Within modules: both positive and negative GIs. Between modules, mostly negative GIs.
- Advantage over clustering methods (based on GI profile similarity): pick up modules based on a consistent module-wise GI pattern, even if gene profile similarity is relatively weak, e.g. due to missing values.

Automated identification of pathways from quantitative genetic interaction data [Battle & Koller, MSB, 2010]:

- Model idea: suppose we have a reporter of the trait level, and we do various genetic manipulations and measure the reporter. By comparing the effect of single deletions and double deletions, we could infer the possibility of the two being in the same pathway and order; then find a global network whose structure is consistent with all the pairs.

- Individual pairs: (Table 1) given a pair of genes A, B and the trait T, there are four possible cases:

$$A \to B \to T \qquad B \to A \to T \qquad A \to T \leftarrow B \tag{1.29}$$

  And the fourth case is similar to the third one but with additional dependency between A and B. The expected T of $\Delta A \Delta B$ under four cases can be computed (effectively Conditional Independence statements): e.g. for the case 1, once B is deleted, A and T are independent, so we have:

$$\mu(\Delta A \Delta B) = \mu(\Delta B) \tag{1.30}$$

  Then we could do the Bayesian model selection for each pair: choose one model out of four possible ones. Roughly the score of a model, r, for a pair is the squared error: difference between the expected double deletion trait and the observed double deletion trait, plus some prior term.

- Global network: the problem is similar to infer the BN from a set of CI constraints. Search among all possible networks, and the score of a network G is defined as

$$S(G) = \sum_{r \text{ consistent with } G} S(r) + \sum_{e \in \text{edge}(G)} S(e) \tag{1.31}$$

  where $S(e)$ is the score of the edge e, reflecting how similar the GI profiles of the two nodes are (to place genes with similar GI profiles together).

- Remark: the limitations/issues of the model
  - Pleiotropic effect: assume that A only affects T through B, in general, A might affect T through other processes, so the double knockout phenotype would not be the same as the phenotype of deleting B only. In fact, if A is pleiotropic, but B is not, then $\mu(\Delta A \Delta B) = \mu(\Delta A)$, opposite to the proposed model (see [Phenix & Karen, 2011]).
  - Ambiguity: if A and B affect T only through one pathway, then we would have roughly $\mu(\Delta A \Delta B) = \mu(\Delta A) = \mu(\Delta B)$, and it would be impossible to determine the order.
  - Prior evidence: such as physical networks, is not used.
  - Module structure: if $A \to B$, B is associated with other genes (e.g. PPI), then A is likely upstream of all these genes too.
  - Independence: in the step of global network inference, assuming that each pair r and each edge e is independent. Does this create some bias?

An integrative multi-network and multi-classifier (MNMC) approach to predict genetic interactions [Pandey & Schadt, PLCB, 2010]:

- Motivaiton: predict GIs of gene pairs, using information about genes and the gene pairs. The past work has used synthetic sickness and lethality (SSL) features, which are not desirable (because existing GIs may be sparse).

- Methods: use multiple classifiers. The features are SSL-independent and include:
  - Single featues: such as whether two proteins have PPI, whether their GO terms are similar, etc.
  - Overlay features (2-hop features): by joining single features. Ex. the overlay feature for A and B may be: if A and C has PPI, and C and B are very similar in sequence, then the overlay feature of A and B would be 1.

- Results: most informative features (Table S1, lower KS p values) include common functions (via GO), PPI (edge, comembership of modules), coexpression, gene sequence similarlity/duplication.

The genetic landscape of a cell [Costanzo & Boone, Science, 2010]:

- Methods:
  - Query genes: 1,700 genes, and double deletion with about 4,000 genes. The query genes are chosen by the growth defects of single deletions.
  - Defining genetic interactions: fitness is inferred from the colony size, and genetic interaction: $\epsilon_{ij} = f_{ij} - f_i f_j$. Also use statistical test to call an interaction. The analysis used $\epsilon = 0.08$ (absolute value) as cutoff.
- A map of genes based on similarity of GI profiles: clusters in the map correspond to functional gene modules (Figure 1).
- Dissecting regulatory relationship: eg. Elp-Urm has negative GI, and they together modify certain types of tRNA. It was found that the negative interactors of Elp-Urm are enriched for certain types of AAs. Thus Elp-Urm complex controls a specific subset of tRNAs.
- The factors influencing the number of genetic interactions: (in the order of decreasing correlation) single mutant fitness, multi-functionality, PPI-degree and protein disorder (unstructured proteins), expression level, conservation.
- Genetic interactions between different biological processes:
  - Genes involved in chromatin, transcription, ER-Golgi transport, and Golgi-endosome transport showed a significant number of interactions that bridge diverse functions, suggesting the central importance of information flow and transport.
  - Other examples: cell cycle, metabolic genes have fewer genetic interactions with other processes; ribosome/translation genes have GIs with genes of the same category, and with nuclear-cytoplasmic transport.
- Remark:
  - Specificity of regulation: the general logic of the regulatory example is - suppose R regulates a module M, however, if we divide M into two modules M_1 and M_2, and R has interaction only with M_1 but not the other, it means that the regulatory relation is actually specific to M_1.
  - Different biological processes have different robustness: e.g. cell cycle may be a robust module with many feedbacks, so disrupting a single gene tends to have a small effect, thus this module has fewer GIs with other modules.
  - Some processes are inherently more connected to other processes: e.g. transcription, intracellular transport. How can this be modeled?

Genetic interactions reveal the evolutionary trajectories of duplicate genes [VanderSluis & Myers, MSB, 2010]:

- Motivation: after gene duplication, their function may be diverged, but the common function may still be retained. Using GI data would allow one to dissect which functions are retained/diverged.
- Model: (Figure 1) after duplication, initally the functions are similar, so there is very little GI except the negative GI between the pair. When the functions start to diverge, each of the pair may have GI with other genes with unique functions, thus the pair acquire different GI profiles, with the difference reflecting their different functions. However, the common functions are invisible in the GI data (other genes with this common function would not GI with either of the pair).
- Hypothesis: gene duplicates tend to have fewer interaction because of compensation.
- Hypothesis: the gene pair tend to buffer each other, so pairs are more likely to have GI (the effect is exposed when both are deleted).

- Hypothesis: the GI profiles of pairs may not be similar to each other.
- An alternative scenario of dosage-duplicates: these are duplicates which are maintained by evolution because of the importance of having a high dosage. Many of these dosage-duplicates are ribosome proteins. In contrast to other duplicates, dosage duplicates often have similar GI profiles.

Quantitative analysis of fitness and genetic interactions in yeast on a genome scale [Baryshnikova & Myers, Nature Methods, 2010]:

- Systematic bias in synthetic genetic array (SGA): the most important contributor is batch effect. Other sources of bias: spatial effect (neighboring colonies tend to be of he same size because of gradient of growth medium), row/column effect (the colonies in the boundary are different from others), competition effect.
- Relation to protein complexes (Figure 5):
  - Within protein complexes: both positive and negative GIs, most of them are pure positive or negative. For positive interactions, the majority (39% vs. 6.5%) are within non-essential complexes; for negative interactions, most are within essential complexes (35% vs. 2%).
  - The large majority of GIs do not overlap with PPIs.
- Genetic suppression: positive interactions between protein complexes are found. In some cases, one complex may rescur the phenotype of another complex. Ex. (Figure 6c) Rim101 pathway: when DIP4, VPS24 or VPS4 is deleted (negative regulators), constitutive activation of downstream effects; when RIM101 (upstream of these genes) is deleted, the constitutive activation is prevented, thus cells have a better phenotype.

Quantitative epistasis analysis and pathway inference from genetic interaction data [Phenix & Kaern, PLCB, 2011]:

- Model (Figure 1): as in the Avery-Wasserman paper, we have the model where $S \to X \to Y \to T$, note that the arrow may represent positive or negative regulation.
  - Both S and X may affect Y through other pathways: the effects are denoted as σ_S and σ_X respectively, with σ_Y the effect of Y on T.
  - In addition, X and Y may have signal-independent effect on T, denoted as α_X and α_Y respectively.

  Write T as a function of the state of S, X and Y (binary), then we have a regression with interaction term where the coefficients are related to σ_S, σ_X and σ_Y.
- Rule: when the deletion of one gene masks the signal-dependent effect of deleting another, the masked gene is downstream irrespectively of the pathway architecture.
  - Note: we assume the data is available where T is measured at S = ON and OFF states, and we take the difference between the two states (the effects α_X and α_Y are thus removed).
  - The simple explanation: when the signal-independent effect is removed, removing the downstream gene would not mask the upstream one because the upstream one will have additional effect; on the other hand, removing the upstream gene will always remove the effect of the downstream one.
- Applicaiton: the GAL gene network, both the fitness and expression (reporter gene) data. Show it is possible to infer the regulatory relationship.

Putting genetic interactions in context through a global modular decomposition [Bellay & Myers, GR, 2011]:

- Goal: unlike the PPI network, there is no obvious functional interpretation of a single genetic interaction, either negative or positive. The strategy/goal is the discovery of block patterns, similar to biclustering in gene expression data analysis.
- Method:
  - Use Apriori algorithm to exhaustively discover all biclusters, then use a nonparametric statistical assessment to filter out biclusters. Apply to positive and negative GI network separately.
  - Randomized the genetic interaction network by switching edge targets, thus randomizing the network structure while preserving the degree distribution.
  - Data: [Costanzo10] GI data.
- Coverage of GIs: after removing biclusters with $> 40\%$ overlap, 10,459 negative biclusters and 615 positive distinct biclusters (as compared to 20 negative, on average, and 6 positive for the random networks). Out of 85,714 negative interactions, 49,983 (58%) were contained in a bicluster structure, while 6802 out of 35,858, or 19%, of positive interactions were contained in biclusters.
- Genetic interaction hubs (and possibly GIs in general): specific genetic buffering between functional modules, and those as result of more general instability. The hubs whose interaction profiles are composed of many different modular interactions may reflect functional versatility and a predominance of unstructured genetic interactions may indicate the latter.
- Functional contexts of genes: For each bicluster (two groups), assign a GO function based on enrichment in one group. Found 74 genes had associations with more than 20 distinct processes.
  - Many examples where the genetic interaction data suggest prevalent multifunctionality that is not reflected in the current GO annotations. Many supported by other genomics data.
  - Ex. VIP1 is an inositol pyrophoshate kinase (a signaling molecule), but no other processes assigned. VIP1 associated with several otherwise nonoverlapping biclusters, enriched for a total of 13 distinct functional annotations. Function of VIP1 in DNA replication and repair.
- Discussion: the problems of the Apriori algorithm:
  - The data is discretized, therefore neglecting the resolution provided by SGA
  - the Apriori algorithm only discovers complete bipartite graphs, so missing values can break large structures into numerous smaller structures. This results in wide spread redundancy in the discovered patterns
- Remark: the results are very difficult to interpret. Ex. search VIP1 or RAP1 or MAD1, it appears in a very large number of biclusters (hundreds), some very small. Highly redundant and fragmented.

1.4.3 Chemical genomics

The chemical genomic portrait of yeast: uncovering a phenotype for all genes [Hillenmeyer & Giaever, Science, 2008]:

- Problem: probe the phenotype of yeast gene deletions: in many different conditions, and test if nonessential genes (in lab conditions) show any phenotype.
- Methods:
  - Mutants: $6,000$ heterozygous deletion strains and $5,000$ homozygous deletion strains.
  - Phenotype profiling: 726 treatments in each heterozygous deletion strain and 418 treatments in each homozygous deletion strain. The treatments include:

    - Environmental stress: AA dropout (7 AAs), high pH, heat shock, vitamin dropout, etc.
    - Small molecules: alkylating, antifungal, antihistamine, kinase inhibitors, etc.

- Results:
  - 97% deletions show phenotype in at least one condition.
  - Multi-drug resistence genes (MDRs): involved in endosome transport, vacuole transport and transcription.
  - Co-fitness analysis (group genes by phenotypic profiles): reveal functional clusters. Ex. peroxisome cluster (15 genes) show fitness defect in hydrogen peroxide treatment, high pH (requires a low pH), and oxidative stress.

1.5 Computational Methods of Molecular Networks

Molecular networks:

- The representation of how genes work together in cells/organisms: two main classes, association networks (e.g. co-expression network, or PPI network), and influence networks (e.g. regulatory networks). In practice, often limited to the structure/connection of biological networks (not the quantitative aspects).
- Sources: signaling pathways, transcriptional regulation, protein-protein interactions, metabolic networks, causal/influence networks (e.g. from eQTL data).

Experimental techniques for PPI networks:

- Y2H: sensitive, even for transient interactions. Have been used for e.g. finding genes in MAPK pathway. However, limited to nuclear proteins.

Principles of molecular network analysis:

- Modularity principle: the modules of related genes often share functionality, e.g. all associated with a complex trait/process, with another gene, etc.
- Network association principle: if many associated genes (e.g. PPI, transcription regulation, co-expression) of one gene have certain properties, then this gene is also likely to have that property.
- Remark: the two principles apply to both inference and application of molecular networks.

Inference of molecular networks:

- Inference from quantitative measurements: e.g. PPI networks from mass spectrometry, signaling networks from measurement of phosphorylation states, etc.
- Inference from integration: connections can be found through rules such as: if a gene regulates a protein X, then it is likely that it also regulates a protein Y, which interacts with X.

Application of molecular networks:

- Identification of biological processes/sub-processes: clusters of genes linked to each other, e.g. in different conditions/organisms.
- Mechanism of gene function: a gene usually functions through connection with other genes. This could be recovered through connecting genes (paths) in the gene networks. Ex. explain the regulatory function of one gene over another.

- Basis of complex traits/processes: link genes (or modules) to complex traits/processes through (1) association with the trait/process, e.g. in the amount of genes; (2) linking to other genes already known to be involved in that process/trait.
- Comparison of molecular networks can be used for a number of purposes: e.g. conserved biological processes across species, common modules acrosss different tissues, etc.

1.5.1 Reconstruction of Interaction Networks

Causal signaling networks in human T cells [Sachs & Nolan, Science, 2005]:

- Methods:
  - Data: multi-variable measurement of protein phosphorylation in individual human T cells (thousand of cells per experiment) through multicolor flow cytometry. Perturbations of 11 proteins were applied.
  - Inference: Bayesian network on the 11 variables. Identified 17 edges.
- Results:
  - The predicted network: 17 high-confidence causal edges are identified. Most are known or expected.
  - Simpler approaches: observational data - only 10 edges, or population average (instead of individual cells) - even fewer edges.

Infer nontranscriptional featurs from expression data [Markowetz & Spang, Bioinfo, 2005]:

- Idea: from perturbation data, it is possible to learn nontranscriptional features. For instance, if mutating A and B (signaling molecules) lead to similar transcriptional response profiles, then it's likely that A and B are involved in the same pathway; if the reponse profile of A is the union of those of B and C, then it's likely that A activates two downstream molecules B and C.
- Model: define a set of S genes (signaling genes, which will be perturbed), and E genes (effect genes, whose expression will be measured). The model has two components:
  - Φ describe the signaling network, and assume that the effect can be propagated along Φ (binary), i.e. disrupting one S gene will disrupt its downstream S gene.
  - Θ be the links of E genes to S genes, and also assume binary propagation (if S is disrupted, its target E will be turned off).

  All the data are discretized, and assume the real expression measurement is a simple function of the value predicted by the network (multinomial distribution), then find Φ through MAP estimation.

Transcriptional regulation of protein complexes within and across species [Tan & Sharan, PNAS, 2007]:

- Idea: TFs often regulate groups of proteins with mutual interaction, thus search for patterns in PPI and TI (transcription interaction) networks, where a highly connected cluster in PPI network that share the same regulator(s) in TI network.
- Results:
  - In existing yeast PPI and TI network: protein complexes often not share the same TF (only 9 out of 78 cases). With the new algorithm, 72 new co-regulated complexes are identified.
  - 24 out of 72 complexes are conserved in Drosophila (i.e. forming clusters in PPI, no TI network data is available). Their regulation: (1) 2 complexes: orthologous TFs exist, and have the same motif, presumably conserved regulation, e.g. Hsf1. (2) 11 complexes: orthologous TFs, but different motifs. (3) 18 complexes: no orthologous TFs.

1.5.2 Application ot Molecular Networks

Physical network model [Yeang & Jaakkola, JCB, 2004]:

- Motivation: explain data from single gene knock-out expression profiling: if one gene is affected by a regulator, then there must be some proteins that mediate the effect of the regulator through a singaling/transcriptional cascade.
- Model: the edges correspond to PPI or transcriptional regulation. Suppose a physical network is given from the data. The goal is to infer the presence/absence (as well as the sign) of each edge, i.e. a subgraph, that satisfies the constraints:
  - Interactions: the edges should be supported by protein-DNA interaction data ([Lee, Science02]), or PPI data (DIP).
  - Expression effect of genes: if g_i has an effect on g_j (expression of g_j is affected in the knockout of g_i), then there should be at least a path in the network from g_i to g_j, whose aggregrate effect along all edges (multiplication of signs of edges) is consistent with the observed effect (activation or repression). Furthermore, any intermediate node should affect g_j.

  These (soft) constraints are encoded by the factor graph model, i.e. each constraint is associated with a potential function (satisfication of the constraint would lead to higher values), and the likelihood is the product of potential functions over all constraints. Several assumptions are also made:
  - No need to consider the joint effect of two incomding edges on the node, as only single deletion data is used, i.e. for each constraint, only one path is considered.
  - Consistency of paths: there may be multiple paths, but they are not required to be consistent (same effect). If there exists one path whose effect matches the constraint, then the constraint is assumed to be satisfied. This OR logic is encoded by a potential function that is factorized into terms corresponding to single paths.
- Remark: the objective function tries to find parsimonious explanation using a small number of edges, because the edges are associated with probabilities ≤ 1.

Molecular networks in C. elegans early embryogenesis [Gunsalus & Piano, Nature, 2006]:

- Motivation: functional modules and interconnections through network analysis.
- Methods:
  - Data: PPI between $4,000$ proteins; a compendium of expression profiles; RNAi phenotypic profiles consisting a vector of 45 phenotypes, describing specific cellular defects.
  - Network construction: multi-support network, with PPI, expression similarity based on correlation and phenotype similarity.
  - Finding functional module (or molecular machines): graph-based clustering algorithm.
- Results:
  - The first type of graph-based clusters: high density of links supported by PPI and phenotypic similarity. Often reveal molecular machines, e.g. ribosome, proteasome, vaculor H^+ ATPase.
  - The second type of graph-based clusters: high density of links supported by expression similarity and phenotype similarity, containing few PPIs. Genes that participate in distinct yet functionally inter-dependent cellular processes. Ex. mRNA metabolism (transcription, translation, trafficking).
  - Characterization of genes with unknown functions through the functional clusters: verified by GFP-based cellular localization analysis.

SPINE [Ourfali & Sharan, Bioinfo, 2007]:

- Motivation: address two problems in [Yeang, JCB04], 1) the local optimum is found; 2) to encode the OR logic (at least one path that satisifies the constraint), multiplication of terms, but these terms are dependent to each other.
- Model: suppose the graph is given, where each edge exists with a probability, $r(e)$. The goal is to find the sign of each edge s.t. the maximum number of constraints are satisfied.
  - Objective function: since the edges exist with probabilities, the goal is to maximize: $\sum_{(s,t)} P(K_{s,t} = 1)$, where (s,t) represent the source and target constraint (the perturbation effect), and $K_{s,t}$ is the indicator variable whether (s,t) is satisfied by the edge assignment.
  - $P(K_{s,t} = 1)$: this depends on the edge assignment. Suppose the assignment is given, there are multiple paths that may satisfy (s,t) constraint, the total probability is computed from the probabilities of individula paths through inclusion-exclusion principle (as the paths may share edges).
- Remark: could also formulate as find X, the edge presence variable, and S, the sign variable of edges, s.t. $P(X,S)$ is maximized. Then this probability should have the probability of paths (determined by X), and constraint satisfication (determined by S).

Information Flow Analysis of Interactome Networks [Missiuro & Ge, PLCB, 2009]:

- Motivation: the importance of a gene in a molecular network is better characterized by "bottlenecks", instead of degrees. However, the existing definition, "betweeness", is based on the shortest path (a gene that lies in the shortest path between many pairs have high betweeness), which is flawed (e.g. slightly longer parallel paths are ignored).
- Information flow model: the confidence score of the interaction is viewed as resistance, and the currents flow from every pair of nodes. The information score of a node is defined as the total currents going through the node. Empirically showed that hte information flow score is poorly correlated with degree, and somewhat correlated with betweeness, with substantial difference.
- Validation:
  - Correlation of information flow scores with essentiallity and pleiotropicity in yeast and C. elegans interaction networks.
  - Application to a muscle gene network: first construct muscle interaction networks, then rank all genes by the score and show that the top genes often are important for muscle function.
- Application to module discovery: recursively remove the highest score nodes, until the network disintegrates into a set of modules.

Identifying functional modules in proteinprotein interaction networks: an integrated exact approach [Bioinfo, 2008]; BioNet: an R-Package for the functional analysis of biological networks [Beisser, Bioinfo, 2010]

- Background: Prize-collecting Steiner Tree (PCST). Ex: decide where to build shops, each shop (vertex) has a potential profit value, and each edge has a cost (building the street). Note: both profits and costs are non-negative. Our goal is to maximize the profits, which is the sum of weights of vertices subtract the costs of all edges connecting them.
- Goal: given p-value of each gene, and a network, detect maximum scoring subgraph (MSS).

- Beta-uniform mixture (BUM): fit a distribution of p-values [Pounds and Morris, Bioinfo, 2003] as a mixture of uniform and a special case of Beta distribution (beta = 1). Note: limit beta = 1 s.t. the distribution is monotonically decreasing, when $\alpha < 1$. The parameters of the BUM model are π_0 and a (Beta parameter), and can be estimated by MLE. This allows one to control for FDR.

- Node scoring: given a set of p-values for each gene/vertex, our goal is to have a scoring function on the vertices s.t. the score is positive if its p-value is less than the threshold at a given FDR; and negative otherwise. Our basic idea of scoring a vertex is: LRT of p-value, $S(x) = \log \frac{f_1(x)}{1} = \log a + (a-1)\log x$. To ensure the signs of the scores, we consider $S(x) - S(\tau)$ where τ is the p-value threshold at a given FDR. This leads to adjusted score as: $(a-1)[\log(x) - \log(\tau)]$.

- Maximum-Weight Connected Subgraph Problem (MWCS): find a connected subgraph with the maximum score. Note: some nodes have negative weights, otherwise, one can just find the biggest subgraph.

- Equivalence of MWCS and PCST: the graph will have both positive and negative weights (if only positive, the problem is trivial, just the whole graph). Let w be the minimum weight of all vertices, and we define a new graph, with vertex weight $p(v) = w(v) - w$ and edge cost $-w$. Its easy to show that the PCST and MWCS problem are equivalent.

- Intuition: to solve MWCS, we need to connect as many as positive vertices as possible, while paying for the costs of negative vertices. This is similar to PCST, where we collect profits but paying the costs of linking the nodes. By adding $-w$ in the graph, we shift the costs of negative vertices to edges in PCST.

- Solving PCST by ILP: intuition is to represent whether a vertex is included in the final solution as binary indicator (to be solved), and edges, etc. as constraint. The problem is then cast as ILP problem.

Network-based methods for human disease gene prediction [Wang & Yu, BFGP, 2011]:

- Background: Interactome networks: PPI, metabolic, and TRN. Function networks: transcription profiling, phenotypic profiling and genetic interaction networks.

- The role of hubs: some studies demonstrate that hubs are associated with mulitple phenotypes. However, further investigation demonstrated that only essential disease genes were associated with hubs and were widely expressed, while nonessential disease genes did not demonstrate these characteristics.

- The framework: given a set of seed genes, predict the proximity of any test gene to the seed genes.

- Defining proximity:

  - Direct neighbor counting: the number of seed genes connected directly to the test gene. An application using this method found 10-fold enrichment over random expectations.
  - Shortest path length: A node that has close proximity to multiple seed nodes receives a higher score as a candidate disease gene.
  - Why local measure of proximity may be insufficient? Eg. two proteins are connected by a hub, or by a protein with a low degree, or through more than one shortest path.
  - Global distance measure: diffusion kernel or random walk with restart.

- Global measures perform better than local measures: In a test involving 110 disease families containing 783 genes [Kohler08], The random walk with restart method performs the best, better than the local distance measure methods (direct neighbors, shortest path). In another evaluation, random walk with restart method also performs the best, better than clustering and neighborhood methods.

- Integration of multiple types of genomic data:

  - Endeavour: (1) 10 features of genes: expression, pathway membership, CRM, etc. (2) candidate genes of interest were ranked based on their similarity to known disease genes in each of these features. (3) Combining the ranks of individual features using order statistics. The correct gene, in the validation of 703 disease and pathway genes, was ranked 10th among 100 candidate genes on average.
  - Prioritizer: (1) functional gene networks: Pair-wise functional associations among genes in each feature were integrated into a single functional linkage network, weighted by overall functional associations, using a naive Bayes classifier. (2) Scores of candidate genes: the sum of the weights of the network links to known disease genes. The performance using the integrated functional network (62% success rate) is better than using the PPI network alone (40% success rate).

- Integration of disease phenotype networks and PPI networks: diseases with similar phenotypes often share either a common set of underlying genes or functionally related genes.
  - CIPHER: integrating the phenotype and PPI networks. Disease similarlity profile of a disease p, and gene closeness profile of g, then correlation between the two profiles (Idea: if p is correlated with some diseases, and g is involved in the same diseases, then g is likely to be involved in p). Performance comparable to Endeavour.
  - PRINCE: construct a prior of g-d association, based on if g is related to another d' and the similarlity between d and d'. Then apply the network propagation algorithm. The difference with CIPHER: PRINCE uses global distance measure while CIPHER uses local. PRINCE outperforms CIPHER, and also random walk with restart (however, the opposite conclusion from Saket paper).
  - Random Walk with Restart on Heterogeneous network (RWRH): The random walker is no longer restricted in the gene network (PPI) but is also allowed to jump to the phenotype network. Outperform random walk with restart and CIPHER.

- Disease modules or subnetworks, in which members would share similar functions, expression patterns or metabolic pathways. The expression association of genes and phenotypes can be used.
  - Co-expression subnetworks related to disease: from tissue-specific gene expression.
  - Disease-specific genes, such as differentially expressed genes identified under disease conditions, were mapped to global PPI network. The shortest path subnetwork was then built by including only the nodes in the shortest path connecting the disease-specific genes.

- Future perspective: (1) Better inclusion of phenotype similarity networks. (2) Node removal and edge removal: distinct functional consequences.

- Remark: why global measures are better than local measures? Shared neighbors increase the evidence of association between two genes (e.g. if they belong to the same protein complex). Similarly for genes in cliques.

Discovering regulatory and signalling circuits in molecular interaction networks [Ideker & Siegel, Bioinfo, 2002]:

- Goal: suppose each gene is activated (differential expression) to some extent, find the subnetwork where most genes tend to be highly activated.

- Methods:
  - Subnetwork scoring: (1) each gene, the p value of differential expression (or other properties) is converted to a z score, which is then averaged to give the z score of the subnetwork; (2) calibration againtst background: find the z score of all k-subnetwork in the background, and calculate the corrected z score (normalized).

  - Search for high-scoring subnetworks: the problem of finding the maximum-scoring subnetwork is NP-complete. Thus use simulated annealing, where the basic operations are switching the states of nodes (in or out of the subnetwork): accept a switch with high probability if it increases the score.

Network-based classification of cancer expression profiles [Chuange & Ideker, MSB, 2007]:

- Motivation: classifying expression profiles (e.g. cancer) on single genes has low power. Enhancer by analysis on groups of genes, either common pathways or functional modules in protein networks.
- Background: single-gene based classification of cancer expression profiles misses some key genes, e.g. p53, presumably because the expression of the regulators are often lower (compared with downstream effectors).
- Methods:
  - Data: expression profiles of breast cancer patients, some metastatic, others non-metastatic.
  - Scoring subnetworks: the average expression of all genes in that subnetwork (expression values are normalized), called subnetwork activity, and its discriminative potential.
  - Search subnetworks with highest discriminative potential: greedy search, starting with a single node and add the edges if that increase the score.
  - Significance test: null distribution from (1) random subnetworks; (2) random subnetworks originating at the same node; (3) permutatoin test (on class labels).
- Remark: other network-based formulation, e.g. Steiner tree algorithm.

Breast cancer network [Pujana & Vidal, NG, 2007]:

- Problem: suppose we know some genes involved in a process (cancer), how can we find other genes also involved, and identify the mechanism?
- Idea: we could find many functionally related genes through different types of evidence, PPI, genetic interaction, etc. And the evidence can be combined to rank the related genes. Two additional ideas:
  - Cross-species data: could also be used, e.g. PPI in another species may still boost the evidence of functional relation.
  - Network: two genes may not be directly related, but may be related through a common neighbor.
- Methods:
  - Reference genes in breast cancer: BRCA1, BRCA2, ATM and CHEK2; from these, derive 164 genes that co-express with one of these four genes (with CC > 0.4).
  - Functional networks: using data from yeast, worm, and fruit fly, including PPI, protein complexes, GI, phenoclustering (genes with similar phenotypic profiles), co-expression, expression effect (expression of a gene x is affected by the mutation of gene y).
  - Brca1 centered network (BCN): all genes connected to Brca1 in one or two steps in the functional network. And the genes can be ranked by the supporting evidence (e.g. the number of the types of connections).

Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes, [McGary & Marcotte, GB, 2007]:

- Related work: [A probabilistic functional network of yeast genes, Science, 2004], [An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae, Lee & Marcotte, PLoS ONE, 2007].

- Constructing functional gene network:
  - Training data: for each pair, whether the genes have the same GO
  - Data types: co-expression, PPI, co-citation, GI, etc.
  - Network edge weight: for each data type, compute LLR score: train the model parameters from the training data.
  - The final edge weight: combine the LLR scores of multiple data types: similar to Naive Bayes, however with weighting. The weighting scheme is: the strongest evidence receive weight 1, the second strongest weight 1/2, the third 1/3, and so on.
- YeastNet: 102,803 functional linkages among 5,483 yeast genes, capturing the tendency of the genes to share GO annotation.
- Gene prediction: given a set of seed genes, predict whether another gene is related to the same phenotype. Score: the sum of the weights of linkages connecting the query gene to genes in the seed set. This is the naive Bayesian combination of evidence that the query gene belongs to the same pathway as the seed set genes.
- Evaluation of predictive method: leave-one-out cross-validation in th phenotype data of 100 nonredundant phenotypes.
- Prediction performance: a majority of phenotypes are reasonably predictable, with 70% of the phenotypes predictable at AUC above 0.65. In contrast, for 100 random sets, none reaches AUC above 0.65.
- The impact of gene networks: PPI network: median AUC about 0.6. YeastNet (all data): median AUC about 0.7.

A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans [Lee & Marcotte, NG, 2008]:

- Forty-three RNAi phenotypes: range from gross (for example, sterility, lethality and growth defects) to specific, such as those affecting single cellular processes (for example, mismatch repair defects, apoptosis) or single tissues (for example, vulva development).
- A robust correlation between gene connectivity in the functional network and the frequency of nonviable RNAi phenotypes. The connectivity of a gene, measured for each gene as the normalized sum of log likelihood scores of its linkages. Using the same network (orthologs), found correlation between connectivity and essentiality in mouse.
- Network-based prediction of LoF phenotypes: For a given phenotype, each gene in the worm proteome was rank-ordered by the sum of its linkage log likelihood scores to the seed set of genes already known to show that phenotype. This approach strongly predicts genes with 29 of the 43 phenotypes reported from genome-wide screens, using leave-one-out prediction.
- Example: life span. A previous study reported 29 genes that extend lifespan when inhibited. Of the 50 most connected to these genes, 10 are replicated in an independent screen.
- Testing: retinoblastoma tumor suppressor pathway. The retinoblastoma-SynMuv B pathway acts genetically redundantly with the synMuv A pathway to repress development of the hermaphrodite vulva. Six genes were identified that encode suppressors of the synMuv pathway. The six genes interacted with 62 and 142 genes in the core and non-core of Wormnet, respectively. RNAi against 10 of 50 tested core interactors (20%) and 6 of 124 non-core interactors (5%) appreciably suppressed this phenotype, comparing with 0.9% in random genes. The other genes found in the screen are still highly clustered in WormNet, but not connected to the seed genes.

- Remark: results from leave-one-out predictions are misleading. Not much need of modeling the propagation in the network, as the vast majority of the genes affecting the phenotype are known. In a more realistic setting, one knows perhaps a small fraction of genes affecting the phenotype, thus need to model the indirect effects/links.

GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function [Mostafavi & Morris, GB, 2008; Bioinfo, 2010]:

- Background: GBA algorithm using multiple datasets: (1) weighted networks of individual datasets; (2) combining multiple networks to create a single functional association networks; (3) given a seed gene list, scores each gene based on its proximity to the genes in the seed list.
- GeneMANIA framework: optimizes the network weights and calculates the discriminant values (label of the nodes) separately.
  - Setting the label bias: positive and negative nodes from existing data. For any other nodes, use the average label value of all assigned nodes.
  - Gaussian field label propagation (GRF) algorithm: let f be the labels, y be the label bias (initial labels), find f s.t. (1) the difference between y and f is small; (2) the difference between f of the adjacent nodes is small, weighted by the edge weights.
  - Network weighting: combine multiple datasets. Let W_h be the weight of the data type h, the final weight W is a linear combination of W_h, with weights α_h. α is learned by regression: the weight W of an edge should match the label bias (if two nodes are positives, the edge weight should be high). Since the initial label bias may be sparse, use ridge regression.
  - Network weighting with multiple prediction tasks (phenotypes or GO categories): combine the regression problem of multiple tasks.
  - Sparsification: for each node, only a small number of neighbors are informative. Limit to top 50 does not descrease performance and improve running time.
- Results: when some of the association networks are irrelevant, the prediction performance of equal weighting scheme is degraded.
- Reference: Fast protein classification with multiple networks, [Tusda & Scholkopf, Bioinfo, 2005]

ResponseNet: integrated analysis of genetic, physical and transcriptional data [Yeger-Lotem & Fraenkel, NG, 2009]:

- Motivation: to study a biological response, the genetic screens (the gene manipulation that affects the phenotypes) and mRNA profiling often lead to different genes. How to do integrated analysis that uses the two complementary data to get a more complete picture?
- Model: let *Gen* represent genetic hits, i.e. genes identified by the genetic screen under some perturbation - tend to be regulators, and *Tra* represented genes differentially expressed under the same - tend to be enzymes, etc. The idea is to find the pathways connecting *Gen* and *Tra* through PPI and protein-DNA interactions.
  - Interaction network: two types of edges 1) protein-protein interaction, 2) metabolic enzymes in the adjacent reactions, 3) protein-DNA interaction: transcriptional regulation. Curated from other data set or literature.
  - Finding connections through minimum flow in the interaction network: the objective function (cost) has two parts: total flow (favoring high-confidence edges), and the cost of using only a small number of regulators (all regulators should be used, favoring regulators with strong evidence).
  - Gene ranking: the potential regulators uncovered can be ranked based on the in-flows.

- Results:
  - Validation of ResponseNet algorithm: in Ste5 deletion, and in DNA damage response. The method recovers the known regulators (even if they are not in the list of genetic hits or differentially expressed genes).
  - Analysis of alpha-synuclein toxicity (incuding ROS production and ubiquitin system impairment): (1) Genetic hits: 55 suppressors and 22 enhancers of toxicity, in vesicle-trafficking genes, kinases and phosphatases, transcriptional regulators, manganese transporters, ubiquitin-related proteins. (2) Differentially expressed genes: hundreds of genes, up-regulated genes: oxidoreductase activities; down-regulated: rbosomal genes. (3) ResponseNet predictions: known pathways: ubiquitin-dependent protein degradation, cell cycly regulation and vesicle-trafficking; novel ones: heat shock, TOR pathway, etc.

Integrating Proteomic, Transcriptional, and Interactome Data Reveals Hidden Components of Signaling and Regulatory Networks [Huang & Fraenkel, Sci Signal, 2009]:

- Motivation: similar to the ResponseNet algorithm. To apply the method in broader context, e.g. proteomics data, the requirement of source and sink should be removed.
- Model: a variant of the ResponseNet algorithm (PCST). Instead of connecting nodes (genetic hits and differentially expressed genes) through flows, simply find edges connecting a specified set of terminal nodes. The edges are chosen to minimize the costs (total weight of edges, favoring reliable edges) while penalizing for excluding the terminal node.
- Set up PCST: min. objective functions, which is the cost of excluding relevant genes and including weak edges. (1) Edge: the negative log of the probability weights on the edges as the edge costs. So weak edges have large costs. (2) Vertex: log ratio of node strength. (3) Combine node and edge costs: a parameter β, empirically determined to optimize the module/tree size.
- Method: the model is applied to yeast pheromone response: genes differentially phosphorylated by pheromone (112), and genes differentially expressed (3-fold, 201).
- Results: more pathways than previously known.
  - Core pheromone pathway: Ste2 (receptor), Gpa1, Ste20, Ste12 (TF)
  - Other MAPK pathways: PKC pathway, filamentous growth pathway.
  - Regulation of cell cycle, transcriptional control, cellular polarization, cellular transport.

Predicting essential genes based on network and sequence analysis. [Hwang & Huang, Mol Biosys, 2009]:

- Motivation: The ability to rapidly identify essential genes has been described to be the most important task of genomics-based drug target validation.
- Network properties: clustering coefficient (CCo), betweenness centrality (BC) of a node i measures the ratio of the shortest paths going through a node, Closeness centrality (CC) of a node i is the reciprocal of the sum of average shortest distances from all other nodes in the network to node i, $KL(i)$ is defined as the largest size of cliques that a node i can join, the essentiality index (EI) as the proportion of essential proteins interacting with a node (protein) i, common-function degree (CFK), to measure the amount of common-function adjacent nodes.
- Sequence properties: ORF lenght, strand, conservation.
- Essential vs. nonessential genes: The means of degree (11.55), clustering coefficient (0.16), and betweenness centrality (0.001) of essential genes in yeast are about twice as large as those in nonessential genes. The clique level (KL, the size of the largest cliques a gene belongs to) outperformed other

topological properties. The other two properties NID and CFK also show better performance than degree. The averaged essentiality index (EI), as well as clustering coefficient (CCo), of essential genes are significantly larger than nonessential. However, they perform worse than expected for the top 10% ranked genes although both of them perform well in other regions.

PRINCE: Associating genes and protein complexes with disease via network propagation. [Vanunu & Sharan, PLCB, 2010]:

- Problem: given a query disease, a PPI network, known disease-gene associations (of this and other similar diseases), priorize the genes in the PPI network of the query disease.
- Methods:
  - Idea: from the known causal gene and the gene of similar diseases, find the common neighbors in the PPI network.
  - Data: PPI network - from 3 human proteome data (MS); disease similarity - from literature profiles using MeSH; known disease-gene associations - from OMIM database.
  - Gene ranking: let $G = (V, E, w)$ be the PPI network. First define the prior relevance of any gene to the query disease (q), as $Y(v) = L(S(p, q))$ where $S(p, q)$ is the similarity between two diseases p and q (p is the disease v is associated), and L is a logistic function. Then propagate the flow from v to its neighbors, for any v. Iterative propation until convergence, then the flow to every node v is its score.
- Results:
  - Finding causal genes of prostate cancer, Alzheimer and diabetes.
  - Three cases of protein complexes (densely connected proteins) associated with three diseases.
- Remark: the main problems are:
  - Need to have known causal genes (or of similar diseases) to start with: both limit the application, and in the results, may bias the genes.
  - To apply for linkage/association mapping: an interval (or SNP) may be associated with multiple genes.

AraNet: functional gene network of Arabidopsis [Lee & Rhee, NBT, 2010]:

- Goal: associate genes with traits from functional gene network.
- Idea: network-guided association: if the neighbors of a gene are involved in certain trait, then this gene is also likely to be involved in the trait.
- Methods:
  - Functional gene network (AraNet): 24 distinct types of gene-gene associations including co-expression, PPI, sharing of protein domains, similarity of phylogenetic profiles, and the same types of evidences from orthologs in yeast, fly, worm and human. The total: 1 million linkages among 19,467 genes (73%), with each linkage weighted by the log likelihood of the linked genes to participate in the same process.
  - Trait association: based on the association of network neighbors.
- Results:
  - Evaluation: Known GO process annotations; linked genes tend to share expression patterns (data not used in constructing AraNet); genes involved in traits tend to be inter-linked.

– Application: (1) Candidate genes of seed pigmentation: reverse genetic screening for the candidate genes and measur the phenotypes: 14 out of 90 candidates display phenotypes; (2) Assigning functions to unannotated genes.

The power of protein interaction networks for associating genes with diseases [Navlakha & Kingsford, Bioinfo, 2010]:

- Neighborhood method: a test gene is ranked by the percent of its network neighbors that are associated with the disease.
- Direct interaction (DI) method: a test gene is predicted to be positive if it has DI with k disease genes (k is a parameter, typically 1-3).
- Random walk with restart algorithm [Kohler08]. PRINCE algorithm [Vanunu10]
- Results: Random walk with restart algorithm performs the best.

HumanNet: Prioritizing candidate disease genes by network-based boosting of genome-wide association data [Lee & Marcotte, GR, 2011]

- Construction of HumanNet: 21 genomic datasets from yeast, worm, fly and human, covering PPI, genetic interactions and co-expression.
- Validation of HumanNet with cellular phenotypes: human mammary epithelial cells (HMEC) treated with shRNA, and measure the growth phenotype. The genes in clusters in HumanNet are more likely to have the same phenoptyes.
- Label propagation in HumanNet: when the seed genes are associated with uncertainty (e.g. in GWAS, each gene has some association score, but most of genes are not strong candidates), propagate labels similar to PageRank.
- Application of HumanNet in GWAS: in Crohn's disease and T2D. Run the analysis on WTCCC data, then . By using HumanNet, the ranks of a number of genes in WTCCC data are boosted, e.g. STAT3 - from 17 to 8; JAK2 - from 3139 to 38. The function of these genes (e.g. JAK2) in the diseases can be found in related studies or replicated in later meta-analysis [Barrett08, Zeggini08].

Predicting Protein Phenotypes Based on Protein-Protein Interaction Network, [Hu & Cai, PLoS ONE, 2011]:

- Phenotype data: 1,460 yeast proteins belonging to 11 phenotypic categories, including mating and sporulation defects, cell cycle defects, cell morphology, stress response defects, etc.
- Network data: PPI network from STRING, covering both physical and indirect interactions. All the genes are assigned to 733 complexes (from physical interactions) and 86 pathways.
- Relating phenotypes and pathways/complexes: a gene can be represented by the membership vector of its complexes/pathways. A phenotype is then represented by the average membership vector of all its genes.
- Predicting phenotypes: for any phenotype, suppose we are given a set of seed genes. To predict the effect of a query gene, we find the strength of its interaction with all seed genes, and a total strength (some average interaction with the seed genes) is then calculated.
- Multi-phenotype prediction: for each gene, predict the score for each phenotype, and the result is a ranked list of phenotypes for the gene.

The Impact of Multifunctional Genes on "Guilt by Association" Analysis[Gillis & Pavlidis, PLoS ONE, 2011]:

- Hypothesis: GBA is based on network association to assign gene functions. However, the multifunctionality of genes may create some bias: one can simply assign multi-function genes to any new phenotype, and achieve good performance. Therefore, the network-based prediction is really due to recapturing the multi-function genes, instead of GBA.
- Defining multifunctionality: by the number of GO terms a gene is assigned to, where a GO process is weighted by (the inverse) of its size.
- Predicting GOs using multifunctionality: create a single ranked gene list based on the defined multifunctionality, and use this as a predictor for assigning genes to 100 GO categories. Suprisingly, the mean AUC is 90% across all GO terms.
- Predicting human diseases with multifunctionality: good performance in Alzheimer's disease, schizophrenia, Parkinson's disease and autism: AUC > 0.7. Indeed, across 4069 sets of disease genes from OMIM [59], the average ROC is 0.76.
- Multifunctionality and node degree: modestly correlated. The performance is better than GeneMANIA using PPI network.
- Predicting GOs using node degress in co-expression network: mean AUC is 0.58.
- Predicting GOs using node degress in PPI network: mean AUC is 0.63. Prediction using GeneMANIA: mean AUC is 0.7 (3-fold cross validation, the performance does not change much with higher fold). Furthermore, the variation of the performance across GOs are highly correlated between GeneMANIA and node degree.
- Alternative metrics: use precision (PPV) at top 50 genes on GO prediction. Single ranking using multifunctionality: 4.7%, and GeneMANIA: 3.8%.
- Node degree correction methods:
  - E.g. the first step in GeneMANIA's operation is to attempt to correct for node degree (down-weighting each edge by node degree).
  - Top overlap method for co-expression network: choose only overlapping nearest neighbors, i.e. to have an edge (u, v) in the network, u must be ranked high in the neighbors of v, and vice versa.

  Node degree correction does not seem to solve the bias: e.g. top-overlap method, still preserve the degree rank of genes.
- Suggestions for removing the bias: comparing the ranking the algorithm gave them to the optimal gene ranking from GO. If there is similarity between the rankings, the algorithm-derived predictions cannot be assumed to be meaningful.

The role of indirect connections in gene networks in predicting function [Gillis & Pavlidis, Bioinfo, 2011]:

- Background: the value of indirect connections in GBA. Most of the methods incorporating indirect connections report improvement over GBA between direct connections, although they tend to perform comparably and only slightly better than direct GBA (e.g. fraction of neighbors associated with the disease).
- Hypothesis: in constructing the network, weak connections (e.g. coexpression or PPI) are removed. Most methods incorporating indirect connections simply reconstruct network information that was originally present as "weak" connections and deliberately removed.
- Comparison: (1) Direct GBA: The sum of co-expression values between the training set and the candidate gene was divided by the sum of co-expression values between the genes outside the training set and the candidate gene to determine degree of candidacy. (2) GeneMANIA.

- Level-2 connections in the sparsified network (indirect with one intermediary) are very strongly predicted by the original correlations. median ROC of 0.986.
- Extended GBA: create edges that correspond to indirect links in the original network. The weight of an indirect link at a given level: setting the weighted sum of indirect links at that level equal to the sum of direct links.
- Comparing the performance of GeneMANIA to basic GBA and extended GBA using co-expression network: GeneMANIA performs much better than BGBA, but it performs very similarly to EGBA. The effect of sparsification: in a mouse co-expression network, sparsification dramatically reduces the performance. Mean AUC = 0.72 with the aggregate co-expression network, but the sparsified network (0.5%) performs poorly: GBA 0.56, EGBA and GeneMANIA 0.57.
- PPIN and co-expression network:
  - For co-expression networks, indirect connnections simply recover the filtered weak connections (and that may not be fully recovered).
  - For PPI network: the extended PPIN is acting more like a co-expression network, where the metric represents similarity in behavior across a range of conditions, rather than strict interaction. Adding PPI network to co-expression network provides a small but significant improvement in function prediction.

Ernest Fraenkel talk: [2012]

- Motivation: for a complex trait (cellular phenotype, cancer, etc.), we often have different types of data, most often, these include: genes that are differentially expressed (modified) in different phenotypes, genes that are causally related to the phenotype. The two types of data do not reveal the same set of genes (some are causal regulators, some are effectors), and need a stategy to use both type of data to understand the mechanistic basis of the phenotype.
  - The types of data include: genetic hits (e.g. in yeast), GWAS hits, change of gene phosphorylation, change of gene expression, etc.
  - Example: DNA damage response. The genes from (1) DE genes in DNA damage response; (2) mutants that have low fitness with DNA damage.
  - Background: our knowledge of biological pathways is limited, and physical network is a better framework to organize our knowledge. Ex. in EGFR mutation, find which genes have different phosphorylation: a small fraction in known EGFR pathway, about 40% in other KEGG, the rest not in any known pathway.
- Methods: suppose we have a physical network (PPI, transcriptional regulation), and genetic hits, differentially expressed (DE) genes related to a phenotype. The idea is to link the genetic hits to the DE genes through the physical network.
  - Prize-collecting Steiner tree: The penalty function penalizes any experimental nodes not in the tree; and the cost of each edge is based on our confidence in the physical network.
  - Prize-collecting Steiner forest: the motivation is that some critical nodes may be missing. Use an artificial root node to connect all nodes.
- Glioblastoma gene network: measure the change of phos. in EGFR mutants with different activities; and the change of expression in EGFR mutants. The two genes lists are then analyzed using the Steiner tree method.
  - Transcriptional regulation network: use DNase I HS data, and motif analysis in HS regions (correlation with gene expression).

  - Results: Estrogen receptor, AR, HSP90, P53, etc.

- Remark/questions:
  - The key question is whether the inferred genes (subnetworks) are causal to the phenotype. Ex. for diseases, it is well known that the DE genes may reflect the change of physiology, not the causal genes.
  - Possible bias with degree: the degree of genes. Ex. for damage response, the master regulator MSN2/4 may be connected to many target genes. If MSN2/4 is used, then other (specific) TFs may not be found.
  - Possible bias with redundancy: suppose two TFs combinatorially regulate a bunch of genes, and both are linked to the same kinase, then only one TF will be chosen.
  - Other types of data/networks: metabolic networks and secondary messangers, protein phosphorylation, methylation and acetylation.

Guilt by Association Is the Exception Rather Than the Rule in Gene Networks [Gillis & Pavlidis, PLCB, 2012]:

- Goal: GBA prediction of gene functions. Ask which connections in the networks are necessary and which connections are sufficient to generate function prediction performance.
- Terminology:
  - Functionally relevant edge: is a network edge that connects two genes that share a function.
  - Critical edge: the effect removing an edge has on prediction performance.
  - Exceptional edge: a critical edge for many functional categories.
- Evaluation methods: use simple voting method for any test gene: co-expression with the seed genes vs. co-expression with the rest. Cross-validation.
- Results: A small number of edges among genes sharing a lot of GOs encode most functional information: in mouse, a network constructed with just 100 edges among pairs of genes which share the largest number of GO terms yields an MAP across GO terms of about 0.09. Using the complete network: MAP 0.047, can be matched with a network of only 23 edges among 45 genes.
- Functional information is not distributed throughout the network: in yeast, use multiple networks (YeastNet, PPI). Average precision of the entire network across all GO categories: about 0.1. Adding edges by their average degree of criticality across all GO groups (their exceptionality), quickly improves the performance above that of the full network, at about 300 edges. Significantly higher at 1,000 edges, MAP 0.15.
- Edges involving genes with high node degree (hubs) are less likely to be critical: pruning the network by privileging connections on low node degree genes, yields a network that, even with 1/2 of connections removed, performs similarly to the original network.
- The power law node degree structure in this network was preferentially encoded in connections that contain no known functional information (removing these nodes, no longer scale-free).

Understanding multicellular function and disease with human tissue-specific networks [Greene & Troyanskaya, NG, 2015]

- Intuition: we have a large set of network features (how two genes are related to each other). To construct tissue-specific networks, we need training data for specific tissues to learn how each feature should be weighted. Ex. if we want to learn network in cortex, the co-expression and PPI in cortex, as well as related tissues, should be weighted more. Also a key idea is that, tissues are hierarchical, so other tissues carry information of one tissue at various degrees.

- Creating gold standard data for specific tissues: use genes of the same GO, but also both need to be expressed in a tissue. Use only tissue-specific interactions. Otherwise, we will not be able to learn the weights of features correctly.
- Reconstruction of tissue-specific functional network: use features 980 co-expression, 4 PPI, motif sharing and miRNA targets. Training with Naive Bayes.
- Application in studying tissue-specific function of genes, eg. LEF1, which changes neighbors in different tissues.
- NetWAS: use connectivity with other nodes as features, learn a SVM using known genes. Training data: genes from GWAS with VEGAS analysis.
- Remark: theoretically, if we know the true co-expression of two genes in a tissue, then that's the only co-expression feature we would need. However, this number may not be known, and co-expression in other tissues can help us learn co-expression in this tissue. This may be the key idea of this paper. If this is the case, we can formulate this as the problem of learning co-expression by combining data from multiple tissues.

1.5.3 Network Motifs

Network motifs: theory and experimental approaches [Alon, NRG, 2007]

- Composite network motifs involving both transcriptional and physical interactions: a common one is a negative-feedback loop between two proteins. One arm is a transcriptional interaction and the other arm is a protein-protein interaction. Ex. p53 and Mdm2 loop involved in monitoring stresses and DNA damage in human cells. Composite negative-feedback loops seem to be much more common than purely transcriptional negative-feedback loops.
- Signaling network motifs: Diamonds combine to form multi-layer perceptron motifs that are composed of three or more layers of signalling proteins. Such patterns can potentially carry out elaborate functions on multiple input signals, including generalization of information from partial signals. They also can show graceful degradation of performance upon loss of components.

Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network [Zhang & Roth, J Biol, 2005]:

- Methods: integrated network with five types of interactions: S (synthetic lethal), P (PPI), R (regulatory), X (coexpression), H (homology).
- Three node motifs: FFL with R; a gene regulated by two TFs (P or H); one TF regulates two genes (P or X or H). The other motifs are related to multiple P/X edges, or involve S.
- Four node motifs: compensatory complexes/processes. Two complexes with S edges between them.

1.6 Machine Learning in Genomics

General lessons on deep learning in genomics [personal notes]

- Background: how CNN works? Interlaced convolution and pooling layer. Convolution: feature detection; pooling: feature presence in a region, achieving some "scale invariance". Then after convolution and pooling, we can treat the output of pooling as a higher-level feature (whether some combination of low-level features is present in a larger region), and repeat.
- Design of DNN: relating to biology. Think of what are important sequence features that determine the function.

  - Ex. in protein contact prediction, the residuals and their structure context (e.g. alpha helix or beta sheet) are important, but not the loop regions (both sequences and lengths can vary). The DNN should be able to capture these properties (e.g. invariance of length of loop regions).
  - Ex. in predicting m6A, While the nearby sequences are important, distal intronic sequences may also play a role.

- Feature detection is not limited to simple sequence features such as motifs. We can also use physical features derived from sequences.

- Hyper-parameters of DNN: e.g. receptive field (filter size), depth of networks, should depend on both biological considerations and training data. Avoid overfitting.

- Remark: in CNN, pooling may lose spatial information. Ex. the output of pooling tells if a motif is present in a larger window, but not its specific location.

- Remark: more generally, how can we encode more biology into DNN? Ex. we know that what matters is certain sequence elements (e.g. alpha-helix), but their spacer sequences (loop regions) do not matter.

A primer on deep learning in genomics [Zou, NG, 2018]

- Practical advice on deep learning: (1) Training set: avoid bias. Eg. to predict pathogenic variants, avoid confounder such as location to genes. (2) Good to compare with simpler machine learning models, e.g. SVM.

- Interpreting deep learning: (1) Derivative: mimic in silico mutagensis of input, however, only infiniestimal changes. (2) To overcome the problem, methods exist to find features that explain how prediction is made.

- Resources: clouds and software libraries.

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. [Alipanahi and Frey, Nat. Biotechnol. 2015]

- Background: RNAcompete data have only weak RSS.

- Method: input sequence (14-100nt) and binding score (binary or continuous). Convolution with motif detector, then recitification stage, and pooling, then fed into a neuron network. Train on each data separately.

- Results: 927 deepBind models for 500 TFs and 200 RBPs. Evaluation with in vitro data: better fit of PWM and RNAcompete data. Also evaluation of in vivo CLIP-seq data: significantly better than Ray et al.

- Generally more difficult to predict RBP sequence specificity than TF.

Predicting the clinical impact of human mutation with deep neural networks (primateAI) [Sundaram and Farh, NG, 2018]

- Training data: common variants of 24 chimp genomes (2 or more copies) - total of 300K missense variants. Most would be benign: validated via dN/dS.

- DNN: human and primate common variants as negative; variants absent in ExAC as positive. Sequence alignment (PWM) as features, and separate networks to predict the secondary structure and solvent accessibility from the sequence alone. The total depth: 36 convolution layers with 400K parameters.

- Application to DDD data: use a threshold of 0.8, de novo missense mutation burden increases from 1.5 to 2.2 fold, close to LOF (2.5).

Convolutional neural network model to predict causal risk factors that share complex regulatory features [Sung and Choi, review for NAR, 2018]

- Method: use associated blocks from GWAS as training data, build a predictive model of blocks. The features are epigenomic and gene annotations for each SNP. To predict blocks, use the worst SNP (pathogenic) in a block.
- Annotations: Roadmap and ENCODE. Gene annotations: 301 pathways from KEGG. TFBS motif matches.
- CNN architecture: Input: m annotations for each SNP. Layer 1: K pattern detectors, each can be thought of capturing one aspect of SNP function, e.g. disruption of a pathway; or regulatory activity in a cell type. Layer 2: pathogenicity of a SNP based on K features in Layer 1. Layer 3: max pooling of all SNPs.
- Model training: association blocks from GWAS (7 phenotypes), and shuffled blocks (by features) as controls. For each block, only choose lead SNPs and 30 nearby SNPs with highest association statistics. Minimize binomial log-likelihood: prediction (0 to 1) vs. actual labels. To avoid overfitting: early stopping and pre-training by an autoencoder. Found that using full model is better than linear model: AUC 3-6% better.
- Evaluating feature importance: use model prediction as labels, and train a random forest to predict labels. Then the importance of a feature to the RF performance is assessed. Found that in mental/immune diseases, neural/immune features are most important. Also found pathway enrichment.
- Validation of findings: enrichment of high-scoring variants in various features, including epigenomic data (e.g. BLUEPRINT) not used in training, in eQTL, in DHS footprints.
- Examples (Figure 2 and 5): statistically indistinguishable variants, only one has strong DNN prediction score.
- Lesson: use DNN to learn nonlinear combination of features (even when there is no obvious spatial component, in this case, since adjacent SNPs do not capture any meaningful feature).
- Lesson: to understand the importance of a feature in a DNN, use DNN predictions as labels, and train another (simpler) classifier to predict the labels.
- Remark: training is done with about 300-400 blocks and 2000 features. It is quite striking that DNN works with this few training samples. Also training data is probably noisy: 30 SNPs per blocks, most of which are probably not causal variants.
- Remark: Model architecture: shared layers among phenotypes may help improving the prediction. Ex. Layer 1 could be the same (a large number of parameters). Layer 2: pathogenicity for particular phenotypes.

Protein structure prediction by deep learning [Jinbo Xu, 2019]

- Fragment assembly: get fragment structure from PDB, then assembly with objective function: either data-based or energy based.
- Contact/distance based method. Co-evolution: usually need a large number of sequences (e.g. a few hundred). Define contact and distance matrix: for AA, use C-beta.
- Contact prediction: co-evlution from MSA; or supervised methods (sequence to known structure).
- Direct coupling analysis: a set of direct contact that explain all observed correlation patterns.

- MRF: the edge function depends on specific pairs (function defined on residual pairs). Related to precision matrix problem. Encoding AA: 21 dim. vector. Reduce to Gaussian Graphical model problem. To estimate: L1 penalty.
- PSICOV: fit correlation patterns (co-evolution), then predict contacts. Works better than MI. To predict contacts: norm of correlation among 21 types. Note: sequence weighting to deal with phylogeny among data.
- Deep Convolutional Residual Neural Network for contact prediction, PLCB, 2017. Even trained for soluble proteins, it can work well for membrane proteins, protein complex prediction (contacts in 2 proteins), and domain-contact prediction.
- Convolution: use kernel to scan a part to detect if a patten exists. Use small kernels, and many layers. Residual network: further improve.
- Training: 10K proteins in PDB. Train contact matrix (training data) as image, and each residual pairs as a pixel, pixel level labeling.
- Input: from MSA (1) Sequential features: local structure, frequency, (2) pairwise features: co-evol, MI (small number). Output: labels from true structure.
- Analysis: naive method is to predict labels from local information. But its important to consider global information: predict all labels together. Intuition: may learn AA is in a particular domain/secondary structure, and use the information to predict interaction.
- Extensions: (1) Application to folding: min. energy function. (2) Distance prediction: labels to distance bins.
- Remark: the difficulty of the problem is how to generalize across proteins (protein lengths are different).
- Remark: how do we interpret the network: e.g domains.
- Lesson: in label prediction problem, important to consider the global contexts and all labels together. In CNN, one can use sequence, but also features derived from sequences.

1.7 Text Mining in Biomedical Literature

1.7.1 Statistical Text Mining and Information Retrieval

1. Genes, themes and microarrays [Shatkay & Boguski, ISMB, 2000]

   Problem: (i) for any gene, find its relevant documents and representative terms - themes of this gene; (ii) cluster genes by the similarity of their documents.

   Methods:

   (a) Find relevant documents and representative terms for each gene: the theme is representated by a language model (the probability whethere a term will be used in a document of this theme), and each document in the collection is either generated from this theme or not; if it is theme-related, then any of its terms will be generated by the theme or background (with some probability) $\Rightarrow$ the relevant document set is called the "kernel document" of this gene.

   (b) Gene clustering: the similarity between 2 genes is determined by the similarity between the kernels of the 2 genes. Specifically, a kernel document is represented by a vector (whether a document is in this kernel), and cos. similarity between the 2 vectors is used

2. Keyword extraction for gene lists [Masys & Corbeil, Bioinformatics, 2001]

   Goal: extract signficant terms from annotations of a query gene list.

   Methods:

   (a) Data: AML vs ALL, differentially expressed genes

   Results:

   (a) Common terms: complement activation
   (b) ALL-specific: immunodeficiency, multiple sclerosis, etc.
   (c) AML-specific: ...

3. GEISHA [Blaschke & Valencia, Funct Integr Genomics, 2001]

   Goal: significant terms from literature for a query gene list.

   Methods:

   (a) Data: Eisen clusters (Spellman cell cycle)
   (b) The analysis on each cluster in Eisen dataset: significant terms, informative sentences global as well as case analysis.
   (c) Term clustering, comparion between different clusters

4. Neighbor divergence score for gene group [Raychaudhuri & Altman, GR, 2002]

   Goal: a scheme that scores whether a group of genes are "functionally coherent" based on literature.

   Methods:

   (a) Gene document association: the literature text.
   (b) Score group coherence:
      - Article neighbors: defined based on word usage of articles.
      - Article relevance to a gene group: scored by counting the number of neighbors that have references to genes in the group.
      - Neighbor divergence: determines whether a function is represented in a gene group from the distribution of article scores.

   Results:

   (a) Define 19 gene groups with coherent (known) functions and 1900 random gene groups: classify the groups using the neighbor divergence scores.

5. Literature profiling [Chaussabel & Sher, GB, 2002]

   Goal: create the literature profiles of genes, cluster genes by their profiles and identify terms related to clusters.

   Methods:

   (a) Term selection: choose terms to create the literature profile of genes. The occurrence of a term in a gene (a set of documents) is the percent of documents this term occurs. Criteria:
      - Baseline occurrence must be small: occur in $< 5\%$ of genes
      - Higher occurrence (25% higher than baseline occurrence) in at least 2 genes in the input list
      - For large gene list, remove terms that are present in 50% of genes
   (b) Biclustering of gene-term matrix: simultaneous hierarchicial clustering of genes and terms; visualization; and the highlighted blocks (of both genes and terms) are those of interest.

6. LACK [Kim & Falkow, BMC Bioinformatics, 2003]

   Goal: signficant terms from annotations for a query gene list.

   Methods:

   (a) Input: a query gene list; the background gene list (all genes); a set of terms to be tested. In addition, the annotation of all genes need to be provided.

   (b) Test: for each term, test if its appearence in the query gene list, defined by the number of genes containing this term in their annotations, is significantly higher than that in the background list, via binomial or Poisson test.

   (c) Data: microarray data: genes in some microorganism upregulated by ferrous iron chelator.

   Results: identify some terms (Salmonella Pathogenicity Island 2, SPI2) overrepresented in the 256 genes; confirmed from other sources.

   Remark: suppose one knows the source of input gene list (e.g. those upregulated by X), then the signficant terms could potentially be related to X, this could lead to new biological hypothesis

7. EASE/DAVID [Hosack, Genome Biology, 2003]

   Methods:

   (a) Allow many/customizable classficiation systems, including GO, SWISS-PROT, etc. (any type of gene-concept assoication)

   (b) Statistics: modified Fisher exact test (jackknifing for significance)

   Results:

   (a) 3 microarray gene lists: compared with manually annotated themes, similar results.

   (b) DAVID system: functional annotation (enriched themes), gene functional classification (group genes by their annotations), clustering redundent annotation terms, display gene-term matrix, redirect to related literature, gene name exploration (e.g.ID conversion), etc.

8. TXTGate [Glenisson & De Moor, Genome Biology, 2004]

   Goal: signficant terms from multiple resources for a query gene list.

   Idea: choose different vocabuloary to annotate the gene list - different views of the same list.

   Methods:

   (a) Multiple domain vocabularies: GO (molecular biology), OMIM (medicine), MeSH (medicine), eVOC (cell types, anatomy, developmental stage, etc.); and gene centric vocabularies (HUGO, SGD).

   (b) Each gene is represented by a keyword profile (vector-space representation). Defined as the mean profile of all documents of this gene. Each document is represented by a weight vector: normalized IDF.

   (c) The profile of a group of genes := mean profile of all member genes.

   (d) Gene subclustering: based on vector-space representation of genes (using the terms in vocabuloaries)

   Results:

   (a) Yeast data

   (b) Human data: 200 genes involved in human macrophage activation upon bacterial infection

(c) Genes involved in colon and colorectal cancel (Additional Data File3). The results based on different vocabularies are very different, ex. GO vs eVOC

9. MILANO [Rubinstein, BMC Bioinformatics, 2005]

   Problem: annotate a gene list.

   Methods:

   (a) Input: a query gene list; a user specified term list.

   (b) Procedure: search gene-term pair in two databases: GeneRIF and PubMed and display the result matrix (documents which contain gene-term pairs). Note that the gene search allows gene name expansion (using LocusLink aliases) and filters the noninformative names (≤ 3 letters or English words)

   (c) Data: 148 genes affected by overexpression of p53

   (d) Evaluation: whether the system could recover the known (about 60) p53 target genes

   Remark: no statistical analysis to extract overrepresented themes

10. MeSHer [Djebbari & Quackenbush, Bioinformatics, 2005]

    Goal: term annotation of gene list

    Methods:

    (a) Data: cardiac response in a mouse model of hypertension to angiotensis II treatment [Larkin, 2004]. Top 15 up and down-regulated genes.

    (b) Resource: gene - MeSH term association from literature

    (c) Statistcis: Fisher's exact test

    Results:

    (a) Comparison with EASE analysis: many of the most significant MeSH terms were similar or related to concepts found previously

11. Associative concept sapce (ACS) [Jelier & Kors, Bioinfo, 2005]

    Goal: identify the relations between genes - group genes that share similar literature content.

    Methods:

    (a) ACS: a space of concepts where the distance between concepts is defined based on co-occurrence patterns of the 2 concepts. The pattern consists of single-step relation (cooccurrence between X and Y), two-step relation (co-occurrence between X and Z, between Z and Y), and multiple-step relations. Concepts that are connected by many paths are close in ACS.

    (b) Use ACS for gene clustering/measuring similarity between genes: the distance between 2 genes is the distance between the 2 gene concepts in ACS (the gene symbol itself is a concept)

12. LSI for gene clustering [Homayouni & Berry, Bioinfo, 2005]

    Goal: clustering of genes by their similar literature content.

    Methods:

    (a) Term selection: terms that appear less than twice in a gene-document and in less than two gene-documents were not included

    (b) Construct term-by-gene matrix: the weighting of a term to a gene (a document set) could be any scheme used for IR. In this application, use a log-entropy weighting scheme. The weight of term i in gene j, $m_{ij} = l_{ij}g_i$, where l_{ij} is the local component, defined via term frequency; and g_i is the global component, similar to IDF.

(c) LSI: do SVD of the term-gene matrix $\Rightarrow$ r eigenvectors

(d) Ranking a query document: rank-s approximation (use only s instead of all r eigenvectors) to compute the similarity of 2 documents

13. ConceptMaker [Kuffner & Zimmer, Bioinfo, 2005]

    Goal: gene clusters that exhibit both a significant gene expression as well as a coherent literature profile.

    Methods:

    (a) Construct term-gene matrix: by a variant of TF-IDF weighting scheme

    (b) LSI (SVD projection): under LSI space, could define a profile representing a set of objects; And the distance between two objects (or two profiles or one object and one profile) could be determined using cosine similarity between the two.

    (c) Find gene clusters: coherent expressions and literature (score of the cluster in LSI space)

    (d) Ranking of terms (or documents) wrt a gene cluster: similarity between a term and the cluster profile using cosine score

14. Non-negative matrix factorization (NMF) [Chagoyen & Pascual-Montano, BMC Bioinfo, 2006]

    Goal: cluster genes as well as provide literature justifications of the clusters (what terms are shared in a cluster).

    Methods:

    (a) Data set: (i) gene-document construction: yeast genes with documents from SGD (pre-generated), then concatenate all documents of a gene into a single gene-document. (ii) term selection: a term is filtered out if it didn't appear in at least 4% of genes, or more than 80% of the genes.

    (b) Gene-term matrix: weight of a term wrt a document is determined using TF-IDF scheme $\Rightarrow$ p terms and n genes, p by n matrix

    (c) NMF: the gene-term matrix $V \approx WH$ where W is the semantic features/themes (linear combination of terms - new basis vector) and H is the representation of genes in semantic features. Ex. a gene is rich in semantic features 2 and 3, but low in all other features.

    (d) Gene-document clustering: agglomerative hierarchical clustering using half-square Euclidean distance as a similiarty measure

15. Anni [Jelier & Kors, BMC Bioinformatics, 2007]

    Methods:

    (a) Construct text-based profile for each gene: (i) thesaurus of biological concepts: MeSH + gene names; (ii) define assoication between a gene and a concept (in fact, any two concepts): LRT where H0 = two concepts are indepedent.

    (b) Gene clustering based on text profile: cosine similarity measure; hierarachical clustering; coherence measure of a gene cluster and p-value.

    (c) Selection of concepts for a gene group: any concept is scored by percentage contribution to the average cosine score for the group. Cutoff: $> 0.5\%$.

    Result:

    (a) Control test set: small gene groups with known function + unrelated genes $\Rightarrow$ test if cluster correctly.

(b) Genes differentially expressed following stimulation of the androgen receptor in a prostate cancer cell line ⇒ tightest cluster contains 4 genes associated with secretory lysosomes. Propose a hypothesis: secretory lysosomes are involved in the production of prostatic fluid and that their development and/or secretion are androgen-regulated processes.

16. Anni 2.0 [Jelier & Kors, GB, 2008]

    Goal: a system that mines related concepts: concepts that co-occur with given concepts; similar concepts; etc. In the case of genes, one application is: find concepts that characterize genes; and find genes with similar annotations.

    Methods:

    (a) Concept system: genes; UMLS - including GO, OMIM, etc.

    (b) Concept recognition in documents: for genes, use the tool Peregrine.

    (c) Building the concept profiles of some concepts (e.g. of genes).

    (d) System: supports these functions:

    - Find related concepts of a query concept: the concepts in the profile of the query that have high weigths.
    - Find similar concepts of a query concept: inner product of concepts profiles as the measure of similarity.
    - Clustering of concepts: e.g. gene clustering.

    For all these functions, the supporting documents can be extracted and analyzed.

    Results:

    (a) Case 1. the analysis of genes differentially expressed in cancer patients and healthy people. Hierarchical clustering of the genes; followed by analysis of the concepts of the the leading cluster(s).

    (b) Case 2. literature based knowledge discovery - find the diseases that may be cured by a given drug. Start with the drug name: (i) find the related concepts: IL-12, etc.; (ii) find the diseases related to these concepts.

17. LAMA [Jelier & Mons, BMC Bioinfo, 2008]

    Goal: meta-analysis of multiple studies (where each study identifies a set of differentially expressed genes) - the connections among these studies.

    Methods:

    (a) Gene assoication: each gene is represented by a concept profiles; and the association between two genes is measured by the similarity between the two profiles [Anni 2.0].

    (b) Association between two studies (two gene lists): defined as the number of gene assoications between two lists.

    Results: when gene overlapping or GO overrepresentation (find enriched GOs for each list and then count common GO terms) is used, the connection among difference lists is small. Significantly more commonailities are revealed when the literature-derived analysis is used.

1.7.2 Information Extraction

1. Text mining and its applications [Rzhetsky & Gerstein, Cell, 2008]

   Applications:

   (a) Consistency of data: check if multiple facts are consistent with each other. Ex. “A inhibits B”, “B inhibits A” and “A and B are both active simultaneously.”

(b) Charting the development of sciences: study the ways in which scientists develop and transmit ideas, and collaborate with one another. Ex. how different fields are related.

2. Extracting PPI by regular patterns [Ono & Takagi, Bioinfo, 2001]

   Methods:

   (a) Process compound or complex sentences: two rules:

   - P1 VB1 P2 CC VB2 P3 → P1 VB1 P2; P1 VB2 P3: P1, P2, P3 are proteins, VB1, VB2 are verbs and CC is coordinating conjunction. Ex. STD1 interacts directly with the TBP and modulates transcription of the SUC2 gene → STD1 interacts directly with the TBP; STD1 modulates transcription of the SUC2 gene.

   (b) Recognition of PPI: by regular patterns, e.g.

   - A interact with B: Spc97p interacts with Spc98 and Tub1 in the two-hybrid system.
   - A bind B: the N-terminal of SIN1 is sufficient to bind SAP1.

   Note that it is allowed to have words between pattern elements (entities, verbs, etc).

3. Textpresso [Muller & Sternberg, PLoS Biol, 2004]

   Idea: define entities in the text, and also terms that may suggest relations (e.g. regulatory relation verbs), then text retrieval with these entities and relational terms, e.g. find a sentence that contains both a gene name and a cell type (this gene is thus likely to be located in this cell). Effectively, co-occurrence based information extraction.

   Methods:

   (a) Ontology: two main categories:

   - Biologial entities: genes, transgenes, biological process, molecular function (e.g. DNA helicase, this is different from specific genes), cell types, experimental methods, phenotypes, cellular component, life stage, drugs and small molecules, etc.
   - Relational terms: regulation (enhancer, suppress), association (bind)

   The words in documents are labeled with these categories using regular expressions. Ex. [Ii]nteract(s—ed—ing); gene names: [A-Za-z][a-z][a-z]-
   d+.

   (b) Interface: a user specifies the exact terms (allow Boolean combinations) and categories, and the sentences matching the terms in the query and all categories specified will be output.

   (c) Advanced interface: use query language, consist of several parts: 1) set parameter values (the scope of search, literature, full text or not, etc.); 2) the actual query: search keyword or category, constraint (number of times it must occur), and the results are stored in a variable; 3) the combination of query results (variables) using AND, OR, NOT.

   Results:

   (a) Example query: find in which cells lin-11 is expressed, could be achieved by the query, "lin-11" AND cell_type.

   (b) Evaluation: using the example of genetic interactions.

   - Small scale evaluation: 8 articles, precision about 30%, and recall 62% (not all sentences use the terms defined by Textpresso).
   - Large scale evaluation: precision by judging the correctness of 200 randomly chosen sentences retreived, and recall by estimating the number of sentences discussing genetic interactions (from 200 randomly chosen sentences).

4. Extracting PPI by optimizing syntax patterns via genetic programming [Plake & Leser, ACM Symposium on Applied Computing 2005]

   Methods:

   (a) Components of patterns: proteins, INouns, IVerbs, fixed words (such as certain prepositions) and word gaps.
   - INouns: binding, interaction, regulation, modulation, etc.
   - IVerbs: activate, control, regulate, stimulate, etc.

   (b) 22 syntax patterns: e.g. INoun of Word*(P1) ProteinA(E1) Word*(P2) [by | through] Word*(P3) ProteinB(E2). An example sentence: "... the minimal requirements for induction of PEPCK by PKA and inhibition by insulin ...". P1, P2 and P3 are parameters of gap length; and E1, E2 are Boolean parameters indicating whether protein names can be compound.

   (c) Parameter learning: maximizing the performance of parameter set (parameters of all patterns) via genetic algorithm.

5. Extracting regulatory relations [Saric & Bork, Bioinfo, 2006]

   Goal: extracting regulatory relations from literature.

   Methods:

   (a) Task definition: both transcriptional and translational regulation. Un-specified relation such as "A activates B" will not be counted.

   (b) Corpora: about 50K abstracts of yeast in MedLine, and also abstract for E. coli, B. subtilis and mouse. For yeast, 9137 and 6640 abstracts are chosen for training and evaluation, respectively.

   (c) POS tagging: both syntactical (e.g. determinant - dt, adjective - jj) and semantic taggs. The later includes, e.g. organisms (org), gene or protein symbols (nnpg) from an external lexicon.

   (d) Entity recognition: recognize entities (names) as part of noun phrases. Rule-based approach. Ex. this is one rule for the entity kinase:
   nx_kinase → dt nnpg jj kinase in org
   which matches this phrase "the ArcB sensory kinase in E. coli".

   (e) Relation extraction: rules are defined to match sentences describing regulatory and (de)phosphorylation relationship. Ex. this is one rule for acviation of expression in passive voice:
   expression_activation_passive → nx_expr induced by nx_gene
   which matches "IL-13 expression induced by IL-8".

   Results:

   (a) Entity recognition: very high accuracy 95% is achieved. The errors often come from phrases such sa "telemorase associated proteins" (which is confused with "telemorase protein").

   (b) Relation extraction: about 400 relation instances in the yeast collection. Accuacy: 83% in the evaluation corpus (75 correct in predicted 90 relations). Recall: check 250 out of 44,354 sentences containing two gene names, found 8 relation instances, thus recall is about 30%.

6. PLAN2L: information extraction system for plant [Krallinger & Valencia, NAR, 2009]

   Goal: a system that automatically extract relations and other facts from literature.

   System functions:

   (a) Retrieving sentences covering certain aspects of a gene: four aspects are supported, flowering, root development, leaf development and seed development. Similar to GeneRIF.

   (b) Finding relations a gene is involved in: regulatory relation, PPI, and subcellular localization.

(c) Finding associations between two genes: output sentences mentioning both.

Methods:

(a) Corpus: documents about Arabidopsis, both abstracts and PubMed Central full text.

(b) Gene name recognition: use a dictionary to identify gene names. For disambiguation and scoring the reliability of a given entity normalization, we calculated the document similarity between the context of mention and the corresponding database record.

(c) Gene regulation: STRING-IE system [Saric, Bioinfo, 2006]

(d) PPI: Support Vector Machines algorithm trained on set of manually classified interaction evidence passages derived from a collection used at the second BioCreative challenge.

(e) Subcellular localization: location terms from SwissProt and GO. A location sentence classifier was constructed using a collection of 2264 protein location description sentences.

(f) Cellular and developmental processes: sentence classifiers using SVM.

Chapter 2

Gene Expression Data Analysis

Goal: the central theme of expression data analysis is: from the phenotypes (cell types, stages, strains/individuals/etc.), identify the genes whose expression patterns define/characterize the phenotypes or the transcriptional program of a process/phenotype; and for these genes, identify their regulatory mechanisms: cis-regulatory sequences, trans-regulators, etc.

Principles of expression data analysis:

- Biological foundations:
  - Cells change gene expression to achieve functions (including response to environment, proliferation and differentation, etc). And similarly, the expression of genes are changed in pertubations, including disease.
  - Regulatory programs of cells are designed by evolution s.t. functionally similar/related genes are co-regulated.
- Analytic strategies:
  - A complex transcription program (e.g. in response to some environmental perturbation) can be decomposed into many modules/subprocesses/pathways.
  - Expression profiles are representations of the biological states/processes, thus they are associated with phenotypes (cell types, disease states, etc.). This principle can be applied to identify candidate genes, predict/identify phenotypes, and in genetic analysis as surrogate of phenotypes. Particluarly useful on expression profiles of gene groups (expression signatures), which represent states of biological pathways/modules.
  - Genes with similar/related functions tend to share regulatory mechanisms and as a result, show similar expression patterns.
  - Expression can affected by multiple factors such as environments, phenotypes, circadian rhythm, experimental conditions. Important to consider these additional factors: confounders, or latent information.

Experimental design and data analysis: gene expression (a quantitative trait) may be influenced by many factors, biological (strains, treatment) as well experimental (batch effect, microarray platform), thus design the experiment and analyze data accordingly to reveal the factor of most interest.

- Quality control: the first step of any expression data analysis. This may involve inspections at both level of genes and samples:
  - Gene level: e.g. housingkeeping genes (significant expression level in many conditions), marker genes for specific conditions/samples, etc.

  - Sample level: the samples are clustered according to some meaningful biological criteria, e.g. samples from the same tissue origin are clustered. Note that for sample level analysis: since samples often have very large number of variables, one can first use PCA, then do clustering using PCs (or plotting the samples along PCs).

  The same idea for QC can be applied for other genomic data, e.g. epigenetic modifications.

- Normalization: across different experiments of expression (different samples), there are often systematic bais, e.g. the gene expression in one sample are consistently higher than another. One thus need normalization s.t. the samples are comparable. A common approach: quantile normalization to match two distributions (assuming one is a reference distribution), and the results are expression in the reference distribution.

  - Remark: only in special cases, one needs to normalize by genes (different genes follow the same distribution), e.g. in clustering. Gene normalization would lose information: different genes are expected to express at different levels.

- Differential expression across two conditions/samples: control samples should be identical to the test sample in all aspects except the one that is being looked at, e.g. treatment.

  - Simple data analysis: fold change; if there are enough replicates: 2-sample t-test.
  - Multiple regression and ANOVA: when there are multiple influencing factors, e.g. sex, environment and genetic background, use regressino or ANOVA to analyze the effect of each individual factor.

- Hidden confounders: often, there may be hidden confounders that are not corrected, e.g. diet, environment, etc. that are different from cases and controls. To control for these confounders, in practice, techniques such as PCA (surrogate variables) are used and PCs are controlled. However, these may remove the true biological signals, as genes in biological pathways tend to be correlated but the covariance is removed by the PCs.

  - Ex. RNA-seq data of Scz. brains and controls, when one uses all significant surrogate variables (36 SVs), no gene is diff. expressed, and it was found that no pair of genes has $r > 0.7$. Only at 9 SVs, a significant number of diff. expressed genes.

- Random effects: in addition to the confounders (both known and hidden ones), there may be random effects that affect expression. Ex. when analyzing expression data of multiple tissues, each with multiple individuals. The tissue has a fixed effect, each individual may contribute a random effect: genetic background, environment that influence expression of all tissues in an individual, but differ between individuals. To correct this, use random effect model, where the random effect term is learned from all samples from the same individual.

Finding gene relationship from expression data:

- Gene modules/groups: if the pattern of gene X is correlated with the pattern of gene Y, then X and Y may share the same function. Examples:

  - Genes whose expression pattern correlated with known cell-cycle genes are likely to be involved in cell cycle.
  - Gene x is a known target of TF F, and y has a similar expression pattern as x, then y is likely to be a target of F as well.

  The module structure can be found via clustering, dimensionality reduction, etc.

  - Network analysis: given a gene group, it is possible to extrapolate for more related genes from a gene network (e.g. PPI network).

  - Biclustering analysis: a set of genes may have correlated patterns across some (but not all) conditions/phenotypes. To reveal biclusters, one could order the matrix in a way to reveal the sub-matrix patterns.

- Transcriptional relationship: when combining with other data, it may be possible to infer regulatory relationship among genes. See "Gene Regulatory Networks".

- Hidden regulatory processes: e.g. signaling pathways.

- Summarizing gene groups/modules:

  - Enrichment analysis: e.g. using GO.
  - Intra-module relationship: may be informative, e.g. the hubs of a co-expression module.
  - Network status: what position in the network a module occupies, the most related genes in the network.

Transcriptional program of a process or phenotype: given expression data of a process or multiple conditions/phenotypes, infer the transcriptional program underlying the changes or the differences at the cellular/organism level:

- Association with phenotype: if the expression pattern of a gene X is correlated with the phenotype or stages of a process P, then X is likely to be a candidate, or part of the transcriptional program, for P. The most important application is: genes that are differentially expressed in different conditions/phenotypes. Examples:

  - To find cell-cycle genes, find genes that have cyclic expression pattern (correlated with phenotypic changes).
  - To find candidate genes for some disease, find genes that discriminate disease and normal cells.
  - To find genes for the development of an organ, correlation the spatialtemporal pattern of genes with that organ: the genes that start to express at the time of organgenesis, have gradident distribution centered at the organ precursor cells, etc.
  - Transcriptional program of cell cycle: genes are activated/repressed in different stages (characteristic expression patterns), corresponding to cell cycle phases.

  In many cases, the analysis solves the problem of classification of expression profiles.

- Comparison of expression profiles: when there is structure in the expression data (spatial, temporal, etc.), then the comparison of relative expression level (or pattern of expression) across mulitple conditions/samples, in cases vs. controls (or other kind of labels), may be informative.

  - Example: comparison of transcriptomics of autism vs. control: the difference lies in the relative expression pattern of genes, some region-limited genes become undifferentiable in different regions.

- Comparison of co-expression structure: the co-expression structure may capture the characteristic transcriptional programs of certain processes/phenotypes. Comparison of co-expression structure:

  - Similar co-expression in different conditions/species: conserved responses
  - Different co-expression in different conditions/species: divergent/adapative transcription in the evolutionary context; discriminative of different phenotypes/processes. Ex. different co-expression networks in different regions of brain.

- Gene group-based analysis: the association of modules/groups with phenotypes. Modules could be defined a prior from other sources; however, gene groups related to a specific condition are better inferred from the samples of the correct condition. Ex, if a large number of samples (genetic or other variations) are available, one could use cluster analysis to infer the groups.

Gene group and expression signature-based analysis: [Nevins & Potti, NRG, 2007]. A phenotype (process) is characterized by not individual genes, but by global expression patterns/profiles of gene groups, or signatures. In other words, gene groups are compact representations of processes/biological states.

- Examples:
  - How cell cycle is progressed in terms of underlying processes such as DNA replication, mitosis, etc. For instance: (i) histone cluster is on during S phase → histone proteins are created in the S phase for chromosome replication; (ii) MET cluster is on during S phase → methionine biosynthesis in S phase, methionine may be the limiting resource of yeast cell cycle.
  - Cancer: expression signatures represent the state of pathway activation/deactivation (or disregulation).
- Expression signatures can be used for predictive purposes: e.g. learn signatures of cancer gene expression profiles and then use them to predict cancer.
- Expression signatures are sub-phenotypes of cells/individuals: this allow it to be used for genetic analysis - identification of sequence mutations associated with signatures.

Phenotype-phenotype relation: expression pattern is a "representation" of a phenotype, thus the relations among phenotypes/samples could be stuided via the expression patterns of the corresponding phenotypes/samples. Application: define the distance among samples, so that one could cluster samples/find similar samples/etc. Ex. two drug treatments produce similar expression responses, then they may act via similar mechanisms of action).

Remark:

- The "association principle" (both gene-gene and gene-phenotype) can be applied in both supervised and unsupervised framework.
- The candidate genes found throught correlation with phenotype: can be further filtered through additional evidence: e.g. functional class of the gene; the place where a gene is expressed.

Reference: [Nevins & Potti, NRG, 2007]

2.1 Microarray data analysis

Reference: [Allison, NRG, 2006]

Experimental design: biological and technical replicates

- Necessary to have both replicates, especially biological replicates
- To test differential expression (DE), need ≥ 5 biological replicates
- mRNA pooling from multiple biological samples is necessary when testing DE (better than choose only one sample, if not able to make one sample / array)

Preprocessing: image processing (read signal intensity), data transformation and normalization

- Often use $\log_2$ transformation (to make the data normally distributed)
- Oligonucleotide array (Affymetrix array): RMA, GCRMA
- cDNA microarray: no clear winner.
- Normalization: s.t. the distribution of expression in different samples are identical/comparable (quantile normalization is often used).

Testing of differential expression: from 2 samples, where each sample contains the data for some replicates

- Fold change: not good, ignore the variability
- Methods: t-test, ANOVA, logistic regression
- Shrinkage method: for variance estimation. Use array-wide variance estimation in addition to gene-specific varance
- FDR: to correct for multiple hypothesis testing. Mixture model estimation of FDR. Bayesian interpretation
- Gene class testing. Issues include: continuity of evidence ignored; inappropriate penalization of some gene classes
- Intersection of multiple results, e.g. genes DE in two datasets. Issues: continuity of evidence; intersection-union test
- Resampling test of significance

A comparison of normalization methods for high density oligonucleotide array data based on variance and bias [Bolstad & Speed, Bioinformatics, 2003]

- Intuition: Suppose we have a gene ranked at k-th quantile, and its expression x. Our idea is that x is not directly comparable across arrays. But we replace x by the expected value considering the fact that the gene is ranked at quantile k. This expected value is given by averaging of genes ranked at k across all arrays.
- M-A plot: M (minus) vs. A (addition) plot. For comparison between two arrays, for any gene k, let M_k = difference of log2 expression and A_k = mean of log2 expression. The scatter plot of M_k vs. A_k.
- After proper normalization: M-A plot should show a cloud around $M = 0$: no systematic difference of expression between genes in two samples. Figure: Wiki MA plot.

Use of within-array replicate spots for assessing differential expression in microarray experiments (limma) [Gordan Smith, Bioinfo, 2005]

- Background: in microarray experiments, one usually have multiple arrays/samples, and for each array, several technical replicates (spots of the same gene). Usually researchers would take average across replicates. However, this may lose information.
- Single gene model: let y_{gij} be expression of gene g in sample i and replicate j. We have:

$$y_{gij} = X_i\beta_g + \epsilon_{gij} \tag{2.1}$$

  We assume $\text{Var}\,\epsilon_{gij} = \sigma_g^2$, and for two replicates j and j', we have $\text{Cov}\,(\epsilon_{gij}, \epsilon_{gij'}) = \rho_g$. The estimator (REML) of σ_g^2 and ρ_g can be obtained from within sample variation and across-sample variations (closed form). In the DE analysis, this correlation structure is taken into account (correlated error) in linear model.
- Common correlation model: correlation structure is the same for all genes, i.e. $\rho_g = \rho$. With the estimated ρ, the estimator of variance terms and test statistics for DE would be different.
- Remark: random effect interpretation, we can also interpret the error term as: $u_{gij} + e_{gij}$, where u_{gij} captures the random variation in spot j, and e_{gij} is the true error. The u term is correlated among replicates.

Cluster analysis [D'haeseleer, NBT, 2005]:

- Distance/similarity metrics: Eucledian distance (needs scaling), Pearson correlation coefficient, angular separation all seem to perform reasonably well. Eucledian distance may be more appropriate for log ratio data, whereas Pearson correlation seems to work better for absolute-valued (e.g., Affymetrix) data.
- Popular clustering algorithms: hierarchical clustering, K-means and self-organization map (SOM).
- Hierarchical clustering: single linkage performs poorly. K-means and SOM generally outperform hierarchical clustering. And SOM may outperform K-means, especially when the number of clusters is large.
- Validation of clustering approach by resampling is necessary.

2.2 Gene Signatures and Phenotype Classification

Cancer classification [Golub & Lander, Science, 1999]

- Problem: class discovery (unknown subtypes) and prediction (of new samples) in cancer.
- Background: acute leukemia, two subtypes, ALL (from lymphoid precursors) and AML (from myeloid precursors).
- Data: 11 AML and 27 ALL samples, each of expression of 6,817 human genes, oligonucleotide array.
- Feature selection: selecting informative genes. Let expression profile of a gene be g, the class labels be c, and the correlation measure between g and c be $P(g, c)$. The correlation (informativeness) can be defined by a t-test. Let the average expression of the gene in two classes be $\mu_1(g)$ and $\mu_2(g)$ respectively, and the standard deviation be $\sigma_1(g)$ and $\sigma_2(g)$ respectively. Then $P(g, c) = (\mu_1(g) + \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$.
- Class prediction method: (traditional classification methods may not be good, see Remark) the ideas are: (1) some genes are informative: expression level indicates ALL or AML; (2) combine the evidence of multiple genes with weights reflecting how informative a gene is. The evidence (weight) of a gene on two classes can be defined as the deviation of the gene expression from the mean expression of two classes in the training data (similar to LDA), the weight is $P(g, c)$. Thus the total evidence of a gene is:
$$v_g = P(g, c)(x_g - \frac{\mu_1(g) + \mu_2(g)}{2}) \tag{2.2}$$
The total vote of AML (or ALL) is the sum of v_g for all genes with positive vote for AML (or ALL).
- Class discovery: SOM method (good for a small number of clusters).
- Results:
  - The informative genes: the genes indicating the origin of cell lineage; cell cycle progression; chromatin remodeling and transcription; cell adhesion.
  - The performance of classifier is validated by: (1) cross-validation; (2) on independent testing data with broader samples. Also show that the performance is robust to feature selection: the number of informative genes from 10-200 gives similar results (reported results at $n = 50$).
- Remark: if use common methods such as logistic regression or SVM, the number of samples (37) is much smaller than the number of parameters (thousands of genes), thus difficult to reliably estimate the model.

The Signature algorithm [Ihmels & Barkai, NG, 2002]

- Goal: the modular organization of yeast genome. The basic idea is that genes may be co-expressed in some conditions.
- Methods:
  - Data: compendium of expression profiles. Input genes for detecting modules: MIPS category, highly-clustered genes and sequences with the same motif.
  - Signature algorithm: take an input set of genes, first, find the conditions where they are co-regulated; then from these conditions, identify other genes that are co-regulated with the input genes.
- Results:
  - Validation of the modules: functional annotation, including subcellular localization and mutant phenotypes/fitness.
  - Higher-order relations of the modules: for modules relevant to the same set of conditions, check how they are related. Ex. rRNA processing module and stress-response module are inversely related; mating gene module and G1/S cell cycle module also inversely related.

A global test for groups of genes: testing association with a clinical outcome [Goeman & van Houwelingen, Bioinfo, 2004]

- Model: suppose the outcome is a function of the expression level of every gene in a group of size m. If more genes are highly expressed, we expect the outcome will be more likely to be 1 (i.e. the effects of genes within the group are additive). This can be modeled with a genarlized linear model with h the link function:
$$E(Y_i|\beta) = h^{-1}(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j) \tag{2.3}$$
The null hypothesis is: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$. Directly testing this with LRT would have very high d.f. (if m is large). Thus assume the random effect: β_i are i.i.d. from $N(0, \tau^2)$. Then the null hypothesis is: $H_0 : \tau^2 = 0$. This can be viewed as an empirical Bayesian model.
- Test statistic: suppose X is the $n \times n$ data matrix (n is the sample size), Y is the outcome vector. And let $R = (1/m)XX^T$, $\mu = h^{-1}(\alpha)$ is the expectation of Y under H_0, and μ_2, μ_4 the second and fourth central moments of Y under H_0. The test statistic is:
$$Q = \frac{(Y-\mu)^T R(Y-\mu)}{\mu_2} \tag{2.4}$$
- Interpretation of test: two ways, first, we can write Q as:
$$Q = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{\mu_2}[X_i^T(Y-\mu)]^2 \tag{2.5}$$
Note that: $Q_i = (1/\mu_2)[X_i^T(Y-\mu)]^2$ is the test statistic of the i-th gene and the outcome, so Q is simply the average of Q_i for all genes. Another interpretation is: Q has a high value when the covariance structure of the gene-expression between samples resembles the covariance structure between outcomes (i.e. samples with similar gene expression also have similar outcomes).
- p value: Q is asymptotically normally distributed. When the sample size is small, the p value can be calculated using a permutation method.
- Results:

  - AML vs ALL dataset: choose all genes as the group. The test is highly significant, suggesting that the dataset is easy to classify.
  - Heat shock dataset: if using all genes as a group, not significant. Specific groups are needed.

- Remark:
  - Model intuition: in this case, higher gene expression, or more genes expressed, both would expect to relate to outcome, thus linear model is reasonable.
  - Random effects: when a model has a large number of parameters, to avoid penalty of large d.f., assume that the effects are random. This leads to an empirical Bayesian treatment.
  - Intuition of the test statistic: (1) averaging: test of a group of objects is some kind of average of the test of each object; (2) similarity: the similarity of covariance structure of instances and responses.

Testing association of gene groups [Tian & Park, PNAS, 2005]

- Motivation: test gene group enrichment, needs to take the correlation structure into account.
- Methods:
  - Statistical hypothesis: there are two ways to formulate the null hypothesis:
  - Q_1: the genes in the set show the same pattern of association with the phenotype compared with the rest of genes.
  - Q_2: none of the genes in the set show association with the phenotype.
- Note that when testing Q_1: the correlation of the genes in the set may make their assocation pattern different from the rest.
- Test procedure:
  - For Q_1: suppose $t_1, \cdots, t_B$ is the score of all genes, then the hypothesis is this distribution of the genes in the group should be the same with the rest. More appropriately, the two distributions should have the same mean, thus t-test or Wilcoxon rank test. The null distribution can be generated by permuting $t_1, \cdots, t_B$.
  - For Q_2: the test statistic can be the average association metric of the group. The null distribution: permuting the phenotype labels.
- Remark: in general, when testing a hypothesis concerning the collective properties of a group, there may be multiple ways of defining H_0: (1) none of the members has certain properties; (2) the average property of the group is no different from a random background.

Connectivity Map [Lamb & Golub, Science, 2006]

- Motivation: identify gene expression signatures of drugs and disease states, then use these signatures to study the relations among drugs and diseases.
- Background: why direct expression profile comparison is not good (e.g. to reveal drug response similarity)?
  - Batch effect: the cells grown at the same time tend to have similar profiles.
  - Microarray platforms: strongly affect the similarity measure.
- Data: 164 small molecule perturbations, on breast cancer epithelial cell line MCF7 (also prostate cancer epithelial and other cancer cell lines).

- Connectivity score: for any query signature (the ranked list of up-regulated and down-regulated genes), define the similarity of this signature with the reference expression profile in the database. GSEA: up-regulated genes should also be up-regulated, etc. in the test profile.
- Results:
  - Connection between small molecules: recover drugs in the same class, e.g. HDAC inhibitors; the top results for a druge gedunin (without knowledge of mechanism of action) are HSP90 inhibitors, suggesting that gedunin may work through inhibiting HSP90 pathway.
  - Connection between small molecules and diseases: query the database with gene signature of disease. E.g. the signature of drug-resistent genes (genes differentially expressed in resistent and un-resistent cancer cells) is connected to the rapamycin.

Network-based classification of metastasis [Chuang & Ideker, MSB, 2007]

- Background: expression data from metastasis patients is hard to classify (different marker sets found in different samples), probably because of heterogeneity (vary from patient to patient).
- Methods:
  - Scoring subnetworks: given a subnetwork of genes, (1) compute the average activity of the subnetwork at each condition/sample; (2) the association of the activity and the phenotype label across samples: mutual information. The significance of score is tested from: sampling diffrent sets of random networks; or permuting labels.
  - Classification of samples: (1) search for subnetworks with locally maximum scores; (2) use the subnetwork activity as features to build logistic regression.
  - Data: about 200 samples with 8000 genes.
- Results:
  - Subnetworks: enriched with GO terms, and with the relevant cancer genes.
  - Subnetworks are reproducible across datasets.
  - Subnetwork markers increase classification accuracy of metastasis.
- Remark: the subnetwork activity score can be viewed as a "hidden" feature (e.g. flux through a pathway), that is not observable but may be more biologically relevant. However, the weights of genes are not learned in this work, and a better way should be regression of phenotype labels using the expression of individual genes (which will learn the gene weights and the regression performance would suggest the subnetwork association).

CAGE data analysis [Megraw & Hatzigeorgiou, RECOMBSAT, 2008]

- Background: genome-wide Cap Analysis of Gene Expression (CAGE) data reveal the positions of TSS that are activated by RNA Pol II.
- Aim: the code for transcription initiation.
- Methods:
  - Classification of single peak TSSs: create a TF profile for each sequence (the TF score in the sequence), then logistic regression of single peak sequences and those not.
- Results:
  - About 45% map to single peak TSSs.

- The logistic regressor achieves almost perfect classification: AUC = 0.98. The strongest features confirms the known sequence features: TATA box, Initiator, etc. GC content plays a significant role in transcription initiation.

PhenoProfiler [Xu & Zhou, PNAS, 2009]

- Motivation: Traditional methods fro prediction of phenotypes from transcriptomes are based on classification, not applicable if the training and testing transcriptomes come from different platforms.
- Idea: if the expression of a gene is correlated with the phenotype in the training data, then it should also be so in the testing data. Because this is based on correlation, we avoid directly compare expression data in training and testing set.
- Methods:
  - Signature genes: association of gene expression and phenotype (discriminative feature). If phenotype is binary, could use 2-sample t-static; if continuous, could use correlation.
  - Prediction: instead of predicting phenotype of one sample, predict many samples simultaneously. Define P as the vector of phenotypes of the testing dataset (multiple samples), find P s.t. the expression of signatures genes correlates with P (summing over all signature genes).

Predicting growth rate from expression profiles [Airoldi & Troyanskaya, PLCB, 2009]

- Motivation: identify genes whose expression underlie the growth rates of yeast cells.
- Methods:
  - Growth rates: controlled through nutrient availability in a chemostat.
  - Selecting growth signature genes: expression of some genes are correlated to growth rates, find through linear regression:
    $$Y_g = \alpha_g + \beta_g X_g + \epsilon_g \tag{2.6}$$
    where Y_g is expression of g, and X_g is the growth rate, α_g and β_g (growth effect) are gene-specific parameters. Genes with large β_g are signature genes.
  - Growth rate prediction from a new expression profile: suppose μ is the new growth rate, fit expression data using estimated β_g (similar to Naive Bayes classification, except that regression is used).
- Results:
  - 72 reliable genes are identified with growth specific expression patterns. cis-regulatory analysis identify Msn2/4, Rap1, and Puf4 (mRNA degradation regulator, at 3' UTR).
  - Growth rate is related to the metabolic cycle: minimum when O_2 consumption is minimum and maximum when O_2 consumption is maximum. Instead of correlating with cell division cycle.
  - Growth rate determination: not by simply nutrient and energy availability, rather by the perception of cells to environment conditions, mostly through Ras/PKA pathway.

GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. [Culhane & Quackenbush, NAR, 2011]

- Goal: provide a collection of gene signatures, each associated with some phenotype/disease characterization. The signatures can be used, e.g. for cancer diagnosis/prognosis.
- Data: literature curation of gene signatures reported in several thousand articles. Mostly cancer. Could search gene signatures by gene or by phenotype.
- Analysis: comparison of gene signatures, the extent of overlap.

2.3 RNA-seq overview

Background: splicing

- Spice sites: in the intron of a pre-mRNA, it has a donor site in the 5' end, and an acceptor site in the 3'end. The sites tend to be conserved, in particular, the 5' end has GU, and the 3' end has AG.
- Alternative splicing: the trans-acting proteins bind to the cis-acting sites to produce spcling, and this subject to regulation. Most common mode is exon skipping: e.g. using the donor site of one intron and the acceptor site of the next intron. Other types of splicing: mutually exclusive splicing, intron retention, alternative 5'end, alternative 3' end. Other types of events: alternative promoters, alternative polyadenylation.

Transcriptomics by RNA-seq [Wang, NGS book, Chapter 7]

- Experimental design: factorial design, could detect interaction terms. Ex. two-factor design, cancer vs. normal and treatment vs. control (2 by 2 table). Randomization if possible: e.g. to compare treatment vs. control, randomize the groups wrt. other factors.
- Sample preparation:
  - Removal of rRNA/mRNA enrichment: polyT primer, or rRNA probes to isolate rRNA.
  - Removal of DNA: use DNase.
  - RNA integrity number (RIN): if not significant degradation, in electrophoresis, we should have two large peaks, corresponding to 18s and 28s rRNA. With degradation, only short RNAs.
- Sequencing strategy: randomize by lane and flow cells to minimize technical factors. Ex. multiplexing sequencing (both treatment and control) in the same lanes.
- Data QC: tools such as RNA-seQC.
  - Percent of rRNA reads: ideally small.
  - Duplicate reads: more complex than DNA-seq, because highly expressed genes may generate duplicate reads.
- RNA-seq normalization: RPKM/FPKM as simple metrics. More complex schemes. Ex. DEseq, normalize read counts for each sample by a scaling factor. Let x_{ij} be the read counts of sample i gene j. We obtain the scaling factor for sample i as:

$$\lambda_i = \text{median}_j(x_{ij}/\bar{x}_j) \tag{2.7}$$

  where $\bar{x}_j$ is the geometric mean of gene j. The assumption here is that for majority of genes, there is no DE among samples (so the median is the same, after normalization by read counts). Then the normalized read count is: $\tilde{x}_{ij} = x_{ij}/\lambda_i$. Limma uses quantile normalization.

RNA-seq overview: [Yunlong Liu's lectures; RNA-Seq: a revolutionary tool for transcriptomics, NRG, 2009]

- Applications of RNA-seq: (1) Quantify gene expression level; (2) Transcriptome structure: splicing isoforms; (3) Variants in transcrips: SNPs in coding region, 3'UTRs (miRNA binding), etc. Allele-specific expression, fusion.
- Limitations of microarray technology:
  - Need to know the sequences to design the array

  - Hard to fit a large number of probes in a single array. Ex. for ChIP-chip, need tiling arrays for the whole genome
  - High level of noise due to (1) cross-hybridization, (2) non-linearity and limited dynamic range of detection: for some highly-expressioned genes, the signals reach the saturation level, i.e. increasing the gene expression level wouldn't lead to change of fluroescence level.

- Comparison of array and RNA-seq:
  - For some low-expression genes, high background noise. The correlation between RNA-seq reads and microarray: OK at moderate expression level (0.5), but < 0.2 for low and high ends.
  - Array is still cheaper, so for cases where the advantages of RNA-seq do not matter match, array could still be alive. Ex. clinic market, where the markers are relatively well-defined.

- Considerations for RNA-seq experiment design:
  - Choosing conditions: especially about the control condition/tissue.
  - Quality control of RNA-seq experiments: positive and negative controls. Ex. spike-in experiments. Ex. RPKM in exons vs. introns or intergenic regions (should be very high in exons).
  - Reproducibility: between technical and biological replicates.

RNA-seq sample preparation:

- Overview: a typical RNA-seq experiment, Figure 1. of RNA-seq NRG review. The goal is obtain a cDNA library with adaptors. Several key aspects:
  - Always do reverse transcription to get cDNA. Either do RNA fragmentation, then RT; or RT first, then fragmentation.
  - Ligation of cDNA fragments with common adaptors, which can be PCR'ed. Since a common adaptor is used, PCR is simple, and presumably retain the relative proportion of RNA species in the sample.
  - Sequencing: generally a fixed number of reads in a library.

  Important for understanding/analyzing RNA-seq data: usually we can only detect the relative proportion of RNA molecules. The RPKM number is measured relative to a given library size.

- 0.05-2 μg total RNA (amplification of low-level mRNA possible).

- rRNA removal is important (more than 90%).

- Reverse transcription (RT): (1) by oligo-DT: focus on poly-A transcripts, but may bias towards 3'end (RNA may degrade in the middle); (2) random primers: better for poly-A'ed transcripts, but may have a high proportion of rRNA.

- Maintain strand-specific information: important for novel transcripts.

- Paired-end or not? Good for detecting fusion: if in many pairs, one from one gene, the other from a different gene, then very likely a fusion. Difficult with single-end.

- Implications for RNA-seq comparison: usually we compare the same RNA species across conditions. Since we normalize by the library size, they are comparable. To compare different RNAs is more difficult: the steps in sample preparation can change the relative amount of RNA, such as reverse transcription and PCR (e.g. some RNA species may be easier to copy because of different GC content).

RNA-seq sequence alignment:

- Challenges: many junction reads (introns appear as large gaps). Also considerable error rates.
- Background: Median exon length of human genome is 127bp, thus a significant fraction of reads are junction reads.
- Alignment strategy I - unspliced read aligners: mapped to reference transcriptome, use the same aligners (BOWTIE, BWA, MAQ, etc.). Not allow large gaps. For BWA-based approaches, perform degrades rapidly as the number of mismatch increases. Seed-extension approach: map to a reference transcriptome of a distant species, increase sensitivity.
- Alignment strategy II: spliced aligners. First map reads to exons (exon-first), then for the remaining reads, map part of them (shorter segments) to each of the exons, then connect. Note: can miss reads in exon-intron junctions.
- Stragegy II: seed-extension. First break all reads into k-mers, then map each k-mer, and finally connect k-mers.
- Junction library: from all potential splice junctions.

Using RNA-seq for quantifying gene/transcript expression [Yunlong Liu's lecture]:

- The relationship between the number of reads and RNA level, gene length and sequencing depth:
  - Each copy of a molecule generate many nonoverlapping fragments (Normally there are multiple copies of the same molecule, so overlapped fragments result). The number of fragments generated is proportional to the number of copies and the length of exon. Each fragment has a certain probability to be sequenced (depending on the sequencing depth).
  - RPKM: defined as $R = 10^9 C/(NL)$, where C is the number of reads fall into the exon, L is the exon length and N is the sequencing depth (number of reads, e.g. 30M). This normalizes the length of transcript and the sequencing depth (s.t. different transcripts can be compared).
- Problems with isoforms:
  - For alternatively spliced genes, we need expression level of each isoform (isoform expression). Many reads thus cannot be unique assigned to a single isoform.
  - Isoform estimation and DE detection methods: Cufflinks and MISO.
  - Exon union and intersection methods: exon union: use all reads in all exons presented to compute RPKM. Exon intersection: similar, but use only reads in common exons. Problem: exon union - over-count the length, thus underestimate gene expression (e.g. some exon may occur only in one isoform with very low expression, but we count it in full). Exon interaction: lose information.
- RNA-seq diff. expression:
  - Simplest method: Fisher's exact test. Ex. 16 reads out of 10M vs. 10 reads out of 3M. The problem: information in replicates is not used.
  - Linear model: log-X (counts) is a linear function of lane/batch effect, gene expression level.
  - Poisson or NegBin distribution of read counts.
  - When the genes have multiple isoforms, better to use isoform expression methods.
- Question: amplification of DNA or transcript before sequencing? Note: this is different from bridge amplification during sequencing, which is only for reading the sequence, but each fragment will still generate one read (the cluster from amplification is used for sequencing).

Using RNA-seq for splicing analysis [Yunlong Liu lecture; Shirley Liu lecture]

- Isoform reconstruction: isoform-based transcript reconstruction, or AS event-based approach (for any exon, whether it is included or not). (1) Isoform-based methods: Cufflinks (use reference genome), Trinity (de novo assembly). (2) AS event-based methods: MISO.
- Isoform abundance estimation: for a given isoform set, estimate the abundance, Kalisto, Salmon, Cufflinks.
- Detecting differential splicing events: most methods are based on change of PSIs, focusing on exon skipping events. For reads overlapping exons, consider exon inclusion reads and exon skipping reads, and test the difference of ratios, e.g. by a Binomial model. Methods: MISO, rMATS.
- Alternative splicing by RNA-seq [Wang et al, Nature, 2008, C. Burge lab]:
  - 10 human tissues and five epithelial/breast cancer cell lines. PolyA selection. 10-30M 32bp reads, 60% mapped to genome, 4% to junctions.
  - Results: 92% of human multi-exon genes have multiple isoforms (minor form ¿ 15%) in at least one tissue.
  - 22K tissue-specific transcripts identified (greater than 60%). Variations between individuals: 10-30%, between tissues: 47-74%.
- Alternative splicing by single cell RNA-seq [Tang et al. NM, 2009]: Human blastomere or oocytes. Identified 75% (5,000) more genes than arrays. 8-19% genes are expressed in at least two isoform in a single cell.

Using RNA-seq for detecting gene fusion:

- Example of gene fusion: Bcr-Abl1 fusion in CML [2237408], TMPRSS2-ERG fusion in prostate cancer [Tomlins, Science, 2005, 16254181]
- Advantages of detecting gene fusion using RNA-seq? Many potential fusion events in the genome, RNA-seq helps detect the function/causal ones (which should be expressed).
- Gene fusion caused by chromosome rearrangements: inversion, translocation, large deletions/insertions. Note: if only at RNA level, it is called trans-splicing.
- Why fusion is bad? The proto-oncogene in the fusion may get expressed by the promoter/enhancer of another gene, or the fused gene may acquite a new activity.
- NGS for fusion: paired-end reads for detecting the events, and single-end reads help identify the break point.
- Computational methods: deFuse, FusionMap, TopHat-Fusion.

Using RNA-seq in cancer genomics:

- Background:
  - Very high mutation rates of cancer: could be 10K per genome.
  - TCGA: main phase (next 5 years), 20 tumor types, 500 each. International Cancer Genome Consortium.
- Representation of somatic mutations in cancer: multiple circles, each circle a type of mutations. For example, the inner circles show inter and intra-chromosome rearrangements and CNVs, and the outer circles show the indels, point mutations, etc. (histogram). Ex. small cell lung cancer genome.
- Benefits of RNA-seq for cancer studies:

  - Comparison of gene expression by identifying dysregulated pathways and potential therapeutic targets.
  - Detection of somatic mutations: in genes and fusions.
  - Biomarkers for detection and diagnosis.

- Tips for RNA-seq:
  - Use paired-end if possible
  - Reduce batch effects: also relevant for RNA-seq.
  - If the goal is to identify somatic mutations (SNPs, indels), use WES or WGS. Reverse transcription can introduce FPs, coverage is not uniform, cannot detect large-scale deletions.

- Somatic mutation rates of cancer: [Chromatin organization is a major influence on regional mutation rates in human cancer cells, Nature, 2012]
  - A melonoma cancer cell line: 33,345 somatic base substitutions, 680 small deletions and 303 small insertions
  - Primary prostate cancer: 3,866 putative somatic base mutations (on average 0.9 per Mb)

- Application: triple-negative (ER,PR,HER2) breast cancer. 10-15%, high mortality, BRCA1 are often triple-negative.
  - Significant variations in normal cells. Some of them are due to menstrual cycle.
  - Tumor vs. normal cells: over 5,000 genes are different.
  - Tumor vs. tumor: increaesd likelihood of passenger findings.
  - Gene fusion: detect multiple fusion events, however, they are unique to each patient.

2.4 Transcriptome Inference

Next-generation transcriptome assembly [Martin & Wang, NRG, 2011]

- About 100-1,000 depth of coverage in a typical RNA-seq experiment.
- Challenges of transcriptome assembly: why cannot we use existing genome assembly tools?
  - The sequencing depth of transcripts can vary by several orders of magnitude. Many short-read genome assemblers use sequencing depth to distinguish repetitive regions of the genome, a feature that would mark abundant transcripts as repetitive.
  - RNA-seq experiments can be strand-specific.
  - Transcript variants from the same gene can share exons and are difficult to resolve unambiguously.

- Considerations of RNA-seq library construction: this step may introduce bias in the data
  - Selecting mRNA and other RNA molecules: polyA selection is effective at removing rRNA but will miss ncRNA. Hybridization-based depletion methods reduce the representation of rRNAs and other highly abundant transcripts. However, this step may bias the quantification of highly abundant transcripts.
  - PCR amplification step: a low sequencing coverage for transcripts or regions that have a high GC content. Better to use amplification-free protocol if possible.

- Data preprocessing: three types of artefacts should be removed from raw RNA-seq data. (1) sequencing adaptors which originate from failed or short DNA insertions during library preparation; (2) low-complexity reads; and (3) near-identical reads that are derived from PCR amplification.

- Error correction: use quality scores to filter/trim reads. Correction using k-mers: k-mers that occur in the data set at very low frequencies are probably sequencing errors. Problem: may falsely remove reads derived from rare transcripts.
- De novo assembly: create the de Bruijn graph, collapse the graph, traverse the graph to reveal isoforms. The methods differ in details, eg. whether to do postprocessing to merge contigs and remove redundancy. Tools: trans-ABySS, Trinity, Oases.
- Trinity: (1) greedily assembling a set of unique sequences from the reads and then pooling together sets of unique sequences that overlap. (2) creates an independent De Bruijn graph for each group of sequences and assembles isoforms within the group, which can run in parallel.
- Comparison of de novo vs. reference-based assembly:
  - Require higher coverage: reference-based need 10x coverage while de novo may require 30x coverage.
  - Sensitive to errors and chimeric reads.
  - De novo assembly is able to detect novel transcripts.
- Combined strategy: Align-then-assemble approach. Assembling the data set using the reference genome, followed by de novo assembling the reads that failed to align to the genome. Alternatively, the transcripts that result from the reference-based assembly could also serve as input to the de novo assembly if the de novo assembler supports both long and short reads.
- Combined strategy: Assemble-then align approach. De novo assembly should be performed first, followed by alignment of the contigs and unassembled reads to the reference to extend and scaffold contigs.
- Assessing assembly quality: proposed metrics
  - Accuracy: among the total alignment of assembled and reference transcripts, the percent of bases that are correctly aligned.
  - Completeness: among all N reference transcripts, the fraction that are correctly assembled (the percentage of a reference transcript, that is covered by assembled transcripts, is greater than a threshold, e.g. 80%).
  - Continguity: among all N reference transcripts, the fractions that are correctly covered by a single, longest transcript.

TopHat: discovering splice junctions with RNA-Seq [Trapnell & Salzberg, Bioinfo, 2009]:

- Motivation:
  - Identification of novel transcripts from a reference genome.
  - Computational efficiency: existing tool QPALMA uses a SVM to classify junctions, however it is very slow (in addition, requires good training data).
- Method:
  - Read alignment to the genome, using BOWTIE. Many reads are not aligned at this stage, called Initially UnMapped (IUM) reads. Error control: each read could have multiple alignments (no more than 10), and no more than 2 mismatches in the 5' end (usu. more accurate than 3' end) - additional filter may be possible. Low-complexity reads (more than 10 alignments) are discarded.
  - Assembly using MAQ: clusters of aligned reads will be merged and a consensus will be built. This is called an "island", which is supposed to be an exon. However, there may be errors, thus really pseudo-consensus.

  - Defining exons and potential splice junctions: allow gaps in the islands because of possible low coverage (if two islands/clusters are very close, merge them and call it a single exon). Default is 6bp. To define all potential splice junctions, enumerate all pairs of canonical donor and acceptor sites (within the range of 70 to 20K bps).
  - Map junction reads by seed-extension: for any potential junction sites, define a seed as k-mer up/down-stream of the boundary ($k = 5$ by default). Then map each read to the possible junctions (require a perfect match in the seeds), then do extension in the 5' end of the read (s=28 bp by default), allowing a user-specified number of mismatches.
  - Report the junctions: create non-redundant junctions from all alignments. All spliced alignments with estimated minor isoform less than 15% will be discarded.

- Implementation details: Junction mapping: this is the most expensive step (map all IUM reads to potential junctions). Create a table with keys as all 2k-mers. For each 2k-mer, a list of all reads containing the 2k-mer. Each read can be mapped to many 2k-mers because of the start positions can be different.

- Problems and limitations:
  - For genes with low expression, no or few reads.
  - In the step of defining a single exon: possible that TopHat merge two exons into a single one, and thus miss splice junctions (e.g. from pre-mRNA).
  - Junctions spanning very large introns or those with non-canonical splice sites.
  - Reads with multiple spliced alignments

- Questions/Remarks:
  - Missing splice junctions from pre-mRNA (islands containing introns): could use the quantitative information, exon should have much higher reads than introns.
  - Mapping splice junctions: exact match of seed, and user-specified threshold of extension. Could be improved by an error model.
  - The final report: how the decision is made? Ex. a junction supported by just one read could be due to error.
  - Paired-read data: does TopHat take advantage of that?

Trinity: Full-length transcriptome assembly from RNA-Seq data without a reference genome [Grabherr & Regev, NBT, 2011]:

- Background: de Bruijn graph. A node is defined by a sequence of a fixed length of k nucleotides ('k-mer'), and nodes are connected by edges, if they perfectly overlap by $k - 1$ nucleotides. This compact representation allows for enumerating all possible solutions by which linear sequences can be reconstructed given overlaps of $k - 1$. For transcriptome assembly, each path in the graph represents a possible transcript.

- Intuition: Trinity partitions the sequence data into these many individual graphs, and then processes each graph independently to extract full-length isoforms and tease apart transcripts derived from paralogous genes.

- Inchworm: assemble reads into unique transcript contigs. Intuition: connect highly overlapped reads (use k-mers to index the overlapped part). To do that, build a k-mer dictionary, then use the most frequent k-mers as seeds, and extend in both directions

- Chrysalis: pool contigs into de Brujin graph. Look for contig overlap and connect them

- Butterfly: report possible isoforms from the graph.
- Challenges for genome-independent methods: sequence errors and SNPs (introduce branch points in the graph), the balance between sensitivity and graph complexity (selection of k)

Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels [Schulz & Birney, Bioinfo, 2012]

- Background: The common idea shared by existings pipelines (Rnnotator, STM, Trans-ABySS) is to run an assembler at different k-mer lengths and to merge these assemblies into one. The rationale is to merge more sensitive (lower values of k) and more specific assemblies (higher values of k).
- Contig assembly: use the first stage algorithms of Velvet to create contigs (hashing and graph construction). The later stage algorithms, Pebble and Rock Band, which resolve repeats in Velvet, are not used, which rely on the assumption that the coverage is uniformly distributed, and the genome should not contain branch points.
- Contig correction: similar version of TourBus to remove bubbles. Local edge removal (dynamic coverage filter): for each node, an outgoing edge is removed if its coverage represents $< 10\%$ of the sum of coverages of outgoing edges from that same node (low coverage edges are likely errors). Finally, all contigs with less than a static coverage cutoff (by default 3x) are removed from the assembly.
- Scaffold construction: use spanning single reads or paired-end reads to construct scaffolds (the number of such supporting reads is called support).
- Scaffold filtering: static filter: e.g low support. Connections between two long contigs are based on a stsistic, which estimates how many read pairs should connect two contigs given their respective coverages and the estimated distance separating them.
- Locus construction: Long contigs are first clustered into connected components. To each locus are added the short nodes which are connected to one of the long nodes in the cluster. This is followed by transitive reduction of the loci (removing redundant connections).
- Transcriptome assembly: given the loci, extract transcripts based on graph topology. Some simple cases: chain, bubble and fork - all the transcripts can be easily extracted by enumerating the paths. For complex loci, use dynamic programming to enumerate heavily weighted paths through the locus graph in decreasing order of coverage.
- Merging assemblies from different values of k's (Oases-M): any transfrag in the graph that is identical or included in another transfrag is removed.
- The impact of k-mer length on assembly quality: completeness metric (the percent of reference transcripts recovered at 80%), in the range of k = 19 to 35.
  - Trade-off: large k leads to low coverage (hard to distinguish errors and reads); small k leads to ambiguity in assembly.
  - Highly expressed genes: prefer longer k because coverage is less an issue ($k \geq 27$). Lowly expressed genes: prefer shorter k ($k = 19, 21$). For most k-mers ($k < 35$), at very high expression level, the performance actually drops (even though coverage increases, this is perhaps due to (1) more error reads; (2) more ambiguity - chance overlap among k-mers.
  - In addition, short k-mer assemblies have the disadvantage of introducing misassemblies.

SEECER: Probabilistic error correction for RNA sequencing [Le & Bar-Joseph, Nucleic Acids Research, 2013]:

- Background: Errors in RNA-seq. (1) Biases in the abundance of read sequences due to RNA priming preferences, fragment size selection and GC-content. (2) Sequencing errors. Up to 3.8% for Illumina Genome Analyzer. To deal with errors: read-triming is often used (removing the bad-quality reads).
- Intuition: for any contig, construct an HMM to represent the possible errors, and use the HMM to correct for errors. The error-corrected reads will be used for de novo assembly.
- Method:
  - Choose any read randomly as a seed, and find any reads that share a K-mer with this read (note: for different reads, may overlap at different k-mers). In this set, use cluster analysis (spectral clustering) to find the biggest subset that likely represent reads from a single transcript.
  - Learn parameters of the contig HMMs (similar to profile HMM).
  - Contig extension on both sides: (1) from all unassigned set, align reads to the consensus of the HMM: requires k-mer match (or more). (2) For all the reads that match the consensus in the edges: learn additional columns (this step would involve multiple reads and may extent more than one column). (3) Use entropy at the extended position(s) to determine it they should be accepted (default cutoff is the maximum entropy is 0.6).
- Remark:
  - Read selection: choosing reads randomly at each step may lead to significant redundancy of contigs (e.g. two isoforms E1-E2 and E1-E3, will learn only E1 contig if the reads start at E1). Could use guided selection, e.g. create a connected graph of reads, and choose reads based on the graph. Also the read cluster is created from K-mer match, could try other ideas, e.g. edit distance. Also HMM learning may include weighting schemes (based on edit distance).
  - Read extension: similar to VCAKE - use multiple reads for extension and the consensus of multiple reads is used for the new position(s). However, VCAKE or other k-mer extension methods rely on depth of coverage and/or quality scores, but SEECER does not.
  - Error modeling: learn HMM for each cluster leads to overfitting - esp. a problem for learning transition (I,D states) probabilities because they are rare events. Global model of systematic errors (Phred scores, and indel rates) could help.
  - Paired-end reads: not used in Seecer. Could significantly improve contig mapping.

2.5 Isoform Quantification and Differential Expression

Chapter 8. High-Throughput Count Data [Modern Statistics for Modern Biology]

- Challenges of count data: heteroskedasticity (different distributions for different genes). Systematic sampling biases such as: sequencing depth and gene length. Limited sample size: difficult to use resampling approach, and make it especially hard to estimate overdispersion.
- Dispersion: let r be sequencing library size, and p_i be the true proportion of gene i (number of fragements of i vs. number of all fragements). Then read count of i follows Poisson(rp_i). However, p_i may also vary across biological replicates.
- Normalization: identify the nature and magnitude of systematic biases, and take them into account. Why not just use library size? Size factor in DEseq2 (Figure 8.1): compare reads of genes in one sample vs. reference/median, do robust regression. Then the slope is the size factor. Idea: size factor is determined by majority of genes (expression should not change).
- Assumption: most genes do not change expression. Could test this by checking a set of genes that are not expected to change expression.

- DESeq2:
  - Bioconductor: use "class", e.g. DESeqDataSet.
  - `DESeq()` function: calls three steps, `estimateSizeFactors()`, `estimateDispersions()`, `nbinomWaldTest()`.
  - Results of DEseq can be summarized: Figure 8.4-8.7, p-value histogram, MA plot, PCA, heat-map of top genes.
- Robust regression: replace least-square objective in regression. Use absolution values, or some function to downweigh outliers (M-estimation), or weighted least square. DEseq2: use weighted least square, weights are determined by Cook's distance, measuring leverage of data points.
- GLM for DEseq2: Equation 8.14-8.16. For gene i in sample j, its count K_{ij} follows Gamma-Poisson distribution with overdispersion α_i, and the mean determined by the linear model

$$K_{ij} \sim GP(s_j q_{ij}, \alpha_i) \qquad \log q_{ij} = \sum_k x_{jk}\beta_{ik} \tag{2.8}$$

  where x_{jk} is the k-th treatment of sample j. Test β using Wald-test.
- Empirical Bayes shrinkage: for α and β (optionally). Estimate prior of the parameters by MLE using all genes. Figure 8.10: two genes, one with larger overdispersion. Different posteriors.
- Count data transformation: log2 pseudocount. Variance-stabilizing transformation. Regularized log transformation.
- Outliers: DESeq2 automatically discards genes with Cooks distance above a cutoff. An alternate strategy is to replace the outlier counts with the trimmed mean over all samples.
- Testing fold change above a threshold: banded null hypothesis.

Mapping and quantifying mammalian transcriptomes by RNA-Seq [Mortazavi, NM, 2008]

- Mouse liver, skeletal and brain. PolyA selection. Illumina 25bp reads, about 50M mapped. 90% uniquely mapped reads were mapped to exons.
- Introduce RPKM: Reads Per Kilobase exon model per Million mappable reads, to measure expression level. Use spike-in Arab. DNA (known the quantity), and show that the correlation between the spike-in level and RPKM is 0.99. 1RPKM = 1×10^5 transcripts per 100ng total RNA.
- Power analysis: as the number of mappable reads increase (depth of coverage), the fraction of genes whose measurement is within 5% error of true values (known from spike-in). For highly expressed genes (RPKM > 3K, $n = 24$), need only 1M reads to reach close to 100%. For lowly-expressed genes (RPKM from 3 to 29, $n = 6,000$), require > 40M reads. Most genes have RPKM below 3.
- Question: the accuracy of RPKM measurement. Does it work well for the low expression level? The Spike-in experiment may have relatively high expression.

edgeR: Small-sample estimation of negative binomial dispersion, with applications to SAGE data [Robinson & Smith, Biostatistics, 2008]; Moderated statistical tests for assessing differences in tag abundance [Bioinformatics, 2007]

- Motivation: hierarchical model of rates. Suppose we are estimating the rate of one group of samples, we assume:

$$Y_i \sim \text{Pois}(m_i\lambda_i) \tag{2.9}$$

  where m_i is the library size of sample i, and λ_i the rate of the sample. The reason we use a different λ_i is that even for biological replicates, the rates can be different (biological interpretation: true expression level, true affinity of DNA, etc). Yet, these different rates are related $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, and we are interested in the common underlying rate $\lambda = \alpha/\beta$.

- Background: Negative Binomial (NB) model with a different parameterization. We assume the count $Y \sim NB(\mu, \phi)$, its mean $\mathrm{E}(Y) = \mu$ and variance $\mathrm{Var}(Y) = \mu + \phi\mu^2$. When $\phi = 0$, this reduces to Poisson distribution (no variation across biological replicates). Also the property of NB: let $Y_i \sim \mathrm{NB}(r_i, p)$ (we use a more common notation here), then $\sum_i Y_i \sim \mathrm{NB}(\sum_i r_i, p)$.

- edgeR model: let Y_i be the count of a gene in sample i, for all samples of the same group (e.g. treatment), we have:

$$Y_i \sim \mathrm{NB}(m_i\lambda, \phi) \tag{2.10}$$

  where m_i is the library size, and λ is the fraction of reads mapped to the gene (transcript). The MLE of λ will depend on the joint estimation with ϕ. A special case is $m_i = m$, then MLE of λ is the total count divided by the total library size.

- Estimation of variance parameter ϕ: to estiate ϕ, use conditional maximum likelihood (CML) when $m_i = m$ are equal. The idea is that the model has a nusiance parameter λ, but we can eliminate it through conditional distribution $Y|Z = \sum_i Y_i$ (the total count is a sufficient statistic of λ):

$$P(Y_1, \cdots, Y_n|Z, \phi) = \frac{P(Y_1 = y_1) \cdots P(Y_n = y_n)}{P(Z = z)} \tag{2.11}$$

  Plug in the NB distribution, it is easy to show that the terms dependent on $\mu = m\lambda$ cancel out. Also to maximize the conditional likelihood, we can ignore the terms that are not dependent on ϕ. The result is that we need to maximize:

$$l_{Y|Z=z}(\phi) = \sum_i \log\Gamma(y_i + \phi^{-1}) + \log\Gamma(n\phi^{-1}) - \log\Gamma(z + n\phi^{-1}) - n\log\Gamma(\phi^{-1}) \tag{2.12}$$

- Quantile adjustment of read counts: when m_i are not equal, we create "pesudodata" s.t. the library sizes are about equal. We first obtain m^* as the geometric mean of m_i. Then to transform our data $Y_i \sim \mathrm{NB}(m_i\lambda, \phi)$, suppose we know ϕ, we deterine the percentile of Y_i, and then we create the data points under $\mathrm{NB}(m^*\lambda, \phi)$ that have the same percentiles. Since we do not know ϕ, this is done in an iterative procedure until convergence.

- Hypothesis testing: we could use asympotic tests such as Wald test or LRT or score test, however, at small samples, an exact test is preferred. Our idea is to look at the total count in treatment group vs. the total count in the control group (after quantile adjustment):

$$Z_k \sim \mathrm{NB}(n_k m\lambda_k, \phi n_k^{-1}), \quad k = 0, 1 \tag{2.13}$$

  where n_k is the number of samples. And we test $H_0 : \lambda_1 = \lambda_0$. Again, we use the conditional distribution $Z_1|(Z_1 + Z_0)$, note that the total count $Z_1 + Z_0$ is also NB distribution. The p-value is obtained:

$$p = \sum_{z=z_{\mathrm{obs}}}^{z_1+z_0} P(Z_1 = z|Z_1 + Z_0) \tag{2.14}$$

  The conditional distribution can be obtained similarly to the CML approach above.

- Variance stabilization: [Robinson & Smith, Bioinfo, 2008] use information in all tags to estimate the dispersion parameter of any gene ϕ_g. Use a weighted likelihood approach that approximates Empirical Bayes, maximize this objective function:

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g) \tag{2.15}$$

  where $l_g(\phi_g)$ is given by Equation 2.12, $l_C()$ is the qCML of all tags, and α is the weight. Under this model, when $\alpha = 0$, ϕ_g will maximize $l_g(\phi_g)$, thus it obtains gene-specific dispersion; when α is large, it

is dominated by common dispersion parameters. The problem is then to set the weight α. The idea is to use the normal hierarchical model as the template (combining gene-specific with group information). If we know that $\hat{\phi}_g$ is normally distributed with $N(\phi_g, \tau_g^2)$, and the prior of $\phi_g \sim N(\phi_0, \tau_0^2)$, then from the Bayesian normal model with known variance, the posterior mean of ϕ_g is:

$$\hat{\phi}_g^B = \frac{\hat{\phi}_g/\tau_g^2 + \phi_0/\tau_0^2}{1/\tau_g^2 + 1/\tau_0^2} \tag{2.16}$$

thus the posterior is the weighted combination of the group mean ϕ_0 and MLE. We use the estimated group mean to replace ϕ_0 above, and this allows us to obtain α:

$$\frac{1}{\alpha} = \sum_g \tau_0^2/\tau_g^2 \tag{2.17}$$

We also obtain the estimator of τ_0. In our problem, even though ϕ_g is not normally distributed, we can use the fact that the score statistic $S_g(\phi) = \partial l_g(\phi)/\partial\phi$ is asympotically normal, and its variance is given by Fisher's information. We thus use the score statistics and Fisher information to obtain α.

- **Lesson**: Approximate of Empirical Bayes by weighted likelihood approach. To do this approximation, the key is to determine the weight of group parameters. We set up a reference hierachical model, and try to mimic it. Specifically, find some statistic in our problem that matches the reference model, then apply the reference model to determine the weight.

Statistical inferences for isoform expression in RNA-Seq [Jiang & Wong, Bioinfo, 2009]:

- Model idea: we want to quantify the expression of each isoform. However, with alternative splicing, a read may be mapped to multiple isoforms (even though it is mapped to a unique exon) because isoforms may share exons. The idea is to model the reads mapped into each exon (generally unique), and infer the transcript level.

- Model of exon reads: suppose the number of reads mapped to the j-th exon is X_j. It follows a Poisson distribution with rate λ_j. This rate depends on the exon length and the level of isofroms containing this exon:

$$\lambda_j \propto \left(\sum_i c_{ij}\theta_i\right) l_j \tag{2.18}$$

where i is an index of isoform, c_{ij} is a binary indicator of whether the i-th isoform contains the j-th exon, θ_i is the isoform level and l_j is the length of the j-th exon.

- Model of junction reads: similarly, the X_{jk} be the number of reads mapped to the junction between exon j and k. It follows Pois(λ_{jk}). The rate:

$$\lambda_{jk} \propto \left(\sum_i c_{ij}c_{ik}\theta_i\right) l_{jk} \tag{2.19}$$

- Intuition of the model: we observe the exon-level reads, which depend on the exon-level concentrations. There is roughly a linear relationship between the two types of variables. So roughly, the method is solving a linear model, where y is the expected counts in exons and x the indicator variable of whether exon is in an isoform, and β the parameters (concentrations) of exons.

- Remark: the model does not explicitly model read mapping uncertainty using EM kind of algorithm. Instead, it takes advantage of the fact that the source of ambiguity here is known and can be addressed relatively easily.

- **Lesson**: identifiability analysis that mimics linear model. For the linear model, the identifiability depends on the number of unknowns vs. the number of equations (imagine we are doing method of moment kind of estimation). Here it means the number of isoforms (unknown) vs. exons (expected count in one exon gives one equation).

RNA-seq gene expression estimation with read mapping uncertainty (RSEM) [Li & Dewey, Bioinfo, 2010]

- Motivation: in RNA-seq, a significant fraction of reads are mapped to multiple genes or isoforms, ranging from 17% to 52%.

- Quantifiying gene expression: could use fraction of transcripts, or fraction of nucleotides (τ_i) of the transcriptome, while the former is interesting biologically, for RNA-seq, the latter is easier. Let c_i be the number of reads from isoform i, then $c_i/N \to \tau_i$ for large N.

- Model: we have N reads, $R_1, \cdots, R_N$, and M isoforms (many genes). Our basic idea is that the true source (isoform and position) of a read is unknown and we'll treat it as latent variable. Let G_i be the isoform of read i, $1 \le i \le M$, and S_i be the start position. In addition, we have O_i for its orientation. Let θ_j be the relative expression level (fraction of nucleotides) of isoform j with $\sum_j \theta_j = 1$. Our process of generating N reads (N is given) is: choose an isoform among all expressed isoform in a way proportional to relative expression level, sample a start position of a read and its orientation, and sample the sequence. Then we have this complete likelihod:

$$P(R, G, S, O|\theta) = \prod_i P(G_i|\theta) P(S_i, O_i|G_i) P(R_i|G_i, S_i, O_i) \tag{2.20}$$

  The first distribution is simply proportion to expression level: $P(G_i = j|\theta) = \theta_j$. The second distribution: $P(S_i|G_i)$ is uniform for simplicity, or we can model non-uniform start distribution and $P(O_i|G_i)$ is 1/2 or 0 or 1, depending on protocol. The last one could model sequencing errors: e.g. errors are more often at the last positions of reads. This is implemented with PSSM (a fixed matrix that is used by all reads).

- Inference: EM algorithm. Show that the function is concave if we only infer θ (ie. fixed PSSM parameters). Also to speed-up, for each read, only consider the isoforms that it is aligned to.

- Remark: the probability of start position S_i captures the coverage along the transcript. This is not necessarily uniform: the paper uses sequence position (in transcript) to determine this distribution (bell-shaped, the same for all genes). We could also model how sequence composition affects coverage.

Cutfflinks: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation [Trapnell & Pachter, NBT, 2010]:

- Constructing transcriptome:
  - Call a paired end read as a fragment. First use TopHat to create the splice alignments (map all fragments).
  - Compatible fragments: a pair of fragments that overlap and does not have different implied introns. Ex. fragment 1 has a splice junction that suggests some intron in the middle of fragment, but fragment 2 does not have that, then they are incompatible.
  - Overlay graph: each fragment as a node, each edge a pair of compatible fragments.
  - Find the minimum number of transcripts to explain all fragments: minimum path cover (minimum number of paths that cover all nodes of a graph). Each path then represents a transcript.

- Estimate transcript abundance. Some reads can be from multiple isoforms. Do ML estimation: $P(R|\rho)$, where R is read data and ρ is the abundance.

- Model idea: the basic sequencing process can be described in two parts: (1) Fragment pool: each transcript is randomly split into fragments. The number of fragments from one transcript is proportional to its expression level and depends on its length (roughly linear). (2) Sampling reads from the pool: for simplicity assume complete uniform sampling.
- Fragment distribution: determined from empirial data, $F(i)$ is the density of a fragment of length i (fragment length is determined from paired-end read, so not a constant, typically around 200bp).
- Notation: let R be all reads, and r be one specific read. Let T be the entire transcriptome, and t be one specific transcript. The matrix $A(r,t)$ denotes whether the read r is mapped to t: equal to 1 if yes, 0 otherwise. Let ρ be the level (up to a constant) of all transcripts. Given any read r, we want to determine the probability of $P(r|t)$, the probability that it is generated from t in that specific alignment.
- The probability that this read originates from t: this is determined by the fragment pool model. This should be roughly proportional to $\rho(t)$ and $l(t)$. However, the fragment length is not uniform, so we correct for this. Specifically, a transcript of length l can generate fragments of any length $i, 1 \leq i \leq l$, with probability $F(i)$, so the corrected length is $\bar{l}(t) = \sum_i F(i)[l(t) - i + 1]$.
- The probability that the read with the specific alignment is chosen: let $I_t(r)$ be the implied length of the read. If the read does not match t, then it is 0. Otherwise, it is the fragment length. The probability of this read is: uniform start probability times the fragment length, i.e. $F(I_t(r))/[l(t) - I_t(r) + 1]$.

- Comparison with Scripture (a similar tool): Cufflinks is conservative (using minimum path cover), while Scripture considers all paths in the transcript graph.
- Remark/Question: in the transript level model, the effective length $\bar{l}(t)$ seems somewhat strange. Note that $F(i)$ is close to 0 as i is very large, thus the contribution of large i's in the defition of $\bar{l}(t)$ can be ignored. Therefore, the effective length of all transcripts are about the same same as long as $l(t)$ is sufficiently large (say, greater than 500bp).

Normalization of RNA-seq data using factor analysis of control genes or samples [Risso and Dudoit, NBT, 2014]

- Motivation: ERCC (spike-ins) are not reliable for controlling technical variations.
- Model: let Y be gene expression, X be covariates of interest, and β their effects. Let W be latent factors and α their effects on Y. Use GLM for read counts data:

$$\mathrm{E}(Y|W, X, O) = W\alpha + X\beta + O \tag{2.21}$$

  where O is the offset term. Joint estimation of all parameters are not infeasible. So use multi-step procedure.
- RUVg: use negative control genes, i.e. $\beta = 0$. Let $Z = \log Y - O$, and do SVD on Z. After estimation of $W\alpha$, then plug in the GLM, and do regression analysis to estimate β. However, RUVg may be sensitive to the selection of genes.
- RUVr: do GLM on X, and obtain residuals, then do SVD. This is very similar to SVA.
- RUVs: use a set of samples where X is constant.

limma powers differential expression analyses for RNA-sequencing and microarray studies [Ritchie and Smith, NAR, 2015]

- Overview: linear model, for gene g, let y_g be expression, we have: $\mathrm{E}(y_g) = X\beta_g$. Variance modeling: let y_{gj} be the expression of g in sample j, the variance is $\mathrm{Var}(y_{gj}) = \sigma_g^2/w_{gj}$, where w_{gj} is weight (given). Use EB to borrow information across genes: posterior variance estimator.

- limma analyzes entire experiments together: modeling correlation and share information.
- Empirical Bayes estimator of variance: shared across genes, model mean-variance trend. Relative weighting of the gene-wise and global variance estimators no longer needs to be the same for all genes.
- limma strategy for read counts: log-scale, then model mean-variance relationship. The mean-variance trend is converted by the voom function into precision weights. Remark: treat data as linear, but the variance depends on the mean. Performance: comparable with NB based models.
- Unequal variability: this affects DE analysis, e.g. in tumor vs. normal comparison. Some people use Welchs t-test instead of classical t-test. limma has two strategies: (1) mean-variance trend. (2) estimate precision weights associated with treatment groups. Also combine (1) and (2): estimate mean-variance trend for each treatment group.
- Input of RNA-seq data: counts, instead of RPKM or TPM s.t. limma can estimate mean-variance relationship (voom).
- Preprocessing: (1) Normalization: remove systematic bias. (2) Explore sample relationship: e.g. plotMDS. Sample distance is the average log-fold change between samples. Identify patterns such as batch effects - this would help guide linear modeling. The removeBatchEffect function: can remove the effects of batch or other covariates.
- Linear models: input matrix is genes (rows) and columns (samples). Test for each row. Models can be fit robustly or by least squares, lmFit function. The plotSA function: diagnosis, residual vs. mean expression.
- Sample weighting: For RNA-seq data, the voomWithQualityWeights function combines observation-level and sample-specific weights for use in linear modelling.
- Blocking and random effects: allow samples to be correlated. Estimate the correlation structure across samples (technical replicates, or other related samples) and use that in linear model testing. Use duplicateCorrelation function to constrain that correlations are the same across all genes.
- Testing for DE: eBayes function estimates variance using EB.
- **Lesson**: the key component of limma is modeling of variance: (1) mean-variance relationship. This helps with count data. (2) Use EB to borrow information across genes. (3) Variance (and mean-variance relationship) varies across treatment groups.

Near-optimal probabilistic RNA-seq quantification [Bray and Batcher, NBT, 2016]

- Ref: see `https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html` for explanations.
- Idea: for isoform quantification, we do not need actual alignment, only information about whether a read is assigned to a transcript or not.
- Building equivalence class of any read: the goal is to learn which transcripts a read could possibly come from.
    1. Build de Bruijn graph of transcriptome (T-DBG): from T-DBG, one can obtain all possible contigs/transcripts as continuous paths (Figure 1b).
    2. k-compatibility class of a node (k-mer): all matching transcripts of a node (Figure 1b: circles).
    3. k-compatibility class of a read: intersection of k-compatibility classes of all constitutive nodes of the read (Figure 1c).

4. Skipping redundant nodes of a read: only nodes at the junctions in T-DBG are informative (Figure 1d).

- Quantification by EM: let y_{ft} be the indicator of whether fragment f is compatible with transcript t. Let α_t be the expression of transcript t and l_t be its length. Our likelihood is:

$$L(\alpha) = \prod_f \sum_{t \in T} y_{ft} \frac{\alpha_t}{l_t} \tag{2.22}$$

We can simply the computation. It is easy to show that we can merge the reads mapped to the same transcripts. Let e be an equivalence class (set of transcripts a read may come from): a read could come from each of the transcripts in the equivalence class, so we have:

$$L(\alpha) = \prod_e \left(\sum_{t \in e} \frac{\alpha_t}{l_t} \right)^{c_e} \tag{2.23}$$

where c_e is the reads in the equivalence class e.

2.6 Detecting Gene Fusion

Transcriptome sequencing to detect gene fusions in cancer [Maher, Nature, 2009]:

- Background: Typically, an aberrant juxtaposition of two genes may encode a fusion protein (for example, BCR-ABL1), or the regulatory elements of one gene may drive the aberrant expression of an oncogene (for example, TMPRSS2-ERG).
- RNA-seq analysis of CML cell line: 66.9 million reads of 36 nucleotides in length.
  - Partial alignment to exon boundaries from two different genes: a set of 111 other chimaeras (with at least two reads) and BCR-ABL1 is one of them.
  - If use known fusion junction as the reference sequence, 19 chimera reads. (Figure S1). Note: a number of reads have short partial alignment with one of the two genes, e.g. 7 reads match < 10 bp with one of the two.
- Hybrid technology: short-read sequencing technology for obtaining deep sequence data and long-read technology (Roche 454 sequencing platform) to provide reference sequences for mapping candidate fusion genes.
- Hybrid strategy for fusion mapping: using multiple prostate cancer cell lines.
  - Using long reads: reads that showed partial alignments to two genes were nominated as chimera candidates. Many of these chimaeric sequences could be a result of trans-splicing or co-transcription of adjacent genes coupled with intergenic splicing (neighboring genes - read-through). 428 VCaP candidates (only one read spanned TMPRSS2-ERG)
  - Short-read data on the candidates: TMPRSS2-ERG fusion as one among 57 candidates. Most are FPs.
  - Integrating the long- and short-read sequence data: the single long-read chimaeric sequence spanning TMPRSS2-ERG junction from VCaP transcriptome sequence, buttressed by 21 short reads. One of only eight chimaeras nominated, overall.
- A scoring function obtained by multiplying the number of chimaeric reads derived from either method: TMPRSS2-ERG was ranked the first.

- Overall, the read-through events appear to be more broadly expressed across both malignant and benign samples whereas the gene fusions were cancer-cell specific
- Lessons:
  - Using short-reads: the problem of FPs, found 112 chimeras, and hard to detect BCR-ABL1.
  - Using long-reads: the problem with coverage, e.g. only 1 read for TMPRSS2-ERG in prostate cancer cell line.
  - Possible misleading cases: trans-splicing (?) and read-throughs. Strategies: the locations of genes; comparison with control cells.

Chimerascan: Chimeric transcript discovery by paired-end transcriptome sequencing [Maher & Chinnaiyan, PNAS, 2009]:

- Background: (1) The restricted expression of gene fusions to cancer cells makes them desirable therapeutic targets, e.g. Gleevec for BCR-ABL1 in CML. (2) The lack of known gene fusions in epithelial cancers has been attributed to their clonal heterogeneity and to the technical limitations of cytogenetic analysis, FISH, etc.
- Comparison of single-read vs. paired-read strategy: Figure S1. Three benefits of paired-end: (1) Reduce multiple mapping and FPs; (2) More coverage of fusion junction (both discordant pairs and chimera reads); (3) Single-read: fusion boundary must map to the middle of reads.
- Mapping the mate pairs: (i) mapping to same gene, (ii) mapping to different genes (chimera candidates), (iii) nonmapping, and other categories (e.g. quality control). Overall, the chimera candidates represent a minor fraction of the mate pairs, comprising $\leq 1\%$ of the reads for each sample.
- Strategy for paired-end reads:
  - Category 2: mate pairs align to different genes. Apply filters: min. mate pairs, best unique mapping, etc. Results: encompassing mate pairs.
  - Category 3: single mate aligns to a gene, scan for another read that maps to the junction. Results: spanning mate pairs
  - The output from the two categories form candidate fusions. Further assess if they are read-throughs.
- Example: TMPRSS2-ERG fusion. Single read: using 100nt reads, a total of 17 chimera reads. Paired reads: 552 supporting reads.
- Result statistics: The long read approach nominated 1,375 and 1,228 chimeras, whereas with a paired-end strategy, we only nominated 225 and 144 chimeras in UHR and HBR, respectively. There were 32 and 31 candidates common to both technologies for UHR and HBR, respectively.
- Quantifying the support of fusion: the normalized mate pair coverage at the fusion boundary (mate pairs per million supporting fusion). The highest ones are likely driver changes. Ex. Bcr-Abl1 and TMPRSS2-ERG ranked among highest.
- Lessons:
  - Advantage of pair-end strategy over single-end strategy: better mapping as well as higher coverage (could be more than 10-fold). High coverage achieved by encompassing mate pairs and spanning mate pairs.
  - Importance of normalized coverage of fusion: likelihood of being driver fusion.

deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data [PLCB, 2011]:

- Motivation: (1) consider reads mapped to multiple positions. (2) Using both discordant reads and split reads - "a strong source of evidence for gene fusions in paired-end RNA-Seq data". The paper shows that discordant read analysis followed by split read analysis is better than the reverse procedure of PERAlign [Hu et al.].
- Use pair-ends. First, map the paired ends to the genome, identify discordant clusters (pairs in different chromosomes or very far - and clusters of reads).
- Use Dynamic programming to identify the fusion point.
- The final step: use fragment length to evaluate the event.

TopHat-Fusion: an algorithm for discovery of novel fusion transcripts [Kim & Salzberg, GB, 2011]:

- Motivation: (1) some existing methods are for paired-end reads; (2) computational efficiency: e.g. Trans-Abyss uses full BLAT to detect the discorcantly mapped reads (across fusion junction).
- Intuition: need to identify the fusion reads, part of which map one exon, and part of which map another exon. So map two seeds in a read (use seed-and-extension instead of local alignment, which is slow), and the pattern occurs, then it's a candidate fusion. The challenge is that one read may often map to more than one position, easily creating FP fusion events.
- Method:
  - First map all reads to exons and splice junctions (TopHat). Left with a set of IUM reads.
  - For each IUM read, break into three segments (25bp each) and map each segment. To identify putative fusion events, require a read to have 13bp exact match in both sides of the fusion. Candidate fusion is defined as two chromosomes or sufficiently far in one chromosome. Also remove all candidate fusion events involving multi-copy genes or repetitive sequences (at most two).
  - Fusion contigs: concatenate multiple reads along a fusion event to create 44bp fusion contigs. Then align all reads to the fusion contigs.
  - For each read, find the best alignment: among mapping to exons, splice alignment or fusion. Penalty based on intron, indel or fusion and then mismatch.
  - Filters: at least two supporting reads. Not in genic regions. Repetitive sequences. Require reads to cover a large window around the fusion point (600bp).
  - Fusion scoring: based on the number of reads in each side, etc.
- Remarks: heuristics employed in every stage:
  - Multi-map reads: if a candidate fusion involves multi-map reads (more than 2), it will be discarded.
  - Require exact 13bp match in each side: many reads that have shorter coverage in either side, or reads with errors will be ignored. In addition, for cancer data, somatic mutations may be important.
  - For each read: only choose the single best alignment.
  - Fusion scoring: no statistic model.

ShortFuse: Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs [Kinsella & Bafna, Bioinfo, 2011]:

- Motivation: (1) multiple-mapped reads; (2) transcript abundance encodes information (not just fusion reads).

- Model idea: (1) identify all putative fusion transcripts. (2) Inference of transcript level: thus if the level of fusion transcript is positive, it supports a fusion event. This model uses an extension of RSEM/MISO/Cufflinks models.
- Identify putative fusion transcripts: for each mate pair, if the two map to different genes (discordant pair), it nominates a putative fusion. A few additional conditions are applied. After this step, all reads are mapped to the augmented reference (with the putative fusion transcripts).
- Model of paired-end reads: (1) Probability of choose a transcript T is proportional to its nucleotide abundance. (2) Probability of sampling a fragment starting at a particular position, uniform. (3) Probability of having a fragment with a particular length: empirical distribution. (4) Probability of generating the read: error model, using informaiton of mismatches and the quality scores.
- Problem of ML estimation: Inference using ML by EM. It suffers identifiability problem: when both genes A, B and the fusion transcript $A-B$ are present, even if there is only a single read from the fusion, the MLE will assign all reads to $A-B$. To solve it, use the probabilistically weighted count of reads supporting the fusion (instead of the estimated fusion level). Also, filter: if coverage at the fusion site is less than one-twentieth of the upstream and downstream coverage, we discard the fusion as a probable artifact.
- Remark:
  - Use only paired-end reads, thus no need of considering single reads that span the fusion point. However, the spanning reads could be informative, but are ignored.
  - The model does not take the prior probability of fusion, which should be small, into account. More generally, we could have a higher-level model of probability of fusion and alternative splicing.
  - The statistical inference is to test if the fusion transcript is present: this is different from the problem of estimating transcript level. Using Bayesian model selection could do better.

Dissect: detection and characterization of novel structural alterations in transcribed sequences [Yorukoglu & Sahinalp, Bioinfo, 2012]:

- Goal: current tools for mapping transcript reads to genome are aimed for short-reads. However, readas are getting longer and a long read has a greater possibility of containing segments from more than two exons, complicating the process of mapping such a read.
- Model idea: we are mapping a transcript read to a genome, however, there may be rearragement, splice and fusion events. So the transcript is a stitch of multiple segments that align to (different places of) the genome. Algorithmic formulation of the alignment problem.
- Nucleotide leve model:
  - Given a transcript (read?) T, and a gene to be aligned G (we also consider its complement G' for the purpose of inversion, and a second gene S (for fusion), we try to identify the aligment of T to (G, G', S, S'), denoted as a function f.
  - The alignment score is sum of matches minus the penalty, which sums all inversion, translocation, splice and fusion. The penalty is a function of distance for inversion and translocation, and is a constant for fusion.
  - DP for alignment: for translocation, need an alignment table X_G. To incorporate inversion, we need two alignment tables X_G and $X_{G'}$. To model fusion, we add two more tables, X_S and $X_{S'}$.
- Fragment level model:
  - Suppose we already have a set of fragments, the goal is to produce a fragment chain that align to the genome.

  - Alignment score: sum of the fragment scores, minus the penalty from fusion, inversion, translocation (the chaining of different fragments).

- Application: to detect fusion, Dissect finds the best single-region alignments: if any of these regions cover a user determined percentage of the sampled anchors in the transcript, a double region is not searched for. Only search for double region when there is no good single-region.

- Questions/Remarks: how to deal with overlapping fragments?

2.7 Transcriptome Deconvolution

Digital cell quantification identifies global immune cell dynamics during influenza infection (DCQ) [Altboum and Amit, MSB, 2014]

- Problem: given expression difference in two conditions, m_i for gene i, and a compendium of expression data across many cell types, b_{ij} for gene i and cell type j, our goal is to estimate the relative cell composition change that explains change of expression.

- Assumptions: consider only marker genes (earlier methods use most discriminative genes of cell types). For cell compendium, assume cells in both resting and effector/activation states are available.

- Model: let c_j be the contribution of cell type j, we have $m_i = \sum_j c_j b_{ij}$. Fit elastic net to estimate c_j.

- Results: mouse model, influenza then RNA-seq on lung. Significant increase in specific Macrophage population, in effector but not resting B cells and CD8+ T cells.

- Analysis: suppose we consider one condition a time, and assume that across conditions, transcriptome of each cell type does not change. Then c_j should be interpreted as change of proportion of cell type j.

- Remark: rare cell types and regularization. For rare cell types, change of proportion would be small, and shrinkage may miss them. Could shrink not on percent, but normalized percent, i.e. change of cell proportion divided by mean cell proportion.

- Remark: the compendium expression data may not contain all relevant cells. Ex. compendium has only immune cells while our data has stromal cells. If our goal is to only estimate immune cell changes using the immune markers, its OK to ignore this.

- Remark: application to spatial transcriptomics: do bulk RNA-seq on regional samples, and do scRNA-seq to infer cell types. Then deconvolution of cell types in regional samples.

- Discussion: smoothing the parameter estimates in time-series data. Take advantage of tree structure of cell types (e.g. similar cell proportions of related cell types), but this may be hard.

Robust enumeration of cell subsets from tissue expression profiles, NM, 2015 (CIBERSORT) [Newman and Alizadeh, NM, 2015]

- Background: support vector regression (SVR). Fit as many data points as possible within ϵ. The outlier points define support vectors (Figure S1). SVR objective function: loss function is slightly different from squared error: insensitive to ϵ error; and L2 penalty for model complexity.

- Model: let B be the reference gene signature matrix, and f be the cell type composition, m the expression data of a sample (mixture). Fix SVR $m = f \times B$. Note: f is not constrained, so after fitting, remove all negative terms and normalize to sum 1. Earlier methods use non-negative model. Note: The method is applied only on microarray data.

- Construction of gene signature matrix: 500 genes in 22 immune cell types. Remove irrelevant features using DE analysis. Build a robust signature matrix (low condition number).

- Evaluation with simulation: 4 blood cell lines (known) and mix with 1 colon cancer cell line (unknown). Show that CIBERSORT performs well even when the cancer cell content is high, $> 50\%$.
- Application in solid tumor data: show the results agree with flow cytometry results with some cell types.
- Remark: a major challenge is multilinearity due to similar transcriptome of closely related cell types. Earlier methods have many problems, e.g. winner takes all. The paper deals it via construction of well-conditioned signature matrix and SVR.
- Remark: not explicitly model the unknown content. Can we obtain tumor expression by removing immune cells?
- **Remark**: the model is fit with each sample separately. If we analyze multiple samples jointly, can we borrow information across samples? Ex. cell proportions may not differ greatly across samples.

Comprehensive analyses of tumor immunity: implications for cancer immunotherapy (TIMER) [Bo Li and Xiaole Liu, GB, 2016];Revisit linear regression-based deconvolution methods for tumor gene expression data [Bo Li and Xiaole Liu, GB, 2017]

- Limitations of CIBERSORT: it uses 22 cell types as reference and focus on immune genes in these cells (LM22 genes), and it assumes that none of these genes are expressed in tumor. This is clearly violated: 1/4 of them are not immune specific.
- Ideas of TIMER: (1) focus on immune genes (filtering tumor genes). (2) Reduce colinearity of immune cell gene expression profiles with only 6 main immune cell types including B, CD4, CD8, DC, Macrophage, Neutrophils. (3) Avoid the constraint that the proportions sum to 1: which induce negative correlation of cell fractions.
- TIMER procedure (Figure 1): (1) Estimate tumor purity. (2) Batch effect removal: between TCGA and reference (6 cell types). (3) Purity based gene selection: filter genes show correlation with tumor purity. (4) Selection of immune signature genes (2000): also remove highly expressed (top 1%) genes, which have large variance and dominate the results. (5) Deconvolution for each sample: let Y^g be expression of gene g, and X_r^g be its expression in cell type r. Find f_r¿0s that minimizes $\sum_g (Y^g - \sum_r f_r X_r^g)^2$.
- Note: no constraint that f_rs sum to 1. They are not comparable among different cell types; only comparable across individuals for the same cell type.
- Validation of deconvolution results: simulations; pathology (known neutrphil levels); infiltrating leukocyte fractions from DNA methylation data.
- Association with clinical outcomes: Figure 3a, CD8 T cells often protective, and macrophages often associated with reduced survival. Other cell types: cancer type dependent.
- Possible causes of variation of immune cell fractions across individuals: need to correct for tumor purity. DC, B-cells increase with total mutation burden in some cancer types. Association with expression of specific chemokine-receptor pairs.
- Impact of collinearity: CIBERSORT reports many non-biological correlations. Some due to normalization (cell fractions sum to 1). Figure 1ab in the new paper: $r = -0.4$ for unrelated cell types. But some not: naive and memory B cells negatively correlated $r = -0.7$.
- **Lesson**: some issues important for deconvolution. (1) Reference transcriptomes should match the target data. This is particularly a problem for cancer samples, since reference transcriptome may not have cancer genes. TIMER: remove tumor genes by correlation with tumor purity.(2) Gene selection: highly expressed genes could dominate the results. (3) Colinearity: cells proportions need to sum to 1, which induces negative correlations.

2.8 Single Cell RNA-seq

ScRNA-seq: experimental considerations and data QC

- General design elements: (1) Cut RNA or DNA (for single cell epigenomics experiments) in cells. (2) Loading barcodes to cells. For Droplet experiments, this means fusion of drops, with additional enzymes (e.g. DNA ligase). (3) Amplification: add PCR handle at some step.
- Barcoding cells: important to have one barcode per cell. Need to validate this at the experimental levels: e.g. with two cells in a single droplet/barcode, it is impossible to detect with data analysis alone.
- Amplification/UMI: PCR can introduce bias/noises. Use UMI inserted during reverse transcription. Another strategy: linear amplification (in vitro transcription).
- Experimental approach to validation: Experimentally create mixture of cell populations and test if the fractions can be recovered.
- Data QC and filtering:
  - Test one barcode per cell.
  - Marker gene expression: in specific cell types, often combined with clustering analysis or marker genes in t-SNE plots.
  - Comparison of single-cell results vs. bulk results: pseudobulk (combine all cells; or cells within a cluster) vs. bulk expression data.
  - Filtering bad cells: not enough detected genes. Could be due to low input or low amplification efficiency.
  - Filtering lowly express genes.

Quantiative model of scRNA-seq data [personal notes]

- General model: let x_{ij} be the count of gene j in cell i, and R_i be the library size of cell i. Let λ_{ij} be the true expression level. Our model is:

$$x_{ij} \sim \text{Pois}(R_i \lambda_{ij}) \qquad \lambda_{ij} \sim (1-\pi_j)\delta_0 + \pi_j \cdot \text{Gamma}(\mu_j, \mu_j \phi_j) \tag{2.24}$$

  where π_j is the proportion of cells where gene j is not expressed, and μ_j is the average expression of j across cells, and ϕ_j overdispersion.
- Modeling true gene expression levels: in single-cell data, useful to have informative priors of λ_{ij}.
  - Gene-specific distribution: common in all cells, e.g. scimpute, MAGIC.
  - Dependency on other genes: e.g. by a linear model, SAVER.
  - Latent variable model: function of a smaller number of latent variables. f-scLVM, fastTopics (topic model), DCA.
  - Other models may be possible: e.g. clusters of cells, with different distributions at different cell types for a given gene.
- Modeling read counts: Poisson should be sufficient, no need of Zero-inflated NB. Normalization: (1)Spike-in: however, not perfect. (2) total UMI counts. (3) The scale factor depends on mean expression of genes (sctransform).

Clustering and trajectory analysis in scRNA-seq [personal notes]

- Clustering/distance metrics: Euclidean distance on genes do not work well for single cell expression. A common strategy: reduce dimensions first, e.g. PCA, and then use the PCs to create KNN graphs: two cells are connected only when they are close. So the KNN graph focuses on local distance (similar to t-SNE idea). The KNN graph can then be used for clustering.
- Factor analysis/latent variable approach: factor to gene relationship is linear: how multiple factors work together. Topic model: similar. DNN/VAE approach generalizes factor analysis.
- Manifold approach to clustering and trajectory analysis: one problem is the shortest distance of cells are sensitive to short circuits. Diffusion map: consider average distance of cells, so robust to short circuits.

Stanford CS 262: single-cell RNA-seq. `http://web.stanford.edu/class/cs262/presentations/lecture12.pdf`

- Applications of scRNA: (1) developmental biology: gene expression programs of cell types. (2) Cancer biology: model of cancer evolution (single vs. multiple clones, cancer stem cells), inferring timing of mutations and drivers, assess treatment (what cells are left after treatment). (3) Microbiology: identify rare bacterial that cannot be cultured, detect activities of genes.
- Background: flow cytometry: pass cells one by one, and collect statitistics of light scattering properties of each cell. FACS: cells are fluroescence labeled (e.g. on cell surface marker and antibody), and the machine applies charges to cells based on fluroescence, which allows cells to be sorted.
- Cell sorting: Manual. Microdissection. FACS. Microfluidics: sort antibody-producing B cells from all B-cells (size or charge?). Drop-seq.
- RNA-seq amplification (S45): from RNA, do reverse transcription (RT) and second strand synthesis to obtain cDNA. Then (1) PCR, or (2) Linear amplification: in vitro transcription (IVT), followed by RT.
- Challenge of amplification (Slide 43): one molecular in the sample initially may be amplified to multiple copies, distorting the count. Using UMI to count number of unique molecules.
- PCA: Identify experimental errors, batch effects. Visualize samples.
- t-SNE (S63): the idea of manifold learning, the data points are not clearly separated in Euclidean space, but do transformation, so that they become separable.
- Sequencing depth for scRNA-seq: Higher depth needed to detect rare cell types and lowly expressed genes (from 5M to 50M reads). With PCA: at thousands of reads, see only two cell types; but with 10M reads, see multiple cell types.
- Biological effects can complicate scRNA-seq analysis: e.g. cell cycle. In [Buettner and Stegle, NBT, 2015], show that removing cell cycle effects (ignore the cell cycle related genes?) discovers hidden population of immune cells.
- Result of ScRNA-seq: digital expression matrix, the read count of genes in each cell.

Single-cell sequencing-based technologies will revolutionarize whole-organism science [Shapiro & Linnarsson, NRG, 2013]

- Methods for single-cell isolation: first need to prepare single cell sample (usually via enzymatic disaggregation).
  - Micro-manipulation: e.g. serial dilution. Low throughput.

  - FACS: biased (e.g. cell surface marker) or unbiased (light scattering properties of cells) isolation. High throughput but require large number of cells in suspension.
  - Laser-capture microdissection: cannot control single cells.
  - Microfludic devices.

- Single cell genomics applications:
  - Reconstruction of cell lineages: somatic mutations.
  - Lineage reconstruction of cancer: metastasis origin, from random cells or from specific clones (subpopulations).

- Methods for single cell genomics: whole-genome amplification, library prep and sequencing. Key is high-fidality, low bias WGS. Alternative, single molecule sequencing. Multiplexing using DNA barcodes.

- Single-cell RNA-seq applications:
  - Identification of rare cell types. Ex. adult stem cells. Cricial question is which part of transcritomes are relevant.
  - Study of dynamic changes; and stochatic processes.

- Single-cell RNA seq methods:
  - Amplification is subject to bias or Monte Carlo effect (the stocastic events in the first few cycles of PCR amplified exponentially).
  - UMI: molecular counting.
  - Severe loss in RNA-eq: 80-90% of mRNA were lost.
  - Single cell transcriptome prep: in vitro transcription (IVT), homopolymer tail (amplication of cDNA by PCR), template switching.

The Technology and Biology of Single-Cell RNA Sequencing [Kolodziejczyk & Teichmann, Molecular Cell, 2015]

- Procedure of scRNA-seq: Figure 1.

- Single cell isolation: cell dissociation could affect transcription profiles, e.g. enzyme treatment.
  - FACS sorting, then microtiter plates. Could do index sorting: record the fluorescence and sizes of all cells.
  - Microfludic device: 96 cells/chip. Advantage: nanoliter volume, thus low regent cost. Disadvantage: relatively homegeneous in cell sizes. Capture effeciency may be low for sticky or non-spheric cells.

- Reverse transcription (RT): standard method is polyT priming. Add polyT linked to other sequences (barcode for each cell, primer used in amplication), then reverse transcriptase only synthesize mRNA. For the second strand synthesis: (1) PolyA tailing: add polyA to the new strand (serving as reverse primer) (2) template switching: a special enzyme that adds some sequences to the 3' end (serving as reverse primer).
  - Remark: with RT, only mRNA becomes cDNA (with primers).
  - 80-90% mRNA is lost in this step (not reverse transcribed).

- Amplification: PCR or in vitro transcription (IVT). Primers presumably have been added in the RT step. PCR: the disadvantage is that the amplification ratio depends on base content.

- The use of UMI during RT can better capture quantitative information: one UMI per mRNA molecule, thus all PCR duplicates have the same UMI. Multiple copies of the same transcript have different UMIs (before amplification) - the UMI sequences are random.
- QC: mapping statistics, mismatch rates, fraction of mapped reads, number of detected genes.
- ERCC spike-ins: mainly capture technical noises.
- DE analysis: a popular tool is MAST. For unsupervised analysis: PCA, MDS, principal curve on sc-RNA data.

Design and Analysis of Single-Cell Sequencing Experiments [Grun and van Oudenaarden, Cell, 2015]

- ScRNA-seq amplification: CEL-seq protocol, using T7 promoter to transcribe cDNA into mRNAs. Different protocols differ in their coverage of transcripts: 5, 3 end or whole transcript. The priming strategy determines the coverage (where the primers are located): e.g. CEL-seq only 3 end, STRT-seq enriched for 5 read, and SMART-seq2 whole transcript.
- Spike-in RNAs: ERCC, some biological difference with cellular mRNA (shorter RNA, shorter polyA and not spiked directly into cells), so not a good standard for quantification.
- Design of scRNA-seq experiments: main considerations are number of cells and sequencing/library complexity. (1) Consider bias in cell purification and isolation, e.g. different sizes. (2) Adjust sequencing depth so that every transcript is sequenced at least 3-4 times. Also consider minimizing batch effects: cells from different conditions sequenced on same lanes.
- Read QC: standard tools, e.g. fastQC. Generally remove non-unique reads. Note the quality of gene model can have a large impact (esp. protocols that enrich 5 or 3 reads): possible to refine the 5 and 3 ends using data.
- Cell QC: filter cells with low yield (total transcript count), due to RNA-degration (low-input) or low amplification efficiency. The two can be distinguished using spike-ins: which should be stable across cells. Note: when cell volumes can differ substantially, should only do mild filtering.
- Expression normalization: use TPM or RPKM, defined on all cells or each cell separately? Analysis: read counts of a gene in a cell is influenced by the total transcriptome in the cell, and the relative level of that gene. The total transcript levels can different substantially across cells, which can be corrected using spike-ins (the number of spike-in reads vs. spike-in amount gives the amplification efficiency of a single cell). However, this is not perfect. Often, what matters is the relative transcript level per gene per cell. Downsampling is the recommended strategy.
- Cell type identification from scRNA-seq: PCA: often limit to 2 or 3 PCs. t-SNE: often group outliers. MDS. Other clustering algorithms.
- Challenges of cell type identification: the presence of confounding factors, both technical (e.g. batch effect) and biological (e.g. cell cycle). Difficulty of identifying rare cell types: RaceID method, iteratively refine the clusters by using outlier cells as cluster seeds.
- Identification of marker genes: DE test, some methods are designed for scRNA-seq data, to deal with high drop-outs.
- Studying differentiation dynamics: pseudo-temproral ordering.
- Studying gene expression noise: Promoter bursting creates additional noise (beyond sampling noise: Poisson). Deconvolution of technical and biological noise from the distribution of transcript counts across cells: Use spike-in to estimate technical noise, and infer the biological variability.

Challenges and emerging directions in single-cell analysis [Yuan, GB, 2017]

- Challenges of scRNA-seq analysis: (1) Transcript dropout (2) Batch effect (3) biological factors such as cell cycle, cell size and state.
- Major directions: spatial transcriptomics, multi-omics.
- Learning about differentiation: the cell states defined by transcriptomic patterns are surprisingly continuous instead of forming distinct, transcriptionally defined groups. Need to distinguish natural variations of the same cell types with functional state transitions.

Manifold learning-based methods for analyzing single-cell RNA-sequencing data [Moon and Krishnaswamy, COSB, 2018]

- Background: Why PCA is not enough for dim. reduction? Ex. suppose we have a latent variable, representing cell fate (0 to T). But for many genes, their relationship with the fate is non-linear: increase expression up to time t, and then decrease from t to T. To capture this, PCA will use two latent variables, Z_1 for 0 to t and Z_2 for t to T, and expression is modeled as: $X = Z_1 w_1 + Z_2 w_2$, where $w_1 > 0, w_2 < 0$.
- Manifold assumption of scRNA-seq: single cell transcriptome states vary continuously.
- Data diffusion approach to learn manifold (Figure 1B): compute distance of cells, kernel function on distance (make distance more 0/1 like), graph that represent only local relationships, random walk on the graph. Distance can be then be defined on the manifold (based on random walk).
- Denoising and learning gene-gene relationship: (1) Existing work: impute only zeros; or use linear model to impute (SAVER: learn mean expression of a gene using linear regression from other genes). (2) MAGIC: projection of data points to manifold. Better gene-gene relation with imputed expression.
- Psuedotime (Figure 3): general steps are Gene selection; manifold learning: embedding of cells into the manifold; Cell organization. (1) Monocle 2: a tree embedding, then pseudotime based on distance along the tree. (2) SLICER: KNN graph based on locally linear embedding, then ensemble path finding to learn trajectories. Not assume bifurcation. (3) Diffusion pseudotime (DPT).
- Dimensionality reduction and visualization: PCA not good for visualization, but still good for reducing dimensions (e.g. a few hundred PCs). t-SNE: focus on shorter distance, so global distance in t-SNE plots are not very meaningful; tendency of t-SNE to fragment progression into clusters.
- Clustering: (1) Spectral clustering. (2) Manifold distance based clustering. Ex. EAC-DC method: manifold distance based on MST.

Current Best Practices in Single-Cell RNA-seq Analysis: A Tutorial [Luecken and Theis, MSB, 2019]

- Raw data processing: e.g. Cell Ranger. Construct count (or read) matrix from data: read QC, assign reads to cells, alignment, UMI, etc. Note: a barcode may not match to a single cell, could be doublet or empty cell.
- Cell QC: the goal is to remove dying cells, doublet cells and other problematic cells (cyto. mRNA leaking). Remove outliers based on three metrics: count depth, gene detection rates and percent of mito reads (should be lower than or around 15-20%). Risk of using one metric: e.g. high count depth may be from cells of large size, instead of doublet. Use multiple metrics jointly.
- Gene QC and other QC guidelines: (1) Gene QC: remove transcripts in too few cells. General guideline: expected size of rare cell population, and dropout rate. (2) Revisit QC metrics after downstream analysis.

- Normalization: (1) Global scaling: obtain size factor per cell. Library size, median count depth of genes. Recommend scrun: linear regression over genes: 50% of genes show no DE. (2) Model based scaling: regress out technical/biological covariates (batch, cell cycle, detection rate of a cell) and library size. (3) Gene normalization: for full-length RNA-seq, use TPM. O/w use CPM. Further normalization (e.g. make genes standard normal) may depend on specific downstream analysis. (4) Log-transformation: if downstream analysis assumes normality, use $\log(x+1)$ transformation. However, this transformation can introduce spurious DEGs.
- Regressing out biological and technical covariates: general strategy is to use linear model to regress out the covariates jointly. (1) Biological: cell cycle, sometimes mito. expression as a proxy of cell stress. (2) Technical covariates: count depth. Normalization can address count depth, but it does not work for genes not expressed.
- **Analysis**: read depth of a cell is a scaling factor needed to adjust for, but it is also a proxy of biological factor(s) that vary across cells, e.g. cell cycle, cell states. This may explain why it is not enough to just linear scale reads per cell.
- Batch effect correction: assumption is that cell type and compositions are the same. (1) ComBat method: both mean and variance of a gene depends on the batches (this can be estimated for each gene, and then use EB to shrink). (2) Evaluation of batch effects: Figure 3, how cells are grouped in UMAP. (3) Best strategy is to deal with batch effect experimentally: e.g. mix samples of different individuals, then use genetic marks to determine sample identity.
- Data integration: from different groups/experiments. Cell types and compositions may be different. Remark: batch effect vs. data integration is similar to fixed vs. random effects.
- Expression recovery: be cautious, may be best for visualization, but not testing specific hypothesis.
- Feature selection: often use highly variable genes, measured by variance/mean. 100-5000 genes.
- Dimensionality reduction: PCA, often used as a pre-processing step. Diffusion map.
- Visualization: UMAP, better than t-SNE.
- Summary of data pre-processing: (1) Gene level analysis: DEG, use measured data. (2) Cell type analysis such as trajectory analsyis and clustering: use reduced data.
- Clustering: suppose distance metric is given (often use PCA), then two options (1) Distance based clustering algorithms, e.g. K-means. Correlation based distance seems to work best. (2) Graph partition: KNN graph (each cell is connected to K nearest neighbors, K = 5-100), then Louvain community detection algorithm. Recommended strategy.
- Cluster annotation: (1) Decide level of clusters: cell types or cell states. (2) Visualization (Figure 6): UMAP, but color cells by clusters. Then annotate the clusters. (3) Marker gene derivation: DEG analysis (t-test or rank-sum). Note the ascertaintment problem because clusters are learned from expression data, so p-values are inflated (need permutation). (4) Cell type assignment: comparison of marker genes with classical markers or markers in reference dataset. Note that marker genes of the same cell types may different because cell type composition may be different from reference. (5) Also possible to skip clustering step, and use reference cell types to assign cells (e.g. scmap).
- Compositional changes across conditions: current work uses Poisson GLM for cell counts, however, cell types are not independent. Better to use compositional analysis.
- Trajectory analysis: recommend PCA-based methods (for simple cases) and PAGA. Can be combined with clustering: e.g. show cell type dynamics along pseudotime (by a graph, which cell types becomes other cells - Figure. 7D). Also RNA velocity analysis can give direction. Regress out uninterested biological effects: e.g. cell cycle.

- Gene expression dynamics: how expression changes over pseudotime, e.g. by regression analysis to find such genes with correlation with pseudotime - some may represent important regulatory genes.
- DE testing: should use measured data, not corrected data. Recommend: MAST (fast) and limma-voom.
- Popular platforms: Scater good for QC and pre-processing. Seurat: a variety of tools. Scanpy: Python based. GUI software.
- Future directions: (1) Deep learning: from batch effect correction to downstream clustering and trajectory analysis. (2) Multi-modal single-cell data.

Wishbone identifies bifurcating developmental trajectories from single-cell data [Settyl and Peer, NBT, 2016]

- Method overview: the basic idea is to create KNN graph that represents local similarity of cells. From the graph, one can obtain the distance of a start cell (early cell) to any other cell: pseudotime/trajectory. Two problems: (1) shortest paths are sensitive to short circuits: two distant cells happen to have short distances by chance. (2) The trajectory is noisy for distant cells: errors accumulate.
- Graph construction and initial ordering of cells: from N cells and M markers, use diffusion map to create a low-dim. embedding. This is more robust to short circuits than naive shortest paths. The Euclidean distance in the low-dim. is used to obtain cell distance, and KNN graph. The shortests path from s to all other cells are the initial trajectory, $\tau^{(0)}$.
- Refining trajectory using waypoints: this is to address the problem of error accumulation along the trajectory. The idea is to use multiple waypoints instead of the single s cell. So for each waypoint, we can compute its distance to all cells: the matrix D. However D is not aligned (not starting from the same start cells). Given a random cell as waypoint w: the distance of s to i from the waypoint of w is roughly the trajectory from s to w, then distance w to i. This aligns D, and create a matrix P. Finally, we put weights on waypoints, the results are the refined trajectory $\tau^{(1)}$.
- Identify branch points and assign cells to branches: this is based on the idea that if two way points belong to two different branches, their distance to other cells can be very different; whereas two waypoints in the trunk should be have similar distance. So we can cluster waypoints into branches.
- Application of Wishbone: (1) T cell differentiation: Figure 2, change of marker expression along Wishbone trajectory, and the branch points. (2) ScRNA-seq data of myeloid cells: t-SNE plots with cells labeled by trajectory or branches (Figure 5).
- Remark: selection of markers is critical to the success.
- Lesson: embedding by diffusion map can help better measure distance of cells (avoiding short circuits). Intuition: diffusion distance is based on averaging paths from source to target cells, while shortest path is sensitive to a single short circuit.
- Lesson: use pseudotime in explorative analysis: gene expression dynamics, t-SNE visualization of cells.

Gene Expression Recovery For Single Cell RNA Sequencing (SAVER) [Huang & Zhang, biorXiv, 2017]

- Background: zero-inflated model, not estimate expression of low-abundance genes. Imputation based on bulk RNA-seq: fail to capture cell-to-cell variation.
- Model: the count of gene g in cell c follows Poisson distribution $Y_{cg} \sim \text{Pois}(s_c \lambda_{cg})$, where s_c is the cell size factor (could be estimated as library size of cell, or from spike-in). The key is the prior distribution of λ_{cg}: Gamma prior, with mean estimated using other genes μ_{cg}, and dispersion parameter ϕ_g. SAVER obtains the posterior distribution of λ_{cg}.

- Estimating μ_{cg} and ϕ_g: for μ_{cg}, use Poisson GLM, where explanation variables are log-transformed, normalized read count of other genes. Use LASSO to shrink parameters to 0. To estimate ϕ_g: MLE, considering multiple assumptions of ϕ_g (e.g. constant CV, or constant variance).
- Use SAVER for gene-gene relationship: correlation, but adjusting for uncertainty of expression.
- Use SAVER for DE test: Wilconxin rank-sum test on recovered expression.
- Validation: mouse brain scRNA-seq data, take 3.5K highly expressed genes and 1.8K high coverage cells as reference, then do down-sampling with different efficiency. Assess the results by: (1) correlation of recovered gene vs. reference gene expression; (2) correlation of transcriptome of a cell, recovered vs. reference; (3) gene-gene and cell-cell relationship; (4) power of DE analysis. In all cases, SAVER does better: in particular, when efficiency is relatively high 25%, SAVER recovers expression well because of adaptive weighting.
- Q: Inference process, in estimating μ_{cg} with Poisson regression, the response variable is treated as given? Or fit both LASSO parameters and expression parameters simultaneously?
- Remark: in model fitting, should do NB regression. But the paper uses Poisson regression to obtain γ and $\hat{\mu}_{cg}$, then estimate overdispersion from NB distribution.
- Remark: comparison with scimpute and MAGIC: they use the expression of the same gene in other cells to impute/estimate expression. Intuitively, suppose we have multiple cell types, then expression of a gene in the cells of the same type would be highly informative. This is not implemented in SAVER: the correlation structure between genes γ_{gg} is constant. Idea: fit different covariance structure on different cell subtypes.
- Remark: to better model scRNA-seq data, we need to better understand biology: extent of stochastic variation; whether stochastic variations also correlate between genes; why only a small number of genes are predictive of the rest.

f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq [Buettner and Stegle, GB, 2017]

- Factor model: cell covariates (cell detection rate, etc), annotated factors, unannotated factors. For unannotated factors: some are dense (confounders), some are sparse (biological processes). Write as $Y = XW^T + E$, where X is factor states (across cells) and W factor weights. Use standard normal for X, for W, use spike-and-slab prior. For w_{gk} (factor k effect on gene g), let z_{gk} be the indicator:

$$w_{gk}|z_{gk} = 0 \sim \delta_0 \qquad w_{gk}|z_{gk} = 1 \sim N(0, 1/\alpha_k) \tag{2.25}$$

  where α_k is a measure of the importance/variance explained by factor k. This is the ARD prior: which will shrink w_{gk} towards 0 if α_k is large, but not much if α_k is small. For different types of factors, use different priors for z_{gk}: (1) Annotated factors: mostly determined by whether gene g belongs to the pathway of k; (2) Sparse unannotated factors: z_{gk} has small chance of being 1 (0.01); (3) Dense unannotated factors: z_{gk} has high chance of being 1 (0.99). Allow certain numbers of sparse and dense un-annotated factors.
- Error model: for Drop-seq type of data, (1) Zero-inflated model: model log-read count, but it is zero-inflated normal. The underlying variables F are continusous (normal) with $F = XW^T$, but data Y is generated from F with a high probability of 0. (2) Poisson model of read counts.
- Down-stream analysis: gene set refinement. Estimation of factor relevance by α_k. Imputation: using hidden values F.
- Application to neuronal dataset: projection of cells in dimensions defined by factors. Ex. use two factors for muscle contraction and innate immune system, cells clusters by cell types, e.g. microglia has high value for innate immune system factor.

Robustness QTL mapping in human iPSC [Abhishek Sarkar, 2018]

- Motivation: detecting QTL of gene expression variance.
- Experiment: 53 iPSC lines, Fludigm C1 (96-well plate). 50-280 cells per individual and about 2M reads per cell.
- Problem with PCA: use log-CPM, and do PCA. However, the top PCs are correlated with technical metrics, such as gene detection rate (read depth). Explanation: log-CPM is not mean-centered, and gene detection rate correlates with log-CPM (e.g. a cell with high gene detection rate will have higher log-CPM on many genes).
- Modeling scRNA-seq read counts: let i be individual, j be cell and k be gene. Let R_{ij} be library size, μ_{ik} be the mean expression level. We have two models, both accounting for excess of zero: (1) for gene k in a cell j: read count follow zero-inflated Negative Binomial.

$$r_{ijk} \sim \pi_{ik}\delta_0 + (1-\pi_{ik})\text{Pois}(R_{ij}\mu_{ik}u_{ijk}) \qquad u_{ijk} \sim \text{Gamma}(\phi_{ik}, \phi_{ik}) \tag{2.26}$$

  where u_{ijk} models cross-cell variation. (2) An alternative model is: the read counts follow Poisson distribution, but the rate follows a mixture distribution with zero component.

$$r_{ijk} \sim \text{Pois}(R_{ij}\lambda_{ijk}) \qquad \lambda_{ijk} \sim \pi_{ik}\delta_0 + (1-\pi_{ik})\text{Gamma}(\phi_{ik}, \mu_{ik}\phi_{ik}) \tag{2.27}$$

  Interpretation: a cell either does not express a gene at all, or express with the mean following Gamma distribution. The two models have the same likelihood.
- Comparing the two models: even though the marginal likelihood is the same, the interpretation is very different. FISH data supports the second model. Idea: use spike-in, and house-keeping genes (expected to be always expressed) to compare the two.
- Simulation to assess the power: mean can be generally estimated, however ϕ and π are much harder to estimate, and need larger number of cells.
- Accounting for confounders: add $e^{x_{ij}\beta}$ term to the Possion read count, where x_{ij} are confounders.
- Results: 200 mean eQTL, 100 variance-QTL. However, almost all variance QTL can be explained by the mean effects.
- Discussion: nonlinear dim. reduction using DNN. Deep generative modeling: latent variables, generate the gene expression rates, then Poisson noise.
- Remark: can we distinguish technical dropout and biological variation of gene expression? Technical dropout rates should not vary much across genes, while biological variation does. Also can we use heterozygous genotypes to infer technical dropout rates?
- **Lesson**: (1) PCA can be problemtic: data needs to be properly normalized. (2) Single-cell read count modeling: on-off expression pattern of genes.

Bias, robustness and scalability in single-cell differential expression analysis [Soneson and Robinson, NM, 2018]

- ScRNA-seq repository: conquer. 40 datasets, most are full length.
- Simulated dataset: choose 7 large datasets, and two predefined groups of cells. (1) Null simulation: subsampling from the same population of cells. (2) Signal simulation: 10% genes are chosen to be DE, with fold change sampled from Gamma distribution with mean 4, shape 2. The mean and overdispersion were estimated by edgeR. Filtering: retain only genes with TPM > 1 in at least 25% of cells.

- Type 1 error control: assess proportion of genes with $p < 0.05$. Most methods struggle to get null distribution of p-values. Much better after filtering.
- FDR control (Figure 4ab): SeurateBimod, DEseq2, monocle do not perform well. EdgeR/QLF performs bad before filtering: tend to call lowly expressed genes with many zeros significant, but ok after filtering.
- Power (Figure 4cd): among methods with good FDR control, edgeR/QLF, SAMSeq, DEsingle and voom-limma achieve high power.
- Summary: recommend filtering, and methods edgeR/QLF, MAST, limma and voom-limma. SCDE: not scale well and only do two-group comparison. MAST: single-cell methods, good FDR control even before filtering, and good power; sensitive to whether to include cell detection rate as a covariate.
- Remark: EdgeR/LRT does not perform well in controlling FDR. EdgeR/QLF: quasi-likelihood F-test instead of χ^2 approximation.

Single-cell reconstruction ofdevelopmental trajectories during zebrafish embryogenesis (URD) [Farrell and Schier, Science, 2018]

- Data: early zebrafish development, 12 time points, 38K cells, scRNA-seq data. t-SNE plot of all cells: developmental points are clear drivers, but no obvious patterns.
- Exploratory analysis of each stage (Figure 1C): early time points, not much separation. Later points: clearly separate cell types.
- URD algorithm (Figure 1D): KNN graph, then define a root cell and multiple tips. Do diffusion from the tip to the root: biased random walk s.t. one can only move to the same or earlier time point. From the paths of the tips, define the branch points as where multiple paths merge.
- Q: no single path from a tip to a root, diffusion defines distance but not paths. How are branch points defined?

RNA velocity of single cells [La Manno and Kharchenko, Nature, 2018; presented by Rodrigo]

- Background: in both bulk and scRNA-seq, 15-20% reads are unspliced. In 10x (polyT priming), have discordant priming from the more commonly occurring intronic-polyT sequences (polyT priming > first cDNA strand > intronic polyT priming > reads).
- ODE model of mRNA dynamics (Figure 1B): let u be the unspliced transcripts, and s spliced transcripts. We have $du/dt = \alpha - \beta u$ and $ds/dt = \beta u - \gamma s$, where α, β, γ are transcription, splicing and degradation rate, respectively. For simplicity, scale with β, i.e. divide all rates by β, or equivalently set $\beta = 1$.
- Approximation: for the purpose of estimation and extrapolation (1) ds/dt = constant. (2) u = constant. Note: assuming degradation is slow and splicing is fast, we will have u constant. **Intuition**: number of unspliced reads u gives the transcription rate (normalized by splicing rate), and the ratio of unspliced/spliced reads u/s gives the degradation rate.
- Fitting parameters: steady state assumption across cells. So $u = \gamma s$ and $\alpha = u$. Solving γ by fitting s vs. u over many cells. Use KNN pooling of both cells and genes (correlated expression) to get better estimate of γ.
- **Remark**: the idea is that on average, most genes or cells are in equilibrium, so we can estimate γ. Then for specific genes and cells, we can estimate the velocity using values of γ and u in that gene and cell.

- Variation of γ across conditions: many genes, do not change much u/s constant. But some genes do vary.
- Results: from RNA velocity field, follow (backwards) of velocity vectors will reach the precursor states. Show in the mouse hippocampus data: found radial glia cells/NPCs.
- Q: How the rates and levels are related to read counts? Need to account for intronic poly-T sites, intron length, etc.
- Q: Gene-specific splicing rates. How can this be normalized?

Deep generative model of single-cell RNA-seq data [Abishek, 2019]

- Basic model of read counts in single-cell RNA-seq: let x_{ij} be the read count matrix, i for cell and j for gene. We have: $x_{ij} \sim \text{Pois}(s_i\lambda_{ij}), \lambda_{ij} \sim g_j(.)$, where s_i is the library size, and $g_j(\cdot)$ is the distribution for gene j.
- Prob. PCA: data vector for sample i, $x_i|z_i \sim N(Wz_i, \sigma^2 I)$, where z_i is the low-dim. representation (latent factor) of sample i, and $W = [w_1, \cdots, w_q]$ ($p \times q$ matrix) is the q PCs.
- Poisson matrix factorization: LDA, PCA or NMF for mean of Poisson. $[\lambda_{ij}] = LF$, L is PCs in each cell, and F is loading matrix. Bayesian Poisson factorization: Gamma prior.
- Accounting for noise: zero-inflated NB, the mean of NB is mixture of 0 and LF (matrix).
- Poisson approximation: $x_{ij} \sim \text{Pois}(\exp(\eta_{ij}))$, and $\eta_{ij} = LF$. However, the variance is no long constant. Srebro and Jaakkola, 2003.
- Variational autoencoder (VAE): (1) Decoder: from low-dim. representation (z) to data (x); (2) Encoder: from data (x) to low-dim. representation (z). Inference of model parameters is done via variational Bayes.
- Using VAE in Poisson read counts: let λ_i be the vector of mean expression of all genes in cell i. To model $\lambda_i|z_i$, we use $\pi(Z_i)$ and $\mu(Z_i)$ (mean and proportion of nonzero), both are output of DNN decoder. Add, we have a encoder model of $z_i|x_i \sim N(\mu(x_i), \sigma^2(x_i))$.
- Variational inference: $p(x|z)$ and $q(z|x)$. ELBD: $\text{E}_q[\ln p(x|z)]$. Sample q from MVN, difficult. Answer: univariate normal to approximate. Stochastic objective function (q changes): goal is to min. mean.
- Local optimum problem in DNN: sigmoid function is insensitive to gradient.
- Reconstruction: in PCA, we keep the matrix L and can reconstruct. In VAE: the encoder learners a mapping $x \to z$, and then we can use it to new data x'.
- Disentangled representation, e.g. handwritten digits, both style and content. Solution: in Autoencoder, add additional latent labels, y_i, z_i. Or we could have observed labels y_i. To use it in scRNA-seq: we have additional label y_i (batch). Intuition: add confounder term in x_{ij} model. Extension: make it VAE, change $\mu(z_i)$ to $\mu(z_i, y_i)$. But decoder does not use the labels.
- Results: real scRNA-seq data, best are obtained with NMF, NMF (with Poisson approximation of normal).
- Applications: denoising, interpreting deep embeddings, recovering trajectories.
- Q: any example of non-linearity in scRNA-seq? Why the VAE does not perform as well as other simpler methods?
- **Lesson**: prob. interpretation of PCA. Can be generalized to count data, where mean (or log-mean) is modeled.

- Remark: object recognition in image analysis vs. scRNA-seq analysis, similar problem of batch effect: light background. How is this problem addressed in object recognition?

Single cell RNA-seq denoising using a deep count autoencoder [Eraslan and Theis, NC, 2019]

- Motivation: (1) Nonlinear gene dependency: not captured by SAVER. (2) MAGIC: not account for count data and dropout.
- DNN architecture (Figure 1): learn latent representation of gene expression of a cell, capturing nonlinear gene-gene dependency. (1) Input layer: gene expression matrix. Expression is normalized: divide by size factor (library size divided by median library size), and log. transformed, and z-score normalized. (2) Output layer: three outputs, corresponding to three parameters of gene-specific distribution, dropout rate, mean and overdispersion and NB, π, μ, ϕ. Use exponential for μ and ϕ, and sigmoid for π. Note: three outputs per gene per cell. (3) Hidden layers: 64, 32, and 64 neurons (the middle layer is bottleneck).
- Parameter estimation: loss function is changed to MLE, with NB or ZINB density, also regularization term, ridge prior on dropout.
- Denoising: replace count values with the mean of expression $\bar{M}$ from the output layer.
- Application to cell clustering in simulation: after denosing, the simulated data show better clusters.
- Application of denoising to gene DE analysis: both single cell and bulk RNA-seq available in hESC differentiation. DE analysis is done by DESeq2, assuming NB (without zero inflation). After denoising, several outlier genes in original scRNA-seq DE analysis are removed (Figure 5B). Comparison of DE genes in scRNA vs. bulk: 20 random cells (many times), compare estimated fold changes in scRNA-seq, DCA gives high correlation (0.9), better than MAGIC (0.87) and ScImpute (0.85) and SAVER (0.76).
- Remark: in DE analysis, only used the denoised data (replaced counts). But overdispersion and dropout have already been estimated by DCA, so its strange that they need to be estimated again by downstream tools (DEseq2).

Normalization and variance stabilization of single-cell RNA-seq data usingregularized negative binomial regression (sctransform) [Hafemeister and Satija, 2019]

- Background: main technical variation is sequencing depth variation across cells (number of molecules/genes detected per cell). May vary order of magnitude across cells.
- Why sequencing depth variation is a problem? Two approaches for normalization: (1) We can just normalize expression by library size or size factor of a single cell or pooled group of cells. (2) NB based model of read counts. Rely on per-gene error models. Answer: the relationship of gene expression vs. library size varies across genes, e.g. in cells with large library size, genes may be more uniformly expressed (proportion of genes change depending on total RNA molecules).
- A single scaling factor does not normalize low and highly expressed genes. While UMI counts per gene scale with sequencing depth, the exact relationship vary. Consider 6 gene groups by expression (UMI counts): do log2(UMI/library), the relationship vs. library differs across groups. Ex. for high abundance genes, the relationship is non-linear (expression saturation at high cell UMI counts, so the log2-ratio actually decreases).
- NB regression to remove the effect of sequencing depth: fit the model, let x_j be expression of a gene in cell j, and m_j be the cell UMI count. The model is:

$$x_j \sim NB(\mu_j, \theta) \qquad \log \mu_j = \beta_1 \log m_j + \beta_0 \tag{2.28}$$

Fit the model for each gene. However, the estimates are not stable (from resampling).

- Regularized NB regression: after fitting β_1, β_0, θ for each gene, kernel regression of these values vs. mean gene expression. The results are regularized parameters for each gene.

- Normalized expression data: Pearson residual. Using the fitted regularized NB model above, define the residual of gene i in cell j as:

$$z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}} \qquad \sigma_{ij} = \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}} \tag{2.29}$$

  where μ_{ij} are expected value from the fitted NB model. Show that they are now independent of cell UMI counts (Figure 3A)

- Using regularized NB regression removes the effect of sequencing depth, while preserving biological sources of variation. If there is no additional source of variation (i.e. NB explains all variation), we expect Pearson residual to be $N(0, 1)$. This is the case for most genes (Figure 4AB). Genes with higher than usual variance may represent biological variation: enriched with immune genes. Show that if using Poisson or NB (per gene) model, this is not true. Some genes now have much smaller variance due to overfitting (Figure 5B).

- Downstream analysis: dim. reduction (Figure 6A). In PCA, its clear that if using log2 transformation, PCs are correlated with cell UMI counts. No such effect in PCA from Pearson residuals.

- Downstream analysis: differential expression test. Use t-test on Pearson residuals, on genes detected in at least 5 cells in one of the groups being compared. Results: log-transformation gives 2000 DE genes while Pearson residuals only 11 (should be 0).

- **Lesson**: to obtain gene expression values that are independent of technical factors (sequencing depth per cell), while retaining biological source of variation. To assess results: normalized expression should be independent of technical factors; variance of expression across cells should reflect biological difference.

- Remark: how do we interpret the dependency of scaling factors on mean gene expression? Ex. proportion of high abundance genes (in read counts) is lower in cells with high UMI vs. cells with low UMI. This could be due to an observed confounder: e.g. high-UMI cells is associated with higher cellular activities, leading to expression of a broader set of genes.

PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells [Wolf and Theis, GB, 2019]

- Challenges of trajectory analysis and limitations of current approaches: (1) Incomplete sampling: many cells in the trajectory are missing. (2) Undirectedness: in methods such as diffusion map, random walk is totally undirected, but in reality, cell differentiation is not reversible.

- Construction of PAGA graph: compute KNN graph of individual cells (cell graph). Then cluster cells and define distance between clusters: based on the number of connections vs. expected number. This is the PAGA graph G^*.

- Define pseudotime (PT): use PAGA graph, instead of single-cell graph, which constrains diffusion to the paths in the PAGA graph (corresponding to differentiation). We then have a coordinate system (G^*, d), where d is diffusion PT (DPT) in the PAGA graph.

- Results on Hematopoiesis (Figure 1): show that even with much smaller number of cells, can reconstruct the path: changes of marker genes match expectation.

- Question: how does PAGA compare with Monocle?

Topic modeling with single cell count data (fastTopics) [Peter Carbonette, NHS, 2020]

- Zheng et al PBMC 68K data. Processing: normalization (log-reads), then PCA, and use PCs to do clustering (k-means). Only 2.7% non-zero counts.
- Application of topic model to PBMC data: $k = 11$ topics, then do t-SNE plot. Q: how to relate topics to particular cell types? Use known labels of B cells? Or compare factor expression to B cell transcriptome?
- Fitting topic model: usually EM, but not very good, not converge to min. LL.
- Poisson NMF: no overdispersion or zero-inflation. Adapt computation for sparse data.
- Analysis: each topic is mean read count per gene (over all genes); and each sample/cell is a combination of topics. Remark: need to normalize by library size in each cell (the relationship may vary with mean expression) - simple, total read counts.
- Topic models: read count of gene j in cell i: $X_{ij} \sim \text{Pois}(R_i \mu_{ij})$, where R_i is the library size of cell i. The expected rate $\mu = LF^T$, each row of l sum to 1; each column of f sum to 1: (1) Multinomial selection of topics L_i for each cell, which should sum to 1; (2) Expected proportion of reads for all genes for a given topic: F_j's should sum to 1.
- Optimization: LL with the constraint. EM; and cyclic coordinate descent; also sequential quadratic programming.
- To compare methods: show how LL changes over time. Tip: EM gets to some reasonably good solutions fast; use those as initial solution and use CCD.
- Topic 11: CD14+ monocytes, strong loading in some non-monocytes.
- Q: how can we test DE (differential topics) upon treatment? Do we have uncertainty of fractions?
- Remark: sparsity assumption may not be necessary here. We have a large number of cells to learn the factors. Given each cell, the topics may not be sparse; and for a given topic, the genes also may not be sparse.
- Ref: cisTopic, Gonzalez-Blas, NM, 2019 (for ATAC-seq data). GLM-PCA: similar to Poisson NMF.

Topic model (fasttopics) and Poisson NMF [Peter Carbonette, manuscript, 2020]

- Topic model: let $i, 1 \le i \le n$ be the cell/document index and $j, 1 \le j \le p$ be the gene/word index. Let w_{it} be the gene at read/word t of cell i. Let F be $p \times K$ matrix of factors/topics, as multinomial parameters over gene/word frequencies with $\sum_{j=1}^{p} f_{jk} = 1$. Let L be $n \times K$ matrix of factor to sample loading, as topic distribution of a cell/document, with $\sum_{k=1}^{K} l_{ik} = 1$. Let z_{it} be the topic of read t of cell i. Our model is:
$$z_{it} \sim \text{Mult}(l_i) \qquad w_{it}|z_{it}, F \sim \text{Mult}(F_{z_{it}}) \tag{2.30}$$
where l_i is the topic composition of cell/document i, and F_k is the word/gene composition of topic k.
- Simplification of topic model: we can merge all genes/words in a same cell. Let x_{ij} be the read count of gene j in cell i and $m_i = \sum_{j=1}^{p} x_{ij}$ be the library size of cell i. We have now:
$$P(x|F, L, m) \propto \prod_{i=1}^{n} \text{Mult}(x_i|m_i, \pi_i) \qquad \pi_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk} \tag{2.31}$$

- Poisson NMF and connection with topic model: we can write the model as:

$$x_{ij} \sim \text{Pois}(\lambda_{ij}) \qquad \lambda_{ij} = \sum_k l_{ik} f_{jk} \tag{2.32}$$

The advantage is that there is no equality constraint on the parameters. To relate to the topic model, we can normalize F s.t. each column sums to 1 (word/gene composition of a topic). Then l^*_{ik} (normalized) means topic-specific rate of a document. Let $s_i = \sum_k l^*_{ik}$ be the size factor of cell i. We have:

$$P(x|F, L, s) = \prod_{i=1}^{n} \text{Mult}(x_i|m_i, \pi_i) \cdot \text{Pois}(m_i|s_i) \tag{2.33}$$

Under this model, we have size factor as additional parameters.

- Summary: how library size is incorporated in the models? Topic model: library size is treated as data to be conditioned on. Poisson NMF: library size is modeled as size factor, which is estimated by the model.

2.8.1 Single Cell RNA-seq Technologies

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets: Drop-seq [Macosko & McCarroll, Cell, 2016]

- Technology: The beads are in the droplet. Use microfludic device so that each cell loaded into only one droplet, with one bead. The bead contains (1) constant sequences for using as primers in PCR and sequencing. (2) cell barcode. (3) UMI. (4) Oligo-dT to capture mRNA.
- Capture effeciency: each cell, only 12-15% of mRNAs are captured.
- Q: Strategy of the synthesis of the second strand?

Mapping the Mouse Cell Atlas by Microwell-Seq [Cell, 2018]

- Agarose gel, cells loaded into wells, then washout. Then add beads which contain oligont. with barcodes (400K per bead).
- Results: 400K mouse single cells. Results quality comparable with existing methods.

2.8.2 Single Cell RNA-seq Studies

Mapping kidney cellular complexity [Science, 2018]

- 60,000 single-cell transcriptomes from adult mouse kidney and resolved them into 21 cell types.
- Mapping disease genes to cell types: for one Mendelian disease, all genes are mapped to a particular type of epithelial cell. Also found a single type with complex trait SNPs/genes.
- Lesson: even if an organ has many cell types, a disease may be caused by changes in only one type. Ex. epithelial cells in intestine may be much more susceptible to damages.

Single-cell sequencing paints diverse pictures of the brain [Nature, 2018]

- ScRNA-seq in several regions of mouse brain. Inhibitory neurons are similar; but excitatory neurons are different, related to different functions of neurons: preparing or initializing motion.

Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects [Nature, 2019]

- Computational method to predict lineage-specific TFs: opposing lineage specific TF pairs that show a significant ratio change upon differentiation.
- ScRNA-seq of heart differentiation: identify Hand2 TF. Validation of Hand2 function in heart development using mouse model.

Single-cell transcriptomic analysis of Alzheimers disease [Nature, 2019]

- Data: 48 individuals (24 cases and controls), 80K cells by single-nucleus RNA-seq. Found major cell types: excitatory, inhibitory neurons, astrocytes, oligodendrocytes, microglia, oligodendrocyte progenitor cells.
- DE analysis across all cell types: heterogenous changes among cell types. Most of DEGs are cell type specific. Myelination-related processes were recurrently perturbed in multiple cell types,
- Association of expression changes with disease progression and related traits: find co-expression modules that are associated with related traits. Modules over-expressed with LOAD genes.
- Different sub-populations/cell states: difference in patients vs. control.
- **Lesson**: how to analyze scRNA-seq data of disease vs. controls. Focus on cell type specific changes; module analysis/association with phenotypes; cell state and composition difference between cases and controls.

2.9 Spatial Transcriptomics

Spatial reconstruction of single-cell gene expression data [Satija and Regev, NBT, 2015]

- Problem: suppose we have scRNA-seq data from dissociated cells, and in situ data of a small set of landmark genes, map the cells from RNA-seq experiments to spatial regions. We assume that the expression data in the reference map is binary.
- Model idea: we can assign a cell to its most likely bin, by similarity of expression of landmark genes. To account for uncertainty, we compute the probability that a cell comes from any bin. So we need a model that links digit expression (reference map) with continuous expression (RNA-seq).
- Imputing gene expression: use Lasso to predict expression of a gene based on all other genes.
- Learn expression model for every gene: if a gene is ON (or OFF) in the reference, whats the distribution of its expression in RNA-seq? Use a simple two-component normal mixture data in scRNA-seq data. Fix the mixing parameters as the proportion of bins with ON state. Furthermore, allow cells to flip between ON and OFF states based on average expression of all genes: given a gene, two population of cells (ON or OFF), compute the cluster mean, and for each cell, reassign it to one of the two clusters based on overall distance.
- Constructing initial map: assuming independence of genes and the same parameters for expression models, assign each gene to a bin with normal distributions, computing the posterior.
- Refining the spatial map: to model covariance of genes. Idea: for each bin, we have the initial map, then we use all cells assigned to that bin (based on L2 distance) to estimate mean and variance of MVN.
- Evaluation: Zebrafish embryo data, 64 bins, and 900 scRNA-seq samples. (1) 28 reference cells with known locations, show that the inferred bins are very close. (2) Cross-validation: take one gene out from reference map and predict its spatial expression.

- Application: predict spatial expression of many other genes, then obtain the spatial pattern of these genes, and find rare cell populations (and marker genes).
- Remark: let X_{ij} be expression of gene j in cell i, and Z_{bj} be the binary expression in bin b. Let B_i be the bin of cell i (unknown). The model parameters are (μ_b, Σ_b) for each bin b. The problem is then inference of B from $P(X, Z|B)$. Intuitively, this is a MVN mixture, but the indicator variable B_i can be learned from Z (has a prior that depends on Z).
- Discussion: the problem is data imputation for matrix, could use some general tools such as factor analysis.
- Discussion: Use spatial smoothing for the expression distribution. Spatial resolution of Seurat is limited: relatively large bins because the method requires large number of cells for each bin.

Spatial genomics and single cell lineage dynamics by seqFISH and MEMOIR [Long Cai, HG seminar, 2018]

- Single mRNA molecule detection by FISH: RNA is fixed first, probes bind to target RNA, hybridize, washing remaining probes then imaging. Quantify expression level, preserve spatial information. Femino, Science 1998; Raj, NM, 2008. Ref; Lubeck 2012
- SeqFISH: Lubeck, NM 2014. Multiple rounds of FISH: remove the probes after each round, then do another set of probes. Each molecule is measured in each round: different barcode. Ex. gene 1: yellow, blue, yellow in three round. Gene 3: red, blue, red. Detection efficiency 80%, low FPs <0.5%.
- Note: it is important that RNA is fixed, so in multiple rounds, we can look at the same molecule.
- Analysis: each mRNA is targeted by a different probe. For simplicity, consider $8 = 2^3$ genes/probes. In each round, a probe is linked to a different color. Ex. Round 1: probes 1-4 are Y and 5-8 are G. Round 2: probes 1-2 and 5-6 are Y and 3-4 and 7-8 are G. Round 3: 1,3,5,7 are Y and 2,4,6,8 are G. In each round, the probes are degenerate, but the color information at a spot (mRNA molecule) reduces the possible gene/probe by 1/2 if using two colors. Then after three rounds, if a gene has color YYY, then it must be gene/probe 1; if GGG, then it must be gene/probe 8.
- Statistics of gene expression: TFs are present at 10 copies per cell. Most genes 0-500 copies per cell.
- Hippocampus: two regions specialized in spatial memory and emotion. Dorsal and ventral parts different. Bulk RNA-seq: highly correlated, suggesting continuous gradients of genes.
- ScRNA-seq can be mapped to seqFISH data: use ABA. Correlated expression (collaboration with GC Yuan). Cluster cells by RNA-seq. Found that 30-100 genes (random) can be highly informative. [Thomson and Patcher, 2016]
- Simpsons paradox: comparison of male vs female acceptance rate. Even in each department, similar rate; across all departments, differential rates. Reason: women apply for more difficult departments. Similar issue in single-cell analysis: bulk RNA-seq results wrong.
- Cost of SeqFISH: direct hybirdization, 3000 RNA barcodes per cell, 10K cells/week (spread over hundreds of genes). Cost per spot is limiting. FISH spots can have more information: 20 reads 1 FISH spot.

Techniques converge to map the developing human heart at single-cell level [Nature, 2020]

- Step 1 (Figure 1): do scRNA-seq, and spatial transcriptomics (about 30 cells per location). From this, determine a set of spatially informative and cell type informative genes (96).
- Step 2: In situ sequencing (ISS) of 96 probes. ISS: DNA probes that match the mRNA targets, and contain bar-codes - which can be detected by imaging. This is done at single cell level. Then combine scRNA-seq and ISS to map location of each cell.

Chapter 3

Transcriptional Regulation and Epigenomics

3.1 Transcriptional Regulation and Epigenomics: Overview

Research challenges of epigenetics: [personal notes]

- Epigenetics and gene regulation: what causes epigenetic changes and how they affect gene expression.
- Epigenetics in cell differentiation: how cellular fate is estabalished through epigenetic changes?
- Epigenetic changes in human diseases: what role does epigenetics play in human diseases? What epigenetic changes may cause diseases? How environment affects epigenetic states of cells? What explains transgenerational inheritance of some traits?
- Remark: the foundamental problem is to understand how epigenetic modifications are established and control gene expression. These will help us understand the implementation of cellular differentiation program. Meanwhile, understanding the process in the context of cellular memory and development helps us rationalize the epigenetic mechanisms.

Problems of understanding the role of epigenetics in gene regulation [personal notes]:

- Sequence-functional relationship of cis-regulatory elements: the relationship between TF binding and expression; the role of epigenetic factors (nucleosomes, chromatin motification, etc.); the regulatory grammar; dual role of TFs (sometimes activate, sometimes repress); etc.
- Encoding messages: TFs relay the message (instructions) from signaling events. The message/instructions can be complex: e.g. long-term repression, transient activation, long-term (stable) activation, turn on the promoter, turn off the enhancer, etc. How are these instructions implemented?
- Chromatin structure: organized into domains of active or inactive regions. How are these domains established? What marks the boundaries?
- Regulatory maps: some part of the genome have regulatory functions. Where are these sequences and how do they control gene expression?
- Estabilishing the epigenetic patterns (histone modification, nucleosome positioning, etc.): rules for TFs to create the histone patterns? Ex. some TFs tend to encode the instruction of long-term repression, while other TFs tend to encode the instruction of transient activation.
- Reading the histone code and DNA methylation: how is the histone pattern decoded? Common coactivators/corepressors? Involve chromatin remodeling?

Problems of understanding the role of epigenetics in development and diseases [personal notes]:

- Cellular response systems. Ex. cell proliferation in B/T cells in response to cytokines; cells switch to a different state in the presence of stress; etc.
- Cell type determination, in multicellular organisms, how many cell types are established and maintained? The key regulatory molecules, their upstream and downstream mechanisms, etc.
- Reconstruction of gene-regulatory relationship/networks: causal regulators in a biological process; regulator-target recognition; signaling pathways between signals and TFs; etc.
- The pattern of changes and differences in epigenetic states across different cells and diseases.
- Structure and function of GRNs: the architecture of GRNs: hierarchical structure (master regulators, etc.), the length of cascades; the dynamics of expression (the order of events); etc. The mechanistics is usually investigated in connection with GRN function.
- Genetic architecture of GRNs: sources of expression variations, phenotypic variations, and how they are distributed.
- How the epigenetic changes lead to change of cell types and diseases? Understand the consequence of these changes, both in cis- and in trans-.
- Rules of imprinting and transgenerational inheritance: what genes are imprinted? What traits can be passed through generations without DNA variations?

Molecular mechanism of transcriptional regulation:

- Two levels of control: transcriptional regulation controlling the level of mRNAs/proteins, and allosteric regulation controlling the activities of proteins. Different roles and interactions of the two levels of control:
  - Example: when cells start to grow, need more metabolic enzymes (transcriptional); once they are synthesized, need fine-tuning their activities s.t. different steps/pathways are coordinated and appropriate level of flux is achieved.
  - Analogy: transcriptional - whether build factories or not; allosteric - whether operate the factories (fine-control).
- Function of enhancers: how enhancers interact with promoters? Multiple enhancers? Context dependence of TF function (acting as both activators and repressors)? Functional relevance:
  - Multiple enhancers: for genes with complex expression patterns, may need mutiple separate enhancers, each giving rise to one restricted pattern.
  - Dual function of TFs: TF is a messager (e.g. the presence of a growth factor signal), but the message may be different for different genes (e.g. activate growth related genes, but suppress apoptosis or autophagy genes).
- Epigenetic code:
  - Writing the epigenetic code: What are different types of modifications? How epigenetic modifications are estabilished? What determines sequence-specificity? How TF binding interacts with chromatin modifications? What decides short-term vs. long-term epigenetic changes?
  - Reading the epigenetic code: How epigenetic states affect gene transcription? How are they interpreted by TFs, and other proteins?
  - Different types of epigenetic modifications probably serve different functional requirements: e.g. short-term change of expression or long-term change of expression.

Design goals of gene regulatory systems:

- Basic function at single cell level: control the cellular states in responses to or in anticipation of signals/genetic programs. The actions taken by the cells could be: changing regulatory program (e.g. from starvation state to growth state), changing the cell fate (proliferation, differentiation, death, etc.).
- Multicellular organism: the ultimate goal is to create and maintain a set of specialized cell types. To maintain cell types, need renewal of cells, and a control of cell differentiation in response to the needs.
- Complex expression patterns: a gene may be expressed in various patterns.
  - Single cells: e.g. in all cases, or only in response to specific signals. The difficult case is the need of expression in unrelated conditions, e.g. DNA repair genes need to be expressed in cell-cycle and in response to DNA damage agents.
  - Multicellular organisms: e.g. in specialized cells only, in all cells (housekeeping). The challenging case is the need of being expressed in a set of unrelated cells: e.g. in neurons and in blood cells (different developmental lineages).
- Sensitivity and memory of responses: cells may sometimes need to be very sensitive to subtle signals, or sometimes robust to noises. Also the response/memory may be short-term or long-term, depending on circumstance. Ex. cell fate determination is a long-term decision (irreversible), but response to starvation is short-term (change is reversible).
- Economics/evolution of gene regulation: minimization of unnecessary protein production, reuse of existing modules for new tasks (sharing modules of multiple tasks) - related to specificity issue. Also the problem is an evolutionary one, e.g. cells reuse modules not because it is economical, but because that is how it is evolved.
- Remark: these functional requirements of GRN determine/dictate various aspects of the GRN, from the arrangement of TFBSs, to the organization of GRN in development.

Design of GRN for cellular function and development:

- Organizational principle of GRN: reflect the functional requirements of the GRN. Some principles:
  - Specificity: many genes may be regulated by the same TF, but the expression pattern of these genes can be different.
  - Locality : closely related genes should be coexpressed.
  - Mutual exclusion: if some genes are needed, other ones need to be suppressed.
  - Anticipation/cross-link: expression of some process may signal the activation of others.
- Feedbacks and transcriptional circuits: how to achieve a certain quantitative behavior, e.g. fine-tuning the sensitivity of response to signals? Examples:
  - AA pathway regulation (general control): sensor is uncharged tRNA, and controlled Gcn4. As AA synthesis proceeds, the uncharged tRNAs are reduced (feedback inhibition).
  - Cell cycle - regulation of G1/S genes: sensor is provided by G1 cyclins, and the stop signal is provided by other cyclins, B cyclins.
- Cell differentiation:
  - Master regulators and transcriptional cascade controlling cell fate?
  - How complex expression patterns (e.g. genes expressed in multiple unrelated lineages) are established?

The role of gene regulation in diseases and evolution:

- Regulatory variations: what cause variations in gene expression and epigenetic chagnes?
- How incorrect regulatory (expression or epigenetic) changes lead to human diseases?
- Adaptation of GRNs to new environments: (fine-tuning the systems) and evolution of novel phenotypes (e.g. modularity may greatly facilitate evolution of new behavior of cells). Associating the changes at TFs, CREs, etc. with the conservation/changes of the function of the regulatory systems.

Common analysis in epigenomics: [personal notes]

- Analysis centered on CREs:
  - CRE annotation: promoter, enhancer (weak/inactive, active), insulator, etc. and the cellular context (cell types).
  - Associating CREs with target genes.
  - CRE function and regulation: biological processes of CREs, the regulating factors, etc. One step in this kind of analysis is clustering of CREs (e.g. by their activity profiles across multiple cells)
- Mapping transcriptional regulation and its mechanisms:
  - TFs regulating genes and CREs.
  - TF interactions/co-regultion.
  - How TFs modulate epigenetic marks; or how epigenetic marks influence TF accessibility.
- Modeling gene expression from enhancer/promoter states: TF binding, DHS, histone marks, etc.

Issues of epigenoimcs [personal notes]

- Mechanisms of CREs and chromatin state modifications: the process that turns on/off CREs, what are the respective roles of TFs histone modifications, chromatin remodeling and DNA methylation.
  - Does chromatin accessiblity always imply biological function? DHS together cover 40% of the genome: how to reconcile with evolutionary findings? Similarly, for histone marks (e.g. K27ac) and TF binding: reflect active transcription? Possible explanation [Michael Snyder]: lack of sensitivity in reporter assay (buffering, small difference, need stress to trigger the effect).
  - DNA methylation and gene regulation: DNA methylation in CpG islands near promoters silence gene expression, but it was also found that they overlap with regulatory elements? For organisms without DNA methylation, how are long-term suppression established?
  - Histone modification: might change DNA structure (charges) to affect chromatin accessibility. Or simply markers that are recognized by histone readers?
  - How are heterochromatin or gene occlusion regions established?
  - Enhancer activation: interactions between TFs, chromatin remodeling complexes and histone modifiers. Specificity of general chromatin factors? Model of pioneering factors? In general, how chromatin state switch is accomplished (inactive to poised to active)?
- Mechanisms of chromatin interactions and enhancer-promoter targeting: how do enhancers recognize promoters.
  - Can an enhancer target more than one promoter? Some estimate 50% [Thurman, Nature, 2012], and some only 10% [Sanyal, Nature, 2012]. Estimated average number is 2 [Bussemaker14, FANTOM study]. If an enhancer targets more than one promoter, are they always in different genes (some genes may have multiple promoters)?

  - Spatial organization of chromosomes: do TADs explain most of the interactions?
  - Mapping enhancer-promoter interactions: inconsistency between different methods. Predicted interactions based on DHS across cell types: very small overlap with 5C and ChIA-PET data (4% from ENCODE). Why?
  - How is specificity of E-P achieved? Possible explanation [Michael Snyder]: protein complex bring enhancer and promoter together, likely due to specificity of PPI.
  - What factors mediate the interactions between enhancers and promoters? CTCF and cohesion: insulators, but also mediating enhancer-promoter interactions. Opposite roles?

- How does a DNA mutation affect regulatory activities?
  - Affect binding to TFs and co-factors.
  - Nucleosome binding: point mutations likely have small effect on affinity to nucleosomes.

- Epigenomics in cell differentiation: how are genes expressed in the correct tissues?
  - Mode of transcription regulation: if a gene is expressed in two tissues, say brain and muscle, but not in other tissues, how is this achieved? Model 1: single enhancer with BS of brain-specific and muscle-specific TFs; Model 2: two enhancers, one for each tissue. Which model is correct? Under Model 1: the TFs in different tissues can be different, thus the enhancer can target different promoters, and have different strength.
  - Enhancer-promoter targets can change in different cell types? [IM-PET paper] Do CTCF/cohesin binding change significantly across cell types?
  - Function of ehancers: only functional in stem cells or early stage cells. Will they be needed in terminally differentiated cells?

- Transgenerational inheritance:
  - Mechanism of transgenerational inheritance in species without DNA methylation: e.g worm and fruit fly. Histone modificiation.

- Experimental techniques:
  - Fragmentation of DNA: restriction enzymes or sonication? Ex. in Hi-C, why cannot we use sonication?
  - Hi-C and ChIA-PET: quantitative relation between sequencing depth and resolution? Intuition: many interactions from random contacts (quadratic); if sequencing depth is not high enough, then at small fragments, the counts would always be really small, and it's hard to distinguish signal and noise (random contacts).

Opportunites and challenges of epigenomics [Tackling the epigenome: challenges and opportunities for collaboration, NBT, 2010]:

- Importance of epigenetics:
  - Epigenetics as a basic mechanism of cellular memory: the epigenetic state of a cell is affected by developmental as well as environmental influences, and both of these inputs may leave epigenetic traces that the cell "remembers".
  - Epigenetics as both a response and inheritance systems: some chromatin changes may be transient changes, whereas others are longer lasting. Some chromatin changes are mitotically heritable and can affect somatic tissues, whereas others may even be inherited through meiosis and affect the next generation.

- Application of epigenomics in understanding basic transcriptional process:

  - Maps of functional elements: enhancers, microRNA genes, imprinted loci, etc. Allow "upstream" investigations to identify the transcription factors, regulatory molecules and pathways that initiate, modulate or maintain epigenomic features. May also allow pursuit of 'downstream' investigations to identify genes with similar suites of epigenetic features in particular cell types.
  - Mapping of DNA methylation, histone modifications and noncoding RNAs in the same cells: the cross-talk among these epigenetic regulatory mechanisms.

- Application of epigenomics in cell differentiation and reprogramming:
  - Compared with differentiated cells, the epigenomes of human embryonic stem cells (hESCs) are unusual, especially with respect to DNA methylation. Understanding how the epigenomic state of hESCs changes during the differentiation process is crucial.

- Application of epigenomics in health and disease:
  - Cancer, a number of other diseases: involve epigenetic dysregulation. However, the extent to which epigenetic dysregulation might be a consequence of, or itself lead to, other common disease states is poorly understood.
  - Transgeneration effect: whether or not aberrant epigenetic states can affect subsequent generations is even less clear.
  - In the case of diseases that have a strong environmental component, epigenome-wide association studies that statistically correlate epigenetic variation with disease states or phenotypes could be of great value.

3.2 Epigenomics Background

Putting epigenome comparison into practice [Milosavljevic, NBT, 2010]:

- Comparing epigenomes to map cellular differentiation:
  - Epigenomes from several related cell types: infer the bifurcating branching patterns of the epigenetic landscape.
  - Comparison between two pancreatic cell types, beta cells and acinar cells: epigenomes of beta cells contain H3K27me3 marks characteristic of the endodermal lineage of the pancreatic cells, whereas the gene expression signature of beta cells largely resembles those of ectoderm-derived neural tissues [Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program, GR, 2010]
  - Reconstruction of cellular differentiation history: integrating data obtained by direct measurements on partially differentiated cell types and from reconstructions based on fully differentiated ones.

- Comparing epigenomes to understand genetic variation:
  - Rett syndrome: MeCP2 mutation acts in trans, altering genome-wide patterns of epigenome maintenance.
  - SNPs may affect DNA methylation in cis.

- Computational challenges of epigenome comparison:
  - Different resolutions of different datasets, e.g. BS-seq at single-bp resolution while MeDIP assays offer hundred-base-pair resolution
  - Searches for similarity among epigenomes, e.g. "alignment".

  - Conservation and differences in epigenomic across genomic loci.
  - Variation with biological processes, e.g. development

The NIH Roadmap Epigenomics Mapping Consortium [Bernstein & Thomson, NBT, 2010]:

- Types of epigenomic features:
  - DNA methylation: BS-seq as the primary assay
  - Histone modifications: six major histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3 and H3K36me3)
  - Chromatin accessibility: DNase I cleavage sites
  - RNA-seq: mRNA and small RNA
- Cell types:
  - Stem cells: hESCs, additional stem cell models, including mesenchymal and neural stem cells, and reprogrammed cells
  - Primary cells and tissues relevant to metabolic and cardiovascular disease, cancer, neuropsychiatric disease, aging and other leading health issues. Prioritized cell types include sorted hematopoietic lineages, liver, muscle and adipose, as well as selected cell types from breast and neural tissues.
  - Fetal tissues.

Epigenetic modifications [Epigenetic modifications and diseases, NBT, 2010; Epigenomics reveals a functional genome anatomy and a new approach to common disease NBT, 2010]:

- Major epigenetic mechanisms: DNA methylation, histone modifications and nucleosome positioning. It is important to keep in mind the interplay between epigenetic factors and the many positive and negative feedback mechanisms.
- DNA methylation: almost exclusively in the context of CpG dinucleotides.
  - DNA methylation plays a key role in genomic imprinting, where hypermethylation at one of the two parental alleles leads to monoallelic expression, also in X-chr inactivation in females.
  - CpG islands: The CpG dinucleotides tend to cluster in regions called CpG islands. CpG dinucleotides are usually quite rare in mammalian genomes (about 1%). CpG islands at promoters of genes are normally unmethylated, allowing transcription. About 60% of human gene promoters are associated with CpG islands and are usually unmethylated in normal cells.
  - CpG island shores: located up to 2 kb upstream of the CpG island. Most of the tissue-specific DNA methylation seems to occur not at CpG islands but at CpG island shores.
  - Gene bodies: methylation at the gene body facilitates transcription, preventing spurious transcription initiations. Gene body methylation is common in ubiquitously expressed genes and is positively correlated with gene expression.
  - Repetitive sequences: hypermethylated, preventing chromosomal instability, translocations and gene disruption.
  - DNA methylation can inhibit gene expression by various mechanisms: (1) promote the recruitment of methyl-CpG-binding domain (MBD) proteins. MBD family members in turn recruit histone-modifying and chromatin-remodeling complexes to methylated sites. (2) directly inhibit transcription by precluding the recruitment of DNA binding proteins from their target sites. (3) In contrast, unmethylated CpG islands generate a chromatin structure favorable for gene expression by recruiting Cfp1, which associates with histone methyltransferase Setd1, creating domains rich in the histone methylation mark H3K4 trimethylation (H3K4me3).

  - DNA methylation and DNA methylation-associated proteins also involved in nuclear organization and in the establishment of specific chromosomal territories.
  - Over evolutionary time methylated cytosines tend to turn into thymines because of spontaneous deamination. The result is that CpGs are relatively rare unless there is selective pressure to keep them or a region is not methylated for some reason.
  - Sequence-specific methylation: several mechanisms have been proposed, mainly suggesting interaction of DNMTs with other epigenetic factors. In plants, RNA-directed DNA methylation is a stepwise process initiated by double-stranded RNAs that recruit DNMTs to catalyze de novo DNA methylation of specific regions.

- Histone modification: have important roles in transcriptional regulation, DNA repair, DNA replication, alternative splicing and chromosome condensation.
  - Euchromatin and heterochromatin: euchromatin characterized by high levels of acetylation and trimethylated H3K4, H3K36 and H3K79. Heterochromatin is characterized by low levels of acetylation and high levels of H3K9, H3K27 and H4K20 methylation.
  - Actively transcribed genes are characterized by high levels of H3K4me3, H3K27ac, H2BK5ac and H4K20me1 in the promoter and H3K79me1 and H4K20me1 along the gene body.
  - Many transcriptional co-activators (e.g., GCN5, PCAF, CBP, p300, Tip60 and MOF) possess intrinsic HAT activity, whereas many transcriptional co-repressor complexes (e.g., mSin3a, NCoR/SMRT and Mi-2/NuRD) contain subunits with HDAC activity.

- Interplay between histone modifications and DNA methylation:
  - Example: DNMT3L specifically interacts with histone H3 tails, inducing de novo DNA methylation; however, this interaction is strongly inhibited by H3K4me.
  - Several histone methyltransferases have been reported to direct DNA methylation to specific genomic targets by recruiting DNMTs, helping in this way to set the silenced state established by the repressive histone marks.
  - Histone methyltransferases and demethylases can also modulate the stability of DNMT proteins, thereby regulating DNA methylation levels.
  - DNA methylation can also direct histone modifications.

- Nucleosome positioning: Nucleosomes are a barrier to transcription that blocks access of proteins to their sites on DNA, and at the same time they inhibit the elongation of the transcripts by engaged polymerases.
  - Nucleosomes near genes: the precise position of nucleosomes around the transcription start sites (TSSs) has an important influence on the initiation of transcription. Moreover, the 5' and 3' ends of genes possess nucleosome-free regions needed to provide space for the assembly and disassembly of the transcription machinery.
  - Nucleosome positioning also plays an important role in shaping methylation: [Relationship between nucleosome positioning and DNA methylation. Nature, 2010] nucleosomal DNA was more highly methylated than flanking DNA. These results indicate that nucleosome positioning influences DNA methylation patterning throughout the genome and that DNA methyltransferases preferentially target nucleosome-bound DNA.
  - Nucleosomes were highly enriched on exons, and preferentially positioned at intron-exon and exon-intron boundaries.
  - Histone variants: regulate nucleosome positioning and gene expression. For example, the incorporation of the histone variant H2A.Z protects genes against DNA methylation.

  - Interaction with other mechanisms: the nucleosome remodeling machinery is influenced by DNA methylation and specific histone modifications. MicroRNAs (miRNAs) can also regulate histone variant replacement or interact with chromatin remodeling complexes mediating the exchange of specific subunits.
  - Chromatin remodeling complex: Several groups of large macromolecular complexes are known to move, destabilize, eject or restructure nucleosomes in an ATP hydrolysis-dependent manner. Four families (SWI/SNF, ISWI, CHD and INO80) that share similar ATPase domains.

- Chromatin/genome structure:
  - Lage genomic regions and gene clusters: e.g. globin cluster. Large (tens to thousands of kilobases) genomic regions regulating gene expression are common. In particular, imprinted genes were often organized in gene clusters, often with common regulatory elements, such as CCCTC binding factor (CTCF) binding sites.
  - Heterochromatin: many large regions of heterochromatin modifications have been found, e.g. in inactive X chromosome. Large autosomal regions of heterochromatin modification across Hox gene clusters are highly conserved across species.
  - Frequent intra- and interchromosomal interactions: for example, SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes, across 200kb region.
  - RNAs in chromatin structure: e.g. antisense RNAs appear to establish heterochromatin (spanning $>$ 100 kb) in mammalian genes, independently of Dicer and the post-translational microRNA machinery. A recent discovery is the role of long intergenic noncoding RNAs (lincRNAs) in establishing heterochromatin.
  - Large organized chromatin lysine (K) modifications (or LOCKs): organize the genome into very large blocks (hundreds to thousands of kilobases), some of which are differentiation-specific in their location and extent and correspond to lamin-associated domains (LADs).

- Influence of epigentic networks/chromatin structure on cellular development and genome function: examples:
  - CTCF: (mediates H19 imprinting) plays a general role in defining the boundaries of functional gene regions. Binding by CTCF can block the interaction between enhancers and promoters, therefore limiting the activity of enhancers to certain functional domains. CTCF can also prevent the spread of heterochromatin structures.
  - Polycomb: Polycomb-group proteins are a family of proteins that can remodel chromatin such that epigenetic silencing of genes takes place. Polycomb-group proteins are best known for silencing Hox genes through modulation of chromatin structure.

Epigenetic changes and human diseases [ibid]:

- Epigenetic dysregulation in cancer: [The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores, NG, 2009]
  - Background: alterations in DNA methylation (DNAm) occur in cancer, including hypomethylation of oncogenes and hypermethylation of tumor suppressor genes. However, most studies of cancer methylation assumed that functionally important DNAm occur in promoters, and in cancer, the important DNAm change in CpG islands.
  - Results: most methylation alterations in colon cancer occur not in promoters, and also not in CpG islands, but in sequences up to 2 kb distant, which we term 'CpG island shores'. CpG island shore methylation was strongly related to gene expression.

  - Results: methylation changes in cancer are at sites that vary normally in tissue differentiation. Epigenetic progenitor model of cancer: epigenetic alterations affecting tissue-specific differentiation are the predominant mechanism by which epigenetic changes cause cancer.

- Epigenetic regulation in cell differentiation: promiscuous methylation differences from somatic cells on a genome-wide scale, notably including differences at non-CpG sites. The sites of differential methylation largely overlap: the same sites appear, in normal cells compared with cancer cells, in stem cells compared with differentiated cells and in comparisons of tissues derived from different germ layers.

- Epigenetic modification and autoimmune diseases:

  - ICF (immunodeficiency, centromeric instability and facial anomalies) syndrome: caused by heterozygous mutations in DNMT3B. Several genes regulating development, neurogenesis and immune function have aberrant expression.
  - Systemic lupus erythematosus (SLE): SLE patients have DNA hypomethylation in PRF1, CD70, CD154, IFGNR2, MMP14, LCN2, CSF3R and AIM2 among other genes. It has been recently reported that hypomethylation in SLE is partially mediated by miR-21 and miR-148a that directly and indirectly target DNMT1.
  - Rheumatoid arthritis: Because the transcription factor NF-kappaB binds very poorly to nucleosomal DNA, histone modifications are needed to allow efficient NF-kappaB binding to its targets. Thus, in rheumatoid arthritis, the reduced activity of HDACs plays a key role in regulating NF-kappaB mediated gene expression.
  - SNP in the 17q12-q21 region (associated with a higher risk of asthma, type 1 diabetes, primary biliary cirrhosis and Crohn's disease): lead to allele-specific differences in nucleosome distribution

- Epigenetic variation across individuals: mice from the same litter and living in the same cage, hundreds of variably methylated regions (VMRs) that are highly enriched by functional annotation for key genes in development and embryonic pattern formation. Several VMRs have recently been linked to body mass index.

Mapping Human Epigenomes [Rivera & Ren, Cell, 2013]

- DNA methylation:

  - In the human genome 60-80% of 28 million CpG dinucleotides are methylated. Earlier studies have established important repressive roles of DNA methylation in imprinting, retrotransposon silencing, and X chromosome inactivation. 5mC methylation can exist in non-CpG sequence contexts, and is enriched at the bodies of actively transcribed genes.
  - Methods for profiling DNA methylation: digestion of genomic DNA with methyl-sensitive restriction enzymes (RE), affinity-based enrichment of methylated DNA fragments, and chemical conversion methods. RE-based methods have low resolution, limited by the number of cut sites. Affinity-based method (MeDIP-seq): the resolutions are highly dependent on the DNA fragment size, CpG density, and immunoprecipitation quality of the reagent.
  - Bisulfate sequencing: unmethylated C will be converted to U and then T upon chemical treatment, while 5mC remains as C. Referred to as BS-seq, WGBS or MethylC-seq, the gold standard of DNA methylation. Cost issue: to obtain single bp resolution, need 30x coverage. It was found that only 20% of CpGs are differentially methylated between 30 diverse human cell types, so the idea to reduce cost is to combine capture-based methods with base-resolution DNA methylation assays for targeted mapping.
  - There are multiple methylation states of C. In partciular, 5mC and 5hmC, but not 5fC and 5caC, are both resistant to bisulfite conversion and therefore cannot be distinguished from each other in MethylC-seq data.

- Histone modification:
  - Distinct signatures of histone modification (Table 1). Promoters: H3K4me3. Enhancers: H3K4me1. Active enhancers, H3K27ac. Poised developmental enhancers: H3K4me1/H3K27me3. Polycomb-repressed regions: H3K27me3.
  - Quality of ChIP-seq: specfic antibody binding and minimum cross-reactivity. Some gold standard: e.g. ChIP-string, profile using 500 representative loci.
  - ChIP-seq technologies are mainly limited by: the need for large amounts of starting material, limited resolution, and the dependence on antibodies. Typically 1 millions cells (for histone) or 5 million cells (for TFs) are required, and this poses challenges when studying primary cells and rare populations such as cancer stem cells.
  - Variations of ChIP-seq: (1) ChIP-exo, treat with exonuclease to digest DNA to the footprint of the crosslinked protein. Good to obtain nucleotide resolution and DNA binding motifs. (2) Two ChIP steps in a row, or Sequential-ChIP-seq, uncover histone PTMs on the same molecule or chromatin associated proteins in the same complex. (3) ChIA-PET: long distance DNA interactions mediated by a specific protein.
- Chromatin structure: nucleosome positioning
  - Determinants of nucleosome positioning: favorable DNA sequence composition, ATP-dependent nucleosome remodelers, strongly positioned nucleosomes and DNA bound proteins such as TFs and RNA Pol II (barriers to nucleosome position shifting).
  - Micrococcal nuclease digestion of chromatin followed by high-throughput sequencing (MNase-seq): DNA wrapped around histone octamers or bound by TFs is protected.
  - Another method: use DNA methyltransferase accessibility to footprint nucleosome positions, called nucleosome occupancy and methylome sequencing (NOMe-seq).
- Chromatin structure: open chromatins
  - DNase-seq: take advantage of the protection conferred by tightly wound nucleosomes from DNaseI endonuclease digestion.
  - Formaldehyde-assisted identification of regulatory elements followed by sequencing (FAIRE-seq): open chromatin regions are also sensitive to shearing by sonication.
  - DHSs span 2.1% of the genome per cell type on average and, impressively, all 4,000,000 sites collectively cover about 40% of the genome.
- Chromatin interactions and higher order architecture:
  - Why higher order structure is important? Ex. chromatin in close proximity to the nuclear lamina (intermediate filaments and membrane associated proteins in inner nuclear membrane) tends to be heterochromatic and transcriptionally repressed.
  - 3C, 4C, 5C and Hi-C: the key step is DNA ligation (under extremely dilute conditions) to favor joining of ends. ChIA-PET is a variation of Hi-C which features an IP step to map DNA interactions involving a protein of interest (e.g. RNA Pol II or CTCF). Resolutions: 40 kb in a recent study [Dixon, Nature, 2012], and may be possible to achieve fragment size (4kb).
  - Hierarchy of nuclear organizations: chromosome territories, TADs, sub-TADs and cis-regulatory interactions.
- Annotation of cis-regulatory elements in the genome:
  - Transcriptional enhancers are characterized by the presence of H3K4me1 but not H3K4me3 [Heintzman, NG, 2007].

  - Active vs. poised enhancers [Rada-Iglesias, Nature, 2011]: 7,000 enhancers in hESC, featuring H3K4me1, TF binding and nucleosome depletion. Active enhancers are near genes expressed in hESCs, while poised enhancers are next to genes inactive in hESC but turned on during differentiation. The two differ mainly by H3K27ac.
  - Super-enhancers [Whyte, Cell, 2013]: a small subset of enhancers ($< 1\%$) form large domains up to 50kb, bound by master TFs. These super-enhancers are posited to regulate key genes important for cell identity.
  - Enhancer decommissioning: H3K4me1 and H3K27ac marks for some key enhancers (e.g. pluropotency) must be removed later in differentiation.
  - Algorithms for enhancer prediction [Rajagopal & Ren, PLCB, 2014]: a random forest classifer using features including histone marks (H3K4me1, H3K27ac), DNaseI hypersensitivity, combinatorial TF binding, H3.3 and H2A.Z histone variant enrichment, bound RNA Pol II, and RNA production (eRNAs).

- Annotation of regulatory interactions:
  - Experimental approach: 5C and ChIA-PET currently provide the best balance of resolution and reasonable coverage. Computational approach: correlates regulatory elements and target promoters across many cells types.
  - Cardinality of enhancer-promoter interactions: one to one or many-to-many? By correlative approach, half of putative enhancers regulate more than one TSS [Thurman, Nature, 2012]. By 5C, only 10% of enhancers interacting with more than one promoter [Sanyal, Nature, 2012].
  - The role of CTCF and cohesin: enhancer-promoter interactions often span hundreds of kilobases surpassing one or more CTCF sites.

- Chromatin dynamics during development:
  - Differentially methylated region (DMR) (among different cell types): are enriched at regulatory elements as evidenced by their overlap with DNaseI sites, TF binding sites, and enhancer chromatin marks.
  - Vestigal enhancer: regions depleted of DNA methylation and enhancer chromatin marks in adult tissues but which exhibit enhancer activity earlier in development. This suggests that methylome retains a memory of the past.
  - Gradual expansion of repressed domains during cellular differentiation (H3K27me3 mark).
  - Identify cell state or stage specific master regulators and construct transcriptional networks: motif analysis in cell-type specific enhancers or promoters. Furthermore, enhancer-driven regulatory networks can be constructed by correlating TF expression data with motif analysis.

- Future directions:
  - Technology challenges: nano-sale experiments that require a small amount of cells. Obtain epigenome of single cell types from heterogeneous cell populations; single-cell methods.
  - A better understanding of the functional relationships between DNA methylation, chromatin modification state, or higher order chromatin structure and gene regulation. Ex. to determine whether a epigenetic state is necessary for transcription or merely coincidental.
  - Better define the target genes of the distal regulatory elements such as enhancers: spatial proximity from chromatin interaction data, however, this may not be sufficient for functional regulation. Another strategy: correlating chromatin state or accessibility with gene expression, however, it does not work particularly well when a gene is regulated by multiple tissue-specific enhancers in different tissues.

3.3 Epigenomics Technologies

Validation of epigenomic experimentals: following an experimental study, we often need to do QC, to verify the enhaners. We use enhancer mapping (e.g. DHS or H3K27ac) as an example.

- Biological replicates: ratio of activities of enhancers (e.g. measured by RPKM) across multiple biological samples. The average ratio should be close to 1.
- Sample clustering: suppose we have samples from multiple tissues/collections, then we use enhancer activities as representation of samples, these samples should form clusters.
- Expected properties of enhancers: e.g. genomic locations relative to TSS; conservation within and between species.
- Enhancer activity patterns: generally, many enhancers are active in multiple samples. If enhancers are random, we expect most of them are specific to one sample.
- Correlation with other datasets: the enhancer set should overlap with results from other studies in related cell types.
- Remark: some of the ideas are similar to those for validating gene expression studies.

3.3.1 Chemical Modifications

Bi-S sequencing:

- Bi-S sequencing: treatment of Bi-S will make C become U (read as T), but methylated C (mC) will be protected.
- Post-ate adaptor tagging (PBAT) [Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging, NAR, 2012]. In conventional Bi-S sequencing, adaptor ligation followed by Bi-S treatment; however, Bi-S leads to fragmentation, which damages a fraction of ligated fragments. In PBAT, Bi-S treatment first, then adaptor ligation on fragments of Bi-S.
- Reduced represtation Bi-S sequencing (RRBS): use enzyme to cut CpG sites, thus enriching the CpG sites. This reduces the sequencing cost.

Analysing and interpreting DNA methylation data [Bock, NRG, 2012]

- Relation between DNA methylation and transcription: (1) CpG rich region: correlation; (2) CpG poor: more complex and context-dependent relationship.
- International Human Epigenome Consortium: DNA methylation in 1000 human cell populations
- EWAS: requires methods for detecting differentially methylated regions (DMRs).
- Alignment: (1) 3-letter alignment. (2) Wild-card alignment: genomic C becomes Y, which matches both C and T. Or equivalently, allow T in read to match C in the genome.
- Bis-SNP: GATK model for Bi-S sequencing data.
- Possible experimental errors and QC:
  - Sample material: DNA quality and quantity. Duplicated reads: a sign of low DNA sample.
  - Sample mix-up/contamination. Reconstruct individuals from genotype.
  - Batch effect: hard to avoid. Correction by statistics.
  - Sequencing into adaptor: some DNA fragments are shorter than reads. Adapter trimming.

  - End repair: when restriction enzymes are used (such as in RRBS), introducing unmethylated Cs.
  - Bi-S conversion: incomplete conversion and conversion of methylated Cs. Use spike-in control; CpC can estimate incomplete conversion, which is rarely methylated.

- Analysis is based on CpG methylation table after basic data processing (alignment, QC, methylation at each CpG).
- Visualization: of regions, candidate or random. Color coding or vertical bars for methylation level. Data in BED or WIG format, convert o binary format, then visualize in genome browser, e.g. UCSC (web) or IGV, IGB (local).
- Plotting and summarizing the results: global summary of data, both for QC and for understanding the results.
  - Box plot and violin plot for distribution of DNA methylation level.
  - Scatter plot across sample pairs: replication; similarity and difference for related samples.
  - Hilbert curve: for spatial distribution (how big are the clusters).
  - Hierarchical clustering: show relationship of samples. Note: specialized methods for bimodal distribution were developed.
- Identifying DMRs: typically regions of a few hundred to thousand bps, believed to control transcription repression.
  - Single CpG comparison: t-test, Wilcoxon, linear model. Methods: 92, 96.
  - Regional comparison: genome-wide tiling analysis (non-overlapping) or in candidate regions (e.g. promoters, enhancers).
  - Hierarchical model for single CpG or regional comparison: better estimation of standard deviation of methylation.
- Validating and interpreting results:
  - Visual inspection of top regions: for signs of artifacts (repetitive regions, etc.). Also global pattern: e.g. Q-Q plot of p-values.
  - Experimental verification of selected regions.
  - Replication in a second set of samples.
- Interpretation of results:
  - Overlap of DMRs with genomic annotations: using Genome Browsers such as Galaxy, EpiExplorer.
  - Pathway analysis: GREAT.
  - Enrichment of cell-type specific functional elements: e.g. TFBSs in one cell type.
- Caveats for interpreting results:
  - For testing enrichment of a genome feature in DMRs: CpG density is a confounding factor.
  - For case-control comparison, hidden confounders: cell composition.
  - Genotype could affect DMR: difficulty of interpreting EWAS, could be DMR reflects genetic difference, instead of causal factor of phenotype.

DNA methylome analysis using short bisulfite sequencing data [Krueger & Andrews, NM, 2012]

- Background: CpG often highly methylated, about 60-80% in mammals. Non-CpG generally unmethylated in differentiated tissues (0.3-3%).

- Important QC steps and considerations:
  - Read trimming before alignment: reads with low call qualities, or adaptor contamination.
  - Read alignment: even with 3-letter alignment, the mapping efficiency converges to standard reads for reads longer than 40 bp. Recommendation: 50-75 bp, compromise of mapping and problems with longer reads.
  - Reading trimming after alignment: duplicate reads or reads with very high coverage.
  - Estimating incomplete conversion: use non-CpG cytosines, or Spike-in's. Caution: methylation level of non-CpG may not be low; spike-in's may not be the same as biological samples.

Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome (TAB-seq) [Yu & He, Cell, 2012]

- Background: TET enzyme converts:
  $5\,\mathrm{mC}(-\mathrm{CH_3}) \longrightarrow 5\,\mathrm{hmC}(-\mathrm{CH_2}-\mathrm{OH}) \longrightarrow 5\,\mathrm{fC}(-\mathrm{CH}{=}\mathrm{O}) \longrightarrow 5\,\mathrm{caC}(-\mathrm{COOH})$
  For bi-S, 5caC can be converted to U and read as T.
- TAB-seq: in normal bi-S sequencing, both 5mC and 5hmC will be protected, and C becomes U (T). In TAB-seq, both C and 5mC will be converted to U (T), but 5hmC will be protected. The key idea is to use TET to convert 5mC, but protect 5hmC before using TET.
- Procedure: (1) use βGT to add glucose to 5hmC; (2) TET that oxidizes 5mC to 5caC (but not protected 5hmC); (3) bi-S conversion of C and 5caC to U (T).
- Proof-of-concept: a single DNA with known modification, show expected conversion and nonconversion using sequencing and mass spec.
- Use spike-in to estimate conversion and non-conversion rates: see Figure S2B. Non-conversion rate of 5hmC is about 92%, and for 5mC is 2% and unmodified C is 0.38%.
- Comparison with previous, affinity-based methods: visualization (Figure 2A) in Oct4 locus. Enrichment of 5hmC in affinity-based peaks: 6-fold.
- Estimation of specificity and sensitivity: (1) Experimental validation of 5hmC missed by affinity method. (2) Among all peaks from affinity methods, 81% have 5hmC in TAB-seq.
- Enrichment of regulatory elements in 5hmC: 3-7 fold enrichment of 5hmC in CTCF, DHS and enhancers in mESC data.

BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions [Hansen, GB, 2012]

- QC idea: methylation state (fraction of M vs. M and U) should not depend on read position. The dependency in the plot would suggest problem - filtering the end positions.
- Smoothing: the idea is that methylation level of adjacent sites should be similar/smooth. Let π_j be the level of site j, and $f(l_j) = \pi_j$. Estimate f with a smooth function: $\log f(l_j)/(1-f(l_j))$ be a second degree polynomial. Also when estimating π_j, use local likelihood centered on j. The data (read depth) follow binomial distribution.
- Results: compare the results under 5x using BSmooth vs. the data at 30x, the DNA methylation levels are highly correlated (0.90).
- Remark: two ideas in smooth estimation, first, estimate a smooth function (using polynomial). Next, use local likelihood for each point.

Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data: BEAT [Akman, Bioinformatics, 2014]

- Region-based analysis: aggregate data from consecutive cytosines into regions, and estimate methylation level in each region.
- Model: a simplified version here. Let k be the number of methylation counts (C's) in a region, and r the true methylation level. Define p_+ as FP rate, obtained from methylation in non-CpG; and p_- as FN rate (methylated sites not protected from BiS conversion). Set $p_- = 0.2$ to account for read mapping error, etc. Then:

$$k \sim \text{Binom}(n, r(1-p_-) + (1-r)p_+) \tag{3.1}$$

  To make inference, the prior of r is a mixture of Beta (one for methylated region, the other for unmethylated), and infer the posterior.

A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data: MeTPeak [Cui & Huang, Bioinfo, 2016]

- Motivation: calling M6A peaks is different from ChIP-seq in that we need to normalize by individual transcripts. Ex. if mRNA levels different between IP and input, and we normalize using the entire library, we may see all bins in the IP are enriched with reads. However, with a single transcript, possible that most of reads come from a single M6A site, thus the reads are not the "background" we should use as control. Solve this problem by explicitly distinguish M6A and non-M6A sites, and (implicitly) use non-M6A as control.
- Model: Figure 1. Analyze each transcript separately. Divide it into n connected bins. Let Z_n be the methylation status of bin n (binary). Z_n is modeled by a HMM with transition probability A. X_{mn}, Y_{mn} are read counts in bin n of sample m of IP and input respectively. Let $T_{mn} = X_{mn} + Y_{mn}$. Then $X_{mn}|T_{mn}$ should follow binomial distribution with probability p_n. We assume that p_n follows different Beta-distributions, depending on Z_n:

$$p_n|Z_n = k \sim \text{Beta}(\alpha_k, \beta_k) \qquad k = 1, 0 \tag{3.2}$$

  The parameters are $\Theta = (\alpha, \beta, A, \pi)$, where π is the stationary distribution of HMM. To infer Θ, we maximize $P(X|\Theta) = \sum_Z P(X, Z|\Theta)$.
- Analysis of the model: why it works? The model distinguishes M6A and non-M6A sites. In the non-M6A sites, we estimate α_0, β_0, which captures the read difference between IP and input. This summarizes both the diff. of library sizes, but also the diff. in [mRNA].
- Application and validation: (1) FDR by using permutations (of IP and input). The number of peaks above a given FDR can be used to compare methods. (2) Use the distance of motif matches to the peak center to compare methods.
- Remark: the method does not address the issue of testing differential methylation (DM). It may be possible to test the difference by comparing the posterior of Z_n and p_n between two conditions. Intuitively, if Z_n's are different, then DM. If both are 1, then we compare the posterior intervals of p_n's.
- Remark: another possible improvement, introduce prior of Z_n to depend on the locations, e.g. more M6A near 3' UTR.

A Highly Sensitive and Robust Method for Genome- wide 5hmC Profiling of Rare Cell Populations [Han and He, Mol Cell, 2016]

- Concept (Figure 1A): use the enzyme βGT to add glucose to 5hmC, and link biotin with glucose. Then pull down biotin and do sequencing.

- Correlation of results across replicates and across different DNA amounts: highly significant between replicates, about 0.8 - 0.9; and also high at different DNA levels.
- Comparison of results with previous (gold-standard) results: (1) high Pearson correlation. (2) Similar profiles: distribution of 5hmC near TSS. (3) Centered on known 5hmC sites: enrichment of 5hmC in Nano-Seal reads as a function of distance.
- Library complexity analysis: how number of unique reads change with library size.
- Sensitivity analysis: the fraction of peaks (from Tab-seq) are recovered with Nano-seal at different DNA levels.

3.3.2 Histone Modification

Chromatin signuatures of promoters and enhancers in human [Heintzman & Ren, NG, 2007]:

- Problem: do promoters and enhancers have unique histone modification code (or how are the histones of regulatory sequences modified)?
- Methods:
  - Experiment: human HeLa cells, 30M ENCODE region, use ChIP-chip to detect binding of RNAP and TAF1 (promoter marker), p300 (enhancer marker), H3 (nucleosome density), and five histone modifications: H4ac, H3ac, H3K4me1, H3K4me2, H3K4me3.
  - Clustering of spatial profiles (of histone or protein binding patterns):
  - Prediction of active promoters and enhancers by histone signatures: use known promoters and p300 binding sites as training data.
- Results:
  - Chromatin signatures of promoters: comparing active and inactive promoters (promoters: from known annotation; activity: from mRNA expression). (1) All five histone modifications are increased in active promoters. (2) Nucleosome free regions (NFR) near TSS, as suggested by depletion of H3; and histone modifications in both sides of NFR. (3) RNAP and TAF1 binding in active promoters, and moderate p300 binding.
  - p300 as a marker of enhancers: over 75% of p300 binding sites are more than 2.5kb away from 5' end of the genes. Evidence for enhancers: (1) About 70% p300 binding sites overlap with DNaseI hpersensitive sites (DHSs). (2) Over 60% p300 binding sites contain highly conserved sequences. (3) More than 40% p300 binding sites contain predicted modules based on putative TFBSs (PReMods).
  - Chromatin signatures of enhancers: the features include: (1) H3K4me1: strongly enriched in enhancers (but not in promoters); H3K4me3: lacked in enhancers (but enriched in promoters). (2) Nucleosome depletion also in enhancers. (3) TAF1 and RNAPII were also present at some enhancers (probably from assembly of transcriptional machinery from both promoters and enhancers).
  - Validation of predicted promoters and enhancers: (1) Predicted promoters: existing annotations, CAGE, DHSs. (2) Predicted enhancers: evolutionary conservation, predicted modules, DHSs.

  Remark: the (experimental) evidence of enhancers: reporter assay (Luciferase activity in this paper); evolutionary conservation; DHSs; p300 binding; and histone signatures (from this paper).

Discovery and characterization of chromatin states for systematic annotation of the human genome (ChromHMM) [Ernst & Kellis, NBT, 2010]:

- Motivation: many types of epigenetic data, from histone modifications to DNA methylations, etc. The combination of modification patterns may form a signature of a certain set of sequences, e.g. promoters and enhancers.
- Model:
  - Data discretization: apply a threshold to each type of data s.t. at any sequence window, only two values per data type.
  - Each sequence window (200 bp) belongs to a hidden state, which emits many types of epigenetic data (independent conditioned on the state), with simple Bernoulli distribution. HMM is used to impose spatial clustering of the same type of states, and capture the dependence between these states, e.g. promoter always follows a coding sequence.
  - HMM model selection: use BIC (and backward elimination) to choose the number of HMM states.
- Results: 51 states (Figure 1), corresponding to promoters, active transcription, repression, repeat sequences, coding sequences, splicing, etc. A number of states are very similar except at very few histone marks.
- Remark: main problems of the model:
  - Discretization: using a uniform threshold. Lose information, and does not account for local variation of mappability, GC, etc.
  - States: completely unsupervised, thus slight variation of histone signatures can lead to two different states. Hierarchical clustering with HMM?

Mapping and analysis of chromatin state dynamics in nine human cell types [Ernst, Nature, 2011]:

- Data:
  - Nine cell types: embryonic stem cells (H1 ES), erythrocytic leukaemia cells (K562), B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK) and mammary epithelial cells (HMEC).
  - Chromatin markers: histone H3 lysine 4 trimethylation (H3K4me3), a modification associated with promoters; H3K4me2 (dimethylation), associated with promoters and enhancers; H3K4me1 (methylation), preferentially associated with enhancers; lysine 9 acetylation (H3K9ac) and H3K27ac, associated with active regulatory regions; H3K36me3 and H4K20me1, associated with transcribed regions; H3K27me3, associated with Polycomb-repressed regions; and CTCF, a sequence-specific insulator protein with diverse functions. Also data for H3K9me3, RNA polymerase II (RNAPII) and H2A.Z (also known as H2AFZ) in a subset of cells.
- Six broad classes of chromatin states from chromatin markers: promoter, enhancer, insulator, transcribed, repressed and inactive states. Within them, active, weak and poised promoters differ in expression level, strong and weak candidate enhancers differ in expression of proximal genes, and strongly and weakly transcribed regions also differ in their positional enrichments along transcripts.
- Variation across cell types: regulatory regions vary drastically in activity level across cell types. Enhancer states show frequent interchange between strong and weak, and promoter states vary between active, weak and poised. Promoter states seem more stable than enhancers.
- Regulation of different classes of genes:
  - Clustered active promoter and strong enhancer regions across all cell types: clusters showing common activity and associated with highly coherent functions.

  - Developmental genes seem to be strongly regulated by both enhancers and promoters, showing the highest number of proximal enhancers and diverse promoter states, including poised and Polycomb repressed. Tissue-specific genes (for example immune genes and steroid metabolism genes) seem to be more dependent on enhancer regulation, showing multiple tissue-specific enhancers but less diverse promoter states. Lastly, housekeeping genes are primarily promoter regulated, with few enhancers in their vicinities.

- Enhancer-gene connection: define activity profile of enhancers and expression profiles for genes, and use correlations between these profiles to link enhancers and genes. Use cis-eQTL data to validate some of these predictions: predicted enhancers are significantly enriched with cis-eSNPs.

- Regulators of enhancers:

  - Predicting regulators: cluster enhancers (activity profiles), and then predicted, on the basis of regulatory motif enrichments, sequence-specific transcription factors likely to target enhancers.
  - Acitivator or repressor: correlate a motif score based on motif enrichment in a given cluster, and a transcription factor expression score based on the agreement between the transcription factor expression pattern, and the cluster activity profile to determine activator or repressor.
  - Predicting TFBS: motif instances within enhancer regions in specific cellular contexts. Validate these inferences using a general molecular signature: local depletions in the chromatin intensity profiles.

Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development [Bonn & Furlong, NG, 2012]:

- Motivation: the central problem is how histone patterns are related to the transcriptional states/changes of genes. In particular, interested in what differentiate active transcription vs. dormant enhancers. Some background:

  - H3K4me1, H3K27ac are markers of CRM locations, as defined by DNase HS or p300 binding.
  - However, the relation between histone modification and expression is not clear. Ex. in ESC, H3K27ac indicates transcription while H3K27me3 indicates repression. But in CD4+ T cells, no correlation between these markers and expression.

- BiTS-ChIP technology: to study the problem, we need to look at dynamic data of how transcription changes over time. However, this is difficult with cell-culture system and better done at embryonic development. However, whole-embryo approach is not ideal and we need data from specific type of cells.

  - Technique: the idea is that we create transgenic animals expressing tagged proteins in specific cells of interest (by using tissue-specific promoter or enhancer). Then we dissect the embryo and use FACS to purify the target cell types. ChIP-seq experiments can be then performed.
  - Data: 6-8h of fly embryo, take the mesoderm cells, and performs ChIP-Seq on 5 histone markers and Pol II.
  - Power of technology: if the technology achieves its purpose, then the mesoderm-specific enhancer should show differnt histone pattern than the other enhancers. This is indeed the case.

- Patterns of all enhancers: from CAD database, take 144 enhancers (more than 1kb away from the genes). 111 (77%) were enriched for H3K4me1, and 23 (16%) were enriched for H3K27ac. Pol II in 11 (8%) enhancers, H3K79me3 on 21 (15%) enhancers, H3K27me3 in 95 (66%) enhancers. Overall, most histone markers and Pol II are enriched than background sequences.

- Patterns of active vs. inactive enhancers: compare 22 enhancers exclusively expressed in mesoderm and 31 enhancers expressed outside mesoderm.

  - H3K4me1: no difference beteween the two types of enhancers.
  - H3K27me3: depleted on active mesodermal enhancers. This suggests that it is a repression marker instead of a marker of poised enhancers, as many enhancers marked by H3K27me3 in the mesoderm were active in other cell types at this stage of development but did not become active in mesodermal cells.
  - H3K27ac, H3K79me3, Pol II: significantly enriched in mesodermal enhancers, but not in non-mesoderm enhancers. However, very heterogeneous patterns of these three markers in the enhancers as each individual marker explains only a small fraction of enhancers.

- Spatial-temporal patterns of epigenetic changes:
  - Rationale for studying temporal patterns: since we only have epigenetic data for one stage, to probe the temporal patterns, we compare the epigenetic states of temporally different ehancers: early (before 6-8hr), active and late (after 6-8hr). If an epigenetic change is transient only at 6-8 hr, then it would not be present in E and L enhancers.
  - H3K27ac, H3K79me3, Pol II: correspond to precise timing, they are mostly in active mesoderm enhancers, but not E or L enhancers.
  - Spatial distributions: H3K27ac, H3K79me3 show bimodal distributions around the center of enhancer (defined by ChIP-seq data of TF and motif), suggesting nucleosome exclusion at TFBSs. Pol II does not have this distribution.

- Predictive model of enhancer activation from histone states:
  - Bayesian network model: use 6 markers as input variables, and active or inactive status of 144 enhancers as output variable. Basically regression, but (1) bimodal distribution of the input variables, instead of normal. Modeled with mixture. (2) input variables are discrete (after mixture modeling), thus allow non-additive effects (e.g. PolII and H3K27ac).
  - Results: Figure 5b. AUC of prediction in 4-fold cross-validation is 0.82. The predictive features are Pol II, H3K79me3 and H3K27ac, H3K27me3 was contraindicative and H3K4me1, H3K4me3 and H3K36me3 had no predictive value for activity.
  - Prediction and validation: apply the model genomewise and predict 112 regions. Test 9 regions, all but one show activity in mesoderm in 6-8h.

- Question: need to answer the specificity of the proposed histone markers. The study was performed on enhancers, so K27ac may distinguish active and inactive enhancers; however, it may also appear in other types of sequences (e.g. promoters, coding sequences).

3.3.3 Chromatin Accessbility

DNase I hypersensitivity technology and data analysis:

- Reference: [Sabo & Stamatoyannopoulos, Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays, Nature Methods, 2006; John & Stamatoyannopoulos, Chromatin accessibility pre-determines glucocorticoid receptor binding patterns, NG, 2011]

- Technology: DNase I cleaves the genomic DNA at hypersensitive (HS) sites, in particular, the double hits (two cleavage 'hits' occurring close to each other, i.e. less than 1,200bp) create DNA fragments. These short fragments are isolated by size fractionation on a sucrose gradient, and then followed by sequencing.

- Hotspot identification: the data from the experiments consists of tags (cleavage events). The HS regions are those with high tag density.

- z score of windows: for a 250 bp window, compute its z score. Suppose there are n tags in the window, and N tags in the surrounding 50K bp region. The number of tags in the window follow binomial distribution $\text{Bin}(N, p)$ where $p = 250/50K$. This allows the calculation of z score:

$$Z = \frac{n - Np}{\sqrt{Np(1-p)}} \tag{3.3}$$

  An alternative is to use genomic background to calculate p and the z score, and report the smaller value of the two.
- 'Hotspot': defined as a succession of 250-bp windows, each of whose z-score was greater than 2.
- Refinement of hotspots: in regions of high enrichment, the local background is inflated and can prevent other HS sites from being identified. So after performing the first round of hotspot identification, delete all tags falling in the first-pass hotspots. The hotspots from the first and second passes were combined, and all hotspots were rescored using the deleted background.

- Peak finding:
  - First, neighboring hotspots within 150 bp of each other were merged.
  - Peak finding within each merged hotspot: (1) peaks (150bp) above the ninety-ninth percentile of the density; (2) if none of these peaks, simply peaks with maximum density value in the hotspot.
- Genomic distribution of HS sites: [Figure 4, Sabo2006]
  - Proximal promoter: 6.4% of lymphoid DNase I hypersensitive sites situated within the proximal promoter region (first 500 bp upstream of the TSS)
  - Distal promoters: 29.8% of DNase I hypersensitive sites within up/down-stream 2,500 bp of TSS.
  - More than 50% are located more than 10kb away from the nearest TSS.

p300 predicts enhancer activity [Visel & Pennacchio, Nature, 2009]:

- Methods:
  - Data: p300 ChIP-seq in mouse forebrain, midbrain and limb tissues, in E11.5 embryo.
  - Validation: enhancer activity in transgenic mice also in E11.5 embryo.
- Results:
  - ChIP-seq predicts (at FDR < 0.01) 2,543, 561 and 2,105 peaks in three tissues. Most in a single tissue.
  - p300 peaks are highly predictive of enhancer activity: $> 80\%$ of 86 tested regions show in vivo enhancer activity, most of which in the correct tissue.
  - Between 86% and 91% p300 peaks overlap with regions under evolutionary constraint in vertebrates.
- Discussion: the conservation of p300 peaks is contradictory to the lack of many functional sequences in ENCODE results. Possible explanations:
  - Developing embryo vs cell cultures in vitro.
  - TF binding (ENCODE) may not be fully suggestive of function.

Chromatin accessibility pre-determines glucocorticoid receptor binding patterns [John & Stamatoyannopoulos, NG, 2011]:

- Problem: the causal relation between binding of TFs and chromatin modification events: preexisting chromatin states to permit TF binding or TF binding changes chromatin states?
- Experiment: glucocorticoid receptor (GR) binding upon ligand stimulation. If TF binding is the driving event, then chromatin state (HS sites) will change significantly upon ligand stimulation. If chromatin state preexists TF binding, then HS sites will not change significantly upon ligand stimulation and the TF binding events should map to the HS sites.
- Preexisting states: the great majority of GR occupancy sites (71%, 5,865 sites) were targeted to the 2.1% of the genome defined by preexisting strongly DNase I-sensitive regions. An additional 9% of GR sites are in the weak DNase I-sensitive regions.
- Transcriptional activation: found no clear relationship between glucocorticoid receptor occupancy patterns and transcriptional activation of nearby genes.
- Influence of chromatin context (DHS) and sequence on GR binding:
  - GRBE motif alone is a poor predictor of binding: of 2,296,115 GRBE (15 bp) matches in the genome, only a very small fraction were actually occupied in vivo after hormone treatment. Many GRBEs with a high matching score were not occupied by a glucocorticoid receptor.
  - Define GRBE sequence classes: by the K-mer sequences ($K = 15$). A total of 1,100 classes.
  - Sequence class and chromatin conext on GR binding: (regression problem with two explanatory variables) stratification on sequence classes. The effect of chromatin context on GR binding is measured by CCC (chromatin context coefficient): the fold increase of GR binding of sequences in HS regions vs. non-HS regions (pre-treatment). CCC distribution: half of sequence classes have no binding in non-HS regions (CCC $= \infty$); the rest from 2-fold to 473-fold.

An expansive human regulatory lexicon encoded in transcription factor footprints [Neph and Stam, Nature, 2012]

- Data: 200M reads per cell type, 41 cell types. 1.1M footprints per cell type.
- TF footprints are quantitative measure of binding: quantify the strength of footprints by depletion of reads. For NRF1, among all 4,600 motif matches in DHS, half have footprints. Most motifs in footprints, 89%, overlap with ChIP-seq. Good correlation of footprint strength and ChIP-seq signal.
- Footprints capture TF-DNA interactions: e.g. USF1, the nucl. not in contact with protein are not protected.
- Footprints are enriched with allele-specific chr. accessiblity (ASCA): in some cases, ASCA are due to lack of footprint in one allele.
- Footprint analysis distinguish direct from indirect TF binding: estimate the proportion of ChIP-seq peaks (with motifs) that also have footprints. The estimates vary widely for TFs, and for some TFs, the proportion vary between promoter and distal regions. Do pairwise TF analysis to find TF whose indirect binding are associated with direct binding of another TF.
- Identifying cell-type specific motifs from DNase footprints: 200 new motifs, and they show similar conservation.

DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence [Sung and Hager, Mol Cell, 2014]

- A new method to call footprints: dnase2tf, first test if a candidate region shows depletion (binomial test), then merge adjacent candidate regions if the merged region is better.

- Comparing DHS footprints with ChIP-seq: for motif matches with footprints, good correlation of footprint strength and ChIP-seq. However, a large fraction of ChIP-seq sites leave no footprints.
- Shape of DNase footprints: reflect nuclease cut preference. Evidence: GR binding, same shape of footprints before and after treatment. Also in naked DNA.
- TF footprint protection correlates with DNA residence time of TFs: e.g. CTCF, strong protection/footprints; GR, short-lived binding (¡10s), no protection/footprints. Many TFs have short residence time: e.g. P53, NF-kappa, ER.

Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape (PIQ) [Sherwood and Gifford, NBT, 2014]

- Background: two options of identifying TFBSs (1) Call footprint first, then identify motif matches; (2) Candidate motif matches, then determine if there is a footprint. Choose (2) in this paper: statistically easier problem if focus on motifs (stronger priors).
- Model ideas: motif first instead of de novo footprinting; spatial smoothing; integrate multiple experiments, including time-course (learn the experiment structure).
- Model of background: consider one sample sample, let x_i be the read count of base i, we assume $x_i \sim \text{Pois}(\mu_i)$, and use the Gaussian process for μ_i:

$$\mu \sim N(0, \Sigma) \qquad \Sigma_{ij} = \text{Cov}(\mu_i, \mu_j) = \sigma_0 k_{|i-j|} \tag{3.4}$$

  To account for multiple samples (e.g. time course), further model the correlation of counts across samples using MVN.
- Model of binding: assume TF binding affects the rates of nearby window (400bp). Let m be a motif match, define I_m as the indicator of whether it is a true TFBS. Let β denote the effects of TF (change of chromatin accessbility as well as protection against DNase cut). Our distribution of rates becomes:

$$\hat{\mu}_i = \mu_i + \beta_{i-j} I_m \tag{3.5}$$

  where j is the motif start. The paper considers strand bias.
- Incorporating prior: $P(I_m = 1)$ increases with motif scores and DNase read counts.
- Inference: (1) Estimate all background parameters. (2) Estimate TF-specific parameters and inference of I_m for all TFs. This is done via Expectation Propagation, which evaluates $\text{E}(I_m|D)$ at each step, and compute parameters from the expected values.
- Benchmarking PIQ: from 1300 TFs, about 700 have profiles. Evaluate performance with ChIP-seq: AUC 0.93, for Centipede 0.87, and DGF 0.65.
- Pioneering TFs: estimate the effect of TF on chromatin opening. Two strategies: (1) time-series ATAC-seq, how chromatin accessibility near a motif is changed by binding from the previous time point. (2) Direct assess motif effect on nearby reads. Both (1) and (2) give similar results. About 100/700 TFs show detectable effects as pioneering TFs. However, even for these TFs, they do not bind to a large fraction of their motif matches.
- Settler TFs: rely on open chromatin.
- Questions: (1) a window around a motif match may have other motif matches. How do multiple TFBSs modify the rates, additive model? (2) PIQ uses a prior for I_m, which depends on motif scores and DNase-seq counts, but the latter is data?

Single-cell chromatin accessibility experiment [Sebestian Pott, Dec, 2015]

- Challenges with single-cell chromatin accessibility experiment: For each cell, only one DNA copy, and any open chromatin regions will generate a small number of fragments. Because of Loss of DNA from each step, generally only one fragment survives and will generate only one non-duplicate read (usually that's what matters - as each fragment will be PCR'ed, so duplicate reads should be simply from PCR). So the data is extremely sparse.
- Single-cell ATAC-seq: problem is that if we do not have reads in a region, it could be, lack of sensitivity or the region is closed.
- Background: NoMe-seq, GpC methyltransferase adds methyl group to GpC (in C) in open chromatin, then do Bi-sulfate conversion: unmethylated C will become T, while C in open chromatin will be protected. In general, 75% of DHS contain 5 or more CpGs.
- Single-cell NoMe-seq: extract nuclei → treatment with GpC methyltransferase → FACS sorting of nuclei (one nucleus per cell in 96-well plate, in practice, use 1-1000 cells) → Bi-S conversion → library preparation and sequencing.
  - FACS sorting: each nucleus is fluroescence labeled, and FACS allows cell passing one at a time time, and select the labeled cells.
  - Bi-S efficiency is believed to be very high. For GpC-MT, the conversion efficiency is not known, but use a very high level of enzyme in experiments.
  - Cell cycle issue: cells at different phases have different DNA copy numbers, and thus they generate different FACS (fluroesence) signals.
- Problems of the experiment:
  - Negative controls: would be needed to estimate the false positive rate.
  - Enzyme (GpC-MT) efficiency: if it is far below 100%, then any C in open chromatin will not be protected in Bi-S step, leading to the wrong signal. We need to estimate the efficiency. Ideas: use some positive controls, e.g. single-cell RNA-seq (enhancers near highly expression genes), high intensity regions in bulk-sequencing studies.
- Possible applications of the technology: generally, hard to know the truth of each cell, so estimate properties of populations, e.g. in how many cells, a region is open.
  - Co-regulation of regions: regions that are open or closed together in the same cells.
  - Clustering of subpopulations: similar to the problem of clustering with missing data.
  - Joint analysis of single-cell chromatin and expression data: if separate single-cell experiments of chromatin and RNA-seq, need to align the cells (e.g. by subpopulations). Could also use fluroesence to label certain TFs of the same cells (Yuwen).

3.3.4 Profiling TF-DNA Interactions

Ref: ChIP-chip technology [Buck & Lieb, Genomics, 2004]

Experimental procedure: two main steps, chromatin immunoprecipitation (ChIP) and DNA hybridization

- DNA sonication (into about 1kb fragments) and formalaldehyde cross-link (to fix DNA-protein interaction, may also fix protein-protein interactions)
- IP: enrichment of bound DNA
- Reverse cross-link, DNA purification and labeling
- Hybridization: Cy5 (IP-enriched DNA) vs Cy3 (background DNA: mock IP withno or irrelevant antibodies, otherwise, the same procedure)

In yeast experiment, resolution of about 1kb.

Experimental issues:

- Platform and arrayed elements. The question is what platform should be used, PCR-prodcuts or oligonucleotide array; what elements should be in the array.
  - Only promoter elements (about upstream 1kb) in the array
  - Entire regulatory region of particular interest
  - Oligonucleotide array: should not be too short (20-25bp) because hybridization may be affected by the variation of GC content of the short oligonucleotides.
  - A sample of the genome (non-tiling), e.g. every few kbps. Dififcult to separate the effect of binding affinity and spacing
  - Whole-genome tiling array: preferred

  Other considerations:

  - Arrayed elements shorter than sheared chromatin fragments (1kb) will not increase resolution
  - Measurement is only relative to other elements in the array, therfore, should have enough presumed non-targets in the array.
- Repetition. Some guidelines:
  - Number of replicates depend on the enrichment, e.g. 8-fold enrichment may need only 3 repeats. Increase enrichment by the IP procedure, e.g. antibody specificity.
  - Repeat with independent sample is important: vary as many irrelevant parameters as possible.

Data analysis:

- Degree of enrichment are measured relative only to other regions represented by other arrayed elements. E.g. in different experiments, the DNA amplification ratio may be very different, thus the Cy5/Cy3 ratios are not comparable.
- Data normalization: default normalization is to multiple a constant to Cy5/Cy3 ratio s.t. the median ratio is 0 in the array. This is based on the assumption that the distribution of ratio is symmetric, but in fact, the targets should have a higher ratio, i.e. the distribution should be skewed towards higher values.
- Median-percentile rank: use rank instead of the ratio, if an element is a true target, then its rank should be consistently high in different replicates. For each experiment, obtain the rank of an element, then determine the median rank of each element across all experiments: the true targets should have high median ranks. The cutoff can be determined by inspecting the distribution: should see two peaks (one for non-targets and the other for targets). The method is not very applicable if the number of targets is small.
- Single array error model: take the weighted average of log-ratio of any element across all experiments. The idea of weighting is: the low intensity signal has a higher uncertainty.
- Sliding window approach: for any region, add log-ratio of all elements in this region. The true targets should correspond to peaks because a true target will appear in multiple adjacent elements (thus, the total signal is strong). The confidence of each peak is based on the number of indepedent arrayed elements used to construct the peak.

Protein-binding microarray (PBM): [Bulyk, COBT, 2006; Berger & Bulyk, NBT, 2006]

- Procedure: epitope-tagged or fluroescence-labeled TF binds directly to the double strand DNA elements in the microarray, then detect the bound TF by the antibody to the epitope or fluroescence. The arrayed elements can be short oligonucleotides or PCR prodcuts.
- Studies using PBM: the PBM method is validated by (i) agreement of motifs derived from PBM and those from ChIP-chip; (ii) predicted TFBSs are highly conserved across yeast species.
- Comparison between *in vivo* and *in vitro* methods:
  - The main limitations of ChIP-chip are (i) requires a specfic antibody; (ii) limited resolution.
  - PBM method: (i) arbitrary protein concentrations and arbitrary buffer conditions; (ii) ignore the coactivators, chromatin context; (iii) the fusion protein may not have the same binding property as the original protein
- Main sources of variability of probe intensity (other than the binding affinity of sequences):
  - The position of the binding sites: sites positioned more proximally to the glass surface tend to have lower signal intensities.
  - The orientation of binding sites: that for some TFs there is a preferred orientation.
  - The flanking sequence within which a site is embedded: the presence of additional moderate or low affinity sites besides those under consideration can increase the observed signal intensity.
- E-value: use the fact that a K-mer is represented by multiple probes to remove the possible biases above. This is a modified form of the Wilcoxon-Mann-Whitney (WMW) test. Given a K-mer, we call all probes containing it the foreground, and the rest background, then the intensity of the K-mer is evaluated by the ranks of the foreground vs. the background probes:

$$\text{area} = \frac{1}{B+F}\left[\frac{\rho_B}{B} - \frac{\rho_F}{F}\right] \tag{3.6}$$

  where B and F are the sizes of the background and foreground, respectively, and ρ_B and ρ_F are the sums of the background and foreground ranks. This metric is sensitive to low outliers, as a few features with high ranks could greatly increase ρ_F. So only consider the top half of probes in the foreground and background sets.

Microfludic device for measuring TF-DNA binding [Maerkl & Quake, Science, 2007]:

- Motivation: PBM approach cannot detect transient TF-DNA interactions.
- Methods:
  - Mechanically induced trapping of molecular interactions (MITOMI) approach:
    - Device: 2,000 unit cells, each with a DNA chamber, detection area and a valve. DNA sequences (all variations of consensus) are spotted on the microarry, and the TF molecules (His tagged) are synthesized and localized on the substrate (linked with antibody to His5-tag).
    - Mechanical trapping: Once TF and DNA sequences reach equlibrium interaction, the button in each cell is actuated and creates mechanical trapping: all unbound TF molecules are excluded.
  - Four bHLH TFs: human MAX (A and B isoforms), yeast Pho4 and Cbf1. All 4-bp variations of the consensue sequence.
- Results:
  - Binding enery landscape: PWM fails to predict low-affinity binding, which is often due to non-specific interaction. PWM has only limited success in predicting affinity of high-affinity (56% of all double substitutions). The results thus demonstrate the importance of non-addivity.

- Importance of flanking sequences: Cbf1 and Pho4 have essentially identical consensus sequences, however they bind different targets in vivo. The difference may be due to very different flanking sequences (3bp flanking consensus) of Cbf1 and Pho4.

Simulating ChIP sequencing [Zhang & Gerstein, PLCB, 2008]:

- Aim: a method to determine the threshold for TFBSs in ChIPSeq data.
- Background:
  - ChIPSeq profile: the IP-ed DNA fragments are sequenced, then mapped to the genome. Each nt. in the gneome now has a tag profile: the number of tags that cover this nt.
  - Tag clusters and tag counts: the overlapped tags form a tag cluster and has a tag count. This number is used to select binding sites.
  - Statistical significance of tag counts: suppose the tag count of a cluster m is y_m, we need to determine the P value of y_m. Place n tags (total number of sequence tags in the experiment) randomly in the genome according to some background distribution, then the percentage of clusters whose counts are equal or higher than y_m is the P value.
- Methods: the problem is effectively to decide a good null distribution s.t. P values can be computed.
  - Simulating ChIPSeq data generation: (i) randomly place TFBSs in the genome; (ii) sampling weights in background sequences: uniform or gamma distribution (1-kb blocks, within each black, all nts have the same weights); (iii) sampling weights in TFBSs: power-law distribution; (iv) place n tags according to the weights.
  - Fitting the background distribution: two ways
    - Simulate the entire data and compare with the real data, determine the background and foreground (TFBSs) distribution simultaneously.
    - Use negative control to fit the background distribution.
- Results:
  - Both the background and foreground distributions (sampling weights) are not uniform: not agree with data if using uniform sampling.
  - Uniform background model increases false positive sites.

Quantitative enrichment of sequence tags: QuEST [Valouev & Sidow, Nature Methods, 2008]:

- Aim: identify TFBSs in ChIPSeq data.
- Methods:
  - Density profiles: at the neighborhood of each position i, create density profiles (in both strands) using Gaussian kernel density estimation (effectively, "smoothing" of tag counts). Then the two profiles are combined to form a combined density profile (CDP).
  - Peak calling: the local maximum of CDPs are identified and are candidate binding sites.
  - FDR for the number of peaks: the negative control data is split into two datasets, one for pseudo-ChIP dataset, and the other background. The FDR is estimated as the ratio of the number of peaks predicted in the pseudo-ChIP analysis to the number of peaks identified in the real ChIP experiment.

CisGenome [Ji & Wong, NBT, 2008]:

- Aim: an integrated system from ChIP-chip and ChIPSeq data analysis.

- Methods:
  - System: data processing, binding region identification, visualization of results, statistical summaries, motif analysis.
  - One-sample analysis: all regions with tag counts higher than a threshold. The FDR of that threshold is determined by: modeling the count in non-bound region with a background distribution. It is found that negative binomial distribution is better than Poisson distribution (larger variance).
  - Two-sample analysis: the enrichment of tags relative to the control is used to identify peaks.

3.3.5 Enhancer Activities

Widespread transcription at neuronal activity-regulated enhancers. [Kim & Greenberg, Nature, 2010]

- Data: ChIPseq in primary neuronal cultures, stimulated with KCl: the influx of Ca_2^+, which ten triggers signaling pathways and eventually changes in gene expression.
- Defining enhancers: CBP binding regions that are located within 2kb of H3K4me1-modified region (removing those that also bind to H3K4me3, which could be promoters, about 7%).
- Activity regulated enhancers: few than 1,000 CBP binding sites in unstimulated condition, upon stimulation, 28,000 CBP binding sites (most outside 1kb of TSS). By intersecting with H3K4me1 data (see definition), 12,000 enhancers.
- Properties of enhancers:
  - Dynamic: histone modification is relatively independent of stimulation. Three TFs: CRED and SRF binding - modest induction; NPAS4 - induction by stimulation.
  - Spatial: H3K4me1 modification in both sides (less than 1kb) of the CBP peaks. The CREB, SRF, NPAS4 and CBP binding are within 100bp of the highly conserved center of enhancer domain.
- Enhancer transcription:
  - RNAPII binding: at about 3,000 enhancers, and binding level increase about 2-fold upon stimulation.
  - Transcription: short RNA (< 2 kb) at 2,000 extragenic enhancers, and whose level change about 2-fold after stimulation. Transcription is strand-specific. The expression level is also correlated with the mRNA expression of the adjacent gene, and transcription depends on the promoter (enhancer-promoter interaction).
- Discussion:
  - Histone modification, TFs and P300: histone modification is static (not regulated) and perhaps marks the open chromatin. Some TFs may help maintain the chromatin state/histone modification, thus tend to be static; while other TFs may binds to the enhancer upon signaling (dynamic). P300 marks the active transcription, thus dynamic.
  - Model of eRNA: either eRNA itself is functional, or RNAPII at the enhancer (through promoter-enhancer interaction) is functional, e.g. through recruiting histone methyltransferase.

Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq [Arnold & Stark, Science, 2013]

- Overview: key steps for estabilishing a new technology, using STARR-seq as an example.
  - Concept: self-transcribed regulatory sequences (no need of bar-coding).

  - Details: key steps including transfection of reporter library to host cells (efficiency); activation of enhancers (trans-environment matters) and RNA-seq.
  - Quantification of results: variation could be introduced in each step above, so use control (input DNA).
  - Technical validation: biological replicates.
  - Functional validation: luciferase assay.

- STARR-seq procedure: (1) reporter library preparation: reporter construct consists of minimum promoter, ORF, query enhancer and poly-A site. (2) transfection. (3) RNA-seq.
- Peak calling: have both STARR-seq and control. In controls, only input DNA, no promoter, poly-A, etc, so there shouldn't be expression. For the same DNA segment in genome, the enrichment in STARR-seq over input measures its activity. Test: binomial test, FDR estimated from permutation. Results: 6K peaks at FDR 1.8%.
- QC and validation: biological replicates, PCC is 0.92. A high fraction of peaks are validated by luciferase (81%). Also the peak strength correlates with luciferase activities.
- Epigenomic marks: if the peaks are real enhancers, many of them should have characteristic epigenomic marks. Found that 69% of peaks are in open chromatin. Most of these peaks have expected marks (K27ac, etc.). For the peaks in closed chromatin, show signatures of repressive chromatin in native context.
- Understanding of gene regulation: strongest enhancers in TFs and housekeeping genes. However, ribosomal protein genes show poor enhancer activities probably because they rquire TCT-motif containing promoters.
- Remark:
  - Chromatin marks (DHS, H3K27ac, etc.) are not perfect marks for enhancer activities. How do we use STARR-seq to understand why some potential enhancers are active while others with similar epigenomic marks not?

The power of multiplexed functional analysis of genetic variants, Nature Protocol, 2016

- Procedure: (1) construction of a variant library (i.e., allelic series) of the sequence of interest, (2) delivery of this variant library to an in vitro or in vivo system, (3) the functional assay (i.e., the stratification of variants by function), (4) sequencing to quantify each variant's representation in the context of the assay, and (5) calculation and calibration of functional scores for each variant.
- Construction of library: PCR suffers from polymerase bias. Methods using individually synthesized oligonucleotides are expensive. Microarray based dna synthesis: release probes from arrays.
- Delivery of the library: via episome or insertion into the genome. If use transcribed barcodes as measure, its ok to deliver multiple alleles per cell.
- Functional assay: regulatory variant, cis-linked with a reporter transcript containing a unique barcode for each variant; then targeted RNA-seq (only the ones with barcodes). Only need to sequence barcodes.

Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits [Ulirsch and Sandaran, Cell, 2016]

- MPRA experiment:

  - DNA library: 2700 SNPs in high LD with GWAS hits. Select 3 sliding windows of size 145 bp around each SNP position. For each construct (allele and sliding window), 14 barcodes. Total of 200K distinct plasmids. Each expression vectors contains the test sequence, minimum promoter and barcodes. Sequencing.
  - RNA library: transfection of the plasmids to cells. Isolate mRNA, reverse transcription, PCR on 3' end, then sequencing the barcodes.

- MPRA data: 200K distinct plasmids, about 30-50M reads, on average, each unique barcode has about 100-300 reads. Note: Figure 1B, log2 CPM distribution is wrong.

- Analysis of MPRA data: for each barcode i, define its activity as:

$$a_i = \log_2 \frac{\text{RNA}_i}{\text{DNA}_i} \tag{3.7}$$

  To identify active constructs (regulatory activity), compare $\{a_i\}$ of all barcodes of a sequence with $\{a_i\}$ of the entire background by a Mann-Whitney test. To identify allelic-specific effects, compare $\{a_i\}$ of the reference vs. alternative alleles via two-sided Mann-Whitney test. If there are replicates, do quantile normalization first.

- Active constructs (ACs) and validation: about 4% of all test sequences. Classification using SVM (6-mer), AUC $= 91\%$. Validation: enrichment of epigenomic marks.

- Functional variations (MFVs) and validation: 32 MFVs, in 30% of GWAS loci. Modest effect, median fold change is 2.4. Claim: explain 14-22% of GWAS hits. Validation: enhancers, DeepSea prediction and TF binding.

- Remark:
  - Experiment: only need to sequence the barcodes, so use PCR to amplify the barcodes and do sequencing.
  - Not enough evidence that the MFVs are GWAS variants.

QuASAR-MPRA : Accurate allele-specific analysis for massively parallel reporter assays [Kalita & Pique-Regi, Bioinformatics, 2017]

- Goal: find allele-specific effects. For a test SNP, four data points: reference and alternative allele counts (sum of all barcodes) in DNA and RNA, respectively.

- Baseline statistical methods:
  - Student's t-test: let a_i be the activity (log2 fold change) of a sequence in i-th replicate, do the t-test (mean is 0?).
  - Fisher's exact test: compare four numbers.
  - Binomial test: fix DNA ratio, and test using RNA counts.

- Model: for a test SNP l, let r_l be the ratio of reference vs. alternative in DNA (assume fixed). The idea is that the ratio in RNA, ρ_l, may be slightly different from r_l, even for a null model. We treat it as Beta distribution, and test if the mean of the beta-distribution is equal to r_l. Let R_l, A_l be the ref and alt allele counts, and $N_l = R_l + A_l$, then:

$$R_l | N_l \sim \text{Binom}(N_l, p_l), \qquad p_l \sim \text{Beta}(\psi_l M_b, (1-\psi_l) M_b) \tag{3.8}$$

  where ψ_l is the expected proportion, and $\Psi_l = \rho_l(1-\epsilon) + (1-\rho_l)\epsilon$, where ϵ is the base-calling error rate, fixed at 0.001. The model is then testing if $\rho_l = r_l$ using LRT for a given r_l.

- Meta-analysis: to combine results from replicates, obtain the s.e. of $\beta_l = \log(\rho_l/(1-\rho_l))$, and use standard fixed-effect meta-analysis.
- Results: Figure 1. Fisher's exact test, Binomial test are highly inflated, T test less so, but still inflated.
- Remark: the problem is likely due to the compositional nature of the data. Consider the case of testing active constructs. If a significant fraction of constructs have no regulatory activity, they do not have counts in mRNA. Then all other constructs would show increased fractions in testing.

A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screen [Gasperini and Shendure, Cell, 2019]

- Experimental design: high multiplicity of infection (MOI) improves the power of detecting DE in CROP-seq. Multiplexing within cells (multiple gRNAs per cell) does not affect the effect sizes (Figure 7). In the pilot dataset, using high MOI finds 200 pairs while low MOI (one gRNA per cell, or 40 cells per gRNA) finds 60 pairs.
- Simulation to assess power as function of MOI: simulate data at the level of individual gRNAs. For one gRNA, MOI only affects the number of cells containing the gRNA. Then expression of gene is modeled by Negative Binomial.
- DE analysis on cis-pairs: combine gRNAs targeting the same enhancer. (1) use Monocle 2, negative binomial model. Regress out number of gRNAs/cell, batch, percent of mito. reads (each factor has a modest impact on power). (2) Monocle 2 p-values are inflated: filtering outliers by percent cells expressed, and use NTC results (non-targeting gRNAs) to obtain empirical p-values and do FDR. Also additional criterion (not clearly explained). (3) Focus only on pairs with reduced expression.
- Study design: (1) Positive control: gRNAs target TSS of highly expressed genes; and alpha-globin. (2) Negative control: NTCs, e.g. gene desert.
- Pilot data: 1000 DHSs, 15 gRNAs/cell and 500 cells expressing gRNA. Confirm the effects in positive and negative controls. Show QQ plot vs. NTC distribution: found 145 enhancer-gene pairs.
- Full data: 5000 DHSs, 250K cells. 28 gRNAs/cell and 900 cells/gRNA. Slight inflation in NTC results: used for correcting p-values and FDR. Effect size distribution: mode at 20% reduction of expression (4-80%). Also gRNAs targeting the same enhancer have correlated effect sizes (most highly repressed genes). Results: 664 DHS-gene pairs.
- Validation by bulk RNA-seq: most are replicated with similar effect sizes.
- Characteristics of found pairs: (1) Distance: median to TSS is 25kb, and in 33% not the nearest genes. Note: only test upstream enhancers (concern that intronic sequences directly change transcription). (2) Enhancers: enriched with P300 (OR = 1.8), K4me1 and K27ac (OR = 1.6), some lineage-specific TFs (1.5 fold). However, not enriched with conservation. (3) Co-enrichment of TF pairs: in promoters and enhancers, 6 motif pairs and 24 TF pairs based on ChIP-seq.
- Lesson: a relatively large set of positive and negative controls. Positive: TSS of highly expressed genes. Negative: NTC.
- Lesson: DE analysis leads to inflated p-values. The study use a large number of negative control to obtain empirical null distribution.

Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations [Fulco and Engreitz, NG, 2019]

- Experiment (Figure 1): test effects of enhancers on expression of a target gene. CRISPRi library targeting enhancers, use RNA FISH to sort cells into 6 bins, then sequence gRNA in each bin. Each element is targeted by a large number of gRNAs.

- Data: 28 target genes with enhancers (DHS) in 500kb of each gene, total of 3000 enhancer-gene pairs.
- Summary of results: (1) Among all enhancers, only 10% affects at least one gene. A small percent (20-30%) affects 2 or more genes. Most are within 100kb of genes. (2) Most genes have single enhancer, but a significant fraction have two or more enhancers. In particular, ubiquitous genes have few enhancers, and tissue-specific genes have 2.6 enhancers. (3) Effect size: 80% activate expression, and effect size 5-93% fold change.
- Prediction of enhancer-gene pairs: 141 positive enhancer-gene pairs, and the rest negative.
- Baseline model: distance, Hi-C and co-activation, none works well. For distance: low precision, since most DHSs within 100kb of a gene have no functional effects.
- ABC model to predict enhancer effects (Figure 3B): ABC effect of an enhancer on a gene, is defined as enhancer activity (geometric mean of DHS and H3K27ac read counts) $\times$ hi-C contact score (KR-normalized contact frequency at 5kb resolution) divided by the sum over all enhancers near that gene. ABC scores are shown to work well: AUPRC = 0.7, or recall 0.7, can reach prediction 0.63.
- Analysis: (1) the improved performance is due to: some pairs have high H3K7ac but low hi-C, some opposite. (2) Results are not very sensitive to hi-C data: use average Hi-C over 8 cell types, similar AUPRC. (3) ABC model does not work well for all CREs, e.g. CTCF sites.
- Lesson: a relatively small percent of sequences have functions in a given context/condition, most are within 100kb, and often target a single gene with effects from 5-90%.
- Remark: (1) ABC score is not about prediction of target genes given an enhancer. So histone activity is very important in determining whether a DHS is enhancer or not. (2) Why still need hi-C contact when most interactions are within 100kb? Hi-C contact frequency reflect distance: if close, likely high frequency.
- Remark: its likely that many (if not most) enhancers have some functions in various conditions, but for a certain system, only a small percent may be functional.

3.3.6 Nucleosomes

Interactions among DNA, TFs and nucleosomes [Segal & Widom, NRG, 2009]:

- Binding affinity landscape and transcription:
  - Binding affinity: assume that nucleosomes are constantly changing between different states (ATP hydrolysis from chromatin remodeling complex), thus the nucleosome binding (as well as TF binding) can be considered as equilibrium process.
  - From affinity landscape to transcription: poorly understood - DNA looping, 3D structure, transcription elongation, etc.
- Determinants of nucleosome occupancy:
  - Largely encoded by DNA sequences: model could explain well the in vitro (naked DNA) data, and in vitro data correlates strongly with in vivo data (CC = 0.89).
  - TFBSs could affect nucleosome occupancy: competition between TFs and nucleosomes could affect nucleosome occupancy, e.g. Abf1 and Reb1, two abundant TFs, leads to nucleosome depletion.
  - Statistical position: the effect of nucleosome-disfavoring sequences extend well into its neighboring sequences.
- Nucleosome and TF interaction:

  - Indirect TF cooperativity: in sequences where mutliple sites are close to each other, each of the binding factors separately competes with nucleosomes.

- Distinct modes of regulation in terms of nucleosome binding:
  - High nucleosome occupancy promoters: typically have many TFBSs, high transcriptional noise, high rate of histone turnover, many targets of chromatin remodellers.
  - Low nucleosome occupancy promoters: few TFBss, low transcriptional noise, low rate of histone turnover, few targets of chromatin remodelers.
  - Low noises in low nucleosome occupancy region: perhaps high TF binding (thus low level of switch between occupied and unoccupied states). Also these regions are not often targetted by chromatin remodelers, which might results more rapid transition between states.

Nucleosomes: H2A.Z [Interactive Fly]:

- H2A.Z: a variant of H2A that is evolutionarily conserved. Essential in Drosophila and mice.
- Function of H2A.Z in yeast (Htz1): involved in both activation and silencing of transcription.
  - Htz1p present at HMR assembled a specialized chromatin structure necessary for silencing HMR [Dhillon, Cell, 2000]
  - Deletion of the gene encoding H2A.Z strongly increases the requirement for SNF/SWI and SAGA, suggesting H2A.Z is involved in gene activation [Santisteban, Cell, 2000]
  - Htz1 is enriched in the euchromatic regions and acts synergistically with a boundary element to prevent the spread of heterochromatin [Meneghini, Cell, 2003]
  - H2A.Z would first poise chromatin at an inactive gene by interaction with a transcription-related factor, or chromatin remodelling component. The H2A.Z-poised chromatin would allow a gene to be remodelled and activated quickly [Larochelle, EMBOJ, 2003]
- Function of H2A.Z in Drosophila:
  - Required for euchromatic silencing and heterochromatin formation. An ordered cascade of events leading to the establishment of heterochromatin, requiring the recruitment of the histone H2Av variant followed by H4 Lys 12 acetylation as necessary steps before H3 Lys 9 methylation and HP1 recruitment can take place [Swaminathan, GD, 2005].

H2A.Z nucleosome organization in Drosophila [Mavrich & Pugh, Nature, 2008]:

- Methods: ChIPSeq of both H2A.Z and bulk (both H2A.Z and H2A, i.e. all) nucleosomes in the embryo (8 - 12h) of Drosophila.
- Results:
  - H2A.Z nucleosomes: 85% of genes contain at least one H2A.Z nucleosome within 1kb of TSS. H2A.Z levels correlated with mRNA expression.
  - Nucleosome organization around TSS and the genic region: the canonical pattern (Fig. 10) - -1 nucleosome (bulk); a nucleosome free region (NFR) at the core promoter; and +1 nucleosome (H2A.Z). After +1, a uniform distribution of H2A.Z in the genic region (genic array) at interval of 175 bp (vs 165 bp in yeast).
    * -1 nucleosome: not incorporate H2A.Z as oppose to yeast, which does have H2A.Z in the -1 uncleosome.
    * +1 nucleosome: further downstream of TSS at +135 bp, thus TSS is not buried in +1 nucleosome (as opposed to yeast), suggesting that Drosophila may use other mechansims downstream to control transcription.

  - Motif organization around the nucleosomes: four classes: anti-nucleosomal, nucleosomal, fixed (relative to TSS) and random. The first two classes of motifs reside in nucleosome-depleted and enriched regions, respectively.
  - H2A.Z nucleosome positioning: +1 nucleosomes may be positioned in part by CC/GG-based nucleosome positioning sequences (NPS).
  - Nucleosome organization around 3' end of the genes: similar to 3' end. H2A.Z nucleosomes followed by a NFR at the 3'end, then bulk nucleosomes.

- Conclusion: differnent level of control of DNA acess in chromatin. In general, nucleosomes (H2A) block the DNA access. The core promoter region does not have any nucleosomes to allow an unimpeded access by the transcription machinery. In the promoter regions, replacement by H2A.Z variant opens the chromatin for access by other regulatory proteins.

- Questions:
  - How nucleosome patterns vary in different cell types?
  - The difference of nucleosome patterns upstream and downstream of TSS: upstream sequences do not use H2A.Z (presumably nucleosome free for the enhancers of genes under transcription) while downstreams ones have a periodic pattern of H2A.Z.

3.4 Single-Cell Epigenomics

Single-cell epigenomics: technology and validations [notes]:

- Validation: comparison of peaks vs. peaks from bulk experiments. Clustering, and compare cell-type specific peaks vs. known markers or genes.

Applications of single-cell epigenomics [notes]:

- Linking distal enhancers to target genes: correlation of epigenome and transcriptome. Biological questions: how do multiple enhancers work together? What is the effect of DNAm (outside promoters) on gene expression? To what extent RNA variation is determined by transcription vs. PTR?

- Identifying regulators of specific cell types/states: motif accessibility changes across cell types. Biological question: how do multiple TFs/motifs work together, e.g. are each of motifs active one at a time or simultaneously?

- Genetics: are eQTLs or GWAS variants driven by signals in specific cell types?

Single-cell epigenomics: techniques and emerging applications [Schwartzman & Tanoy, NRG, 2015]

- General challenges of single-cell technologies: isolation, reocvery (yield and efficiency of sequencing libraries) and scalability to large number of cells.

- Single cell information from bulk Bi-S sequencing data: reconstruction of epi-haplotypes (groups of CpG sites), typically 200-500 bps. No long-range information.

- Single-cell Bi-S sequencing: RRBS or PBAT to generate indexed sequencing libraries. Identify single chromosome methylaome in diploid cells.

- Data analysis: controlling for noises/confounders (negative controls should be used). Missing data and imputation (difficult for methylation, but possible for other epigenomic structure).

- Applications of single-cell epigenomics:
  - Defining cell niches, rare cell types.

  - Dissecting heterogeneous cell populations. Ex. tumor. Studying single-cell drug response.
  - Correlating epigenomic mechanisms.
  - Integrating with SC RNA-seq.

Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution [Shema and Buenrostro, NG, 2019]

- Comparison of epigenome and transcriptome: sources of discrepancy. (1) Multiple epigenomic states (Figure 1): repressed > poised, with no change on transcription. (2) Transcriptome changes with no epigenomic changes: e.g. stress response.
- Temporal trajectories of the epigenome: by ordering single cells in psuedotime, we have data similar to time-course data. This allows us to study the possible causal factors, e.g. enhancer preceding promoters? Or whether TF controls a gene.
- Cellular heterogeneity in gene regulation: (1) TF cooperation or competition: best to address at single cell level (Figure 2b). (2) Enhancer-promoter looping: may differ between cells, e.g. two enhancers controlling the same gene, at different cells, 0, 1 or 2 enhancers in contact with promoter (Figure 2c). Looping may change expression burst frequency.

Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state (Drop-ChIP) [Rotem and Bernstein, NBT, 2015]

- Idea: index (barcode) chromatin fragments from each cell before IP.
- Technology: MNase and cell in one drop, and DNA barcode with adaptor (PCR handle) in another drop. Fusion of drops and enzymatic buffers (DNA ligase), facilitated by electrostatic field. ChIP, amplification and lib prep and sequencing.
- QC and fine-tuning of technology: control flow rate and other parameters s.t. most of time, each cell is loaded with only one barcode. 10 times more barcode drops than number of cells s.t. each barcode is used only once.
- Validation by Cell-barcode ratio (Figure 2c): one barcode-one cell in 87% cases.
- Single-cell profile vs. merging of multiple cells vs. bulk ChIP-seq profiles: visual inspection show high concordance. Specificity: 50% reads in single cell data are in bulk peaks. Sensitivity: 800 peaks per cell, so sensitivity about 5%.
- Overall agreement of Drop-ChIP and bulk data (Figure S3): combine all peaks/reads from single cell data, and comparison with bulk. Most peaks are shared, and very high correlation of FPM.
- Possible to distinguish cell types using just a few hundred peaks (Figure 4c). Recover mixed proportion of cells (known mixture beforehand).
- Learning biology of subpopulations: obtain profiles of subpopluation of cells by combining data from all cells in a subpopluation.

Single-cell chromatin accessibility reveals principles of regulatory variation [Buenrostro and Greenleaf, Nature, 2015]

- Background: ATAC-seq, load Tn5 transpoase with adaptors, then the enzyme will cut open chromatin regions, inserting the adaptors. PCR amplification and sequencing. Need 50K cells vs. millions of cells for DNase-seq.
- Experiment: (1) Fludigm, within each chamber, lysate cells and do ATAC. Each Tn5 transposase is loaded with a cell barcode and adaptor. (2) PCR and library prep for each cells. (3) Pool the library and do sequencing.

- Data: about 200 cells, library size 10k - 100k reads per cell (mean 73K), and 20-50% are in open chromatin peaks (found from bulk experiments).
- Validation: (1) Correlation of aggregate reads from all scATAC-seq cells vs. bulk: $R = 0.8$. (2) Feature of peaks: distribution of aggregate single-cell reads highly centered near TSS.
- Analysis: For any peak in a single cell, usually 0 or 1 read. Data is very sparse: e.g. 10% of promoters are represented in a single cell. Analysis is performed at the level of a group of predefined peaks (e.g. TF motif, within a larger genomic region): for the group, obtain the count matrix (number of reads in each peak in each cell). Obtain the sum of reads of the peak set and assess the deviation by downsampling. Additionally, compute an overall variability score across all cells.
- Identifying TFs with high variability, also TF combinations.
- Co-variation of genomic regions: define genomic bins (median size of 135kb), and assess the deviation of each bin in each cell. Then correlation of the deviations of any two genomic bins in the same chromosome.

A single cell's open chromatin [NM, 2016]

- Idea: to minimize DNA loss (fragments from DNase cutting), added large excesses of circular carrier DNA. PCR adaptors do not ligate to circular DNA.
- Reference: Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature 528, 142146 (2015).

Single cell NOMe-seq [Seb Pott, eLife, 2017]

- NOMe-seq: GpC is usually unmethylated. After adding the enzyme, only GpC in accessible regions become methylated. Then do Bi-S conversion and sequencing.
- ScNOMe-seq procedure: treatment with enzyme, nuclei sorting in 96-well plate, Bi-S conversion, library prep. and sequencing.
- Results: 5-10% DHS covered by some reads at single cells. 50% DHS from bulk are accessible (GpC) in single cells.
- Detecting TF footprints: with 20 cells, found CTCF footprint. Validate by correlation of CTCF motif scores with footprint strength (percent of reads with methylated GpC).
- Collaboration with Salk: novel library prep. to increase genome coverage. 3000 cells, neurons in cortex. Number of unique reads >5M / cell.
- Grouping cells: in neurons, high level of non-CpG methylation, which in gene body correlate with gene expression. So obtain average gene body non-CpG methylation to group cells by t-SNE. Found neurons group by layer.
- Discussion: clustering in non-neuron cells, can still use average methylation in 100kb bins.

ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data [Schep and Greenleaf, NM, 2017]

- Goal: estimate if a motif is over-represented in accessible regions in a given cell vs. average cells. In other words, define motif characterization/representation of individual cells.
- Method: a pre-defined set of peaks, and motif information in peaks (present or not). Map reads/fragments to peaks for a given cell. Count number of reads mapped to all peaks with motif in that cell. To assess its statistical significance, compute Z-scores. Compare this number with the number of reads mapped to background peak set: chosen to match GC content and average accessibility (across all cells).

- Remark: the background set matches average accessibility. So if a motif is important in a large fraction of cells in the data, then it would not find the signal.

Strength in numbers from integrated single-cell neuroscience [NBT, 2018]

- Background: typical single-cell protocols require fresh biopsies. For archived samples, use single-nuclei.
- Experiment: sort human brain cells, then do single-nuclei RNA-seq and chromatin accessibility profiling (THS-seq, similar to ATAC-seq).
- ScRNA-seq analysis: 35 clusters of brain cells.
- Constructing chromatin accessibility profiles of clusters: harder to cluster THS-seq data because of limited dynamic range. So map the THS-seq cells to RNA-seq clusters, then merge all those cells to create the chromatin profiles for each RNA-seq cluster.
- Search for TFs for each cell type: find motifs overrepresented in differential chromatin regions.

Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain [Lake and Zhang, NBT, 2018]

- Experiments: 6 post-mordem brains, frontal and visual cortex; cerebellum. Single nuclei RNA-seq (snDrop-seq) and scTHS-seq (similar to ATAC-seq, more specific). Total of 60K cells (30K each type of data). For scDrop-seq: 6K reads per cell or 700 genes. For scTHS-seq: 10K reads per nuclei.
- Clustering of snDrop-seq: (1) normalization: first use library size (total number of reads), then estimate variance per gene (variance depends on mean), and use variance to normalize expression. (2) PCA: top 150 PCs from 2000 most variable genes. (3) Clustering: KNN graph, and community detection algorithm.
- Clustering of scTHS-seq: similar, but limited dynamic range (read count), use truncated Poisson.
- Validation of cell types: heat-map of expression of marker genes for the subtypes of neurons and non-neuronal cells.
- Joint analysis: obtain chr. accessibility profiles on refined cell types. To map each cell of scTHS-seq to cell subpopulations defined from snDrop-seq: (1) Build a predictor of gene to DHS site: use DE genes, and Diff. active sites. The classifier use locations of sites, etc. as features. (2) A cell lineage tree from snDrop-seq: for each branch point, learn DE genes, and use the corresponding sites to assign a cell to one of the two branches.
- Remark: the clustering analysis is based on normalized reads, which takes variance of gene expression into account.

Cicero Predicts cis-Regulatory DNA Interactions From Single-Cell Chromatin Accessibility Data [Pliner and Trapnell, Mol Cell, 2018]

- Method: Merge 50 cells and compute correlation of enhancer and promoter.

A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility [Cusanovich and Shendure, Cell, 2018]

- Dataset: sci-ATAC-seq on 13 adult tissues in mice. Total of 1-10K cells/tissue, and 8-23K reads/cell.
- QC: mostly based on bulk data (pooling cells). Replication across experiments. Comparison with earlier bulk data. Insert size distribution.
- Joint peak calling across 13 tissues: LSI clustering in cells of each tissue, then call peaks in each cluster. Merge peaks: total of 400K peaks. Create cell x peak matrix.

- Clustering: t-SNE, and Louvian clustering: 30 clusters. Then repeat for these clusters: total of 85 patterns (Figure 2A-C). Some patterns match tissues, but some are mixed. Also find cluster-specific peaks (DA peaks).
- Assign cell types to clusters: use cell-type specific expression. Use Cicero to assign peaks. (1) Promoter accessibility: find genes of cluster-specific promoters. (2) Gene activity scores: sum of the accessibility of all DA peaks assigned to a gene.
- Matching scRNA-seq and sc-ATAC data: assuming cells are labeled with types in scRNA-seq, our goal is to assign cell type labels to ATAC cells. For ATAC-seq data, create gene activity scores. Then we have the common space of variables (genes). We stack cells from RNA and ATAC data, and do PCA (Figure 3c) on all cells together. Then do KNN classification: for each ATAC cell, find 20 nearest neighbors of RNA cells, and use majority vote. Note: use promoter accessibility instead of gene activity scores also get good results.
- Motifs underlying each cell type/cluster: train CNN (Bassat) of OCRs of 85 cell types. Learn motifs using first level filters: 280 out of 600 filters. Create filter/motif-cell type matrix: remove one motif/filter a time, and assess drop of classification performance for each cell type. Projection of motifs to single cell space: let C_{ij} be the score of motif j to cell i, we compute it as $\sum_k A_{ik}B_{kj}$ where A_{ik} is the cell-site matrix, and B_{kj} site-motif matrix (presence). Then normalize C_{ij} by max.
- Chromatin accessibility specialization of cell types in tissues: even for the same cell type, their chromatin profiles may differ across tissues. Ex. from all DCs, clustering - 6 clusters, some exclusively from lung.
- Chromatin accessibility changes during hematopoiesis: do Monocle2 of cells, five main branches. Compare the accessibility profiles of branches vs. known cell types. F4 for erthyroids, F2 for lymphoids and the rest for myeloids. Ex. hemoglobin locus: co-accessibility of enhancer vs. promoter only strong in the erthyroids branch of the tree.
- GWAS h2g analysis: apply S-LDSC on 85 cell types.
- Remark: assigning ATAC-cells to RNA cell type labels, use majority voting, which does not adjust for different cell type proportions.
- Lesson: (1) Matching ATAC cells to RNA cells: could use gene activity scores to create a common feature space. (2) Motif analysis: projection of motifs to single cells.
- **Lesson**: to study development and differentitation, focus on activation of lineage-specific genes along the tree, in particular, check activation in particular branches.

cisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-seq Data [Gonzalez-Blas and Aerts, NM, 2019]

- ScATAC-seq: no UMI, duplicated reads usually removed.
- Model: LDA model. Each cell is associated with a set of topics (with probabilities sum to 1). Each topic is represented by a set of words (present or not): so the unknowns are z_{it}, whether word i is in topic t or not. The inference problem is to infer the posterior distribution of z and topic distribution in cells. Use Collapsed Gibbs sampling: to sample z_{it} given all other values of z's.
- Model selection: for number of topics, typically 5-50. Also hyperparameters: Dirichlet prior parameters.
- Cell clustering and cell state identification: each cell is represented by its topics. Can cluster or visualize cells by the probabilities or Z scores using t-SNE, UMAP, etc. Test enrichment of epigenomic features (suppose they are given) in a given cell: compute the ranking of regions in each cell based on reconstruction (imputation of regions based on topics of the cell), then test enrichment using AUCell.

- Topic investigation: Cell level view. Show the percent/enrichment of a topic in all cells/cell types, Figure 1e.
- Topic investigation: Region level view. (1) Enrichment of epigenomic signatures in a topic: if binary, use GREAT or similar; if ranking of regions in topics, use AUCell. (2) Motif enrichment, and differential motifs between cell types (MAST).
- Validation of method: using blood cell scATAC-seq data with known cell labels (from FACS). Evaluate cell type clustering by ARI (Figure 1cd)
- Application to human brain: (1) cell clustering (t-SNE), and annotate cells, using epigenomic signatures derived from earlier study (Figure 2a). (2) Comparison of topics with epigenomic signatures (motifs) using heat-map (Figure S10e).
- Application to mouse prefrontal cortex: annotate cell types/clusters using scRNA-seq data. Learn about 250 regulons from gene expression using SCENIC, then map regions to genes (closest genes), and obtain epigenomic signatures of regulons/cell types.
- **Lesson**: annotation of cell types/clusters from single cell epigenome data using two strategies: (1) epigenomic signatures of known cell types: from bulk or other sources. (2) Using scRNA-seq data: map peaks to genes, then annotate topics/regions by genes.

Assessment of Computational Methods for the Analysis of Single-Cell ATAC-seq Data [Chen and Pinello, GB, 2019]

- Procedures of scATAC-seq analysis: four modules: define regions, count features, transformation and dim. reduction. Methods different in these four aspects (Figure 2 as a summary).
- ChromVAR: features defined as motifs/k-mers.
- Cicero: use gene activity scores (reads in extended promoter regions) as features.
- Causnovich2018: Latent Semantic indexing (LSI). Define regions by calling peaks in initial cell types/clusters, then merge peaks. Count features and normalization by TF-IDF matrix. Then do SVD.
- SnapATAC: use uniform sized bins as features. Normalization of counts by library size of cells, then PCA.
- Simulation study: use 6 FACS-sorted cell types, and their peak sets. Randomly generate reads in each cell according to its assigned cell type. Compare methods by clustering vs. gold standard (known cell types), also visualize by UMAP. Results: SnapATAC and Causnovich2018 perform the best using ARI metrics of clustering, and cisTopic is close.
- Evaluation in real data: use FACS cell types; or marker gene expression (evaluate Gini index); or tissue source of origin.
- Performance comparison summary: SnapATAC and CisTopic the best across all real data, Causnovich2018 slightly behind.
- Problem of rare cell types: the first step, defining regions, is important. If use pseudo-bulk (most methods), then peaks in rare cell types may be missed at the very beginning of analysis! Recommend Causnovich2018 approach, learn the peaks in an unsupervised way.

ArchR : An integrative and scalable software package for single-cell chromatin accessibility analysis [Granja and Greenleaf, 2019]

- Input: BAM files. Stored in disk in HDF5, more compact. A group of HDF5 files are known as Arrow files, stored in memory.

- QC of data: use TSS enrichment scores to filter low-quality cells. Remove doublet cells: use synthetic doublets, projection on embedding, and do prediction using nearest neighbors.
- Determine peak sets: (1) SNAPatac: use 5-kb bins, too large. Signac: use pre-defined set. (2) Use 500bp bins. Arrow files still smaller than SNAPatac.
- Dim. reduction: use LSI. For very large dataset, construct LSI using a subset of cells, then project the rest.
- Methods for benchmarking gene activity scores: use matched single cell ATAC-seq and RNA-seq data. Create low-overlapping cell aggregates (100 cells each and 500 aggregates), and for each aggregate, obtain gene activity scores. Then map cells to scRNA-seq cells using Seurat, and do label transfer. This maps every scATAC cell to a scRNA-cell. Then obtain the cell aggregate by gene expression matrix. Evaluation metrics: correlation of genes across aggregates; or correlates of aggregates across genes.
- Gene activity score models and evaluation: use PBMC and bone marrow datasets with matched single cell ATAC-seq and RNA-seq. Evaluate a number of models to predict expression from genes. Best model (Figure 2ab, Figure S9): centered at TSS, exponential decay, and extended gene body. Also validation using paired bulk ATAC and RNA data from hematopoiesis (Figure S9k-m): use down-sampling in ATAC and RNA data to create pseudocells (10K fragments each). Similar results.
- Psudo-bulk analysis and validation: in the hematopoiesis dataset, find 21 clusters. Pseudo-bulk in each cluster. Overlapping with bulk data, and motifs of relevant TFs found.
- Identifying positive driver TFs: ChromVAR scores of TFs - variation of motif accessibility across cells. Positive drivers: correlation of gene activity scores (TF) with motif accessibility (in UMAP, Figure 3e). Footprint analysis (Figure 3f-h).
- Peak co-accessibility: Figure 3i. Method: use low-overlapping aggregates, based on Cicero. Define aggregate single cells using low-dim. similarity. Then create 500 aggregates that do not overlap (all with 80% overlap are filtered). Then compute Pearson correlations of two features across these aggregates, using log2-normalized cell aggregate.
- Integration of scATAC and scRNA: using hematopoiesis data. Alignment: choose top 2000 variable genes in scRNA-seq, and impute expression with MAGIC. Use CCA to find anchor cells, and extract expression from matched cells. Show concordance of gene activity scores and actual expression of marker genes (Figure S14a).
- CREs of gene expression: Figure S15a, correlation of peak accessibility and gene expression, similar to peak co-accessiblity analysis, except that expression data is from aligned cells from scRNA. Identified 70K significant linkages. Q: use all cells from scATAC, or just cells with good anchors?
- Cellular trajectory analysis: defining cells and trajectory (1) User-supplied backbone: e.g. one cluster for stem cell, one for progenitor, one for differentiated cells. Define cluster mean. In B-cell example: defined as HSC, LMPP, CLB, pre-B and B. (2) Define trajectory: clusters along the trajectory (using cluster mean), then fit smooth.spline function. (3) Identify individual cell positions along this trajectory based on min. Eucledian distance from a cell to the manifold.
- Dynamic and correlation analysis with cellular trajectory: group cells across the trajectory. Smooth any feature of interest along the trajectory by user defined smoothing window. Could also do correlation analysis using low-overlapping aggregates. Q: How to choose cells to define a feature along the trajectory?
- Trajectory analysis with B cell example: trajectory (Figure 4e), coordinated changes of peaks and genes using significant peak-gene links. Example loci: Figure 4fg. TF analysis along B cell trajectory: same trend of gene activity score (TF), gene expression and motif accessibility changes (Figure 4h).

Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimers and Parkinsons diseases [Corces and Montine, NG, 2020]

- Data: (1) Bulk ATAC-seq in 7 regions, 30 samples. (2) ScATAC-seq in 4 regions, 10 samples. (3) HiCHIP: use H3K27ac to enrich interactions. 65% of interactions involve OCRs in both anchors, 25% involve OCR in one anchor.
- Bulk ATAC analysis: large regional variations. About 150K peaks.
- ScATAC-seq: about 24 clusters, including ext. and inh. neurons, microglia, astrocytes, etc. Comparison with bulk: about double the number of peaks (300K). Most new ones are cell-type specific.
- Defining cell-type specific peaks: use feature binarization to define cell-type specific peaks. 6 main cell types: for each peak, compute its mean and s.e. of log-CPM at each cell type. Consider one peak, let $(\mu_1, s_1), \cdots, (\mu_6, s_6)$ be mean and s.e. of the peak in 6 cell types (ranked from highest to lowest). Then check if the first group is higher than the second, μ_1¿$\mu_2 + s_2$, and so on, until a break point is found (condition no longer satisfied). Then all groups above the break points are 1 and the rest 0. If no breakpoint found, the peak is discarded.
- Results of peak clustering analysis: Fig. 1h. about half are in single cell types, and the other shared among several types of neurons or among astrocytes, OPCs and oligodendrocytes. Use testing to select 220K peaks.
- Motif analysis in cell-type specific peaks: show matching motifs in each cell type; also show footprints.
- Refining cluster analysis: take only neurons, then do clustering, find 30 neuronal clusters.
- GWAS enrichment analysis with S-LDSC: AD show enrichment in microglia cells, SCZ ext. and inh. neurons (Fig. 2c).
- SNP annotations: overlapping with peaks. Delta-SVM effects and ASC in bulk ATAC-seq.
- Prioritization of candidate SNPs: PICALM locus of AD GWAS, one SNP is in a peak specific to oligodendrocytes. HiCHIP and co-accessibility to PICALM promoter. SNP also disrupt KLF4 motif from gkm-SVM analysis.

Cardiac Cell Type-Specific Gene Regulatory Programs and Disease Risk Association [Hocker and Bing Ren, biorxiv, 2020; Alan presentation, group meeting]

- Background: atrium and ventricle (push blood from heart to body). Common disease: Atrial fibrillation: arrthymia, atrium and ventricle not synchronized.
- Data: single-nuclei ATAC-seq in normal heart, 4 chambers. Median of 2.8K reads per cell. Total of 79K cells in ATAC and 35K in RNA.
- Data processing: normalization, RNA-seq use log2-CPM, ATAC-seq uses TF-IDF (however, IDF probably makes no difference). PCA/SVD. Then clustering and DEG analysis.
- Clustering: 9 from ATAC (SnapATAC), and 12 from RNA (Seurat). Annotate cell types from promoters in ATAC. Some discrepancy (Figure 1F): some cell types identified from RNA-seq, e.g. myofibroblast not identified from ATAC (genes in RNA-seq not correlated with promoters in any ATAC-seq clusters).
- Remark: possible explanations: (1) Sample difference > low proportion of some cell types. (2) Expression are enhancer driven. (3) Enhancer priming? But this is mature cells.

- Mapping CREs: 270K CREs, 4.7% of genome. 67% are known from earlier studies on all human tissues. 19K high cell-type specificity. Clustering the CREs: do GO enrichment and motif analysis. For enriched motifs: correlation of [TF] with accessibility of target motif matches (motif enrichment), across cells of all types. Ex. SP1 motif strongly enriched in macrophages, and SP1 is exclusively expressed in macrophages (Figure 2F). Using [TF] data can distinguish TF families with similar motifs: GATA.
- DA peaks between chambers: use edgeR on cells. Found about 10K are cell-type specific. Use co-accessiblity to link DA peaks to genes. Often associate with chamber-specific expression of genes. Candidate TFs for specific chambers.
- Associate CREs with heart failure: H3K27ac data from heart failure patients. Show that a large fraction of differential enhancers are from CM cells.
- Enrichment of CREs of GWAS risk variants (Figure 5A): S-LDSC on 5 heart-related traits including AF and CAD, stroke. AF: enrichment (by Z-scores) in atrial CM and ventricle CM. CAD, stroke, heart failure: no enrichment.
- Identifying putative causal variants: Fine-mapping at 111 loci using L = 1. 38 variants with PPA > 0.1 in CREs. Candidate: enhancer in intron of KCNH2 (K+ channel), co-accessible with KCNH2 promoter, total PPA of two SNPs = 0.28.
- Functional study: reporter assay to show allele-specific effects. Then in iPSC derived CM cells, CRISPR deletion of the enhancer > gene expression change > cardic cell physiology change (change of Action Potential).
- Q: how [TF] correlation with ATAC-seq is done? The paper does not do ATAC-RNA integration.
- Discussion: DEGs and DA analysis after clustering. This may seem circular, but the effects may be small. Clustering results are driven by many genes, and DEG testing is done on one gene a time. So imagine if we remove the tested gene from clustering, the results would be likely very similar.

3.4.1 Single-Cell Multi-omics

Single-Cell Multiomics: Multiple Measurements from Single Cells [Macaulay and Voet, TIG, 2017]

- Experimental approach for DNA and RNA-seq from the same single cells: I. whole cell lysis, then (1) RT to obtain cDNA and DNA; then use mRNA specific protocol for amplification (second strand synthesis and IVT) and PCR for DNA. (2) Physical separation of RNA and DNA amplification: use beads with poly-T to capture mRNA, e.g. G&T-seq. II. Physical separation of nucleus and cytoplasm: gentle disruption of cell membrane, but not nuclear membrane. Ex. scTrio-seq.
- Application of single cell DNA and RNA-seq: correlate CNVs with gene expression. Detecting and confirmation of SV events. Joint discovery of SNVs and detecting RNA editing. Detecting de novo eQTL by ASE (a new regulatory mutation drives ASE in nearby genes).
- Application to cancer: lineage reconstruction.
- Experimental approach for single-cell epigenome and transcriptome sequencing: DNA methylation via BiS-seq or RRBS and RNA-seq. Rely one physical separation of nucleus and cytosol. Ex. scM&T-seq.

Joint profiling of chromatin accessibility and gene expression in thousands of single cells [Cao and Shendure, Science, 2018]

- Single cell combinatorial indexing (SCI): how it differs from Drop-seq? Cells are sorted in a plate with many wells, each well will be loaded with one index. Then in the second round, each cell will have another index. The combination of two indices, e.g. 96 x 356, allows each cell to have an unique barcode.

- Sci-CAR: Fig. 1, in the plate, do RT and tranposition, then indexed PCR (for barcodes), then pool cells and do library prep.
- Lung cancer cells: RNA, 3K UMIs; ATAC, 1.5K UMIs per cell.
- Kidney cells: data about 13K co-assayed cells. Difficulty: cannot cluster with ATAC-seq alone because of read sparsity.
- Identifying cell-type specific chromatin:(1) Clustering cells by transcriptome; (2) Create pseudocells: similar transcriptome. (3) Group ATAC-seq data of pseudocells: very separate in t-SNE (Fig. 3C).
- Linking distal CREs with gene expression:use 200 psuedo-cells, do LASSO of expression vs. CREs within 100kb.
- Lesson: the main benefits of single-cell multi-omics, include the abilities to: (1) Create profiles of chromatin accessibility in each cell type: it may be hard to define cell type with scATAC-seq alone with low read coverage. (2) Study motif accessibility and [TF] changes across cell types and psueotime. (3) Linking distal CREs with genes: using pseudo-cells.

Integrative single-cell analysis [Stuart and Satija, NRG, 2019]

- Coupling cytometric measurement with scRNA-seq: index sorting by FACS, then scRNA-seq in microliter wells (Figure 2a).
- Joint profiling of single cell DNA/epigenome and mRNA: oligo-dT primer for amplification of mRNA; or selective amplification of DNA vs. mRNA. Adding Bi-sulfate treatment to do single-cell methylation measurement.
- Profiling proteins and RNA at single cells: epitope-tagging (Figure 2c). To profile a protein, e.g. cell surface marker, link antibody to the protein with DNA-barcode and polyA, then the DNA-barcode is RTed and is sequenced in scRNA-seq. The DNA barcode level reflects levels of the protein/surface marker.
- Lineage tracing: insert gRNA target sites in the genome, and put under the control of inducible promoter. During development, CRISPR/Cas9 create random mutations, which lead to lineage signature. Since they are transcribed, they can be sequenced by scRNA-seq. Similar to CROP-seq.
- Multi-modal single cell data analysis: joint clustering of multiple modalities, MOFA, multi-view method.
- Classification of cell types: (1) Mapping scRNA-seq data to reference: scmap. (2) Mapping other single-cell modality to scRNA-seq: then learn cell type specific epigenome or proteome profiles.
- Joint analysis of multiple single-cell modalities: MATCHER, 1D psuedotime that generates both RNA and epigenome data. LIGER: use integrative NMF. Interesting findings: DNA methylation changes often lag behind changes in gene expression [MATCHER paper].

Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity (LIGER) [Welch, Cell, 2019]

- Goal: integrate multiple single cell datasets with common features (genes). Some factors are shared across datasets, while others are dataset-specific.
- Model assumption: we assume that data are determined by shared factors. Some factors have the same effects on the genes in two datasets, but other factors have different effects. Ex. shared factors may represent processes such as protein synthesis, while unique factors represent cellular stress response (sensitive to environment/experiment). Let the effects of shared factors be W, and the unique factors be V_1, V_2 for the two datasets, respectively. In the two datasets, the latent factors are independently sampled, so there is no sharing, we denote them as H_1 and H_2 respectively.

- Model: we have iNMF model, $E_1 = H_1 \times [V_1, W]^T$ and $E_2 = H_2 \times [V_2, W]^T$, where V_1 and V_2 represent dataset-specific factors. The model minimizes the squared error and the penalty $\lambda \|H_i V_i\|^2$. The penalty term can be understood as PVE explained by dataset-specific factors, which we would like to minimize.

- Clustering: use reduced dimensions of per cell and done on each dataset separately. For each cell, we assign it to a single factor with max. loading, denoted as $F(i)$. Define for each cell i, $FN(i)$ the histogram of $F(i)$ for all its neighborhood cells in its own dataset. The difference between i and j is based on $FN(i)$ and $FN(j)$, then do community detection. Intuition: two cells are close if all their neighbors (in its own dataset) belong to the same clusters.

- Benchmarking performance: compare alignment (between datasets) and agreement (whether neighbors are consistent in separate and joint analysis). Comparable to or better than Seurat in two blood cell datasets. In t-SNE plot, cells of different datasets are mixed, instead of separate by batches.

- Joint analysis of scRNA-seq and scDNAm in brain: clustering analysis (Figure 6). Define 4 main cell types (via markers), then zoom in to cluster on each of main cell types - 37 clusters in total.

- Comparison of gene expression and DNAm: obtain expression and DNAm levels with each of 37 clusters, then compare across clusters. (1) MECP2 and TET3: gene expression correlate with global DNAm across cell types. (2) TF vs. target sequences (DMRs): for each TF, find top 10 best DMRs based on anti-correlation of [TF] and DNAm. (3) Cis-analysis: anti-correlation of DMRs and nearby gene expression.

- Q: Visualization: can show cells from different datasets in the same plot. Is this based only on shared factors?

- Lesson: to show that batch effect does not drive results, see if cells are separate by batches or cell types.

- Remark: to capture different covariance structure in two datasets, we could use the same weight matrix W, but different factors Z. Ex. we can imagine in patients vs. controls, the weight matrix represent factor (biological processes/TFs) to gene matrix the same, but the factors (e.g. TF expression) differ between conditions.

Comprehensive Integration of Single-Cell Data (Seurat v3) [Stuart and Satija, Cell, 2019]

- Data preprocessing and normalization: (1) Log. normalization, then normalize gene expression per gene across all cells. (2) Variance stabilization: learn mean and variance relationship across all genes - fit a quadratic curve. The variance of a gene is then predicted from the curve based on its mean expression. Then do z-score transformation, and estimate the variance - this represents over-dispersion not explained by mean-dependent variance. Use this to rank genes and choose top 2000 genes.

- CCA: let X and Y be expression data of two datasets (normalized with mean 0 and s.d. 1). Find the vectors u and v s.t. Xu and Yv maximally correlated.

$$\max_{u,v} u^T X^T Y v \text{ s.t. } \|u\| \leq 1, \|v\| \leq 1 \tag{3.9}$$

  Note: in typical CCA, views are paired, e.g. image and text. In this case, we note that features are "paired", i.e. the same gene is measured in two studies. So we treat each gene as a sample, and its two measurements are two views. So the matrix X is $p \times m$ and Y is $p \times n$, where p is the number of genes, and m, n number of cells. CCA thus projects genes into low dimensions, and try to correlate them in two views.

- Remark: The standard CCA requires paired samples with both types of data. This CCA version allows one to use data from separate studies.

- Projection: after CCA, cells of reference and query data can be projected to the same space. Q: Does the method allow dataset-specific factors?
- Defining anchors by Mutual Nearest Neighbors (MNN): for each cell in one dataset, we find its neighbors in the other dataset. A cell A in the reference and B in the query dataset are a pair of anchor cells if: B is in the neighbors of A in the query data, and A is the neighbors of B in the reference data.
- Anchoring scoring: for a pair of anchors A and B, we can define the neighborhood of each cell in both reference and query, and assess the overlap of four neighborhoods.
- Batch correction: once we match anchors, we can compare expression difference of matched cells. The difference represents batch effect. So we can apply the correction vector to the data.
- Application to multiple scRNA-seq data: also to make the data more noisy, remove one cell type in each dataset. Figure 2: without the strategy, data are clustered by batches/platforms. Also show anchor scores help.
- Anchor weighting: given a query cell, we can define the strength of association of the cell c to an anchor in its neighborhood in the query space a_i. It is basically the distance in the factor space, then do weighting (by anchor quality/scores) and normalization (by the number of anchors in the neighborhood). Results are W_{ci} matrix, anchors by cells: its sum to 1 for all anchors.
- Label transfer: suppose we have labels of all anchors in the reference cell. Given a query cell, we obtain all its neighboring anchors, and the predicted label is just the weighted prediction of all anchors. Let L be the class label of all anchors, we have then predicted labels of all query cells are LW^T, where the prediction scores sum to one for each cell.
- Application of label transfer strategy (Figure 3B-D): show that the strategy works better than other methods that classify a cell only based on its measurement, but not use neighbor information. Note: dim. reduction could do CCA; or do PCA on reference, then projection of query to the reference PC space.
- Application to scRNA-seq and scATAC-seq integration (Figure 3EF): use full-length scRNA-seq data. (1) Pre-processing scATAC-seq data: define gene activity scores from scATAC-seq (Cicero). (2) Results: advantages in identifying more refined clusters in scATAC-seq, assigned to 17 clusters, and some inhibitory neurons into 4 subtypes by using RNA-seq information. (3) Pseudo-bulk ATAC-seq profile: validation with known profiles followed by motif analysis.
- Discussion: about scRNA-seq and scATAC-seq matching, sometimes a cluster in one modality is missing in another. Could be due to data quality issues or method artifact. Also gene activity scores may not work well for genes not regulated by chromatin.
- Lesson: use mean-variance relationship to stabilize variance, i.e. similar to a hierarchical model, use mean to obtain the prior of variance for each gene.

MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data [Argelaguet and Stegle, bioRxiv, 2019]

- MOFA model notations: suppose we have multi-omics bulk or single cell data. We have M views $m, 1 \leq m \leq M$, e.g. one view may be gene expression, another DNAm. For each view, we have a set of features $d, 1 \leq d \leq D_m$. We also have factors k, $1 \leq k \leq K$.
- MOFA model idea: same factors (shared) that have effects on features of different views (RNA-seq or peaks). What is different across views: the effect of a factor on a feature may be different - we allow a factor to have different effect/weight distributions in different views. How does this help with factor analysis? Ex. two samples have similar factors, then we expect their data across multiple views are all similar. So the model tries to do joint dimensionality reduction across all views. However, the model does not directly model how views (e.g. DNAm and mRNA) are related to each other.

- MOFA model: for a given view m, the data model $Y_m = ZW_m^T + \epsilon_m$, where Z is $N \times K$ factor matrix, and W_m is the $D_m \times K$ weight matrix. The prior model of Z and w_{kd}^m:

$$Z_{nk} \sim N(0,1) \qquad w_{kd}^m \sim (1-\theta_k^m)\delta_0 + \theta_k^m \left(0, \frac{1}{\alpha_k^m}\right) \tag{3.10}$$

  where we use spike-and-slab prior for the factor effects and ARD prior for the effect size. This can be rewritten as the product of normal and Bernoulli: let s_{nk} be the indicator variable, we have:

$$(\hat{w}_{kd}^m, s_{kd}^m) \sim N(\hat{w}_{kd}^m|0, 1/\alpha_k^m) \cdot \text{Ber}(s_{kd}^m|\theta_k^m) \tag{3.11}$$

  We have hierarchical prior for α_k^m and θ_k^m, Gamma and Beta prior respectively.

- Remark: the importance of a factor varies with different views, by using different priors/sparsity parameters, θ_k^m. Within features of a view, the prior is shared. See Figure 1 in Supplement. Ex. a factor may explain large variation in DNAm, but less in RNA.

- Error model: see MOFA paper [Argelaguet, MSB, 2018]. For counts, let y_{nd} be the count of gene d in cell n, model it as:

$$y_{nd} \sim \text{Pois}(\lambda(z_n \cdot w_d)) \tag{3.12}$$

  where $\lambda(x) = \log(1 + e^x)$. With this model, when x is small, $\lambda(x)$ is close to 0. When $x > 1$, say, $\lambda(x) \approx x$. However, this error model is sufficient to capture variation in data: once Z and W are given, the expected rate is given - there is no variation/errors for each gene.

- MOFA+ model: we additionally have different groups of cells. We assume a factor is sparse, and may be active in only a certain group of cells. Ex. mouse E3.5 vs. E7.5, in two conditions (cell groups), factor 1 activity, which represents differentiation (ectoderm vs. endoderm), differs. This is captured by modeling Z_{nk}, factor activity in sample/cell n, with different prior distributions for samples/cells n in different groups. Let g be the group sample n belongs to. We have similar prior for the factors:

$$(z_{nk}^g, s_{nk}^g) \sim N\left(z_{nk}^g|0, \frac{1}{\alpha_k^g}\right) \cdot \text{Ber}(s_{nk}^g|\theta_k^g) \tag{3.13}$$

- Remark: despite that the prior distribution of factors are different, all the samples are still projected on the same space. Ex. two set of samples/cells with very different factors are likely to belong to different clusters.

- Application to single-cell RNA-seq data across multiple groups: scRNA-seq of mouse brain in three time points (10x data). Learn factor activity on each time points (Figure 2a,d). Assigning cell types: show factor activity change over time, matches cell type changes during development (Figure 2c).

- Application to single cell DNA methylation data: views are brain regions $\times$ three CpG locations (promoters, gene body, other), and cell groups are defined by brain regions.

- Lesson: even for the same dataset, we can consider them as different views, if the factors are expected to act differently on different views. Ex. DNA methylation data: promoters, gene body and enhancers.

- Remark: the model learns how different views are related to each other by learning common factors acting on multiple views in the same samples. This is different from the CCA approach of Seurat v3, where the data could come from different cells.

High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell (SNARE-seq) [Shen and Kun Zhang, NBT, 2019]

- Method: single nuclei (Fig. 1), droplet based.

- Mouse cortex data: 5K nuclei, with 350 UMIs in RNA and 2.5K UMIs in ATAC.

- Remark: number of UMIs per cell is lower in RNA vs. DNA. Why RNA-based clustering still helps?

3.4.2 Single Molecule Epigenomics

Single-molecule regulatory architectures captured by chromatin fiber sequencing (Fiber-seq) [John Stam, Science, 2020]

- m6A-MTase (M6A-IP-seq): Use DNA methyltransferase to label Adenosine in accessible chromatin. Then use IP to pull down m6A sites, and do sequencing. Show that the results agree with DNase-seq.
- Fiber-seq (Figure 2A): do enzyme treatment and PCR-free library construction, and do PacBio sequencing. The methylated As can be read directly from PacBio. So the results are single molecule chromatin fiber with chromatin accessibility information. Median length: 10Kb.
- Results of Fiber-seq: Figure 2E for an example, DNA accessible regions, dense m6A. Also periodic methylation in inter-linked regions between nucleosomes.
- Neighboring DHS: usually all or none fashion at a single DHS, and co-activation is common for adjacent DHS.
- Nucleosome positioning: primarily determined by DHS boundaries instead of DNA sequence preference.
- TF occupancy: could detect individual TFBS (e.g. CTCF, Figure 5) from the methylation data.

3.5 Regulatory Sequence Analysis

Methods for studying function of cis-regulatory sequences:

- Expression of CRS: in reporter constructs. Luciferase reporter assays in cultured cells; transgenic assays in mouse or zebra-fish/medaka for mammalian enhancers.
- Binding of transcription factors or other co-factors with CRS: ChIP-chip or ChIP-seq.
- Nucleosome binding of CRSs: nucleosome depletion is a sign of function.
- Chromatin modification: epigenetic signatures for function (histon acetylation/methylation).
- Evolutionary pattern of sequences across species: conservation and positive selection.

Regulatory sequence prediction:

- Motif based approach:
  - Training of the recognition model: from a given set of related sequences (co-regulated genes, the targets of the same TFs, etc.), learn the model that represents the cis-regulatory features, including the motifs involved and how they are organized in a CRM.
  - Application of the recognition model: use the model, e.g. a set of motifs, to predict the locations of CRMs in new sequences.

  The first step may be skipped, if the relevant motifs are known.
- Recognition model training: this is often equivalent to finding a set of overrepresented motifs in the input sequences.
  - Overrepresentation test: test the statistical significance of the number of occurrences of the motif(s) in the input sequences. CLOVER, CREME, ModuleSearcher.
  - Probabilistic modeling: the relevant motif should maximize the probability of generating the input sequences, assuming some process of sequence generation. cis-Module, EMC-Module, Gibbs Module Sampler, PhyloGibbs-MP.

  - Discriminative/regression approach: the relevant motif(s) should allow one to discriminate the sequences (with the background) or relate the sequence contents to properties. DiRE, DISCOVER, Lever, LRA.

- Recognition model application:
  - TFBS clustering: test the significance of the TFBS clusters. cis-Analyst, MCAST, MSCAN, Target Explorer, ModuleScanner.
  - Probabilistic modeling: maximize the probability of the target sequence. Ahab, Cister, Cluster-Buster, COMET, PhylCRM.
  - Discriminative/regression approach: apply the trained discriminative/regression model.

- Comparative genomics approach:
  - Conservation: Gumpy, rVISTA
  - Accelerated evolution: Ha-CNS

Remark:

- Probabilistic vs regression approach: the regression approach possesses a few advantages, including the easiness to add new features and less dependence on assumptions (e.g. about background sequence model). Therefore, with more data available and more important features recognized, the regression approach may be more promising.

- Motif detection/testing - the role of control: to claim a motif is overrepresented in some set of sequences, need to have appropriate control. Ex. to find the motif of a TF using the ChIP data of the TF, we consider different controls:
  - Random un-bound sequences from the genome: insufficient, as the bound sequences may be enriched with promoters, which tend to have different sequence characteristics from other genomic sequences.
  - Randomly permutated bound sequences: insufficient, as the permuted sequences destroy the general promoter signals, thus the TF-specific motif may be missing.
  - Comparison of strong vs weak bound sequences: may still be biased, if the two groups of sequences have different fraction of promoters.

  The general strategy is: identify possible confounding variables, and choose the background sequences that have the same (statistically) distribution of confounding variables as the foreground sequences. In sequence analysis: GC content, repeats, and sequence types (promoters, UTRs, etc.) are important confounding variables.

Reference:

- Motif-based approach: [Papatsenko & Levine, Nature Methods, 2005], [GuhaThakurta, NAR, 2006], [Hannenhalli, Bioinfo, 2008], [Tan & Ravasi, Genomics, 2008], [Van Loo & Marynen, Briefings in Bioinfo, 2009]

- Comparative genomics: [Visel & Pennacchio, Semin. Cell Dev. Biol., 2007]

3.5.1 Motifs and Binding Profiles

Sequence logos [Schneider & Stephens, NAR, 1990]:

- Aim: display the specificity/consensus sequence of a motif.
- Methods:
  - Height of a position: the information contenct, defined as:

$$R = 2 - H = 2 + \sum_b f(b) \log_2 f(b) \tag{3.14}$$

    where H is entropy, b is a base and $f(b)$ is the frequency of b.
  - Height of a nucleotide: proportional to frequency

$$h(b) = f(b) \cdot R \tag{3.15}$$

JASPAR [Sandelin & Lenhard, NAR, 2004]:

- Curation: a PWM is chosen by several criteria, including (i) experiment must have meet the standard; (ii) at least 5 sites; (iii) non-redundent (no more than one profile for a particular factor).
- Results: 111 profiles (as of 2004), currently about 130 profiles.
- Tools: the computational framework called TFBS.

Context-specific independence mixture modeling for positional weight matrices [ISMB, 2006]:

- CSI model: this notion of adapting model complexity to the data is known as context-specific independence (CSI). We present an extension of the conventional mixture framework by learning an explicit dependency structure between the components of a PWM mixture. The basic ideais to reduce the number of parameters by representing binding site positions with little variability in the different components by the same distribution.
- The advantage of the CSI model: in a conventional mixture random sequence deviations will cause the parameters in the different components for the same position to vary slightly, even if there is no meaningful variability on the sequence level. Therefore, learning a CSI structure does not only yield a more parsimonious model, as less parameters are required, but also increases robustness for noisy data.
- CSI model: Figure 2. a) Model structure for a conventional mixture with 5 components and four RV. Each cell of the matrix represents a distribution in the mixture and every RV has an unique distribution in each component. b) CSI model structure. Multiple components may share the same distribution for a RV as indicated by the matrix cells spanning multiple rows. In example C2 ,C3 and C4 share the same distribution for X2.

Quantifying similarity between motifs (TOMTOM) [Gupta & Noble, GB, 2007]

- Measuring motif similarlity: the sum of column-wise similarity function. Supported functions are: Pearson correlation coefficient (default), log-likelihood ratio, KL divergence, etc.
- Statistical significance of results: the main contribution is to derive the null distribution of the score function, and compute its E-value, defined as p-value multiplied by the size of the target database, the expected number of times that the given query would be expected to match a target as well or better than the observed match in a randomized target database of the given size.

Drosophila AP factor motifs by B1H [Noyes & Wolfe, NAR, 2008]:

- Aim: characterize the binding specifities of AP factors.
- Methods:
  - Factors: 35 different AP factors, which represent nine different DNA-binding domain families, including zinc fingers, homeodomains, bHLH, bZIP and winged helix, etc.
  - B1H technique: can choose selection stringencies (5 and 10 mM 3-AT), higher stringency selection yielded a more constrained motif, due to a greater enrichment of the highest affinity sites.
- Results:
  - B1H motif specificities are in most cases consistent with previously determined specificity data, where available. Ex. Bcd, Kr, Tll.
  - B1H motifs of Cad, Gt and Kni are much more specific than FlyReg ones.
  - Validation: for a given TF, the correlation of motif counts (of all CRMs that are expressed in a position) and expression of the TF at that position (similar to [Schroeder04]).
- Remark: the motifs are generated from a small number of selected binding sites (20 to 30 on average) and are typically collected at a single stringency, thus may not detect low-affnity BSs.

Measuring PWM similarity [Pape & Vingron, Bioinfo, 2008]:

- Problem: given two PWMs, measure its similarity. The earlier methods use KL-divergence between the 2 columns and take some kind of summary (mean, max, etc.) over KL of all columns
- Methods: natural measure. If PWMs A and B are similar, then the hits of A and B sites should be correlated.

A Feature-Based Approach to Modeling Protein-DNA Interactions [PLCB, 2008]:

- Review of existing methods: several models were developed to capture such dependencies.
  - In the first class, the dependencies between neighboring positions are modeled using a Markov model of some order. A recent representative of this class is the permutated variable length Markov model (PVLMM) of Zhao et al., which incorporates two major improvements: it searches for the best permutation of the motif positions, and it reduces the number of parameters by using a context tree representation for the Markov model representation.
  - The second class of models was proposed by Barash et al., who developed a Bayesian network approach to represent higher order dependencies between motif positions. Ben-Gal et al. extended this approach by using a context dependent representation of the conditional probability distributions. However, the Bayesian network representation, due to its acyclicity constraints, imposes nonnatural restrictions on the motif structure, and its conditional probability distributions limit the number of dependencies that can be introduced between positions in practice.
  - Another class of TF binding specificities models that is complementary to the above two is a mixture of models.
- Model: Our approach is based on describing the set of sequence properties, or features, that are relevant to the TFDNA interactions. Features may be binary (e.g., C at position 2, and G at position 3) or multivalued (e.g., the number of G or C nucleotides at positions 14), and global features are also allowed (e.g., the sequence is palindromic).
  - A representation of Markov networks, which is often referred to as log-linear models [52], is a natural framework for compact representation of a distribution as a set of feature functions.
  - Given a dataset of N aligned i.i.d TFBSs, our aim is to optimize the data likelihood over all possible models.

- Model learning:
  - We reduce the feature space using a Binomial test to evaluate the statistical significance of features that span more than one position
  - In order to control for model complexity and to achieve a sparse representation, we use the L1-Regularization suggested by Lee et al. [55], which penalizes models linearly by their sum of weights.
- De novo motif finding:
  - In the first, we extract all sequences of length K (referred to as K-mers) and greedily grow motifs (defined by a set of OR and AND operations on a set of K-mers) that are discriminatively enriched in the positive set over the negative set.
  - In the second step, each enriched KMM is used for extracting aligned TFBSs from the positive set, from which a motif model, FMM or PSSM, is learned.

3.5.2 Motif Discovery

CREME [Sharan & Karp, Bioinfo, 2003]:

- Problem: a set of putative regulatory sequences, find the enriched motif sets (from a candidate list of all motifs).
- Outline: statistical score of the number of occurrences of a motif cluster (a set of distinct of motifs), itemize all motif clusters. Evaluation by the expression of genes containing the significant motif clusters (subset of the input genes).
- Methods:
  - Sequence preparation: only conserved sequences in human-mouse alignment.
  - Testing motifs: the initial set of motifs will be filtered to choose significant motifs (individually). The count of motif occurrence is tested by assuming Poisson distribution; or randomly sample the set number of promoters, and count the motif occurrences. The p-value cutoff is 0.001.
  - Motif cluster counts: the number of sequence segments where each motif occurs at least once.
  - Testing motif clusters: a motif cluster is significant if its count is significantly more than expected by the counts of individual motifs. Enumerate motif clusters, and for each cluster, do permutation test: randomly permute the "labels" of motif hits in the sequences and count the motif cluster occurrences. Multiple hypothesis correction by Bonferroni with cutoff p value 0.05 (after correction).
- Results:
  - Application to human cell cycle genes (to all 874 cell-cyle related genes, only 376 have conserved promoters): several significant motif clusters are found, and genes containing the same motif cluster have more coherent expression than the group overall.
- Remark: one often assumes that the similarly regulated genes share the same cis-regulatory mechanism. But in fact, there may be multiple such mechanisms for the same set of genes. This need to be taken into account, especially if the input gene set is large, e.g. when applied to all genes up-regulated in a tissue.

CLOVER [Frith & Weng, NAR, 2004]:

- Problem: given a set of related sequences (which are potentially regulatory), test if a motif is significantly overrepresented.

- Methods:
  - Scoring motif presence in a set of sequences: (i) scoring a single sequence: calculate the average LLR score of all predicted sites in the sequence; (ii) scoring the sequence set: assume i sequences are related to the motif, compute the product of the sequence scores; and average over all values of i.
  - Evaluating statistical significance: four ways of creating the null distribution and calculate the p value: nucleotide shuffling; collect di-nucleotide frequencies and generate random sequences; shuffling of PWM and compute the score of random PWM on real sequences; choose background sequences from genomes and randomly sample length-matched sequences. In each set of null distribution, randomize 1000 times.
  - Multiple testing: use p value 0.01 as cutoff. Claim that multiple testing is not an issue: some motifs have p value 0 (among 1000 randomizations), and some motifs are significant among all backgrounds.
- Results:
  - Comparison of two simple methods based on contingency tables: (i) motif counting: the total number of motif counts in positive vs background sets; (ii) sequence counting: the number of sequences containing the motif in positive vs background sets.
  - Testing in Drsophila segmentation CRMs: choose 19 CRMs and predict the overrepresented motifs including Kr, Hb, etc.

Promoter motifs in Drosophila [Down & Hubbard, PLoS CB, 2007]:

- Aim: discover motifs in Drosophila promoter sequences.
- Methods:
  - NestedMICA: a program for motif discovery. Features: infer multiple motifs simultaneously; use Nested Sampling approach for motif search; a family of background models (instead of a single MC), where each base is generated by one of several possible Markov chains.
  - Procedure: 200 bp sequences flanking the 5' of genes in arm 2L. A total of 422 kb sequences from 2,424 genes. Each motif starts with fixed length 12, postprocess to remove uninformative columns.
  - Known motifs: (i) 10 motifs from [Ohler & Rubin, GB, 2002]; (ii) FlyReg: 30 motifs; (iii) SELEX and primary publications (extended JASPAR): 172 motifs out of which 62 are from Drosophila TFs.
  - Gene expression: in situ data. Each gene is annotated by ImaGO terms.
  - Comparison of PWMs: (i) define divergence between two distributions (Eq. 3), and the distance between two motifs is defined by the sum of distance between corresponding positions. For two motifs of different length, find the lowerest score among all possible alignments. (ii) significance of the query motif and target (the best match in a motif set): random shuffle of the query motif and the obtain the score (best score with all motifs in the set) - empirical p values.
- Results:
  - Motif dictionary from Drosophila promoters: 120 motifs. Removing redundency (and known motifs) givs 87 novel motifs.
  - Comparison with known motifs: 8 matches to Ohler motifs; 7 to FlyReg; 14 to extended JASPAR. Total of 25 matches with known PWMs.

  - Conservation of motifs: a motif is conserved if there is a correlation between motif scores and the conservation (fraction of sites that are conserved in Dmel, Dsim and Dyak). 78 out of 120 motifs show significant correlations.
  - Association of motifs with gene expression pattern: measured as the number of times each motif is associated with an ImaGO term (association: if the motif appears in the 200 bp promoter sequence of a gene with that term). Test the significance (empirical p values) by permuation of motif labels. Found 25 motifs show significant association with expression patterns.

- Remark: the majority of known developmental TFs are not found, suggesting that they do not bind preferentially to the 200 bp upstream of the target genes.

Discovering gapped motifs [Chen & Li, PNAS, 2008]:

- Problem: discovering gapped motifs: motifs with several degenerate positions.
- Methods:
  - Idea: favor motifs with three additional patterns in the sequences (i) found in promoters with high binding intensity (instead of bound vs background); (ii) have positions that are repeatedly matched by similar patterns; (iii) evolutionarily conserved.
  - : Scoring motifs: S_d - the score from enrichment in positive vs negative sequences; S_p - the score from frequent occurrence in highly ranked sequences; S_c - the score from conservation. The final score is defined as: $S = S_d(S_p)^a(S_c)^b$, where a and b control the importance of positional and conservation scores.
- Results:
  - Including positional scores is helpful for gapped motifs; conservation scores helpful for ungapped motifs.

Amadeus [Linhart & Shamir, GR, 2008]:

- Methods:
  - Evaluation data: 42 gene sets from human, mouse, fly and worm (26 TFs and 8 miRNAs): 25-2338 genes per set. Mostly from ChIP-chip or ChIP-seq data.
  - Method: (1) count k-mers in positive and reference promoters (chosen to be all promoters of that species); (2) hypergeometric (HG) or binned HG test (controlling for GC content and length); (3) combine significant k-mers to PWMs.
- Remark: use the reference promoters (instead of Markov models for the background sequences) is important; and use binned test further improves the performance.

DREME: motif discovery in transcription factor ChIP-seq data [Baily, Bioinfo, 2011]

- Motif finding: Regular expression (RE) as representation of motif. Exhaustively search all RE motifs: for each motif, count its presence in the positive and negative sets (only count the number of sequences containing a motif, not the number of motif occurrences, because of the issue of self-overlapping motif), and perform Fisher's exact test to determine significance.
- From the results, all the motif instances will be used to construct a PWM.

3.5.3 Predicting Enhancers from Sequences

Cister, COMET and Cluster-Buster [Frith & Weng, Bioinfo, 2001; NAR, 2002; NAR, 2003]:

- Problem: a set of PWMs, scan for the binding site clusters.
- Methods:
  - All three methods employ the same HMM for sequences (background states and motif states), and the score of a sequence segment is roughly the LLR score. The difference lies in the approximation of the score, and the computational approximation for locating the cluster (within a larger sequence, without using a fixed-length window).
  - Cister: does not directly predict motif clusters, but returns a probability curve indicating the probability that each basepair in the sequence lies within a cluster, using the linear-time Forward-Backward algorithm.
  - COMET: not the full log-likelihood ratio, just the Veterbi algorithm (i.e. the most likely arrangement of motifs within the subsequence). An advantage of Comet is that it calculates E-values to indicate the statistical significance of its predictions.
  - Cluster-Buster: employing a linear-time heuristic which attempts to return the same cluster predictions as the full quadratic-time algorithm.

MCAST [Bailey & Nobel, Bioinfo, 2003]:

- Problem: given a set of motifs, scan for TFBS clusters.
- Methods:
  - Model: similar to COMET, using HMM to score a putative module: the hits, the gap penalty. The difference: (i) Non-conventional HMM, s.t. it only emits statistically significant motif hits; (ii) a fixed background model.

MSCAN [Johansson & Lagergren, Bioinfo, 2003]:

- Problem: scan a sequence for clusters of TFBSs, from a set of relevant PWMs.
- Methods:
  - For a sequence window (fixed length): find the optimal k-hits, i.e. k non-overlapping sites with dynamic programming. Then among all values of k, choose one with the lowest p value.
- Results: performance similar to Cister and COMET in muscle and liver data.

Target Explorer [Sosinsky & Honig, NAR, 2003]:

- Problem: scan a sequence for clusters of TFBSs, from a set of PWMs.
- Methods:
  - Scoring a cluster: The score for an entire cluster is calculated as a sum of individual scores for the minimal required number of sites for each TF.
  - Cluster cutoff: The cut-off score for an entire cluster is taken as the sum of cut-off scores for individual sites.

Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects [Noto & Craven, Bioinfo, 2006]:

- Problem: given two sets of sequences, one set CRM and the other not. Classify a sequence, utilizing the CRM structure: the rules of motif interactions.

- Model: design HMM that represents the logical and spatial rules of CRM. Use one state for one motif and connectivity between motifs represent rules. For example:
  - AND: use both motifs are used
  - OR: choose either state in the path
  - Strand preference: emit either a motif or its rev. complement with different probabilities
  - Order: motif states appear in a certain order
- Inference: learn both structure and parameter simulaneously
  - Structure: at each step, each operator (AND, OR, etc.) is applied to the best solution and estimate parameters for each new structure. The highest-scoing one will be chosen for the next round.
  - Parameter: Discriminative learning approach for HMM [Krogh, IEEE, 1994]
- Results: comparison with [Segal & Sharan, RECOMB, 2004]
  - Yeast ChIP-chip data: classify a promoter sequence (positive if bound).

PReMod [Blanchette & Robert, GR, 2006]:

- Aim: identify the CRMs in the human genome, using motifs in Transfac
- Methods:
  - Motifs: 481 Transface PWMs
  - Expression data: 79 tissues from Gene Atlas.
  - Identification of putative TFBSs: matches in human genome (regions with mouse and rat alignment), then matches conserved in mouse and rat will receive additional scores.
  - Score a module: choose 5 TFs that give the highest scores, then assign a P value on the total module score.
- Results:
  - In silico validation of predicted CRMs (pCRMs): overlap between pCRMs and known CRMs, known DNaseI hypersensitive sites.
  - Experimental validation by ChIP-chip of ER and E2F: 3% of predicted ER pCRMs and 17% of predicted E2F4 pCRMs are bound in ChIP-chip results.
  - Genes that are the targets of the same TFs (according to pCRMs) are more likely to share expression pattern: 27 out of 229 TFs show significant correlation; 595 out of 26,016 possible motif pairs show significant correlation.

DISCOVER [Fu & Xing, Bioinfo, 2009]:

- Problem: given a set of sequences (probably long), and a set of PWMs, predict the locations of CRMs and TFBSs.
- Methods:
  - CRF model: let x_i represent the i-th nucleotide, and y_i be the hidden state of the i-th nucleotide. The state could be background, CRM or one type of TFBS. The goal is to predict y from x:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \exp\left[\lambda \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})\right] \tag{3.16}$$

where $\mathbf{F}$ is the feature set and λ is the parameters (weights of features). The feature value is defined at each position i, typically, the sequence content of this nucleotide and its adjacent sequences (such as PWM match), the property of this neighborhood (whether a nulecosome is bound here), etc.

  - Feature set: PWM matches; state transition (density of TFBSs, expected length of CRMs); conservation - PhastCons score; nucleosome occupancy; repeats (could be used to filter out repeats: suggests the background state); distance to TSS; etc.
  - Model application: train the model paramaters in annotated data (CRMs and TFBSs labelled in a set of sequences) and apply to new sequences.

- Results:
  - Outperform Cister, Cluster-Buster, Ahab, MSCAN and Stubb: on a set of annotated Drosophila TFBSs, evaluated with LOOCV.

Chromia: CHROMatine based Integrated Approach [Won & Wang, GB, 2010]:

- Goal: prediction of TF binding sites/regions with motif matches and histone modification patterns.
- Methods:
  - Data: ChIP-seq data of 13 TFs and 8 chromatin marks in mouse ES cells.
  - Model:
    * Ideas: (1) within TFBSs, higher motif scores, and higher histone scores. (2) Histone scores should reflect the pattern of histone modification: bimodal distribution around the peaks of TF binding.
    * Binding region prediction: score every 100 bp bin. For every bin, do a LRT where H_0 is the background state, a single-state HMM and H_A is either the promoter bound region or enhancer bound region, represented by a 3-state HMM (the first and third for the bimodal pattern and the second the dip). At each state, emit the motif score and histone modification intensity. For emission probabilities, use a mixture of M distributions (different levels of motifs/histone modification within TFBSs).
- Results:
  - Different TFs have different preferences: some prefer promoters (e.g. E2F1, c-Myc) and some prefer enhancers (e.g. Oct4, Nanog) and others have no preferences.
- Remark: generative model not for sequence itself, but for the properties of sequences: TF matching, histone modification, etc. Modeling sequence properties would be an important paradigm for sequence analysis with more functional genomic data. In Chromia, if not model the bimodal histone pattern, it would be a simple Naive Bayes model: generate features from labels (though a mixture model is used to allow some variations of feature distribution).

3.5.4 Regulatory Analysis of Groups of Genes/Sequences

ModuleScanner and ModuleSearcher [Aerts & De Moor, Bioinfo, 2003]:

- Problem: given a set of sequences and PWMs, search for the overrepresented motifs (CRM model) and locate the regulatory sequences.
- Methods:

- Module scanning: the score of a sequence wrt. a CRM model (a set of motifs) is defined as the product of the scores (LLR) of all putative binding sites, with penalty for motifs not occurred in the sequence.
- Module searching: search for best CRM model among all possible ones using $A*$ tree.
- Sequence preparation: only conserved blocks in upstream 10k in human-mouse alignments.

ModuleMiner [Van Loo & Marynen, GB, 2008]:

- Aim: discover CRMs and relevant factors in a set of co-regulated genes.
- Background: ModuleSearcher [Aerts & De Moor, Bioinfo, 2004], search for the best combination of motifs in all windows of a set of co-regulated genes. The motif set is evaluated by the best window, where a window is scored as the total of LLR scores of all motif matches.
- Methods:
  - TRM is defined as a set of motifs (may include multiple instances of one PWM, maximum 6) and several parameters: the length of CRM, Boolean parameters of whether motifs can overlap, etc. To score a TRM wrt. a set of co-regulated genes:
    - Each gene is assigned a score according to its match to the TRM: one PWM in the TRM can occur only once, and the score is the maximum score of all CNSs (conserved non-coding sequences).
    - The "fitness" of a TRM is defined as: score all genes in the genome, and the ranks of the co-regulated genes. Use order statistics to assign a probability to the combination of the ranks of all co-regulated genes.
  - TRM search: use a genetic algorithm to find the TRM with the highest fitness.
  - TFBS sets: 3 sets are considered: (i) PWM matches in human-mouse conserved regions, at least 75% PID over a min. of 100 bps; (ii) (i) + binding sites occurring in both human and mouse; (iii) (ii) + allowing possible mouse TSS differences.
- Results:
  - Validation of ModuleMiner with 12 muscle enhancers: LOOCV.
  - Detection of CRMs in microarray clusters: Gene Atlas data of 140 human and mouse tissues. CRMs are often close to TSSs.
  - Detection of CRMs in embryonic developmental gene sets: 5 data sets. CRMs are depleted in TSS regions.
- Remark/Criticims:
  - TRM model is restrictive: only allow six PWM instances (may have duplicates), thus at most six sites can be used.
  - Use of conservation information: very stringent critieria of CNS; no mechanism of using information in multiple species, especially at BS level.
  - TRM search: expensive, need to search a large space.

PhyloGibbs-MP [Siddharthan, PLoS CB, 2008]:

- Methods:
  - Features of PhyloGibbs-MP (vs PhyloGibbs): module prediction, prior motifs, discriminative motif finding.

  - Motif/TFBS finding experiment: recovery of known TFBSs. Test the performance by AUC. Compare PhyloGibbs-MP-n8 (at most 8 colors), PhyloGibbs-MP-n3 (at most 3 colors) and other programs.
    * Yeast: 466 known sites in SCPD in 200 genes. Input: upstream 1K of 200 genes.
    * Fly: REDFly binding sites. The specific factors chosen from the TRANSFAC database were Abd-B, Adf1, Cf2, Dfd, Eip74EF, Stat92E, Su(H), Ubx, bcd, dl, ftz, hb, ovo, pan, sna, z. Surrounding sequence was selected, such that the total length of the sequence was 250 bp per binding site. Species: mel, sim, yak, ere.
  - CRM prediction: expand each REDFly CRM with neighbor sequences and get 234 stretches of DNA that were at least 10000 bp long and contained at least one annotated CRM. Prior motifs: 73 motifs in FlyReg. Species: mel, sim, yak, ere.

- Results:
  - TFBS prediction experiment: PhyloGibbs-MP-n8 performs the best. Better than PhyloGibbs because in PhyloGibbs, a motif is required to have at least one site. This becomes a problem when there are many colors. Better than PhyloGibbs-MP-n3 because the data set has many different types of motifs.
  - CRM prediction experiment: (i) Program variables of PhyloGibbs: prior motifs significantly better than no prior motifs; 2 species (mel-yak) is slightly better than 4 species. (ii) Comparison of Stubb (mel-yak), Cluster-Buster, cisModule, EMCmodule vs PhyloGibbs-MP (prior, 2species): similar performance of PhyloGibbs-MP and EMCmodule; at lower spec. level, Stubb has higher sens. than other programs. The results do no strongly support PhyloGibbs-MP.

- Remark:
  - In CRM prediction experiment: both Stubb and Cluster-Buster are run with input = 73 motifs. Not make sense.
  - Why PhyloGibbs-MP with mel-yak better than mel-sim-yak-ere? Possibly due to lineage-specific changes: at higher score cutoff (higher spec.) - 4-species version has higher sensitivity as few lineage-specific changes are to be found in this ragen; at lower score cutoff - 4-species version fails to uncover some sites, thus lower sensitivity (consistent with the pattern found).

Assessment of composite motif discovery methods [Klepper & Drablos, BMC Bioinfo, 2008]:

- Methods:
  - Task: a set of sequences and PWMs, identify the involved motifs and predict module locations.
  - Datasets: 10 different datasets, including muscle and liver sequences.
  - Evaluation: both module location and module composition (which motifs are involved).

- Results:
  - The performance varies a lot across different datasets. The general trend: CisModule performed poorly on most sequence sets, Cister and Stubb usually scored somewhere in the middle, while CMA, ModuleSearcher, MSCAN and Cluster-Buster were often found among the top scoring methods on each set.
  - CMA and ModuleSearcher were clearly best at identifying the correct motif types involved in the modules, and they were also the only methods capable of coping with large and noisy PWM sets. The other PWM-reliant methods appear to be more suited for detecting modules with some prior expected composition than for discovering completely new and uncharacterized modules.

Charactering functions of mouse TF motifs [Jaeger & Bulyk, Genomics, 2010]:

- Methods:
  - Motif data: all 8-mers of 104 mouse TFs in [Badis09].
  - Gene set analysis: mouse tissue-specific gene expression clusters from [Zheng, JBiol04] and GO annotation terms. A total of 1371 gene sets.
- Results:
  - Motif enrichment in conserved developmental enhancers: search for 8-mers (enrichment) in tissue-specific enhancers from [Pennacchio, Nature06] and [Visel, NG08]. Validate the enriched motifs using TF expression data from Allen Brain Atlas (slightly different developmental stages than the enhancer data). Ex. eye - Nr2f2, Six6, Gata3; midbrain - E2f2, Sox4, Rfx3, Zic3.
  - Conservation of motifs: the PMW bound 8-mers tend to have lower substitution rates in upstream/downstream 10k bp regions.
  - Motif enrichment in mouse tissue-specific expression clusters or GO gene sets: found 285 significant pairings using PhylCRM-Lever. Some have literature support, e.g. Sp4 in brain; Tcf3 in AP axis patterning. Cannot clearly distinguish similar motifs: e.g. Myf6 (muscle) motif is associated with neuron development, probably due to neuronally expressed TFs with similar specifities such as Neurog1/2.

3.5.5 Sequence Properties of Enhancers

Binding site arrangement in fly devlopmental CRMs [Makeev & Papatsenko, NAR, 2003]:

- Problem: preference of distance between adjacent binding sites?
- Background: many known examples where distance and/or orientation of BSs affect the regulatory function of promoter and enhancer sequences.
- Methods:
  - Data: fly CRMs of 5 TFs: Bcd, Cad, Hb, Kr and Kni.
  - Non-randomness test: the inter-site distance is not random, tested by the expected number of inter-site distances assuming the binding site distribution is random (Bernoulli trial at each position).
  - Periodicity testing: Fourier analysis of distance distribution - the frequency domain, find the strongest period
- Results:
  - Site distribution is not random: the fraction of sites having spacing in the range 50-60 bps is larger than expected.
  - Periodic signals in arrangement of a single motif: helical spacing in Bcd - 10 bp; Hb - 11 bp and Cad. No signal for Kr. Number of sites of Kni is too small for analysis.
  - Periodic signal in motif combinations: no signal for Bcd and Hb; Bcd-Kr pair, 17 bp (opposite side of DNA helix).
- Conclusion: Composite elements (binding site pairs in a certain arrangement) may be important structural units for constructing more complex regulatory sequences.

Spatially conserved word pairs [Chiang & Eisen, GB, 2003]:

- Idea: syn. TFBS pairs will be both spatially closed, and conserved in multiple species

- Methods:
  - Identify the conserved word pairs: a word is defined as a hexamer
    - Find the jointed conserved word pairs: the words pairs whose conservation is more than the independent conservation of each word itself (chi-square test)
    - Closely spaced word pairs: permutation test for close spacing. Address the issues like sequence composition variation in the genome (not use Poisson distribution)
  - Association of word pairs with expression change: genes containing a word pair identified should have closer expression change in an experiment than overall gene set. Test the distribution (hisogram) of expression of the word-pair associated gene set vs all gene set via KS test
- Results:
  - Among 2.16m hexamer pairs, find 8,452 jointly conserved word pairs in 4 yeast species
  - Close spacing test: find 989 pairs, called word templates
  - 314 word-pairs are selected by the association with expression change.

Rule-based learning for CRM [Hvidsten & Fidelis, GR, 2005]:

- Problem: find rules of motif combinations that are characteristic of expression profiles
- Methods: the main procedure consists of:
  - Constructure motif profiles (1 if a motif is present, 0 otherwise) for all genes
  - For each gene, divide all other genes in 2 sets, depending on whether a gene has similar expression to this gene. Learn Boolean rules that discriminate the two sets.
  - Rule filtering: only rules that are general (cover several genes) and accurate (genes with this rule have similar expression) are considered.

Discovering functional transcription-factor combinations in the human cell cycle [Zhu & Church, GR, 2005]:

- Problem: given humand and mouse genomes, and a set of TF profiles, identify the pairs of interacting TFs
- Methods:
  - Find all putative TFBS of some motif (anchor motifs): search PWM matches in the human-mouse conserved region
  - Extract the neighboring region of all putative sites in both human and mouse $\Rightarrow$ motif finding in these regions with AlignACE
  - Selection of motifs that are statistically enriched among all candidates
  - Functional testing: genes with the same pair should have correlated expression
- Results: human homotypic and heterotypic TF pairs

A universal framework for regulatory element discovery across all genomes and data types [Elemento & Tavazoie, Mol Cell, 2007]:

- Problem:
  - Predict motifs assoicated/responsible for gene expression pattern (either clustering or up-/down-regulation). Want to be able to assoicate motif directly with expression, unlike previous approaches which separate the two and look at only statistical patterns.

  - Learn regulatory logic: position bias, orientation bias, motif-motif co-occurrence and motif-motif colocalization (likely interaction)

- Methods:
  - Motif representation and search: regular expression, and search in the space of all regular expression to maximizes the quality measure (assoication with expression profiles). For any motif, call the presence/absence pattern of all genes as its "motif profile".
  - Gene expression profile: clustering $\Rightarrow$ non-overlapping clusters; or obtain expression up/down value
  - Association between a motif and an expression profile: the assoication of motif presence/absence pattern of a gene with its expression profile (can be multiple values), measured by mutual information (MI). The significance of MI is assessed by randomization test.
  - Position bias and orientation bias: mutual information between position/orientation and cluster index
  - Motif interactions and co-occurrence: (i) functional interaction between motifs: the MI of two motif profiles; (ii)co-localization: if there is functional interaction, treat two motifs as a single unit (Boolean AND), and test the association of this joint motif with the expression profile.
  - Expression data: yeast - expression compendium consisting of 173 microarray experiments; fly - gene groups in BDGP in situ data; human - Gene Atlas
  - Sequence data: for each gene, take 1kb upstream TSS and 300 bp downstream.

- Results:
  - Yeast:
    * 78 non-overlapping clusters $\Rightarrow$ 17 DNA motifs, out of which 14 match known ones.
    * Each motif is associated with at least one cluster and many motifs explain/are assoiciated with multiple clusters. Some motif is under-represented in the related cluster.
    * Cases of position bias (Fig S6, S7) and orientation bias (Fig S8) are observed
    * Cases of motif interaction, actually mutual exclusion, are also observed (Fig S10)
    * Comparison with AlignACE: AlignACE has a higher false positive rate (at default parameters) and AlignACE finds motifs for each cluster, thus the motifs are often redundent since one motif may be associated with multiple clusters.
  - Human: 73 predicted DNA motifs, out of which 20 match known ones in Transfac or Jaspar.

- Remark:
  - Limit to only immediate upstream (1kb) and downstream sequences. Too restrictive for higher organisms such as fly and human.
  - Two tests of motif relationships: (i) functional interaction: can be viewed as a way of selecting motifs based on how often they target the same genes; (ii) co-localization: viewed as treating the joint motif as a new feature.

Prediction of synergistic transcription factors by function conservation [Hu & Collins, GB, 2007]:

- Problem: (1) find TFs that have homotypic clustering; and TF-pairs that have heterotypic clustering; (2) apply these learned rule to predict TFBS

- Methods: a true TFBS cluster (either homo. or hetero.) should be enriched in the genome and the cluster tends to be more conserved

- Enrichment of TFBS cluster in single genome: LOD_{co} score, defined as the log-ratio of the frequency of promoters containing this cluster and the promoters not containing. Only the scores that are significant (through comparison with random shuffled sequences) will be considered in the next step.
- Conservation of TFBS cluster: find nubmer of promoters containing the cluster in 2 species respectively, and the ortho. promoters that contain the cluster in both species. Test the significance through hypogeometric test $Rightarrow$ scaled to LOD_{og} score
- For each distance constraint, there is a pair of LOD scores. Correlation of the two LOD vectors.
- Prediction of TFBS cluster: for a promoter containing the TFBS cluster, assess the significance of this cluster given the distance between the two

- Results:
  - Learn 51 homo. TFs from 234 Transfac motifs; out of which 7 match known homo. TFs (total 15 known)
  - Prediction of E2F homo. cluster: higher spec and sens than rVista and EEL.
  - Case study (Fig 6): E2F sites in human and mouse promoters. Cases where in the ortho. promoters, E2F sites are in very different positions (relative to TSS).
- Remark/Criticims: the problems of the approach include:
  - Unable to generalize to multiple species
  - Correlation of LOD scores: intutition that a TFBS cluster should be both enriched and conserved, but the LOD correlation seems not very statistically solid

Binding site orienation and spacing [Shultzaberger & Eisen, PLoS ONE, 2008]:

- Problem: Met4 regulates target genes by employing two TFs, Cbf1 and Met31. In other words, the Met4-dependent activity is determined by the affinities and arrangement of Cbf1 and Met31 binding sites, but how?
- Methods:
  - Multi-site information: consider a multi-site consisting of one Cbf1 (A) and Met31 (B) site, the information score of this multi-site (SI) is defined by:

$$SI = R(A) + R(B) - GS(d) - OS(o) \tag{3.17}$$

    where $R(A)$ and $R(B)$ are the score of two sites respectively measuing the binding affinity, $GS(d)$ measures the contribution from spacing (d) and $OS(o)$ measures the contribution from orientation (o). $GS(d)$ is defined as:

$$GS(d) = -log_2 n(d)/n + e(n) \tag{3.18}$$

    where $n(d)$ is the number of occurrences at spacing d and n is the total number of occurrences over allowed values of d, $e(n)$ is a small sample correction value. And $OS(o)$ is defined as:

$$OS(o) = -log_2 n(o)/n + e(n) \tag{3.19}$$

    Note that only orientation of one TF is considered here.
  - Procedure for learning rules of orientation and spacing: if a rule (e.g. close spacing) is important, then genes with high scores under this (and affinity) rules will have high expression change in microarray experiments. So any rule is evaluated by: rank all genes with this (and affinity) rules, the average expression change of top-ranked genes in Met4-stimulated microarray experiments.

- Results:
  - Orientation: ordering of Cbf1 and Met31 is important, but not orientation.
  - Distance to TSS: within 450 bases of TSS.
  - Optimal spacing between Cbf1 and Met31 is 9 to 68 bases.
  - Met 4 activation model: Cbf1, separted by 9 to 68 bases, Met31 in either orientation, then less than 450 bps to TSS

Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data [Jothi & Zhao, NAR, 2008]:

- Motivation: call binding events from ChIP-seq data.
- Homotypic clustering in binding data: among three TFs examined (CTCF, NRST, STAT1), a good fraction (37-71%) of binding regions contained more than one motif within the 200-bp region using the program MAST.

Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers [Gotea & Ovcharenko, GR, 2010]:

- Hypothesis: HCT is an important feature of human/mammalian genomes, and functionally important.
- Methods:
  - Data: human, mouse, chicken and fugu genome, 701 PWMs from JASPAR and TRANSFAC.
  - HCT prediction: 10K top-scoring HCTs (through a HMM) in the genomes, eliminate HCTs longer than 3kb and more than 25% TFBSs in the coding regions. At least four sites with distance between adjacent sites less than 500 bp. Further restrict to those HCTs found in both human and mouse, inside an ECR (evolutionarily conserved region) - about 10% of all HCT. Most PWMs (208) are represented by more than 500 HCTs.
- Results:
  - Distribution of HCTs: 38.6% are in promoter regions, the rest evenly distributed between intergenic (30.7%) and geneic regions (excluding promoters).
  - Functional evidence of HCTs: significant overlap with ChIP-chip/-seq data of TF binding and p300 regions, with known developmental enhancers, and TFBSs in conserved HCTs evolve under purifying selection.
  - Coverage of HCTs: evolutionarily conserved HCTs occupy nearly 2% of the human genome, more than half of the promoters contain HCTs and almost half of the 487 experimentally validated developmental enhancers contain them as well.
- Remark: the HCT criteria are very conservative: at least four sites, in ECR, HCT in both human and mouse. So the functional enrichment of these conserved HCT is expected. On the other hand, these sequences represent a small fraction of functional regions, e.g. less than 1% of ChIP-seq regions overlap with conserved HCT for several TFs (Table 1).

Discovering homotypic binding events at high spatial resolution [Guo & Mahony, Bioinformatics, 2010]:

- Motivation: from ChIP-seq data, discover distinct, but spatially-close binding events.
- Discovery of joint binding events (less than 500 bp): in GABP ChIP-seq data, GPS discovered about 450 single binding events, and 122 joint events.

3.6 Transcription Factor-DNA Interactions

Absence of a simple code: how transcription factors read the genome [Slattery and Rohs, TIBS, 2014]

- Background: DNA major and minor grooves.
- Base read out: mainly determined by interactions between AA residuals and DNA bases, through H-bond and hydrophobic interactions. The H-bond donor/acceptors are base-specific in major grooves, but not in minor grooves. So base read out is driven by the major groove.
- Shape read out: DNA bending, unwinding, DNA shape particularly in minor grooves and shape-dependent electrostatic potential.
- Most TFs use both base and shape readout: Figure 3. Ex1: two alpha helices in the TF interact with DNA major groove, and other parts recognize flanking bases in the minor groove. Ex2: shape readout: DNA bending of central spacer.
- Modeling TF-DNA interactions: to extend beyond simple motif recognition, MRF model, k-mer regression model, di-nucleotide model, also several models incorporating structural features. A thermodynamic model of protein-DNA interaction from Stormo group: energy terms from interactions between adjacent bases.
- Different modes of TF-DNA binding: variable spacing between half-sites, multiple DBDs, multimeric binding, alternative structural conformations.
- Why do TFs recognize specific sites, not all motif matches? Histone post-translational modifications (PTMs). DNA accessibility is a major determinant; however, caution about this: ChIP-seq experiment uses cross-linking, which favors open chromatin regions.
- Different types of TFs: Figure 7A-C. (1) Pioneering TFs: can open inaccessible regions. (2) Settler TFs: bind to all motifs in accessible regions. (3) Migrant TFs: bind to a subset of motif matches in accessible regions. Likely controlled by co-factors.
- Only a fraction of TFBSs from ChIP-seq have regulatory functions. Figure 6D. Regulatory functions are defined as: driving gene expression and respond to change of [TF].
- What may cause non-functional TF binding? (1) Non-specific binding in open chromatin regions (which are likely to be cross-linked in ChIP-seq). Evidence: high occupancy regions are gone using cross-linking free ChIP protocol; stronger peaks are more likely functional. (2) Cell type mixture: ChIP-seq uses millions of cells, so the strength of peaks may not reflect true binding strength. Ex. a high-affinity sites in only a small fraction of cells may be low peaks.
- Possible mechanisms of non-functional binding (Figure 7D): non-specific binding via electrostatic interactions (DNA: negative charge, TF: positive charge), binding to homo-oligomer tracts, indirect binding/tethering.
- How do multiple TFs work together to control gene expression? Enhancersome and billboard models (Figure 8). Both may involve cooperative TF interactions: former may be direct protein interactions, and the later may be nucleosome-mediated cooperativity or cooperative interactions at other levels (e.g. multiple TFs acting on transcriptional machinery).
- Evolutionary forces on different models of TF interactions: enhancersome may be suited for switch-like behavior of enhancers, and may be important when [TF]s are low.

- What leads to cell-type specific TF binding? Some TFBSs are not tissue-specific: e.g. some ER binding sites are present in multiple tissues, and they are high-affinity ER response elements. Generally, tissue-specificity may be controlled by chromatin environment, which is related to change of [TF] and co-factors. (1) Change of expression of co-factors. ESC example: Isl1 is broadly expressed, but its binding sites are refined by another TF. Expression of two TFs lead to two fates (two kinds of motor neurons). Both TFs form complex with Isl1. (2) Change of expression of TFs with similar binding specificities. GATA switch during HSC differentiation: from GATA2 binding to GATA1 binding, driven by upregulation of GATA1.
- Open problems: (1) How histone PTM affect TF binding? (2) How single-cell genomics can help us understand TF binding?

Kinetic model of TF-DNA interaction [von Hippel & Berg, PNAS, 1986]:

- Goal: how specificity of TF binding with its target sequence is achieved? With a small number of TF molecules, many competing sites in the genome, etc.
- Background: 2 modes of TF-DNA interaction: (i) specific binding using H-bond; (ii) non-specific electrostatic interactions. TF protein could take one of the two conformations. Note that non-specific interaction is important to get strong enough binding.
- Idea:
  - DNA contains many sites (of different affinities), but linear order is not important. Thus could view one site simply as one molecule with specific affinity. The system is equivalent to a system of TF molecules interacting with different classes of BS molecules: the kinetic approach
  - Number of sites of each affinity class can be approximated via random genome approximation; and the binding affinity could be modeled by a geometric distribution (each mismatch carries a constant penalty)
- Model:
  - Coupled equilibrium: a system of reactions

$$R + S_i \rightleftharpoons RS_i \tag{3.20}$$

    Let R_T be the total number of molecules of R (repressor), K_i be the association constant of the reaction for RS_i and D_i be the total number of sites S_i. Then one can show:

$$R_T = R_F + \sum_i \frac{K_i D_i R_F}{1 + K_i R_F} \tag{3.21}$$

    where R_F is the number of free R molecules. For a specific site s, we could compute its fractional occupancy θ_s by solving x, defined as $\theta_s/(1-\theta_s)$, from the following equation:

$$R_T = \frac{x}{K_s} + \frac{xD_s}{1+x} + \sum_{ns} \frac{xD_i}{x + K_s/K_i} \tag{3.22}$$

    Analysis: simplify by the approximation $x >> 1$ and $K_s/K_i < x$ for strong pseudosites and $K_s/K_i > x$ otherwise. See Equation (4) of the paper
  - Binding: plug in K_j and D_j into the equation of x. We have:

$$D_j = f_n(j) = 2N \left[\binom{n}{j} \left(\frac{1}{4}\right)^{n-j} \left(\frac{3}{4}\right)^{j} \right] \tag{3.23}$$

where n is the size of TFBS. The result is in Equation (8) of the paper, by applying the approximation that $x << K_s/K_{ns}$. The key step is to show that:

$$\frac{x}{x + K_s/K_j} \approx \frac{x}{x + d_j} + x\frac{K_{ns}}{K_s} \tag{3.24}$$

Analysis/Interpretation: to ensure specificity, the average number of TFs lost to non-specific binding ($\bar{m_s}$ in the paper) must be small $\rightarrow$ n should be large and d should be large $(5 - 100)$.

Free energy of binding [Berg & von Hippel, TIBS, 1988; JMB, 1987]:

- Model: consider position i of a site, let $b_{\max}$ be the best base at this position, $p(b)$ be the genomic frequency of the base b, and $f_i(b)$ be the frequency of b at position i of the count matrix of the sites. Then the mismatch energy (free energy of binding, relative to the strongest site) is:

$$\lambda\epsilon_{i,b} = \log\frac{p(b)}{p(b_{\max})}\frac{f_i(b_{\max})}{f_i(b)} \tag{3.25}$$

  Note that the mismatch energy is generally positive, suggesting that a site is not as good as the strongest site.

- Idea: consider all sites (with length L) of energy E, view it as a microcanonical ensemble: L particles with total energy E. Compute the frequency of a particle(position) being a certain energy.

Estimating binding energy from base frequencies [Heumann & Stormo, ISMB, 1994]:

- Problem: given a set of sites of some TF, estimate the energy matrix of the binding (s.t. the energy of any future site can be computed). Suppose $E(S)$ is the energy of a site S, which is the sum of energy of each position. Let $W_i(b)$ be the energy of base b at position i, need to determine the matrix W.

- Idea: maximize the probability of binding of all known sites.

- Model: the probability of binding of a sequence S is given by $P(S) = e^{-\beta E(S)}/Z$ where Z is the partition function and $\beta = 1/kT$. Need to:

$$\max\prod_{j=1}^{N}\frac{e^{-\beta E(S_j)}}{Z} \Leftrightarrow \min U = \sum_j \beta E(S_j)/N + \log Z \tag{3.26}$$

  Assume the genome is random, we approximate the partition function:

$$Z = \prod_i \sum_b p(b)e^{-W_i(b)} \tag{3.27}$$

  where $p(b)$ is the genomic frequency of b. Plug in the frequency of b in the known sites, $f_i(b)$ into the equation of U, we have:

$$\min\sum_i\sum_b W_i(b)f_i(b) + \ln\prod_i\sum_b p(b)e^{-W_i(b)} \tag{3.28}$$

  Because we are free to choose the reference state of energy, we could choose it s.t. $Z = 1$. Thus the problem becomes a constrained minization problem:

$$\min\sum_i\sum_b W_i(b)f_i(b) \quad \text{s.t.}\prod_i\sum_b p(b)e^{-W_i(b)} = 1 \tag{3.29}$$

  The solution is given by:

$$W_i(b) = -\ln\frac{f_i(b)}{p(b)} \tag{3.30}$$

  Thus, the energy of a site is the negative of its LLR score, or $E(S) = -LLR(S)$.

- Equivalence of Heumann-Stormo model and Berg-von Hippel model: under the Heumann-Stormo model, the mismatch energy is given by:

$$\beta\Delta E(S) = \beta E(S) - \beta E(S_{\max}) = \sum_i \ln \frac{p(S[i])}{p(S_{\max}[i])} \frac{f_i(S_{\max}[i])}{f_i(S[i])} \tag{3.31}$$

  which is exactly the equation given by Berg and von Hippel.

MatrixREDUCE [Foat and Bussemaker, Bioinformatics, 2006]:

- Problem: given genome-wide binding data of a TF, learn the binding specificity of this TF and the biophysical model.
- Methods:
  - Binding intensity of a sequence (extended region): several assumptions:
    - Additive energy of positions for any site, thus parameterize the motif by the energy (relative to some reference sequence, S_{ref}) at each position, w_{jb}, not need background distribution.
    - The fractional occupancy of any site is far from saturation, thus the occupancy is equal to the product between the factor concentration and the binding association constants (which allows some parameter to be merged).
    - The total occupancy of a region is equal to the sum of fraction occupancy of all sites in this region.
  - Measurements and parameter fitting: the test intensity, I^{test}, is proportional to the total occupancy plus a background intensity:

$$I^{\text{test}} = \gamma N(U) + \alpha^{\text{test}} \tag{3.32}$$

    where $N(U)$ is the occupancy of a region U. The control intensity is only equal to background intensity α^{control}. Finally, there is an error term. We have:

$$\frac{I^{\text{test}}}{I^{\text{control}}} = \beta N(U) + C + \epsilon \tag{3.33}$$

    where β, C are constants, and ϵ is the error. Plug in the equation of $N(U)$ and merge constants in $N(U)$ with β (new constant F), the free parameters become: w_{jb}, F and C. The parameters are fit by minimizing the square error, summing over all peaks/genes in the dataset.
  - Procedure: consists of several step: (i) construct seed matrix using word count approach; (ii) parameter optimization with quasi-Newton method; (iii) k-fold cross-validation; (iv) comparison of learned PWMs with results obtained independently.
- Results:
  - Comparison with information theory based approach: using the predicted binding energy and the experimentally measured binding energy. MatrixREDUCE vs BioProspect and MDScan: overall similar. However, those two methods depend on the background distribution and on the threshold used to label a region as bound.

TRanscription factor Affinity Prediction: TRAP [Roider & Vingron, Bioinformatics, 2007]:

- Methods:
  - Model: suppose the free energy of the strongest site is 0, the free energy of a site is defined by the sum of mismatch energy of all positions, scaled by λ. If the base at position i is α, the

most frequent base is max, and the frequency of bases in the PWM is $m_{i,\alpha}$ and $m_{i,max}$, then the mismatch energy is given by:

$$\log(\frac{m_{i,max}}{m_{i,\alpha}}\frac{b_\alpha}{b_{max}}) \tag{3.34}$$

The total occupancy of a sequence with length L is given by the expected bound molecules:

$$\langle N\rangle = \sum_{l=1}^{L-W} p_l = \sum_{l=1}^{L-W} \frac{R_0 e^{-\beta E_l(\lambda)}}{1+R_0 e^{-\beta E_l(\lambda)}} \tag{3.35}$$

where E_l is the energy of site starting at position i, and $R_0 = K(S_0)[TF]$ is the product of TF concentration and the binding affinity of the strongest site.

- Parameter esimation: the free parameters are:
  - λ: the scale of the mismatch energy, determines how strongly variations in the target sequence will be penalized. For large λ, the mismatch penalty is small; and small λ, large penalty.
  - R_0: small R_0, low occupany, depends linearly on R_0.

  The parameters are chosen by maximizing the correlation coefficient of the predicted $\langle N\rangle$ and the observed R/G ratio.

- Results:
  - Parameter range: the prediction is relatively insensitive to the parameters. Empirically choose $\lambda = 0.7$, and $\log R_0$ from linear regression with motif width W (it is found that R_0 depends roughtly linearly on W):

$$\log R_0 = (0.6 \pm 0.1)W - (6 \pm 2) \tag{3.36}$$

  - Prediction of TF targets (use threshold to decide if a region is bound or not in ChIP-chip data): measured by ROC curve area. TRAP is better than hit-based methods for most TFs. In this experiment, use $\lambda = 0.7$ and $R_0(W)$.
  - Prediction of regulators (again from ChIP-chip data): evaluated by the rank of known regulators in the prediction. TRAP better than hit-based method. Again, no parameter training by ChIP-chip data.

Conserved TF-binding affinity [Ward & Bussemaker, Bioinfo, 2008]:

- Motivation: TF targets can be predicted by binding, however, binding may not be functional. The idea is to predict functional TF targets through conserved binding.
- Outline: show that the predicted TF targets are more likely to be functional using functional genomic data: they are more likely to respond to change to [TF], etc.
- Methods:
  - Affinity: from [MatrixREDUCE], however, the binding parameter need not be estimated (see below). Called N^U.
  - Conserved affinity: Minimum affinity of the promoter in 4 yeast species. Called N^C.
  - TF-TF interactions: (i) Affinity co-occurrence: Spearman correlation between affinities of two TFs (across all promoters); (ii) Affinity co-conservation: Spearman correlation between conserved affinities of two TFs.
- Results:
  - Most predicted binding affinity for TFs is not conserved.

  - Conserved affinity correlates with regulatory susceptibility: the genes with high conserved affinity are more likely to change expression in TF deletion microarray measurements, and are more likely to have large functional binding (the proportion of binding that contribute to expression, learned from the MA-Network program).
  - Conserved affinity correlates with nucleosome depletion.
  - Affinity co-conservaton provides evidence for TF-TF interactions.

- Remark:
  - Assume the binding affinity is proportional to [TF], thus all sequences of the same TF share the same constant term [MatrixREDUCE], thus the binding parameter need not be estimated (from ChIP-chip data).
  - If not train the binding parameter, then the conserved affinity score is effectively the network-level conservation score of programs such as STUBB or Cluster-Buster.

Statistical normalization of TF-sequence binding [Manke & Vingron, PLoS CB, 2008]:

- Problem: given a sequence, want to predict which TF binds it, then need a statistical normalization of predicted binding affinity comparable across different factors.
- Methods:
  - Distribution of binding affinities from random sequences: the binding affinity is summation of affinity at each site, so one can use Fourier transform (similar to MGF) to obtain the distribution.
  - Parameterization of the distribution: generalized extreme value distribution (GEV) fits reasonably well the empirical distribution, but parameters are TF specific.
- Results:
  - For 567 human promoters with known TFs: rank all TFs wrt to each promoter and compare if the ranking is consistent with the known factors.
- Remark: the likelihood ratio of a sequence is an approximation of total binding affinity (likelihood ratio of a site is roughly the binding affinity or exponential of energy). But it does not have χ^2 distribution, and is not comparable across different TFs, as shown in this paper.

Affinity Density [Hazelett & Weiss, Bioinfo, 2009]:

- Problem: given the binding data of target sites of some TF, predict its genome-wide targets.
- Outline: a method based on signal processing that approximate binding affinity via integrating the site strength, the number and density of sites. Evaluation of the method by assessing the specificity and sensitity of the predicted targets, using known targets, the likely targets from expression data and from RNAi experiment.
- Method:
  - Scoring affinity: use the known sites with known affnity, and signal process technique called Hann filter.
  - Conservation of affinity in two species: minmum score in two species must be greater than a threshold.
  - Programs to compare: TargetExplorer and GenomeSurveryor.
- Discussion: the problem of "aliasing", for methods scoring cluster of motif matches. Sensitive to window size, binding site cutoff.

- Remark: two issues not clear with the comparison with competing programs: gene assignment (Affnity Density assigns regulatory regions 1kb surrounding genes), and conservation (not used by TargetExplorer and GenomeSurveryor, but the authors said the conserved motif search is used with the two).

Interaction-dependent TF binding [Wang & Hannenhalli, RECOMB-RG, 2005]:

- Problem: do TF-TF interactions contribute to binding, and how to identify them?
- Methods:
  - Linear regression: for each sequence, define the feature of a TF as (approximately) the total occupancy of this TF to this sequence. Then linear regression on the occupancies of TFs, assuming each TF contributes independently. The response variable is the transformation of the p value in the ChIP-chip data.
  - Model training: divide the data into training (50%), model selection (25%) and testing (25%). For each dataset (of one TF), choose the TF that maximizes the fit in the model selection data, add to the motif-set (for this TF). Finally, compare the model with non-interaction model in the test data.
  - Negative control: do the same procedure of model training except that a random TF is added to the motif set (instead of the best-fit one).
  - Data: 90 yeast TFs with ChIP-chip data and PWMs.
- Results:
  - Interaction model significantly improves non-interaction model (with only 2 TFs in the motif set). And predict 377 synergistic interactions and 68 antagonistic interactions.
  - Validation of predicted TF interactions: using the known TF-TF interactions in yeast cell cycle.

TF co-localization hotspots [Moorman & van Steensel, PNAS, 2006]:

- Aim: how different TFs bind to genomic regions? How are their binding sites/targets organized?
- Methods:
  - Data: DamID binding data of seven transcription factors with diverse physiological functions, five cofactors, and two heterochromatin proteins at 1-kb resolution in a 2.9 Mb region of the Drosophila melanogaster genome (Adh-cactus region).
  - Testing co-localization of two factors: Pearson correlation of binding profiles.
  - Defining hotspots: SOM analysis of all regions and K-means clustering, divided into 8 "chromatin types", and define type 1 as hotspots, which are bound by all factors except two.
- Results:
  - TF co-localization: strong correlation with almost any pair of TFs; visual inspection of hotspots. Not due to non-specific protein-DNA binding because Gal4 does not co-localize with other proteins.
  - Co-localization may be due to protein-protein interaction instead of DNA binding: (i) in type 1 regions (hotspots): enrichment of binding sites of some factors, but no Bcd, EcR, USP; (ii) mutant of Bcd DBD (DNA binding domain) resembles wt Bcd, while Bcd DBD alone shows very different binding.
  - Genes with hotspots tend to have higher expression level, measured by microarray experiments.

ChIPModules [Jin & Farnham, GR, 2006]:

- Problem: analysis of ChIP-chip data, find related TFs and the binding model
- Methods:
  - TFBS identification: find conserved TFBSs using Transfac PWMs and human-mouse alignment (must be conserved in both human and mouse)
  - Feature selection: find colocalized motifs by enrichment test in the neighborhood regions. For each sequence, find its best binding site of the experimental TF, then extract its neighborhood region within distance Δ (a program parameter) and search for motifs in the neighborhood region vs background by a hypergeometric test.
  - Binding model: use CART to classify bound vs non-bound sequences using TFBS presence (0/1) as features (only those features selected in the previous step).
  - Data: ENCODE regions; core promoters bound by E2F1
- Results:
  - ENCODE E2F1 ChIP data: find 24 PWMs that are enriched (at distance threshold 270 bps, which is determined by 10-fold CV). CART learns the best co-factors as AP-2α, NFAT, LBP1, ELK1 and EGR.
  - Comparison of models: presence of TFBS, presence of conserved TFBS and CART model to classify the sequences - ROC. CART model is significantly better.
  - ChIPModules performance in different datasets: measured by specificity and sensitivity of both training and testing data with 10-fold CV (Table 2).
  - Experimental validation of E2F1 and AP-2α co-localization (predicted by ChIPModules): ChIP-chip assays of E2F1 and AP-2α in 14,000 human promoter regions, and assess the overlap of regions bound by two factors.
  - Application of ChIPModules to a large promoter dataset (OMGProm database): 10 predicted ChIPModules targets are chosen for PCR analysis and most of them show high E2F1 and AP-2α binding.

Statistical methods for cooperative binding [Datta & Zhao, Bioinfo, 2008]:

- Problem: given ChIP data of two TFs, test if the two TFs bind cooperatively.
- Methods:
  - Log-linear model: given a pair of TFs, classify each gene into one of 4 categories, depending on if it is bound by the TFs: $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. This 2 by 2 table can be analyzed with the log-linear model:
    $$\ln(F_{ij}) = \mu + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12} \tag{3.37}$$
    where λ_i^1 and λ_j^2 are effects of factors f_1 and f_2 respectively, and λ_{ij}^{12} is the interaction term.
  - Mixture model with hidden states: the actual states are not observed, but only the binding intensities or p value of binding. The data of each gene is thus treated as (p_1, p_2) where p_1 and p_2 are p-values of the two factors. The state is a hidden variable with 4 possible values. The probability distribution of p value given the state is modeled by using local FDRs.
  - Data: ChIP-chip data of 204 TFs of yeast in different conditions: rich medium, aa starved, etc.
- Results:
  - Predict cooperative interactions of yeast TFs: many of the predicted interactions in cell cycle are supported by literature. The results are largely similar to those using log-linear models with p value cutoff.

- Validation by expression data: the average pairwise corrleation of the target genes of the cooperative pairs. Higher in mixture model results in simple thresholding.
- Cooperative binding is condition dependent.

Statistical learning of TF-DNA interaction [Zhou & Liu, NAR, 2008]:

- Methods:
  - Feature extraction: the motif feature of a sequence is defined as the sum of motif scores of top $m = 25$ sites in the sequence, where motif score of a site is the likelihood ratio score (first-order Markov chain as background).
  - Statistcal learning methods: predict response variable, Y, (log. of the ChIP intensity ratio) of a sequence (S), which is transformed to a set of features ($\mathbf{X}$). Note that Y is a continous variable. Linear regression, ANN, SVM, MARS, boosting and BART.
  - Validation: positive data from bound regions and negative data from random unbound region with the same length distribution. 10-fold cross validation using Pearson correlation as the performance measurement in the test data (CV-cor). Report the average correlation, and its standard deviation across 10 test samples.
- Results:
  - Human Oct4 data: baseline method (linear regression with single motif) - CV-cor 0.446; best methods are boosting (0.581) and BART (0.600).
  - Learned new motifs by BART: sequence features are chosen by the posterior inclusion probabilities. The learned motifs are validated by literature through the functions of the new motifs and the known TF-TF associations. Uf1h3b motif is similar to Klf4 motif (reprogramming factor); conservation of Uf1h3b motif in one known Sox2-Oct4 site; Nfy_Q6 is involved in ESC and associated with E2F.
  - Learned non-motif features: short di- or tri- nucleotides. Significantly affect the CV performance, e.g. removing all the non-motif features, reduces CV performance b 12-15%.
  - Validation with mouse Oct4 data: use the trained model from human to discriminate 1000 bound regions and 2000 random upstream sequences with identical length distribution. Measure performance by classfication errors. Reduced FP rate vs the baseline method - scanning using the Oct4/Sox2 composite motif.

Binding specificities of 104 mouse TFs [Badis & Bulyk, Science, 2009]:

- Methods:
  - 104 mouse TFs: covering 22 structural classes. DBD fusion with GST. Out of 104 TFs: 25 have known matrix from TRANSFAC or JASPAR; 14 have orthologs with known matrices.
  - TF binding specifities: PBM of binding affinities of all 8-mers. E-values are used to represent the strength of binding between a TF and a 8-mer (between 0.5, best, to -0.5, worst).
  - Seed-and-Wobble algorithm to search for secondary motifs: first find the primary motif, then among the k-mers of high signal intensity that are not explained well by the primary motif, search for the secondary motifs.
- Results:
  - Similar TFs may not have similar specifities: (1) Irf4 and Irf5: AA sequence identity greater than 60%, bind the same high affinity sites, but different low-affinity sites; (2) members of the same or related families: e.g. Sox and Sox-related families have qute different binding specifities - Sox family AACAAT, and Tec/Lef family TCAAAG.

- Many TFs have alternative motifs: nearly half of 104 TFs. Most secondary motifs represent different affinity classes (weaker binding). Different cases of secondary motifs:
  - 19 cases of positional interdependence: among two adjacent or distant nucletode pairs.
  - 1 case of variable spacer length.
  - 16 cases of combinations of positional interdependence and variable spacer length.
  - 5 cases of "alternate recognition interfaces": secondary motifs are very different from the primary ones.

High Resolution Models of Transcription Factor-DNA Affinities Improve In Vitro and In Vivo Binding Predictions [Agius & Leslie, PLCB, 2010]:

- The unique sequence in each probe is a 36-mer, and the probe set is mathematically specified to contain all possible 10-mers as subsequences. We used the probe data as labeled training examples, i.e. pairs $(x, y) =$ (sequence, intensity), for learning a function $f(x)$ that predicts binding intensity from (36-mer) sequences.

- Supervised learning strategy: we used K-mer based string kernels for representing the similarity of double-stranded probe sequences on the PBM, and we trained support vector regression (SVR) models to directly learn the mapping from probe sequence to binding intensity from PBM training data

- Kernel: computes a similarity between probe sequences based on inexact matches to k-mer features, allowing up to m mismatches, where we count mismatches in the alphabet of dinucleotides. A practical choice: considering 13-mer sequences, allowing up to 5 mismatches, and operating in the first order alphabet of dinucleotides.

Nucleosome-mediated cooperativity between transcription factors [Mirny, PNAS, 2010]

- Background: possible mechanisms of flexible TFBS organization (1) cooperativity mediated by simultaneous contact with BTM. However, if flexible TFBS happens in sequences not activating transcription (not all TFBSs are functional at a moment), this mechanism is insufficient. (2) One TF has an effect on changing chromatin structure/accessibility, or nucleosome eviction by chromatin remodeling.

- Model: a sequence (less than 150 bp) with n TFBSs. Key assumption: nucleosome is either bound or not. The whole system exists either in nucleosome-bound states, $N_0, \cdots N_n$ or open states $O_0, \cdots, O_n$, where i is the number of TFBS occupied by TF. Components of the model:
  - DNA binding by nucleosome: $N_i \Leftrightarrow O_i$, with equilibrium constant L, in the range of 100-1000.
  - TF binding: to both nucleosomal DNA (N) or naked/open DNA (O), so $N_i \Leftrightarrow N_{i+1}$ with constant K_N, or $O_i \Leftrightarrow O_{i+1}$ with constant K_O. The affinity of open DNA to TF is much higher, $K_O << K_N$ by 0.1 to 0.001.
  - TF concentration: impact K_O, and typically, [TF]$/K_O$ is 1-5.

- Behavior of the model: highly cooperative. In the TF occupacy vs. [TF] curve, high Hill coefficient (switch-like behavior). Intuition: binding of some TFBSs will shift the whole system from N states to O states, and effectively remove nucleosome, making the DNA more accessible to the remaining sites.

- Allosteric effects: Histone modification. Small change of L by histone modification can have a large effect on the TF occupancy.

- Number of TFBSs: when it is too small, not much cooperativity. Needs a minimum of 3-6 TFBSs every 150 bp.

A Linear Model for Transcription Factor Binding Affinity Prediction in Protein Binding Microarrays [Annala & Nykter, PLoS ONE, 2011]:

- Goal: given PBM data (35-mer probes) in one experiment, predict the intensity in a second, replicate experiment.
- Linear model: consider are all 4-mer, 5-mer, 6-mers, and 2000 7-mers and 1000 8-mers, use binary features (allow overlap): $h_{sk} = 1$ if the K-mer k is in the probe s and 0 otherwise. The design matrix is strand-specific, thus reverse complement K-mers are considered separately. The model:

$$p = H\alpha + \epsilon \tag{3.38}$$

  where p is probe intensity (log-transformed and mean-subtracted) and α is a vector of K-mer affinity contributions. LASSO is not used because of computational constraint.
- Probe noise model: DNA microarrays have been reported to have roughly linear probe noise, i.e. the noise leve σ^p is proportional to the intensity I_p, $\sigma_p = \beta I_p$. If technical replicates are available for a PBM experiment, it can be a good idea to estimate the coefficient β and solve the linear system using a weighted least squares approach.
- DREAM challenge: predict the intensity in the second (target) array. Average Pearson and Spearman correlations of 0.624 and 0.624, across the 86 paired PBM samples, using preprocessing and quantile normalization.
  - E-value based method: Spearman correlation of 0.53.
  - Highest median intensity K-mer (HMIK) in the probe's sequence: Pearson correlation 0.515, Spearman correlation 0.418.
  - 8-mer based HMIK predictor performed better than the PWM motif models at predicting binding affinities [Chen & Morris, RankMotif++, Bioinformatics, 2007].
- Effect of parameters on the performance:
  - Preprocessing: the average accuracy only sees moderate improvements from preprocessing (0.603 without any preprocessing).
  - K-mer length: the accuracy does consistently improve as the K-mer length approaches 8 bases. Roughly 1000 highest median intensity 7-mers and 8-mers are enough to achieve saturation in terms of accuracy. We also tried using only 6-mers and regularized 7- and 8-mers, but again the results were worse than our original model.
  - Probe noise model: no significant impact on prediction accuracies. This result may be explained by the observed lack of strong correlation between probe noise level and intensity.
  - Gapped K-mers: the inclusion of 500 gapped 8-mers to the model did improve prediction results in a statistically significant manner, although the improvements were very small (average Pearson correlation 0.624 to 0.626).
  - Strand specificity: important, without which, the average Pearson correlation across all 86 PBM samples dropped significantly from 0.62 to 0.56.
- Discussion:
  - The solutions are not unique: e.g. a 4-mer is often a linear combination of four 5-mer columns. We tried to avoid this issue by first learning a 4-mer model, then learning a 5-mer model based on the residual, then a 6-mer model on the new residual and so forth, but the performance is worse (0.614).

Quantitative analysis demonstrates most transcription factors require only simple models of specificity [Zhao & Stormo, NBT, 2011]:

- Background: energetically the situation appears much simpler, with individual base pairs often contributing approximately independently to the total binding energy.
  - Although deviations from strict independence are common, the nonindependent contributions tend to be of smaller magnitude compared with the independent contributions.
  - The physical intuition is that transcription factorDNA recognition is primarily based on complementarity between the sequence-dependent positioning of hydrogen bond donors and acceptors in the grooves of the double helix and those of the amino acids on the surface of the transcription factor.
  - Because most mutations change the shape of this network of hydrogen bond donors and acceptors locally, their effects are also mostly local.
- Background:
  - In PBM experiments, signal intensity of a probe in theory should be directly proportional to the probability of the transcription factor binding to the sequence of that probe. In practice, however, the relationship is not so straightforward owing to a number of factors such as background signal, position effect and influence of flanking sequences.
  - Drawback of E-values: (1) the intensities of probes containing the same 8-mer are highly variable, likely due to the effects of the confounding factors (e.g. flanking sequences). (2) More serious problem: low affinity sequences that partially overlap the binding site tend to appear on the same probes as high affinity 8-mers, often resulting in artificially high E-scores. Using PHO4 data, found that a substantial fraction of 8-mers with high E-scores are low affinity sequences that only appear to be enriched due to their flanking sequences.
- Methods: BEEML-PBM. Biophysical model, with correction of experimental artifact/bias (model probe intensity data as function of actual affinity of binding)
  - Biophysical model: the binding probability is related to the sequence. Suppose we have a sequence S_i at the position j of a probe, the probability that the sequence is bound is:

    $$P(j) = P(S_i) + (1 - P(S_i))P(\bar{S}_i) \tag{3.39}$$

    where $P(S_i)$ is the probability S_i is bound, and is determined by the energy of S_i, and $\bar{S}_i$ is the reverse complement of S_i.
  - Positional effect: (where the K-mer is in the probe) binding prob. is the average binding prob. over all possible positions of binding. Denoted as $F_{\text{pos}}(j)$.
  - For each probe, need to consider the background signal (i.e. only a certain fraction of signal is due to protein binding): at probe i, the probability that probe intensities in bin i is generated by TF binding is given by

    $$W_i = \frac{O_i - B_i}{O_i} \tag{3.40}$$

    where O_i is the observed number of probes in bin i, B_i is the expected number of probes in bin i from background distribution.

  We could then have the regression model: the binding probability of the probe i is:

  $$F(i) = \sum_j P(j) F_{\text{pos}}(j) \tag{3.41}$$

  And the objective function is:

  $$O(\epsilon, \mu) = \sum_i W_i (Y_i - a - cF(i))^2 + \lambda \sum_b \sum_k \epsilon(b,k)^2 \tag{3.42}$$

where $\epsilon(b, k)$ is the energy matrix, and a and c are parameters. The probe intensities Y_i are z-transformed.

- Results: We evaluate PWM performance by its ability to predict transcription factor binding preferences on a different PBM design. PBM experiments are performed using two arrays with different probe sequences, but both contain all possible 10-nt-long binding sites. The main finding: a single BEEML-PBM PWM is usually sufficient to provide excellent quantitative descriptions of PBM data.
  - The PWM estimated from replicate 1 performs very well on replicate 2 data (Fig. 1a), $R^2 = 0.91$. In contrast, the primary PWM found by Badis et al. does not capture Plagl1 binding specificity (Fig. 1b), $R^2 = 0.47$, leading the authors to the conclusion that multiple PWMs are required.
  - This holds true for most of the 41 transcription factors identified by Badis et al. as having clear secondary binding preferences. Figure 2a shows that in all but seven cases, a single PWM explains $> 90\%$ of the experimental variability, defined as the reproducibility of 8-mer median intensities (R2) between replicates.
  - We find that the BEEML-PBM PWMs are usually shorter than the PWMs found by Badis et al., and that those PWMs are often consistent with the EMSA results. With a few exceptions, the simple PWM model performs very well, supporting the hypothesis that the energetics of transcription factor-DNA recognition is generally simple.
  - The suboptimal estimation of the PWMs in previous studies can be accounted for by the lack of a biophysical model for transcription factor binding and the use of summary statistics, such as E-scores and Z-scores. This can be corrected by taking into account the specific characteristics of PBM data and maximizing the fit to the intensity data directly.
- Comparison with other methods: BEEML-PBM improved the predictions in every case (compared with UniPROBE PWMs), the resulting model has many fewer parameters than the SVR model, and each parameter has a specific biophysical interpretation (e.g., a binding energy contribution of a specific base-pair to the transcription factor-DNA interaction).

Determination of transcription factor binding [Cheung & Ruan, NG, 2011]

- Co-localization of Estrogen-receptor (ER) α and FoxA1 binding sites: through motif analysis of ChIP-seq data of ER-α.
- Importance of FoxA1 in determining ER binding: knockdown of FoxA1 in breast cancer cell lines significantly reduces ER binding. Even in ER-sites that do not co-localize with FoxA1 sites, reduced ER binding is observed.
- Model of FoxA1 action: nucleosome-dense regions where ER-a is bound are more likely to be bound by FoxA1; nucleosome-free regions where ER-a is bound are less likely to be bound by FoxA1. This suggests a model where FoxA1 is a "pioneering" factor that help displace nucleosome and ER-binding.
- Lesson: multiple models of combinatorial interactions among TFs. One model is the "pioneering factor" model: one factor helps change chromatin structure s.t. another factor can bind.

Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors [Wagih and Frey, biorxiv, 2018]

- Data: about 32K ASB variants with $p < 0.01$ (binomial) across 80 TFs. Also 79K non-ASB variants: $p > 0.5$ with at least 10 reads. Highly skewed distribution of number of ASB variants per TF (Figure 1e): CTCF has 7000 ASB; and 51 TFs have at least 20 ASB.
- Classification of ASB vs. non-ASB: CADD, Eigen and GWAVA, nearly random performance.

- PWM scoring: delta raw scores the windows with biggest difference of PWM scores between two alleles. delta track: difference of max. score windows for each allele. The latter is better: a low-scoring wild type sequence harboring a variant in an important position of PWM will have high delta raw score.
- Performance comparison: DeepSEA log-FC performs the best, and slightly better than DeepBIND and gkmSVM. In general, model AUC (in predicting peaks vs. non-peaks) works well for many TFs, but their ASB AUC can vary greatly.
- Possible impact of other factors: number of PPI partners - TFs with many PPIs show worse performance. DNA methylation: reduces performance.
- Remark: (1) deep learning methods consider the context of ASB: thus if a variant is not in a TFBS, the change of scores by two alleles is likely small. (2) Why model AUC does not predict ASB AUC? Sequence context can help classify TFBSs vs. non-TFBSs, but may not help with predicting ASB.

3.6.1 DNA Structure and Dynamics

DNA breathing: possible impact on transcription factor - DNA interaction [Boian Alexandar, meeting, 2020]

- Background: breathing is spontaneous base flipping/DNA denaturation. It is not limited to DNA, proteins can also display breathing.
- EPBD model: polymer model of DNA. Each monomer/bp n, let y_n be the distance of two strands. There are two types of strings: within a monomer (i.e. base-pairing), between adjacent monomers. We consider the free energy of the polymer $H(y) = \sum_n V(y_n) + W(y_n, y_{n-1})$, where $V(y_n)$ is the potential energy of monomer n, and $W(y_n, y_{n-1})$ the potential of adjacent monomers. $V(y_n)$ depends on base-pairing: with AT pair (3 H-bonds) being more "stiff" string. For the second interaction:

$$W(y_n, y_{n-1}) = \frac{K_{n,n-1}}{2}\left(1 + \rho e^{-\beta(y_n - y_{n-1})}\right)(y_n - y_{n-1})^2 \tag{3.43}$$

  where $K_{n,n-1}$ is given by the dinucleotide sequence (extended PBD model).
- Simulation and sampling of EBPD model: use Langevine MD and MCMC. Advantages of LMD: information about trajectory, and bubble duration time (ps scale).
- Implications: (1) DNA bubble formation depends on DNA sequences, with higher DNA stability (e.g. AT-rich sequences) showing lower bubble formation. (2) Propagation of effects: if we change bubble of one base, it may expand to sequences 100bp away.
- Representation of bubble formation: focus on equilibrium distribution, let L be length (number of bases covered), A be amplitude and n be position of the bubble. We can represent using the PDF of $P_n(L, A)$. Could also model bubble duration if using LMD.
- Experiments showing importance of DNA breathing on transcription and TF-DNA interactions: (1) One example: add a few bases near (10bp) a motif, changes DNA bubble and TF interaction. (2) Jablensky, Mol. Psychiatry 2012: DNA bubble 86 bp away from motif can change TF binding. (3) Fis1-DNA affinity: a few bases in the motif not important for binding (base mutation has no effect), but adding DNAm, changes bubble, and change TF binding.
  Note: EPBD model can incorporate effects of DNAm.
- DNA breathing on circular DNA: DNA bubbles generally make DNA structure more flexible, acting as hinges that facilitate DNA looping.
- Discussion: the importance of nucleosome. DNA-histone interactions have time scale of ∼ms scale, but bubble ∼ps scale, so it's OK to ignore nucleosomes.

- Discussion (from Alan Bishop): (1) the stacking potential in the EBPD model captures primarily entropy, similar to vibrations. (2) Time scales are important: opening has to be slow for protein attachment or engaging.
- **Remark**: is it possible that DNA bubbles are important for chromatin folding?

DNA breathing dynamics distinguish binding from nonbinding consensus sites for transcription factor YY1 in cells [Alexandrov and Usheva, NAR, 2012]

- Background: TF-DNA interactions involve direct points-of-contact (DRS) and induced fit (indirect recognition). The latter is sensitive to flanking sequences.
- DNA breathing dynamics: simulation to obtain DNA bubble probabilities (BP) and bubble life time at the DRS.
- Flanking sequences of YY1 consensus sites can influence binding: PLG promoter and other examples, binding of oligonucleotide in vitro, but some sites do not bind in vivo. The unbound sites have lower BP (less than 1/4) than bound sites at DRS.
- Single SNP in the flanking regions can change YY1 binding: a site in a region associated with SCZ, SNP 30bp upstream of YY1 DRS. The difference is the lower BP of the DRS of the unbound YY1 site.
- Model: YY1 DRS, multiple point of contacts between the protein and the template DNA strand, so higher BP is expected to facilitate binding. However, for other TFs that make contacts with both DNA strands, suppressed BP may favor TF-DNA binding.
- Remark: the paper does not fully establish that the difference of YY1 bound and non-bound sites is due to difference in BP - could be other differences in the flanking sequences.

Binding of Nucleoid-Associated Protein Fis to DNA Is Regulated by DNA Breathing Dynamics [Nowak-Lovato and Alexandrov, PLCB, 2013]

- Background: DNA breathing that yields relatively long bubbles is a feature of mammalian core promoters. High propensity of bubble formation correlates with TF binding sites and affect TF binding
- Background: Fis-DNA binding requires or induces DNA bending. Rules of binding: palindromic motif, 2 inclusion rules -7G and -3A/G, and exclusion rule -4A.
- FIS1 vs. FIS2 binding sites: FIS1 much stronger binding than FIS2. Both satisfy all rules. Sequence difference of 27 bases: 5bp difference in the middle (core binding region, but not direct contact). Simulation: higher bubble formation at FIS1 than FIS2.
- Characterizing strong FIS binding in vitro: 58 high affinity sites, do MCMC (faster version) to obtain opening profile, measured by average base displacement over nearby positions. Strong binding is characterized by high opening in -2 to +2 region (characteristic opening profile or COP): Figure 3.
- Classification of FIS binding sites: (1) Training sets: for positive sequences, choose sites that have high correlation with COP. (2) SVM: presumably using opening profile at each position as features. (3) Results: AUC somewhat higher than two other programs which use structural features of DNA.
- Remark: the classification experiment seems circular as the training examples are chosen to match the COP.
- Remark: COP is defined by particular pattern of opening profiles (bubble in the middle 5 bp, not in direct-contact bases). Not clear if this can be generalized to other TFs.

A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder [Duan and Gejman, AJHG, 2014]

- Background: MIR137 is one of the strongest GWAS loci in SCZ. The miRNA also targets a number of ASD genes.
- Rare variant near MIR137 locus: sequence 2K cases and 2K controls. Rare variant burden in promoter and enhancer variants.
- Single SNV: A>T variant drives the rare variant burden (not explaining all burden), 11 in cases and 0 in controls.
- Functional study: reporter assay confirms the function of mutation and 3C shows the interaction with MIR137 but not the two other adjacent genes.
- The A>T variant reduces YY1 binding: the variant is 5 bp upstream of YY1 motif. EMSA confirms the reduced binding. DNA breathing simulation shows a large effect of A>T on bubble probability.
- Lesson: a variant near TF motif can affect TF binding, and DNA breathing might be one mechanism.

Quantitative modeling of transcription factor binding specificities using DNA shape [Zhou and Rohs, PNAS, 2015]

- Background: four DNA structural features, see DNAshape paper, Figure S1. Minor groove width (MGW), Roll, Propeller Twist and Helix Twist. For each base, compute its four features using 5 bases centered around it.
- Method: PBM data, to incorporate DNA shape, combine 1-mer feature and 4 shape features: 1mer+shape model. For each position, we have five features: whether it is a particular base, whether it has the four shape features. Do SVM training.
- Results: 1mer + shape model is significantly better than 1mer model (Figure 1B). The improvement differ between different TF families. Also larger improvement on gcPBM data, which has less positional bias and capture information on genomic flanking regions.
- Models with 1mer + 2mer and 1mer+2mer+3mer are similar to 1mer+shape. However, the k-mer models work less well with smaller datasets because of higher model complexity.
- Learned feature weights reflect TF structure: Figure 5C, the alternating signs of the weight of Roll feature match the angles in X-ray structure.
- Question: in the 1mer model, does it require to find the motif location? Is the model based on the single/best PWM or some kind of sum of PWM scores?

3.7 Enhancers: Mechanisms and Modeling

Reference:

- Genomic views of distant-acting enhancers [Visel & Pennacchio, Semin. Cell Dev. Biol., 2007; Nature, 2009;]
- Enhancer function: new insights into the regulation of tissue-specific gene expression [Ong & Corces, NRG, 2011]
- Functional and Mechanistic Diversity of Distal Transcription Enhancers [Bulger & Groudine, Cell, 2011]

Function of developmental enhancers:

- Modularity of transcriptional regulation by enhancers:

  - ApoE: At least six distinct sequence elements flanking this gene control different aspects of APOE expression, including kidney expression by promoter, two liver-specific enhancers in downstream of the gene, a skin enhancer, two multiple tissue enhancers directing gene expression to adipocytes, macrophages and brain astrocytes, and a distal brain-specific enhancer. Each of these discrete elements are on the order of several hundred basepairs in length and are scattered across 42 kb.
  - Nkx2-5 (Csx): at least five distinct elements target Nkx2-5 gene expression to specific sub-regions of the developing heart as well as to non-cardiac tissues and it has been suggested that this regulatory complexity played a important role in the evolution of the multi-chambered mammalian heart.

- Spatiotemporal precision of developmental enhancers: the high spatiotemporal precision of single enhancers - in combination with their modular mode of action - has allowed complex gene expression patterns to evolve. These single elements tend to be more restricted in their tissue specificity than the mRNA expression patterns to which they contribute. Example: Hoxd11.

Identification of enhancers:

- Comparative genomics:
  - Conserved non-coding sequence were found not to be randomly distributed in the genome. Instead, they are located in a highly biased manner near genes active during development. Ex. approximately 90% of the tissue-specific p300 peaks identified by ChIP-seq in developing mouse tissues overlapped regions that are under detectable evolutionary constraint.
  - Lack of conservation in ENCODE sequences: the molecular marks of a surprisingly large proportion of sequences in the ENCODE regions suggested that regulatory functions are not, or are only weakly, conserved.
  - Lack of evolutionary conservation in developmental enhancers: e.g. eve2 enhancer, Pax2 enhancer in Drosophila, and pho2b enhancer of zebrafish, similar expression pattern, but not conserved.
  - Multi-species TF binding data: e.g. only 10-22% TF binding is conserved in mammals in CEBPA and HNF4-α targets in liver.
  - Limitations: (1) indicative of function, but it is not necessarily indicative of enhancer activity. (2) Conservation cannot predict when and where an enhancer is active in the developing or adult organism.

- DNaseI hypersensitivity: usually short (100-300 bp) regions of genomic DNA from which nucleosomes are excluded due to the binding of transcription factors.

How enhancers function: chromatin modification

- Nucleosome dynamics and chromatin remodeling:
  - Many functional enhancers have unstable nucleosomes containing the histone variants H3.3 and H2A.Z, and replacement of these nucleosome may facilitate the access of TFs to promoters and enhancers. Ex. in prostate cancer cells, some enhancers contain H2A.Z, upon androgen stimulation, AR replaces H2A.Z-containing nucleosomes and lead to an increase in flanking H3K4me2 nucleosomes.
  - The presence of dynamic nucleosomes reflects the activity of the DNase I hypersensitive sites.

- Chromatin signatures of enhancers: histone modification patterns
  - Promoters: active promoters defined by Pol II and TBP-associated factor 1 (TAF1). Marked by nucleosome-free regions with flanking H3K4me3. The occupancy is largely invariant across cell types (?)

  - Enhancers (as defined by p300 binding): highly enriched in H3K4me1, H3K4me2 and H3K27ac. The pattern is highly cell-type specific and correlate with tissue-specific gene expression pattern.
  - Enhancers (as defined by DNase I hypersensitivity): in CD4+ T cells, associated with all three H3K4 methylation states and the histone variant H2A.Z. In addition, subsets of putative enhancers contain distinct histone modifications, including H3K9me1, H3K18ac, H3K9ac and H3K14ac. The discrepency may result from the fact that p300 binds only a subset of enhancers.

- Effect of enhancer activation:
  - Co-activators: Enhancers were frequently associated with p300/CBP and/or TRAP220 (also known as MED1). The location of cell-line-specific distal DNase-I-hypersensitivity sites correlates with p300 binding at these sites.
  - Enhancers affect downstream transcriptional events: such as RNAP release. Ex. H3S10ph is required for activation of cytokine-induced genes, HSP and Myc-targets. The adaptor protein 14-3-3 binds to H3S10ph nucleosomes and help release RNA Pol II from paused states.

How enhancers function: stepwise enhancer activation during development [Ong & Corces, NRG, 2011]

- Factor relay model: TFs bind to enhancer sequences in a stepwise fashion during development. Ex. in ESC, enhancers of genes related to B-cell differentiation are bound by SOX2 and FOXD3. SOX2 contributes to the establishment of the H3K4me2 mark and FOXD3 represses transcripion in ESC. As ESCs differentiate into pro-B cells, recruitment of the lineage-specific SOX4 to SOX2 binding sites is required for enhancer activation.

- Epigenetic signals for activation potential: epigenetic modification may exist long before gene activation. At each stage, the epigenetic patterns are interpreted by cell-type specific TFs. The outcome may involve gene expression or alteration in epigenetic information that modifies its response capacity and narrows down cell differentiation choices.
  - Ex 1. in ESC, H3K4me1 and H3K27ac are associated with active enhancers, and H3K4me1 and H3K27me3 are linked to inactive enhancers (poised enhancers). During differentiation, H3K27me3 is replaced by H3K27ac and enhancer is activated.
  - Ex 2. in HSC, H3K4me1, H3K9me1 and H3K27me1 marks are associated with the enhancers of differentiation genes prior to their activation.
  - Ex 3. in MCF7 and LNCaP cells, many enhancers are marked by H3K4me2. FOXA1 swithces on distint transcriptional programs in two cells by binding to ER or AR binding sites in enhancers.

How enhancers function: linking enhancers with promoters

- Background: chromosome conformation capture (3C) assay and its derivatives:
  - 3C: mapping enhancer-promoter interactions. Similar to ChIP, the 3C approach relies on formaldehyde crosslinking to capture DNA-DNA interactions directly in intact cells or cell nuclei.
  - More advanced methods such as 4C (linking with chip). Example: beta-globin gene locus control region (LCR) makes reproducible tissue-specific contacts with other loci predominantly located on the same chromosome but in some cases dozens of megabases away from the LCR.

- Long-range interactions:
  - Chromatin looping vs. tracking (scanning) model: the looping model is supported by the experiments using 3C and its derivatives.
  - Enhancer-promoter interaction is not limited to genes located in cis on the same chromosome: Ex. olfactory H enhancer has been shown to interact with multiple olfactory receptor genes both on the same and different chromosomes.

  - Transcription factories: a nuclear subcompartment that is rich in RNA polymerases and TFs where dispersed genes gather to become active.

- Cohesins:
  - Cohesin interaction with CTCF. CTCF organizes global chromatin architecture by mediating intra- and interchromosomal contacts.
  - Cohesin may stablize enhancer-promoter interactions.
- Non-coding RNA (eRNA): may have some role in promoter-enhancer interactions.

Role of enhancers in human diseases:

- Non-coding SNPs from GWAS: in 40% of cases (472 of 1,170) no known exons overlap either the linked SNP or its associated haplotype block (LD), suggesting that in more than one-third of cases non-coding sequence variation causally contributes to the traits under investigation.
- Examples of enhancer mutation leading to diseases:
  - Thalassaemias: the main symptom is anemia, resulting from imbalance of α- and β-globins. Many of these globin chain imbalances were due to deletion or chromosome rearrangements that resulted in the repositioning of distant-acting enhancers required for normal globin gene expression.
  - Limb development defects (preaxial polydactyly): limb-specific long-distance enhancer ZRS (also known as MFCS1) of SHH, located in the intron of a neighboring gene (1M bp away). Approximately a dozen different single-nucleotide variations in this regulatory element have been identified in humans with preaxial polydactyly and segregate with the limb abnormality in families.
  - Hirschsprung's disease: an enlargement of the colon, caused by bowel obstruction. An enhancer sequence located in intron 1 of proto-oncogene RET was identified and found to contain a common variant contributing more than a 20-fold increased risk for Hirschsprung's disease than rarer alleles in this element.

A shared architecture for promoters and enhancers [NG, 2016]

- GRO-cap experiment: detect unstable nascent RNA (5' cap) with higher sensitivity.
- Similarity between promoters and enhancers: pairs of transcripts in a divergent manner with relatively tight spacing of 110 bp. TF in the middle, and similar binding of Pol II, TBP, TFIIB.
- Difference in post-transcriptional regulation: promoters give rise to a stable transcript in one direction and an unstable transcript (uaRNA) in the antisense direction, enhancers give rise to unstable transcripts (eRNAs) in both directions. Due to: presence of early polyadenylation signals and the absence of splice sites in the enhancer transcripts.

Pervasive Chromatin-RNA Binding Protein Interactions Enable RNA-Based Regulation of Transcription [Cell, 2019]

- ChIP-seq of many RBPs: find often RBPs actively bound in chromatin regions, especially promoters.
- Extensive co-association between TFs and RBPs. Ex: RBP25 and YY1 colocalization. RBM25 depletion attenuates all YY1-dependent activities, including chromatin binding, DNA looping, and transcription.
- Possible model: RBPs may enhance network interaction through harnessing regulatory RNAs to control transcription.

3.7.1 Physical and Statistical Models of Enhancers

Statistical mechanical model of TF-DNA interaction [Gerland & Hwa, PNAS, 2002]:

- Model:
  - Single TF molecule
    Assumption: (i) no free TF molecule, i.e. each TF molecule must bind with some position in DNA; (ii) each TF binds with two modes with energy $E(\vec{s})$ and E_{ns} respectively.
    Analysis: all micro-states include TF binds with $1, 2, \cdots, N$ sites with one out of two possible modes, the partition function is given by:

$$Z = \sum_{j=1}^{N} e^{-\beta E(\vec{s_j})} + N e^{-\beta E_{ns}} \tag{3.44}$$

    where $\beta = 1/k_B T$. Then the probability of TF binding to a specific target sequence t is:

$$P_t = \frac{1}{1 + e^{\beta(E_t - F_b)}} \tag{3.45}$$

    where F_b is the total partition function of the genomic background. The background can be viewed as random sequence, thus the expectation of F_b is given by:

$$F_b = -k_B T \overline{\ln Z_b} \approx -k_B T \ln \overline{Z_b} \tag{3.46}$$

    $\overline{Z_b}$ has two parts, one from specific interaction and the other from non-specific interaction:

$$\overline{Z_b} = \overline{Z_{sp}} + N e^{-\beta E_{ns}} \tag{3.47}$$

    And assuming N sites are independent and generated from multinomial sampling:

$$\overline{Z_{sp}} = N \prod_{i=1}^{L} \left[\sum_{s \in \{A,C,G,T\}} e^{-\beta \epsilon_i(s)} p(s) \right] \tag{3.48}$$

    where L is motif length and $p(s)$ is the genomic frequencey.
  - Multiple TF molecules
    Idea: treat the target site and its bound TF molecule, if any, as an open system which exchanges particle (the TF molecule) and energy with environment (genome).
    Analysis: apply the Gibbs distribution (bound and un-bound states differ by energy E_t and by one TF molecule):

$$P_t = \frac{1}{1 + e^{\beta(E_t - \mu)}} \tag{3.49}$$

    where μ is the chemical potential of TF. Let n be the total number of TF molecules, because $N >> n$, the genome can be viewed as a dilute solution, then we know that μ should be $k_B T \ln n$ plus some constant. When $n = 1$, we should reduce to the previous case, so the constant must be F_b. We have:

$$\mu = k_B T \ln n + F_b \tag{3.50}$$

    Rewrite the equation of P_t:

$$P_t = \frac{1}{1 + e^{\beta(E_t - F_b)}/n} \tag{3.51}$$

    One could define the threshold TF concentration as: $\widetilde{n} = e^{\beta(E_t - F_b)}$.

- Remark/Question: if there are multiple copies of the same target sequence in the genome, then TF could not distinguish these sites. The binding probability therefore should be reduced (by the number of identical copies).

Bacterialphage lamda O_R control system [Shea & Ackers, JMB, 1985]:

- Problem: the maintanence of the lysogenic state of phage lambda: normally OFF, triggered by external signal
- Process: let R be cI repressor and C be cro, P_R is the promoter of cro and P_{RM} be the promoter of cI repressor. There are 3 operator sites involved in controlling P_R and P_{RM} (see Fig. 1)
  - signal $\rightarrow$ recA
  - recA leads to degradation of R
  - R binds to O_{R_1} and O_{R_2}: P_R is OFF
  - C binds to O_{R_3}: P_{RM} is OFF
  - C binds to O_{R_1} and O_{R_2}: P_R is OFF The system is at lysogenic state when P_{RM} is active (R is repressor, as long as it is high, it will be lysogenic).
- Thermodynamic framework for modeling gene regulation: consider a DNA sequence segment (e.g. a promoter), let s be one state of the sequence (the occupancy states of its binding sites as well as any variables needed to chacterize the interactions of bound proteins), we want to compute P_s, the probably of state s. The key is:
- **Physical idea:** treat the DNA sequence and all bound proteins as a microscopic system and everything else the environment. At equilibrium, the system must have the same chemical potential of all its proteins as the environment.
- Let ΔG_s be the free energy of s, including all DNA-protein, PPI, as well as DNA looping, etc. For simplicity, we assume there are only two types of proteins A and B, and there are n_A and n_B molecules of A and B in the state s repsectively. The chemical potentials of A and B can be determined from the environment because of equlibrium: $\mu_A = k_B T \ln[A] + \mu_{A_0}$ and similar for B, where μ_{A_0} is a constant. Apply Gibbs distribution:

$$P_s = \frac{1}{Z} e^{(-\Delta G_s + \mu_A n_A + \mu_B n_B)/k_B T} \propto [A]^{n_A} [B]^{n_B} e^{-\Delta G_s / k_B T} \tag{3.52}$$

  The equation can be easily extended to the case involving more type of molecules. Basically, each bound molecule will contribute its concentration to the product.
- Model: several assumptions/ideas (O represents empty/unoccupied state)
  - cooperativity of R binding at O_{R_1} and O_{R_2}
  - P_R is ON when $O_{R_1} = O$ and $O_{R_2} = O$
  - P_{RM} is ON when $O_{R_3} = O$

Transcriptional activation by cooperative binding [Gibson, Theoretical and Population Biology, 1996]:

- Motivation: explain the sigmoidal/switch-like behavior of transcriptional response vs [TF]. What determines the threshold location and range (narrowness)?
- Physical idea: switch-like behavior is caused by cooperative DNA binding.
- Model:

– Binding: suppose there are n sites in the sequence, consider a configuration c, if there are k_c sites occupied, its free energy of binding is given by:

$$\Delta G_c = k_c u + a_c w \tag{3.53}$$

where u is the energy from binding of a single TF and w is the energy from cooperative binding between bound TF molecules. The term a_c measures cooperativity: its simplest form is $a_c = k_c - 1$. It may be necessary to use a stronger form of cooperativity (in the case of hb activation by Bcd). The probability of c is given by:

$$f_c = \frac{x^{k_c} \exp(-\Delta G_c / RT)}{\sum_c x^{k_c} \exp(-\Delta G_c / RT)} \tag{3.54}$$

– Transcriptional response: let the transriptional rate of c is r_c, then the total response is:

$$\text{response} = \sum_c r_c f_c \tag{3.55}$$

An additive relationship of r_c is given by: $r_c(0) = 0$ and

$$r_c(k_c) = \alpha + \beta(k_c - 1) \tag{3.56}$$

Note that all configurations with the same number of sites occupied can be merger with a single "state".

– Threshold width: defined as the ratio of the two [TF] required to produce a specified activation of target gene, e.g. 25% to 75% of maximal activity. Threshold width is reduced by the number of binding sites of activator, but not cooperative interaction energy (once it reaches a certain level). Require 5 to 6 sites to achieve minimum threshold width.

- Results:

  – Activation of hb by Bcd: response curve, dependence on the number of Bcd sites. Fig. 3.

Transcriptional activation by synergy [Wang & She, JMB, 1999]:

- Motivation: sigmodial behavior of transcription activation; synergy is reduced at high level of GTFs.
- Physica idea: synergy of transcription - multiple bound activators simulatenously contact the general transcription factors (GTFs) in the BTM. Furthermore, the additional synergy from assembly of GTFs: e.g. one bound TF contact TFIIA, another TFIIB, yet another TFIIF, then additional interactions between TFIIA and TFIIB, and between TFIIB and TFIIF.
- Model:

  – Commitment of promoter: Z molecule binds with DNA independently. Let Z be the activator (ZEBRA), and S_J^Z be the state of promoter where J molecules of Z are bound. Suppose there are N sites, and Δz be the energy of binding of Z to DNA, we have:

$$(N - J) e^{-\Delta z / RT} [S_J^Z][Z] = (J + 1)[S_{J+1}^Z] \tag{3.57}$$

This is derived from detailed balance of two states J and $J+1$ in the Markov chain of the system (see below).

– Nucleation/assembly of holoenzyme: each bound Z molecule could interact with some GTF with energy Δf, and any additional molecular of Z could contribute to an energy Δg (assembly of GTFs that are caused by nucleation on Z. Let $S_J^Z Z_M^F$ be the state where J molecules of Z are

bound to DNA, and M out of J interact with GTFs, and $[F]$ be the concentration of GTF, then we have the balanced equation for $M = 1, 2, \cdots, J-1$:

$$(J-M)e^{-(\Delta f+\Delta g)/RT}[S_J^Z Z_M^F][F] = (M+1)[S_J^Z Z_{M+1}^F] \tag{3.58}$$

For $M = 0$, we have:

$$Je^{-\Delta f/RT}[S_J^Z Z_0^F][F] = [S_J^Z Z_1^F] \tag{3.59}$$

- Recruitment: ZEBRA-GTF complex is tethered to the core promoter with interaction energy Δp. In addition, when $M > 1$, there is an additional interaction Δq. Thus we have:

$$[S_J^Z Z_1^F P] = [S_J^Z Z_1^F]e^{-\Delta p/RT} \tag{3.60}$$

And for $M > 1$:

$$[S_J^Z Z_M^F P] = [S_J^Z Z_M^F]e^{-(\Delta p+\Delta q)/RT} \tag{3.61}$$

- Transcription: the level of transcription is proportional to the total $[S_J^Z Z_M^F P]$ over all J and M.

- Results:
  - Experiment system: ZEBRA (from EBV) and human cell extracts and template bearing multiple ZEBRA binding sites. Change the number and affinity of ZEBRA binding sites and measure the transcription response.
  - Synergy: measured by the ratio of transcription response of a template with n sites wrt. n times the response of the template with 1 site. Synergy is not monotonic with [F]. Maximized at some intermediate level of [F].

- Remark:
  - Markov chain formalism/analogy of chemical reactions: if substances of a system of reaction can be viewed as different states of one system/complex, then the reactions among substances can be modeled as a Markov chain, where state transitions correspond to reactions between different substances.

Combinatorial transcription logic [Buchler & Hwa, PNAS, 2003]:

- Goal: explain how different arrangements of TFBS could generate different transcription logic
- **Principle**: if a system consists of two indepedent subsystems, then the partition function of the system is the product of the partition functions of the two subsystems. In particular, some subsystems may be reused in a complex system, and this property could simplify the computation of partition functions.
- Background: (TF binding with individual site) 2 micro-states: bound or unbound. The Boltzman weigth of unbound state is 1 and of the bound state is $q = [TF]/K$, where K is the dissociation constant, the probability of TF binding or fractional occupiency of the site is:

$$P = \frac{Z_{ON}}{Z_{ON} + Z_{OFF}} = \frac{q}{1+q} \tag{3.62}$$

- Physical idea:
  - The protein-protein interactions (TF-TF or TF-RNAP) stabilizes or distablize certain micro-states, thus affecting the probability of RNAP binding.
  - Three types of protein-protein interactions between protein i and j: repression ($\omega_{ij} = 0$), no interaction ($\omega_{ij} = 1$), activation ($\omega_{ij} = \omega$). In addition, there could be strong activation, when $\omega_{ij} = \Omega$.

- Model: suppose there are L sites of a promoter sequence, let σ_i be the occupiency of site i: $\sigma_i = 1$ if bound; $= 0$ if not bound. Let q_i be the weight of the TF_i when bound with site i, and ω_{ij} be the weight of interaction between TF_i and TF_j. Then the Boltzman weight of the micro-state $\sigma_1, \cdots, \sigma_L$ when RNAP is not bound is:

$$W(\sigma_1, \cdots, \sigma_L) = \prod_{i=1}^{L} q_i^{\sigma_i} \prod_{i<j} \omega_{ij}^{\sigma_i \sigma_j} \tag{3.63}$$

  The partition function of the micro-states where RNAP is not bound is:

$$Z_{OFF} = \sum_{(\sigma_1, \cdots, \sigma_L)} W(\sigma_1, \cdots, \sigma_L) \tag{3.64}$$

  The Boltzman weight of a given micro-state when RNAP is bound is: $W(\sigma_1, \cdots, \sigma_L)Q(\sigma_1, \cdots, \sigma_L)$ where the factor Q is the weight contributed by RNAP (its binding with DNA and its interaction with other bound TFs). Q is 0 if one of the repressor site is bound; also assume that when some activators are bound, they can only interact with RNAP independently (one at a time). So:

$$Q(\sigma_1, \cdots, \sigma_L) = q_p \prod_{i=1}^{L} [1 - \sigma_i \delta(\omega_{p,i}, 0)] \left[\omega \sum_{j=1}^{L} \sigma_j \delta(\omega_{p,j}, \omega) \right] \tag{3.65}$$

  Note that if $\sigma_i = 0$ for all i, then the second term in the above equation should be 1.

- Design of transcription logic: (i) change distance between TFs to control TF-TF interaction, e.g. place two sites very distant to turn off cooperativity; (ii) place the TFBS wrt proteins to control TF-RNAP interaction, e.g. place a site inside RNAP promoter to block RNAP access. Examples:
  - AND gate: extra favorable interaction between A and B
  - OR gate: no cooperative interaction between A and B
  - NAND gate: both A and B are inside core promoter, thus either of them could shut down RNAP
- Computation of partition function: consider EQ gate as an example (Fig. 5c). There are 4 subsystems whose weights are: $Q^+_{R_1}$, $Q^-_{R_1}$, $Q^+_{R_2}$ and $Q^-_{R_2}$. In ON micro-states: P is bound, S is not bound, and both R_1 and R_2 could bind (independently); in OFF states: P is not bound (R_1 and R_2 bind independently) or S is bound (P not and S interacts with R_1 and R_2).

Transcriptional control in Drosophila [Reinitz & Sharp, ComPlexUs, 2003; Janssens & Reinitz, NG, 2006]:

- Motivation: earlier evidence shows that eve stripe2 and 3/7 are not completely determined by MSE2 and MSE3. In particular, the 1.7kb upstream sequence can direct stripe 7 at low levels. Want to build a predictive model of 1.7kb (which contains MSE2) that has both stripe2 and low level of stripe7.
- Physical idea: (i) activator binding presents AF (adaptor factor) molecules for RNAP; (ii) binding of activators is modified by direct competition and quenching.
- Model:
  - Basic fractional occupancy of site i by factor a: $K_i\nu^a/(1 + K_i\nu^a)$ for site i, where ν^a is the concentration of factor a.
  - Direct competition of site i (bound by factor a) by factor b binding with site j by factors a and b: $K_i\nu^a/(1 + K_i\nu^a + K_j\nu^b)$.
  - Quenching: a quencher b at position k could reduce the fractional occupancy of site i by activator factor a, $f_{i[a]^A}$, by $1 - q(d_{ik})E_b f^Q_{k[b]}$, where d_{ik} is the distance between i and k, and $q(d)$ is a function of distance. E_b measures the strength of the quencher b.

- Number of AF sites available from bound activator molecules: $N = \sum_a C_a \sum_i F^A_{i[a]}$ where $F^A_{i[a]}$ is the final fractional occupancy of site i by a.
- Number of bound AF molecules: $M = f_{AF}N$, where $f_{AF} = (K_{AF}[AF]^n)/(1 + (K_{AF}[AF]^n))$
- Transcriptional rate: $\approx \exp(-(\Theta - QM))$ for some constants Θ and Q.

- Methods:
  - Expression of 1.7kb sequence: lacZ reporter mRNA level at different stages of the embryo, from C13 to 8 time points in C14: T1-T8.
  - TF expression: from earlier studies [Jaeger & Reinitz, Nature, 2004]
  - Regulatory sequence: upstream 1.7kb, the first 1.1kb contains diffuse binding sites, followed by MSE2. It contains 17 footprinting sites and other sites are predicted using PATSER with varying p-values. For Gt (no PWM), the putative sites are found by matching with known sites.
  - Model fitting: for each TF a, two parameters - the binding affinity K_a, and the strength of activation C_a or repression E_a. For Gt, one affinity per site. In addition, two parameters for basal expression level Θ and R_0. The model is fitted with (expression of all TFs) vs (lazZ expression), a total of 406 data sets, 7 times points with 58 (35% to 92%) AP positions per time point. The root-mean square error of expression is to be minimized.

- Results:
  - Temporal expression pattern of 1.7kb reporter: initially (from C13) a broad anterior pattern, and restricts to stripe2 and 7 at T5, then starts to decline (endogenouse Eve does not decline), almost disappears in T7 and T8 (not used for model training).
  - Reproducing the expression pattern of 1.7kb sequence: 17-site model (only verified BS) vs 34-site model (add predicted sites). 34-site model is better to reproduce two important features: (i) posterior expression; (ii) the correct positioning of the stripe 7.
  - Analysis of stripe 7 expression: finding which sites (and TFs) are important for its expression level and the anterior, posterior boundaries. Two types of analysis are done:
    - Contribution of each site: $C_a F^A_j[m_j, n_j, a]$ for activator site j of TF a or $1-q(d_k)E_b f^Q_{k[m_k,n_k,b]}$ for repressor site k of TF b.
    - Mutations *in silico*: the most important sites have the largest effect on the expression.

    Find that Cad is the activator and Gt, Tll important for the anterior and posterior boundaries respectively.
  - Stripe 2 expression: confirmed the earlier studies. Do various *in silico* mutations: deletion or insertion or enhancerment of individual Bcd or Hb sites; Gt mutants and compare the results with experimental results (only qualitatively, whether a certain mutation increase/decrease the expression level, whether it expand the expression domain).

- Criticisms:
  - All the sites are independent, as if the TF molecules occupy sites simultaneously. If there is cooperativity between two activators, then the binding of one is modified by the binding of the other, and vice visa. So binding of the sites are no longer independent.
  - Quenching: not consider the complexity of TFBS arrangement, e.g. quenching by Q on an A site may be blocked by another site between A and Q sites. Furthermore, the authors claim that quenching by multiple repressors for any activator site is important (the effect is multiplicative), as activator sites are neighbored by about 6 repressor sites on average in 1.7kb sequence. However, there is evidence that a single or few repressor sites are sufficient for suppressing activation, e.g. [Arnosti & Small, Dev, 1996; Gray & Levine, GD, 1996; Kulkarni & Arnosti, Dev, 2003].

- Independent verification (data not used for training) is limited: only qualitative, and for most cases, the effect of enhancing/deletion an activator site should be rather easy to reproduce. For repressor sites, only a single experiment involving Gt mutant.

- **Remark:**
  - Principle of analysis: understand how design → function, in particular, extract important design features, without which the function cannot be correctly implemented. Build model(s) to understand the design and make new and testable predictions/hypothesis.
  - Function: focus on important features of expression pattern including the boundary positions, the relative expression level.
  - Design features: this paper is focused on the important sites and TFs wrt. any sequence. The techniques of analyzing the design include: the comparison of models with different sites, the analysis of contribution of the individual sites according to the model, *in silico* mutations of the sequences/sites or TFs.

Transcriptional activation: cooperativity and synergy [Veitia, Biol. Rev., 2003]:

- Problem: explain the sigmoidal transcriptional response (STR) - cooperativity or synergy?
- Physical idea: both cooperative binding to DNA and synergy from simultaneous interaction of bound TF molecules to BTM can contribute to STR.
- Model:
  - Cooperative binding: let PF_i be the state of promoter where i molecules of F (activator) are bound. Then the probability of PF_i over all promoter configurations is:

$$Y_{PF_i} = \frac{A_i[F]^i}{1 + \sum_i A_i[F]^i} \tag{3.66}$$

    A_i is given by:

$$A_i = \binom{n}{i} \prod_{j=1}^{i} K'_j \tag{3.67}$$

    where K'_j measures the assoication constant of a specific configuration with $j-1$ site becoming a specific configuration with j site. Thus non-cooperative binding means: $K'_1 = K'_2 = \cdots$ and cooperative binding means: $K'_1 < K'_2 < \cdots$.
  - Transcriptional synergy: suppose K_{BTM} is the assoication constant of one bound F with BTM, then the association constant from a configuration of i bound F molecules is K^i_{BTM}. The transcription response is proportional to the total of BTM engaged in complexes with PF_i. Thus:

$$\mathrm{TR} = f\left[\sum_{i=1}^{n} K^i_{\mathrm{BTM}} Y_{PF_i}\right][P_T][BTM] \tag{3.68}$$

- Results:
  - In the absence of cooperativity: could generate switch-like behavior with narrow threshold. Furthermore, with a wide range of K_{BTM}, from 10^3 to 10^{10} M^{-1}, the transcription is dominated by the most populated promoter states.
  - In the presence of cooperativity: sigmoidal transcription. Increase the number of binding sites: shift the threshold to lower [TF] and increase the sharpness of the threshold. However, as n increases, cooperativity mainly affects location, but not threshold width. Furthermore, the cooperativity increase cannot reduce the threshold width when it reaches a certain level.

- Remark: the correction of synergy may be necessary if BTM is touched by only one or a few F molecules at a time. Let $g_i = \Delta G_i^o / \Delta G_1^o$, then in these cases, $g_i < i$. Furthermore, even with "pure" synergy, one may need to have a constant term: $\Delta G_i^o = i\Delta G_1^o + \Delta G^S$, where ΔG^S is a entropic term of the change of freedom of F molecules bound to DNA.

Transcriptional regulation by the numbers [Bintu & Philips, COGD, 2005]:

- Problem: a model that relates expression to [RNAP] and [TF]
- Idea: analyze all possible micro-states of the system and identify which micro-states (class of micro-states) would lead to the desired condition. Apply Boltzman distribution.
- RNAP model: DNA of size N, P molecules of RNAP, assume that binding to the promoter site is specific and all other non-specific binding are treated equally. There are two classes of micro-states: (I) all P RNAP molecules bind to ns sites; (II) P_1 RNAP molecules binding to ns sites and the remaining RNAP molecule bind to the promoter. The total partition function is:

$$Z_{tot}(P) = Z(P) + Z(P-1)e^{-\epsilon_{pd}^S / k_B T} \tag{3.69}$$

  where $Z(P)$ is the partition function of micro-states where P sites bind to ns sites and ϵ_{pd}^S is the energy of specific promoter binding. $Z(P)$ is given by:

$$Z(P) = \binom{N}{P} e^{-P\epsilon_{pd}^{NS} / k_B T} \tag{3.70}$$

  where ϵ_{pd}^{NS} is the energy of ns binding. Plug in $Z(P)$ an $Z(P-1)$, we have the probability of binding is:

$$p = \frac{Z_I}{Z_{II}} = \frac{1}{1 + \frac{N}{P} e^{\Delta\epsilon_{pd}/k_B T}} \tag{3.71}$$

  where $\Delta\epsilon_{pd} = \epsilon_{pd}^S - \epsilon_{pd}^{NS}$.

- A general formula of binding probability: the effect of a regulator to RNAP binding can be captured by a single paramter F_{reg}, the binding probability is:

$$p = \frac{1}{1 + \frac{N}{PF_{reg}} e^{\Delta\epsilon_{pd}/k_B T}} \tag{3.72}$$

- Activator-RNAP model: suppose there are additional A activator molecules, there are four classes of micro-states: (I) all RNAP and activator molecules bind to ns sites; (II) one RNAP molecule binds to promoter and all others to ns sites; (III) one activator molecule binds to activator site and all others to ns sites; (IV) one RNAP and one activator molecules bind to promoter and activator sites and all others to ns sites. Let $Z(P, A)$ be the partition function of the micro-states where P RNAP and A activator molecules bind to ns sites, then:

$$Z_{tot}(P,A) = Z(P,A) + Z(P-1,A)e^{-\epsilon_{pd}^S/k_B T} + Z(P,A-1)e^{-\epsilon_{ad}^S/k_B T} + Z(P-1,A-1)e^{-(\epsilon_{pd}^S + \epsilon_{ad}^S + \epsilon_{pa})/k_B T} \tag{3.73}$$

  And the binding probability is given by:

$$p = \frac{Z_{II} + Z_{IV}}{Z_{tot}} \tag{3.74}$$

  The partition function $Z(P, A)$ can be found through multinomial coefficients: number of ways of dividing N sites into 3 groups: bind with RNAP, bind with activator and no binding. The $F_{reg}(A)$ is given in Equation (8) of the paper.

- Cases: let A be an activator site, R be a repressor site, and P be RNAP site
  - Activation: 4 possible mechanisms
    - simple activation: $A \to P$
    - cooperative activation: the site A_2 helps A_1: $A_2 \to A_1$, $A_1 \to P$
    - dual activator sites: both sites A_1 and A_2 can help RNAP binding: $A_1 \to P$, $A_2 \to P$
    - synergistic acviation: cooperation and dual activation: $A_1 \to P$, $A_2 \to P$, $A_2 \to A_1$
  - Repression: similar to activation
    - simple repression: $R \nrightarrow P$
    - cooperative repression: the site R_2 helps R_1: $R_2 \to R_1$, $R_1 \nrightarrow P$
    - dual repressor sites: both sites R_1 and R_2 can suppress RNAP binding: $R_1 \nrightarrow P$, $R_2 \nrightarrow P$
    - synergistic repression: cooperation and dual repression: $R_1 \nrightarrow P$, $R_2 \nrightarrow P$, $R_2 \to R_1$
- Remark: limitations of the current model
  - Only consider repressor sites that overlap with P
  - No effect of repressor on activator sites

Neurogenic gene expression in fly [Zinzen & Papatsenko, Curr Biol, 2006]:

- Motivation: both rho and vnd are genes expressed in ventral neurogenic ectoderm (vNE), but their expression patterns are slighly different. Explain the patterns and the subtle differences in terms of their enhancer structure.
- Model:
  - Enhancer state (site occupancy): an enhancer contains multiple modules, where each module is a set of tightly linked TFBSs (e.g. a DTS element of linked Dorsal, Twist/Snail sites). The following assumptions are made:
    - A module is active iff it is bound by an activator (for DTS both Dorsal and Twist) and no repressor is bound.
    - An enhancer is active if any of its modules is active.
    - There exist cooperativities between TFs: Dorsal-Twist, Dorsal-Dorsal, Twist-Twist and Snail-Snail.

    The independent modules reflect distance dependencies that are important, for example short-range repression.
  - Kinetic of mRNA: let I_M be the maximal transcription rate from the promoter, k be the strength of the enhancer, P_{Enc} be the probability of the enhancer being active, then:

$$\frac{d[mRNA]}{dt} = sI_M(1 - e^{-kP_{Enc}/I_M}) - r[mRNA] \tag{3.75}$$

    When kP_{Enc}/I_M is small, the rate is mainly determined by P_{Enc}. At steady state, we also have the steady state level is propoortional the the rate, thus $[mRNA] \propto P_{Enc}$.
- Methods:
  - Sequence data: rho, vnd enhancers (about 500bp) from earlier predictions (cluster of motifs).
  - Expression data: rho, vnd DV expression from mRNA; and the protein levels of Dorsal, Twist and Snail are measured by antibodies of these proteins.

- Model training: each TF has an affinity (K), which is the same for all sites of this TF; and cooperativity for some TF pairs. Simultaneous fitting of both rho and vnd expression (same affinity, but each enhancer has its own cooperativity), by maximizing Pearson correlation.

- Results:
  - Arrangent of Dorsal and Twist sites are important for vNE expression: the inversion of a Twi/Sna site in vn abolishes its expression in vNE.
  - The comparison of rho and vnd enhancer sequence and expression: rho is expressed more dorsally than vnd. At the sequence level, vnd has two DTS and more Dl, Twi, Sna sites than rho.
  - Mechanisms of the expression difference between rho and vnd: via changing parameter values, find:
    * Dl-Twi cooperativity: affect dorsal border of rho expression
    * Twi-Twi cooperativity: affect border of rho expression
    * Changing the number of modules: increase and expand expression
    * Sna-Sna cooperativity: affect level of expression in the mesoderm (ventral side)

    This leads to the hypothesis that: Dl-Twi cooperativity is stronger for rho enhancer, which has closer linkage, and thus lead to more dorsal expression. The vnd enhancer has 2 DTS and more sites, thus supposedly higher expression, but this is reduced by the stronger Sna-Sna cooperativity (closer in vnd than in rho).

- Remark:
  - The design features of the system: number and TFBS affinity, TFBS arrangement, and TF gradients. The effect of TFBS arrangement is studied by fitting free cooperativity parameter in different enhancers, and testing how its value is determined by TFBS arrangement.
  - According to the authors, the enhancer activation, i.e., site occupancy, is the rate-limiting step because (i) it is less efficient than interactions between bound TFs and RNAP complex (or cis interactions between enhancers and promoters); and (ii) there are abundant components of the transcriptional machinery.

Enhancer response to antagonistic gradients [Zinzen & Papatsenko, PLoSCB, 2007]:

- Model:
  - One activator site and one repressor site: ON iff the activator is bound, but not the repressor: $Z_{ON} = K_A[A]$, thus
$$p = \frac{Z_{ON}}{Z_{tot}} = \frac{K_A[A]}{(1 + K_A[A])(1 + K_R[R])} \tag{3.76}$$
  - Binding site arrays: ON iff some activator is bound and no repressor is bound. The partition function of a binding site array of M sites for TF X, without cooperativity, is:
$$\Psi_X^M = (1 + K_X[X])^M \tag{3.77}$$
    With cooperativity:
$$\Psi_X^M = 1 + \sum_{k=1}^{M} \binom{M}{k} C_X^{k-1} (K_X[X])^k \tag{3.78}$$
    A special case of cooperativity (all-neighboring model, e.g. from lateral diffusion):
$$\Psi_X^M = 1 + \sum_{k=1}^{M} (M - k + 1) C_X^{k-1} (K_X[X])^k \tag{3.79}$$

If there are M activator sites and N repressor sites, then by our assumption:

$$p = \frac{\Psi_A^M - 1}{\Psi_A^M \Psi_R^N} \tag{3.80}$$

– Direct competition: an array of A/R sites (each site can bind with A or R). It is unclear how cooperativity is affected by mixed states. Simplification: all-neighboring model for activator or repressor:

$$\Psi_{AR}^M = 1 + \sum_{k=1}^{M} (M - k + 1)[C_A^{k-1}(K_A[A])^k + C_R^{k-1}(K_R[R])^k] \tag{3.81}$$

Then:

$$p = \frac{\Psi_A^M - 1}{\Psi_{AR}^M} \tag{3.82}$$

– Multi-module enhancers: for a long enhancer, split into multiple independent modules (where the activation of individual module is given above), then the enhancer is active iff some module is active. Let p_i be the probability of successful state of the i-th module, then:

$$P_{Enc} = 1 - \prod_i (1 - p_i) \tag{3.83}$$

– Short-range repression: in general, modules are not completely independent, one module can suppress another module if they are close. If A is under short-range repression of R, then when R is bound, A will not be completely shut down, instead, its statistical weight will be reduced (comparing with the case when there is no short-range repression) by δ. In other words, the statistical weight associated with the state where both A and R are bound, $K_A[A]K_R[R]$ will be split into: $\delta K_A[A]K_R[R]$ and $(1-\delta)K_A[A]K_R[R]$, where the former is active and the latter not. Example, one activator and one repressor site with short-range repression:

$$p = \frac{K_A[A] + \delta K_R[R]K_A[A]}{(1 + K_A[A])(1 + K_R[R])} \tag{3.84}$$

In particular, one repressor in a module can affect an adjacent module. Consider another example involving 2 modules a and b where a has one activator and one repressor site and b has one activator site. Then without short-range repression, the promoter is active if either a or b is active. The partition function, $\Psi^{ab} = (1 + K_A^a[A])(1 + K_R^a[A])(1 + K_A^b[A])$, and the total weight of all inactive sites is: $\Psi_{off}^{ab} = 1 + K_R^a[R] + K_A^a[A]K_R^a[R]$. So

$$p = \frac{\Psi^{ab} - \Psi_{off}^{ab}}{\Psi^{ab}} \tag{3.85}$$

In the case of short-range repression, some mixed states will be active (fraction δ of the weight of those states). These mixed states include those R is bound in a and A is bound in b:

$$p = \frac{\Psi^{ab} - \Psi_{off}^{ab} - (1-\delta)(K_A^b[A]K_R^a[R] + K_A^a[A]K_A^b[A]K_R^a[R])}{\Psi^{ab}} \tag{3.86}$$

– Direct competition coupled with short-range repression: the effect of repressor is only to compete with the activator for the same site, but it does not affect other activator sites in the enhancer. The mixed states are considered to be successful under this model. For example, an array of M A/R sites without cooperative binding: active if there is at least one occupied activator site (regardless of the repressor sites).

$$p = \frac{(1 + K_A[A] + K_R[R])^M - (1 + K_R[R])^M}{(1 + K_A[A] + K_R[R])^M} \tag{3.87}$$

  - Long-range repression: the repressor bound in one enhancer can shut down the promoter, i.e., all enhancers of this gene. For example, if there are enhancers, each of contain repressor sites, then the promoter is active iff one enhancer is active, and the other enhancer is not repressed (could be empty).

- Remark:
  - The models are primarily qualitative: the enhancer is active if some activator is bound and no repressor is bound. For example, the fact that the different activators may have different strengths is not modeled.
  - The short-range repression model: not fully specified, e.g the general case of an array of modules (what happens to an activator site flanked by two repressor sites, etc.).
  - Inconsistency of the concept of module and short-range repression: within a module, any repressor bounding will turn off the module, and the model allows short-range repression between modules. However, if accepts the idea of short-range repression, then the module length must always be less than the range of short-range repression (100-150bps), but this is clearly not the case.
  - The interpretation of short-range repression model: part of a mixed state is active (with fraction δ). It is not clear whethere δ depends on activator-repressor pair or repressor only since no biochemical interpration is not provided in the paper. One possible nterpretation is: there are two forms of activator, in the presence of repressor, only some portion is active. This is not equivalent to the model where the repressor reduces the activator binding.

Genetic switch in EBV [Werner & Aurell, PRE, 2007]:

- Problem: how genetic switch of the C promoter in EBV is acheived?

- Model: the sequence has two types of binding sites E, for EBNA-1 and O, for Oct-2. Suppose in the state s, the number of E and O sites occupied are: n and k respectively. The number of coopeartive interactions among E sites is n_1, and the number among O sites is k_1. Thus the free energy of the state s is:
$$\Delta G_{n,k,n_1,k_1} = nE_E + kE_O + n1E_{E_1} + k1E_{k_1} \tag{3.88}$$
The number of states s with n, k, n_1, k_1 parameters is denoted as $\xi(n, k, n_1, k_1)$. The transcriptional rate of a state is binary: it is active if eight or more E are bound and inactive otherwise. The total transcriptional rate (or probability of transcription) is the fraction of active states.

- The cooperative interactions among E and O sites: assume that any two adjacent E (or O) molecules can interact (regardless of whether some O site is between). Thus n_1 can take any value from 0 to $n-1$, and similar for k_1. The interaction between E and O sites (steric hindrance): assume one occupied site can block its adjacent site of the other type. Specifically, consider two models: single-blocking: only one neighboring site is blocked; and double-blocking: both neighboring sites are blocked.

- Methods:
  - Data: Activity of the C promoter at different number of binding sites.
  - The cooperativity of the system is defined as the Hill coefficient: $\frac{d\log(P/(1-P))}{d\log[E]}$ at $P = 0.5$, where P is the probability of transcription.

Predicting expression from regulatory sequences in Drosophila [Segal & Gaul, Nature, 2008]:

- Physical idea: expression depends on configurations: the weighted average of expected expression under each configuration. The weight of each configuation is given by its Boltzman weight.

- Model:

- Weight of configuation: let c be a configuration: occupancy state of the CRM (all its sites). Suppose c contains k bound sites and let the TF at position i be $f(i)$, the Boltzman weight of c is:

$$W(C) = \prod_{i=1}^{k} \tau_{f(i)} LR(i) \prod_{i=1}^{k-1} \gamma(i, i+1, p(i+1) - p(i)) \tag{3.89}$$

  where τ_j is the concentration of TF j, $LR(i)$ is the likelihood ratio of site i, and $p(i)$ is the position of i, the function $\gamma(i, j, d)$ represents the cooperativity between site i and j whose distance is d.

- Expression: let E be the expression, then:

$$P(E) = \sum_{c} P(E|c)P(c) \tag{3.90}$$

  where $P(E|c)$ is the expression when the configuation is c, given by a logistic regression:

$$P(E|c) = \text{logit}\left(w_0 + \sum_{i=1}^{k} w_{f(i)}\right) \tag{3.91}$$

  where w_j is the expression contribution of TF j.

- Inference:
  - Compute the partition function $Z = \sum_c W(c)$ by dynamic programming
  - Compute expression: summing over c by sampling c from $P(c) = W(c)/Z$.
  - Parameter fitting: minimize the L^2 error of predicted expression

- Methods:
  - Motifs: motifs from [Schroeder & Gaul, PLoS Biol, 2004], including TorRE.
  - TF expression: from [Myasnikova & Reinitz, Bioinformatics, 2001].
  - CRM sequences and expression: for training, 44 CRMs from [Schroeder & Gaul, PLoS Biol, 2004]; for validation, 11 from [Ochoa-Espinosa & Small, PNAS, 2005] and 15 from [Sinha]. The expressions of both TFs and CRMs are taken at mid-blastoderm: about 20mins into cycle 14.
  - Validation: in addition to the two indepdent data set, also use 10-fold cross-validation (expression at each AP position is classified as ON or OFF with some threshold) and compare the results with those obtained using random PWMs.
  - Factor occupancy: the occupancy of a TF at some site at some AP position is defined as the probability that this site is occupied, marginalizing over all possible configuations c, or equivalently, the sum of probabilities of all configurations where the site is occupied. Computed by DP.

- Results:
  - BS strength: (i) weak BSs contribute significantly, contributing about half of the total factor occupancy; (ii) strong BS only model has low predictability.
  - Homotypic clustering and cooperative binding: (i) sigificant local clustering of BSs of the same factor by testing how often a pair of sites occur in proximity; (ii) (short-range) cooperativity sharpens segmentation: models without cooperativity have lower predictability.
  - Roles of TFs: maternal factors (bcd, cad, torRE) are activators; zygotic gap factors (hb, gt, Kr, kni, tll) are repressors. No context-dependent function found for hb. Verified by the correlation between factor occupancy and the expression (activator sites tend to be occupied at the positions where a CRM is expressed).

- TFBS arrangement: examine the occupancy of sites of different TFs - CRM generally contains one or 2 types of activating inputs (choose among bcd, cad and torRE) and multiple repressive inputs; the choice of activator(s) entails the choice of appropriate repressors.
- Testing affinity-threshold model: no correlation between the bcd affinity (the maxmium occupancy of bcd along any AP position) and the posterior border of target module expression → module expression boundaries seem to be determined as much by repressive gap gene input as by attentuation of maternal activation.
- Sequence-specific competition or occlusion: the sites for different TFs seldom overlap thus, competition or occlusion does not play a major role.

- Remark:
  - Fractional occupancy of a site is a better definition of functionality than binding energy, allowing one to explore site arrangement and the association between a TF and a CRM.
  - Design features: TF roles; strength of sites; site arrangement: homotypic and heteropic clustering (cooperativity), overlapping sites (occlusion), etc.

Gene regulatory function of yeast Pho5 promoter [Kim & O'Shea, Nat Struct Mol Biol, 2008]:

- Goal: the gene regulatory function (GRF) - how promoter output depends on the input [TF]?
- Methods:
  - Pho5 promoter: (Figure 1) TATA box, two binding sites of Pho4 - UASp1 (low affinity) and UASp2 (high affinity); three nucleosomes, TATA box is covered by Nuc-1, UASp2 by Nuc-2 and another Nuc-3.
  - Experimental system: Pho5 promoter is activated by the TF Pho4, which is tagged with YFP; and produce CFP. Thus the GRF is represented by how CFP depends on YFP level. The GRFs of multiple variants of Pho5 promoters are measured, where different variants have different number/strength of Pho4 sites.
  - Characterization of GRFs: maximum expression level, threshold ([Pho4] required to achieve half-maximal level) and sensititive (Hill coeff. of the GRF).
  - Model: the key assumption is Pho4 bound in the adjacent site will help displace Nuc-1 bound to TATA box. Consider the case of only one Pho4 site and TATA box (Figure 4a), the reactions are:
    * Binding of Pho4 to its site: k_{assoc} and k_{dissoc};
    * Association of Nuc to TATA with the Pho4 site not occupied: k_{reass} (the dissociation can be ignored);
    * Association of Nuc to TATA with the Pho4 site occupied: k_{remod} and k_{reass}.

    Thus k_{remod} measures the strength of dissociation helped by Pho4 binding.
  - Model fitting: fit the four dimensionaless parameters with the data: dissociation constants of the Pho4 sites (one exposed, one nucleosomal), the remodeling constant (for displacing nuc.), and the nucleosome reassociation constant.

- Results:
  - Qualitative characterization of GRFs of Pho4 promoter variants: different variants have different maximum expression. Find inverse correlation between maximum expression level and nucleosome occupancy (via ChIP on histone H3).
  - Model performance: (i) fit the data well (via comparing the three parameters of GRFs); (ii) reproduce the qualitative patterns: the threshold is determined mainly by the affinity of the exposed site; the nucl. site has a greater influence on maximum expression than exposed site.

- Discussion: the nucleosome diversify the gene expression profile - with nucl. on only one site.

Thermodynamic model for random promoter library [Gertz & Cohen, Nature, 2009]:

- Goal: predict gene expression from promoter sequence.
- Methods:
  - Synthetic promoters: TFBS combination + basal promoter + reporter gene (GFP) inserted into the yeast genome
  - Expression measurement: each promoter, measure fluorescence of 25,000 cells by flow cytometry. The expression level of a promoter is measured as the fluoresence level per unit cell volume (because cells have different volume, which needs to be normalized), via regression. Also can estimate biological replicate variance (of different cells): 35%.
  - Libraries of synthetic promoters: L1 library - Mig1, Gcr1 sites + spacer; L1-test library: L1 and weak Mig1 sites; etc.
  - Model: based on [Shea & Ackers, JMB, 1985] and [Buchler & Hwa, PNAS, 2003]. Suppose there are two TFBSs, then the interactions between: each TFBS with its TF; TF-TF interaction; and TF-RNAP interaction. Allow negative (unfavorable) interactions, and only adjacent bound TF molecules can interact (no distance dependenc, just interact or not).
  - Model fitting:
    * Regression tool: nlinfit (nonlinear regression) tool of MATLAB. Confidence interval were estimated using nlparci of MATLAB.
    * Treat long (> 2 building blocks) and short promoters differently (two groups have different biological replicates).
    * Parameterization: start with fitting TF-RNAP interaction, then TF-DNA interaction, then TF-TF interactions (with estimated TF-RNAP interaction, but neural TF-DNA interactions, which do not make significant difference).
- Results:
  - L1 analysis (with test libraries): spacer-RNAP interaction; cooperativity between two Mig1 molecules - independently verified by relating number of Mig1 sites and expression (repression). Explain about 40-50% variation of data.
  - L1 weak analysis: (i) weak sites behave as strong sites if there are strong sites nearby because of cooperativity: single strong site $<$ strong + weak $\approx$ 2 strong sites; (ii) 6.7 fold lower affinity of weak site, consistent with PWM analysis, which predicts 9-fold lower affinity. Weak sites are overrepresented in promoters, tend to co-occur with strong sites.
  - Prediction of Mig1 targets: rank promoters by their expression for all genes in yeast. Predict more targets than (i) PWM strong match only; (ii) known targets.
- Discussion:
  - Strong and weak sites together provide better sensitivity to changes in signals, thus a common regulatory mechanism.
- Remark:
  - DNA-TF and TF-RNAP parameters are tightly linked in the model and are hard to deconvolute.

Grammar of short-range repression [Fakhouri & Arnosti, MSB, 2009]:

- Goal: the rules about quenching efficiency, e.g. how it depends on distance.

- Model: only consider the case where there is a single activator (all synthetic constructs share the activators, thus they could be lumped together). Two parts, **F**, the weights of configurations, and **E**, the expression contributions.
  - Weights: the weight of a configuration where adjacent activator and repressor are bound simultaneously is reduced. Suppose there is a single activtor A, the weight will be reduced by $1 - q(R)$ for any adjacent repressor R (multiplicative for multiple repressors), where q is the quenching efficiency, which may depend on distance, etc. Cooperativity between repressor sites are allowed (always the same distance).
  - Expression contribution: only activator contribute.
- Methods:
  - Data: 12 synthetic constructs with 4 activators (Dorsal) and different number and arrangement of repressors (Gt). Measure the expression patterns (expression in neuroectoderm, with gaps created by Gt repressors). Note that some constructs contain 340bp spacer sequence between binding sites and the promoter, but they show identical expression patterns (with those without spacers), so they are not included in the analysis.
  - Model: assume distance dependent quenching efficiency and put distance into bins (e.g. all distances within 10bp range - the same quenching efficiency).
- Results:
  - The effect of cooperativity: optimal value at about 4. Not clear how significant it is.
  - The quenching efficiency: not monotonic to distance. Possible phasing effect.
- Remark:
  - Not a general framework for working on an arbitrary number of sites: single activator site, all repressor sites have the same affinity, quenching efficiency is not regularized (should be a relatively smooth function), cooperativity is defined on a fixed distance.
  - Interpretation of results: very sparse data points (12 different distance values with each used probably a few times), not clear how reliable the estimated parameters are.
  - Hypothesis about the quenching efficiency: nucleosome structure (only close in 3D - dependent on nucleosomes, can Gt interacts with corepressor).

Optimization of thermodynamic models [Bauer & Bailey, Bioinformatics, 2009]:

- Aim: choose the optimization method for thermodynamic models.
- Methods:
  - Thermodynamic model: [Reinitz03] model, minimizing RMSE.
  - Simulated annealing (SA): simple geometric cooling or LAM cooling; neighborhood function (proposal function to sample from the neighborhood) - uniform sampling from a certain radius within the constrained interval.
  - Gradient descent (GD): multiple random starts; stopping condition is defined by the change of parameters or the change of function values; search in constrained space by introducing transformation on the parameters.
  - Experimental data: (i) [Janssens06] data; (ii) [Segal08] data, with 10 CRMs (randomly chosen) where predictions are GOOD or FAIR; (iii) synthetic data from [Janssens06], generated from the correct model using the estimated parameter values.
  - Evaluation: 5-fold cross validation, Pearson correlation (CC) as the metric.

- Results:
  - SA with simple geometric cooling schedue is as good as LAM schedule.
  - SA vs GD: SA is significantly better. In [Segal08] data, CC under SA reaches about 0.65, while under GD only about 0.5.
  - The optimization landscape: extremely flat near the true solution (synthetic data), s.t. GD is not able to converge to the true solution even with small perturbations. Also the landscape contains many local minimum, so that GD never converges to the true solution with 100 runs (47 cases to the trivial minium where the expression is 0).

Daul role of TFs in segmentation [Bauer & Bailey, 2009]:

- Problem: do some TFs have dual role, i.e. activator in some CRMs and repressor in others in segmentation network?
- Outline: if the hypothesis is true, the model allowing dual roles will show significant improvement of data fitting. Explore the mechanism of dual role: SUMOylation.
- Methods:
  - Reinitz model: 18 parameters, each TF has two parameters - binding and transcription; and 2 other parameters: basal transcription and saturation.
  - Segal model: each TF has three parameters - binding, transcription and cooperativity; each CRM has a parameter - basal transcription; motif PWMs are also free parameters. Total 344 parameters.
  - Role determination: under Reinitz model, for each CRM, try different configuratoins (i.e. each TF could be an activator or repressor) and choose the best one (some variations of this scheme is used). If a TF is assigned as an activator in 2/3 CRMs, then it is a solid activator; same for repressor; otherwise, it will be assigned a switching role.
- Results:
  - TF roles: Cad as solid activator; Tll and Kni as solid repressors; and the rest switching.
  - Performance improvement with dual role for Kr and Hb: improve correlation from 0.25 to 0.35 and 0.37.
  - Correlation of prediction with SUMOylation: only those factors predicted as switching roles have SUMOylation site in their protein sequences. And SUMO is known to be a protein that modifies the role of a TF.
- Criticism: the evidence is weak in some aspects:
  - Exclusion of the contribution of other TFs: the authors show that the other motifs are not overrepresented in the CRMs where the TFs show switching roles. However, there are potentially a large number of TFs with uncharacterized motifs; and the overrepresentation test generally tends to be conservation.
  - Performance improvement with dual role model: the improvement involves not just a few parameters, but also the configurations (for each CRM, the role of a TF) are free parameters.
  - The cutoff by 2/3 to assign TF roles is arbitrary.
- Remark: the question of why a TF acts in different roles in different enhancers is not addressed. What is the rule?

Thermodynamic model of gene regulation fro the Or59b olfactory receptor in Drosophila [Gonzalez and Altafini, review for PLCB, 2018]

- Motivation: each olfactory neuron expresses only one class of OR. How is this achieved?
- Or59b CRE: Figure 1B. Two Hox sites bound by two co-activators, and one Pou site bound by one of the two (another subunit). E-box: bound by Fer1, which recruits RNP.
- Notations: two co-activators A (Acj6) and B (Pdm3) and subunits A1, A2, B1, B2. Activator Fer1: C. RNA Pol II: R. The system has 48 possible configurations: two Hox sites of A and B (2 x 2), Pou sites (A, B or not bound), Ebox (x) and Pol 2 binding (2).
- Model of normal state: (1) Positive cooperativity between A1 and A2; between B1 and B2. (2) Positive cooperativity between C and R. (3) Negative coop. between A1 and B2 and A2 and B1: short-range repression. (4) Epigenetic effect: when A or B binds Pou, the site E-box becomes accessible to C. Model this as: let q_C be the weight of Fer1 binding to Ebox, we multiply this by h_1 (small) when none of A or B is bound to Pou (do not multiply h_1 when Pou is occupied).
- Model of mutant state: change chromatin state to make it more open. (1) Effect on E-box binding by C: when Pou is bound, no difference; when Pou is not bound, since chromatin is more open now, we have larger h_1 (> 1). (2) Interaction of binding of A and B and C binding: change $w_{A_1A_2}$ and $w_{B_1B_2}$ by a constant h_A or h_B (Pdm6, but not Acj6, double binding inhibits Fer1 binding to E-box)
- Experimental data: Table 1, expression level (categories) under different mutational states of TFBSs, and of Suv39.
- Fitting the model: using normal state and Suv39 heterozygous mutant. Do sampling of parameters, and hence prob. of RNP occupancy. Validation using Suv39 knockout, and different TF concentrations.
- What we learned? Model parameters Figure S3 and Table S3 (epigenetic). Main and somewhat unexpected findings: (1) Sun39 mutant: Ebox1 site is generally sufficient to drive expression (E6, E14). (2) Pdm3 double binding may hinder Fer1 binding to Ebox (state E8), likely due to spatial competition. Acj6 double binding no such effect (E12). These are captured by the parameters h_A and h_B. (3) W.t. sequence: expression in heterozyg. mutant is not consistent (E16).
- Summary: modeling of epigenetic effects (1) Pou binding by A or B makes E-box open: parameter h_1 that is small when A or B not bound, but 1 if Pou is bound. Also h_1 changes with Suv39 mutant (higher). (2) Spatial competition of B double binding on Fer1 binding to Ebox (only when chromatin becomes very open): multiply h_A or h_B to cooperativity parameters $w_{A_1A_2}$ and $w_{B_1B_2}$.

3.7.2 Design principle of promoters and enhancers

Specificity and robustness of TF binding [Sengupta & Shraiman, PNAS, 2002]:

- Model:
  - Specific binding of TF: the binding energy of TF with a sequence x is: $E(x) = \epsilon \cdot x$ where the vector ϵ represents site-specific binding energy. The probability of binding $p(E)$ is $[1 + \exp((E - \mu)/k_BT)]^{-1}$ could be approximated by a step function: $p(E) \approx 0$ if $E(x) < \mu$.
- Inference:
  - Estimation of ϵ and μ and prediction of TFBS: given known sites $x_1, \cdots, x_n$, find ϵ and μ with $\|\epsilon\| = 1$ and maximum μ^2 s.t. $\epsilon \cdot x_i \leq \mu < 0$.

Plasticity of cis-regulatory input function [Mayo & Alon, PLoS Biol, 2006]:

- Problem: given a cis-regulatory sequence, when the sequence changes, how its cis-regulatory input function (CRIF) changes? Specific questions include, e.g., small mutations always lead to small changes of CRIF? Can all possible CRIFs be reached from point mutations?

- Model:
  - Characterizing CRIFs of lac promoter: the promoter is a function of two ligands, [cAMP] and [IPTG]. To simplify the analysis of CRIFs, represent a CRIF by four plateaus corresponding to the saturating levels of two ligands: I - neither of them is saturating; II, III - one of them is saturating; IV - both are saturating. Or, use the three ratios: I/IV, II/IV, III/IV, as a representation of CRIF.
  - CRIF function from thermodynamics: similar to [Buchler03] model, the promoter is active if RNAP is bound or both RNAP and CRP are bound (they will interact). The set of all configurations also include: none of RNAP, CRP, lacI is bound; only CRP is bound; and only lacI is bound. In addition, the level of active CRP is a Hill function of [cAMP], and the level of inactive lacI is a Hill function of [IPTG].
- Results:
  - Random point mutations on lacI, CRP and RNAP binding sites, then map their CRIF. Could generate a number of CRIFs, corresponding to Boolean AND, Boolean OR, and intermediate functions between AND and OR.
  - Theoretically map parameter space (binding affinity to lacI, CRP and RNAP) and phenotype space (defined by the three pamameters) by using the promoter function. The main features are:
    * Uniformly spaced parameters lead to non-uniform distribution in phenotype space, i.e. some phenotypes are simply more likely than other ones, in particular, some CRIFs cannot be reached via point mutations in the three sites.
    * The wild-type CRIF appears to reside in a region where mutations could easily lead to new functions, thus plastic in the sense that many mutations do not ruin the input function completely, but rather result in a new computation.
- **Remark:**
  - A major problem is to understand the sequence-function relationship of regulatory sequences, or in general, the structure-function relationship of (complex) biological systems/networks. Ex. whether small changes of structure always lead to small changes of function; how easy it is to evolve new functions; etc.
  - The theoretical models can help address these problems: by changing parameters in the model (or simulating sequences/systems), and predicting behavior of the new sequences/systems. In simple cases, this is just bifurcation analysis of dynamic systems.

Bacterial promoter architecture [Hermsen & ten Wolde, PLCB, 2006]:

- Problem: baceterial promoters often have complex organizations: homotpyic arrays of the same sites; overlapping sites; etc. Why?
- Idea: in silico evolution of promoters, and observe the features of the resulting sequences.
- Methods:
  - Model of cis-regulatory function: basically [Buchler03] model.
    * Cooperativity: between any two TF molecules (could be of different types) bound within 3bp.
    * Activation: only the site nearest to the core promoter will interact with RNAP, called primary (activator) site. The other sites are called auxiliary sites.
    * Repression: steric hinderance of RNAP. The site that overlap with RNAP is the primary repressor site.
  - Evolutionary algorithm of selecting promoters:

    - The goal function: different logic gates.
    - Evolution: 500 individuals with random starts (both promoter sequences and amino-acids); fitness is defined as the least square error between outputs and goals (over all the range of TF inputs); at each step, 20% top ones are selected and rest removed and replaced by copying from randomly chosen promoters from the selected ones.

- Results:
  - Homo-cooperativity: for activators, auxilary sites are stronger than primary sites to achieve steep response; for repressors, the opposite case.
  - Hetero-cooperativity: used whenever a response should be conditional on the presence of more than one TF. Example: AND gate.
  - Competition between modules: provides more complex responses. Ex. NOR gate (Figure 4): when both TF1 and TF2 are low: the strong TF1, or TF2 site is occupied (without cooperative binding with other sites), and activates RNAP; when either of them is high: it binds with the homotypic array of sites, which overlap with RNAP.

- Remark: 3bp for cooperativities may be too constrained. The design of NOR gate depends on this (or generally, the fact that a TF can be both an activator and repressor): a high affinity site vs an array of low-affinity homo-cooperative sites.

How to tune an enhancer [PNAS, 2016]

- Experimental system: library of variations of Otx-a, a 69-bp neural enhancer of the Ciona Otx gene.
- Altering the spacing between ETS and ZicL sites by as little as 1 bp, or reversing the orientation of a single binding site can strongly affect expression.
- The tissue-restricted enhancers have either high-affinity ETS/ZicL sites or optimal spacing, but not both: Optimizing both affinity and syntax causes ectopic expression.

3.7.3 Gene Expression by Epigenomics

Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. [Front Cell Dev Biol. 2014]

- Motivation: given ChIP-seq data of TF binding and/or histone modification (HM) around genes, can we predict the expression of a gene?
- Model: we have data in one cell type, for each gene, we have its expression (y), and the explanatory variables, TF association strength, HM, etc. Usually focus on region near TSS, e.g. some weighted sum of TF binding (higher weight if close to TSS). Then we train the model using many genes.
- The difference in models lie in how the features are defined (e.g. weighting according to TSS), whether to use features in the gene body, and so on.

Correlation of TF binding and expression [Liu & Clarke, JMB, 2002]:

- Problem: can expression be predicted from TF occupancy?
- Methods:
  - Occupancy of TF: suppose there are n sites in a sequence, the occupancy of TF can be defined as the probability that at least one site is occupied: $P = 1 - \prod_i (1 - \theta_i)$ where θ_i is the fractional occupancy of site i, computed from the predicted K_d of this site and the TF concentration.

  - Correlation with expression: classfy all genes as regulated by the TF or not from expression data. Then predict target by using the predicted TF occupancy.

- Results:
  - The TF occupancy is a better predictor of regulatory target than the number of binding sites defined by threshold.

ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells [Ouyang & Wong, PNAS, 2009]:

- Goal: predict absolute expression level in ES cells, from the ChIP-seq binding of multiple TFs. Two specific challenges:
  - How to assign ChIP-seq peaks to target genes?
  - Linear model assumes additivity of TF binding; however, multiple TFs may work synergistically. Ex. a factor A may work either with B or C, when A-B is bound, it achieves activation; when A-C is bound, it achieves repression.

- TF association strength (TFAS): assigning ChIP-seq peaks to its nearest genes is problematic. Use a quantitative (partial) assignment, TFAS, and show that it improves the correlation between binding profile and expression.

- Principal component (PC) regression: the observation is that multiple TFs often bind to the same regions, thus a significant correlation among the TF binding variables. Use PCA, and each principle compoment represents one pattern of co-binding. There could be multiple co-binding patterns (i.e. PCs), and the expression is a linear function of PCs.

- **Remark**: when the variables have complex dependency/synergistic interactions, we could model the underlying combinatorial patterns as latent variables; and then inference (regression) can be performed on these latent variables. In particular, this could model the variables which could have multiple modes of action on the response variable (e.g. either increase or decrease, depending on the context).

Predicting CRM expression from TF binding [Zinzen & Furlong, Nature, 2009]:

- Problem: one can obtain genomewide TF binding profiles through ChIP-chip or ChIP-seq, how to map these data to expression patterns?

- Idea: expression patterns (categoried, instead of numbers in spatial-temporal scale) are functions of binding affinties to multiple relevant TFs. Build a classifier from these features.

- Methods:
  - ChIP-chip data of TFs in mesoderm development (muscle differentiation): Twi, Tin, Mef2, Bap and Bin. Five profiles at each stage, measure five stages.
  - CRMs: 200 bp around ChIP-chip peaks with additional processing (combine clusters of peaks). Average length 270 bp, about half are multi-peak regions bound by multiple TFs.
  - Binding-expression mapping data: (1) expression data: CRM activitity database (CAD), from REDFly and this experiment; (2) choose 310 ChIP-CRMs for which CAD data is available as training data, i.e. they overlapped with one of the CAD enhancers (ChIP-peak heights as TF occupancy/features). Furthermore, expression classes are manually defined: five patterns - three single-tissue classes, and two complex classes containing two expression domains.
  - Testing data: 35 ChIP-CRMs are assayed for expression patterns, by transgenic reporter assay.

  - Classification problem: train SVM for each of the five expression classes. The SVM uses Gaussian radial basis function (RBF) kernel. The hyperparameters are: the penalization coefficient (C) and the kernel precision (γ), these are chosen by the LOOCV performance.

- Results:
  - Prediction accuracy: out of 35 CRMs, 25 were correct, 5 partially correct and 5 failed. Many errors in the somatic muscle class, probably reflecting the lack of other specific regulators.
  - Many different TF occupancy could lead to similar expression patterns, in terms of: the type of TFs, the duration of binding and the intensity of the ChIP signal.

- Remark:
  - Classification algorithm: (1) CRMs have different length, normalize the ChIP-signals by length? (2) Biological interpretation of SVM classifiers: SVM with RBF kernel correpsonds roughly to the constraints on the TF occupancy (in a range) - the decision surface is a hypersphere.
  - The success depends on: (1) high-quality CRMs and ChIP-chip signals; (2) expression classes are chosen s.t. they likely represent basic building blocks of complex spatial patterns, or common patterns.

Histone modification levels are predictive for gene expression [Vingeron, PNAS, 2010]

- Motivation: build a model of histone modifications (HM) at promoters to gene expression. From the model, learn which HM are important, and how general the model is (can we apply the model learned from one cell type to a different one).

- Data: 38 histone marks, 1 histone variant, and 2 controls (non-specific antibody ChIP-seq) in CD4+ T cells. Microarray or RNA-seq experssion data. Note: controls correlate with the open chromatin and can be relevant covariates.
  - Histone mark data: use tag counts in 4K promoter regions as independent variables, do log-transformation.
  - Expression data: use log-transformed microarray expression values.
  - Remark: the model is fit over all genes in one cell type, so no need of normalization by library size (cell types).

- Model: regression model. Use 10-fold cross validation to assess model performance.

- Main results: high correlation between HM and expression. Best single-HM model: H3K27ac, $r = 0.72$. Best three-HM model (H3K27ac, H3K4me1, H4K20me1), $r = .75$. The best three or four features perform almost as good as the full model.

- The best features differ between two types of promoters: high-GC promoters (HGP, most promoters) and low-GC promoters (LGP).
  - HGP: H3K27ac (TSS), H4K20me1 (TSS and gene body).
  - LGP: H3K4me3 (TSS), H3K79me1 (gene body).
  - Remark: the HMs enriched in TSS are associated with transcription inititation, while the HMs enriched in gene body are associated with transcription elongation.

- The model fit in CD4+ T cells performs well in a different cell type (other T cells), even on genes with diff. expression levels.

- Lessons:

- Data processing step: when fitting the model across genes, it is important to normalize by genes (e.g. same promoter length); when fitting across cells, normalize by cells (e.g. library size). A general idea is to use quantile normalization.
- Statistical model: may consider a model that takes biology into account, e.g. different types of promoters have different models.
- Different HMs may measure different steps of transcription (init., elongation, etc.). For promoters, the key HM may be H3K27ac and H4K20me1.

CHROMIA: Genome-wide prediction of transcription factor binding sites using an integrated model [Won & Wang, GB, 2010]:

- Model: Chromia integrates continuous (eight histone modifications) and discrete data (DNA sequence) in its model. It converts the discrete sequence data to continuous PSSM score signals. The binned histone modification and PSSM score are used as an input to the HMMs.
- Model training: regions centered at TSS and p300 binding sites were selected to train HMMs for promoters and enhancers, respectively.

Bayesian network analysis of targeting interactions in chromatin [van Steensel & Ideker, GR, 2010]:

- Three types of proteins in the chromatin: (1) histones; (2) DNA binding factors (DBFs), which typically recognize specific sequence motifs, and (3) proteins that do not contact DNA directly, but interact with DNA via other proteins, which we will refer to as chromatin proteins.
- Motivation: comparison of the binding maps of multiple proteins can be informative. For example, if two proteins have highly similar distributions, this may indicate that the two proteins share a common targeting mechanism, or that one protein recruits the other. Conversely, mutually exclusive distributions suggest that the two proteins may be targeted by different, incompatible mechanisms, or that one protein prevents the other protein from binding.
- Data: maps of 43 broadly selected chromatin components in Drosophila. Six histone modifications and one histone variant (H3.3), nine DBFs, and 26 chromatin proteins, such as classic heterochromatin proteins, Polycomb Group (PcG) proteins, nucleosome remodeling factors, high mobility group proteins, histone modifying enzymes, cofactors, and several other types.
- Targeting interaction $X \to Y$: defined as an interaction between two chromatin components X and Y, such that the presence of X at a specific set of genomic loci promotes the association of Y with these loci.
- Results: uncover distinct mechanisms that determine the genomic binding patterns of the heterochromatin components HP3 (also known as LHR) and SU(VAR)3-7, and we demonstrate that the nucleosome remodeling protein Brahma (BRM) has a central role in the targeting of various DBFs.

Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development [Kaplan & Eisen, PLG, 2011]:

- Modeling TF binding: generalized hidden Markov models to infer the occupancies of one or more transcription factors across any DNA sequence given the concentration of these factors and their DNA binding preferences. Basic model leads to correlation coefficient about 0.4 (prediction vs. in vivo binding). Adding competition made little difference.
- Measure the influence of chromatin state:
  - Modeling nucleosome occupancy: integrated the exact position of nucleosomes into our model at single nucleotide resolution to enable the competition between transcription factors and nucleosomes in binding DNA to be modeled. Nucleosome model by (1) a sequence-specific model

of nucleosome binding, that takes into account presumed preferences for certain DNA sequence feature; (2) a sequence-independent model of nucleosome binding, where nucleosome are viewed as long space fillers.

  - Direct genome-wide measurements of DNA accessibility obtained from DNase I digestion of chromatin in isolated blastoderm embryo nuclei.

- Incoporating chromatin accessibilty in generalized HMM: a non-uniform prior probability of regulatory binding along the genome. Nucleosome modeling adds no benefit, while DNase I accessibility improves correlation to 0.6-0.9.

- Cooperative TF interactions: gHMM have limited ability to model the broader context of binding events, thus added a second, sampling-based, phase to our computational model. However, these cooperativity parameters only improved the predictive power of the model by $< 5\%$.

CENTIPEDE: Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data [Pique-Regi & Pritchard, GR, 2011]:

- Motivation: predict TFBSs, by integrateing cell- or tissue-specific experimental data such as histone modifications and DNase I cleavage patterns with genomic information such as gene annotation and evolutionary conservation.

- Model: for any motif match i, we need to decide if it is a true TFBS or not. Let G be the sequence information, including conservation, distance to TSS and motif score, and Z be the TFBS indicator (1 if yes, 0 otherwise), and D be the tissue-specific experiment data (histone or DNase data). The basic model is: $G \to Z \to D$. The first part of the model, the prior probability $Z = 1$ given G is:

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 \cdot \text{PWM score} + \beta_2 \cdot \text{Cons. score} + \beta_3 \cdot \text{TSS distance} \tag{3.92}$$

  where $\pi_i = P(Z_i = 1|G_i)$. The second part of the model: first we assume that if there are multiple experimental measurements (e.g. histone and DNase), they are conditionally independent given Z_i. For some type of data (e.g. histone), only the tag counts in the region around the motif match (200 bp) matter, let it be R_i, and we have:

$$P(R_i|Z_i = k) = \text{NegativeBinomial}(\alpha_k, \tau_k) \qquad k = 1, 0 \tag{3.93}$$

  For other type of data (DNase), the actual position of the tag matters (Figure 4 and S8), as different TFs may have distinct spatial pattern of hypersensitive sites (e.g. TF binding may protect DNA from DNase action). Let $X_i = (X_{i1}, \cdots, X_{iS})$ be the tag count (usually 0 or 1) at every position near the putative site (S is the window size). Then we model X_i as a multinomial distribution:

$$P(X_i|R_i, Z_i = 1) = \text{Multinomial}(R_i; \lambda_1, \cdots, \lambda_S) \tag{3.94}$$

  For the case $Z_i = 0$, we assume a multinomial distribution with uniform probabilities: $\lambda_s = 1/S$.

- Inference: unsupervised training for each TF, essentially a mixture model of Z_i. Parameter estimation is done by EM algorithm, summing over Z_i's. Once the parameters are estimated, infer Z_i conditioned on G_i and D_i.

- Validation and analysis: use ChIP-seq data of 6 TFs in LCL as the test dataset. Using prior genomic data plus DNase-seq data obtain mean AUC = 98.11%. The histone data are informative but do not provide additional predictive power for TF binding when DNase I data are included in the model.

- Large-scale predictions: a genome-wide map of 827,000 transcription factor binding sites in human lymphoblastoid cell lines, which is comprised of sites corresponding to 239 position weight matrices of known transcription factor binding motifs, and 49 novel sequence motifs.

- Remark:
  - The conceptual problem of CENTIPEDE model is that TF binding leads to tissue-specific pattern of DNase (or histone data). In [Chromatin accessibility pre-determines glucocorticoid receptor binding patterns, NG, 2011], the opposite was found to be true: the DNase pattern predetermines TF binding.
  - However, the spatial pattern of DNase tag counts around TFBSs is likely to be caused by TF binding.
  - Statistical concerns: the PWM score only influences the prior, so the specificity (wrt a particular TF) is questionable.

FIMO: Epigenetic priors for identifying active transcription factor binding sites [Partida & Bailey, Bioinfo, 2012]:

- Model: opposite of the CENTIPEDE model, where the epigenetic data are treated as prior, and the sequence information is the data (generated from the motif). Let z_i be the TFBS state of the i-th motif match, and y_i be the epigenetic data (tag count at i), we have $P(z_i|y_i)$ as a monotoically increasing function, linear function in the paper (Equation 1). Let x_i be the sequence of the motif match, then $P(x_i|z_i)$ is modeled as a typical product-multinomial distribution. To infer z_i, we have:

$$\log \frac{P(z_i = 1|y_i, x_i)}{P(z_i = 1|y_i, x_i)} = \log \frac{P(z_i = 1|y_i)}{P(z_i = 0|y_i)} + \log \frac{P(x_i|z_i = 1)}{P(x_i|z_i = 0)} \tag{3.95}$$

  where the first term is the log prior-odds, and the second the usual log-likelihood ratio score of sites.

- Inference: note that the model involves only one parameter β, which is used for the prior distribution. β can be interpreted as the number of TFBSs in the genome, and set as 10K in the experiments.

- Evaluating TFBS prediction methods: in the ChIP-seq data of six TFs in LCL, the gold standard data, only the putative sites with PWM score above a cutoff is accepted. Multiple methods are compared: PWM score only, Histone as prior, DNase as prior, and CENTIPEDE (Table 1).
  - DNase prior and histone prior significantly better than PWM score alone; and DNase prior better between the two priors.
  - CENTIPEDE outperforms in most cases with AUC measure, in particular, significantly better than other methods with sensitivity at 1$ FP rate.
  - Surprisingly, the simple method of ranking all putative sites (filter by PWM scores first) by DNase scores outperforms CENTIPEDE.

3.8 Chromatin Looping and Enhancer-Promoter Interactions

Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data [Dekker & Mirny, NRG, 2013]

- Background: cross-linking, bond that links one polymer chain to another. In proteins, the most common type of cross-linking is disulfate bonds. Some small molecules can be used as crosslinking agent, e.g. HCHO can link two proteins/DNAs when they are physically close, forming a cross-link -CH2- called a methylene bridge: `http://publish.uwo.ca/~jkiernan/formglut.htm`

- Different types of technologies:
  - 3C: yields a long-range interaction profile of a selected gene promoter or other genomic element of interest versus surrounding chromatin

  - 4C: generates a genome-wide interaction profile for a single locus
  - 5C: (many-by-many) combines 3C with hybrid capture approaches to identify up to millions of interactions in parallel between two large sets of loci.
  - Hi-C: (all-by-all) not anchored on a single locus of interest but instead generate matrices of interaction frequencies that can be represented as 2D heat maps with genomic positions along the two axes.
  - Resolutions: (1) 5C: higher resolution, could be 4kb. (2) Hi-C: 100kb resolution with several hundred million read pairs. To increase the resolution by a factor of n, one must increase the number of reads by a factor of n^2. Note: in recent papers from Bing Ren lab, the resolution is 5-10kb.

- Hi-C [Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Lieberman-Aiden & Dekker, Science, 2009]. Method: Figure 1.
  - Cross-linking of DNA (chromatin segments) in close proximity.
  - Cut with restriction enzyme (or sonication)
  - Fill ends with biotin
  - Ligation: form a hybrid DNA molecule (circular)
  - Purify and shear DNA, pull down biotin. Then a DNA segment containing biotin is a hybrid DNA: each part from one of the two interacting chromatin segements.
  - Paired-end sequencing of the pulled down DNA: identify the junctions.

- Overview of computational approaches:
  - Identify pairs or sets of loci that interact more frequently than would otherwise be expected, which points to chromatin looping or specific co-location events
  - Restraint-based modelling and approaches that model chromatin as a polymer: use all of the interaction data to build ensembles of spatial models of chromosomes.

- Interpreting data from 3C-based technologies: what does "chromosome interaction" means? The data report on the relative frequency in the cell population by which two loci are in close spatial proximity. These could be due to a number of reasons
  - Direct and specific contacts between two loci, mediated by protein complexes.
  - Indirect co-localization of pairs of loci to the same subnuclear structure, such as the nuclear lamina, nucleolus or transcription factory.
  - Nonspecific result of the packing and folding of the chromatin fibre, as determined by other (nearby) specific long-range interactions or other constraints.
  - Random (nonspecific) collisions in the crowded nucleus.

- Studying chromatin looping and linking regulatory elements to target genes.
  - Some examples of interactions of specific loci studied by 3C: CFTR locus, beta globin and LCR, MYC gene.
  - Approach to detecting chr. interactions: (1) The baseline of interaction frequencies: obtained from the entire dataset. This led to an estimate for the baseline interaction frequency for each genomic distance. (2) Looping interactions were then identified by detection of signals that are significantly higher than this baseline.

  - Topologically associating domains (TADs): a prominant feature of chromesome organization. Many interactions within a TAD, and relatively few interactions between different TADs. Ex. high-resolution 5C map of X-chromosome inactivation centre reveals a series of TADs. Median size of TADA around 400-500 kb.
  - Featurs and possible mechanism of TADs: generally invariant across cells, can be transcriptionally active or inactive, and the boundary encoded by boundary elements. Boundary regions are enriched in CTCF-bound loci.

- Some insights about gene regulation from the studies of chromatin looping/interactions:
  - One abundant class of long-range interactions involves promoters looping to sites bound by the insulator protein CTCF.

In search of the determinants of enhancer-promoter interaction specificity [van Arensbergen & Bussemaker, Trends Cell Biol, 2014]

- Mapping physical interactions:
  - Active promoters were found, on average, to contact 4-5 enhancer-like elements. The majority of these elements are located within 500 kb from the interacting promoter, with an estimated median distance of about 125 kb. Only a fraction of looping distal elements contact the nearest promoter: 27-60%.
  - Active enhancers were found to contact approximately two promoters on average. Note: earlier studies show that enhancers can control multiple neighboring genes.
  - ChIA-PET survey of loci bound by RNA polymerase II found that promoters often interact with other promoters; some of these promoters are able to act as enhancers for their partner promoter.
  - Enhancer-promoter interactions may be highly tissue-specific in human cells. However, in a 4C study of 100 fruit fly enhancers, the interactions appear to be stable between mesoderm and whole embryos at two stages. Treatment of human fibroblast cells with TNF-a also left the enhancer-promoter contacts largely unaltered even though 800 genes showed differential expression.
- Hierarchy of enhancer-promoter interactions: indirect contact, direct contact, functional contact.
- Caveats of the physical interaction data:
  - Due to the size of the DNA fragments analyzed (4 kb on average), elements annotated as promotersare in fact promoters plus several kb of flanking DNA.
  - A sizeable fraction of the encounters reported by the current methods reflect indirect contacts: e.g. all belong to a large “transcription factory”.
- Correlation between enhancers and promoters:
  - The ENCODE study: correlation between DHSs at one site and DHS at TSS, across 79 cell types. A modest fraction (4%) of these overlapped with physical interactions identified using 5C or ChIA-PET, suggesting that correlations can arise in multiple ways, including through indirect and noncausal associations. Similar approach: the chromatin state of enhancers was correlated with that of promoters.
  - FANTOM study: atlas of eRNAs in 800 tissues. Promoters were linked to approximately five enhancers, and enhancers were associated with approximately two promoters. Among the inferred intarctions, 21% were supported by physical contacts based on ChIA-PET.
- Mechanism 1: biochemical compatiability

  - Ex. Drosophila gsb and gsbn genes, the enhancers are located in a 10kb region between the genes, yet the activations by enhancers are specific. Possible mechanism: TADA box or DPE (downstream promoter elements) in the core promoter may determine enhancer specificity. For example, Caudal binding prefers DPE to TADA-containing promoter.
  - Note however that there is a competition between promoters/enhancers, so the preference does not reflect absolute compatability or not. Ex. IAB5 enhancer preferentially activated a TATA-containing promoter but could activate a DPE-containing promoter if no TATA promoter was available.
  - Mechanism of IAB5 enhancer specificity: selective activation of AbdB by IAB5 depends on a 255-bp element in the proximal promoter of AdbB (no TADA box).
  - Enhancers near ribosomal protein genes are compatible with TCT-containing promoters, but not with other promoters (TADA or DPE).

- Mechanism 2: chromosome architecture
  - Chromosome architecture generally favors close genes. In a fly experiment, most enhancers regulate neighboring genes [Start group, Nature, 2014].
  - Long-range interactions may require assistance from architectural proteins, such as CTCF. The combination of CTCF and cohesin is preferentially found at genomic locations that exhibit constitutive (cell-type invariant) looping contacts. Other players include mediator complex, ncRNAs, which can determine specificity of enhancer-promoter interactions.
  - Topologically associated domains (TADs) represent another, albeit related, architectural feature that can help direct enhancers to the right target promoters.

- Mechanism 3: insulator elements
  - Insulator elements: DNA elements that can prevent the activation of a promoter by an enhancer when placed between them. In Drosophila, five distinct insulator complexes have been identified; in vertebrate, mainly CTCF.
  - The looping model of insulator: two or more insulator sites physically interact with each other, thereby establishing loops that alter the 3D conformation of the chromatin fiber in a manner that affects the ability of enhancers to interact with promoters. Evidence: contact between CTCF sites.
  - The decoy model, insulators interact directly with an enhancer or promoter, thereby interfering with enhancer-promoter encounters.

- Mechanism 4: Local chromatin composition
  - Histone modification patterns (e.g. H3k27ac or H3K4me1) and p300 binding: may depend on history of the local chromatin.
  - Lamina-associated domains (LADs): over the scale of 100 kb to 1 Mb. Usually suppress transcription and promoter-enhancer contact.

Chromatin Domains: the unit of chromosome organization [Dixon and Ren, Mol Cell, 2016]

- Evidence of chromatin domains: TADs match with the replication sites (RS).
- Nature of TADs: TADs are hierarchical in nature, there are sub-TADs. While TADs tend to be cell type invariant, loops and sub-TADs can change with tissues. TADs also exist at individual cells.
- Function of TADs:

  - Co-regulation of multiple genes: many examples of gene clusters, e.g. cytochrome genes, OR, protocadherin genes. Within TADs, the regulation may be non-specific, leading to possible co-regulation of related genes. This may be adaptive.
  - Blocking spread of transcription and set the boundaries.

- Consequences of TAD disruption: examples in human diseases and cancer.

- Three main forces that affect chromatin organizations:
  - Chromatin fiber movement.
  - Attraction due to binding sites of some diffusible modules. Cohersin complex: SMC1, SMC3 and RAD21.
  - Insultion: due to CTCF or other proteins. CTCF sites are in 90% of TAD boundaries.

- Clues to infer TAD mechanisms: disruption of CTCF binding has a large consequence on TADs. Cohersin, while essential for enhancer-promoter interactions, do not have a similar role in TAD formation. CTCF sites are present in TAD boundaries, but there are far more CTCF sites within TADs.

- Hand-cutoff model: Cohesin form rings, and CTCF connects the two TAD boundaries together.

- Extrusion model: cohersin-CTCF complex moving in opposite directions in chromosome, forming the extrusion, until they reach convergent CTCF motifs.

- Insulation-activation model (this paper): CTCF and housekeeping gene TSS place nucleosomes. And short nucleosome spacing makes chromosomes stiff, preventing interactions (chromosomes are harder to bend); while the chromosome regions with fewer nucleosome have more interaction.

- Questions/Remark:
  - What are the functions of CTCF sites within TADs?
  - One can directly test the prediction under insulation-activation model, e.g. the nucleosome density vs. TADs.

The Three-Dimensional Organization of Mammalian Genomes [Bing Ren, ARCDB, 2017]

- Compartments: A - transcriptionally more active, B - less active. Different in histone marks. Also compartment switching occurs during cell differentiation, 36% during stem cell differentiation.

- TAD distribution: TAD boundaries agree well with DNA replication domains.

- Mechanism of TAD formation: boundaries are enriched with CTCF and housekeeping genes. 75% and 33% of TAD boundaries have CTCF or housekeeping genes within 20kb. (1) Loop exclusion model: Figure 4c, cohesin ring/chromosome motor complex moves along the chromosome until it reaches convergent CTCF motif, then stop. Polymer simulation supports this model. However, it doesnt explain all TAD boundaries. (2) Active transcription can help define TAD boundaries: histone acetylation may disrupt nucleosome folding.

- Sub-TADs: less conserved across cell types, reflecting cell-type specific E-P interactions (mediated by Mediator/Cohesin). Chromatin loops anchored by CTCF and Cohesin served as the units of gene regulation: termed insulated neighborhood, may constrain E-P interactions.

- Chromatin loops and distal interactions: random collision is very rare for distal sites. What we have learned about distal interactions: often involve CTCF; E-E and P-P also common. Types of interactions: (1) invariant across cell types, set up TAD boundaries. Often convergent CTCF. (2) More cell-type specific, often enhancer interactions, and associated with cohesin and mediator.

- 3D genome changes during ESC differentiation: TAD boundaries are similar, but TAD activity (intra-TAD and inter-TAD interactions) changes: many from compartment B to A; and large increase of transcription.
- Tissue-specific organization of 3D genome: FIRE (frequently interacting regions), about 200kb in size, with interactions. Cell-type specific regulation.
- Changes of TAD in gene dysregulation and diseases: both cancer and congenital diseases. (1) Deletion of TAD boundaries: ectopic activation. In Gioloma, activation of oncogene. Also change of CTCF is frequent in cancer. (2) Inversion: may lead to ectopic activation or decrease of expression, depending on the relative position of E, P and boundary. (3) Duplication: SOX9 example (Figure 5c). Within TAD of E, may activate SOX9 > sex reversal. Involving boundary: may not have effect or activate a nearby gene.

Understanding the 3D genome: Emerging impacts on human disease [Seminars in Cell and Developmental Biology, 2019]

- Role of genes in 3D genome: (1) CTCF and cohesin. CTCF deletion experiment: CTCF required for CTCF-mediated chromatin loops and TADs, but not AB compartment. (2) YY1 (a TF): chromatin structural protein that promotes chromatin looping between promoters and enhancers. (3) LncRNAs: Xist, Firr important.
- Chromatin domains and gene expression: lamina-associated domains (LADs) heterochromatin.
- How noncoding genetic variation may lead to gene expression changes? (1) chromosome translocation can lead to trans-splicing. (2) SNVs in enhancers: Hi-CHIP experiment: 700 SNVs of AIDs, interacting with 2500 genes through chromatin interactions.
- Mechanisms of diseases by disruption of 3D genome (Figure 3): DNA replication changes can lead to problem of cell cycle; or gene expression changes.
- Gene expression disruption by SNVs: SNVs in enhancers may alter chromatin interactions. Ex: immune diseases, CADs, prostate cancer.
- Gene expression disruption by SVs (Figure 5c): deletion of TAD boundaries lead to enhancer adoption or hijacking. Inversion can also lead to enhancer adoption.
- DECIPHER study [Ibn-Salem, GB, 2014]: 922 deletions in DECIPER database, linked to to monogenic diseases that are associated with genes within or adjacent to the deletions. Up to 11.8% of the deletions could result in TAD disruption and lead to enhancer adoption.
- Limb/EPHA4 locus study [Lupianez, Cell, 2015]: rearrangements associated with limb malformation. Use genome editing in mice: these changes lead to ectopic long-range promoter-enhancer communications. This happens only when the boundary elements are disrupted.
- Enhancer hijacking in cancer: Ex. glioma, epigenetic change of TAD boundary (reduced CTCF binding) leads to enhancer hijacking in PDGFRA gene. CTCF/cohesin-binding sites are often mutated in cancer. CNVs may help formation of cancer-specific TADs.

Topological domains in mammalian genomes identified by analysis of chromatin interactions [Dixon and Ren, Nature, 2012]

- Overview: high resolution (deep sequencing) map of chromatin interactions by Hi-C reveals TAD structure in the genome. The problems: functions of the TADs, and the possible mechanism of establishing TADs.

- Method: identification of TADs in the genome. Define directionality index at each region: the bias of interacting with upstream sequences vs. downstream ones, or vice versa. From the indices, use HMM to partiontion the genome into domains.
- TADs: 2,200 domains in mouse ES cells, with median size of 880 kb (covering 91% of the genome). Also data in human ESC and IMR90 fibroblasts.
- TAD and heterochromatin: enrichment of CTCF binding sites (insulator elements), heterochromatin marks. The boundaries represent the end points of heterochromatin spreading.
- TADs across different cell types: about 50-70% are conserved across cell types. The cell-type specific interacting regions are enriched for diff. expressed genes ($> 20\%$). TADs also highly conserved across human and mouse comparison (ES cells).
- TAD boundaries are enriched with: promoter marks (eg. H3K4me3), CTCF sites, housekeeping genes and tRNA genes. For CTCF sites: consider how often they are associated with TAD boundaries within a distance. See enrichment vs. random expectation at various distance cutoff from 2kb to 200kb (Fig. S10): highest at 2-fold at 20-40kb. Using 20kb as distance cutoff, 4.8K out of 32K (15%) CTCF sites are associated with TAD boundaries (Fig. 2c).

Iterative correction of Hi-C data reveals hallmarks of chromosome organization [Imakaev and Mirny, NM, 2012]

- Motivation: the observed counts between pairs are due to both real signal, random polymer looping and biases (e.g. GC content, RE density).
- Model: our goal is to estimate the "true" relative contact probability due to random polymer looping (which should remove the biases).
  - Given the locus i and j, the observed contact is O_{ij}. Let E_{ij} be the expected count, then O_{ij} follows Poisson distribution with rate E_{ij}.
  - The rate E_{ij} depends on the bias at both loci as well as the true relative contact prob. due to random looping. We write it as $E_{ij} = T_{ij}B_iB_j$, where T_{ij} is normalized so that it sums to 1 for any given locus, i.e. $\sum_j T_{ij} = 1$. It can be shown that the factorization holds true.
  - We fit the model by MLE: the free parameters are B_i and T_{ij} subject to the constraints on T_{ij}.

Transcription factor and chromatin features predict genes associated with eQTLs [Wang & Wernisch, NAR, 2013]

- Background: method of mapping E-P pairs
  - Chromatin conformation capture: limited by low resolution
  - Co-variation between gene expression and enhancer histone marks: H3K4me1 occurs ubiquitously along the genome, there are too many potential enhancers that can be mapped to a particular gene.
  - eQTL: limitation is the SNPs in LD, thus low resolution.
- TFBSs, histone marks co-localize with eQTL: 29 TF ChIP-seq from LCL, 221K CREs. Choose 10K eSNPs: 1303 CREs positioned within 1000 bases of eSNPs.
- Co-expression between TFs and target gene expression: expression of TFs across 38 distinct hematopeitic cell types, then compare with expression of the genes near CREs (that are bound by a TF). The correlation of TF-gene pairs (a gene, and a TF bound to a CRE near the gene) depends on the distance between CRE (bound by the TF) and gene: at 1kb distance, higher correlation; beyond 1kb, the average R^2 is 0.18.

- Creating evaluation dataset: the 369 CREs that contain eQTL SNPs, we examined all genes within 1 Mb, some are targets (associated at $p < .001$) and the rest are non-targets.
- Other four features:
  - TF co-occurrence: in ChIP-seq data, TF may bind to both promoter and the target enhancer if there is E-P interaction.
  - Open chromatins at promoters.
  - The amount of overlap of TF GO annotations with those of nearby genes.
  - CTCF insulator sites block enhancers from interacting with gene promoter: so use CTCF markers located between CREs and test genes.
- Random forest classifers using all 6 features (genomic distance, TF-gene co-expression, and four additional features): AUC = 0.90 when using only distance and 0.96 using all features (many non-targets, so AUC is high).
  - Application of the classifer to other cell types: use model trained from LCL, test on fibroblasts and T-cells. Still better than single feature (genomic distance) - the difference is relatively small though.
- Performance of the method for distal enhancers: classifying genes located at least 150 kb away from a CRE. AUC of the method is 0.91 while using genomic distance is 0.75. At a distance of > 150 kb, the presence of insulator markers between the gene and CRE seems to accurately distinguish target genes from non-targets.
- Remark/Question:
  - For defining TF-gene co-expression feature, if a CRE is bound by multiple TFs, how coexpression is defined?
  - The approach relies on the expression and GO of TFs, but this is not often available.
- Lessons:
  - If a CRE targets a gene, then the TFs bound to the CRE should be co-expressed with the gene. This is a weak predictor though.
  - Insulator sites between an enhancer and non-target is a good feature.
  - Genomic distance is still the most important feature in general.

Global view of enhancer-promoter interactome in human cells (IM-PET) [He and Tan, PNAS, 2014]

- Limitations of experimental approaches: HiC has low resolution (1Mb), ChIA-PET has high FN rates, spatial proximity is not equal to functional regulation.
- Four features:
  - Enhancer-promoter correlation (EPC): Pearson correlation between enhancer histone signature (translate to an enhancer score), and FPKM of the gene. Across 11 cell lines. Correlation in real pairs is about .2-.3, while the false pairs is sligtly above 0.
  - Transcription factor and target promoter correlation (TPC): the correlation between the expression of TFs that bind to an enhancer and the activity of the target promoter. Correlations: .15 vs 0. To define the relevant TFs: for each enhancer, choose the best five TFs, based on motif scores and TF expression level, then correlate the TF expression and gene expression across multiple cell lines. The final TPC is the average correlation over the top five TFs.

  - Coevolution of enhancer and target promoter (COEV): sequence similarity and conserved synteny. For sequence similarity, compute the conservation of both enhancers and promoters across 14 mammalian species. The COEV score is defined as the product of enhancer conservation, promoter conservation and conserved synteny.
  - Distance constraint between enhancer and target promoter (DIS): there is a monotonic decline in the frequency of EP pairs with increasing distance. The distance distribution of real EP pairs is significantly different from that of nearest pairs and that of nonspecific interactions that arise due to random collision of chromatin fiber.

- Classification:
  - Training data: 2,000 real and noninteracting EP pairs from published ChIA-PET data for K562 and MCF-7 cells. Negative training set: for each enhancer, first sample a promoter based on the random distribution of contact frequency, then select the nearest promoter. The random contact frequency, as a function of distance s, is derived from linear polymer theory [Capturing Chromosome Conformation, Science, 2002]:

$$f(s) = k \cdot (s^{-3/2}) \cdot e^{-1400/s^2} \tag{3.96}$$

    where k is a constant reflecting the efficiency of the cross-linking reaction (note that k does not matter for relative frequencies).
  - Random forest (RF), use five-fold cross-validation to assess accuracy. Also trained SVM and logistic regression: RF is found to perform the best (significantly better than the other two).

- Prediction of EP pairs in 12 human cell types: validation using high-resolution HiC, ChIA-PET and eQTL. All perform better than existing methods.
  - 161,999 active promoters in these cell types using RNA-Seq data and GENCODE annotation of transcripts

- Enhancer-promoter tissue specificity:
  - On average, 2.9 promoters per enhancer. An example enhancer: constitutively active in four cell types. However, its target promoter(s) varies across the cell type. Furthermore, the expression specificity of the predicted targets is consistent with the predicted EP specificity.

- Cohesin can mediate chromatin looping without the involvement of CTCF: through analysis of cohesin sites within EP pairs.

Connecting the regulatory genome [John Stam, NG, 2015; Pollard, NG, 2015]

- Principle: the epigenomic state of the intervening genomic region extending from a distal element to its target gene differs markedly from the intervening segments between the same element and non-target genes.

- TargetFinder: use Hi-C data from [Rao, 2014] for training/evaluation. Three types of features: conservation, epigenomic states and motifs. Algorithm: random forest. The most informative features are "window features", those about the intervening sequences between enhancers and promoters - use average or density or the height of the top one does not make a large difference (the informative features are relatively sparse).

- Informative features:
  - Features indicate true interactions: P300, JUN, marks of heterochromatin, makrs of paused RNA polmerase.

  - Features indicate no interactions: cohesin complex, marks of open chromatin, marks of active promoter.

  Some interesting features that could generate hypothesis: enrichment of CTCF and RAD21 in the vicinity of target promoters. enrichment of the 'connector' factor HCFC1 at distal CREs.

- Performance: best when training and testing in the same cell type, but a large drop when doing the analysis on a different cell type.

Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts [Ay and Noble, GR, 2014]

- Motivation: mid-range interactions, 50kb-10Mb range for complex euk. genomes. Contacts in this range may occur due to random looping of the DNA, instead of formation of specific chromatin loops.
- Background: two sources of problems for analyzing contact data
  - Ranomd polymer looping of DNA: e.g. even by chance, two adjacent loci are more likely to have contact with each other than two distant loci.
  - Experimental bias: genomic characteristics such as GCcontent and mappability, as well as technical aspects of the assays such as cross-linking preference, fragment length, and circularization length.
- Contact probability: its relationship with the genomic distance is complicated: it depends on the sequencing depth, resolution, genomic distance of interest. Empirically, it was found that the relationship can be approximated by a power law, however the constant depends on aforementioned parameters.
- Baseline: discrete binning approach. The idea is to compute the null contact probability, then the significance of the observed contact count can be tested by binomial test.
  - Contacts between different chromosomes: the contact probability is uniform. Suppose we have M distinct pairs, and observe N counts. Any count between the two chr. has probability $1/M$ to be in any pair, thus the contact between any specific pair follows $\text{Binom}(N, 1/M)$.
  - Contacts within a chromosome: the contact probability depends on the distance. Assume the contact probability is uniform within any distance range (i.e. step-wise constant). Then within any distance range (bin), e.g. 100kb -105kb, apply the above approach: count the number of distinct pairs and the number of observed counts fallen into the range.
- Ideas of a new method:
  - The baseline approach is not a smooth function, thus not desirable.
  - Among all contacts, some are from null distribution, some are real (i.e. not due to random polymer looping). The real ones should be removed.
  - Bias need to be taken into account in constructing the null model. These bias may include GC content, the density of RE (restriction enzyme) sites, etc. The bias changes the null distribution.
- Method:
  - Defining contact probability: suppose we have N contacts in a range of interest (e.g. intra-chr., or all mid-range contacts), the contact probability is the probability that any randomly chosen contact from N falling into any specific pair.
  - Initial spline fit: Suppose we have a certain number of distance bins. For the i-th bin, let d_i be the distance of this bin (in fact, use average distance of all pairs of loci falled into this bin), and p_i be the contact probability of this bin. We want to fit a smooth curve, i.e. a spline using (d_i, p_i). To obtain p_i, let h_i be the average number of contacts in all pairs in the i-th bin, then $p_i = h_i/N$, where N is the total number of contacts (reads) in the range of interest (all mid-range contacts).

  - Refining the null model: use the initial spline and the binomial model (null contact probability of any pair is read from the spline) to test all pairs, and remove pairs whose p-value is less than $1/M$, where M is the total number of locus pairs. Then use the rest of contacts to fit the spline.
  - Correct for bias: use the ICE model in the raw counts, and obtain bias for each locus. Given two loci, l_1 and l_2, let p_{raw} be the raw contact probability from lookup in the spline, we multiply the raw probability by b_1 and b_2, where b_1 and b_2 are the bias estimated from the ICE method.

- Fit-Hi-C boosts statistical power: Figure 3C, comparing with the binning method, Fit-Hi-C improves the number of significant contacts at a given FDR, by 6-46%. Also, analysis of control data set suggests that Fit-Hi-C control FDR appropriately.

- Validation using known promoter-enhancer contacts:

  - Known contacts from ChIA-PET experiment. ChIA-PET measures the chromatin interactions mediated by a specific protein of interest. A recent dataset has ChIA-PET of RNA Pol II in ESC. Fit-Hi-C captures 77% of enhancer-promoter contacts and 73% of all contacts reported by ChIA-PET at 5% FDR. The rate is higher than the discrete binning approach.
  - Enhancer-promoter pairs. Data of enhancers, active or poised, predicted from histone marks, and gene expression in ESC. The contacts predicted by Fit-Hi-C are enriched with the pairs of active enhancer-gene expressed in ESC, but much less so with pairs of poised enhancer-gene. Also, genes whose promoters contact with at least one active enhancers have higher expression.
  - Cell-line specific contacts. Some genes, Gucy1a3 and Gucy1b3, are expressed in mCortex, but not in mESC. This tissue-specific expression can be explained from contacts in 3C data. There are many contacts between the promoters of the two genes and distant regions in mCortex, but not many in mESC.

- The number of high-confidence contacts in different sequences/regions: using ChromHMM and Segway, the insulators, heterochromatin, and binding peaks of pluripotency factors (Oct4 and Nanog) have more contacts in ESC than promoters and active enhancers (about twice).

- Correlation of high-confidence contacts with topological domains and replication timing:

  - Significant contacts more likely to happen within domains.
  - Significant contacts more likely to have similar replication time.

Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application [Li and Yijun Ruan, BMC Genomics, 2014]

- Experimental protocol (Fig. 2): in the beads, have linker sequences. After ChIP step, linker ligation with the bound protein-DNA complex. Suppose the complex has two distal DNA sequences: red and green, then the free ends are ligated with the linkers.This step will produce two kinds of sequences: red-linker1-linker2-red; or red-linker1-linker2-green. Then reverse crosslinking and RE digestion, sequencing.

- Data analysis: self-ligation PETs used to identify binding sites of the protein. Inter-ligation PETs used to define interactions: use clusters of PETs. Procedure: Fig. 3, linker filtering, PET mapping, classification of self- and inter-ligation PETs.

- Remark: the use of different proteins determine what type of interaction we profile. If use RNP, we will enrich for E-P interactions. If use CTCF, enrich for TAD or CCD domains.

Genome-wide map of regulatory interactions in the human genome [Heidari and Synder, GR, 2014]

- Background: Hi-C experiments require very high sequencing depth to reach high resolution. Currently, 20-50 kb. ChIA-PET enriches interacting loci, thus need much lower depth, can accurately detect interactions at < 5 kb.

- Data: ChIA-PET in myelogenous leukemia human cell line (K562) of six factors, (H3K4me1, H3K4me2, H3K4me3, H3K27ac) as well as POLR2A and a component of the cohesin complex (RAD21).
  - Mapping the interactions: obtain null distribution of interaction frequency from a control (resampling) dataset that retains
- the same distribution of PET distances as the observed data set.
  - Average size of regions: 3K and average distance: 120 kb.
- Comparison with previous data (validation of results)
  - Binding peaks: highly consistent with DHS peaks.
  - A total of 44% of CTCF regions, 36% of promoter regions,and 21% of enhancers were involved in at least a single interaction.
  - Remark: in ChIA-PET data, a lot of peaks may be detected (pulled down by Ch-IP), but not all of them are involved in interactions.
  - Most interactions (97%) occur within TADs.
- Factors enriched at interacting loci: Cohesin, CTCF, ZNF143, and HOT (high-occupancy targets of TFs) regions. ZNF143: transcription activator, which often colocalizes with CTCF and cohesin. Binding an 18-bp motif located on the core promoter region.
- Lessons: Nature of ChIA-PET experiments: ChIP define the binding peaks; PET part define interactions. Not all peaks are involved in interactions.

CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription [Tang and Yijun Ruan, Cell, 2015]

- Experiment: ChIA-PET, mediated by CTCF and RNP II in 4 cell lines, including LCL, HeLA.
- Define allele-specific interactions: find allele-specific CTCF anchors, and then phase the haplotypes. Some anchors can be phased > phased interactions; some non-phase, see Fig. 4B. About 300 allele-specific interactions (called haplotype-biased).
- Example: among 350 AS interaction anchors, 39 are located at CCD boundaries. One example in Fig. 5B: near CCD bounary, T allele stronger CTCF binding and C allele weaker. T allele: several loops, C allele, no loops.
- Asthma SNP: rs12936231 at ORMDL3 locus. SNP disrupts CTCF motif, and show AS CTCF binding, and loops in only one allele.

Enhancer-promoter interactions are encoded by complex genetic signatures on looping chromatin. [Whalen and Pollard, NG, 2016]

- Data: 6 cell lines from ENCODE, interacting and non-interacting E-P pairs (enhancers: several hundred bps). Training data from chromatin contact and capture Hi-C data.
- Machine learning method: linear SVM, decision tree and boosting tree (best performance). Also compare E/P features vs. Extended enhancer (3kb)/P features vs. E/P/Window: the EE/P is close to to E/P/Window, and much better than E/P, suggesting that windows, especially the sequences near enhancers determine E-P interactions.
- Ex. of E-P interaction: the active enhancer skips several active promoters in the middle. The skipped promoters do not have binding of Rad21 and CUX1.

- The important predictive features of E-P interactions are: (1) Binding of structural proteins (CTCF, Rad21) and cofactors (Cux1) in enhancers and/or promoters. (2) Activating histone marks H2AZ and H3K9ac, and elongation marks H3K36me3: likely markers of productive E-P interactions. (3) Repressive marks and PRC2: intervening sequences are not available for binding. (4) Sequence-specific TFs/proteins: Zn-finger proteins as adaptors (binding with CTCF and lineage-specific TFs bound at promoters), other TFs that ofter determine if enhancer/promoter is in active vs. poised state. The TFs that provide predictive boost are often cell-type specific.
- Remark: binding of sequence-specific TFs are good predictors of true E-P interactions and active/productive transcription.

The 4D nucleome project [Dekker, Nature, 2017]

- Genome organization in nucleus: the genome is compartmentalized in active and inactive spatial compartments; within each compartment, folding of chromatin fibres brings together loci and regulatory elements that are otherwise distant.
- Technology development: single-cell Hi-C, imaging in live cells, mapping RNA-DNA interactions; alternative cross-linking with bivalent photo-activated crosslinkers.
- Computational analysis: (1) Integration of complementary approaches. (2) Results are from ensemble of cells. Model development that takes advantage of single cell data. (3) Dynamic models of chromatin structure over time.

Function and regulation of 3D genome architecture in development and disease [Francois Spitz, 2019]

- Part I. Background: Myc in HSC differentiation. (1) Myc controls the balance of renewal and differentiation. (2) Myc enhancer cluster (BENC), 1.3Mb away, strong effect on inducing Myc expression. Phenotype copy LOF of Myc. (3) BENC: deletion of individual enhancers often have partial effect of expression, and the effects vary with cell types. Amplification and increased accessibility of BENC in AML patients.
- Exp: insertion of reporters (regulatory sensors) in genomic positions. Find regulatory domains: similar expression patterns within the domain. Ref: Ruf, NG, 2014. Symmons, GR, 2014. Q: "expression patterns refer to tissue-specific expression?
- Part II. Mechanism of TADs. Cohesin: deletion by Nibpl, which is cohesin loading factor. Effect: compartments not affected, but TADs are lost. CTCF binding to boundary is reduced.
- CTCF KO: increase contact between interTAD regions. NIPBL KO: reduce contact of intraTAD regions.
- Loop exclusion model: cohesin creates extrusion, then scan the genome, until reaching CTCF sites.
- TAD loss does not affect regulatory potential of enhancers.
- Tissue-specificity of sub-TADs and microdomains: cohesins could be preferentially loaded to active enhancers.
- Live imaging to study internal TADs. Discussion: single-cell E-P interactions. E-P interactions are relatively stable.
- Part III. Possible mechanisms of E-P interaction specificity: promoter competition for the same enhancers, promoter may also help facilitate loading. Promoter specificity not that important (e.g. FGF8 locus), but relative locations are more important.
- Additional mechanism may stabilize E-P interactions: some models, e.g. brmodomain.

- Prediction of E-P interactions from 1D. Q: what are features? Sequence features and epigenomic information.
- Part IV. Evolutionary principles of 3D genome organization: developmental enhancers may have larger TADs. Gene dense regions: might be more conserved. No TADs (or less established) in fruit fly.
- Hyp: TADs evolve as a response to genome expansion. Q: large non-vertebrate genomes: selection for TADs?
- Specificity issue: e.g. chicken and mouse enhancers of FGF8, in both cases, enhancers located in gene B activates only FGF8. However, if put chicken region in mouse genome, the enhances will activate both genes.
- Discussion: TF activity: opposite effects. An activator could be converted to repressor upon Wnt signaling.

TAD fusion score: discovery and ranking the contribution of deletions to genome structure [GB, 2019]

- Background: TAD boundary detection often vary substantially across methods.
- Model of contact frequency: divide genome into 5kb bins. Interaction freq. between two bins: bias of two bins x distance (1D) / insulators in between. Fit the model parameters with the observed contact frequencies.
- Model of effect of deletion: change of distance and insulators.
- TAD fusion score: defined as the expected change of contact frequency from deletion.
- Validation of regions with high TAD fusion scores: negative selection.

3.8.1 Computational Modeling of Chromatin Structure

Computational approaches from polymer physics to investigate chromatin folding [Bianco, COCB, 2020]

- Possible bridging proteins: Pol II, YY1, Mediator, PRC. Some such as TFs, Pol II, HP1 and mediator were observed in phase-separated condensates.
- Application of PRSIM model: to capture Hi-C data of Pitx1 locus, to show tissue-specific expression correlates with spatial configuration.
- Limitation: SBS does not model loop exclusion.
- Loop exclusion model: motor or driven by thermal diffusion, or transcription-induced supercoiling. Difficult with LE model: (1) Elimination of CTCF makes the boundaries fuzzier, but not remove them. (2) Both cohesin and CTCF depletion do not affect much gene expression. (3) Hard to explain simultaneous many-body interaction between multiple genes and enhancers, which is easier under SBS model.
- Both SBS (homotypic interactions) and LE models act together to shape chromatin organization. Support: Francois, Nature, 2017 paper; HiP-HoP model.
- Challenge of LE model: TAD boundaries are highly variable in single cells and that cohesin depletion does not affect TAD formation in single cells but only at the population average level. Ref: Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells [Bintu and Xiaowei Zhuang, Science, 2018].

Mechanistic modeling of chromatin folding to understand function [Brackey, NM, 2020]

- Background: higher resolution Hi-C reveals FIREs.
- Two kinds of models: (1) Inverse: take Hi-C data, learn 3D structure/constraint to fit the Hi-C data. (2) Mechanistic: take protein binding data, generate 3D structure. Use Hi-C data to verify the structure.
- Remark: ML models use hi-C data to train a model from DNA sequences, protein binding (and other features) to Hi-C/3D structure. It is different from both types of models.
- Background: MD simulation, atom behavior is governed by classical mechanics, Newtons Laws. The force field is computed from quantum chemistry. Only realistic for small system, e.g. a small protein unit. Numerical simulation: deterministic simulation, discrete time steps, and update position every step using Newtons Law.
- Using MD for larger system such as a chromosome: (1) Simplify solvent: Brownian or Langevin dynamics, represent as drags, thermal jostling (from water). (2) Coarse-grained DNA: groups of atoms as a unit, then do MD simulation. Interaction potentials are often: phenomenological. For chromsomes, may include steric hinderance (no overlapping atoms), polymer bending stiffness.
- Langevin dynamics (Box 1): adding to force field two other terms: (1) viscous drag, $-\xi v$, where v is velocity. (2) random kick from water.
- Bead-Chain model: each bead may represent several kb. Simple behavior: without other forces, a polymer configuration would look like a random walk. We can answer questions such as, What is the probability that the polymer will be forming a loop?
- Remark: the important difference with PRISM is that there is no need of homotypic interactions among the same factors. Phase transition is induced by increasing of DNA binding site densities by factor binding.
- Phase separation of proteins/binders: similar to Polycomb cluster and transcription factories. However, in reality, more dynamic. Refined model: allow turnover of binders to different states, binding or non-binding (e.g. by PTM).
- Loop exclusion model: the SBS model does not explain CTCF motif bias. So need exclusion.
- Note: bead-chain, and loop exclusion models all work for large-scale structure, but not gene scale predictions of E-P interactions. Also we have FISH data, at the single-cell level, of chromatin structure.
- HiP-HoP model: use both bridge diffusion and loop exclusion, from data of OCR and CTCF. Also allow variable linear compaction or chromatin fiber thickness (amount of DNA unit length), varied with H3K27ac - less compact fiber facilitates interactions. Output: population/ensemble of locus conformation (one per cell). Applied to Pax6 locus: validation using single cell level FISH data, show that the model predicts the E-P distance.

An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization (3D-GNOME) [Szalaj, GR, 2016]

- Background: in ChIA-PET, three types of PETs (tags): (1) singletons - long-range, but more like background level. Singleton contains information of chromatin structure: correlate with hi-C data. (2) PET clusters: long-range interactions. (3) Self-ligation PETs: local sites. Advantage vs. Hi-C: higher resolution.
- Overview: contact frequency normalization to get 2D heat-map, then 3D structure, and visualization and interpretation tools. Data represented as a hierarchy: top level, segments, about 2Mb in size; then anchors from ChIA-PET. Sub-anchors: regions between two anchors, use 5-7 subanchors per loop. Note: CCD sizes vary greatly 10kb to 7Mb, so cluster CCDs as segments to have more more uniform sizes

- Model overview: use Simulated Annealing to sample configurations $\{r_{ij}\}$'s, according to the energy function $E(r)$. The energy function has two parts: $E_{\text{polymer}}(r)$ and $E_{\text{data}}(r, d)$, where d represents experimentally determined pair-wise distance. Use a standard Harmonic potential, $\sum_{i,j}(r_{ij} - d_{ij})^2$. The interpretation is: r represents "equilibrium" distance, and d the actual distance. The energy function is thus the energy of a spring model.

- Low-resolution model: use CCD as segments (beads in SBS model) - better than uniform bins. Then construct low-resolution models using MDS or SA (polymer model). This model can also be used for hi-C data.

- High-resolution model: based on PET clusters. Within each segment: use anchors and loops (sequences between two interacting anchors). Use anchor interactions to constraint the polymer model, to minimize energy function. The function has several terms: (1) Anchors: Distance energy/harmonic potential; (2) For sub-anchors (sequences in the region between anchors), distance energy; (3) Sub-anchor bending energy; (4) CTCF orientation.

- Remark: (1) comparing with PRISM, 3D-GNOME uses loops defined by anchor proteins to learn the structure. (2) Use Simulated Annealing/MC methods to sample structure. Comparing with MD: requires equilibrium assumption, but more computationally efficient.

Polymer physics predicts the effects of structural variants on chromatin architecture (PRISM) [Bianco and Nicodemi, NG, 2018]

- Model: SBS model. A chromosome is modeled as N beads. Each bead is colored (binding factors): color 0 represents background bead - self-avoiding (steric hinderance), other n types of beads, each representing one binding factor. Binding of beads by factors can bridge different regions. Each factor has two parameters: concentration and interaction energy with beads. The model displays phase transition from unstructured to compact chromatin folds when concentration and interaction energy increase. Let c_i be the color of bead $i, 1 \leq i \leq N$. Given the parameters, and the placement of binding sites c, we can obtain the ensemble of chromosome configurations. From these, we can compute contact frequency between beads.

- Inference: use Simulated Annealing to find the best chromatin structure that explains the observed contact matrix from input. The objective function H measured difference of contact frequency between predicted vs. experiment C_{ij}. The model also has a penalty λ to penalize too many binding sites. The algorithm iterate these steps: i) take a polymer model with a given arrangement of binding sites; ii) derive the thermodynamics ensemble of its 3D conformations; iii) evaluate its corresponding contact matrix; iv) compare it with Hi-C data. It changes polymer model (binding site placement) at each step until convergence using SA procedure.

- Computational approximation: given binding site placement c, to compute contact frequency still need simulations to obtain ensemble of chromosome configuration. Use Mean-field approximation, results similar to full MD ($r > 0.9$).

- Validation using Hi-C data of EPHA4 locus: found 16 to 21 factors. Type I factors: located within TADs; type II factors: cross-TADs. Cannot simply map factors to ChIP-seq data. One factor may correlate with CTCF. Results in good agreement with experimental data (Fig. 2).

- Incorporating CTCF: add CTCF as binding factors, and use CTCF sites (convergent pairs). Not help with overall performance, help predict interactions between CTCF, but add some interactions not supported by data.

- Prediction of effects of SVs on EPHA4 locus (Fig. 3): three events, (1) a large deletion changing TAD boundary: ectopic activation of PAX3. (2) a similar deletion (> 1Mb), smaller, not including TAD boundary, increased interaction with PAX3, but much weaker. Consistent with experimental

observation of PAX3 not activated. (3) Inversion involving TAD boundary (1.1Mb): ectopic activation of Wnt6.

Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci (HiP-Hop) [Buckle and Gilbert, Mol Cell, 2018]

- Basic Bead-Chain model: the position of bead i, r_i is governed by Langevin dynamics:

$$m_i \frac{d^2 r_i}{dt^2} = -\Delta U_i - \gamma_i \frac{dr_i}{dt} + \sqrt{2k_B T \gamma_i}\eta_i(t) \tag{3.97}$$

  where U_i is the force field, the second term is the drag and the last one random noise. The U_i term models steric hinderance of nearby beads, and bending of chromatin. The steric interactions of adjacent and non-adjacent beads are modeled by WCA potential, which depends on diameters d and distance r_{ij} between two beads i and j. The bending term depends on a constant and the angle between three adjacent beads.

- Bridge diffusion model: each TF molecule (only one kind) is also treated as a bead, and the location is tracked. The number of TF molecules is given (known), and there is an interaction term of TF binding to site (not consider [TF] as a separate parameter). Using truncated LJ potential in the force field term to model TF interaction (attractive) with its site. Additionally, each TF randomly switch from binding to non-binding states (switching rate is a parameter).

- Loop Exclusion (LE) model: choose two random beads, and put a "string" to connect the beads. The string limits the distance between two beads. We allow beads to random move to next bead (with a given rate). The number of extruders is given/fixed. If one LE reaches a correctly oriented loop anchor bead, movement of that side of the spring stops, and the other side will keep moving until reaches another anchor.

- Variable Compaction Chromatin model: adding additional springs between next-nearest neighboring beads. This has the effect of making the chromatin thicker. In regions with H3K27ac, no springs.

- Setting model parameters: (1) Resolution: 1kb bins as bead. (2) Timescales: diffusion constant, mass. (3) TF switching rate and number of molecules, LE unbinding rate and extrusion velocity. (4) Persistence length of the polymer. (5) Parameters of the potential terms. Because of the running time, do not search for parameters.

- PAX6 locus: epigenetic features, choose 3 cell lines with PAX6 expression, ON, OFF and HIGH. In HIGH/ON, see higher H3K27ac; CTCF and Rad21 largely similar across cell lines, but additional CTCF/Rad21 sites at PAX6 promoter in HIGH/ON.

- DNA accessibilty better than H3K27ac to model the positions of factor binding sites. Compare with Capture-C data.

- Adding Heteromorphic Chromatin fibers improve model fitting with capture-HiC data: allow fiber to depend on H3K27ac.

- Switching off individual features and compare with experimental data: show that TFs give rise to promoter-enhancer interactions, while LEs generate domains.

- Structural variation at Pax6 locus: found 8 main classes of structure (from clustering), including no interaction of E-P at Pax6 promoter; only upstream enhancer interactions; or both up- and dow-stream enhancer interactions. Or cluster by E-P interactions (Fig. 5E): in OFF cells, mostly no E-P; in HIGH or ON, mostly upstream with promoter; or upstream + downstream with promoter.

- Structural variation matches single-cell transcriptional heterogeneity of Pax6.

- Lesson: H3K27ac regions may be outside OCRs, see Fig. 4A, yellow (H3K27a) is much larger than red (ATAC).
- **Question**: how is bivalent or multi-valent binding of TFs modeled? Ref: Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains [NAR, 2016].

Spatial chromatin architecture alteration by structural variations in human genomes at the population scale (3D-GNOME 2.0) [Sadowski and Yijun Ruan, GB, 2019]

- ChIA-PET targeting CTCF and Pol 2 data in LCL. 40K CTCF-mediated interactions. Median length of fragments (anchors): 2.7Kb. Interactions: median length of 100kb. Chromatin-contact domains (CCDs): about 2000 in the genome. General agreement with hi-C data.
- Genetic data: SVs from 1KGP. Also ChIP-seq data of about 15 LCL samples.
- CTCF orientation: among clusters, >70% have CTCF motifs of unique orientation in both anchors. Among them, 65% have convergent orientation.
- Allele-specific CTCF variant (Figure 2): asthma-associated SNP in a CTCF site, change CTCF motif. The site is in chromatin loops of multiple interactions. The SNP is associated with different CTCF peaks (from ChIP-seq of multiple individuals) and is eQTL of 4 nearby genes: ORMDL3, GDDMB, etc.
- Algorithm to build chromatin models: basic model (Figure 3ab). Extension to predict the effect of SVs on individual chromatin loops. Consider SVs that overlap with anchors:
  - Deletion: anchors will be removed. If the anchor is in the CCD boundary (outmost anchor), the CCD will be fused with the nearby CCD.
  - Duplication: anchors and all the contacts will be duplicated.
  - Inversion: consider CTCF motif orientation.

  SVs that miss anchors will only change the length of chromatin loops.
- Ex. of application of 3D-GNOME: deletion of a CTCF site, leads to ectopic interaction (decrease of 3D distance) of an enhancer with TAL1 promoter, an oncogene (Figure 3d).
- Ex. of a SV (Figure 4): about 5-10kb deletion of CTCF anchor site, that is the boundary of an insulated neighborhood. Model predicts that deletion leads to activation of one gene (decrease of 3D distance) - confirmed by expression change; and reduction of another gene.
- Duplication and inversions of CTCF anchor sites: lead to change of CTCF binding signal in ChIP-seq data.
- Conservation (frequency of SVs) in CTCF anchor sites: (1) CTCF anchors intersected with active enhancers/promoters: tend to be conserved. (2) CCD boundary: not well-conserved. CTCF motifs in ChIP-seq peaks outside CCD boundaries are more conserved than those in the boundaries.
- CTCF anchor sites in GWAS: focus on blood cell traits (HP) and AIDs (AI). CTCF anchors with enhancers/promoters, more enriched of GWAS SNPs, comparing with enhancers/promoters (small difference in anchors in enhancers, 2-fold higher in anchors in promoters).
- Examining SVs using eQTLs (only based on SVs): 450 samples in gEUVADIS. Assess eQTL on all genes in the same CCD. Found 234 unique SV-eQTLs modifying expression levels of 192 genes.

- Enrichment of functional elements in SV-eQTLs: RNAP anchors are enriched, but not CTCF anchors. Possible explanation: large negative selection of SVs disrupting CTCF anchors, while SNVs are tolerated. So enrichment of GWAS signal in SNVs targeting CTCF anchors, but not see enrichment of eQTLs in SVs.
- Lesson: (1) Allele-specific CTCF (SNV) can change chromatin interactions, and is a mechanism of eQTL affecting multiple genes. Also see enrichment in GWAS SNPs. (2) Prediction of SV impact on chromatin interactions: possible based on polymer models. (3) SVs can disrupt insulator neighborhood, and change gene expression. Some examples, but not necessarily enrichment in eQTLs, which are most likely to be common variants.

3.8.2 Prediction of Enhancer-Promoter Interactions

Reconstruction of enhancertarget networks in 935 samples of human primary cells, tissues and cell lines (JEME) [Cao and Kevin Yip, NG, 2017]

- Model idea: to predict E-P in a sample, we can use enhancer and promoter features of that sample. The key idea is to add global information: if E and P are correlated across many samples, so E likely regulates P. To use that information for a particular sample, we consider $\beta_j \cdot x_{ij}$, where β_j is the effect and x_{ij} the state of enhancer j in sample i. This term should capture sample-specific enhancer effects, but we also consider this term relative to total expression y_i in sample i. This leads to the idea of using the residual $y_i - x_{ij}\beta_j$ as the enhancer-gene feature for sample i.
- Data: enhancer features from 127 Roadmap samples and 900 FANTOM samples (eRNA from CAGE). For Roadmap data, union of ChromHMM enhancers, histone marks and DHS peaks to define candidate enhancers. Each enhancer is associated with several enhancer features: H3K4me1, H3K27me3, H3K27ac, DHS, CAGE. Training data: use ChIA-PET data.
- Using multiple enhancers of a gene better explain variation of gene expression across a large set of samples.
- JEME: (1) Fit multiple regression (LASSO) of candidate enhancers (within 1 MB) vs. gene expression, one enhancer feature a type. (2) Classification of E-P in each sample using: (1) how much gene expression is explained by an enhancer, (2) distance between enhancer and TSS, histone marks of enhancer, TSS in the sample.
- Validation: cross-validation, enrichment in TADs or chromatin contact domains (CCDs), depletion of intervening CTCFs.
- Properties of E-P connections: for FANTOM5, 50-80% of genes specifically expressed in a sample group are regulated by enhancers only in that sample group.
- Classification of enhancer-promoter association patterns: multiple enhancers of a gene, independent regulation, always co-regulate the same gene, and mutually exclusive.
- Remark: even if the effect of a CRE on a gene is global, and does not change, its specific effect in a sample/cell type will vary depending on the state of the CRE in that sample/cell type.

Quantitative prediction of enhancerpromoter interactions (3DPredictor) [GR, 2020]

- Problem: given epigenomic data, predict E-P interactions.
- Problems of existing methods: TargetFinder, random-split strategy is problematic, since windows are not independent (split into training and testing data).

- Model: machine learning, use Hi-C data as training set. Features: gene expression, CTCF binding, chr. accessibility, genomic distance. Important features: epigenetic characteristics of the region between interacting loci; orientation of CTCF sites.
- Application to study genome deletion in EPHA4 locus: apply 3Dpredictor to deleted sequences, reproduce Pax3 and enhancer interactions.

3.9 Reconstruction of Gene Regulatory Networks

Techniques/data:

- Gene expression data: under different conditions and treatments and of different genetic background (e.g. mutants).
- DNA-protein binding experiments (ChIP-chip, etc.)
- Regulatory sequences and TF motifs.

Principles:

- Transcript relations: the co-expression of genes can be used to infer the GRN. First, the correlation bewteen genes and putative regulatory proteins may represent true (direct or indirect) regulatory relations. Second, the correlation among genes may allow one to identify transcriptional modules, thus simplifying inference.
- Regulatory sequences: the motif content of regulatory sequences may suggest the regulatory relation between TFs and target genes.
- Ideas for limiting networks:
  - Incorporating other types of data: e.g. PPIs - prefer networks which assign interacting proteins to the same clusters. This can be achieved by MRF with pairwise potential. TF-binding data: which assign genes to regulators according to ChIP- data.
  - Perturbations: e.g. gene knockout.
  - Dynamics of systems.
- Design the experiment to test specific hypothesis concerning the GRN. This include:
  - If x is a target of TF f, then change of f expression $\rightarrow$ change of x expression
  - If two TFs $f1 \rightarrow f2$, then direct target of $f1$ will not be affected by mutation of $f2$

Regulatory relationship among genes:

- Regulator binding to the promoter/enhancer sequences of the target gene: note that the binding events may be condition dependent.
- The promoter/enahncer sequqnces: contains the binding sites of the regulators.
- The effect of perturbation of regulators on the expression of the target gene: in some cases, need to consider the relative expression (differential expression) of target genes.
- Association between regulator expression and target expression: in general, this is not causal, but may be combined with other evidences, or if there is enough data.
- Genetic mapping (eQTL).

Functional similarity between genes / identifying gene groups:

- Similarity of expression profiles: often need functional labeling, e.g. specific to only one signal, involved in specific processes, etc.
- Similarity of CRSs: sharing similar TFBSs.
- Gene functions: involved in the same processes/subprocesses.

Functional interaction among proteins:

- Protein-protein interactions: among related TFs, signaling proteins, etc.
- Physical networks: the transitive relations among proteins.
- Co-regulation of similar target genes.
- Synergistic effect on target gene expression.
- Synergistic or other types of effects on phenotypes: genetic interactions.

References: [Friedman, Science, 2004], [Workman & Ideker, Science, 2006], [Bonneau, Nat Cell Biol, 2008], [Amit & Regev, Science, 2009], [Kim & Regev, Science, 2009], [Przytycka & Slonim, BriefBioinfo, 2010]

3.9.1 Using Gene Expression Data to Reconstruct GRN

Using Bayesian Networks to Analyze Expression Data, [Friedman, J Computational Biology, 2000]

- Equivalence class of BNs: BN only represent a set of conditional relationship (CI), so if two graphs have the same set of CI's, then they are not distinguishable, and equivalent. Ex. if $A \to B \to C$ is one BN, then it is equivalent to $C \to B \to A$ as both essentially describes $A \perp C|B$.
- Theorem: Two DAGs are equivalent if and only if they have the same underlying undirected graph and the same v-structures (i.e. converging directed edges into the same node, such as $A \to B \leftarrow C$).
- PDAG: an equivalence class of network structures can be uniquely represented by a partially directed graph (PDAG), where undirected edge means that some members of this class have $X \to Y$ while others $X \leftarrow Y$. The PDACs fundamentally limit our ability to make inference of causal relationship.
- Inference of BN: the optimization problem is NP-hard. In practice, use MCMC.
  - To make MCMC more efficient, during each update, only recomputes the likelihood of part of the network that the update has affected.
  - Ideas for speeding up MCMC: derive CI from the preliminary anaysis, and then only search for networks that satisfy the constriants.

Module Networks [Segal & Friedman, NG, 2004]:

- Problem: given a set of regulators and a set of genes, find the partition of genes (modules) and how the regulators control the expression of each module (regulatory program).
- Methods: see Fig. 5 for the output/goal of the algorithm
  - Data: more than 400 regulators, and a large number of microarray samples of yeast.

- Model: n variables, M samples. Assume the n variables can be grouped into K modules, where all variables of one module have the same parents in the Bayesian Network. Effectively, one can treat all variables of a module as a single formal variable for that module, and the module variables form a usual Bayesian Network. Define T as the module network template (the Bayesian network on the module variable), and A as the assignment of variables to modules, and θ be model parameters. To make inference, compute the likelihood function on the ground Bayesian network (expand a module variable to all its actual variables):

$$P(D|T, A, \theta) = \prod_{j=1}^{K} \left[\prod_{m=1}^{M} \prod_{A(X_i)=j} P(x_i[m]|\pi_{M_j}[m], \theta_{M_j|\pi_{M_j}}) \right] \tag{3.98}$$

  where π_{M_j} is the parent of the module M_j.
- Expression model: the condition probability distribution (CPD) of a gene given the parent of the module where the gene belongs to. This function is specified by a regression tree defined on the expression of the regulators.
- Inference: learn both modules and regulatory programs simultaneously. At M-step, learn the regression tree and the model parameters; at E-step, reassign genes to the best modules.

- Remark:
  - To reduce search space, group genes into clusters/modules.
  - Different from other Segal work in: the expression is determined by the control regulators, not modelled by a normal distribution

The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo [Bonneau & Thorsson, GB, 2006]:

- Problem: given a set of predictors (TF level or environmental stimuli), predict the level of expression of genes.
- Methods:
  - Data: 268 mRNA microarray in a wide range of genetic and environmental perturbations. 24 conditions for validation (not used for training). 2,404 genes of which 124 are TFs. Each condition is compared with a reference condition that is identical in all 292 experimens. Genes that are not significantly changed (significant changes in less than 5 conditions) are removed. Normalize all genes s.t. variances equal to 1.
  - Predictors: 72 TFs (100 TFs show significant changes, then group TFs with correlation > 0.85); and 10 environmental factors (EFs).
  - Biclustering: cMonkey that groups genes and conditions based on - (i) expression correlation in subsets of conditions; (ii) sharing cis-regulatory motifs; (iii) functional or metabolic associations. Results in 300 biclusters covering 1,755 genes. In addition, 159 individual genes with unique expression profiles.
  - The basic kinetic equation of expression: let X be predictors and Y be response variables, i.e. expression of individual genes or biclusters (mean expression of all genes in a bicluster).

$$\tau \frac{dy}{dt} = -y + g(\beta Z) \tag{3.99}$$

    where g is a logistic or truncated linear function. βZ describes the effect of predictors:

$$\beta Z = \sum_i \beta_i z_i[x] \tag{3.100}$$

where $z_i[x]$ is some function of x. To model steady-state, simply set $\frac{dy}{dt} = 0$, and it becomes a standard regression. To model time series data, replace the derivative with finite difference, or effectively, use the predictor values at the prior time point (about 6 hours earlier).

  - Encoding factor/predictor interactions: for example, if $\beta Z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 \min(x_1, x_2)$, then $\beta_1 = 0, \beta_2 = 0, \beta_3 = 1$ approximates AND function. Thus, choose βZ be a linear function of x_i and $\min(x_i, x_j)$.
  - Model fitting: fit the model for each gene or bicluster. The performance metric is RMSD. The fitting consists of two steps: (1) Exhaustively evalute all single and pair-wise interactions: save the top five single influences and top two pair-wise ones. (2) Fit the model with the chosen predictors, using L1 shrinkage or LASSO.

- Results:
  - The predicted network consists of 1,431 regulatory influences on 300 biclusters and 159 individual genes. Average 3 regulatory influences per bicluster/gene.
  - RMSD in training and testing data are similar: mean RMSD - 0.37/0.37 (train/test) for biclusters, and 0.75/0.67 for genes (significantly higher). The effects of different modeling components: no temporal modeling - 0.40; single predictor per bicluster - 0.43; no interactions - 0.41.

- Questions:
  - Environmental factors: how are they quantitified, or just binary?
  - Biclustering method (cMonkey): use cis-regulatory motif to improve biclustering, but for the new species, where does the motif information come from?
  - Biclusters: trained only on conditions where the genes form clusters? And how to predict new conditions?

GWTM: Genome-Wide Transcriptional Modeling [Barenco & Hubank, MSB, 2009]:

- Motivation: the mRNA degradation rate strongly affects the transcript profile, thus incorporating the degradation rate in the calculate of transcript data.

- Methods:
  - The transcript level is determined by both synthesis rate and degradation rate:

$$\frac{dX_j(t)}{dt} = B_j + S_j f_j(t) - D_j X_j(t) \tag{3.101}$$

    where $X_j(t)$ is the level of transcript j at time t, B_j is the basal rate, $f_j(t)$ is the activities of the regulator at time t, S_j is the synthesis rate and D_j the degradation rate (measured through mRNA time course after blocking transcription). The first derivative is obtained from the time course by Laplace interpolation. This allow one to infer the production rate: $B_j + S_j f_j(t)$ for any transcript, called G_j.
  - Ananlysis of G_j profiles: (1) clustering G_j: using graph-based clustering (find cliques in correlation graph); (2) identify TFs associated with clusters by motif enrichment, and whether known targets of the TF are enriched; (3) find additional targets by correlating G profile with the activity profile - the average G_j of all genes in a cluster.
  - Data: transcriptional response to DNA damage in human T cell line (MOLT4) using time-course microarray.

- Results:
  - Three main clusters: Cluster 1 - NF-κB, c-Jun; Cluster 2 - p53.

- About 70% of 200 most up-regulated genes can be assigned to one of the three activitity profiles.
- Evaluation: the transcript level of predicted targets after knocking down the TFs.

- Remark: because of experimental noises and other issues, it may be better to focus on the most prominent expression profiles (or production profiles, as in this work), extract the relevant regulators, and then identify additional targets.

3.9.2 Combining Regulatory and Transcriptome Data for GRN Reconstruction

Predicting regulatory targets [Beyer & Ideker, PLoS CB, 2006]:

- Goal: given a TF, predict its targets by integrating all relevance evidences.
- Methods:
  - Diverse types of evidences: for a regulator X and a gene to be tested Z, the evidence supporting $X \to Z$ includes the following
    - Binding (B): if the sequence of Z to the factor X in ChIP-chip experiments
    - Sequence (S): if the promoter of Z (2000 bp) contains the binding site of X, by using the PWM of X. Scored by log-likelihood ratio (LLS).
    - Orthologous sequences (O): if the promoter of the 4 related yeast species contains the bindng sites. Add the LLS of 4 species together.
    - Coexpression (C): initally identify the cluster and TF relationship (which clusters are related to which TFs). Then if the expression of Z is highly associated with a cluster of X, then Z will have a high C score.
    - 2hop scores: if X regulates some gene Y, and Y is related to Z in some way, then Z will receive a score from combining the score of X regulating Y, and the score of Y being related to Z. The XY score is always based on binding, and the YZ score includes: coexpression (E), physical interactions (P), phylogenetic profile (Y) and fusion (Z). It is not clear how YZ scores are exactly defined, taken from an earlier study [25].
  - Integrating evidences: Naive Bayes classifer. For each type of evidence, get the score distribution of positive and negative training data (pooling all known TF-target pairs), and the contribution of this evidence (suppose it is E) is:

$$LLS(E) = \log \frac{P(E|L)}{P(E|\bar{L})} \tag{3.102}$$

    where L stands for positive hypothesis and $\bar{L}$ stands for negative hypothesis. The total score is:

$$LLS = \log \frac{P(L)}{P(\bar{L})} + \sum_E LLS(E) \tag{3.103}$$

    Note that the evidence from different types can be combined with a weight factor (empically found to be optimal at 1).

- Remark/Criticisms:
  - To define 2hop scores, how does the intermediate gene Y is chosen? In particular, if there are multiple intermediate genes?
  - The 2hop scores rely on the knowledge of known targets in the form of binding. However, binding data is not generally available. The coexpression score (C) depends on knowing the relationship between cluster and assoicated TF, again, may not be available.

- Converting score to a probability: if the training data is available, then one can estimate the score distribution of positive and negative data, which can be used for computing the probability of an observed score under positive and negative models. In particular, this could allow one to use statistical methods, such as Naive Bayes. An alternative method is to treat a score simply as a feature, and use the general classification method.

SEREND [Ernst & Bar Joseph, PLoS CB, 2008]:

- Goal: for any TF, predict its targets
- Idea: use the known targets to train a classifier of the targets of this TF.
- Methods:
  - Classifier: use logistic regression.
  - For any TF, its target will have two sets of features: (i) expression profile (one feature per experiment): used to build an expression classifier; (ii) motif match (1 feature): used to build motif classifier. The two sets of features are combined to build a meta-classifier.
  - Self-training: because no negative data is available, and in fact many unlabeled genes are likely to be targets. Relabel a gene to positive if its score is above some threshold. Repeat the process.
  - Evaluation: by ChIP-chip data; and by new experiment to test dynamics of prediction.
- Remark:
  - Meta-classifier: how to merge evidence (of different types)
  - Self-training procedure: use unlabeled data to improve classification.

TRANSMODIS [Yu & Wang, PLoS ONE, 2008]:

- Problem: given the motif and expression of genes across multiple conditions, predict the target genes.
- Methods:
  - Mixture model: for a given set of genes, some are targets and other ones are non-targets. The promoter sequence of target genes should match the PWM of the motif (assuming the best site is used), and the sequence of non-target should match background. The expression of target genes should follow the same Gaussian distribution while that of non-target genes should follow a different Gaussian distribution. Furthermore, assume that the targets have consistent and significant expression changes in all experiments studied, i.e. the same mean and variance of targets and non-targets under all conditions.
  - Procedure: in the first step, use the genes whose sequences match the core motif to train the model. In the second step, use the trained model to classify genes not containing the motif (to find additional targets). No parameter estimation in the second step.
  - Processing of expression data: the data in each experiment must be normalized to have mean 0 and variance 1. The Guassian parameters of non-targets are fixed to 0 and 1.
  - Unequal variance problem: the target and non-target expression have different variances. Thus, a very large or small value will always be classified as the one with larger variance. Intuitively wrong. Saft-guarding procedure.
- Remark/Criticisms:
  - Expression model: assume that all target genes follow the same expression pattern, too restrictive. Furthermore, assume that the distributions in all conditions are identical, even more restrictive.
  - Sequence model: only the best site is used.

Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions [Rodelsperger & Robinson, NAR, 2011]:

- Method: assume we have enhancer data (from ChIP-chip/seq of some TF), built a binary random forest classifier for the problem of deciding whether a single gene is an enhancer target based on the four features: genomic distance from the gene to an enhancer, CSS (conserved synteny score) between gene and enhancer, PPI distance and GO similarity between the enhancer-binding protein and gene.

LPS-induced transcriptional program in macrophages [Ramsey & Shmulevich, PLoS CB, 2008]:

- Problem: find the TRN of LPS-induced genes, in the form of (f, C), where f is some TF and C is a cluster of genes.
- Methods:
  - Expression data processing: filtering genes with minimum absolute expression; testing differential expression; transformation of data by SDR (signed difference ratio).
  - Gene clustering: K-means with Euclidean distance, model selection by BIC.
  - TLC: time-lagged correlation between a TF f and a gene. The overall association between f and a cluster C is found by combining p-values of f and any gene in C.
  - Promoter scanning: for any motif m, scan 2kb upstream of each gene in the cluster C. Sigificance of m wrt. C is tested by Fisher exact test (the number of genes matching m in C, the number of genes in C, and those two numbers in the entire gene set. Choose cutoff 0.01). The association between f and C, $p^{\text{scan}}(f, C)$ is defined as the minimum over all $p^{\text{scan}}(m, C)$ for all m that are associated with f.
  - Integration of TLC and promoter scanning with Fisher's method $\Rightarrow$ network of (f, C).
- Results:
  - From expression data (upon LPS stimulation): about 2,000 significant genes, grouped into 32 clusters
  - Build a network of 36 TFs and 27 clusters.

Systematic identification of mammalian regulatory motifs' target genes and functions (PhylCRM) [Warner & Bulyk, Nat. Methods, 2008]:

- Problem: find the mapping between gene sets and motif combinations.
- Methods:
  - PhylCRM: given a sequence, find the regions enriched with some specified motif combination. Specifically:
    - Single motif score: HB model with LRT
    - Multiple motifs: need to deal with overalpping cases. If a site is associated with multiple overlapping motifs, then the motif with higher score will be chosen for this site.
    - Total score of a region: sum of scores of all sites in this region
    - Alignment gaps and missing TFBS in some lineages: remove those lineages, i.e., the score will only depend on those lineages without gaps and are conserved TFBSs
    - Statistical significance: evalutated by computing the distribution under null model.
    - Boolean combination of motifs (?)
  - Lever: given a list of gene sets and a set of motifs, evaluate all gene set-motif combination pairs (GM pairs). For any GM pair, do the following:

    - * Extract CRMs: highest scoring regions with PhylCRM.
    - * A motif combination (MC) is evaluated by whether it distinguishes the given gene set (foreground) and the rest (background): AUC score
    - * Statistical significance: AUC score → Q value. For a single GM pair, the significance of AUC can be defined by permutation test: permutating gene labels (foreground or background). Since we are testing many GM pairs simultaneously, we only need to permutate the entire gene sets, say N times. Then the AUC of a pair can be calculated by the distribution of AUCs in N permutations (each permutation - a matrix of AUCs).

    A matrix (heat-map) of gene-set and MC can then be drawn.
  - CRM data for evaluation: (i) Wasserman 27 muscle CRMs (learned using human-mouse conservation) - 75kb containing each CRM; (ii) the collection of 46 known sarcomeric genes (each should contain a CRM sequence). The size matched background set.
  - Myogenic expression data: time course expression data of human skeletal muscle cell differentiation. Clustered into 14 groups with K-means algorithm, then intersect with GO categories (that are enriched in these clusters) to get 101 gene sets. 4 known myogenic motifs: MRF, MEF2, SRF and Tead.
  - Motif collection: 174 human regulatory motifs (with sequence conservation) from [Xie, Nature, 2005], identified 4-kb proximal promoter regions.

- Results:
  - Comparison of PhylCRM with Stubb and Comet: PhylCRM uses 8 mammalian species, Stubb only human-mouse and Comet human only. Results in Table S1: (i) Wasserman data: all three programs are similar without phylogeny, and PhylCRM and Stubb are similar with phylogeny; (ii) sarcomeric data: without phylogeny, PhylCRM similar to Comet (AUC about 0.60) , but much better than Stubb (0.50); with phylogeny, PhylCRM (0.74) better than Stubb (0.59).
  - Validation of methods with 4 myogenic motifs: test enrichment of the 4 motifs in 14 clusters: significant enrichment in up-regulated clusters (C0 to C5), especially the MRF motif alone. However, group all up-regulated gene (or down-regulated genes) show litter enrichment.
  - Analysis of 174 motifs in 101 gene sets:
    - * Discovery of other known motifs involved in myogenesis, such as AP-1, Elk-1 and Pitx2.
    - * The properties of the discovered TRN: (i) some gene sets (unexpectedly) share common motifs; (ii) certain motifs, e.g. NF-Y regulates a large number of genes.
  - Experimental verification of 6 predicted CRMs: 4 genes close to known myogenic genes, 2 others are distal ones (¿10kb). Tested by: (i) binding to muscle factors via ChIP-QPCR: 4 proxmal CRMs show strong binding; (ii) expression in muscle cells via luciferase assay.
  - Functional annotation of uncharacterized motifs: e.g. TGACATY can be annotated as being involved in the regulation of plasma membrane genes.

- Remark:
  - CRM prediction with given motifs: Wasserman data is biased because they are identified using human-mouse conservation. The difference (15%) of performance between Stubb and PhylCRM is identical in the case of no phylogeny and with phylogeny. Thus the reason Stubb is not as good is probably due to the weak sites, but not the conservation.

Simultaneous clustering of promoter sequences and expression: kimono [Holmes & Bruno, AAAI, 2000]:

- Motivation: earlier procedures do clustering based on expression and then see if the genes of the same cluster have similar promoters. Better to find clusters of genes that have (a) similar expression patterns **and** (b) similar promoters.

- Model: K clusters of all given genes, each gene has both sequence and expression data. Each gene is assigned to some cluster, and after the assignment, generate both data:
  - Expression model: normal distribution with cluster-specific mean and variance
  - Sequence model: fixed-length, ungapped and nonrepeating model of [Lawrence & Wootton, Science, 1993] with cluster-specific motif
- Inference: the cluster index of each gene, the motif of each cluster and the mean, variance of expression of each cluster
  - Gibbs sampling step: for each cluster, sample the alignment for any gene conditioned on the alignment of all other genes in this cluster
  - Expectation step: set the cluster index according to the posterior probabilities
  - Maximization step: set the cluster center according to MLE

Genome-wide discovery of transcriptional modules from DNA sequence and gene expression [Segal & Koller, Bioinformatics, 2003]:

- Motivation: simultaneous expression clustering and identify enriched motif combinations.
- Methods:
  - Expression model (M $\rightarrow$ E): given a module assignment M, the expression E follows the normal distribution with mean and variance specific to M.
  - Module assignment model: which module a gene belongs to is determined by its regulatory sequence. (1) S $\rightarrow$ R: the sequence S determines if a motif is present (R). The contributions of scores of all windows in the sequence. (2) R $\rightarrow$ M: the motif presence pattern (R) determines which cluster the gene belongs to (M). Multi-class logistic regression: each module has a specific motif profile.

MA-Networker [Gao & Bussemaker, BMC Bioinformatics, 2004]:

- Aim: binding targets are not necessarily expressed. Reveal the relationship between binding and expression.
- Methods:
  - Transcription factor activity profiles (TRAPs): let E_{gt} be the expression (log-ratio) of gene g in experiment t, and B_{fg} be the binding measurement of factor f to the promoter of gene g, then:

$$E_{gt} = F_{0t} + \sum_f F_{ft} B_{fg} \tag{3.104}$$

    where F_{ft} is the activity of factor f in experiment t. Use backward selection to eliminate uninformative transcription factors from the model.
  - Gene-TF coupling: Pearson correlation between the expression profile of the gene and the TRAP of the factor. For a TF, the genes are classified by: B-, unbound; B+/C+, bound and coupling genes; B+/C-, genes that are bound but do not couple.
  - Data: about 750 expression data, and binding data of 113 TFs.
- Results:
  - Coupled and uncoupled targets: 37 of 113 TFs are chosen. On average 58% of significantly bound genes are classified as significantly coupling genes.

- Validation of coupled targets: (i) enrichment of specific GO cateogories in B+/C+, but not B+/C-. (ii) B+/C+ genes are more likely to response in TF deletions. (iii) The promoter sequences of B+/C+ genes are more enriched with known motifs than B+/C- genes.

- Remark:
  - The binding measurement should also be condition-specific.
  - Can the results explain that TF activities are sometimes positive, sometimes negative? And why?

Bayesian error analysis model - BEAM [Sun & Zhao, PNAS, 2006]:

- Problem: given gene expression and protein-DNA interaction data, infer the regulatory relations (the existence and strength) among genes and TFs

- Methods:
  - Model: the expression of the i-th gene is determined by the activitities of all relevant TFs and their contributions to the expression. Based on an equilibrium between [TF], the TF-DNA complex and mRNA level:
$$\log(y_i) = \sum_j a_{ij} \log(\beta_j) + \text{error-term} \tag{3.105}$$
    where y_i is the expression of i-th gene, β_j is the activity of j-th TF and a_{ij} is the contribution (stochiometric coefficient) of j-th TF to i-th gene. Variables: $a_{ij} = r_{ij} b_{ij}$, where r_{ij} is the indicator varaible of i-j interaction, and b_{ij} is the binding between i and j, both r_{ij} and b_{ij} are assumed to be known from Chip-ChIP data (to caputre some uncertainty here, r_{ij} is not entirely determined by binding data, rather is determined proabilistically)
  - inference: MCMC

DREM: Reconstructing dynamic regulatory maps [Ernst & Bar Joseph, MSB, 2007]:

- Model:
  - The expression of a gene across n time points: modeled as a HMM, where the state represents the expression level and the transition represents change of expression (up or down). The states of the HMM constitute a binary tree.
  - The transition probability is a logistic function of the feature of a gene (binding by TFs), and for each time point (except the last one) and for each split (each of the hidden state), there is a logistic classifier (determine how state will change in next time point).

- Model learning:
  - Parameter learning: likelihood
  - Structure learning: learn the states of the HMM. The intuition is that only split if necessary.

- Remark:
  - Many hidden states, and thus many classifiers (one per state except those in the last time point). No regularization on classifier and no structure of these classifiers. In particular, in late time point, genes have been split and there are small number of genes per hidden state, thus not possible to learn classifiers. Maybe possible to aggregate, e.g. classifier is defined on whether a gene is up or down, regardless of its current level/history of up/down.
  - Because of these properties, good for a transcriptional response to a signal/stimulus, but not to developmental dynamics, e.g. stem cell differentiation.

The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line [FANTOM, NG, 2009]:

- Problem: TRN controling the growth arress and differentiation of PMA-stimulated THP-1 cells.
- Methods:
  - Promoter expression by CAGE: TSS usage (measure the first 20 bp of mRNA). Allow to identify promoter sequences (CAGE promoters). Expression profile of mutliple time points following PMA stimulation.
  - Motif Activity Response Analysis (MARA): (1) TFBS prediction from Transfac and sequence conservation using MotEvo, also use positional preferences (relative to TSS) as prior for predictions. (2) Regression of motif content and expression through motif activity: p for promoter, t time point, m for motif, and $N_{p,m}$ is the number of sites of m in p.

$$e_{p,t} = \text{noise} + c_p + \tilde{c_t} + \sum_m N_{p,m} A_{m,t} \tag{3.106}$$

  - Regulatory links: the regression tells the motif activity profile, $A_{m,t}$. Correlate this of a motif with the expression of a target promoter, if $Z > 1.5$, then call this a link.
- Results:
  - TRN: 30 core motifs and 199 predicted regulatory edges. The activity profile of 30 motifs can be grouped into 9 clusters, some are down-regulated, etc.
  - Function of target genes: targets of down-regulated motifs: cell cycle related; targets of up-regulated motifs: immune response, cell adhesion, etc.

PTMM: Post-Transcriptional Modification Model [Shi & Bar-Joseph, JCB, 2009]:

- Motivation: the activity of a TF may be determined by mRNA level, or can be inferred from its target expression (if post-transcriptionally modified).
- Model:
  - Let $G_i(t)$ be the expression of the i-th gene at time t, $T_j(t)$ be the activity of the j-th TF at t (hidden), and w_{ij} be the weight of j on gene i. Then, we have:

$$G_i(t) = \sum_j w_{ij} T_j(t) + N(0, \beta_d^2) \tag{3.107}$$

    where β_d^2 is the error variance. We assume that there is an indicator variable, Z_j for each TF j, with its prior a Bernoull distribution with parameter ρ. If $Z_j = 0$, i.e. not post-transcriptionally modified, then $T_j(t) \sim N(G_j(t-1), \tau_d^2)$ (mRNA level with time delay); if $Z_j = 1$, assume $T_j(t)|T_j(t-1) \sim N(T_j(t-1), \gamma_d^2)$.
  - Regularization and prior knowledge: the variables are (o) - observed expressed, **h** - hidden TF activity, **z** - indicator variable, the parameter are: W - weights, and θ - all other parameters. Want to maximize the penalized likelihood, which has two penalty terms: (1) the total absolute values of weights (L_1 penalty), to encourge most weights to be zero; (2) the penalty if w_{ij} does not agree with the prior TF-binding data.
  - Data: yeast expression data under various conditions. Also evaluation with DNA-damage expression data, where the additional ChIP-chip data can be used for validation, instead of as prior.

SDREM: Linking the signaling cascades and dynamic regulatory networks controlling stress responses [Anthony Gitter thesis defense; Gitter & Bar Joseph, GR, 2013]:

- Reference: [Discovering pathways by orienting edges in protein interaction networks, NAR, 2010]
- Motivation: given the time-series gene expression data of stress response, how can we find the TFs? Suppose we know some of genes that are potential "sensors" (e.g. knockout has phenotype in stress conditions), can we use these "sensors" to priortize TFs, e.g. favor TFs that are well-connected to the "sensors"?
- Network edge orientation model: given the sources and targets, and a network, orient each edge so that there are many paths from the sources to the targets. This is formulated as the objective function:

$$\sum_p I_s(p)w(p) \tag{3.108}$$

  where p is a path, and $I_s(p) = 1$ if it is satisfied (link source and target with the edge orientation) and 0 otherwise, $w(p)$ is the weight/confidence of path based on the confidence of PPI network.
- Model: two components, first, the dynamic expression model with DREM; second, assess if the TFs reported by DREM are well-connected to the sources. Specifically,
  - Given the TFs reported by DREM: assess its connectivity. For a single TF, this is basically the sum of the weights of all paths ended at this target TF. Also compute the connectivity scores of random targets to get the score distribution.
  - Use the connectivity scores of TFs as the priors of TFs in DREM search.
- Remark/questions:
  - Problem of physical network for mapping signaling events: most signaling events are transient, thus the corresponding interactions may be transient as well. Not clear how good the methods for mapping PPI actually recover those signaling interactions.
  - Constraints on edge orientation: e.g. kinase phosphorylates other genes; signal flows from membrane receptor to downstream proteins. Incorporating such contraints on edge orientation may help.
  - Comparison with network flow algorithms: the main benefit seems to be that the direction of each edge in the graph is fixed (but unknown). Not clear what is the benefit.
  - Sensor problem: it is possible to do analysis using only TFs, without sensors. Simply find the nodes connected to all TFs. How good this will recover the sensors?

Inference of transcriptional regulation in cancers (RABIT) [Peng Jiang and Xiaole Liu, PNAS, 2015]

- Background: TFs are important for tumor, e.g. E2F1 and FOXM1 are overexpressed, and drive proliferation; several other TFs drive tumor metastatisis.
- Data: 686 ChIP-seq data from 150 TFs on 90 cell types. Define TF targets by weighting the number of BSs by their distance to TSS. Also correct for covariates that may influence expression level, including gene CNA, promoter DNA methylation, etc.
- Model for single tumor sample: for gene i, let its diff. expression (vs. normal) be Y_i, let B_{ij} be the j-th background factor (e.g. promoter DNA methylation) of gene i, and R_{ij} be the regulatory potential of TF j on gene i. We have:

$$Y_i = \sum_{j=1}^{p_b} B_{ij}\beta_{b,j} + \sum_{j=1}^{p_r} R_{ij}\beta_{r,j} + \epsilon_i \tag{3.109}$$

  Or in vector form $Y = B\beta_b + R\beta_r + \epsilon$. For each tumor sample, we can then obtain the regulatory activity of a TF by the t-score of its $\beta_{r,j}$. To improve the model: choose the best ChIP-seq data for each TF, and do forward variable selection on TFs.

- Cross-sample inference for a tumor type: for each TF, correlate its expression and somatic mutation with its t-scores (regulatory activity) across all samples. Only significant TFs are retained.
- Results: (1) 20-30 TFs across 32 tumor types. (2) Using RBPs: using motifs to define RBP targets. Also find significant RBPs, e.g. RBFOX1.
- Remark: may find many TFs driving DE of many genes, but it does not mean that the TFs drive tumorigenesis.

Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases [Macbach & Bergmann, NM, 2016]

- Reconstruction of GRN:
  - TF-enhancer mapping: use FANTOM data for enhancers, search for conserved motif instances in these CREs. Weights defined as the confidence of TF-motif instance.
  - Enhancer-gene mapping: distance based weight, and the activity of level of gene and enhancers.

  Evaluation of TF-gene edges: relatively high, close to ChIP-seq results.
- Define connectiveity of gene pairs in network: p-step random walk kernel (probability of moving from one gen to another in p steps).
- Testing enrichment of a gene group in network: among a network in a tissue context, assess whether genes from a disease (GWAS) tend to be more connected. For the top genes in GWAS, compute mean network connectivity and assess its significance by permutation. Results: relevant cells show enrichment, but not much better than simpler approaches, e.g. PPI, or general (non-tissue specific) networks.
- Remark:
  - Construction of GRN: use TF and conserved motif instances can obtain a reasonable prediction of TF binding. Combine this with gene and enahncer activity levels.
  - Network enrichment analysis: pair-wise connectivity seems to be a very indirect measurement. Also the connectivity may be sensitive to network structure: e.g. the presence of a single hub can dramatically changes the connectivities of many pairs.

Modeling gene regulation from paired expression and chromatin accessibility data [Duren and Wong, PNAS, 2017]

- Model outline: let TG_l be the expression of target gene l, its expression is primarily determined by the activity status of elements, Z_i for element i. The activity status of an element is determined by the status of occupancy of chromatin regulators, or CRs (e.g. P300), denoted as $C_{i,j}$, the occupancy of CR j on e_i. And the status of CRs depend on TF expression, TF motif.
- Model: (1) the CR occupancy model, $C_{i,j}$ is determined by all the TFs with motif in e_i, or show PPI with CR j. Let TF_k be expression of TF k, TFS_k be TF expression specificity (across tissues), $B_{i,k}$ be the presence of motif k in e_i, and O_i be the openness of e_i, we have:

$$\text{logit}(P(C_{i,j}=1)) = \eta_{l,0} + \eta_{l,1} \sum_k \eta_k \left(TF_k TFS_k B_{i,k} O_i\right)^{1/4} \tag{3.110}$$

  (2) The element activity model: Z_i depends on C_{ij} for all CR's:

$$\text{logit}(P(Z_i=1)) = \alpha_{i,0} + \sum_j \alpha_{i,j} C_{i,j} CR_j \tag{3.111}$$

(3) The gene expression model: only elements with $Z_i = 1$ contribute, and in these elements, consider all TFs with motif presence.

$$TG_l = \beta_{l,0} + \sum_l \beta_{l,i} Z_i \left(\sum_{k \in MB_i} \gamma_{l,k} B_{i,k} TF_k \right) + N(0, \sigma_l^2) \tag{3.112}$$

- Assumptions: CR only determines activities of elements, but not expression level directly. Expression is determined by TFs bound in active elements.
- Analysis: how do we learn so many parameters? (1) Assuming we know all active elements, then we can just regress gene expression vs. TF expression and motif in the active elements, and this gives gene-specific $\gamma_{l,k}$ and $\beta_{l,i}$. (2) Suppose we know $C_{i,j}$ (which is similar to Z_i if there is only one CR, e.g. P300), then we can regress $C_{i,j}$ vs. TF expression and motif to obtain η_k.
- Possible issues with PECA include:
  - Over-parameterization: the intermediate variables $C_{i,j}$ may not be identifiable.
  - TF activities can be condition-specific, and may only be weakly correlated with expression. This is not modeled.
  - Functional form: arbitrary, and scale may not be correct, e.g. multiplication of $B_{i,k}$ with TF expression is weird.

Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility [Lamparter and Kutalik, PLCB, 2017]

- Motivation: learn the effects of TFs in regulating chromatin accessibility.
- Motif accessibility score: for each motif, count the number of instances in open chromatin regions in a sample. Now have the motif activity matrix: rows corresponding to motifs and col- umns corresponding to cell lines. Then quantile normalize across samples for each motif. Regress out the first PC, then standardized on both rows and columns.
- Model (Figure 1): for each motif, y is n-dim. vector of motif accessiblity score, where n is number of samples. Let x_i be expression of gene i (TF or other genes, which are controls), regression of y vs. x_i. Add random effect term to account for cell line relatedness: if gene expression are highly similar between two lines, then its likely that the motif accessibility scores are correlated.
- Results: chromatin accessibility regulator (CAR) rank of TFs, from ENCODE data. At FDR 0.1, found 25% TF subfamilies have chromatin effects. Strong enrichment of pioneer factors.
- Some TFs: gene expression are not good indicators of activity; instead downstream genes are better. Ex: GRs have low CAR ranks for GR motif, but downstream genes have higher CAR ranks.
- Remark: could use the association of gene expression and chromatin accessibilty to search for co-factors of TFs and chromatin remodeling genes.

Integrated analysis of motif activity and gene expression changes of transcription factors (IMAGE) [Madsen and Mandrup, GR, 2018]

- Motif collection: about 800 non-redundant motifs. Note: some TFs bind with multiple modes. Supplement with computationally predicted C2H2 family motifs, and homology-based motifs.
- Motif scoring: use p-values better than LRT. Optimal p-value cutoff suggested at 5E-4.

- Two-step model: define all candidate enhancers of a gene as those within 100kb. Step 1, regress enhancer occupancy vs. motif activity and motif presence in enhancers. Let O_{SE} be occupancy of enhancer E in sample S (centered and scaled), A_{SM} be motif activity of M in sample S, and N_{ME} be count of motif M in enhancer E (centered). We have:

$$O_{SE} = \sum_M A_{SM} N_{ME} \tag{3.113}$$

  In step 1, determine target enhancers of motifs by motif presence and leave-out analysis: removing motif and see if enhancer occupancy prediction accuracy changes significantly. Step 2, regress gene expression vs. motif activity. Let E_{SG} be expression of gene G in sample S, D_{EG} be distance weighting of E wrt. G, T_{ME} be indicator of whether E is a target of M (from step 1). Then:

$$E_{SG} = \sum_M A_{SM} \cdot \sum_E D_{EG} T_{ME} N_{ME} \tag{3.114}$$

- Analysis: how this model works? suppose we have data from condition 1 to 2, if enhancers containing M show higher occupancy, it suggests that motif activity of M goes up in condition 1 to 2.
- Validation: in adipocyte differentiation dataset, for several TFs with ChIP-seq data, show that predicted motif targets have higher read counts.
- Application to a differentiation dataset: found 100 new motifs and matched TFs. Most TFs also show differential expression. Choose 6 TFs with expression patterns, and do K.O. 5 out of 6 show adipocitye phenotype.
- Remark: limitations include, not incorporating [TF], thus from motifs to TFs may be hard (in the paper, selection of TFs uses [TF] information). Each enhancer has the same effect, other than distance weighting.

The cis-Regulatory Atlas of the Mouse Immune System [Yoshida and ImmGen, Cell, 2019]

- Data: low input RNA-seq and ATAC-seq on 80 immune cell types. About 500K OCRs, including 14K in TSS.
- LMM to estimate contribution of chromatin accessibility to gene expression (Figure 2):
- Linking OCRs to genes: for half of genes, at least one OCR associates with expression within 1Mb, generally positive association. Mostly close to the genes: 50% within 13kb of TSS (Figure 3C). Among these genes, $> 70\%$ have more than one OCRs, but they are often correlated with each other (Figure 3D). Found in 500 genes, clear evidence of multiple independent signals (Figure 3G): some OCR explains NK cell activation, one OCR explains B cell activation.
- Determining roles of TF on chromatin accessibility: define TFBS accessibility scores: OCRs containing a motif vs. background OCRs. Correlation of accessibilty scores vs. [TF] (Figure 5B-E): some non-linear relationship, 61 activators and 18 repressors.
- Role of TFs in immune cell differentiation: (1) TF enrichment in cell type specific OCRs (Figure 6C). (2) Clustering analysis/patterns of TF binding OCRs across cell types (Figure 7DE): e.g. some Pax5 targets are specific in B cells, some constitutive.
- Discussion: possible repressive mechanism of TFs, hit-and-run, instruct stable histone marks or DNA methylation.
- **Lesson**: the relationship of OCR and gene expression: most often within 100kb, often nearby OCRs have correlated activities.
- **Lesson**: basic principle to learn TF functions in cell type differences: if a TF drives cell type, its motif should be enriched in OCR of that cell type. Correlation of motif accessibility vs. [TF] can reveal the role of TF as activator or repressor.

3.9.3 Perturbation Based Reconstruction of GRN

Genetic reconstruction of a functional transcriptional regulatory network [Hu & Iyer, NG, 2007]:

- Motivation: ChIP-chip data does not identify the functional targets. A functional target of a TF should have differential expression when TF is deleted.
- Methods:
  - Experiment: gene expression of 263 TF knockout strains. A gene is defined as a (initial) target of a TF if its expression is changed in the TF deletion strain.
  - Refinement of TRN: remove indirect edges (using the idea of graphs).
  - ChIP-chip data: [Lee, Science, 2002]
- Results:
  - Low overlap between deletion targets and ChIP-chip targets: about 3% overlap (but statistically very significant). However, with better data, higher overlap. Example, of the 354 Rap1 targets in the high-quality ChIP-chip data set and the 537 targets we defined, 144 were shared, compared with only 71 targets shared between our data set and the large-scale ChIP-chip data set.
  - Refined TRN: 45% of all genes were significantly regulated by at least one transcription factor; 138 transcription factors had more activated targets than repressed targets, whereas 114 transcription factors had more repressed targets than activated targets.
  - Motif prediction: for 106 transcription factors, of which 61 had no previously defined binding motifs, identified new high-confidence binding motifs.
  - Function of TFs: GO analysis of TF targets, identify several new predictions of TF functions. Ex. Atf1 targets are enriched with chaperon, experimentally show that $\Delta atf1$ strain has growth defect at high temperature.
  - Direct vs indirect regulation: not observe any indicatione that indirect transcriptional or post-transcriptional regulation was responsible for the effects of a transcription factor deletion (testing by comparing target sets of different TFs: if indirect regulation, should observe large overlaps). This suggests that during normal growth, regulation by a transcription factor is not propagated appreciably via extended cascades involving other transcription factors or by indirect regulatory steps.
  - Binding alone may not be functional: many promoters occupied by transcription factors under normal growth conditions are not actively regulated, pointing to extensive control of transcription at a step distinct from transcription factor binding. E.g. 28% of the targets of HSF1 that were clearly occupied by Hsf1 and activated by temperature shift were not activated by overexpression of Hsf1.

TRN mediating mammalian pathogen response [Amit & Regev, Science, 2009]:

- Problem: TRN responsive to infections by pathogens, in particular, how specificity to different pathol-gen is achieved.
- Idea: first identify a list of candidate regulators, then do pertubation on each regulator: a true target will change expression in response.
- Methods:
  - Initial TRN reconstruction on transcriptional profiles: a modified version of Module Networks: L2-regularized linear regression of the mean profile of a module with a combination of candidate regulators.

  - Candidate regulator list: all regulators identified by modified Module Networks and additional regulators from CRE analysis (motifs enriched in the responsive genes).
  - Marker gene list: (select genes for expression measuring using nCounter) genes whose expression are most discriminative of different pathogen responses.
  - Perturbation: shRNA knockdowns on 125 candidate regulators, and measure the expression of marker genes through nCounter.
  - Determining regulatory links: a target should change expression in response to regulator knockdowns: reduce expression for activator knockdown and increase expression for repressor knockdown. To assess the significance of expression changes: (1) control for gene-specific noise by comparing changes to perturbation of control shRNAs; (2) control for shRNA-specific noise by comparing to changes in the expression of control genes.

- Results:
  - Transcriptional profiling of immune dendritic cells (DCs) in response to five different pathogenes (multiple time points): two main groups of responsive genes: anti-viral and inflammatory, inaidtion, IPC-specific and shared (across all five).
  - TRN: 125 regulators, 118 marker genes. On averague, one gene is activated by 14 regulators, and repressed by 5.
  - Analysis of the regulatory control programs of anti-viral response and inflammatory response: 33 regulators for each response, and the rest common to both. Cross-inhibition between the two control programs.
  - Inflammatory response: feed-forward loops which may ensure response to persistent, instead of sporadic signals. Anti-viral response: a two-tiered circuit involving feedback and feedforward loops, implicating a module of cell-cycle regulators.

3.10 Integrated Epigenomic Analysis

The PsychENCODE Project [Akbarian & Sestan, NN, 2015]

- Disease focus: ASD, BP and SCZ.
- Brain regions, developmental periods and cell types:
  - Regions: cerebral neocortex, hippocampus, amygdala, caudate nucleus, nucleus accumbens and cerebellar cortex.
  - Developmental periods: developing brain including prenatal and early postnatal brain.
  - Cell types: neurons vs. non-neurons using fluorescence-activiated nuclear sorting. Also use iPSC-derived neurons and CNON cells (cultured neuronal cells derived from olfactory neuroepithelium).
  - Rhesus macaque and chimpanzee.
- Genomic data: (1) H3K4me3, H2K27ac and ATAC-seq; (2) 3C and Hi-C; (3) Enhancer activity (STARR-seq); (4) Methylation: BiS-seq; (5) Nucleosome occupancy: NOMe-seq; (6) Ribosome profiling, RNA-seq, micro-western.
- Data analysis: (1) CREs and genes; (2) QTL and GWAS overlap; (3) case-control comparison of expression and epigenome.
- Functional characterization of findings using mouse and iPSC-derived neurons.

An integrated encyclopedia of DNA elements in the human genome [ENCODE, Nature, 2012]

- Summary of findings:
  - The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.
  - Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.
- Promoter analysis:
  - Two types of promoters: (1) broad, CpG rich TATA-less promoters; (2) narrow, TATA-box-containing promoters. Distinct histone modification patterns and selective TFBSs.
  - Predictive model of gene expression (CAGE): Two-stage model: (1) classification by random forest; (2) linear regression: if the label from the classification algorithm is 0, predict 0, otherwise, use the linear model. The performance is evaluated by cross-validation.
  - Remark: the two-phase model is necessary to fit the data: many genes have expression close to 0; and the expression of the rest vary quantitatively (Figure 2A, many genes have observed expression below 1e-5).
  - Histone modification as features to predict expression in K562 cells (leukemia cell line): AUC=0.95, regression r = 0.78. Classification: H3K9ac, H3K4me3, H3K4me2; regression: H3K79me2, H3K9ac, H3K4me3. Repression markers, if removed, have small impact on explaining expression variation.
  - TF binding as features: classification AUC=0.89, r=0.62. Most of this correlation could be recapitulated by the aggregate binding of transcription factors.
  - Remark: the model does not say enhancers are not important, in fact, enhancers act to change promoters states, to change expression, so promoter information could be sufficient to explain expression.
- Histone modification patterns around TF binding sites:
  - Clustered aggregation plots (CAGT) for unsupervised HM pattern discovery: (1) for a target mark, first classify the entire profile (target mark values across all sites) into high and low categories (by the strength of signal), and focus only on the high-profile. (2) Clustering analysis for the high-signal profile: k-means (a large number of potentially redundant clusters) followed by hierarchical clustering (merging), with possible flipping (two strands). For the k-means, the distance is defined as one minus the Pearson correlation coefficient.
  - Around TFBSs, histone markers typically show spatial, asymmetric patterns. Ex. H3K27me3 (repression marker) near CTCF binding sites: in 17.6% cases, flanking H3K37me3 signals, which can be further divided into 6 patterns (found by an unsupervised tool).
- Asymmetry is found for most histone markers, but not for DNase. Nucleosome is strongly asymmetric near TSS.
  - Hypothesis: transcription factors can form barriers around which nucleosomes and histone modifications are arranged in a variety of configurations.
- TF co-association analysis:
  - TF co-association: given the set X and Y, define the test statistic as $\|X$ intersect $Y\|/\|X\|$ (asymmetric), then calculate its significance. This is done through sampling to get a background distribution of base overlap ratios for random regions given the distribution of binding peaks in the two datasets. Block sampling instead of bp sampling. Also the segmentation component of the sampling procedure was to take into account the fact that TF binding sites are not uniformly distributed in the whole genome, but rather highly correlated with DNase signals. (Idea: random sample X and Y peaks, within the DNase regions.)

  - About 3,000 significant pairs in tie 1 and 2 cells. Some associations are general, some specific to promoters/intergenic regions.

- Genome segmentation: use histone markers, DNase, Pol II.
  - Segmentation by ChromHMM: (1) discretization of signals; (2) HMM, emitting multiple independent Bernoulli RVs. 200bp windows.

- Segmentation by Segway: (1) Motivation: control the duration; missing data and higher-resolution (bp level). DBN structure is more flexible, e.g. modeling of multiple hidden RVs. (2) Model: Q_t is the segment label (enhancer, promoter, gene body, etc.). The observation $x_t|Q_t$. Instead of modeling $Q_t|Q_{t-1}$ (HMM), we assume a variable J_t that controls whether Q_t changes or not; and J_t depends on Q_{t-1} and C_{t-1} (the countdown variable that controls segment length).
  - Integration of the two segmentation methods established a consensus set of seven major classes of genome states. TSS, PF (promoter flanking regions), T (transcribed, H3K36me3 transcriptional elongation signal), two enhancer states E and WE (weak/inactive enhancers) - differ in activation markers such as H3K27ac, CTCF (multifunctional, not just insulator sites), R (repressed, H3K27me3 polycomb-enriched regions).
  - CTCF-binding-associated state is relatively invariant across cell types. E and T states have substantial cell-specific behaviour.
  - RNA associated with the states: Polyadenylated RNA is heavily enriched in gene bodies. promoter-associated short RNAs.
  - DNA methylation pattern in seven states: hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

- SOM clustering of segments:
  - SOM: nonlinear generalization of PCA.
  - Based on 72 features (12 ChIP-seq and DNase-seq assays in six cell types). Rearrange the segments in 2D space s.t. the similar ones are placed close to each other.
  - Using SOM clusters for functional analysis of genes: Many of the ultra-fine-grained state classifications revealed in the SOM are associated with specific gene ontology (GO) terms. Ex. TSS clusters enriched with immune response or TF activity.
  - High-resolution clustering: ex. two neighboring SOM units, corresponding to TSS clusters, both have similar H3K27me3 in ESC, but different level in HUVECs. The two have functions in neuronal cells and body patterning, respectively.

- Questions/comments:
  - TF co-associations: different models: physical interactions or independently acting factors? Ex. distance and orientation preference.
  - Segmentation algorithm: training? How to control for the number of states s.t. they correspond to promoters, active enhancers, etc.? But we have other readouts, expression levels, DNA methylation. Can we create a supervised model that relates the distinct histone patterns (latent variables) with expression, DNA methylation, TF binding?

Predicting cell-type-specific gene expression from regions of open chromatin [Natarajan & Ohler, GR, 2012]

- Data: DHS of 19 cell types from ENCODE, in each cell type, DHS occupies about 2% of genome. 28% of DHS overlap with CTCF.

- DHS patterns:
  - DHS in TSS, gene body and intergenic regions. Median size around 300 bp, with DHS in TSS bigger.
  - GC content of DHS around TSS is much higher (median 0.8) than DHS elsewhere (0.28).
  - Only 14% of DHS are cell-type specific (only in one cell type, or overlap less than 50% with other DHS).
  - However, DHSs at TSS are generally not cell type specific (less than 1%). Many genes have open TSS across all 19 cell types (8393).
- Different classes of genes: pattern of GC and chromatin accessibility
  - Define gene expression in each cell type (microarray data): Z score for each gene from the expression across all cell types.
  - Constitutively active genes: 168 genes with $Z < 1.7$ in all cell types. Up-regulated genes (UR): high Z score in one cell type, and moderate Z in other cell types, and similarly Down-regulated (DR) genes.
  - GC content: Constitutively expressed genes displayed a particularly high CG content in their proximal promoter regions
  - Accessibility of TSS: UR genes in a small number of cell types likely maintain a closed chromatin conformation until cellular processes require up-regulation. In contrast DR genes may be viewed as constitutively expressed genes that are repressed in a single cell type.
- Classifying gene expression patterns:
  - For each cell type: we define several classes, UR, constitutive, etc, and the goal is to classify different classes. For each gene, we define features as: TF association to that gene. The TF association is defined by scanning PWMs to the DHS of that gene (either the best DHS, or two separate features: one for distal, one TSS). To link DHS to gene, use the nearest pair.
  - Results: using only features from TSS, the classifier (sparse logistic regression) performance is close to random. Using distal DHS improves performance, median AUC across all cell types is 0.73.
  - Also find putative regulators that distinguish different classes in a given cell type.
- Lessons:
  - Genes often have broad expression, as evidenced by open chromatin at TSS in half of genes across all cell types. Enhancers tend to be more cell-type specific, but overall most of DHS are active in more than one cell type.
  - GC content highly correlate with gene expression.

Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions [Sheffield & Furey, GR, 2013]

- Motivations/problems: from DHS and gene expression data of many cell types
  - Identify the DHSs of similar activity profiles (i.e. active in similar tissues)
  - Find the TFs regulating the clusters of DHSs
  - Associate target genes with DHSs
- Data: DHS and expression data of 112 samples, represeing 72 unique cell types and 15 tissue lineages.

  - DHS data were generated in two labs, so use ComBat to remove the batch effect.
  - DHS data were quantile normalized for each sample.

- DHS cluster analysis:
  - For each DHS, define its activity profile across all samples, then do SOM clustering.
  - SOM clusters: for multi-cell type clusters, the multiple cell types involved often are known to be related, there are exceptions though.
  - Properties of SOM clusters: promoter overlap, CpG island and conservation, these properties tend to correlate. Ex. SOM clusters overlapping with promoters tend to be less conserved.

- Identifying TFs of DHS clusters: de novo motif analysis of each SOM cluster and compare the results with JASPAR.

- Target genes of DHSs:
  - Pearson correlation of quantile normalized DHS and expression data. Each DHS is compared to all genes with 100 kb.
  - Defining significance: to obtain the null distribution of the Pearson correlation, for each gene, use random DHS across the genome to derive the null distribution. Only $P < .05$ DHS-gene pairs are considered significant.
  - Among all 2.7M DHSs, 20% correlate significantly with at least one gene in 100 kb. The majority of these DHSs (71%) correlate with only 1 gene, but some can be as large as 44.

- Lessons:
  - Heterogeneity of regulation: generally, DHSs (CREs) in one sample are very heterogenenous in terms of regulatory mechanisms, so it is better to divide them into different groups. One way of doing that is through activity profiles across multiple samples.
  - Representation/visualization technique: in analysis across many samples, group samples by broad tissue origins, e.g. brain, stem cell, endothelial, fibroblast, and these can be color-coded in a figure, e.g. activity profile.
  - Representation/visualization technique: Tie-plot, representation of relationship between multiple genomic elements (CREs, genes).
  - Representation/visualization technique: scatter-plot to represent multiple observations, say (x, y), where each observation may be a cluster, a functional element, etc. And by using size/color of dots, the scatter plot can also represent (x, y, z)'s.

A promoter-level mammalian expression atlas. [FANTOM Consortium, Nature 2014]

- Background: Most genes have more than one TSS, and the regulatory inputs that determine TSS choice and activity are diverse and complex.

- CAGE: cap analysis of gene expression `http://fantom.gsc.riken.jp/protocols/basic.html`. Method: start from mRNA, reverse transcription to obtain cDNA (digest mRNA), then cap-trapping to obtain 5'end sequence (linker). The last step is single molecule sequencing (no PCR amplification) of 27 nt sequence (called tag). One sample, sequencing of 4M tags.

- Data: 573 human primary cell samples, 152 human post-mortem tissues and and 250 different cancer cell lines. Also a large number of mouse samples, including developing tissues.

- CAGE peak identification: map tags to the genome, in general, multiple peaks in one sample. To define all alternative TSS, use DPI (deconvolusion peak identification) to all samples. Figure 1b shows an example of a gene with 6 peaks (spanning about 250bp) in 4 cell types/lies, with different cells prefer different peaks.

  - Peak quantification: normalized tag per million reads (TPM), using edgeR.
  - A total of 185K peaks across all samples (105K at TPM > 10) and about 73K peaks in all primary cells.

- Promoter architecture: some promoters are broad, some sharp, based on the spread of TSS along the genome. Found periodic spacing of WW motifs with a 10.5 bp repeat downstream of the dominant TSS (WW suggests nucleosome binding).

- Tissue specificity of transcripts:

  - 185K peaks: non-ubiquitous (cell-type restricted, 80%), ubiquitous-uniform (housekeeping, 6%) or ubiquitous non-uniform (14%). Ubiquitous: detected in more than 50% of samples (median > 0.2 TPM) and uniform as a less than tenfold difference between maximum and median expression.
  - Gene set enrichment: in housekeeping genes, ribonucleoprotein complex and RNA processing. In ubiquitous non-uniform genes, enriched for cell cycle genes.
  - An average primary cell sample expressed a median of 8,757 including peaks.

- Features of cell-type specific promoters:

  - CpG island (CGI) promoters are often assumed to be ubiquitious, but found that 50% of CGI and most non-CGI-non-TATA promoters (92%) had non-ubiquitous expression profiles.
  - Cell-type-specific promoters of both classes were enriched for binding of cell-type-specific transcription factors and both classes tend to have proximal high-specificity enhancers.

- Cell-type specific TFs:

  - Defining TF cell type specificity: based on their promoter expression in the sample relative to the median across the collection.
  - In any primary cell type, median of 430 (306 to 722) TFs were expressed at 10TPM or above. The number of robust promoter peaks per TF was similar to coding genes (4.8 compared to 4.6).
  - A clear connection between tissue-specific TF expression profiles and relevant phenotypes. Ex. in mouse inner ear hair cells: about half of top genes have some phenotype (knockout in mice or human diseases).

- Expression clustering of promoters and functional annotation: 120K promoters clustered into 4,000 groups, then co-expression between these groups (Figure 4). To do functional analysis: enrichment of ontology terms in the samples containing each of the promoter in a cluster.

- Questions:

  - If a gene is transcribed from alternative peaks in one sample, will there be one or multiple transcripts, or proteins?
  - Cell type specificity analysis of genes are done at the promoter level. How would the results change if we look at the gene level? Probably more ubiquitious.

- Lessons:

  - The cancer cell lines generally fail to cluster in a sample to- sample correlation graph with their supposed cell type or tissue of origin, and express more TFs than primary cells.
  - Enrichment of TF expression in a certain cell type can be used to assess the functional importance of the TF.

An atlas of active enhancers across human cell types and tissues. [FANTOM, Nature 2014]

- Using eRNAs to establish active enhancers:
  - Bidirectional capped RNAs is a signature feature of active enhancers: overlay CAGE tags on ENCODE enhancers (H3K27ac and H3K4me1 with p300 binding). On average, the inititation in two directions are separated by 180 bp. On this basis, we identified 43K enhancers across 808 human CAGE libraries.
  - Validation: 70% of enhancer candidates show activity in reporter assay. In contrast, only 20-30% of untranscribed candidates from ENCODE (based on DHS, or histone marks) were found to be active in reporter assay.
- Inititation and fate of transcribed enhancer RNAs:
  - Unspliced and typically short (median 346 nucleotides). No evidence of associated downstream RNA processing motifs. Most are non-polyadenylated. Resembling antisense promoter upstream transcripts (PROMPTs).
  - RNA Pol 2 inititation, de novo motif analysis revealed sequence signatures in CAGE-defined enhancers closely resembling non-CGI promoters.
  - Capped enhancer RNAs might be rapidly degraded by the exosome.
- Comparison of CAGE enhancers, DHS and H3K4me1/H3K27ac-defined enhancers in the same cell types:
  - CAGE-defined enhancers were strongly supported by proximal H3K4me1/H3K27ac peaks (71%) and DHSs (87%).
  - Only 4% of DHSs overlapped CAGE-defined enhancers. Only 11% of H3K4me1/H3K27ac loci overlapped CAGE-defined enhancers.
  - H3K4me1 and H3K27ac supported only 24% of DHSs distal to promoters and exons.
- Clustering of cell types by enhancer expression: generally grouped functionally related samples together. Although fetal and adult tissue often grouped together, two large fetal-specific clusters were identified: one brain specific and one with diverse tissues.
- Tissue specificity of enhancers:
  - The majority of detected enhancers within any facet (mutually exclusive cell type and organ/tissue groups) are not restricted to that facet. Exceptions: higher fraction of specific enhancers include immune cells, neurons, neural stem cells and hepatocytes amongst the cell-type facets, and brain, blood, liver and testis amongst the organ/tissue facets.
  - Enhancers are more generally detected in a much smaller subset of samples than mRNA. In any cell type, the number of detected expressed gene transcripts (> 1 TPM) is 19-34 fold larger than the number of detected enhancers.
  - Also a set of ubiquitious enhancers (200 or 247, defined by primary cell or tissue facets): expressed in the large majority of facets. They are 8 times more likely to overlap CGIs and they are twice as conserved.
- Linking enhancers to TSS:
  - Pairwise expression correlation (normalized TPM) between enhancers and TSS (within 500kb): use Pearson correlation. 64% of enhancers have at least one correlated TSS, and nearly half (40%) of enhancers were linked with the nearest TSS.
  - Overlap with ChIA-PET data: 20% overlap of ChIA-PET pairs, comparing with 4% for pairs defined by DHS correlations.

  - On average, a RefSeq TSS was associated with 4.9 enhancers and an enhancer with 2.4 TSSs.
  - Predictive model of expression: use 10 nearest enhancers of a TSS as predictors, quantify the importance of a predictor by proportional contribution to the variance. Focus on 2.2K TSS with high predictive power, $R^2 \geq 0.5$. Using shrinkage model: only one to three enhancers are necessary to explain the expression variance observed in the linked gene, and generally proximal enhancers are more predictive than distal ones.
  - Enhancer redundancy: defined as multiple enhancers of the same gene whose expression patterns are correlated. Correlation of level of redundancy of a TSS with its expression level (additive effect).

- Enrichment of disease-associated SNPs in enhancers: over-represented in regulatory regions to a greater extent than in exons.

- Questions:
  - Defining enhancers: how? Does it take into account the features such as the peaks are separated by around 200bp?
  - Hypothesis: high activity enhancers evolve sequence features similar to non-CGI promoters to better attract Pol 2, which facilitate E-P interactions and transcription inititation.
  - Exact relationship between H3K27ac and active enhancers (CAGE-defined)? The overlap seems too low (11%).
  - Association of one enhancer on average with 2.4 TSS: different genes? To what extent this is due to correlation of TSS?

- Lessons:
  - Among DHS, only a small minority correspond to active enhancers. H3K27ac loci probably have a higher fraction: at least 11% (probably much higher).
  - Tissue specificity of enhancers: most are promiscuous (multiple facets), a small group (200 or so) are ubiquitous.

3.11 Non-coding RNAs

3.11.1 Small Non-coding RNAs

On the art of identifying effective and specific siRNAs [Pei & Tuschl, NM, 2006]

- Criteria for designing siRNA targeting a gene: effective and specific (miminize off-target effects).

- Additional details of siRNA mechanism: 21-23 nt RNA with two 3' overhangs. An siRNA is designed for full complementarity with target mRNA. RISC cleaves the mRNA at a site precisely 10 nt upstream of the nt opposite of 5'-end of the guide strand.

- Choosing target sequences in mRNA: generally in coding sequences, but the UTRs may also be OK. Other constraints: target orthologs in more than one species; all possible splice variants of a gene.

- Consideration of sequence composition: the sequence could influence the efficiency of the three main steps: siRNA-duplex stability, RISC loading/assembly (require cleavage of the passenger strand), mRNA binding.
  - Asymmetry of siRNA duplex: determined by the sequence composition - thermodynamic stability.
  - GC content (30-52%): low GC may desatabilize siRNA duplex and reduce mRNA binding; but high GC may impede RISC loading.

  - Center of duplex: prefer low internal stability, likely because of the requirement of cleaveage of passenger strand and release.
  - Other single nt positional preference.

- Consideration of accessbility of target mRNA: local secondary structure (stem-loops) may restrict the accessbility of RISC, and reduce siRNA efficacy.

- Specificity: two kinds of off-targets (imperfect base-pairing), (1) share complementarity over half of the siRNA sequence; (2) 6 to 7 nt perfect match in 3'UTR with positions 2-7 or 2-8 of guide strand of siRNA. The current strategy is to have mismatches with any undesired targets.

Bioinformatics of siRNA Design [Hakim Tafer, MMB, 2014]

- The strand with the lower 5 stability was preferentially incorporated into the RISC.

- The relative stability of the siRNA ends was a major determinant of the functionality of siRNAs.

- The impact of target mRNA structure by thermodynamic analysis: the totoal binding energy between siRNA and mRNA is the sum of breaking energying (mRNA secondary structure) and the hybridization energy (base-pairing).
  - Problems: free solutes, RISC ignored.
  - siRNA efficiency is directly correlated to the target site accessibility.

- Sequence based siRNA design:
  - 21 nt duplexes with 2 nt overhang and low GC content were highly efficient
  - siRNAs must be asymmetric in order for the guide strand to be introduced in the RISC (weaker 5 end binding)
  - Reynolds et al [NBT, 2004]: a lack of structure in the guide siRNA favors repression efficiency. Position-specific nucleotide preferences found (also in other studies).
  - Machine-learning methods: ANN or Lasso. Recovering the asymmetry as well as the presence of U at position 10 of functional siRNAs.

- Accessibility-Aided siRNA Design:
  - Sirna: sample the secondary structure of target mRNA, then estimate the accesbility of sites. Combine accessibility criterion with other rules. In an extension, use the energetic cost instead of accessibility.
  - OligoWalk: similar to Sirna, but uses exact computation (partition function) instead of sampling.
  - RNAxs: similar Accessbility model of OligoWalk. The best two design criteria turned out to be **asymmetry** of the siRNA and **target site accessibility**. The best combination is: these two plus **self-folding** (foldingenergy of the siRNA [15]), and **free-end** (folding structure of the siRNA [15]) criteria.

3.11.2 Long Non-coding RNAs

lncRNA [John Rinn lecture on YouTube, 2014]

- lncRNA: similar to protein coding sequences, it has multiple segments stitched together, and polyA tail. Locations: intronic, antisense (in coding), bidirectional, intergenic. About 8,000 lncRNAs. Typically low expression and low conservation.

- Functions of lncRNA in reprogramming. lnc-ROR is essential for reprogramming of fibroblast to stem cells.

- Screening for functions of lncRNA: from 400 disease loci where no protein-coding genes are present, use filters including no expression and mouse orthology, find 18 candidate lncRNA loci. KO in mouse shows that 3 are required for life.
  - FENDRR KO in mouse: lethal, expressed specifically in lung, and may be invovled in ACD/MPV (a disease).
- Molecular mechanisms of lncRNA: four types of interactions, A: RNA-protein, B: RNA-DNA, C: protein-DNA, D: RNA-RNA. lncRNA can act through these modes:
  - RNA helps protein bind to DNA through RNA-protein and RNA-DNA interactions (A + B).
  - Scaffolding: RNA links multiple proteins, which then bind to DNA (A + C).
  - RNA-protein complex (A+C): e.g. ribosome.
  - Complex of protein, RNA and DNA (A + B + C): e.g. telemerase.
- Studying function of FIRRE: a lncRNA in X-chr.
  - Cellular localization: use visulization, in nucleus.
  - Experimental technique to determine RNA binding to DNA: RNA-DNA hybrid $\rightarrow$ capture using probes bound to RNA $\rightarrow$ washing (remove RNA) $\rightarrow$ sequencing DNA.
  - RNA binding spread to 5M bp across X-chr, and also in 5 other chromsome (physically close).
  - K.O. of FIRRE: the 5 loci no longer localized, and the stem cells become sick.
- Summary of lncRNA function: RNA-mediated organization of chromtins. The significance include: enhancer-promoter interactions, and gene-gene proximty, which may influcence cell fate.

An atlas of human long non-coding RNAs with accurate 5' ends [Hon & Forrest, Nature, 2017]

- Background: lncRNA have low abundance, not highly constrained, unstable, and often originate from promoters or enhancers. At some loci, function lies not in the transcribed units, but the act of transcription itself.
- Define LncRNA atlas: FANTOM CAGE associated transcript model (FANTOM CAT). Link CAGE clusters with transcripts, resulting in transcripts with accurate 5' ends. Total 50K transcripts, and 27K are lncRNAs.
- Classes of lncRNAs: about 70% originate in DHS. 40% are e-lncRNAs: DHS have enhancer features. 30% overlap with DHS with promoter features: among them, most (6K) are divergently transcribed from promoters of mRNA (divergent p-lncRNA) and about 1.7K are transcribed from unknown promoters (intergenic p-lncRNA).
- LncRNA conservation: study both TIR (initiation region) and exon conservation. In general, they are less conserved than mRNAs, and TIRs are more conserved than exons. But show signs of selection. Also a large fraction of TIRs, 74% for e-lncRNA and 56% for intergenic p-lncRNA, overlap retrotransposons. Also compare the activity across tissues: 50% p-lncRNAs are conserved across mammals, while only 20% e-lncRNAs are.
- LncRNA expression specificity: tissue-specific, 11% (5,600) lncRNAs are expressed in a facet.
- Relation to GWAS: find cell-type enriched genes and trait-associated genes (PISC). Then find significant cell type-trait pairs. Some of these pairs, e.g. middle temporal gyrus and ASD is driven by 18 lncRNAs.
- Enrichment of e-lncRNA of eQTL: also find correlation between e-lncRNA and mRNA that are linked by eQTL. Similar correlations were found between mRNA-mRNA and lncRNA-lncRNA, suggesting some general mechanism of transcriptional coregulation.

- Q: For e-lncRNA and intergenic p-lncRNA, often have pairs of transcribed units?
- Remark: DHS regions are enriched with enhancers, and they are known to be conserved. So conservation signal here does not prove the functional significance of lncRNA. Similar things can be said about GWAS/eQTL enrichment.

Non-coding RNA: More uses for genomic junk [Nature, 2017]

- Recruitment of CBP: The acetyltransferase enzyme CBP can be recruited by transcription factors (TFs) to regions of chromatin, however, CBP has limited enzymatic activity owing to auto-inhibition.
- Importance of eRNA: production of RNA by Pol II close to CBP can relieve this auto-inhibition, through RNA-CBP binding.
- Another reference: RNA Binding to CBP Stimulates Histone Acetylation and Transcription [Cell, 2017]

CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells [Liu & Lim, Science, 2017]

- CRISPRi library: targeting TSS of 17,000 lncRNAs. Test in 7 human cell lines including iPSCs.
- A considerable number (500) of the tested lncRNAs influenced cell growth. In almost all cases, though, the function was highly cell typespecific, often limited to just one cell type.

Chapter 4

Post-transcriptional and Translational Control

Post-transcriptional regulation in corticogenesis: how RNA-binding proteins help build the brain [Pilaz and Silver, Wiley Interdiscip Rev RNA, 2015]

- Process of cortical development (Figure 1): (1) NPCs: Neuroepithelial cells (NE) undergo symmetric cell division; Radial glia cells (RGCs) undergo assymmetric cell division, RGCs + IPC or neuron. (2) Neuron migration along the basal process (from IPCs) as a scaffold: from both RGCs and IPCs, and differentiation during migration. (3) Excitatory neurons reach the Cortical plate. Young neurons pass early neurons in migration, so deep layers form first/earlier.
- Key processes affected by RBPs (Figure 3): NPCs - proliferation (cell cycle), differentiation and apoptosis. Neurons: migration, differentiation, maturation and apoptosis.
- Translation control by FMRP: (1) Translation repression: at the synapses. Alleviating FRMP repression allows quick production of proteins. (2) NPC (IPC) differentiation: FMRP KO mice have more IPCs. Likely mediated by Profilin1 (PFN1). (3) Neuron migration: mediated via N-cadherin. (4) Translation activation of NOS1, a regulator of synapse.
- Translation control by EIF4E: EIF4E family protein may both promote or inhibit translation. In one case: repression of Neurog1/2 and NeuroD1, lineage differentiation factors of neurons. Model: NSCs are preloaded with mRNAs encoding differentiation factors, but their translation is repressed by EIF4E.
- RNA localization: (1) at the synapse is important for neuron functions, e.g. memory formation. (2) Asymmetric localization in NPCs: fate determinant, mediated via Cyclin D2.
- Lesson: the function of a RBP may be highly heterogeneous, at the biochemical level, can activate or repress translation; at the physiological level, may affect multiple processes such as NPC differentiation, and neuron migration. However, the mechanisms of these heterogeneity are unclear.
- Lesson: PTR is important for corticogenesis in several aspects such as (1) Differentiation: repression of translation of differentiation factors. (2) Localization in synapse: fast responses.

The role of RNA processing in translating genotype to phenotype [Manning, NRM, 2017]

- Principle: genetic variants affect cis-acting elements in RNAs, then affect RNA processing. The key is to identify these cis-acting elements, and their functions on different aspects of RNA processing.
- RNA processing at two levels (Figure 1): (1) Pre-mRNA level: splicing and 3 end processing. (2) mRNA level: stability, localization and translation.

- Splicing: core bases in the exon-intron boundary, consensus splice sites including branch sites, intronic/exonic splicing enhancers/silencers (ESE, ISE, ESI, ISI). The splicing elements are often bound by splicing factors, e.g. RBFOX1.
- 3 end processing: alternative polyadenylation (APA). May result from alternative splicing of the terminal exons (different 3 UTRs), or from APA at the last exon (different 3 UTR length).
- MiRNA-mediated regulation: (1) Changes of miRNA expression have been associated to diseases. (2) Genetic variants may change cis-elements: different 3 UTRs (e.g. from APA), or miRNA binding sites.
- RNA stability and structure (Figure 4): riboSnitches may affect the RSS of cis-elements in RNA, affecting miRNA or other trans-acting factors (e.g. make miRNA binding sites inaccessible).
- RNA translation (Figure 4): genetic variants may change the core RP binding sites in 5 UTR, or RSS (hairpins can delay translation). Synonymous SNVs are likely important because they are less evolutionarily constrained. Ex. CFTR gene: the 3-base deletion removes one AA, but also changes the last base of one AA (syn. change). Shown that the syn. change affects the degradation of mRNA (binding of one protein).
- RNA localization: variants may affect binding of RBPs involved in transport, e.g. transport of BDNF to the distal dendrites.
- Approach for mapping cis-acting elements: in addition to CLIP-seq, ASE can be useful. Not all ASEs are driven by CREs regulating transcription.
- Lesson: Importance of cis-acting elements, binding sites of RBPs or miRNAs or basic machinery for 3 end processing/translation. The change of RSS by genetic variants should be generally evaluated at the cis-elements (e.g. whether a RSS change affect binding site of miRNA - the change may happen elsewhere, not in the exact miRNA binding site). RSS change may also change the rate of translation (not mediated via binding of specific trans-acting factors).
- Lesson: RBPs are heterogenous groups of proteins, involved in various aspects of RNA processing: 3 end processing, core splicing machinery, splicing factors, stability regulation (e.g. NMD), transport/localization, RISC, and so on.

Finding function in mystery transcripts [Nature, 2016]

- Problem: functions of lncRNAs. Ex. in ESC, identified 2000 lncRNA, 148 are exclusive to ESCs. Hints that they are functional: evolutionary conservation.
- Genetic manipulation: CRISPR-cas9. Often lncRNA are less sensitive to small changes of sequences, so need large cuts. Often may not find observable effects.
- RNA interactomes: cross-link RNA and proteins by formaldehyde or UV, then mass-spec. Found large structure from Xist complex.
- Structure: use chemical probes (SHAPE). Attach acetyl group to flexible regions, which will block reverse transcriptase. So the DNA sequences will be many fragments instead of long strands.

RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins [Mao & Sun, NAR, 2016]

- Background: CLIP-seq and RIP-seq are two major technologies for measuring RBP-mRNA interactions. In human, about 15% SNVs are predicted to alter RNA secondary structure, called riboSNitches.
- rbSNV annotations: (1) RBP binding: 112 CLIP-seq datasets. (2) RBP motifs. (3) RiboSNitches: from the RNAfold program. (4) Impact on miRNA-RNA interactions. (5) eQTL.

- Classifying functional SNVs: using combination of these annotations. Strongest evidence: eQTL + RBP binding + other evidence (motif, riboSNitch or miRNA).

A Deep Learning Approach for Learning Intrinsic Protein-RNA Binding Preferences [Ben-Bassat and Orenstein, review for ISMB, 2018]

- Background: in vitro RBP-RNA binding, RNAcompete experiments, 240K synthetic RNAs of length 30-40 nucleotides.
- Background: methods for modeling RBP-RNA binding (1) RNAcontext: sequence information as PWM, structural information represented as a vector of the preferences to each structure context. (2) DeepBind: CNN, but sequence alone, 16 filters of length 16. (3) RCK: k-mer based model, both on sequences and on structure.
- Encoding structural information: for each position, predict its probabilities of being in five structure contexts, stem-loop (paired), hairpin, inner-loop, multi-loop, external. Structure can thus be represented as 5 numbers at each position.
- CNN: sequence and structural similar models. Convolution layer: 256 filters, 128 of length 5 and 128 of length 11. One pooling layer and one fully-connected layer (or 2).
- RNN: bidirectional RNN.
- In vitro results: validation using an independent RNAcompete data. (1) RNN slightly better than CNN. (2) CNN better than DeepBind, but not from structure (small effect), but from more filters and different lengths.
- In vivo results: validation with eCLIP-seq. Similar performance of all methods.
- Extracting sequence and structural features: for each filter, find the subsequences that lead to the strongest activation, then PFM on the strongest subsequences.
- Remark: the CNN/RNN may not properly capture the RNA structure. Ex. only a sequence motif in a particular structure context (e.g. hairpin) is possible to bind to some RBP. The two seems to interact in the CNN only at the fully connected layer (one convolution layer for each type of input).

Regulatory discrimination of mRNAs by FMRP controls mouse adult neural stem cell differentiation [Liu and Reitcher, PNAS, 2018] Commentary: Fragile equilibrium between translation and transcription [PNAS, 2018]

- Background: 750 RBPs in human genome, 200 of them have been implicated in diseases. FMRP is known as translation repressor.
- Experiment: k.o. of FMRP in NPCs, then do RNA-seq and ribosome profiling (measure translation efficiency, TE).
- Six groups of genes based on mRNA and TE changes (Figure 1): translation up, buffering up (translation up, and mRNA down to buffer); translation down, buffering down; mRNA up and mRNA down. About equal sizes (300-600 genes per group). Translation or mRNA changes are defined by simple rules, fold change > 1.2 and nominal $p < 0.05$.
- Different groups show different GO pathways or cellular components: e.g. synaptic gens are enriched in translation up and buffer up groups; genes related to cell adhesion and neurogenesis were enriched in the mRNA down group (defective neuron differentiation in FMR1 KO).
- Comparison of mRNA features of six groups: 5 UTR, 3 UTR, CDS length. Some groups show difference.

- Possible models of why mRNA can change: (1) FMRP may stabilize mRNA, via reading M6A (competing with other reader). (2) miRNA regulation: binds to miRNA sites. (3) Indirect effects on TFs. (4) Negative feedback levels on mRNA when translation changes.

A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation [Board and Seeling, Cell, 2019]

- Background: APA is regulated by PAS (polyadenylation signal). Each PAS has 6-mer core (CSE) with upstream and downstream sequence elements (USE and DSE).
- Experiment: 3M reporters. Design: variation of CSE, USE and DSE replaced by random sequences.
- APARNET: Figure 2A. Output: isoform produced from each PAS. Two convolution layers, interlaced with subsampling layers, a fully connected layer and logistic regression output.
- Interpreting APARNET: regulatory code of APA. Layer 1: known motifs of RBPs, e.g. CPSF. Layer 2: larger motifs or motif interactions, e.g. a RBP motif flanked by poly-T.
- Prediction of APA across cell types: even though training is done on data not tissue-specific, APARNET works well in data across many tissues (from GTEx). Only 3% of APA sites are tissue specific.
- Prediction of SNVs affecting APA: validation in Geuvadis data. Some examples: Figure 6D.
- Lesson: (1) Use CNN to understand regulatory grammar. (2) Post-transcriptional regulation may not be as tissue-specific as transcriptional regulation.

4.1 RNA Splicing

Research problems of RNA splicing [personal notes]

- Profile isoform variations across cell types and conditions.
- Understand cis- and trans- regulatory mechanisms of splicing. Q: do we have deep intronic splicing elements? What are good experimental technologies for mapping intronic elements?
- Understand the contribution of splicing dysregulation to human diseases.

The Expanding Landscape of Alternative Splicing Variation in Human Populations [Park and Xing, AJHG, 2018]

- Complex alternative splicing (AS) patterns: alternative first/last exons, multiple events joined together, e.g skipping of 2 consecutive exons.
- Technology for mapping AS: short reads. Long range coupling of exons can be lost (Figure 2B): e.g. A-B-C-D-E and A-C-E are two possible isoforms; but with short reads, one may determine A-B-C-E as an isoform.
- Technology for mapping AS: synthetic long reads, many mRNA pools s.t. each pool contain only one isoform (few molecules per pool, so on average, a gene has only one isoform), which can be reconstructed by de novo assembly.
- Quantifying AS: one strategy is isoform quantification. Another strategy is to focus on events: most common Percent Spliced In (PSI) for each exon.
- Splicing QTL: treat PSI of an exon as quantiative trait and do association. Modeling the errors of PSI can lead to improvement.

- Allele-specific AS: require expressed exonic SNP. Remark: ASE could be due to AS-AS.
- Example of sQTL: SP140, variant in exonic splice site that causes exon skipping. Another example, intronic variant disrupt 5 splice site and exposes a downstream splice site, leading to truncated variant (NMD).
- Experimental technology for variant effect: mini-gene, insert exon of interest, and measure PSI. MPRA.
- Prediction of AS events from sequence features: some informative features are exon/intron length, divisiability by three of introns, splice site strength.

MISO: Analysis and design of RNA sequencing experiments for identifying isoform regulation [Katz & Burge, NM, 2010]:

- Suppose we want to test if E is involved in an AS event (exclusion). If E is included in an AS event: we have A-E-B isoform, and if not, A-B isoform. The reads in the junctions A-E and E-B are inclusion reads, supporting E; while the reads in the junction A-B are exclusion reads, rejecting E. However, constitutive reads are still informative: if there is no splicing, the number of reads in A and B would imply the number of reads in A-E and E-B (or vice versa).
- A probabilistic model: infer the transcripts (AS events - PSI level of the isoforms) from the read data. Also could use the model for statistical testing (BF) of differential splicing (in two different conditions). The model has two parts: (1) the fragment pool: the number of fragment from one isofrom is proportional to its level and the length. (2) the sequencing step: the reads are uniformly chosen from the fragment pool.
- Notation: consider the n-th read, let R_n be its read alignment, and $I_n = k$ be the isoform it originates from. Let $m(k)$ be the number of mappable positions in the k-th isoform. Let Ψ be the fragment abundance, with Ψ_k be the abundance of the k-th isoform, which is proporotional to the percent (level) of the k-th isoform multiplied by its length.
- Single-end model: the prior distribution Ψ follows Dirichlet distribution. For the n-th read, the distribution $I_n|\Psi$ follows multinomial distribution. The probability of read sequence follows the uniform distribution, i.e. $P(R_n|I_n) = 1/m(k)$.
- Paired-end model: this is similar to the single-end model except that in the last part, we will model both R_n (read alignment) and λ_n the read length. We have:

$$P(\lambda_n|I_n = k, \mu) \sim N(\mu, \sigma^2) \tag{4.1}$$

  where μ is a parameter for empirical fragment length distribution. And $P(R_n, \lambda_n|I_n = k, \Theta) = 1$ or 0, depending on whether the read is consistent with the k-th isoform at position R_n (with length λ_n).
- Comparison with Cufflinks: overall model is similar, difference in the length model: how the number of fragments depends on the length of transcript.

The human splicing code reveals new insights into the genetic determinants of disease [Xiong and Frey, Science, 2015]

- Model: (1) predict the PSI for each exon. (2)Use sequence features near the exon: 300bp around exon, and 300bp in the upstream intron and 300bp in the downstream intron. Collect 1400 sequence features in these areas, including: 200 known motifs, 300 novel motifs, short motifs (1-3bp), transcript features including length, RSS; and nucleosome occupancy features. Note: exons tend to have higher nucleosome occupancy. (3) Train 2 layer NN to predict Ψ (high or low percent of inclusion) for each tissue from sequence features. The hidden layer has 30 variables, and the output layer 16 output (number of tissue types for model training)

- Training the model: 10K exons with evidence of alternative splicing. For each exon, we have the fraction of spliced exon Ψ. On average, AUC is 95.5%.
- Insights from the model: (1) mostly encompass the collective effects of known RBPs. (2) the same features can influence Ψ differently in different tissues.
- Application to predict splicing effects of SNVs: find $\delta\Psi$ for the SNV of each tissue, the aggregate $\delta\Psi$ across all tissues. Most high-scoring SNVs are intronic. Most SNVs with large effects (20k) are close to splice sites, but 465 intronic SNVs are more than 30 bp away.
- Validation of splicing prediction of SNVs: among 465 intronic SNVs, disease SNVs are 9 times enriched than CVs.
- ASD: 5 cases and 12 controls. Find genes predicted to be mis-regulated, and show that these genes are enriched with high brain expression, etc.

RNA editing in nascent RNA affects pre-mRNA splicing [GR, 2018]

- RNA editing: occurs in chromatin associated RNAs, prior to polyadenylation.
- Linking between RNA editing and splicing: 500 editing sites in the 3 acceptor sequences that can alter splicing of the associated exons. Also exons: splicing is modulated by RNA secondary structures that are recognized by the RNA editing machinery.

Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells [Gupta and Tilgner, NBT, 2018]

- Experimental approach (Figure 1A): full length cDNA library, with UMI and cell barcode. Then do both short-read (10x) sequencing and long-read sequencing. The cell barcodes are shared. Short read would provide cell type/cluster information.
- Data: mouse brain, clustering and cell type assignment. Assign cell types by marker genes. QC: confirm cell type proportions are similar across replicates.
- Long-read (PacBio) processing: the key is to correctly assign cell barcodes. Find polyA (9) tail in the within 200 bp of either read end assuming 2/30 error rate. 62% reads have T9. Overall, 58% poly-A containing reads can be assigned unique cell barcodes. Long read alignment: STARlong.
- Nanopore sequencing: lower rate of detecting cell barcodes. Homopolymer error is a problem.
- Discovery of novel isoforms: (1) Known splice sites: require all exons, and known splice sites. 10K. (2) Allow novel splice sites. Total 16K novel isoforms.
- Validation of cell-type specific isoforms: (1) Bulk RNA-seq: confirm the presence of isoforms. (2) Cell sorting and then do bulk-RNA seq.
- Example: Bin1, 6 new exons, A1-6. Clear cell type specific pattern (Figure 2f).
- Analysis of pattern of alternative splicing: coordinated alternative exons (two exons both expressed or not across cell types). Most are adjacent exons.
- Remark: important analysis steps (1) Long-read assignment to unique cell barcodes. (2) Validation of novel isoforms: bulk-RNA seq. (3) Characterization of cell-type specificity of novel isoforms.

Predicting Splicing from Primary Sequence with Deep Learning (SpliceAI) [Jaganathan, Cell, 2019]

- Model: use known isoforms (pre-mRNA transcripts) from GENCODE. From these isoforms, extract splice donors and acceptors. Take sequences up- and down-stream 5kb (test different length) of the base, predict its label: donor, acceptor or neither. Network architecture: ResNet.

- Evaluation of splice donor and acceptor site prediction with independent testing data (Fig. 1E): PRAUC 0.98. Using 1kb on each side obtains PRAUC 0.97, suggesting most signals are not too far.
- Evaluation in GTEx RNA-seq data using exon PSI: exons consitutively spliced in or out have scores close to 0 or 1; AS exons have scores from 0.1 to 0.9.
- Making sense of the prediction model (Fig. 1D): do in silico mutations, and assess how the acceptor site/donor site is affected by a nearby mutation. Clearly: the site itself is important (score 0.7), but other sequences can also be equally important.
- Importance of nucleosome occupancy: implicitly captured by the model. Fig. 1G: spliceAI scores correlate with experimentally measured nucleosome occupancy.
- Prediction of mutation effect (Figure 2A): compare reference pre-mRNA transcript vs. alternative transcript, prediction of acceptor site score and donor site scores in nearby 100 bases. Take the max. of acceptor/donor scores (both gain and loss) across all nearby sequences as the score of that mutation.
- Validation (Figure 2BC) using private mutations and novel splice junctions: GTEx, use reads to estimate novel splice junctions. Private mutations with high spliceAI scores are enriched with novel splice junctions: OR = 10 - 40.
- Estimating sensitivity of detecting crytpic splice variants (mutations outside acceptor/donor sites with effects): using novel junctions (Fig. 2F). About 71% in sites withine exons, and 40% in intronic positions.
- Cryptic splice sites: effects are likely tissue-specific. Using private cryptic splice sites. Fig. 3A: control, single isoform; individual with variant, two isoforms, and the ratio differs between tissues. In general, strong splice sites (score close to 1): not tissue-specific. Moderate: often tissue-specific effects.
- DNM burden of cryptic splice variants: overall OR = 1.3 for ASD, and 1.5 for DDD (Fig. 5A). With score 0.8, OR = 1.7 (Fig. 5C).
- Remark: Tissue-specific AS is reflected in the predicted acceptor/donor probabilities, with values close to 0 or 1 being constitutive isoforms. Similar, the tissue-specificity of variant effects is also reflected by the scores.

Deep-learning augmented RNA-seq analysis of transcript splicing (DARTS) [Zhang and Yi Xing, NM, 2019]

- Motivation: detection of differential splicing (DS) across conditions is difficult for lowly expressed genes. Use priors to improve the detection.
- Building prior model of DS: training sets are pairwise comparison of RNA-seq, and for each exon, a label of DS or not. Use these exon-label pairs as training data. Features: 2.5K cis-sequence features including conservation, motif, splice site strength, RSS. Trans- features: 1.4 RBPs in the sample compared (x2 for two samples). Train DNN: 4 layered.
- Detecting DS by Bayesian hypothesis testing (BHT): for each exon, count the reads that are skipped vs. exon-inclusion reads. The number of skipped reads follow Binomial distribution, with parameter PSI. Detecting difference of PSI between two conditions. H_0: the prior of the difference is small, H_1: larger. The prior variance is given.
- DARTS: the combined model use prediction from DNN as prior.
- Evaluation of DARTS in ENCODE/Roadmap data: training on all pairwise comparisons, and testing one leave-out sample. The model trained on both datasets performs best: AUC > 0.8.

- Application to EMT RNA-seq data: using only RNA-seq data (DARTS-flat) detects 77 events. Adding prior: 52 more events.
- The paper mostly discusses exon skipping, but training is also done on other events: alternative 5 and 3 splice sites, intron retention.

4.2 RNA Modification

RNA Epigenetics [Dominissini & He, The Scientist, 2016], Gene expression regulation mediated through reversible m6A RNA methylation [NRG, 2014]

- 140 RNA modification, with methylation of adenosine at the N6 position (M6A) being the most common in both mRNA and lncRNA.
- Experiment: M6A-seq, use m6a antibody, then sequencing.
- Distribution of M6A sites: total about 12,000. Most often in 3'UTR and slight enrichment in 5' UTR. 87% of exonic M6A sites are in long exons, > 400 nucleotides. Likely involved in splicing of long-exon transcripts - single isoform genes are relatively unmethylated.
- M6A motif: A/G-A/G-ACU, where the middle A is methylated. Motif of METLL3/14.
- Erases: FTO, ALKBH5. FTO: oxidative DNA and histone demethylation.
- Writers: METTL3, METTL14, and WTAP (intially discovered as a splicing factor).
- Readers mediate the effects of M6A, processing/splicing and export (nucleus), translation and decay (cytosol).
  - YTHDC1 regulate RNA splicing and nuclear RNA exportation.
  - YTHDF1 regulates translation: targets increase ribsome binding.
  - YTHDF2 promotes mRNA degradation (more than 3,000 cellular RNA targets).
- M6A conservation: many peaks conserved between human and mice, and distribution pattern/motifs conserved between yeast and human.
- Indirect reading mechanism: change of mRNA secondary structure. M6A weakens the duplex RNA, and may change the accessiblity of RNA-binding motifs (RBMs) - known as M6A switch. Evidence: m6A alters the local structure in mRNA and lncRNA to facilitate binding of HNRNPC, a nuclear RBP responsible for pre-mRNA processing [25719671]
- Hypothesis of M6A function: (1) Methylation-mediated sorting: marker to gruop and synchronzie cohorts of transcripts for fast-tracking mRNA processing; (2) affects cell-state transition.
- Phenotypic effects associated with M6A: ESC development, stress response and cancer.

M6A modification: mapping and control [Discussions and notes]:

- PAR-CLIP: METLL3/14 bound protein, pull down mRNA (cross-link, IP), then sequencing and calling. The same GGACU motif is detected.
- M6A-seq: usually find narrow peaks, < 200 bp. There are regions with large read counts, but relatively flat peaks.
- Possible to do isoform-specific calling: if the isoforms have distinct 3'UTR. Possible to have intronic M6A sites. But difficult to detect because of pre-mRNA are transient.

- RNA modification happens co-transcriptionally: interaction between RNA Pol 2 and METL3/14. This is similar to splicing (but splicing coudl also happen post-transcriptionally). The specificity of M6A may be controlled by TFs [Chuan's talk].
- De-methylation: happen mostly in nucleus. The enzymes could also move to cytoplasm, but function unclear.

Function of M6A [Discussions and notes]:

- Recognition of M6A by reader proteins: direct binding of reader proteins with M6A via YTH domain. Prefer M6A-motif (GGAC).
- Background: RNA-binding proteins (RBP). Some in nucleus: involved in splicing and export. Other RBPs: in cytoplasm.
- Export of mRNA: could have specificity, which is affected by M6A. Knockdown of reader protein for export (?), could observe accumulation of M6A in nucleus.
- Does a mRNA transcript get recognized by only one reader? Chuan's hypothesis: "fast track" model, sequential recognition. Faster export, faster translation, etc. M6A at different parts of a gene could have different functions. Ex. 5 UTR or coding: translation. 3 UTR: RNA stability.
- Hypothesis: M6A often occurs in long exons, possibly avoiding abortion of transcription prematurally.
- Role of YTHDF2 in development: maternal-zygotic transition (reduced expression of maternal RNA and increase of zygotic). YTDH2 KO shows delay of transition.
- M6A in Cancer: FTO is an oncogene. ALKBH5 correlates with glioblastoma poor outcome. METTL14: R298P reduced MT activity, it acts in dominant negative fashion. In many patients, lower level of M6A in endometrial cancer. Downstream target: AKT pathway.

Open questions about M6A modification [personal notes]:

- What controls M6A specificity in a cell? Motif, mRNA secondary structure, TFs?
- M6A readers: how they recognize target mRNAs? Do different readers recognize the same M6A or different M6A or different genes (one M6A site has only one function)? How do M6A readers work: recruiting additional RBPs?
- M6A switch: how widespread? Does it work independently of reader proteins?
- Effect sizes of M6A on export, splicing, translation and decay: how much faster? Similar across different genes?
- How does M6A pattern vary between cell types? Can we rationalize the difference (e.g. M6A targets are cell-type specific)? What may control the differences?
- Demethylation: if in nucleus, only affect newly transcribed mRNAs?

RNA modifications and structures cooperate to guide RNA-protein interactions. [Lewis and Kalsotra, NRMCB, 2017]

- Challenge: RBP binding motifs are common, what determines the specificity of their binding? Hypothesis: RNA modifications and structure determine RBP specificity.
- Evidence of M6A in splicing: depletion of writers and erasers has large effects on splicing. M6A colocalization with polyA (multiple sites), and with exonic splicing regulatory sequences.

- Possible mechanisms of M6A effects on splicing: (1) reader YTHDC1, recruits SRSF genes (splicing factors). (2) M6A increases binding of HNRNP gene, which changes splicing.
- M6A in mRNA degradation: M6A in 3 UTR can have opposing effects (1) reader YTHDF2, which facilitate transport of mRNA for processing, i.e. degradation. (2) Binding by HUR, which stabilizes mRNA by blocking miRNA binding.
- M6A in translation: (1) Reader YTHDF1, increases cap-dependent translation. (2) In stress conditions: normal transcripts cannot be transcribed (cap-dependent). In stress-related mRNA: M6A in 5 UTR. YTHDF2 translocates to nucleus, and prevent FTO effects. YTHDF2 recruits eIF3, and leads to cap-independent translation.
- RNA structure and modification can affect RBP and miRNA binding: (1) some RBPs prefer binding in a particular RNA secondary structure, e.g. RBFOX2. (2) RNA structure can help loop RBP binding sites to their target sequences, e.g. RBFOX2 binding site that is distal (500bp from exons), but brought close to the exon by hairpin loop. (3) M6A structure switch: M6A affects base pairing of a stem loop, and expose binding site of HNRNPC. Thousands of such M6A switches were found.
- Effect of RNA structure on RNA processing: start codon accessibility (single strand) promoters translation, uncapped mRNA can be stabilized by stem-loop in 3 UTR.
- Remark: M6A effects may not depend on specific readers. As long as M6A affects binding of some RBPs, it could have some biological functions.

High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis [Schwartz and Regev, Cell, 2013]

- Background: M6A appears to be static across cell types and stimuli in mammals, but only during meiosis in yeast.
- Refining experimental protocol of M6A-seq: (1) Shorter fragments. (2) A yeast strain with no RNA methyltransferase. Results: 3k peaks, but half are present in input or deletion strain. Total of 1.1K true M6A sites. Resolution: the distance of the center of peak to consensus is 3 nt (18 if including the sites removed by deletion strain).
- Majority of M6A sites contain the consensus, but only 1/40 of the consensus sequences are methylated.
- Association of features with M6A sites: (1) Additional sequence info: 73% are U in the +4 position and 63% are A in the -4. (2) Enriched near stop codon. (3) Secondary structure: less stable. To assess the RSS of a M6A site, choose 50bp window around the site, then use RNAfold to predict mRNA secondary structure, and compute the MFE (minimum free energy) of the window. Shuffle the window to obtain null distribution and the Z-score of each site.
- Classification: use logistic regression with 1-3 features. Positive set: 700 sites with the consensus; Negative set: 10K non-methylated sites surrounding the consensus. Use sequence feature alone: AUC 0.84; use position bias: AUC 0.686; position bias + RSS: AUC 0.696. Use all: AUC 0.867.
- Conservation of M6A: (1) Strong conservation at gene level: 200/600 methylated genes are conserved. (2) Substantial turnover of M6A sites: but still 54 sites are exactly conserved, and 60 within 100 bp (10 and 4 times enriched vs. chance expectation, respectively).
- Remark: RSS analysis: may need to determine the RSS of entire transcript. However, this can be complicated in the presence of alternative splicing. In yeast, little AS, so not a problem.
- Lesson: M6A is determined strongly by the sequences in the flanking region of consensus.

N6-methyladenosine Modulates Messenger RNA Translation Efficiency [Xiao Wang and Chuan He, Cell, 2015]

- Model: YTHDF1 + m6A > YTHDF1-m6A > increase TE. Experiments in HeLa cells.
- Direct binding of YTHDF1 to m6A: PAR-CLIP on YTHDF1 (cross-linking leads to mutations). Note: PAR-CLIP has high resolution. YTHDF1 targets heavily overlap with m6A sites (50%), and share the same motifs and YTHDF1 sites are very close to the motif locations.
- Effect of TE by YTHDF1: RP and RNA-seq under YTHDF1 knockdown. TE is reduced by DF1: in DF1-non-targets, 55% reduced TE, and in DF1-targets, 75% reduced TE (Figure 2A,C). Note: this does not prove YTHDF1 has direct effect on TE.
- Tethering experiment: show YTHDF1 has direct effects on TE. Create a reporter tethered to N-terminal domain of YTHDF1: higher TE when tethered (Figure 4D).
- Dependency of YTHDF1 effect on m6A: (1) knockdown of METTL3. TE is lower in YTHDF1 targets comparing with non-targets (Figure 3A). (2) Also YTHDF1 knockdown changes distribution of m6A: more likely to be in mRNA, rather than ribosome-bound mRNAs. Note: (1) alone does not prove the model, its possible that METTL3 k.d. changes some gene, which modifies DF1 activity. But combining (1) with direct binding and (2) leads to much stronger evidence.
- Effect of YTHDF1 and YTHDF2: two proteins have 50% common targets. Focusing on these targets: DF1 and DF2 have independent effects (DF1 on TE and DF2 on stability).
- Lesson: to establish the causal chain $A > B > C$, need to confirm multiple steps, $A > B$, $A > C$. Often the key is to show the dependency of $A > C$ effect on B. The key experiment here is: YTHDF1 effects on TE depends on METTL3.

SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features [Yuan Zhou and Qinghua Cui, NAR, 2016]

- Single-nucleotide resolution of M6A: data of 5 tissues. Training data: 8K positive and 65K negative sequences, chosen around the actual M6A site matching DRACH motif (for both positive and negative). To avoid positional bias: more likely to choose non-m6A near a M6A site.
- Features: (1) Positional nucleotide around the M6A site/motif. (2) KNN feature: sequence similarity in the adjacent window (among top K similar sequences, how many are positive) (3) To capture sequence information independent of relative position to M6A: use d-spaced nucleotide pairs (d from 0 to 3).
- Secondary structure feature: use RNAfold, encode RSS at each position as hairpin loop, interior loop, paired, etc.
- Classifier: random forest, full transcript mode (including pre-mRNA) or mRNA mode.
- Performance: positional features important. Found positional bias in distal positions, e.g. -29, -28, +29. Many distinct rules from random forest, suggesting that M6A sites form diverse clusters. KNN feature: little improvement over position. Nucleotide pair feature: substantial improvement (AUC from 0.8 to 0.89, AUPRC from 0.37 to 0.52). Most features do not show strong positional bias.
- Contribution of secondary structure: weak accuracy, cannot further improve. Experimental RSS data can be helpful, but not available.
- Full mode vs. mRNA mode: full mode significantly better, supporting the importance of intronic sites (m6A modification in pre-mRNA stage).
- Tissue-specificity of M6A: the model trained and tested on the same tissue better than cross-tissue performance. The difference sometimes can be large, e.g. AUPRC 0.42 vs. 0.28.

- Remark: limitations of training set: (1) only use DRACH matches - missing other m6A sites and some matches may be FP sites. (2) Not cell-type specific.

Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation [Solbodin and Agami, Cell, 2017]

- Experiment for measuring translation efficiency (TE) of different promoters: promoter-reporter-barcode constructs, then stable transfection in MCF7 cells. Isolate mRNA and sort by surcose gradient (polosomal profiling): different amounts of ribosomes. Then sequencing.
- TADA box promotes TE: positive correlation of TADA promoters and TE. Introducing TATA box to non-TATA promoters increase TE (surcose gradient and protein/mRNA).
- Transcription rate positively affects TE: (1) Figure 4A: Genome-wide positive correlation of transcription rates (GRO-seq) and TE (Ribo-seq and RNA-seq). (2) CPT (inhibitor of RNAP II) reduces mRNA, and even more on protein: in reporter, and 700 genes across the genome (Figure 4BC). CPT also reduces TE in polosomal profiling experiment.
- Effects of transcription rates on m6A: lower transcription likely increases m6A:
  - Non-induced reporter mRNA (lower transcription): much higher levels of M6A.
  - Reducing transcription rates by mutated TATA-box or elevated MgCl: higher deposition of m6A in reporter (Figure 5B).
  - CPT treatment enhances M6A in total mRNA in cells by 2 fold (Figure 5C). Similar results in RNAP mutant (Figure 5E)

  Also find evidence of MTC interaction with RNP II with CPT treatment.
- Inhibition of METTL14 directly increases TE in non-induced mRNA (in induced mRNA, less M6A to start with). Adding adenosine also reduces TE.
- Model/lesson: (1) TATA box provides natural variation of transcription initiation rates. (2) M6A levels vs transcription and translation: slower transcription favors M6A (easier to recruit MTC by RNP); M6A inhibits translation.
- Discussion: some papers reported positive effect of M6A on translation rate, however, these are M6A in UTRs. Here: m6A in CDR likely impedes translation.
- Discussion: RNP II pauses may increase the change of MTC engagement, consistent with the enrichment of M6A near start and stop codons.
- Remark: the evidence of transcription rates affect m6A is not conclusive: (1) only one reporter mRNA; (2) Genomewide results rely on perturbation of RNAP (CPT or RNAP mutant), which may have many other effects.

The RNA N6-methyladenosine modification landscape of human fetal tissues [Xiao, NCB, 2019]

- Data: m6A profiles MeRIP, in 8 fetal tissues (2nd trimester): brain, liver, lung, kidney, heart, stomach, placenta and skeletal muscle. Many novel m6A peaks: e.g. in kidney, 50K peaks, half of which not reported before.
- Distribution: large fractions in introns and intergenic. Ex. brain: 31% in introns and 14% in intergenic regions.
- m6A variation across tissues: 21 samples cluster by tissue types (Figure 1d). Brain and placental very different.

- Differential m6a: defined as two fold difference, limited to transcripts expressed in all 8 tissues. Found about 500-1000 tissue differential m6A peaks. About half of tissue-differential m6As were found in introns.
  - Remark: likely significant underestimate. Limit to constitutive transcripts, and 2-fold difference is relatively large change.
- m6A in lincRNAs: total of 4K lncRNAs, half are in e-lncRNAs.
- M6A levels positively correlate with expression stability/homeostasis across tissues.
- Enrichment in eQTL: 1.34 to 1.77 fold enrichment of m6A in eQTL from matching tissues.
- GWAS enrichment (Figure 5e): FET using GWAS catalog. Overall 2.2 fold enrichment, $p < 0.001$. Strongest enrichment include: heart in body measurements, cardiovascular measurement, lipid (4 fold); liver in body measurement, hematoloigical measurements, CVD, lipid (4-7 fold), kidney in CVD and lipid (4-6); lung in body measurement, immune and lipid (3-4).

REPIC: A database for exploring N6-methyladenosine methylome [Liu and Mengjie Chen, GB, 2020]

- Background: existing databases do not consider cell-type specificity.
- Data: m6A/MeRIP-seq from 672 samples of 49 studies, covering 61 cell lines or tissues in 11 organisms. 53% are from human. Also DNase-seq and histone ChIP-seq data.
- Computational pipeline: removing adaptor and QC (Cutadapt), remove rRNA (HISTAT2), genome mapping (HISTAT2). Duplicate reads: prevalent, not removed. After obtaining mapped reads, do peak calling (MACS2, exomePeak, MeTPeak), visualization (Bigwig).
- Comparison of peaks called by methods: only 13% of MACS2 peaks have Jaccard index > 0.5 with exomePeak, and even lower with MeTPeak. exomePeak and MeTPeak much higher overlap.
- Cell and tissue specificity: samples are generally clustered by cell types (Figure 4).
- Visualization: genome browser powered by GIVE (Sheng Zhong group). Display both epigenomic and m6A tracks.
- Remark: a main issue is which peak calling tool to use.

N6-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription [Jun Liu and Chuan He, Science, 2020]

- Background: in mESC, METLL3 KO is embroynic lethal, but not YTHDF2 (decay regulation) KO. In contrast, YTHDC1 KO show similar pheontype as METLL3.
- METLL3 KO mESC show increased transcription (nascent RNAs), and open chromatin (amount of ds DNA break) - Figure 1. The results are similar in YTHDC1 KO, but not in DF2 KO.
- METTL3 KO: largest decrease of m6A in chromsome-associated RNAs (caRNAs), by more than 50% using LCMS. Also confirm using MeRIP-seq. Focusing on chromatin associated regulatory RNA (carRNA), including promoter-associated RNAs, eRNAs and repeat RNAs: about 15-30% have m6A, and about 60% of these are regulated by METTL3. Also changes in m6A levels negatively correlate with expression of carRNAs. Model: METTL3 regulates carRNAs by m6A, which leads to lower expression or decay.
- YTHDC1 facilitates the decay of a subset of these m6A-modified RNAs: DC1 KO show no change of m6A in mRNAs, but in a subset of caRNAs: 20-30% of eRNAs, paRNAs and 60% of repeat RNAs. For repeat RNAs, effect strongest in LINE1, which is important for chromatin and transcription. Note: DC1 is a reader, yet it affects m6A/A levels by change stability of m6A-transcripts.

- M6A effects on transcription of genes: (1) METTL3 KO leads to change of gene expression (overall activation). For genes up-regulated: tend to have upstream carRNAs marked with m6A. (2) METTL KO: increase of H3K4me3 and H3K27ac; and increase of YY1, EP300 binding.
- Model: Figure 4K, a subset of carRNAs are methylated, and degradation is promoted (via YTHDC1). This leads to reduced activation of CREs and downstream gene transcription.

Chapter 5

Biological Systems

5.1 Yeast

1. Amino acid starvation response [Natarajan & Marton, MCB, 2001]

   Background: activation of Gnc4 by AA starvation: uncharged tRNA → Gcn2p, a translation initiation factor kinase → derepression of Gcn4 translation.

   Methods:

   (a) Conditions: WT cells, gcn4Δ cells, under 3AT treatment (histidine starvation).

   (b) Gcn4-target genes: genes induced by 3AT by at least 2-fold in WT cells, but not in gcn4δ cells.

   Results:

   (a) About 1,000 genes are induced and repressed by at least two-fold, respectively, by 3AT treatment. Among these, 539 are Gcn4 targets (induced), and repression of some genes are also dependent on Gcn4.

   (b) A large fraction of Gcn4 targets (about 50%) contain Gcn4 motif in the promoter sequences. And genes containing Gcn4 motifs are much more likely to be induced instead of reduced by Gcn4. Thus Gcn4 seems to induce targets by direct binding, while repress other targets through indirect interaction.

   (c) RP regulation by Gcn4: repression of 90 RPL and RPS genes depends on Gcn4. Probably through indirect regulation as the RP genes lack Gcn4 motif.

   (d) AA biosynthetic genes: 73 genes, for all pathways except Cystine, a large propotion or most are Gcn4 targets. Ex. in histidine pathway, all except His6; and in pyruvate family, all except Ilv5. No histidine specific response: histinde pathway induction no stronger than other pathways.

   (e) Other Gcn4 targets: 16 vitamin/cofactr biosynethic genes, 8 mitochondrial carrier proteins (shuttling between mitochondrial and cytosol), 4 AA precursor biosynthetic genes, 5 purine biosynthetic genes, and 6 peroxisomal genes.

   (f) 26 TFs are Gcn4 targets: including regulators of AA and purine biosynthesis (Arg80, Leu3, etc.), TCA cycle intermediates (Rtg3).

   (g) Role of Gcn4 on glycogenesis: both genes of glycogen breakdown and storage are induced by Gcn4, but the gcn4Δ cells accumulate glycogen, suggesting the net effect of Gcn4 is glycogen breakdown (perhaps by providing AA biosynthesis precursors).

2. Pertubation analysis of yeast galactose metabolism [Ideker & Hood, Science, 2001]

   Problem: the regulatory control of galactose utilization pathway.

Summary: genetic perturbation of the pathway, measure the gene expression and infer the mechanism.

Methods:

(a) Perturbation: mutant of each of the known players of galactose pathway, treat with galactose and measure expression

(b) Expression: measured with cDNA microarray for mRNA and with isotope-coded affnity tag (ICAT) and tandem mass spectrometry.

Results:

(a) Comparison of mRNA and protein expression: moderate correlation, $r = 0.61$. There are 15 genes whose mRNA level did not change while protein levels changed significantly.

(b) Hypothesis generation in galactose pathway: e.g. mutation of Gal7 and Gal10 (enzymes) reduced expression of other Gal enzymes, suggesting that the products of these enzymes modify the activities of regulatory proteins.

(c) Identification of related genes: clustering of expression profiles reveal genes with expression profile similar to galactose genes; and some of them are Gal4 targets.

Remark: pertubation analysis (both genetic and environmental) as a way to reveal the regulatory control.

3. Stress response of C. albicans [Enjalbert & Whiteway, MBC, 2003]

   Background:

   (a) C. albicans has a commensal relationship with warm-blooded organisms and thus would be expected to live in a relatively stable environment in terms of temperature and osmotic conditions. In contrast, oxidative stress could be a frequent challenge for C. albicans cells as they are targeted by macrophage cells.

   (b) The Candida lineage appears to have initiated more than 150 million years ago. This divergence is emphasized by the presence of 2000 genes in C. albicans that have no homologues in S. cerevisiae.

   Methods: transcriptional profiling of C. albicans upon temperature and osmotic and oxidative stresses.

   Results:

   (a) No C. albicans genes were significantly induced in a common response to the three stresses: there is considerably more overlap between the lists of stress-repressed genes compared with the stress-induced genes (reducing growth rates).

   (b) The presence of Msn2/4 regulation in C. albicans: Candidate homologues of the Msn2p and Msn4p transcription factors exist in C. albicans. Although STREs (STress Response Elements, recognized by Msn2/4) are present in the promoters of many C. albicans stress responsive genes, their number of copies per promoter is reduced compared with the equivalent S. cerevisiae genes.

   (c) Not able to detect stress cross-protection in C. albicans.

4. Application of rFBA in E. coli [Covert & Palsson, Nature, 2004]

   Problem: the transcriptional regulatory network of metabolism in E. coli.

   Methods:

   (a) rFBA: Logic statement of regulatory process over time; and FBA, with extra constraints from gene expression level.

   (b) Data: 1) phenotype data: different E. coli strains under different environments (different carbon and nitrogen sources); 2) expression data: expression data of widetype, mutants (of six relevant genes) under oxygen response (aerobic-anaerobic transition).

Results:

(a) Starting model (from literature): 79% accuracy in phenotypic predictions, but low in expression predictions.

(b) Refining the model: update regulatory rules from expression data (in particular, using mutant expression). No change on phenotypic prediction, but substantially better on expression predictions.

Q. how are regulatory rules represented in the method and applied in the model? Does the method need expression data to make phenotypic predictions, if so, for a new mutant (where expression data is not available), how would it predict phenotype?

5. Transriptional regulatory code of yeast [Harbison, Nature, 2004]

Aim: the TRN of yeast. Note that the ChIP experiments only reconstruct TRN under certain conditions, so to extract TRN with under all potential conditions need to use sequence information.

Methods:

- Data: GWLA for each of the 203 TF of yeast (covers almost all yeast TFs). 147 TF hits more than 10 probes. Also 84 TFs in at least 1 of 12 conditions (selected by the known role of TF in that condition).
- Motif discovery: run multiple motif-finding programs in bound regions, then do clustering and stringest statistical tests ($P \leq 0.001$ using hypergeometric test). Additional criteria including the requirement for conservation across three of four related yeast species.
- TRN reconstruction:
  - ChIP-chip binding with $P \leq 0.001$.
  - Choose matches that are conserved in multiple yeast species.
- Co-occuring motifs: each promoter is: bound by both factors, by only one of them, or by none. Test the enrichment using hpergeometric test.

Results:

- Motif finding: 65 TFs with unique high-confidence motifs. Comparison with known motifs: for 21 TFs, no prior information on motifs
- Validation of predicted TRN: using literature data. E.g. six well-studied cell cycle transcriptional regulators bind to the promoter for YHP1 , which has been implicated in the regulation of the G1 phase of the cell cycle.
- Promoter architecture: Fig. 3
  - Single regulator: binding site of a single factor. Typically involved in a common biological function.
  - Repetitive motifs: repeats of a particular binding site sequence. Possible functions: necessary for stable binding by the regulator, e.g. Dal80; also permit a graded transcriptional response, as has been observed for the HIS4 gene16.
  - Mmultiple regulators: many of the genes in this category encode products that are required for multiple metabolic pathways and are regulated in an environment-specific fashion.
  - Co-occuring motifs: imply that the two regulators physically interact or have related functions at multiple genes.
- Condition dependence of TF binding: the binding of TF and target sequence is highly dependent on the condition (Fig. 4)
  - Condition invariant: bind the same set of targets in different conditions, e.g. Leu3. May be regulated by allosteric regulation of TF (e.g. Leu3 by the metablic precursor of leucine).

- Condition enabled: bind only in certain conditions, e.g. Msn2. Nuclear translocalization may be one mechanism for achieving the extreme case of condition-specific regulation.
- Condition expanded: for some targets, bind only in certain conditions, e.g. Gcn4, regulated by increasing protein stability.
- Condition altered: bind different targets in different conditions, e.g. Ste12, achieved through interaction with differnt other TFs (whose DNA binding specificites may be different).

6. Transcriptional control in the metabolic network of yeast [Ihmels & Barkai, NBT, 2004]

   Hypothesis: the expression pattern of genes involved in metablic networks are related to the role of genes in the metabolic networks.

   Methods:

   (a) Data: metabolic pathways from KEGG; gene expression from 1000 datasets: environmental and genetic perturbations, and many natural processes.

   (b) Co-regulation: measured by Pearson correlation; also the modular structure of the genes is determined by a bi-clustering algorithm.

   Results:

   (a) Linear pathway enzymes are often co-regulated.

   (b) Isozymes: for divergent junctions, often only one branch is correlated with the incoming branch; or if both in and out branch have isozymes, distinct outgoing branches wither ofter coexpressed with alternative isozymes.

   (c) Other co-expressed genes: transporters and TFs are often in the same modules as enzymes.

   (d) Hierarchical modular structure of co-regulation: top processes are: AA biosynthsis, protein biosynthesis and stress response.

7. Topological changes of TRNs in yeast [Luscombe & Gerstein, Nature, 2004]

   Idea: map TF-target networks in different conditions, and see how the regulations are different, e.g. how often a TF is condition-specific.

   Methods:

   (a) Data: yeast expression profile of cell cycle, sporulation, diauxic shift, DNA damage and stress response. The network (static) is constructed by ChIP-chip data and literature, including both TF-target and TF-TF interactions.

   (b) Identify active subnetwork in any condition: (1) TF is present in high level; (2) differential expression of targets; (3) link between the two; (4) if an interacting TF is active, then the TF is active, do it recursively.

   Results:

   (a) Comparison of endogenous (cell cycle and sporulation) and exogenous (resposne to environmental changes) networks: in endogeneous networks, fewer targets per TF, more complex TF combinations, longer regulatory cascades, many feedforward loops (in constrast, coregulated genes, SIM and MIMs are common motifs in exogenous). Suggesting that sub-networks have evolved to produce rapid, large-scale responses in exogenous states, and carefully coordinatated in endogenous conditions.

   (b) Permanant hubs (always active) are multi-functional TFs or regulating house-keeping genes, most TFs (78%) are transient, i.e. most important in one condition.

(c) Rewiring (changing of regulatory targets) is common: even for permanent hubs. This allows TFs to be used in many conditions (thus single TFs cannot achieve specificity). In contrast, only 51 out of 360 TF pairs are used in multiple conditions.

8. Just-in-time transcription [Zaslaver & Alon, NG, 2004]

   Hypothesis: the design of transcriptional regulation of metabolic networks follows certain rules as a consequence of evolutionary optimization.

   Model: consider a linear metablic pathway with three enzymes. Assume:

   - The pathway is controlled by a single repressor protein R, which is functional only when it is bound by the ligand, P (the end product). The active form of R is thus: $R_T \frac{P}{P+Kr}$, where R_T is the total level of R, assumed to be constant.
   - The enzyme level is determined by: 1) transcription, which is determined by the probability that the active R is not occupied: for E_i, this is: $\beta_i \frac{1}{1+R/k_i}$; 2) degradation or reduced concentration due to cell growth.
   - The enzymatic reactions: follow M-M kinetics and the molecular concentration is also reduced by cell growth.

   Minimization of the evolutionary cost function: the protein production cost and the quickness to reach a target flux (product). The solution predicts $\beta_1 > \beta_2 > \beta_3$ and $k_1 < k_2 < k_3$, which also suggests the activation order and the maximum level.

   Methods: measure promoter activity of all enzymes in a pathway (GFP) under conditions: 1) all AAs are present; 2) one AA is removed; 3) no AAs. Sample every 4min or 8min. A library of 52 reporter strains representing 50% of known AA biosynthesis genes. Note that: the promoter activities measured in this way already reflect the post-translational modifications of the regulatory proteins.

   Results:

   (a) Overall dynamics: AA addition generally reduce expression of that AA. Cross-activation and -repression is observed: e.g. addition of glutamate reduces expression of arginine biosynthsis enzymes.

   (b) Depletion of arginine, methoioine, serine (unbranched pathways): activation order and maximum activation level (normalized) follows the order of genes in the pathway.

   (c) Theoretical prediction: consistent with the experimental observation.

9. DNA damage response [Workman & Ideker, Science, 2006]

   Background:

   (a) Signaling pathways. Tel1 and Mec1 aggregate at DNA lesions, and active signaling cascades that include Chk kinases, which in turn trigger transcription and transcription-independent responses, including activation of DNA repair and cell cycle arrest.

   (b) Differentially expressed genes: by MMS treatment, several hundred.

   (c) Genes required for recovery from MMS damage through deletion studies. Little overlap between the differentially expressed genes and the sensitive genes.

   Methods:

   (a) Selecting TFs involved in DNA damage response: union of four sets of TFs. Differentially expressed TFs (T), target genes (ChIP) that are differentially expressed (B), sensitive genes (deletion will lead to lower fitness in MMS condition, S), and literature (L). Total 30 TFs are chosen.

   (b) TF-binding in MMS treatment, comparison with normal conditions to define condition specific binding.

(c) Deletion buffering: If a gene is differentially expressed in widetype, but not under TF knockout, we call this TF-gene pair deletion buffering (a regulatory relationship).

(d) Explain deletion buffering by physical networks: find the parsimonious paths from TF to the affected gene.

Results:

(a) Differential binding upon MMS treatment: six factors bind significantly more genes (Cad1, Pdr1, Ino4, Rim101, Uga3, Dal81) and eight less (Yap5, Rtg3, Hsf1, Swi5, Crt1, Ash1, Msn4, Fzf1).

(b) Combinatorial interactions among TFs: two TFs whose target gene sets overlap. The pattern changes with MMS treatment, e.g. cell cycle genes normally work together no longer do so after MMS treatment. A prominant combination appearing after MMS treatment: Ino4, Dal81, Mcm1, Rim101, Ecm22, Rpn4 and Uga3.

(c) Deletion buffering: 341 pairs found. E.g. Rfx1, a transcriptional repressor, and RNR2, RNR3, RNR4, synthesis of new nucleotides during DNA repair, also consisent with binding patterns (Ctrl1 bind to the promoters of these genes in normal conditions, but not after MMS).

(d) Explaining deletion buffering: only 37 pairs can be explained be direct ChIP-chip interaction, by applying the physical network model, add 68 buffering events.

(e) The network of DNA damage response (Figure 5): regulatory cross-talk among the processes of DNA replication and repair, cell cycle, stress responses and metabolic pathways. Mainly among TFs in different processes, e.g. Ace2 (cell cycle) regulates Gcn4; Ino4 (phospholipid metabolism) regulates Swi5 (cell cycle). For some genes, regulated by TFs involved in different processes, e.g. Rnr2, Rnr4 regulated by Yap6 (carbohydrate metabolism), Crt1 (damage response), Swi6 (cell cycle).

10. TOR and PKA targets [Chen & Powers, Curr Genet, 2006]

Methods:

(a) TOR targets: genes whose expression is significantly affected by rapamycin treatment (under YPD or minimum medium, MDN).

(b) PKA targets: genes whose expression is changed by bcy1 deletion.

Results:

(a) LYS (and branched chain AA) are co-regulated by TOR and PKA: LYS expression is repressed by rapamycin in MDN (but not in YPD, where LYS expression is not needed). Also, this rapamycin effect disappears in bcy1 mutant. It was claimed that the TOR effect on LYS expression is independent of known regulators: Mks1, Lys14, etc.

(b) Classification of genes affected by both TOR and PKA:

- Biosynthesis (RP, LYS, LEU, etc.): co-regulated by both, and the activity of one pathway can compensate the other.
- Glucose fermentation: both pathways seem to be required.
- Respiration: genes of mitochondrial function. TOR is required for expression of these genes, but expression repressed by PKA.

Remark: the TOR, PKA effects on LYS, LEU, etc. may be attributable to the change of flux through α-KG or pyruvate. The experiment showing TOR effect on LYS is independent of Mks1: in Mks1 mutant, rapamycin represses LYS expression (similar to w.t.). However, if rapamyin affects glycolysis and α-KG flux, then even in Mks1 mutant, this sam effect can affect LYS expression.

11. Strategies of regulation of ribosomal genes [Levy & Barkai, PLoS ONE, 2007]

    Motivation: how is ribosomal biogenesis regulated? Internal feedbacks (the internal nutrient states), or direct sensing of environmental conditions? Normally, the two strategies are coupled, as favorable environmental conditions generally lead to favorable internal states.

    Background:

    (a) Internal feed-backs in bacterial: rate of ribosomal biogenesis increase with the square of growth rate. Purine NTP levels directly regulate rRNA transcription. Also the stringent response: induced by uncharged tRNA, repress the transcription of genes associated with translation.

    (b) In yeast: the rate of ribosomal biogenesis is growth rate dependent, but cannot decouple the internal vs external mechanisms.

    Methods:

    (a) Decoupling internal and external states: Adh1 mutant cannot fermentate glucose, thus grows better in glycerol (nonfermentable) than glucose. If internal feed-back dominates, ribosomal biogenesis gene expression will be higher in glycerol than in glucose; if the environmental cues dominate, then expression will be higher in glucose.

    (b) Temporal kinetics of expression and growth rate: the instantaneous growth rate can be measured.

    Results:

    (a) Ribosomal biogenesis gene expression is highly responsive to changing environments: the carbon source (fermentable or not), temperature, growth rate (nutrient availability, which changes in culture).

    (b) Using Adh1 mutant: the expression of ribosomal biogenesis genes and stress genes, are tuned to environment.

    (c) Temporal kinetics: the initial gene expression responses preceded the change of growth rates.

    (d) Growth-related genes: find genes whose expression correlate with growth rate. Most groups show correlation only in specific conditions. It may be that correlation with growth is observed only for genes involved in rate limiting processes.

12. Functional specifity of ribosomal proteins (RPs) [Komili & Silver, Cell, 2007]

    Hypothesis: RP paralogs may have different functions.

    Methods:

    (a) Bud-site selection: to determine the function of RPs in bud-site selection, use Ash1 mRNA localization as marker. Ash1 mRNA is localized to the bud sites.

    Results:

    (a) Bud-site positioning: of 15 genes implicated in bud-site selection (the mutants have phenotype in bud-site selection), 14 have duplicates, but only one paralogy is required.

    (b) Transcriptional profiling: of the mutant strains and compare with w.t. yeast. The genes affected by deletion of paralogs are often different.

    (c) Phenotypic profiles: for one RP and its paralog, compare their phenotype profiles, including: cell size, bud-site selection, telemere length, sensitivity to stress and drug sensitivity. They are often different.

    (d) Cellular localization and assembly of paralogs: often different.

Discussion: "ribosome code" as an additional level of control of gene expression. Different combinations of RPs, posttranslational modification of RPs and different forms of rRNAs allow calibrated translation of specific mRNAs.

13. Comparative genomics of yeast stress responses [Gasch, Yeast, 2007]

    Ecology of yeasts:

    (a) Free-living yeasts: commonly found in tree exudates, plant roots and surrounding soil, on ripe and rotting fruits, and in association with insect vectors that transport them between substrates.

    (b) Other species exist in close association with animal hosts, in either a commensal or a pathogenic framework. Pathogenic species must contend with the defence mechanisms of their host, in particular the reactive oxygen and nitrogen species generated by the immune system.

    Environmental stress response (ESR):

    (a) Scer and Spom: similar ESR genes, and similar dynamics (transient response and then a new steady-state level of expression). A small number of genes involved in fatty-acid metabolism were induced in Scer, but repressed in Spom.

    (b) Both Scer and Spom ESR include feed-backs: the response regulators (Msn2/4 in Scer and Atf1 in Spom) are induced, as well as the regulators of the TOR/PKA pathway (inhibition of the stress response).

    (c) Calb: a much weaker common-stress response (relatively small number of genes are induced common to all stresses). But still functional links among the ESR orthologs in Calb: weak but consistent similarities of expression, enriched with STRE (stress response elements, recognized by Msn2/4), and inhibitory relation between PKA and ESR.

    Function of ESR:

    (a) Cross-stress protection: entirely dependent on Msn2/4 in Scer and Atf1 in Spom, suggesting the importance of ESR.

    (b) Trade-off between stress protection and restoring normal growth when stresses are moved or cells acclimated to the conditions.

    Regulation of ESR:

    (a) Scer: diverse stresses $\rightarrow$ Msn2/4 (nuclear transloation) $\rightarrow$ ESR. Different signaling pathways may converge at Msn2/4 (e.g. Mec1 for DNA damage). Rpd3 (HDAC) may also play a role in ESR. Osomotic stress $\rightarrow$ Sln1, Sho1/Msb1 (two pathways leading to) $\rightarrow$ Hog1 (MAPK) phosphorylation and nuclear translocation $\rightarrow$ Sko1 (TF) $\rightarrow$ osmotic defense genes. Hog1 is phosphorylated by many other stresses, but failed to translocate to nucleus, thus its effect is limited. Oxdiative stress $\rightarrow$ Yap1 $\rightarrow$ oxidative defense genes (in ESR or additional genes).

    (b) Spom: diverse stresses $\rightarrow$ Sty1 (Hog1 ortholog) nuclear translocation $\rightarrow$ Atf1 (Sko1 ortholog) $\rightarrow$ ESR. The upstream signaling: from three routes, (1) oxidative stress $\rightarrow$ Wis4 (MAPKKK) $\rightarrow$ Wis1 (MAPKK) $\rightarrow$ Sty1; (2) diverse stresses $\rightarrow$ Win1 (MAPKKK) $\rightarrow$ Wis1 (MAPKK); (3) heat shock $\rightarrow$ phosphatases, inhibiting Sty1. Oxidative stress $\rightarrow$ Pap1 $\rightarrow$ oxidative defense genes.

    (c) Difference between Scer and Spom regulation: Hog1 is a specialized regulator in Scer, and Spom does not have Msn2/4 ortholog. The specialization of Hog1 probably occurs in the lineage leading to Scer.

    (d) Calb: Hog1 plays a role in general stress-resistence, similar to Styl1. The downstream regulators are not clear, one candidate is Mnl1 (Msn2 ortholog).

    Common themes in ESR regulation:

(a) Common response to diverse stresses: implemented by condition-specific signaling pathways and general factors (Msn2/4, Rpd3 in Scer and Sty1-Atf1 in Spom).

(b) Feed-back mechanisms for ESR suppression: TOR/PKA pathway for negative feedback, and the positive regulators for positive feedback.

(c) Condition-specific TFs in addition to common TFs: e.g. oxidative stress, Yap1 in Scer and Pap1 in Spom.

14. Arginine biosynthesis in E. coli [Caldara & Cunin, JBC, 2008]

Background: Arginine biosynthesis and pyrimidine de novo biosynthesis pathway (UTP and CTP) (Figure 1). Two pathways share the metabolite CP, synthesized by CPSase (a key enzyme subject to regulation by Ornithine and UMP) from Glutamine.

Problem: the regulation of both pathways s.t. the stable output of one pathway when another pathway is perturbed.

Model: only the steady-state behavior of the system, and only metabolite measurements are available. Simplifications/key ideas:

- The linear steps can be compressed into a single step: only one enzyme is subject to regulation/rate limiting.
- Transcriptional regulation: not explicitly modeled because only steady states (of metabolites) are measured. The enzyme level will be treated as parameters in the model (the $V_{\max}$ terms in Michaelis-Mentor kinetics).
- Allosteric regulation of key enzymes: may take multiple inputs. For CPSase, the effects of regulators (Ornithine and UMP) are modeled by affecting binding of substrates (ATP).

Model fitting: literature curated kinetic parameters, the rest by fitting the steady states under minimum condition, and Arginine-rich medium.

Results: wild type, Arginine-rich medium, ArgR mutant, feedback-resistant (disruption of metabolic regulation: end-product inhibition) and double mutants of both genetic and metabolic control. Two key findings (validations):

- Under Arginine-rich medium (vs. minimum medium): accumulation of CP, but CTP/UTP about constant. In general, the CTP/UTP level is relatively insensitive to free Arginine pool (under other conditions). This is mainly ensured by the pyrimidine salvage pathway (UMP is regenerated/interconverted from nucleotide pathways).
- Disruption of either genetic or metabolic control: modest increase of arginine (5 to 6 fold); of both leads to 80 times change.

Remark:

(a) How pyrimidine salvage pathway helps maintain constant CTP/UTP under CP accumulation?

(b) The roles of specific regulatory mechanisms are not well characterized: e.g. how important Ornithine activation of CPSase is? Does it help to create CP when a sudden increase of flux in arginine pathway depletes CP?

15. Genetic variation of stress sensitivity and gene expression in S. cerevisiae [Kvitek & Gasch, PG, 2008]

Problem: different yeast strains must adapt to respective niches, thus the phenotypes, in particular, stress sensitivities may be different. And if so, what is the basis?

Background:

(a) Genetic variation in natural populations of S. cerevisiae: little geographic structure, low sequence diversity perhaps resulting from the low rates of out-crossing between strains (thus lower effective population size).

(b) Yeast phenotypes that have been characterized: wine making, thermotolerance, spoulation efficiency, drug sensitivity, morphological traits, etc.

Methods:

(a) Data: 52 strains from: European vineyards, commercial wine- and sake- producing strains, clinical, fruit substrates, oak-tree soil, lab strains and three other close species. Stresses analyzed: minimum medium, heat shock, presence of toxic drugs, oxidizing agents, osmotic stress, etc. Expression is profiled under normal medium (the basal level of expression should affect the chance of survial in a sudden stress).

(b) Testing selection of stress sensitivity (as a quantitative trait): compare genotypic variation and phenotypic variations. The ratio P/G vs. 1 indicates positive, negative selectin or neutrality.

Results:

(a) Stress sensitivity variation: (1) most vineyards and lab strain grow modestly well in most of the environments; (2) great variations in some conditions: copper sulfate, sodium chloride, freeze-thaw (probably niche-specific), and least variations in minimum medium lacking AAs, high temperature, and non-fermentable acetate.

(b) Gene expression variation: (1) many of variations are due to S288c (lab strain)-specific expression; (2) CNV also affects some variations, 2-5%; (3) genes with variable expression in at least one non-lab strain (about 1/4 of all genes) are enriched in TATA genes, genes with paralogs, but under-enriched with essential genes; (4) genes with the largest variations (in at least 3 of the 17 non-lab strains) are enriched with: AA metabolism, sulfur metabolism and transposition.

(c) Association between gene expression and stress sensitivity: (1) copper sulfate resistence: CUP1 (metallotheeionein); (2) NaCl resistence: Na^+ homeostasis and/or osmolarity maintenance. Sensitivity to cell wall-damaging drug: genes involved in mitochondrial function and translation, ATP synthesis.

Discussion:

(a) Possible adaptive selection of stress sensitivity: e.g. copper sulfate resistence in vineyard strains, copper has been used as an antimicrobial agents in vineyards for a lont time.

(b) Sources of gene expression variation: some of them are due to the noisy nature of expression process, e.g. TATA-element containing genes may tend to have a larger change of expression with environmental or genetic perturbations.

16. Hog1 network in yeasts [Capaldi & O'Shea, NG, 2008]

Problem: yesat responses to different stresses. The role of general factors, Msn2/4, and other specific factors?

Methods: mutant cycle analysis.

(a) To study the interaction of two genes H (Hog1) and M (Msn2/4), create the mutants: ΔH, ΔM and ΔHM, and compare gene expression under these constructs (and widetype). The idea is: if H and M have cooperative interaction, then the change of gene expression in ΔHM cannot be explained by the change of expression induced by ΔH and ΔM combined.

(b) Formalism: infer the H, M, and Co components on gene expression from expression data under all genetic backgrounds.

Results:

(a) Mutant cycle analysis of yeast response to KCl osmotic stress: both independent activation by Hog1 and Msn2/4 alone, and a Co component.

(b) Analysis of Msn2 nuclear import following KCl stress: Msn2 import is activated by Hog1, but there are also other pathways that may lead to Msn2 import.

(c) Downstream targets of Hog1: mutant cycle analysis of Sko1/Hot1, the Co component is highly correlated to H component, thus Hog1-dependent gene induction occurs almost entirely through Sko1 and Hot1.

(d) Stress response to high glucose: activation of a subset of genes targeted by Hog1, the general stress response by Msn2/4 is not activated.

Discussion: the logic of the network architecture (Figure 5) is:

- KCl stress: cells need to immediately respond by post-translational modification, and initiate transcriptional response to prepare for future high osmality and damage (disruption of PPI and protein-DNA interaction). Thus need to activate general stress response program mediated by Msn2/4.
- Glucose stress: the damage is more limited, thus only need to activate Hog1 controlled genes, but not Msn2/4 stress program.

17. Activity motifs [Chechik & Koller, NBT, 2008]

Problem: the temporal pattern of enzyme activation in conditions of metabolic shifts?

Methods:

(a) Temporal activity motif (TAM): describe the order of activation of enzymes in a branch or linked branches, e.g. forward activation (the downstream enzymes are activated later).

(b) Data: microarray data under metabolic changes, 63 times coureses plus 13 new ones, data collected at 15, 30, 60, 120 or 240 min.

Results:

(a) Enriched TAMs: (1) forward activation to produce metabolic compounds efficiently; (2) backward shutoff to rapidly stop production of a detrimental product and (3) synchronized activation for co-production of metabolites required for the same reaction.

(b) Distribution of enriched TAMs: mostly in central carbon metabolism, not found in AA biosynthesis.

18. Oxygen response of E. coli [Yang & Tang, Thesis, 2009]

Problem: how does cell regulate gene expression in response to aerobic/anaerobic transition?

Background: during aerobic/anaerobic transition (O_2 depletion), E. coli reestablish the metabolic flux via two main regulators, FNR and ArcA:

- ArcA: activated by initial transition. Activated by member kinase ArcB, which is activated by electron carrier.
- FNR: activity directly regulated by O_2. Only activated by persistent condition.

Transcriptional response of genes:

(a) Genes up- and down-regulated by the transition: e.g. TCA cycle, oxidative phosphorylation.

(b) Comparison with FBA predictions of metabolic flux changes: largely consistent, but some flux changes are achieved through allosteric regulation, e.g. pentose phosphate pathway.

(c) Not all differentially expressed genes are regulated by FNR and ArcA. Only the branch to TCA cycle, branches PDH and PTA-ACK.

Control by FNR and ArcA: need to meet the design need of responding to different O_2 levels: activating different target genes. This is achieved by regulation of FNR and AcrA by O_2, forming a switch, and by combinatial regulation of FNR and AcrA of target genes (e.g. some are positively regulated by one, but negatively by the other).

Chapter 6

Mathematical Modeling of Biological Systems

6.1 Biochemical Background

This section is based on [Alon, Introduction to Systems Biology]

1. Time and spatial scales [Table 2.1]

   Cell volume:

   - E. coli.: 1 μm^3 (10^{-15} L)
   - Yeast: 1000 μm^3
   - Mammalian cell: 10,000 μm^3

   Protein concentration: a useful relation is: 1 molecule / cell $\approx$ 1 nM. The number of protein molecules within a E.coli. cell (18304323): in cytosol, the average number is 526.

   Time scale:

   - Diffusion time of proteins across cell: 1 msec, 10 msec, and 100 msec for E.coli, yeast and mammalian cells.
   - Equlibrium binding of small molecules to proteins: very fast, from 1 msec in E.coli to 1 sec in euk. cells.
   - Equlibrium binding of TF to DNA: 1 sec.
   - Transcription: 1 min in E. coli and yeast, 30 mins in mammalian cells.
   - Translcation: similar to transcription.
   - mRNA life time: 2-5 mins in E. coli; 10 mins to hours in euk. cells.

   Principle: typically, the reactions that only involve non-covalent interactions are much faster than the ones that involve breaking and formation of chemical bonds. Thus for a system of both types of reactions, it can be assumed that the faster reactions are at equilibrium.

   - Transcription: TF-DNA binding is fast and assumed to be in equilibrium and the rate-limiting step is the process of mRNA sysnthesis from DNA.
   - Enzymatic reactions (Michaelis-Menten kinetcis): the first step of enzyme association with substrate is fast, and the slow step is the actual convertion of substrate to production, which often involve chemcal bonds.

- Metabolic shift: a metabolic system can be assumed at steady state: the levels of intermediate do not change (this is different from equlibrium of every reactions). In the case of environmental changes, new steady states of the metabolic network can be established quickly, often within minutes.

2. Binding of ligand and macromolecules [Appendix]

   System: ligands (usually small) and macromolecules. The ligand regulates the activity of the macromolecules, when bound. The goal is to quantify the extent of activation/repression of macromolecules, which is usually degree of occupancy by ligands.

   Single ligand binding: suppose P is the macromoclue, and L is the ligand. If the reaction is fast, we assume the existence of the equilibrium:

   $$P + S \rightleftharpoons PS \tag{6.1}$$

   We are interested in the occupiency of P, i.e., the fraction of P that is bound with S. This is given by:

   $$\frac{[\mathrm{PS}]}{P_T} = \frac{[\mathrm{S}]}{K + [\mathrm{S}]} \tag{6.2}$$

   Multiple ligand binding: if multiple L molecules may bind to P, then there may exist cooperativitity, and the result is modified by the Hill equation:

   $$\frac{[\mathrm{PS}]}{P_T} = \frac{[\mathrm{S}]^n}{K^n + [\mathrm{S}]^n} \tag{6.3}$$

   where n is the Hill coefficient. The mechanism of cooperative binding may be explained by MWC model: P may exist in two states - active or inactive. Binding of L modifies the probability of switching between the two (more L molecules bound will increase the probability of switching).

   Application: this simple system can be invoked to explain. The key assumption is: this reaction is fast relative to other reactions in the system, thus safe to make the equlibrium assumption.

   - Enzyme substrate binding: the enzyme occupancy is related to the substrate concentration.
   - TF induction by signaling molecules: the fraction of active TF is related to the amount of signaling molecules.
   - DNA binding by TFs: the fraction of DNA (binding sites) is realted to the number of TF molecules.

3. Enzyme kinetics: Michaelis-Menten equation [Appendix]

   System: substrate is converted to product, catalyzed by the enzyme. Understand how the synthesis of product depends on the amount of enzyme and the concentration of substrates.

   Analysis: the reaction:

   $$E + S \rightleftharpoons ES \rightarrow E + P \tag{6.4}$$

   Suppose the rates of the first part are k_{on} and k_{off} respectively and the rate of the second reaction is ν. The first reaction is assumed to be much faster than the second, thus [ES] is in steady state. Solving this for [ES], we obtain:

   $$\frac{d[\mathrm{P}]}{dt} = \nu[\mathrm{ES}] = \nu E_T \frac{[\mathrm{S}]}{K_m + [\mathrm{S}]} \tag{6.5}$$

   where $K_m = (\nu + k_{\mathrm{off}})/k_{\mathrm{on}}$.

6.2 Transcriptional Regulation

This section is based on [Alon, Introduction to Systems Biology]

1. Concepts of transcriptional regulatory network (TRN) [Section 2.1-2.3]

   Cells as information processing machinery: a cell needs to respond properly (change expression of genes) to environmental and internal signals such as:

   - Physical parameters: temperature, osmotic pressure, etc.
   - Signaling molecules from other cells.
   - Nutrients.
   - Harmful chemicals such as those causing DNA damage.

   This problem is solved by: using TFs as sensors of signals (i.e. the activities of TFs will change in response to these signals), then TFs act upon the target genes to change their expression.

   Separation of time scales: the allosteric regulation of TFs and binding of TFs to DNA are much faster than transcription and translation.

   Remark: separation of time scales is a general concept, that is important in many systems.

2. Modeling activator and repressor function: kinetic approach [Appendix]

3. Combining signaling and transcriptional regulation [Appendix, Section 2.3]

6.3 Gene Networks & Design Principles

1. Overview of gene network modeling

   Modeling gene networks by logical (Boolean) networks:

   - Reference: [Dynamics of DNA damage induced pathways to cancer, reviewed for PLCB, 2013]
   - Defining the behavior of complex biological networks: e.g. in the context of cancer, what matters is the cell fate: proliferation or apoptosis. Thus the activities of anti- or pro-apoptotic proteins are crucial variables.
   - Constructing the networks: important questions/decisions may include:
     - Direct vs. indirect interactions: suppose the true model (direct interacion) is $A \to B \to C$, if it does not distinguish direct and indirect interactions, the model would be the direct model plus the edge, $A \to C$. This would lead to a different prediction, e.g. A would still affect C if B is knocked out.
     - Assigning function of genes (how gene activities are related to cell behavior).
   - Logical analysis of complex networks: dependency matrix (pairwise relationship) and steady state analysis. A key question is how would logical functions are defined when a node is influenced by multiple ones?
   - Evaluation of model predictions: using the implications of the model - indirect effects, the changes of gene activities upon perturbation of some other genes.

2. Gene regulatory networks in cellular level

   The lac operon in E.coli. [Kuhlman & Hwa, PNAS, 2007]:

   - Problem: lac operon is activated by CRP (cAMP Receptor Protein), which is activated by cAMP, and repressed by lacR, whose effect is reduced by lactose. Understand how lac expression (lacZ) is affected by cAMP and lacR (or agents that affect lacR).

- Experimental observations: the experiments measure the expression of lacZ as a function of [cAMP] and [IPTG] (similar to lactose, reduce the effect of lacR). In particular, to make the system simple, use mutants that remove the effect of:
  - lacY: lacY causes positive feedback of lac operon expression, to simplify, always use ΔlacY.
  - Endogenous cAMP: use the mutant ΔcyaA and cAMP in the medium.
  - cAMP degradation: need mutation of cpdA.

  Therefore, the triple mutant TK310 (ΔcyaA ΔcpdA ΔlacY), is the primary mutant for analysis of combinatorial control of IPTG and cAMP. The main observations that need to be explained are:
  - Hypersensitivity to IPTG.
  - The effect of CRP (or cAMP): increase the expression, but also increase the sensitivity of lac promoter to IPTG.

  The main feature of the curves is the cooperativity (how sensitive lacZ activation is to the level of cAMP or IPTG), which can be measured by Hill coefficient.
- The network of interactions:
  - CRP activates lac expression and lacR represses expression by binding to operator site (O_1).
  - cAMP activates CRP by forming cAMP-CRP complex (only the complex is active for lac operon).
  - cAMP level is further influenced by transport between intra- and extra-cellular environment (the effect of AC and cAMP degradtion is ignored).
  - IPTG binds with lacR: inactivates DNA binding of lacR.
- Model:
  - CRP activation: [CRP*] (active level of CRP) is determined from the equlibrium of the reaction:
    $$cAMP + CRP \rightleftharpoons CRP* \tag{6.6}$$
    Note that only intracellular [cAMP] is relevant, so need to further consider the transport process (assuming it is simply proportional to [cAMP] in the medium).
  - lacR activity: two lacR dimers form a tetramer and binds to operator. If both of the two units are bound by IPTG, the tetramer cannot bind to its operator site. The levels of active forms, [lacR*] (one unit is active) and [lacR**] (both units are active), are determined from the equlibrium among all possible molecular species of LacR-IPTG complex.
  - lacZ activation by CRP: follow the thermodyanmic model in [Buchler03], the effect of a single activator is:
    $$G \approx G_0 \frac{1 + \omega[A]/K_A}{1 + [A]/K_A} \tag{6.7}$$
    where ω is the interaction of A and RNAP.
  - lacZ repression by lacR: the effect of a single repressor is:
    $$G \approx G_0 \frac{1}{1 + [R]/K_R} \tag{6.8}$$
    lacR tetramer could also bind with the other two operator sites (O_2 and O_3). When this occurs, the tetramer binds with O_1 and O_2 (or O_3) simultaneously through DNA looping, thus is more favorable than binding with O_1 alone (cooperativity). This effect can be included by adding two looping terms to the denominator, which depends on the local concentrations of repressor at O_2 and O_3, given that the repressor is already bound to O_1.

- lacZ expression as determined by CRP and lacR: if they are independent, the effect can be multiplied. For example the joint effect of a single activator and single repressor is:

$$G \approx G_0 \frac{1 + \omega[A]/K_A}{1 + [A]/K_A} \frac{1}{1 + [R]/K_R} \tag{6.9}$$

  CRP can bind cooperatively with lacR, i.e. bound CRP bends the DNA and effectively increases the local concentrations of repressor to the sites O_2 and O_3.

- Analysis: the parameters of the model are either taken from earlier studies, or fitted with the observed ligand-response curve. The main insights obtained from applying the model are:
  - The effect of CRP: CRP is an activator in that it increases gene expression at each IPTG level. However, it increases gene expression in a nonlinear way such that the maximal fold-change and the sensitivity of the IPTG response are both enhanced.
  - Hypersensitivity of IPTG response resulted from multiple factors: (i) allosteric coupling between inducer and operator binding (Hill coefficient about 1.5); (ii) DNA looping increases cooperativity (Hill coefficient about 2); (iii) CRP-mediate DNA bending further enhancers the looping-induced cooperativity (Hill coeffecient 2.5 to 3).

  The distribution of cooperativity to multiple factors may reflect a general principle to achieve all-or-none type response common in biological systems.

Dynamics of yeast leucine biosynethsis pathway [Chin & Li, PLoS Biol, 2008]:

- Problem: leucine biosynethsis pathway, activated when leucine is depleted in the environment. Two features/functional requirements of this system:
  - When leucine is not needed (leucine level is already high): the system should stop the pathway.
  - When leucine is needed (leucine level becomes low): the system should quickly react and start the pathway.

  How would such requirements are met?

- Background: (Figure 1) features of the system (including metabolic reactions and regulation of the enzymes):
  - Linear pathway from pyruvate to leucine, where the upstream steps are shared by valine and isoleucine pathway.
  - The end product leucine inhibits the enzymes Leu4 and Leu9, the first enzymes that are unique to leucine pathway.
  - The intermediate product, αIPM activates the TF, Leu3, which serves as the transcriptional activator of all enzymes in the pathway (in combination with the TF Gcn4, which is activated by starvation of general amino acid). This makes recovery faster: increasing αIPM will further activates the downstream enzymes (in addition to the fact that the increase of substrate, αIPM, will lead to more prodcut)

- Experimental method: the protein abundance following the Leucine depletion is monitored by an automated system:
  - Genes are tagged GFP.
  - Individual cells can be separated in a population, and the expression of proteins measured, by flow cytometry.

  In this system, it is found that because of cell division, mother cells and daughter cells have different protein response profiles, so separate them and analyze only mother cells.

- Model intuition: intuitive description of the change of system upon Leucine depletion:

  - Release of Leu4 from leucine inhibition, in addition to transient induction of upstream genes (could be induced by Gcn4 because of general amino acid starvation): leads to a quick response in αIPM synthesis.
  - The accumlation of αIPM leads to activation of Leu3, which activates enzymes, particularly Leu1 and Leu2.
  - More synthesis of leucine restores the balance.
- Model formulation: two assumptions are made:
  - The upstream enzymes and products have constant levels. This is justified from experimental observation that their changes are much smaller than changes of downstream enzymes: Leu1 and Leu2.
  - The system has multiple types of reactions: enzymatic reactions of metabolites, transcriptional regulation, and ligand-protein interactions or post-translational modification of TFs/enzymes. It is assumed that the last type of reaction is considerably faster than the first two, thus could assume that they always stay in equilibrium.

  Based on these assumptions, the changes of enzymes Leu1, Leu2 and three metabolites, αIPM, βIPM and Leucine can be modeled by a system of ODEs:
  - Leu1, Leu2: basal transcription; transcriptional regulation (activation) by Leu3, which itself is activated by αIPM; and degradation.
  - Leu4 and Leu9: the active form of the two enzymes depends on the level of Leucine (post-translational modification). Not explicited modeled, instead, their effects are incorporated in the level of product, αIPM, they generate.
  - αIPM, βIPM and Leucine: production from the upstream step, and usage from the downstream step (exxcep Leucine), and degradation.
- Results: differential response of upstream and downstream enzymes (as defined by αIPM): the upstream genes have a small fold change (2 to 4 fold), and is faster; the downstream genes (Leu1 and Leu2) have much larger changes (more than 20-fold). Thus Leu3 only strongly affects Leu1 and Leu2 but not the other enzymes.
- Analysis of the model:
  - Basal expression level of Leu1 and Leu2: increase them will lead to quicker recovery of leucine, however, this strategy is not optimal since it will create waste when leucine is abundant.
  - Rate of transcription from Leu1 and Leu2: increase them will lead to quicker response, without affecting the steady-state level of leucine. Thus it suggests that it is possible to separately tune steady-state and dynamic behavior.
  - Alternative design: Leu4/9 also strongly controlled by Leu3. The drawback of this design is: Leu4/9 level will strongly depend on αIPM, which is required for Leu3 activation; however, αIPM itself is the product of Leu4/9. Thus the response would get slower.
- Remark: the advantages of GFP tagging and flow cytometry, comparing with mRNA microarray: (i) measure protein expression level; (ii) able to measure individual cells, important when the population is heterogenous.
- Remark: design principles emerged from this study:
  - Negative feedback can be used to stop unnecessary actions. This is used by leucine to inhibit Leu4/9, when leucine level is already high.
  - Positive feedback can be used to increase the response time. This is used by αIPM to activate Leu3, which further activates the pathway.
  - Cross regulation: when some components are involved with multiple processes, their regulation will be more constrained. This happens with the genes shared in multiple pathways.

- Modularity: the system is designed so that each part of functionality can be separately tuned without affecting other parts. This also applies to static and kinetic behavior.

- Remark: Modeling multiple types of reactions in a system:
  - Enzymatic reactions: Michaelis-Mentor kinetics provides the standard formulation:

$$\frac{d[P]}{dt} = k_2[E]_0 \frac{[S]}{K_M + [S]} \tag{6.10}$$

    where k_2, K_M are constants, $[E]_0$ is the total enzyme concentration and $[S]$ is the concentration of substrate, which itself is a variable. Could assume that the reduction of $[S]$ over time due to synthesis of P also follows the above equation, with the opposite sign.
  - Transcriptional regulation: for simple kinetic models, use Hill equations for the rate of transcription (where the ligand is transcriptional activator). Of course protein degraduation need to be considered.
  - Ligand-macromolecule interaction/post-translational modification: assumed to be faster than the other two types, thus only consider equilibrium. Let $[L]$ be the ligand concentration, the fraction of bound macromolecules can be specified with Hill equation (n is Hill coefficient):

$$\theta = \frac{[L]^n}{K^n + [L]^n} \tag{6.11}$$

3. Gene regulatory networks in development

   Canalization of gene expression in Drosophila blastoderm [Manu & Reinitz, PLoS Biol, 2009]:

   - Problem: what is the mechanism of canalization: the reduction of phenotypic variation during development from genetic and environmental variations, i.e. the ability of an individual to produce a robust phenotype?
   - Background:
     - Extensive variation in expression levels, domain borders and time and order of the appearance of individual domains. Ex. the variation of the Bcd gradient can be measured by the standard deviation of the boundary (defined by some threshold), 4.6% egg length (EL).
     - Variations in the expression pattern is significantly reduced over time. Ex. Variation of gap gene borders is about 1% EL.
   - Model: the concentration of factor a of nucleus i is denoted as v_i^a. Three phases:
     - Interphase: protein systhesis, protein diffusion between neighboring nuclei and protein decay.

$$\frac{dv_i^a}{dt} = R_a g\left(\sum_{b=1}^{N} T^{ab} v_i^b + m^a v_i^{Bcd} + h^a\right) + D^a(n)\left[(v_{i-1}^a - v_i^a) + (v_{i+1}^a - v_i^a)\right] - \lambda_a v_i^a \tag{6.12}$$

       In this equation, $N = 6$ is the number of zygotic genes; T^{ab} represents a regulatory interaction matrix: the regulatory effect of gene b on gene a; h^a is a threshold parameter; $g(\cdot)$ is the sigmoid function; R_a is the maximum synthesis rate; $D^a(n)$ is the diffusion parameter, which depends on the number of nuclear divisions before the current time t; and λ_a is the protein decay rate.
     - Mitosis: no protein synethesis.

$$\frac{dv_i^a}{dt} = D^a(n)\left[(v_{i-1}^a - v_i^a) + (v_{i+1}^a - v_i^a)\right] - \lambda_a v_i^a \tag{6.13}$$

     - Division: cell division, thus protein concentrations in the daugher cells are identical to those in mother nucleus, and the number of nuclei doubles.

- Results:
  - Variations of Bcd and egg size are reduced by gap gene cross regulation.

4. Robustness of gene networks

   Robustness principles:

   - Challenge: in many biological systems, perturbing the networks, including gene deletions/mutations, environmental perturbations, stochastic expression, etc. have little effect on phenotypes. Why? This is a fundamental problem of the genotype-phenotype map.
   - Compensation: fundamentally, the perturbation on a functional gene is compensated by other changes in the network. And this may happen at multiple levels: gene-level, pathway-level, and cell-level.
   - Gene duplicates/shared function: duplicated genes may still retain some commen ancestral function despite a high-level of divergence.
   - Within pathway compensation: genes may compensate each other in the same pathway. Ex. a metabolic pathway, when one gene is reduced, the end product is affected, which feedback upon other genes in the pathway s.t. the final outcome is not affected.
   - Between pathway compensation: two related pathways may compensate each other. Ex. a cell may utilize glucose or galactose, or sugar vs. fatty acid. In each case, when one pathway is disrupted, the cell may switch to another pathway so that the final outcome is similar (energy product in this case). Another example: AA transport and AA biosynthesis may compensate each other.
   - Role of feedback loops: they serve as sensors of the perturbations, and activate the necessary compensatory mechanisms. So this is not an independent mechanism of compensation.
   - Evolution of compensation: cells evolve different ways of solving the same task (e.g. ATP production) in a highly variable environment. Once evolved, these networks of different pathways (for often similar/related tasks) provide compensation in cases of genetic/environmental perturbations.

   Effects of gene deletions: what determines the consequence, essentiality vs. robustness?

   - Redundancy/gene duplication: genes with paralogs tend to be less essential. [Gu & Li, Nature, 2003], [Kafri & Pilpel, NG, 2005].
   - Network compensation or genetic buffering: metabolic reflux [Wagner book]. The importance of buffering can be demonstrated by multiple knockout experiments [Costanzo & Boone, Science, 2010]. Alternatively, one can say that negative genetic interactions demonstrate the network buffereing/compensation as a mechanism of robustness to single mutations.
   - Environmental specificity of gene function: some genes are dispensable in nutrient-rich conditions [Papp & Hurst, Nature, 2004]. In general, organisms keep many genes for functioning in a variety of conditions, however, the majority of genes are essential in some conditions [The Majority of Animal Genes Are Required for Wild-Type Fitness, Ramani & Fraser, Cell, 2012].
   - Network topology - connectivity within a network: hub genes (which presumably important for maintaining network connectivity) tend to be more essential [Jeong & Oltvai, Nature, 2000]. Additional evidence may come from protein evolutionary rates [Evolutionary Rate in the Protein Interaction Network, Science, 2002], or from module-level analysis [Song & Singh, PLCB, 2013]. However, the exact importance of this mechanism is somewhat controversial:
     - Experimental bias: [Yu et al, High-quality binary protein interaction map of the yeast interactome network, Science, 2008]. Y2H tends to find interactions of essential genes.
     - Essential protein complexes: see below.

- Essential protein complexes: the genes in important large protein complexes tend to be essential [Zotenko & Przytycka, PLCB, 2008], offering an alternative explanation of the correlation between degree and essentiallity.

Simplified Models of Biological Networks [Sneppen & Semsey, Annu Rev Biophy, 2010]:

- Goal: the potential of negative and positive feedback, as well as their combinations, to generate robust homeostasis, epigenetics, and oscillations.
- Negative feedback loops: The logic of negative feedback makes it ideal for stabilizing systems and minimizing fluctuations. Thus, negative feedback is associated mostly with maintenance of homeostasis.
  - Examples of posttranscriptional modification: small molecule end product inhibition, riboswitches: the metabolite binds to and inactivates the mRNA of the necessary enzyme.
  - Examples of transcriptional regulation: negative autoregulation of a TF.
- When negative feedback is delayed, in time it can give rise to oscillations.
- Negative Feedback in Stress Response: one usually finds that negative FLs in stress response systems involve protein-protein interactions in which a TF regulates the production of a protein that in turn catalyzes the proteolytic inactivation of the TF.
  - Classical example: The p53-dependent response to DNA damage in mammals involves inactivation of p53 by binding to Mdm2. Similar feedback motifs are found in the heat shock and in the SOS response systems in E. coli.
- Positive Feedback: Switch-Like Responses and Bistability.
- Feedback loops in regulation of metabolic networks:
  - The feedback typically involves the formation of a complex between a small molecule and a TF that regulates enzymes that, in turn, affect metabolism of the small molecule.
  - Typically, a transcriptional regulator (R) senses the intracellular concentration of a particular small molecule (s) and regulates transcription of the transport proteins (T) using one FL, facilitating the influx of the small molecule. With the second FL, R controls transcription of enzymes (E) responsible for the metabolism of the small molecule. Miultiple logical structure of the two loops, depending on whether s activates or inhibits R.

Biological Robustness: Paradigms, Mechanisms, and Systems Principles [Whitacre, Frontiers in Genetics, 2012]:

- Functional redundancy:
  - Saturation effects: as seen in the sensitivity of reaction flux toward enzymes with high catalytic activity.
  - Local functional redundancy in genes and metabolic pathways: distributed forms of robustness.
  - Functional redundancy: does not require identical elements and instead often arises between molecular species, developmental pathways that are redundant only within particular contexts (called degenerate).
  - Examples of functional redundancy: (1) glioblastoma cancers, therapies targeting the EGF receptor are thwarted by the co-activation of alternate receptor tyrosine kinases (RTK) that have partial functional overlap with the EGF receptor. (2) In S. cerevisiae, the adhesins gene family expresses proteins that typically play unique roles during development, yet can perform each other's functions when expression levels are altered.
- Feedback and regulatory complexity:

  - Integral feedback control system: e.g. bacterial chemotaxis signal transduction networks. Nested and partially overlapping feedback loops that are not easily deconstructed into isolated, single reference point, controllers may confer robustness.
  - Robust regulatory control can spontaneously emerge in networks of positively and negatively reinforcing regulatory interactions. For instance, in simulations of GRN, Siegal and Bergman (2002) found gene expression patterns became more robust as the number of regulatory connections in the network is randomly increased. This and other studies have shown that robustness can emerge in GRN models without direct selection for robustness.
  - Random Boolean networks: network stability improves if the average number of regulatory factors is increased for each gene or if regulatory interactions become more redundant.
  - Diffuse regulatory control (many weak regulatory interactions) has been proposed to explain stable flux patterns in metabolic networks (Csermely, 2004), meta-stable organization of protein interaction networks.
- Robust architectures:
  - Topology: the integrity of a scale-free network (SFN) is robust to the loss of randomly selected nodes, which typically have low connectivity, yet is fragile to the removal of highly connected hubs.
  - A bow-tie architecture: many inputs are fed into a central core that is the unutilized to produce many distinct outputs. It is believed to contribute to the robustness of many biological networks. In metabolism, the bow-tie architecture provides a formal description of the large fan in of catabolized nutrients that produce a small number of activated carriers (e.g., ATP, NADH) and precursor metabolites (the knot of the bow-tie), that then fan out in the synthesis of numerous building blocks (e.g., nucleotides, sugars, amino acids) and eventually larger macromolecules.

Lethality and centrality in protein networks [Jeong & Oltvai, Nature, 2000]:

- The relationship between node degree and network connectivity: when nodes are ranodmly removed in a yeast PPI network, the network diameter does not change significantly. In contrast, when the highly connected nodes (hubs) are removed, the diameter increases rapidly.
- Correlation between node degree and essentiallity: the likelihood that removal of a protein will prove lethal correlates with the number of interactions the protein has. Only some 0.7% of the yeast proteins have more than 15 links, but single deletion of 62% or so of these proves lethal.

Role of duplicate genes in genetic robustness against null mutations [Gu & Li, Nature, 2003]:

- Motivation: in gene knockout experiment, most of the genes have no phenotype. Can this be explained by gene duplication or distributed robustness (compensating pathways)?
- Singletons tend to have higher fitness effect than duplicates (Figure 1): e.g. 12.4% in duplicates versus 29.0% in singletons for lethal effect, 64.3% in duplicates versus 39.5% in singletons for no effect.
- Sequence similarity and compensation: for genes with highly simliar duplicates (measured by K_A), the compensation rate is high.
- Remark: the growth conditions of the experiment affect the rate of compensation: e.g. as more growth conditions are included, the fraction of genes with no fitness effect upon deletion is lower (from 49% to 39% for singletons).

Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast [Papp & Hurst, Nature, 2004]:

- Method: FBA to determine the flux through the enzymes in different conditions. The model has high accurancy, e.g. it predict the fitness effects of 88% of single-gene deletions under nutrient-rich growth conditions.
- Gene dispensability in various conditions: the model predicts that 68% genes should have zero enzymatic flux under nutrient-rich conditions, consistent with experimental observations. In contrast, 37-68% of the seemingly dispensable genes are environmentally specific, and 76% of genes are predicted to catalyse reactions that are essential under specific conditions.
- Role of gene duplicatio: explain 3.8-17% of gene dispensability. Gene duplication probably does not evolve protection against intracellular noise: if so, one would expect more important (essential genes under flux analysis) genes tend to have isozymes, but this is not what is expected. Instead, gene duplications are correlated with high flux, suggestin dosage is one reason of selection of duplicated genes.
- Remark: need to avoid circular reasoning. Ex. the duplicated genes, even really important, tend to have lower fluxes because the flux is distributed across duplicates, then obviously, the method would not predict those genes to have high flux.

Transcription control reprogramming in genetic backup circuits [Kafri & Pilpel, NG, 2005]:

- Dissimilarly expressed genes: can show backup capacity when one is deleted. This is caused by reprogramming of expression of the other gene. Ex. Acs1 is repressed by glucose, but when Acs2 is deleted, Acs1 acquires an Acs2-like responsiveness to glucose.
- Promoter architecture of backup-providing paralogs: maximal backup coincided with intermediate levels of motif sharing.
- A simple model of transcriptional reprogramming: (Figure 6) both enzymes participate in the same pathway, but when one is deleted, the end product is reduced. This activates the transcriptional controller, which then induces the expression of the paralog.

Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy, [Kafri & Pilpel, PNAS, 2007]:

- Background: duplicated genes could be maintained for a long evolutionary time. For example, duplicate genes evolve more slowly than singletons, despite an initial increased evolutionary rate. What are the characteristics of retained duplicated genes?
- Results: redundant partners are significantly more frequently associated with the so-called protein network "hubs".
- Remark: hub genes are often in essential protein complexes, which may contain more duplicates for reasons such as dosage, functional divergence, etc.

Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality [Zotenko & Przytycka, PLCB, 2008]:

- Hypothesis 1: the removal of hubs disrupts the connectivity of the network, thus the correlation between hubs and essentiallity is due to the maintainence of network connectivity.
- Hypothesis 2: the majority of proteins are essential due to their involvement in one or more essential protein-protein interactions that are distributed uniformly at random along the network edges.
- Test: if essentiality were related to maintaining network connectivity, then one would expect essentiality to be better correlated with these centrality measures than with the node degree. Results: not the case, no better measure is found than node degree.

- Explanation: there exists protein complexes that are densely connected. Essential proteins are enriched in certain protein complexes, called Essential COmplex BIological Modules (or ECOBIMs). Explanation: certain functions that are essential to the cell, for example, transcription regulation or cell-cycle regulation, rely on large multiprotein complexes.
- We found that the fraction of essential proteins among non-ECOBIM hubs is, depending on the network, only 13 to 35% (close to genome-wide average).
- Some ECOBIMs: ribosome assembly, vesicle transport, RNA splicing, etc.

Backup in gene regulatory networks explains differences between binding and knockout results [Gitter & Bar-Joseph, MSB, 2009]:

- Motiviation: in TF knockout experiments, about 3% of TF targets in ChIP-chip show differential expression. And a similar proportion of differentially expressed genes are bound by the TFs. What explains the low overlap of binding and expression targets?
- Data cleaning: (1) high quality ChIP data partially explain the low overlap (for some TFs, two fold); (2) removing genes bound or regulated by many TFs (thus likely not TF-specific) improve the overlap.
- Gene duplication (redundancy): for genes with high level of duplication (paralogs), the overlap is very low (close to 0). The ratio is increased to 12% for genes with virtually no paralogs. Also verified that for some TF pairs, their targets in ChIP-chip indeed overlap substantially (e.f. Fkh1 and Fkh2) and they bind to the targets in a competitive fashion.
- Indirect effects: incoprating 2 steps of indirect effect in the physical network (a gene is indirectly affected by a TF A if A is linked to another TF B, and B binds the gene) leads to an overlap of 22%.
- Remark:
  - Feedback loops between TFs: when one TF is knocked out, it activtes another TF to compensate. However, no evidence of this mechanism in this dataset.
  - Evolution of TF duplication: originally evolves for functional specialization. Ex. originallly all genes are regulated by the same TF, but it may advantageous to fine-tune a subset of genes, and this cannot be accomplished by promoter evolution alone (e.g. fine-tune expression at certain signal, but maintain the same basal expression level).
  - Consequence of TF duplication: the TFs still have similar binding motifs, and similar transcriptional activities, thus there is a certain level of redundancy in the system, as a by-product of function specialization.

Need-Based Up-Regulation of Protein Levels in Response to Deletion of Their Duplicate Genes [DeLuna & Kishony, PLoS Bio, 2010]:

- Methods: about 600 paralog gene pairs, use GFP-fusion to measure the expression of a gene under two conditions: wild type vs. knockout of paralog.
- Compensating change of expression: 15% (29) of the detectable duplicate genes are significantly up- or down-regulated in the paralog-deletion strain grown in rich medium. Most are up-regulated, as expected. Most of these changes are transcriptional. The functions of proteins that show responsiveness are very diverse: metabolic enzymes, cell cycle proteins, heat shock proteins, Golgi proteins.
- Evidence that the upregulation is adaptive (need-based): (1) paralog responsive pairs are highly enriched with synthetic sick/lethal (SSL) pairs; (2) paralog responsiveness depends on specific environment conditions: in the condition where the gene is not necessary, no responsive upregulation (e.g. lys20 expression not changed in lys21 deletion under condition with sufficient lysine).

- Mechanism of responsiveness: end-product regulation of gene expression, direct regulation between paralogs (e.g. Hxk1).

Genes Confer Similar Robustness to Environmental, Stochastic, and Genetic Perturbations in Yeast [Lehner, PLoS ONE, 2010]:

- Methods: single gene deletions (nonessential genes), then measure the strain's: genetic (mutational) robustness - the variability of fitness in different mutations, environmental robustness - variability under chemical or other treatments, stochastic robustness - phenotypic variance across individuals.
- Results: all three are highly correlated. Genetic hubs (identified in synthetic lethal networks) protect genes from all three perturbations. They tend to be chromatin remodeling complexes and molecular chaperons (e.g. Hsp90)

From Hub Proteins to Hub Modules: The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization, [Song & Singh, PLCB, 2013]:

- Hypothesis 1: essential proteins are more connected than non-essential ones within the same modules (essentiality is partly due to the influence of genes within modules, i.e. the local function of genes).
- Hypothesis 2: modules linked with more other modules have higher essentiality, i.e. essentiality at module level is due to the influence of modules on other modules.
- Background: it has been proposed that, in yeast two-hybrid networks, the observed relationship is due to study bias favoring the determination of interactions of essential proteins [Yu & Vidal, Science, 2008].
- Intraprocess interactions are a main factor in the relationship between protein essentiality and interaction degree
- Essential proteins are more central within essential protein complexes.
- Essential complexes and processes tend to have higher cross-talk degree in a module-level network.

5. Distributed design in biology

   Rules for Biologically Inspired Adaptive Network Design [Tero & Nakagaki, Science, 2010]:

   - Background: slime mold grows to capture food source, eventually a network is established to take food from multiple sources.
   - Model: suppose there are n food sources, the objective of slime mold is to grow a network that connects these nodes efficiently (minimum-spanning tree). This algorithm is implemented by the negative feedback: the rates of streaming stimulate an increase in tube diameter, whereas tubes tend to decline at low flow rates. The flow rate is inversely proportional to the distance bewteen two nodes: thus tube will tend to grow in shorter edges (large flow rate).
     - Intuition of the algorithm: at each step, suppose K nodes have already been connected (optimally), then the node outside K with the shortest distance to the K nodes has the maximum flow, thus highest growth rate. This is similiar to the greedy MST algorithm.
   - Remark: many biological systems are essentially solving a computational problem, often distributional. Other examples: fish schooling behavior, honeybee foraging (voting), neuronal synpatic pruning, etc. See `www.algorithmsinnature.org`.

   Algorithms in nature: the convergence of systems biology and computational thinking [Navlakha & Bar Joseph, MSB, 2011]:

- Problem: biological systems (at different levels) often need to solve a problem without any centralized control. This poses special challenges, e.g. lack of information on what choices others might make. What solutions do biological systems evolve?
- Coordination among players: in fly development, a cluster of identical cells will become a mixture of sensory organ precursor (SOP) cells and some other cells, and the two others are connected in a certain fashion s.t. SOP cells are only connected to non-SOP cells.
  - This is similar to the maximum independent set (MIS) problem: elect a set of local leaders in a graph such that all other nodes in the graph are connected to a member of the MIS and no two MIS members are connected to each other. However, traditional algorithm relies on information such as degree of nodes, which is N.A. for individual cells.
  - Biological solution: with probability that increases exponentially over time, each node that has not already connected to an MIS node proposes itself as an MIS node. This leads to a selection pattern in which denser areas in the graph are assigned MIS nodes early on and less populated areas are assigned at later stages.
- Collective decision making:
  - Building consensus: e.g. honeybees choosing a new home, fishes forming groups to avoid predators.
  - Synchronization: e.g. of fireflies.
- Resource allocation/transport: e.g. in taking food from multiple sources in slime molds [Tero & Nakagaki, Science, 2010].
- Fault tolerance in a network: a network faces the challenge of (random or targeted) node failures (from various sources). The design of the network should make it tolerant to these failures. Ex. gene duplication in GRN.

6.4 Modeling Special Systems

1. DNA looping [Vilar & Saiz, COGD, 2005; Saiz & Vilar, PNAS, 2005; Vilar & Saiz, COSB, 2006]

   Thermodynamic model:

   - Positional free energy: the free energy of binding/interaction depends on the concentrations of the different components through positional free energy, which has the form, for molecule A:

     $$\Delta G_{pos} = \Delta G^0_{pos} - RT \ln N_A \tag{6.14}$$

     where N_A is the number of A molecules. This is derived from the probabilitic argument: the probability of finding a molecule in a restricted space (where it can bind with the target, e.g. DNA) is proportional to concentration; on the other hand, this probability can be written as a Boltzmann distribution, as a function of positional free energy. Typical values of ΔG^0_{pos} is around 15 kcal/mol.
   - Free energy of looping: ΔG_l, should be part of the free energy of a state s where DNA looping is involved. For example, an enhancer with two operator sites O_1 and O_2, the free energy of state where both sites are occupied by a single dimer is:

     $$\Delta G_{O1-O2} = \Delta G_{O_1} + \Delta G_{O_2} + \Delta G_l \tag{6.15}$$

     where ΔG_{O_1} and ΔG_{O_2} are free energy of binding of O_1 and O_2 respectively.

   DNA looping in lac operon:

- Short loops: the main determinant is DNA elasticity (DNA bending, twisting, etc.). The system being studied is lac operon where the repression by lacR depends on the distance between two operator sites of lacR, O_1 and O_{id}. The promoter is activated if the main operator O_1 is not occupied. This allows on to express ΔG_l, the free energy of looping, in term of the efficiencies of repression, R_{loop} (when looping is allowed in experiment) and R_{noloop}:

$$\Delta G_l = -RT \ln \frac{R_{\text{loop}} - R_{\text{noloop}}}{R_{\text{noloop}} - 1}[N] \tag{6.16}$$

  where $[N]$ is the concentration of the repressor. The main findings are:

  - At spacing less than 200 bps, the repression effect as well as ΔG_l is phase-dependent. Fourier analysis showns that there are two periodicities: one main period about 10.9 bp, and the other weaker one at 5.6 bp.
  - The range of ΔG_l is $7.5 \sim 10$ kcal/mol, with the amplitude of fluctuation (2.5 kcal/mol) similar to the free energy of interaction between TFs. Thus the transcriptional effect is strongly dependent on precising DNA positioning.

  The dependence of ΔG_l on distance is made more complex by the possibility that the protein molecules can take different conformations to dimerize (to accomodate different spacings) [Swigon & Olson, PNAS, 2006].

- Long loops: the main determinant is the loss of entropcy when two DNA regions are tied together [Hanke & Metzler, BiophsJ, 2003]. Phase dependence disappears at distance more than 400 bps. The free energy can be approximated by:

$$\Delta G_l = \Delta G_{l_0} + \alpha RT \ln(l/l_0) \tag{6.17}$$

  where l_0 is a reference length and α is a constant. The experimental values are $\alpha \approx 1.24$ and $\Delta G_{l_0} \approx 4.72$.

2. DNA structure

   Bubble formation in DNA [Rapti & Bishop, PRE, 2006]:

   - Aim: given a DNA sequence, the probability that a bubble (from denaturation) is formed at a certain position.
   - Background: the distance between two bases of a pair can deviate from the optimal distance determined by H-bonds because of thermo motion. If this distance is larger than a threshold for an extended sequence, then a bubble is created in this region.
   - Model: a DNA sequence of length N, the distance (deviation from equlibrium distance) at any position n is y_n, the sequence can exist in a micro-state defined by $(y_1, y_2, \cdots, y_N)$. The energy of a micro-state is sum of the following two forms of energy over all positions:
     - Morse potential: $V(y_n)$, from H-bond at position n.
     - Anharmonic potential: $W(y_n, y_{n-1})$, from the stacking interaction between adjacent bases (the deviation in position n and $n-1$ causes a relative shift of poisition of bases at the two positions, thus create additional potential).

     The partition function of the system is computed from Boltzman distribution, integrating over all $y_n, 1 \le n \le N$.
   - Remark: the transfer integral method for computing partitition function, over continuous variables. An analogy of transfer matrix method where variables are discrete.

3. Neuronal networks

   Strategies of synaptic pruning: [Navlakha & Bar Joseph, lab meeting]

- Problem: in development, neuron networks form edges from learning. What are the principles/algorithms of determining which edges are connected?
- Model: start from a random network, learning consists of stimulation in the form of (s, t) (source-target) pairs (for message passing), and the edges are grown or pruned s.t. the future message passing is easier. The main goal is: (1) efficient for message passing; (2) robust: more paths between any two pairs; (3) maintanence cost (number of edges).
- Analysis: how growing or pruning strategy, and the parameters of the strategy (threshold for adding/removing edges) influence the results of learning.
- Remark/lessons:
  - Neuron development: (1) genetic determinism: e.g. six layer structure, overall conserved brain organization, etc. (2) plasticity: learning shapes the neuronal connections.

Synchrony in stochastic neuronal networks [DeVille, Bulletin of Math Biol, 2009]:

- Problem: firing pattern in a neuron network: are the firings synchronized? How often do they fire? etc.
- Model:
  - Assumptions: neurons are connected as a graph. Each neuron may be in one of $1, 2, \cdots, K$ states. At each time point, if a neuron reaches state K (firing threshold), it will fire, leading to: first, each of its neighbor has probability p to be promoted (state variable adds 1); second, its own state will reset to the basal level.
  - Behavior characterization: firing dynamics (number of firings in each time point) and firing rate (average firing per time point); synchrony (whether multiple neurons fire together in multiple time points); residence time (the average time a neuron stays in the rest state); etc.
  - Mathematical formulation (deterministic): when the number of neurons, N, is large, the system is described by deterministic equations, similar to Markov Chain. The frequency distribution of states at any time point, is a function of the distribution in the previous time point.
  - Mathematical formulation (stochastic): analogy with random graphs, since when a neuron fires, it will have a probability to activate one of its neighbor, which in turn will activate its neighbor, etc.
  - Simulation: simulate the dynamics of the neuron networks, under different settings (parameters and network topology).
- Analysis: the goal is to study the behavior pattern (as described before) of the system, and how it depends on the structural features of the system, most importantly, the probability p and the network topology. The main results are:
  - When p is large, the system fires synchronously; when p is small, asynchronously; when p is between, the system switches between synchrony and asynchrony, similar to a two state Markov chain. The intuition is: when p is large, it is similar to a random graph with large connectivity, which will have a giant component; when p is small, the largest component of a random graph is exponentially smaller than the size of the graph $O(\log n)$.
  - Firing rate is not a monotonic function of p, when p is larger than some value, the firing rate actually reduced (because of synchrony).
  - Firing pattern depends on the network topology (small-world, scale-free, random, etc.): e.g. scale-free networks tend to have few firings, probably related to the fact that in these networks, there are a few number of hubs.
- **Remark:**

- For a system to be analzyed, recognize what are the important behavior/function features. Ex. for a neuron network, it may be its synchrony; for a GRN, it may be the quantity of a key protein; etc. For a very complex system, characterize the behavior of a system with relatively simple variables, that are tied to the imporant behavior features.
- The goal of the analyis of system is generally: characterize the behavior of system, and how it depends on the structural features of the system. These structural features can often be described by a few parameters, hence, bifurctation analysis, etc.
- Whenver possible, come up with an intuitive understanding of the results.

4. Synthetic biology

   Synthetic promoters in plants [Venter, TIPS, 2006]:

   - Goal: design promoters that drive expression of transgenes. Common requirements are:
     - Constitutive expression
     - Inducible expression: by environmental cues, biotic and abiotic stress (e.g. pathogens, drought), hormones, chemicals.
     - Tissue-specific and development-stage specific.
   - Methods: in general, do not know precise sequence-expression relationship, so will rely on expression classes. For instance, we have expression profiles corresponding to multiple conditions, and need to create promoters that drive expression at at both conditions. A possible procedure:
     - For expression class, construct the promoter model (which motifs are important, any constraints)
     - Create a new promoter that is the union of the models for both conditions.