

Contents

2 Sequence Alignment	3
2.1 global alignment	3
2.2 local alignment	3
2.3 Affine gap alignment	3
2.4 Match scores	3
3 Short read assembly	3
4 Sequencing	3
5 Normalization	4
5.1 Methods	4
5.2 Transformation	5
6 DE gene	6
6.1 Problem	6
6.2 Statistical models	6
6.3 Hypothesis testing	6
6.4 log-likelihood ratio test	7
6.5 limitations	7
6.6 Multiple hypothesis testing	7
7 Clustering	8
7.1 Motivations	8
7.2 hierarchical clustering	8
7.3 Model based clustering	8
7.3.1 Partitional clustering	8
7.3.2 Gaussian mixture models	8
7.4 Graph based clustering / Iterative partitioning	9
7.5 Cross-validation	9
7.5.1 internal validation	9
7.5.2 external validation	9
7.5.3 hypergeometric distribution	9
7.6 Bi-clustering	10
7.7 Importance of ordering	10
8 Classification	11
8.1 Types of classifiers	11
8.2 Weighted voting	11
8.3 Bayes classifier	11
8.4 Naive Bayes classifiers	11
8.5 Feature selection	11
8.6 SVM	12
8.7 Challenges in computational analysis of omics data for development of molecular signatures	12
9 Single Cell	13
9.1 Dimensionality reduction	13
9.2 Graph-based method	14
9.3 Supervised dim-red: Neural networks	15
9.3.1 Additional NN architecture: Siamese	15
10 ChIP-Seq	16
10.1 Background	16
10.2 Questions related to TF binding	17
10.3 ChIP-seq	18

11 Cis-regulatory motif	21
11.1 Background of transcription factor binding motif analysis	21
11.2 Finding regulatory motifs	21
11.3 Motif finding based on EM	21
11.4 Transform a PWM into log likelihoods	22
11.5 Comparative genomics - PhyloCon	22
12 Haplotype inference	23
12.1 Genetic polymorphism	23
12.2 SNP genotyping	23
12.3 Haplotype inference	23
12.3.1 Probabilistic model for haplotype inference	23
12.4 Phase	23
12.5 Linkage disequilibrium (LD)	24
13 Population structure	26
13.1 Background	26
13.2 Inferring pop structure from genotype data (nowadays mostly just PCA)	27
14 Linkage Analysis, GWAS(Genome Wide Association Study)	29
14.1 Background	29
14.2 GWAS	29
14.2.1 single SNP association analysis	29
14.3 multimarker association test	30
14.4 using reference datasets for genotype imputation	30
14.5 Epistasis	31
14.5.1 Detecting epistasis	32
14.6 Population struture and association analysis	32
14.7 computationally solve the problem	32
15 eQTL (expression quantitative trait locus)	33
15.1 limitation of likage analysis and GWAS	33
15.2 terminologies	33
15.3 human vs model organism	34
15.4 allele-specific eQTL mapping	35
15.5 allele specific eQTL mapping	35
15.6 limitations	36
16 scRNA-seq & CRISPR	37
16.1 Background	37
16.2 co-expression eQTLs	37
16.3 CRISPR	37
16.4 array vs pooled screening	38
17 Comparative genomics	39
18 Epigenetics	40
19 3D genomes	41
19.1 Genome organization	41
19.2 nuclear organization	41
19.3 Hi-C	41
19.4 3D genome in different resolutions	42
19.5 what establishes chromatin interaction	42
19.6 Computational methods	42
19.6.1 Normalization	42

2 Sequence Alignment

2.1 global alignment

- Needleman & Wunsch
- the usual DP u know

2.2 local alignment

- free ride to any starting position + ending pos

2.3 Affine gap alignment

- 3 layer graph

some tricks:

- if only allow gaps of certain length, take max over all possible gap length

2.4 Match scores

- log-odds scoring
- PAM model
- BLOSUM

3 Short read assembly

4 Sequencing

5 Normalization

Assumption

Assume the subset of the genes in the sample are expressed at the same total level across all cells / samples

5.1 Methods

- **RPKM - read per kilobase of transcript per million reads of library**

- Corrects for coverage, gene length
- 1 RPKM ~ 0.3 - 1 transcript / cell
- Comparable between different genes within the same dataset
- packages: TopHat / Cufflinks

$$RPKM = \frac{\text{no. of reads mapped to gene G} \cdot 10^9}{\text{total no. of mapped reads} \cdot \text{length of gene G}}$$

- **FPKM - fragments Per Kilobase of exon model per Million mapped fragments**

- similar to FPKM
- for pair-end seq, $FPKM = 2RPKM$

$$FPKM = \frac{\text{no. of fragments mapped to the exons} \cdot 10^9}{\text{total no. of mapped reads} \cdot \text{length of the exon}}$$

- **TPM - transcript per million**

- Normalizes to transcript copies instead of reads
- Longer transcripts have more reads
- RSEM, HTSeq
- steps:

1. $RPK = \frac{\text{read count of gene}}{\text{length of gene}}$
2. scaling factor = $\sum RPK/10^6$
3. $TPM = \frac{RPK}{\text{scalling factor}}$

- additional assumptions

1. values are exactly the same between runs (genes could have variable values)
→ quantile normalization
2. values are normally distributed w same mean and var across samples
→ scale factor normalization
3. assume some genes have stable values over runs (rank invariant)
→ invariant set normalization

- **TMM - Trim Mean of M**

- high level idea: remove extreme values before normalizing
- intuition:
sample A and B both have 100 genes sequenced to the same depth, 90 genes in A and B are expressed at about the same level, last 10 genes expressed at extremely high level in B
→ could appear as the first 90 genes expressed twice as high in A than in B, which does not make sense
Reason: fixed amount of sequencing real estate

- observed counts of gene g in experiment k: $E[Y_{g,k}] = \frac{\mu_{g,k} L_g}{S_k} N_k$, where
 - * $S_k = \sum_g \mu_{g,k} L_g$
 - * $\mu_{g,k}$ true expression level of gene g in experiment k
 - * L_g length of gene g
 - * N_k total no. of reads for experiment k
- $M_g = \log \frac{Y_{g,k}/N_k}{Y_{g,k'}/N_{k'}}$
- $A_g = \log Y_{g,k}/N_k + \log Y_{g,k'}/N_{k'}$
- trim off genes with extreme M and A values (in the paper, took 5% and 30% as cutoff)
- compute ratio based on all other genes
- * needs to estimate the differences between non-DE genes and thus need to have roughly same number of genes up- or down- regulated between 2 conditions. (roughly balanced)

5.2 Transformation

- While ratios are useful, not symmetric
→ hard to visualize different changes
- use a log transform, and focus on the log ratio

$$y_i = \log \frac{R_i}{G_i}$$

- Empirical studies have also shown that in microarray experiments the log ratio of (most) genes tends to be normally distributed

6 DE gene

6.1 Problem

- log-fold change isn't ideal (there are problems)
- noise from technology too
- expression being double the amount → how reliable? → variation estimation
- easy to estimate variation if large biological models, but often only 2-3 replicates available
- Solution: often assume some mathematical model

6.2 Statistical models

- Gaussian

– mean μ , var σ^2

A couple methods based on counts...

- Binomial

–

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

– mean: np , var: $np(1-p)$

- Poisson

–

$$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

– mean = var = λ

- Negative Binomial

– common when data has variance $>>$ mean (overdispersed)

– defined as the number of successes in a seq of Bernoulli trials before **some number** of event r occurs, eg. no. of trials until 3 heads

– Thus appropriate for modelling biological replicates

–

$$P(x = n) = \binom{n-1}{k} p^k (1-p)^r$$

where r : no. of failures, k : no. of successes until r failures, and $n = k + r$

– mean: $rp/(1-p)$, var = $rp/(1-p)^2$

– can be re-written in a format similar to Poisson dist.

– Poisson assumes same mean and variance, whereas NB assumes larger variance

– → dispersion parameter α s.t.

$$\text{mean} = \lambda, \text{var} = \lambda + \alpha\lambda^2$$

– NB = Poisson when $\alpha = 0$

6.3 Hypothesis testing

- H_0 : mean expression of the gene under two conditions are the same
- p-value: how likely it is to see the data we observe under H_0
- eg. one sample t test, two sample t test, non-parametric rank test, χ^2 test

6.4 log-likelihood ratio test

- Compute the likelihood under H_0 and H_1 . ie.

$$\Pi_{i \in A} P(x = i | \text{model param}) \Pi_{i \in B} P(x = i | \text{model param})$$

- log-likelihood ratio = $\log \frac{L1}{L0}$ (note we are assuming equal variance under both hypothesis)
- degree of freedom: no. of free parameters - 1

6.5 limitations

- assume specific probabilistic model
- need many replicates
- multiple hypothesis testing issue

6.6 Multiple hypothesis testing

- eg. 1 trillion monkeys, one of the randomly typed out shakespeare
- Bonferroni Correction
 - very conservative
 - may cause us to miss out genes
 - adjusted p-val = original pval / no. of genes testing
- FDR
 - 100 genes identified w p-val 0.05, then 5 genes are probably falsely discovered
 - **more stuff from hw2???**
- Permutation based methods
 - idea is to determine the prob. of seeing the real data in a random sample
 - divide by no. of permutations done
 - can be a problem when the no. of samples small
- time series data are often too hard to be considered, not always possible

7 Clustering

7.1 Motivations

- clustering genes → markers, functional assessment, expression pattern recognition
- clustering cells → cell type identification / classification
- clustering both → associate subset of genes with subset of experiment
- **distance measure:** (dissimilarity) $\in \mathbb{R}$
 - eg. euclidian dist, 1-correlation coefficient

7.2 hierarchical clustering

- dendograms!
- bottom-up approach
- algorithm:
 1. each item → a cluster
 2. iteratively select the two closest cluster and merge
 3. stop when only 1 cluster left

7.3 Model based clustering

7.3.1 Partitional clustering

- non-hierarchical, each instance is placed in exactly one of k non-overlapping clusters
- eg. k-means
- algorithm:
 1. Decide on a value for K, the number of clusters.
 2. Initialize the K cluster centers (randomly, if necessary).
 3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.
 4. Re-estimate the K cluster centers, by assuming the memberships found above are correct.
 5. Repeat 3 and 4 until none of the N objects changed membership in the last iteration.

7.3.2 Gaussian mixture models

- algorithm:
 1. decide k
 2. initialize w random param
 3. E-step: assign probabilistic membership to all input samples

$$p_{i,j} = p(C = i|X_j) = \frac{p(C = i|X_j)p(C = i)}{\sum_k p(C = k|X_j)p(C = k)}$$

- 4. M-step: re-estimate parameters based on probability membership
- 5. stop when change in param in one iteration < some threshold
- the importance of initialization

7.4 Graph based clustering / Iterative partitioning

1. construct the neighborhood graph
2. assign weights to the edges (similarity)
3. partition the nodes using the graph structure

	Bottom up	Model based	Top down
Running time	naively, $O(n^3)$	fast (each iteration is linear)	could be slow (matrix transformation)
Assumptions	requires a similarity / distance measure	strong assumptions	general (except for graph structure)
Input parameters	none	k (number of clusters)	either k or distance threshold
Clusters	subjective (only a tree is returned)	exactly k clusters	depends on the input format

- Number of clusters:
 - find “knee” in objective function

7.5 Cross-validation

Recall that GMMs is a generative model. We can compute the likelihood of the left out data to determine which model (number of clusters) is more accurate

7.5.1 internal validation

- without external data
- **Coherence**
- inter-cluster similarity and the average intra-cluster similarity
- Requires the definition of the similarity / distance metric
- **Stability**
- whether clusters are resistant to minor perturbation
- measure similarities between two clustering runs, each with randomly sampled subset of objects (80-90%)

7.5.2 external validation

- matched to known data
- eg. mySigDB

7.5.3 hypergeometric distribution

- used to compute significance of overlap between two sets
- if l elements overlap, the p-val is

$$p = \text{prob}(l > \hat{l}) = \sum_{j=l} \Pr(\text{exactly } j \text{ matches})$$

7.6 Bi-clustering

find subsets of genes that behave similarly among a subset of experiments

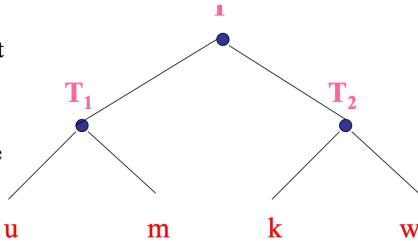
7.7 Importance of ordering

- clustering tree
 - could have 2^n leaf ordering
- goal: give a leaf ordering that maximize similarity of adjacent elements
- let Φ be search space, $v_1 \dots v_n$ leaves, S similarity matrix

$$\max_{\phi \in \Phi} \sum_{i=1}^{n-1} S(v_i^\phi, v_{i+1}^\phi)$$

for any pair of leaves (u, w)
which could be on different
corners (leftmost and
rightmost) of T .

For a leaf $u \in T$, $C_T(u)$ is the
set of all possible corner
leaves of T when u is on
one corner of T .



$$L_T(u, w) = \max_{m \in C_{T_1}(u), k \in C_{T_2}(w)} L_{T_1}(u, m) + L_{T_2}(k, w) + S(m, k)$$

- time complexity: $\Theta(n^3)$
- space complexity: $O(n^2)$

8 Classification

8.1 Types of classifiers

- generative
 - eg. mixture models
- discriminative
 - eg. log regression

8.2 Weighted voting

- v = vote
- decision boundary = $\frac{v_{win} - v_{lose}}{v_{win} + v_{lose}}$

8.3 Bayes classifier

- class 1 $\rightarrow X \sim N(\mu_1, \sigma_1)$
- class 0 $\rightarrow X \sim N(\mu_0, \sigma_0)$

$$p(class=1|X) = \frac{P(X|class=1)P(class=1)}{P(X|class=1)P(class=1) + P(X|class=0)P(class=0)}$$

- pick the highest among $p(class1|x)$ and $p(class0|X)$, decision boundary is the loglikelihood of the two

8.4 Naive Bayes classifiers

- assume each X_i independent

$$p(X|class=1) = \prod_i p(X_i|class=1)$$

- aka assume $X_i \sim N(\mu_i^1, \sigma_i^2)$
- simple but works decently well

8.5 Feature selection

- H0: gene not predictive, H1: gene predictive

8.6 SVM

- max margin classifier

SVM optimizes the following:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \varepsilon_i$$

subject to the following inequality constraints:

For all x_i in class +1

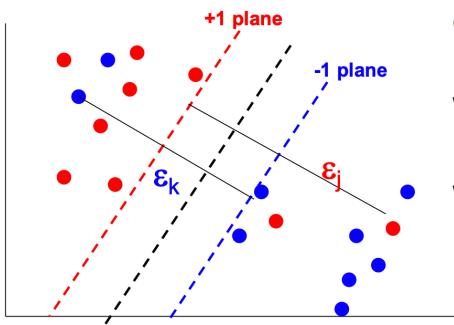
$$w^T x_i + b \geq 1 - \varepsilon_i$$

For all x_i in class -1

$$w^T x_i + b \leq -1 + \varepsilon_i$$

For all i

$$\varepsilon_i \geq 0$$



-

8.7 Challenges in computational analysis of omics data for development of molecular signatures

- ez to develop predictive model → easy to believe a model is good when it's not
- several theoretical/practical problems exist
- most cannot be validated on an independent cohort, unfortunately :/

9 Single Cell

Goals:

1. cells differentiates into sub-celltypes
2. unknown celltype discovery

9.1 Dimensionality reduction

Motivation:

1. high dim data often has lower dim representation w/o much reconstruction error
2. lower dim representation can often represent info about high dim pairwise dist.

Types of dim-red:

- **Global methods**
 1. all pairwise dist equally impt
 2. lower dim pairwise dist fit high-dim ones
 3. often use magnitude or rank order
- **Local methods**
 1. only local dist reliable in high dim
 2. more weight on modelling local dist correctly

Methods:

- PCA
 - finds directions with largest variance
 - minimize squared reconstruction error
 - equiv to liner autoencoders
 - Steps of PCA
 1. \bar{X} : mean of all samples(usually rows), adjust $X \rightarrow X' = X - \bar{X}$
 2. covar matrix $C = X'^T X'$
 3. find eigenvectors and eigenvalues of C , ie. all pair of \vec{v}, λ st. $C\vec{v} = \lambda\vec{v}$
 4. eigenvalues can be used to calculate percentage of total variance for each component

$$v_j = 100 \frac{\lambda_j}{\text{total eigenvalue}}$$

This is non-parametric method, do not insist on a parametric encoding function

- Multi-Dimensional Scaling
 - arrange low dim points to minimize diff between pairwise distances in the high and low D space
 - a possible approach: start w a random vector, perform gradient decent
 - **is there something to do with PCA?** then we don't need iterative method
- Sammon (non-linear autoencoder)
 - with extra layers, much more powerful than PCA, but can be slow to converge, and can get stuck on local optima
 - Multi-Dimensional Scaling(MDS) can be made non-linear by giving higher weights to smaller distances, a popular formula is

$$\text{cost} = \sum_{ij} \left(\frac{|x_i - x_j| - |y_i - y_j|}{|x_i - x_j|} \right)^2$$

where x is high-dim dist, and y is low-dim dist

- still slow and get stuck on local optima

9.2 Graph-based method

- address uniform circularity
- Isomap is a dim-red technique based on graphs
 - each datapoint is connected to k nearest neighbor in high-dim
 - edge weights = euclidean dist
 - approx of distance = shortest path in contracted graph
- Probabilistic local MDS
 - local distances are more impt than non-local ones
 - in this way all local distances are given equal importance
- stochastic neighbor embedding(SNE) has a probabilistic way to decide if a distance is local
 - convert global distances into probability of one datapoint picking another datapoint as its neighbor
(what defines a neighbor tho) - still about isomaps?
 - each point in high-dim has a conditional probability of picking any other point as its neighbor
 - distribution (some sort of Gaussian) is over high-dim distances (if high-dim coords unavailable, a similarity / dissimilarity matrix may be used)
 - $p_{j|i}$ is the prob. of picking j given starting at i in high-dim.
$$P_{j|i} = \frac{e^{-2d_{ij}^2/2\sigma_i^2}}{\sum_k e^{-d_{ik}^2/2\sigma_i^2}}$$
 - having the probabilities potentially allow us to throw away the raw high-dimensional data
 - evaluation done using pairwise distance in low dimensional map (shows how well the lower dim representation models high-dim data ig)
 - $q_{j|i}$ is the prob of picking j given starting at i in low-dim
 - compute the Kullback-Leibler divergence between prob. in the high-dim and low-dim spaces (why not just use dist in high dim?) - more space / time efficient
 - nearby pts in high-dim should be close in low-dim
- picking σ used to compute p - the radius of the gaussian
 - different radius is needed in different parts of the space to keep the no. of neighbors constant
 - big radius \rightarrow high entropy for distribution over i 's neighbors
 - small radius \rightarrow low entropy
- Symmetric SNE
 - simpler than stochastic
 - works best if different procedures are used for computing p 's and q 's.
 - compromise: no longer guarantees that if using same dimension will produce optimal solution
 - turn conditional prob into symmetric pairwise probabilities
- Optimization methods for SNE
 - simulated annealing could lead to better global optimization
 - add Gaussian noise to y location in each update
 - spend longer time at noise level where global structure start to form
 - t-SNE - use Gaussian at many (infinite) spatial scales, cheaper as we don't have to exponentiate anymore (why????)
- UMAP (uniform manifold approximation)
 - like stretching a rubber sheet s.t. distance to nearest neighbor == 1 for each datapoint

9.3 Supervised dim-red: Neural networks

- last few layers have much fewer values than inputs
- use intermediate layers as lower-dim representations
- can easily add prior biological knowledge, such as protein interactions or transcription factors
- essentially, some nodes in the hidden layer are same as before, others are based on biological info

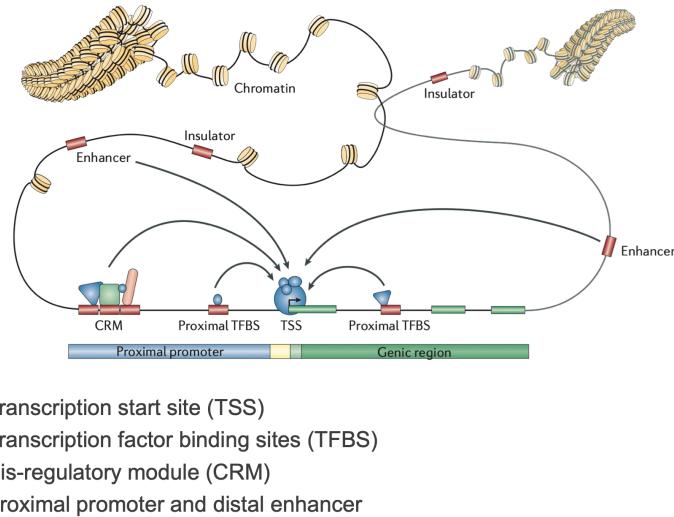
9.3.1 Additional NN architecture: Siamese

- supervised, but not trying to maximize training accuracy
- input: whether each pair is similar
- output: binary label of similar / not similar
- thus directly optimize dim-red layer for KNN

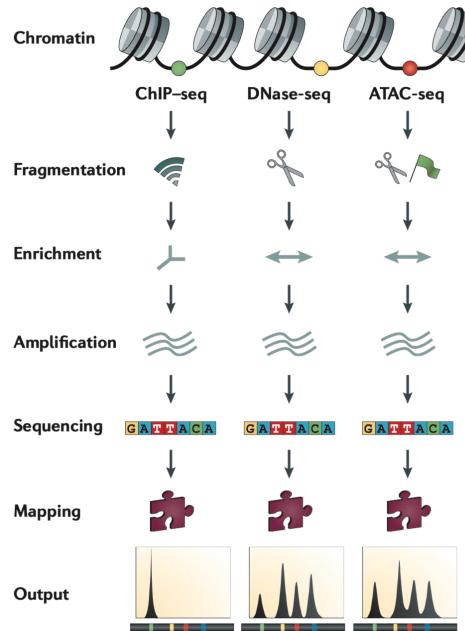
10 ChIP-Seq

10.1 Background

- transcription factors bind to specific locations on the chromosome
- the binding is highly specific
- this has impact on gene regulation



- The core promotor regions have about 300 TFs, (general transcription machinery, required for the transcription of most things), another 1500 TFs for others (proximal enhancer/promotor/silencer, only affect some genes)
- comparisons of different whole genome enrichment technologies



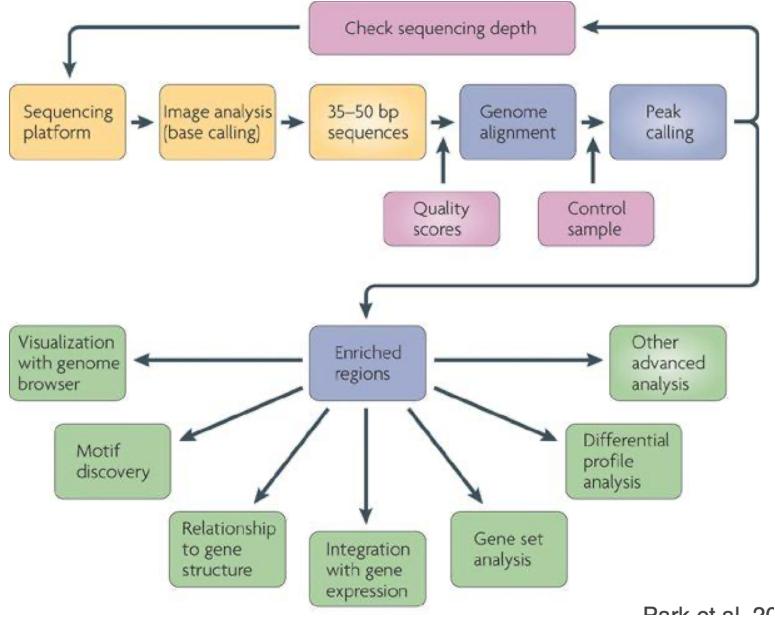
10.2 Questions related to TF binding

- Where do they bind
 - some almost always bind to proximal promoters
 - others could bind to many regions
 - How does specific binding work
 - exists a consensus motif (most common sequence)
 - looks like this
- 
- shows which nucleotide is most abundant at each position, represented as **Position Weight Matrix (PWM)**
 - **Motif:** a **recurring** sequence with biological importance
 - * can be constant or have a few variable elements
 - * often serves as binding site for TFs /proteins
 - * often very short (5-25)
 - * often distant from gene
 - * often has inexact repeating pattern (challenge for identification)
 - sometimes also an effect of protein-protein interactions,
ie. protein conformation change upon interactions etc?
 - to determine binding site, often uses **Protein Weight Matrix**, but it is over-simplifying
- How to identify where they bind
 - genes regulated by the same TF are likely to have the same motif in regulatory region
 - to find genes regulated by the same TF → **comparative genomics** could be very useful
 - eg. knock-off of specific TFs → lower expression of a set of genes
 - group of genes that are co-expressed across many set of experiments are also likely to be regulated by same TF
 - challenges
 - * we don't know the exact motif seq
 - * we don't know how far it will be
 - * motifs can differ slightly across
 - * how to distinguish it from just "random" sequences that happen to repeat?
 - Multiple sequence alignment (MSA)
 - motif-finding based on EM algorithm (MEME-suite - Multiple EM for Motif Elucidation)
 - How is it involved in gene regulation
 - Is it useful for gene regulatory network?

10.3 ChIP-seq

- technology:
 - ChIP: chromatin immunoprecipitation
 - studies protein interaction with DNA
 - able to map global binding site precisely for any protein of interests
- 1. chromatin immunoprecipitation + high throughput sequencing
- 2. detect genome-wide location of TF and other binding proteins in labs
- 3. find all DNA seq bound by TF-X
- 4. try to learn the regulatory mechanisms of a TF or DNA-binding protein
- General strategies to call ChIP-seq peaks
 - ChIP-seq yields distributions for tags from forward and reverse strands
 - overlap of the two can be observed
 - actual binding site of TF should be between the 2 distributions
 - from the difference between the 2 peaks → formulate a single **peak summit**
- MACS: model-based analysis for ChIP-seq
 1. map reads using Bowtie2
 2. get ChIP-seq reads around but may not contain binding site
 3. sequence are from ends of randomly chopped segments – > hopefully overlap at binding site
 4. produce 2 adj. set of read peaks located about $2 \times$ fragment length away
 5. **shift distance**: dist between read peaks at which will find true peak
 6. automatically subtract control to define a final set of peaks
 - input:
 - bandwidth : sonication size
 - mfold: high-confidence fold enrichment
 - slides $2 \times$ bandwidth window across genome to find regions with tags $> mfold$ enriched compared to a random tag
 - random sample 1000 high quality peaks, separate their Watson Crick tags, and align by midpoint
 - > 2 peaks, shifts = $d/2$
 - tag distribution *Poisson*
 - MACS uses a dynamic parameter λ_{local}
 - **P-val of peaks?**
 - **FDR?**
- Downstream analysis
 - to identify motif for TF using ChIP-seq peaks – **MEME**
 - to find out what the sequence motifs resembles – **TomTom**
 - to find peak regions of known motifs – **FIMO**
 - look up biological pathways or functions of target genes of TF – **GREAT**
- Practical pipeline
 - overview: TF-binding → co-factors → DNA histone modifications → DNA DS break

- goal: converge NGS reads to signals / peaks tracks → infer potential TF binding regions



- fastq → FASTQC → bam/alignment

→ wig:

- narrow → narrow peak caller
—calculate QC metrics



- broad / mixed → broad peak caller —calculate QC metrics



- QC good → differential enrichment
- QC bad → combine replicates (IDR)

- Steps: (eg. ENCODE)

0. experiment design

- replicates (more replicates >> greater coverage)
if more TF targets → multiple hypothesis testing???
- choice of control
 - open-chromatin regions fragmented more easily than closed ones
 - copy number variation in cancer samples
 - potential non-specific antibody binding
 - repetitive regions could cause alignment errors
 - MOST DO CONTROL AS PART OF THE INPUT
 - some also do mock IP as control (DNA obtained from IP by a mock antibody such as IgG)
- * **matched control** is required for downstream analysis

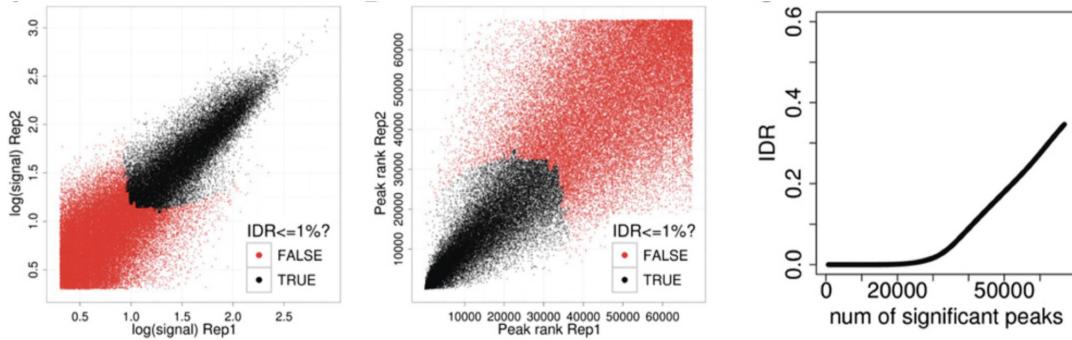
1. input data & QC

- fasta format → FastQC

2. Alignment and filtering

- alignment tools: bowtie, bwa → output bam
- QC = % mapped
- filtering tools: samtools, picard

- QC = non-redundant fraction
3. Peak calling
- narrow peak
 - broad peak
 - QC: fraction of reads in peak (FRiP)
 - * FRiP values correlate positively and linearly with the number of called regions
 - QC: strand cross-correlation
 - * shift vectors with each other and calculate correlations
 - * plot shift size w correlation
 - Irreproducibility discovery rate (IDR)
 - * avoid of choices of initial peak caller cutoffs
 - * modelling peaks pairs from replicates as belonging to two groups: reproducible group and irreproducible group



- visualization
 - * UCSC genome browser etc
 - * IGV genome browser
- eg. let $\lambda_{a,b}$ be the observed read mapping coverage of the peak region.
 - How to calculate the P-value of an individual peak?
 - How to calculate an empirical FDR of the identified peak for each P-value?
 - P-val: $P(x > \lambda_{ob})$
 - We use the same parameters to find ChIP-seq peaks over control and control peaks over ChIP (that is, doing a sample swap) to have the empirical FDR. Here the empirical FDR is defined as # of control peaks / # of ChIP-seq peaks.

11 Cis-regulatory motif

11.1 Background of transcription factor binding motif analysis

- basic for TF binding, where they bind
- Position weight matrix (PWM)
- *De novo* motif binding
 - comparative genomic approach → find motifs within same species

11.2 Finding regulatory motifs

- goal: given a set of genes that are probably regulated by the same TF, find the TF-binding motif in common
- challenges:
 - we don't know the sequence of motifs
 - it could be located near or far
 - motifs could differ slightly from one to another
- assumptions:
 - functional part of genome evolve more slowly than non-functional part → greater selection pressure
 - sequence alignment → identify conserved region
 - look for functional features in conserved regions
 - comparative genomics rely on being able to detect similar regions across genomes

11.3 Motif finding based on EM

- E step: fill in the expected values of the missing variable
- M step: regular MLE using the value computed in the E step and values of the other variables
- guarantee convergence to local optima
- eg. MEME uses EM to iteratively refine PWM and identifying sites for each PWM
 1. estimate motif model (PWM)
 - start with a k-mer seed (random, or specific)
 - build PWM by incorporating some background frequencies
 - eg. build a matrix with 4 rows and k cols where each col corresponds to an index
 2. identify examples of the model
 - for every k-mer in input, identify its prob. given the PWM
 - eg. score all k-mers using a sliding window with the current matrix
 3. re-estimate the motif model
 - calculate new PWM based on the weighted frequencies of all k-mers in input sequence
 - eg. more likely k-mer from step 2 count more towards new PWM freq. the new PWM is estimated based on the weighted contribution of all k-mers
 4. iteratively refine the PWMs and identifying sites until convergence
- a bit more formally
- input:

- a set of sequences $s_1 \dots s_N$
- partition s into k-mers for a fixed k
- set of letters, usually assumed to be $\{A, T, C, G\}$
- params of the mixture model: $\lambda_1\theta_1 + \lambda_2\theta_2$
 - TF motif model θ_1
 - background model θ_2
 - model probabilities: λ_1, λ_2

11.4 Transform a PWM into log likelihoods

- to score a single site s for motif M , we use $\Pr(s|M)$, which is the PWM
- easy to calculate the probability for any seq. and motif
- to take bg freq into acct, divide by bg freq, and log to make it additive

$$\log_2 \frac{\Pr(s|M)}{\Pr(s|M_{bg})}$$

A. TATA
B. CAGG

Solution

A. The scores only differ for the last 2 positions. We have $\log_2(\frac{0.8}{0.4}) + \log_2(\frac{0.6}{0.4}) = \log_2 3.5$ for A and $\log_2(\frac{0.05}{0.1}) + \log_2(\frac{0.25}{0.1}) = \log_2 3$ for B and thus A has larger score.

- To include background probability:

• score = sum of the corresponding entries in the PWM for each position and each letter

11.5 Comparative genomics - PhyloCon

- Align two profiles
- each col \rightarrow base count $(n_A, n_C, n_G, n_T) \rightarrow (f_A, f_C, f_G, f_T)$
- log-likelihood ratio = $\sum_{\text{base } b} n_{bj} \ln(f_{bi}/p_b)$ measures likelihood of observed data from col J being generated by dist. estimated by column i
 * not all regulatory elements are conserved

12 Haplotype inference

12.1 Genetic polymorphism

- SNPs
- structural variation
 - copy number variation
- minor allele frequency: freq of less frequent allele

12.2 SNP genotyping

- SNP array:
 - pro: cheaper
 - con: incomplete set of genetic variation
- WGS
 - complete set of genetic variation
 - more expensive

12.3 Haplotype inference

- Genotype → haplotype
- from only knowing AA or Aa to knowing which chromosome it comes from → Phasing
- recombination breaks haplotype over time

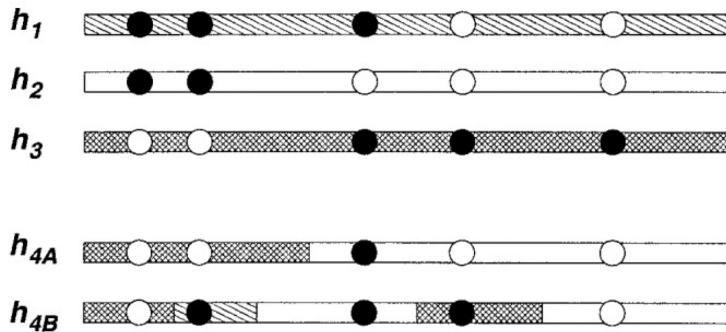
12.3.1 Probabilistic model for haplotype inference

- Construct an HMM $\Pr(\text{haplotypes} \mid \text{state}, \text{genotypes} \mid \text{observations})$ that models the observed genotype sequences as observed, and models the unknown haplotypes as unobserved variables
- Haplotypes can be inferred as haplotypes that maximize $p(\text{haplotypes} \mid \text{genotypes})$: Viterbi algorithm!
- transition prob: models recombination events
- emission prob: generate genotype, sometimes models mutation

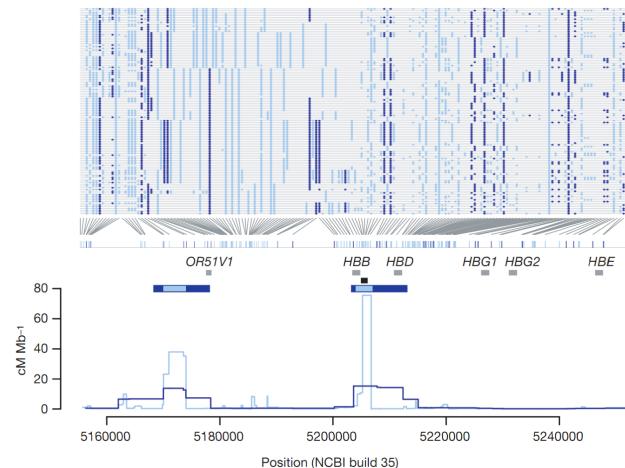
12.4 Phase

- parameter estimation in PHASE
 - Gibbs sampling
 - EM
- Genotype imputation
 - Given: SNP array genotype data
 - some has missing data / untyped SNPs
 - much cheaper
 - use HMM
 - more data, more accurate for imputation
- Beagle / SHAPE-IT are most commonly used software

- h_1, h_2, h_3 : unobserved ancestral haplotypes
- h_{4A}, h_{4B} : unobserved haplotypes for individuals
- Circles: alleles, mutations

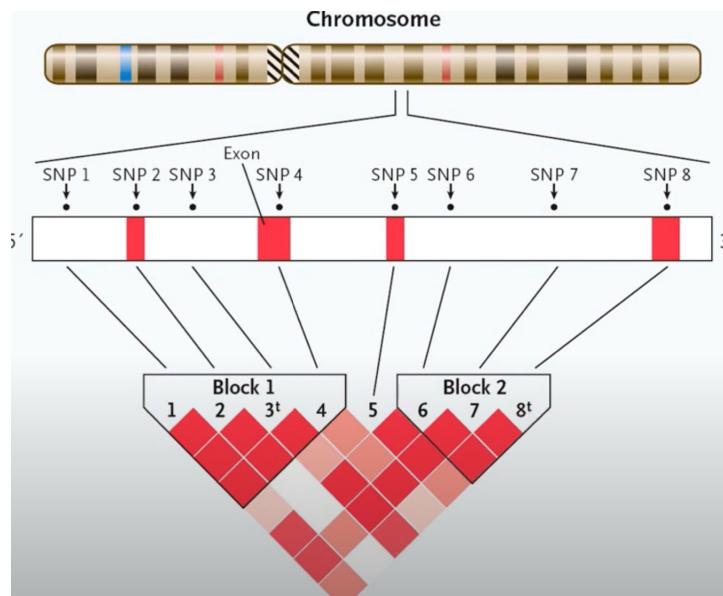


Haplotype Structure and Recombination Rate Estimates: HapMap I vs. HapMap II



12.5 Linkage disequilibrium (LD)

- reflect rs between alleles at diff loci
 - linkage equilibrium: no linkage, not coupled
 - disequilibrium → linkage
- LD: allelic association measure
- calculating LD:
 - assume independence, calculate expected frequency
 - $D_{AB} = P_{AB} - P_A P_B$, where A, B are SNPs, and this is calculated for one pair of alleles, eg. $P_{TG} - P_T P_G$
- calculate for all loci (SNPs)
- spot-light figures
- neighboring loci shows more linkage
there could be LP blocks on the chromosome, where there's stronger linkage within block and weaker across block
each block is like a voting district / Marker regions



- They are **Tag SNPs**
- pretty standard practice
- dense genotyping more expensive
 - get the LP
 - use inference on sparser genotyping
 - multi-phase procedure

13 Population structure

13.1 Background

- def: set of individuals with distinct genetic variations
- eg. ancestral history, lactose intolerance
- Hardy-Weinberg equilibrium
 - Under random mating, both allele and genotype frequency remain const
 - Current gen:
 - * $D + H + R = 1$
 - * $p = \frac{D+H}{2}$
 - * $q = \frac{R+H}{2}$
 - Next gen:
 - * $D' = p^2$
 - * $H' = 2pq$
 - * $R' = q^2$
 - * $p' = \frac{p^2+2pq}{2} = p^2 + pq = p$
 - * $q' = \frac{2pq+q^2}{2} = q^2 + pq = q$
 - given population data, can test if it holds (often chi-square test)
 - * testing is recommended
 - * if fail, means something is going on
 - * or genotyping error
 - * want HWE to hold for control group
 1. compute allele freq from observed data
 2. compute expected genotype freq
 3. compute test statistic (deg of freedom 1)
 - Due to **Genetic drift**, even when assumptions of HWE hold, HWE may not hold
- genetic drift
 - change in allele freq due to random sampling
 - all mutations eventually drift to 0 or 1 eventually
 - is neutral
- Wright's F_{ST} (Wright-Fisher model)

$$\binom{2N}{k} p^k q^{2N-k}$$
- Ways how populations evolve
 - population divergence
separated into subpopulations with independent selection and drift
 - admixture
mixing of population

13.2 Inferring pop structure from genotype data (nowadays mostly just PCA)

- Mixture model

- cluster individual into K populations
- does not model admixture**
- cluster individual into populations
- probability model for mixture of C Gaussians

$$p(x) = \sum_{i=1}^k p(x|c=i) \cdot p(c=i)$$

– c : labels, $p(c)$ label freq, multinomial

– x : genotype / alleles,

$$p(x|c) = \prod_{i=1}^j p(x_i|c)$$

, assume independence

– can then learn the model with EM

– Inference: $p(c|x)$ to infer cluster label

- Admixture model

– modern species are mixture of ancestral populations

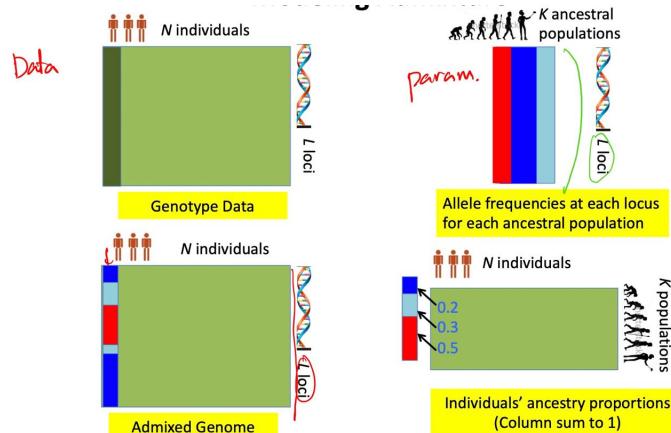
– genome consists of contributions from multiple ancestral populations, eg. Asian, African, Caucasians

– want to model it as a mosaic of variant

– Assumptions

* no linkage disequilibrium

* SNPs are iid



– for each individual $i : 1 \rightarrow n$:

* sample θ_i from $\text{Dirichlet}(\alpha)$

* for each loci $j : 1 \rightarrow L$

· sample $Z_{j,i}$ from multinomial θ_n

· sample $X_{j,i}$ from β_{k_j} for k chosen by $k = Z_{j,i}$

– still have to go back to β

– instead of sampling p_c of population it gets p_c for each individual

- Evaluations
 - probabilistic model
 - is generative process
 - explanatory, descriptive, and interpretable
 - computation can be very expensive
- PCA
 - fast, although not too much information
 - input: $N \times L$ for N individuals and L loci
 - good: easy visualization
 - bad: doesn't give intuition about what is going on, no allele frequency info

14 Linkage Analysis, GWAS(Genome Wide Association Study)

14.1 Background

- Genome polymorphisms
 - human genealogy → SNPs → haplotypes
 - useful markers for studying disease association
 - finding genetic markers that are likely to be linked with the disease locus, instead of finding the disease locus itself (cuz it's hard)
 - making use of linkage (dis)equilibrium, find those loci with $r < 0.05$
- linkage analysis
 - **family data**
 - more likely to have linkage data
 - Effective for rare diseases
 - Low resolution on the genomes (only a few recombinations)
 - **Parametric linkage analysis:**
 - * need to specify disease model
 - * Highly effective for Mendelian disease caused by a single locus
 - * usually large pedigree
 - * Founder probabilities (founders: parents not in population)
 - thus need to assign distributions to their genotypes
 - often done with Hardy-Weinburg
 - genotype of the founder couple assumed independence
 - * children get their genes according to **Mendel's law**
 - from genotype → phenotype
 - * complete vs incomplete penetrance

$$\Pr[\text{complete} | DD] = 1$$

$$\Pr[\text{incomplete} | DD] < 1$$

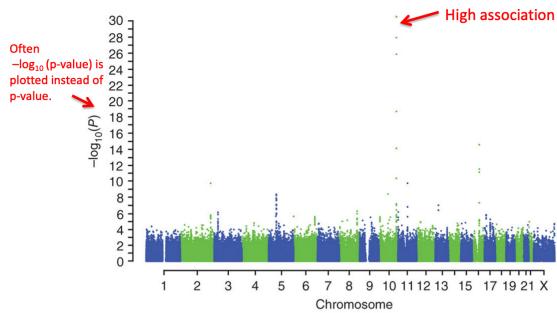
* Probability of prodigee = product of the prob. of the genotype and phenotype of all individuals

$$(2 \times .01 \times .99 \times .7) \times (2 \times .01 \times .99 \times .3) \times (1/2 \times 1/2 \times .9) \times (2 \times 1/2 \times 1/2 \times .7) \times (1/2 \times 1/2 \times .8)$$

14.2 GWAS

14.2.1 single SNP association analysis

- **unrelated individuals**
- Easier to find a large number of affected individuals
- Effective for common diseases
- Relatively high resolution for pinpointing the locus linked to the phenotype
- Genetic model for association:
 - additive models for minor alleles
- Correcting for multiple hypothesis testing
 - Bonferroni correction → overly conservative
 - Permutation test → computationally expensive



14.3 multimarker association test

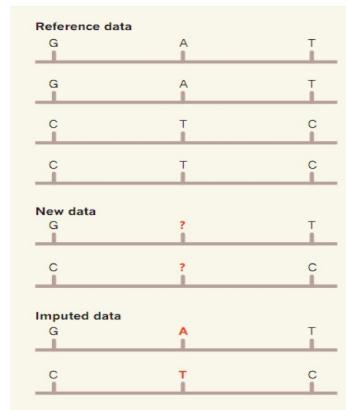
- a haplotype of multiple SNPs is a better proxy for a true causal SNP than a single SNP
 - Exploit the linkage disequilibrium structure in genome
- Form a new allele by combining multiple SNPs for a haplotype

SNP A	SNP B	Auxiliary Markers for Haplotypes
0	0	1 0 0 0
0	1	0 1 0 0
1	0	0 0 1 0
1	1	0 0 0 1

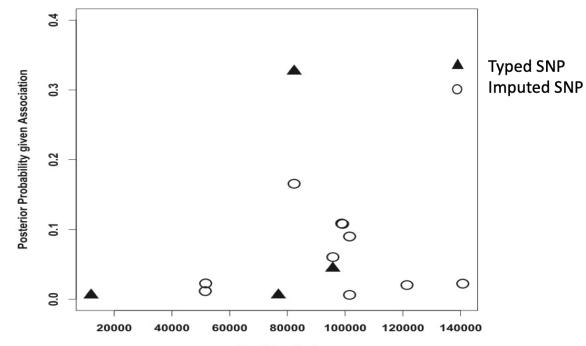
- test haplotype for association

14.4 using reference datasets for genotype imputation

- reference data: dense SNP data - from dataset eg 1000 genome project
- new data = sparser SNP

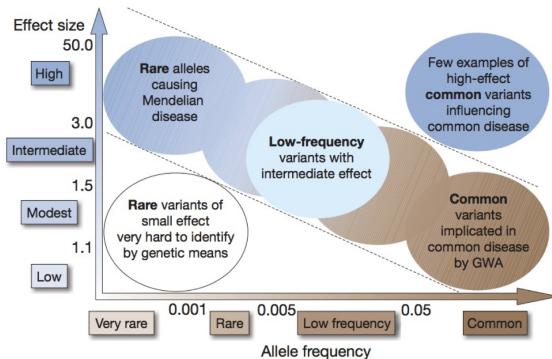


- leverage LD - data after imputation w ref data



- imputation based method - everyone is doing it rn

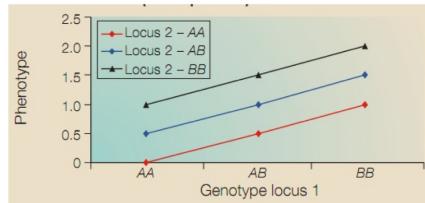
- association to rare variants (GWAS only mostly apply to common variants)
at least a lot more confident when allele is common
ppl often combine multiple rare alleles along a gene and compare gene
- underlying: common allele - common disease / common hypothesis – assumption
usually 5% considered “common”, 0.5% - 5% considered “low frequency variant”, else “rare variant”
- still quite a lot of variant to analyze



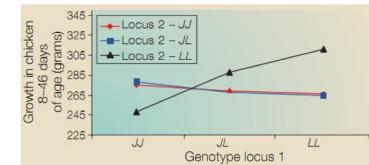
- feasibility of detecting disease loci

14.5 Epistasis

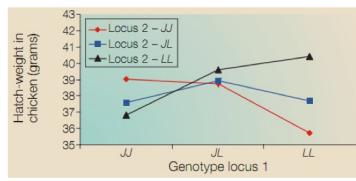
- effect of one locus masks another loci



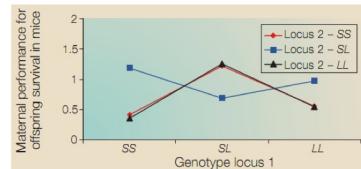
(a) no epistasis



- Dominant epistasis
- One locus in a dominant way suppresses the allelic effects of a second locus



- Co-adaptive epistasis
- Genotypes that are homozygous for alleles of the two loci that originate from the same line (JJ with JJ, or LL with LL) show enhanced performance.
- Almost no marginal effects: average effect of JJ, JL, LL do not differ



- Dominance-by-dominance epistasis
- Double heterozygote (LS, LS) deviates from the phenotype that is expected from the phenotypes of the other heterozygotes.
- Double heterozygotes have a lower phenotype than expected.

(b) epistasis

- hard to detect - only if the interacting SNPs are considered jointly
- also suffer from multiple testing problem

14.5.1 Detecting epistasis

- Epistatic effects of SNPs can often be detected only if the interacting SNPs are considered jointly – The number of candidate SNP interactions is very large
- For J SNPs, $J \times J$ SNP pairs need to be considered for epistasis
- In general for J SNPs and K-way interactions, there are $O(JK)$ candidate interactions • Computationally expensive to consider all possible groups of interacting SNPs
- For a reliable detection of K-way interactions, a large sample size is required
 - Multiple testing problem

14.6 Population structure and association analysis

- population structure can cause false positives
 - samples in case are often more related
 - Any SNPs more prevalent in the case population will be found significantly with the case
 - may capture both disease and population related SNPs
 - ideally have equal proportion of population in case and control
- Family-based studies
 - The effect of population structure can be controlled by the use of parents' genotypes
 - In practice, collecting genotypes from multiple individuals in a family can be hard.(e.g., late-onset diseases)
 - more for rare, family specific traits
- Population-based design
 - Data collection is easier for a large number of unrelated individuals than families.
 - The control samples can be reused in different studies.
 - for general, common traits

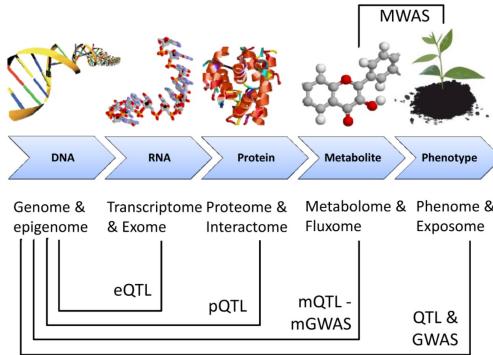
14.7 computationally solve the problem

- population based method
- **Eigenstrat:** PCA based method
 1. inferring ancestry - PCA applied
infer continuous axis
 2. removing ancestry effects
genotype at candidate SNP adjusted by amounts attributable to ancestry along each axis
scaling factor: regression on genotype - ancestral info
[missing eqns from notes](#)
 3. association tests
- linear mixed model

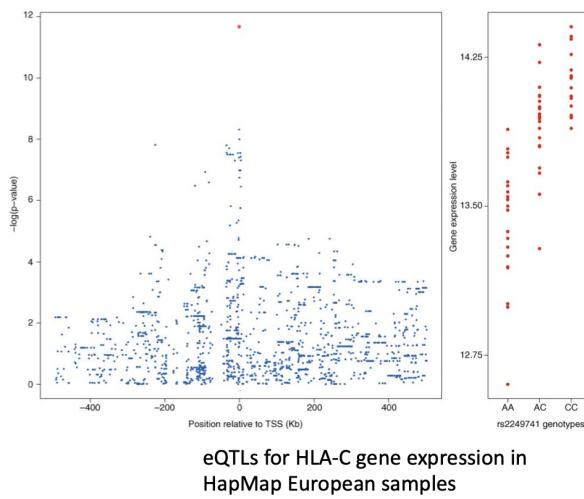
15 eQTL (expression quantitative trait locus)

15.1 limitation of linkage analysis and GWAS

- we often know the genetic loci
- but don't know the molecular mechanism



- find connection between DNA and mRNA
- eQTL mapping

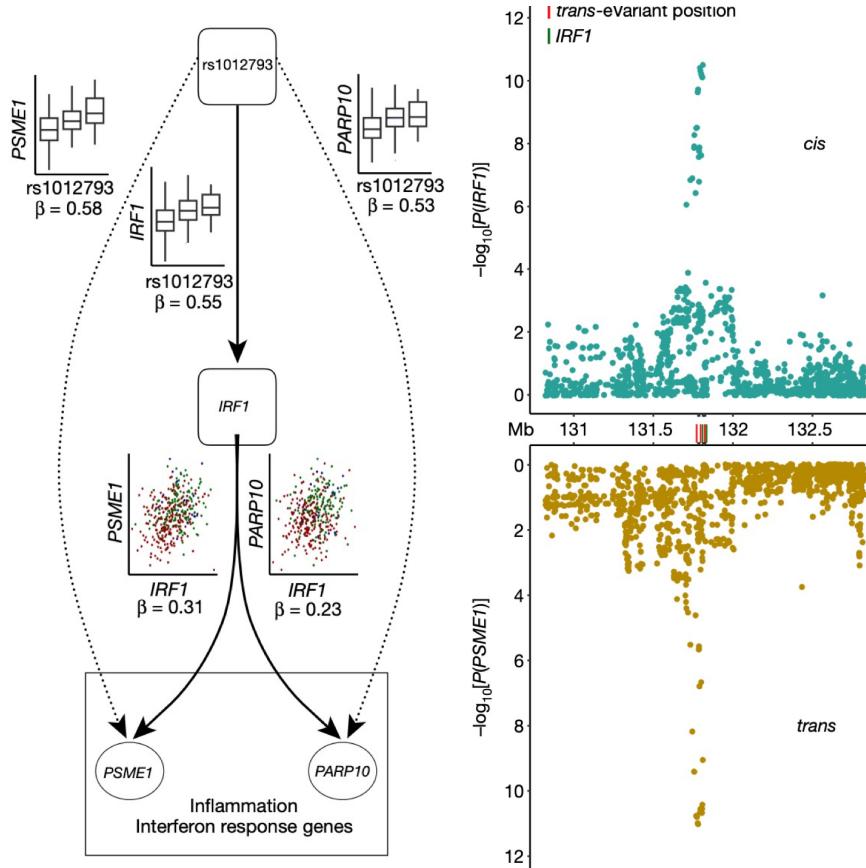


- genotype tissue expression (GTEx) project
 - goal: characterize molecular function of human genome
 - postmortem samples from normal, non-disease tissues - reference data
- popular tool to study genetic basis of expression for
 - diff tissue types
 - diff diseases

15.2 terminologies

- **eGene** genes whose expression is affected by eQTLs

- **cis eQTL** in the genome, the eQTL is located **near** the eGene
 - E.g., mutations in the promoter region of a gene influence the expression level of the gene
 - but enhancers can be far away, may look like trans
- **trans eQTL** eQTL is located **far away** (or on a different chromosome)
 - E.g., mutations in the transcription factor gene
 - * cis regulatory can have downstream effect → trans



- eg. From GTEx thyroid expression levels
 - SNP rs1012793 affects expression of IRF1 in cis and PSME1 and PARP10 in trans

15.3 human vs model organism

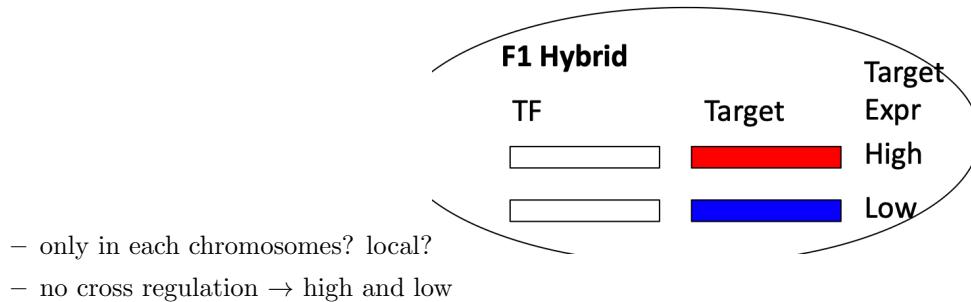
- human harder to collect data
- model organisms can be cultured in a lab
- Yeast recombinant inbred lines, instead of population of unrelated individuals
 - two founder strains, BY and RM
 - Mate the founders in a lab for generations
 - Genotype/expression profile for 110 progenies
 - In a follow-up study
 - The same founders, 1000 progenies, whole-genome sequencing, RNA-seq
- effect of recombination
 - shuffle genome through mating
 - More generations of mating means more shuffling through recombinations and higher resolution for eQTL mapping

15.4 allele-specific eQTL mapping

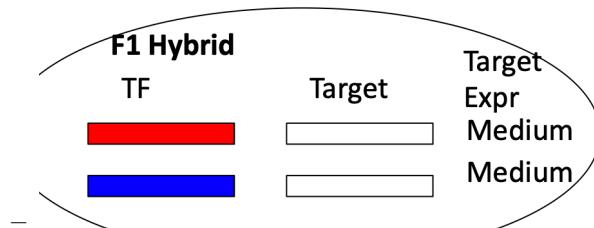
- cis-acting elements - DNA sequences in the regulatory region of the gene(e.g.,TF binding sites)
- trans-acting elements - RNA and proteins that interact w cis
- problem:
 - assumed cis = local and trans = distant, but not the case
 - exists nearby trans and distant cis
- fix: **allele-specific expression quantification w transcriptome RNA-seq**
 - diploid: maternal / paternal allele diff, RNA-seq can capture this
- idea: between two species
 - Studied cis/trans regulatory divergence between two species – First study that makes use of allele-specific gene expressions to learn about gene regulatory network – Examines genome / transcriptome data for two parent species and their F1 progeny
- idea: extension to population genome / transcriptome data

15.5 allele specific eQTL mapping

- cis-trans regulatory divergence in two species
- cis / trans, but location not very helpful
- Only cis-regulatory divergence between two species



- trans-regulatory divergence:



- both are medium → got averaged out
 – since there are trans-regulation
 – one transcription factor is less efficient than the other

- ratio between the high and low in parents vs. ratio in F1 progeny
- ratio = 1 → trans-regulatory divergence
- $y = x \rightarrow$ cis-regulatory divergence

- most genes are influenced by a mix of the two
- regulatory genes tend to be trans-acting
- structural genes: Usually terminal nodes with no trans- acting effects

15.6 limitations

- only examine 2 species and offspring (sample size too small)
- can do on population
- Population-based study: allele-specific eQTL mapping
- Cis-acting eQTLs = cis-regulatory variation
 - linear model wrt alleles (eg, A, C, T, G)
- Trans-acting eQTLs = trans-regulatory variation
 - linear model wrt genotypes (eg, AA, Aa, CG)

16 scRNA-seq & CRISPR

16.1 Background

- experimental validation of eQTL w CRISPR perturbation
 - genome editing w CRISPR, followed by scRNA
- scRNA made possible
 - Cell-type specific eQTLs
 - Co-expression eQTLs

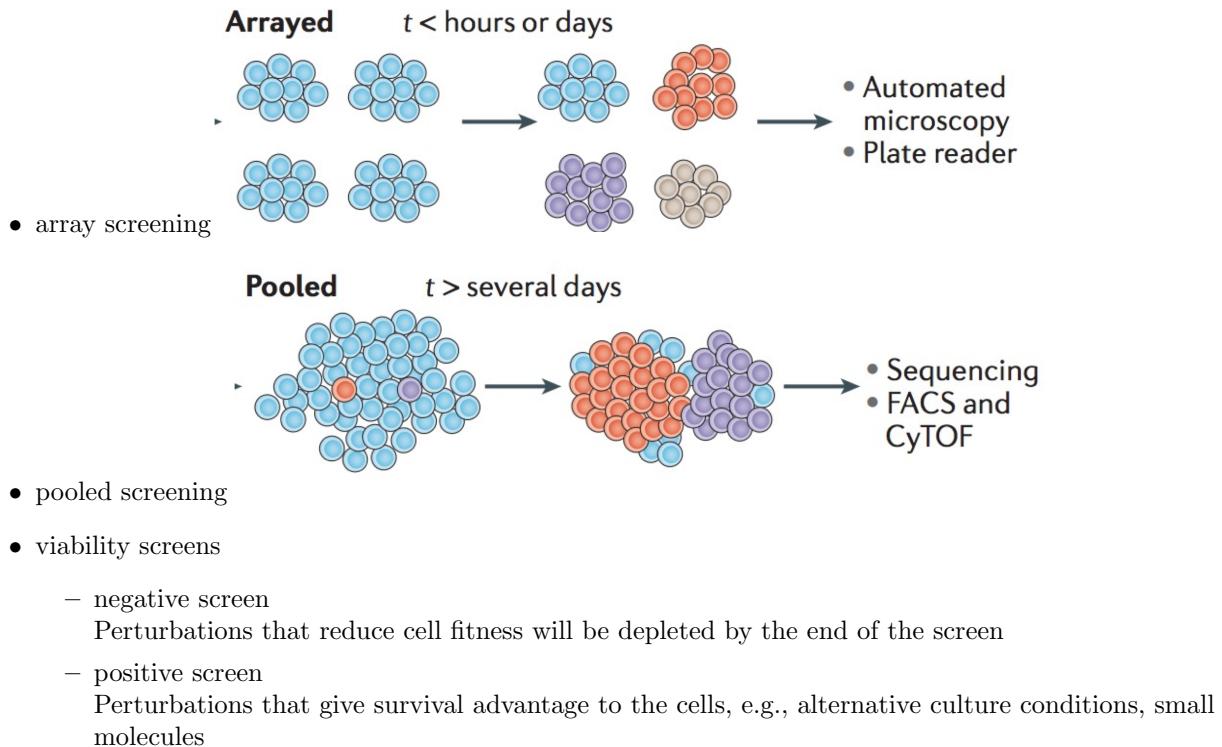
16.2 co-expression eQTLs

- one vector for expression and one vector for genotype
- coexpression eQTL: study eQTLs that influence “co-expression” between two genes
- → personal gene regulatory network

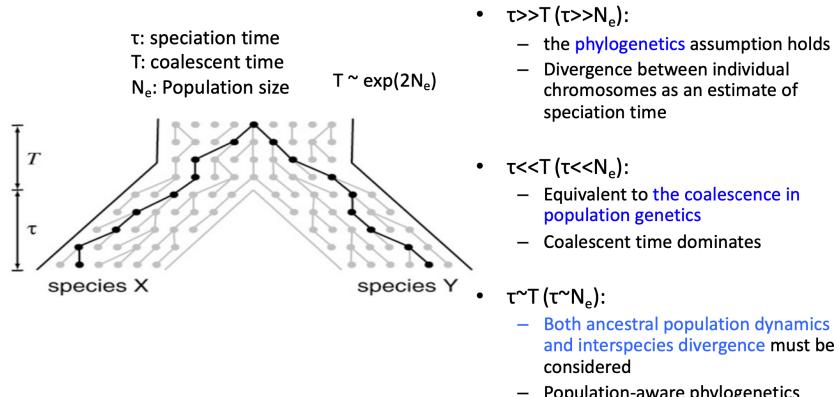
16.3 CRISPR

- genetic screening
- three components
 - Perturb:Knockout, RNAi,CRISPR
 - Model:primary cells,organoids,mice
 - Assay:measuring phenotypes of interest, RNA-seq,scRNA-seq
- Cas9 molecule
 - a component of the type II CRISPR bacterial adaptive immune system
 - DNA double-strand break(DSB) by Cas9 is repaired by the endogenous DNA DSB repair pathways
- sgRNA (single guide RNA)
 - guide Cas9 to the correct place
 - 20bp complimentary to the target DNA
- CRISPR knockout
- CRISPR interference
 - repress gene of interest
- CRISPR activation
 - overexpress the gene of interest

16.4 array vs pooled screening



17 Comparative genomics



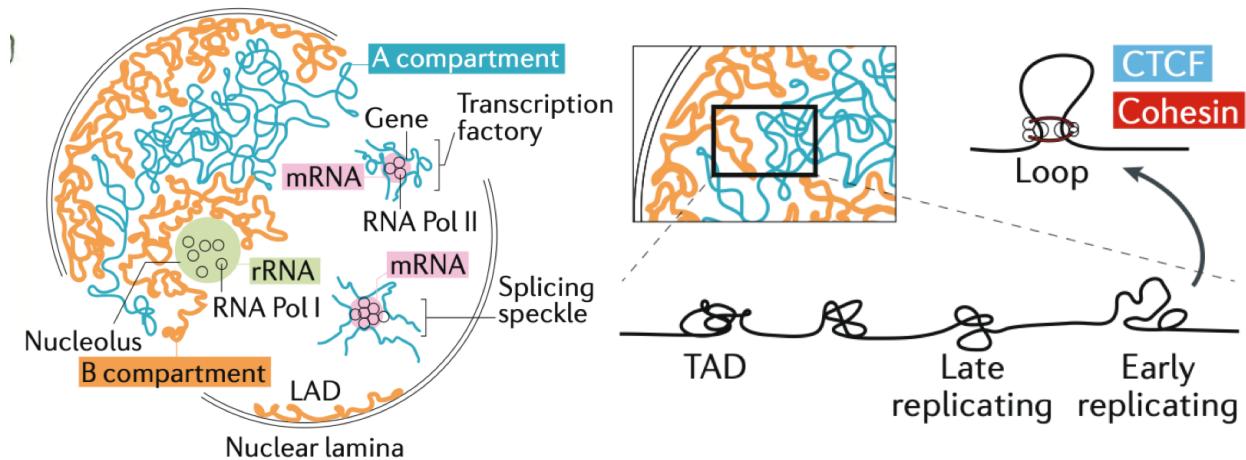
18 Epigenetics

19 3D genomes

19.1 Genome organization

- genome is not linear (ofc)
- nucleosomes → DNA packaging
- higher order genome organization

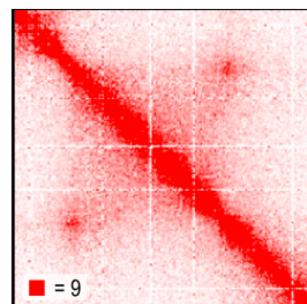
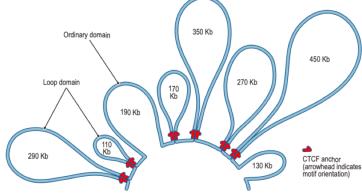
19.2 nuclear organization



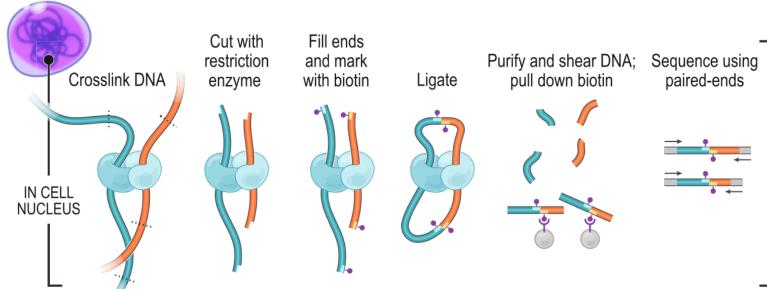
- A compartment: blue region
- B compartment: orange region

19.3 Hi-C

- DNA-FISH (fluorescence in situ hybridization)
 - 2 probes w diff colors → measure spatial dist
- 3C and its derivatives
 - 3C: one vs one
 - 4C: one vs all
 - 5C: all vs all
 - Hi-C: genome wide all vs all
- Contact maps
 - Chromatin structure with loops
 - Hi-C contact map



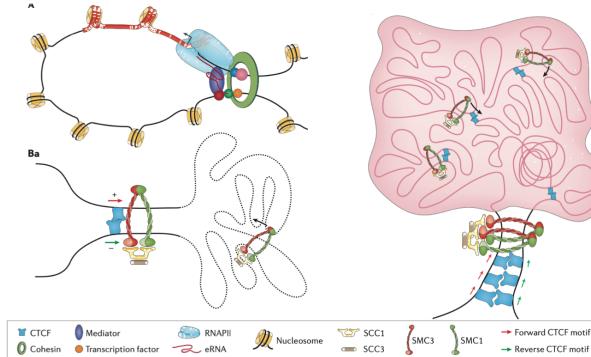
- take genome, bin it by length
- rows and col s represent non-overlapping bins across genome
- each entry = count of read pairs that connect corresponding bin pair
- bin size: no standard rule



19.4 3D genome in different resolutions

- cell cycle and single cell

19.5 what establishes chromatin interaction



- In a TAD, multiple loop-extrusion complexes are dynamically producing new loops and multiple such complexes are halted at the TAD boundaries by closely spaced **CTCF proteins**.

- what if some key players are perturbed?

- perturbation with CRISPR etc

19.6 Computational methods

19.6.1 Normalization

To correct bias in the contact map

- **Explicit approach** – assume some known bias
- **Implicit approach** – assume no known bias and each locus receives equal coverage
- **Vanilla coverage (VC) approach**
 - something like devide by row sum and col sum
 - **Doubly stochastic matrix** - each row sum to 1, each col sum to 1
 - normalization as matrix balancing problem: turn a matrix into doubly stochastic

- idea: make the diagonals even

$$O_{ij} = B_i B_j T_{ij}$$

$$\sum_{i=1, j \neq i}^N T_{ij} = 1$$

- **Directionality index** – identify TADs

- At TAD boundary – highly imbalanced
- comparable upstream and downstream within TADs
- use HMM to find TADs
 - * hidden states: true directionality biases
- A compartment – more open stuff
- ix