Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

High-dimensional genomics assays & data

Keegan Korthauer

12 January 2022

with slide contributions from Paul Pavlidis

Last time: CHD8 & ASD

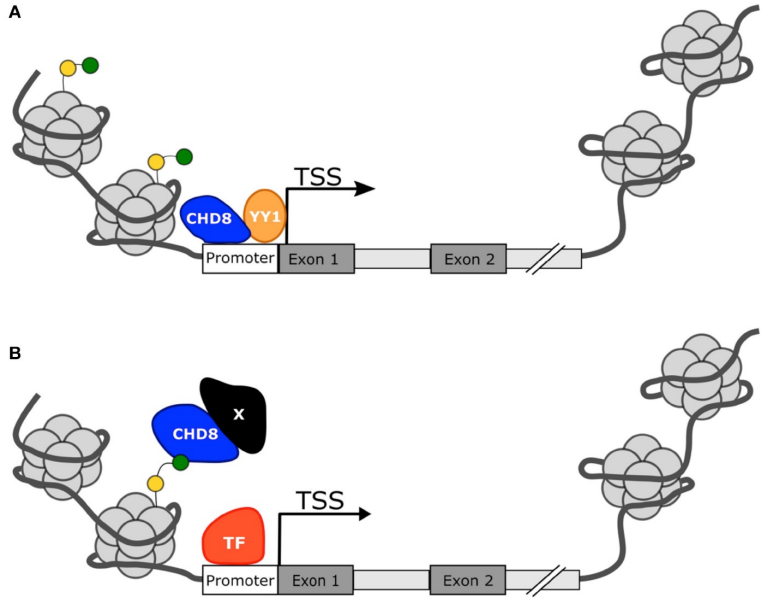
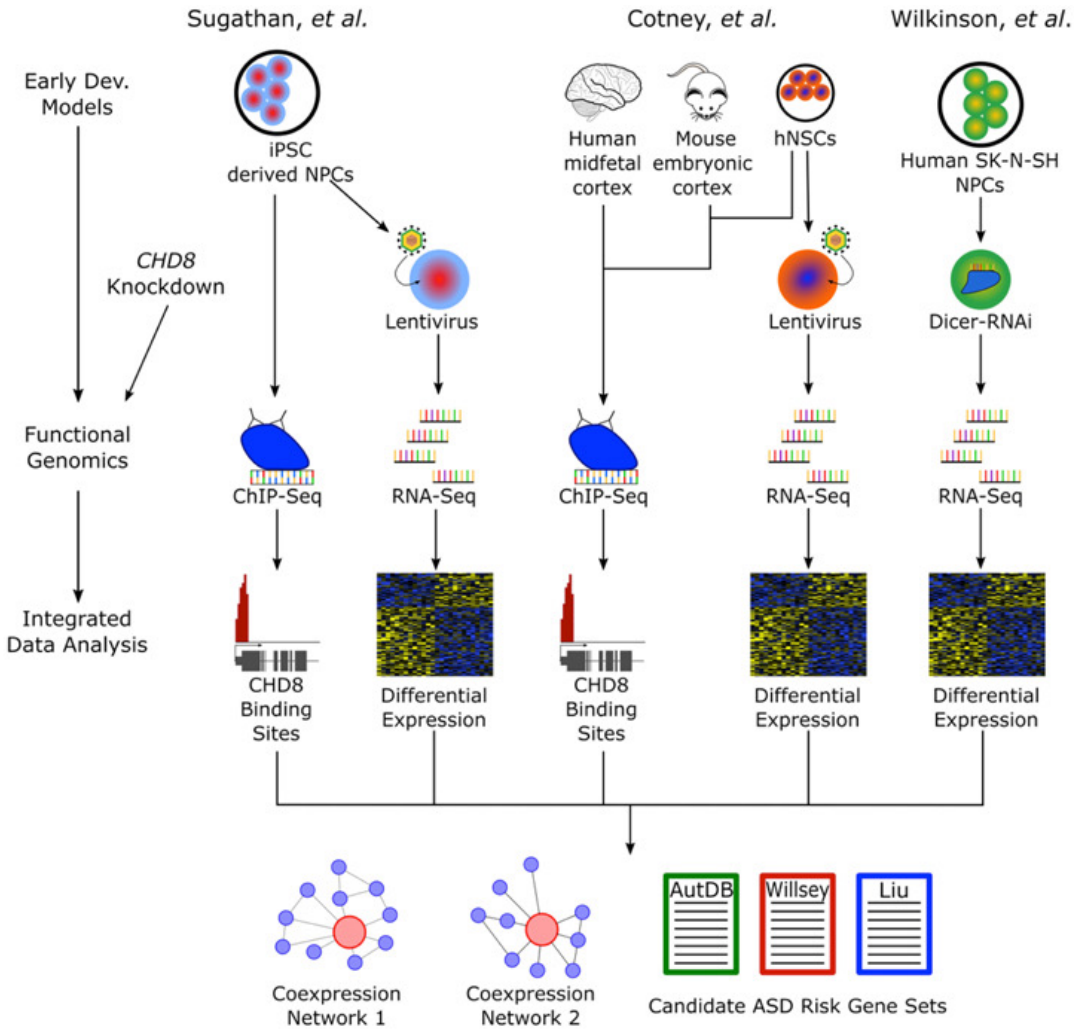


FIGURE 3 | Proposed mechanisms for CHD8 transcriptional activation. (A) CHD8 is most commonly found near active transcription start sites with histone modifications H3K4me3 (green circle) or H3K27ac (yellow circle). CHD8 may directly activate genes by directly binding near the transcriptional start site and promote transcription factor activity or recruitment. **(B)** CHD8 may indirectly activate genes through interactions between modified histone sites and other co-regulators to make chromatin more assessable.

https://www.frontiersin.org/articles/10.3389/fnins.2015.00477/full

- Found several **very rare** SNVs in CHD8 linked to ASD
  - <0.5% individuals with ASD
- Now one of best-established ASD-associated gene
- Hypothesis: CHD8 regulates transcription by binding to DNA and/or binding to other proteins which are bound to the DNA

What genes does CHD8 regulate, and what happens when it is broken?



https://www.frontiersin.org/articles/10.3389/fnins.2015.00477/full

Figure is from a 2015 review, summarizing three (high-dimensional) studies:

- Look for where CHD8 binds to DNA genome-wide (ChIP-seq)
  - Thousands of binding sites, much in/near promoters
- Decrease CHD8 in cells/animals, see what happens to all genes (RNA-seq, etc.)
  - Thousands of RNAs go up or down in levels
- What changes are 'functional'? Some of the "regulatees" are other ASD-associated genes

Learning Objectives

Part 1: What is high-dimensional data?

- Motivation to collect "high-dimensional data"
- Basic mechanics of **sequencing-based** and **microarray-based** assays

Part II: A few common applications

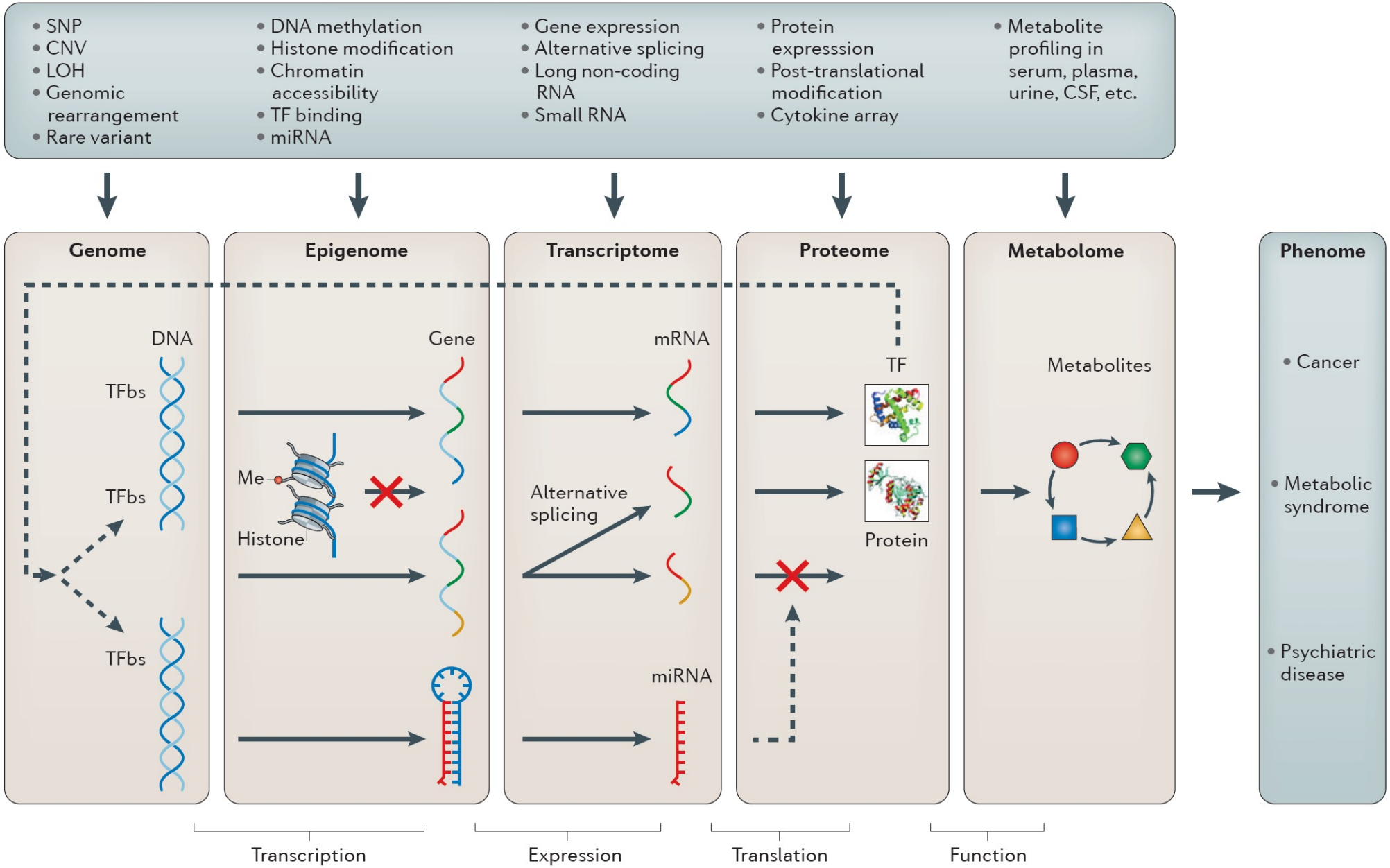
- Understand the main goal and general mechanism of each
  - DNA genotyping (SNPs)
  - DNA methylation (bisulfite conversion + seq or arrays)
  - DNA-protein interactions (ChIP-seq, CUT&Tag)
  - RNA quantification (RNA-seq)

Collecting data the **low-dimensional** way

- Pick one variable (e.g. "Expression of Gene X") and study it under various conditions
- Usually "specific hypothesis-driven" ("We hypothesize CHD8 activates expression of Gene X")
- Use assays that address only that question; publish
- Repeat this for another variable
- Powerful, but knowledge accumulates slowly and synthesis is difficult
  - Scattered around the literature
  - Experiments rarely done the same way

Factors in the move toward "systems biology"

- Limitations of the "one thing at a time approach" – how do the parts work together?
- Technology enabling increasingly detailed analyses – measure many "things" at once in a single experiment ("High-throughput", "High-dimensional" or "High-content")
- Hypotheses/questions with a "non-specific" flavor: e.g. "What genes does CHD8 regulate?"
- Criticisms of the approach include
  - The experiments rarely give easy answers
  - Sometimes just generating "biomarkers" without gaining insight.
  - "Fishing expeditions"



Ritchie et al. 2015 doi:10.1038/nrg3868

Motivating genomics studies

Hypothetical situation:

- Patients with disease subtype A have poor prognosis, while subtype B is more treatable
- Telling A from B is difficult using "conventional" means
  - Cells look the same, etc. – we only find out by seeing what happens to the patients over time

Questions:

- Can we characterize the differences better? – "Biomarkers"?
- Can we find new targets for drugs or for diagnosis?
  - Drug targets are often proteins, encoded by genes
- Are there other subtypes that can be discovered? ("Clusters")

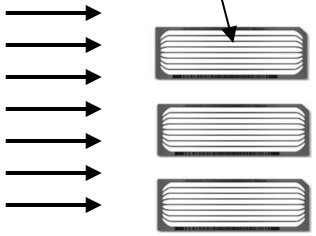
A prototype genomics experiment

One of these gives us measurements for every gene, in a sample

Each value =
Measurement of one gene one sample

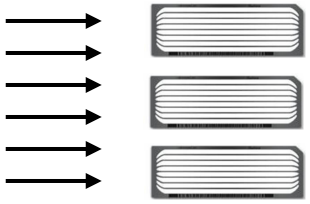
Samples

Condition A



Assays
(RNA-seq)

Condition B



One row per gene (thousands)

One column per sample

	Sample_ANAN001A	Sample_ANAN001B	Sample_ANAN001C	Sample_ANAN001D	Sample_ANAN001E	Sample_ANAN001F	Sample_ANAN001G	Sample_ANAN001H	Chd8.e14.S12	Chd8.e14.
Chaf1a	5.14026690	4.94132189	5.268524908	5.01515985	5.10076733	5.18254532	5.342874653	5.001467423	3.9772641	3.67448
Chaf1b	5.05766571	4.84498569	5.161571400	4.95561977	5.11963327	5.03959384	5.303529019	4.929144632	3.8600541	3.89531
Champ1	4.92706308	4.99119785	4.827593493	4.86168682	4.94216633	4.88877583	5.048075565	4.995995825	4.8214729	4.71430
Chchd1	5.69996648	5.88249076	5.909221367	6.01544790	6.05453151	6.06921009	5.815513688	5.881654420	6.1969070	6.71826
Chchd10	1.11700906	0.49317050	1.116364761	1.27092205	1.12375825	1.00988912	0.682043996	0.986449184	2.8404661	3.53580
Chchd2	6.78273388	6.67454533	6.732675821	6.83279072	6.66261581	6.67311273	6.575583851	6.577197792	6.6420195	7.08060
Chchd3	5.61699744	5.58864447	5.907230366	5.84502976	5.72150173	5.87171081	5.773155752	5.735732695	5.6340849	5.84485
Chchd4	4.10286451	4.07586682	4.172176698	4.20916334	4.22867557	4.08707810	4.139775324	4.388060374	4.2502683	4.80832
Chchd5	3.54432563	3.45838524	3.518695206	3.60317244	3.49872809	3.46075772	3.437265256	3.582862373	3.1988098	3.56479
Chchd6	4.57081930	4.49929334	4.774061716	4.73109677	4.68342709	4.68180195	4.537803502	4.685548893	5.0209376	5.97639
Chchd7	2.23767729	2.01107578	2.437207583	2.39914645	2.34875903	2.25555800	2.175308145	2.016273058	2.4733083	2.40635
Chd1	4.01045860	4.13387024	4.012145049	4.00767747	3.99458108	4.17295236	4.254927644	4.085858447	3.4109000	2.69280
Chd1l	2.73363537	2.56222676	2.986071071	2.58070453	2.68835020	2.74507339	2.931090880	2.709829062	1.6415284	1.66515
Chd2	3.76274816	3.96230946	3.584988345	3.62793189	3.75557716	3.82964474	3.909333621	3.787623558	3.9513823	3.26288
Chd3	6.36257941	6.68590320	6.259273949	6.06500151	6.25379415	6.33096694	6.489224423	6.457524071	6.9395850	6.83240
Chd3os	4.13677912	4.32714634	4.263762955	4.48615428	4.46227526	4.29868779	4.314635974	4.621897735	5.5944277	6.01407
Chd4	6.83257292	7.10301114	6.758219332	6.64615706	6.78564424	6.83088917	6.966181279	6.946714128	6.5825751	5.98749
Chd5	1.77163980	2.46621236	0.989334271	1.21640480	1.78227341	1.77048444	1.893826496	1.943095972	3.9584141	4.01819
Chd6	3.84791505	4.18495272	3.501762770	3.52597323	3.79659225	3.74093596	3.761387007	3.777791958	4.1765415	3.44387
Chd7	4.42114801	4.63795328	4.072470139	4.22240724	4.44294712	4.32464976	4.275048783	4.393796974	3.7707527	2.92205
Chd8	4.18812046	4.95888622	4.435364587	3.89659418	4.12444302	4.65387109	4.142757199	4.748590846	4.6852528	4.18897
Chd9	3.03756558	3.29237429	2.686909319	2.95489741	3.06326253	3.04111376	3.044780033	3.049053139	3.4462653	2.97209
Chek1	4.00909974	3.94056593	4.159375537	3.95076578	4.20545370	4.13196543	4.237381171	3.931389171	2.5700125	2.17708
Chek2	3.06760070	2.99850687	3.251924401	3.01167477	3.23215290	3.16039796	3.344098120	2.894302116	2.1716760	1.73388
Cherp	5.59225606	5.59869607	5.266960310	5.20757383	5.40443195	5.37405243	5.596488885	5.526052741	4.9632305	4.49207

...

This happens to be a transcriptome data set, but a lot of data sets we encounter can be put into this kind of matrix representation. Meaning of rows and values will vary.

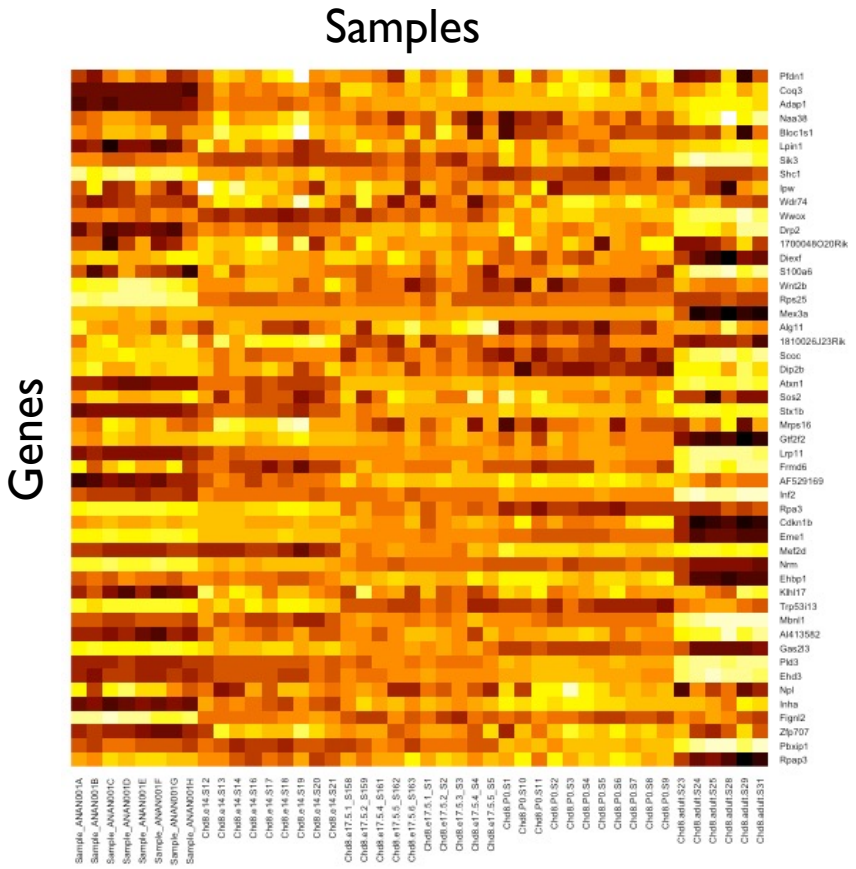
Alternative representation as colours

Samples

Genes

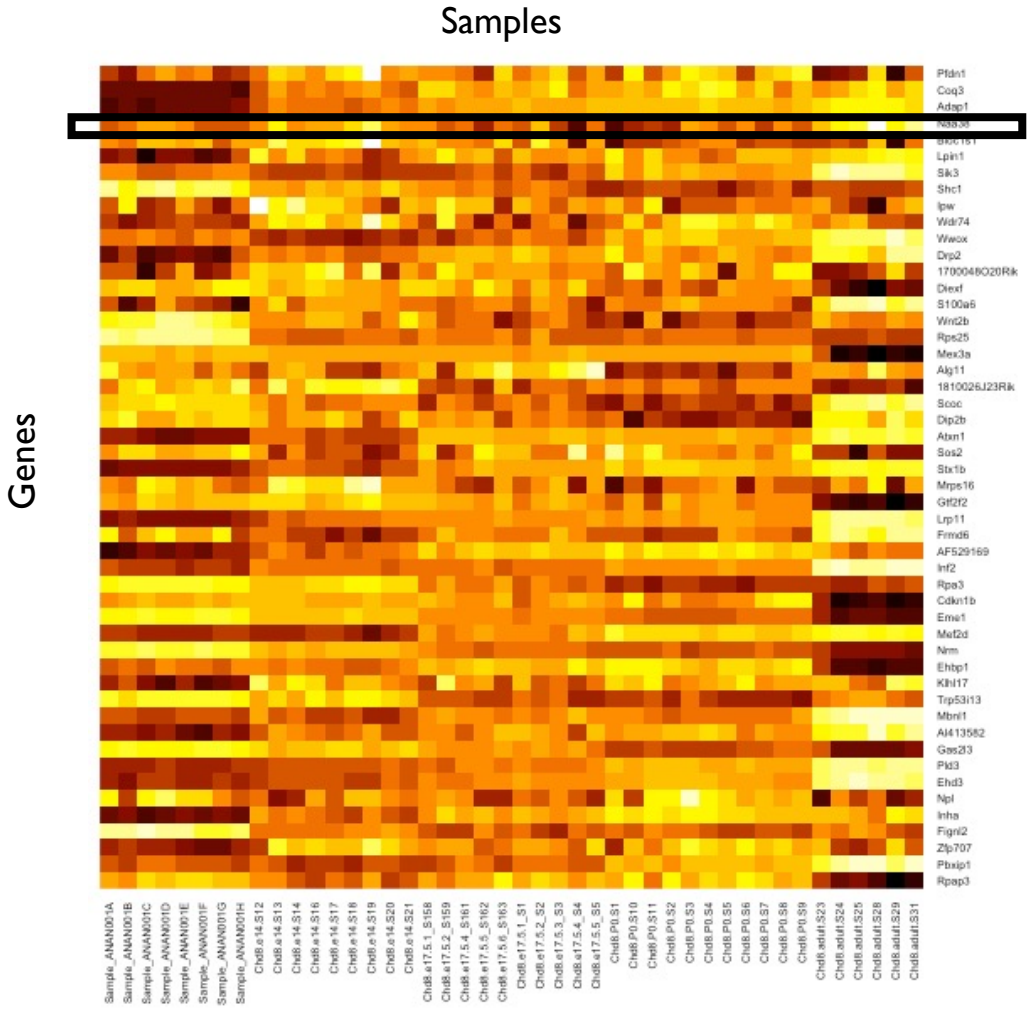
	Sample_ANAN001A	Sample_ANAN001B	Sample_ANAN001C	Sample_ANAN001D	Sample_ANAN001E	Sample_ANAN001F	Sample_ANAN001G	Sample_ANAN001H	Chd8.e14.S12	Chd8.e14.
Chaf1a	5.14026690	4.94132189	5.268524908	5.01515985	5.10076733	5.16254532	5.342874653	5.001467423	3.9772641	3.67448
Chaf1b	5.05766571	4.84498569	5.161571400	4.95561977	5.11963327	5.03959384	5.303529019	4.929144632	3.8600541	3.89531
Champ1	4.92706308	4.99119785	4.827593493	4.86168682	4.94216633	4.88877583	5.048075565	4.995995825	4.8214729	4.71430
Chchd1	5.69996648	5.88249076	5.909221367	6.01544790	6.05453151	6.06921009	5.815513688	5.881654420	6.1969070	6.71826
Chchd10	1.11700906	0.49317050	1.116364761	1.27092205	1.12375825	1.00988912	0.682043996	0.986449184	2.8404661	3.53580
Chchd2	6.78273388	6.67454533	6.732675821	6.83279072	6.66261581	6.67311273	6.575583851	6.577197792	6.6420195	7.08060
Chchd3	5.61699744	5.58864447	5.907230366	5.84502976	5.72150173	5.87171081	5.773155752	5.735732695	5.6340849	5.84485
Chchd4	4.10286451	4.07586682	4.172176698	4.20916334	4.22867557	4.08707810	4.139775324	4.388060374	4.2502683	4.80832
Chchd5	3.54432563	3.45838524	3.518695206	3.60317244	3.49872809	3.46075772	3.437265256	3.582862373	3.1988098	3.56479
Chchd6	4.57081930	4.49929334	4.774061716	4.73109677	4.68342709	4.68180195	4.537803502	4.685548893	5.0209376	5.97639
Chchd7	2.23767729	2.01107578	2.437207583	2.39914645	2.34875903	2.25555800	2.175308145	2.016273058	2.4733083	2.40635
Chd1	4.01045860	4.13387024	4.012145049	4.00767747	3.99458108	4.17295236	4.254927644	4.085858447	3.4109000	2.69280
Chd1l	2.73363537	2.56222676	2.986071071	2.58070453	2.68835020	2.74507339	2.931090880	2.709829062	1.6415284	1.66515
Chd2	3.76274816	3.96230946	3.584988345	3.62793189	3.75557716	3.82964474	3.909333621	3.787623558	3.9513823	3.26288
Chd3	6.36257941	6.68590320	6.259273949	6.06500151	6.25379415	6.33096694	6.489224423	6.457524071	6.9395850	6.83240
Chd3os	4.13677912	4.32714634	4.263762955	4.48615428	4.46227526	4.29868779	4.314635974	4.621897735	5.5944277	6.01407
Chd4	6.83257292	7.10301114	6.758219332	6.64615706	6.78564424	6.83088917	6.966181279	6.946714128	6.5825751	5.98749
Chd5	1.77163980	2.46621236	0.989334271	1.21640480	1.78227341	1.77048444	1.893826496	1.943095972	3.9584141	4.01819
Chd6	3.84791505	4.18495272	3.501762770	3.52597323	3.79659225	3.74093596	3.761387007	3.777791958	4.1765415	3.44387
Chd7	4.42114801	4.63795328	4.072470139	4.22240724	4.44294712	4.32464976	4.275048783	4.393796974	3.7707527	2.92205
Chd8	4.18812046	4.95888622	4.435364587	3.89659418	4.12444302	4.65387109	4.142757199	4.748590846	4.6852528	4.18897
Chd9	3.03756558	3.29237429	2.686909319	2.95489741	3.06326253	3.04111376	3.044780033	3.049053139	3.4462653	2.97209
Chek1	4.00909974	3.94056593	4.159375537	3.95076578	4.20545370	4.13196543	4.237381171	3.931389171	2.5700125	2.17708
Chek2	3.06760070	2.99850687	3.251924401	3.01167477	3.23215290	3.16039796	3.344098120	2.894302116	2.1716760	1.73388
Cherp	5.59225606	5.59869607	5.266960310	5.20757383	5.40443195	5.37405243	5.596488885	5.526052741	4.9632305	4.49207

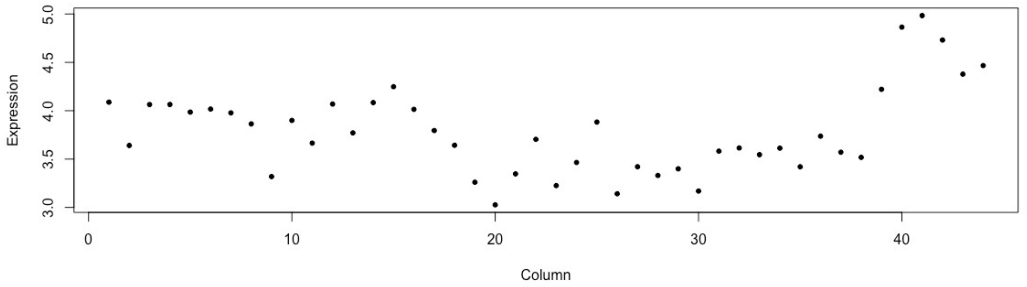
…



- Each square represents the expression level of one gene in one sample
- In this scheme, lighter colours mean higher levels of gene expression (“activity”)
- Only a portion of the data is shown

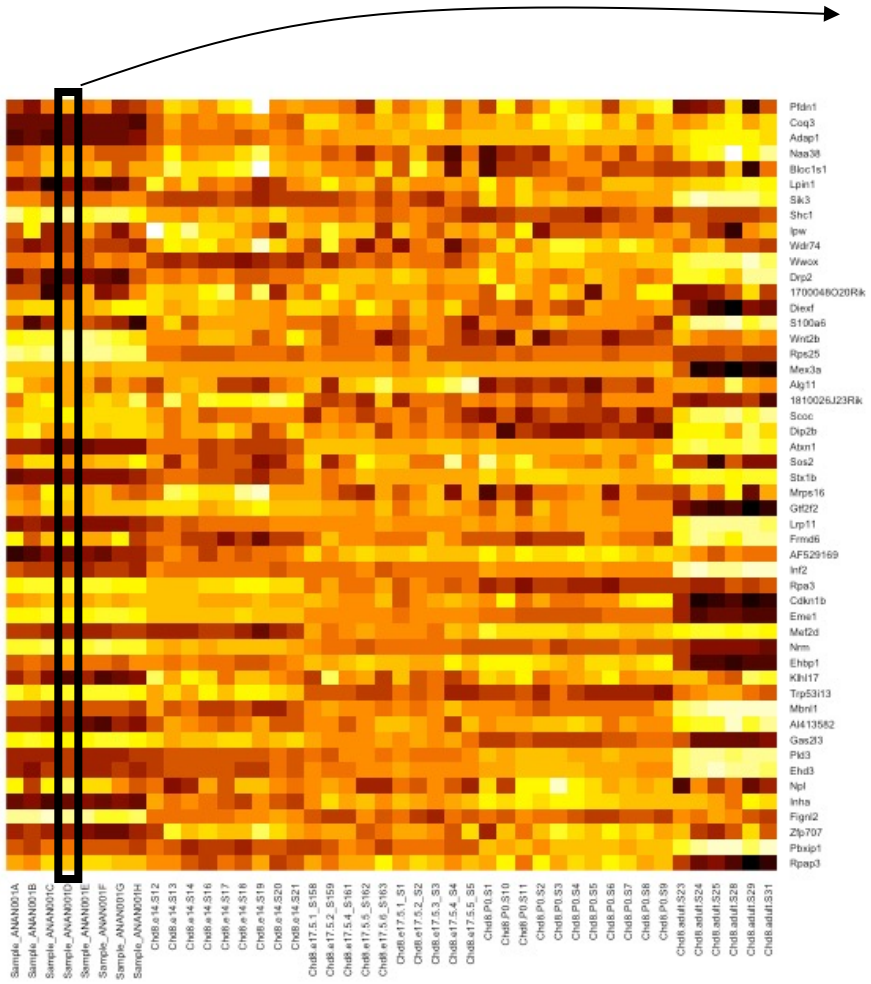
Profile for a single gene



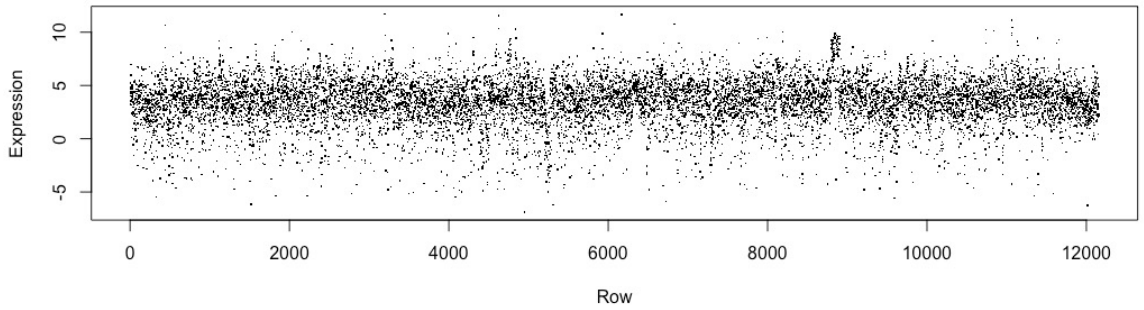


- The order of columns (samples) is not necessarily important – we can rearrange without affecting the data as long as we keep track of which sample is which
- Better yet – order or group by a meaningful factor/covariate

Profile for a sample



(for many, many more rows)



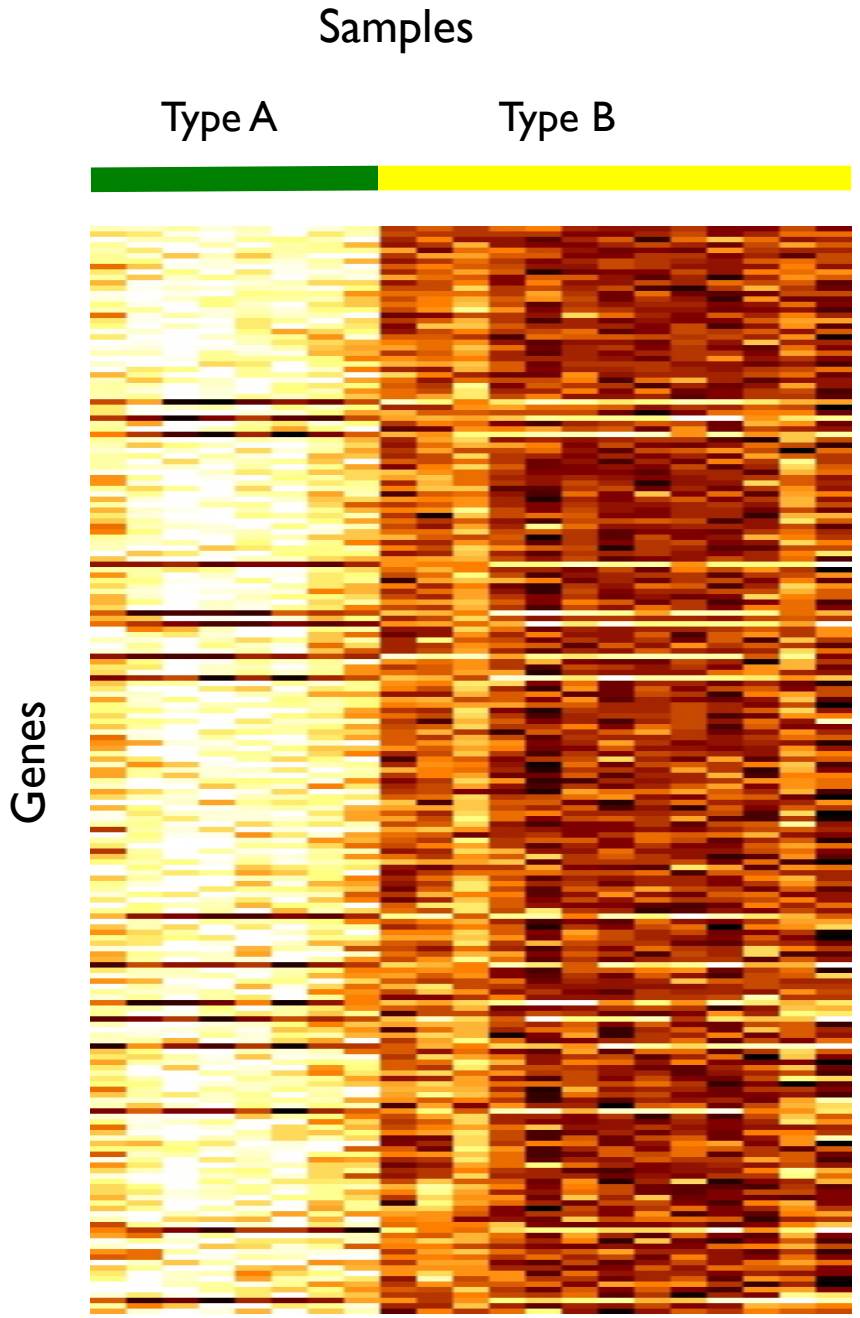
- (We usually wouldn't plot 12000 points like this)
- The "shape" (bumps, dips) of this isn't particularly meaningful – suggests the rows aren't in a completely random order in this example.
- Could rearrange the order of rows without changing the data (of course keeping track of which row is which)
- Could also order by genomic position (may be more meaningful for other data types with continuous signal e.g. copy number, methylation)

One type of analysis

- For this data set, I've ranked the genes by how different they are between types A and B (t-statistic P-value) - "Differential expression"
- Only the first few genes are shown (out of thousands)

- Though it can be a lot more complicated, a lot of "high-dimensional" studies boil down to something like this, at least in part

What's the big deal?



A few pitfalls and challenges
(why we need this course)

- Signals can be small relative to non-signal: data are **noisy** with finite sensitivity
- **False negatives** are often a given, and **false positives** are a major danger
- Need to address outliers, batch effects and other systematic **artifacts** (can dominate biological signal)
- Dealing with (and exploiting) biological and statistical dependencies – e.g. genes are not independent
- Getting just a list of "hits" isn't enough – can we understand something more about the "system"?
- Data sets (and questions) can be much more complex than these simple examples; perhaps most interestingly when you have multiple data types for the same samples (e.g. DNA sequence, DNA methylation and RNA levels)

High-dimensional assays

In this course most experiments we discuss involve one of two basic technologies (or both):

Microarrays for:

Single nucleotide polymorphism genotyping (SNP array)

DNA methylation quantification (methylation array)

Copy number variation (array CGH)

RNA quantification (expression array)

DNA-protein binding (ChIP-Chip)

Pause for Q1

Sequencing-based assays for:

DNA genotyping (targeted, exome, or whole genome seq)

DNA methylation quantification (Bisulfite-seq)

Copy number and structural variation (targeted, exome, or whole genome seq)

RNA quantification (RNA-seq, scRNA-seq)

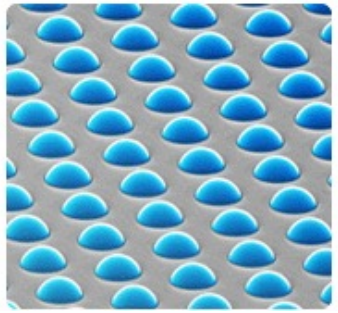
DNA-protein binding (ChIP-seq, CUT&Tag-seq)

- A common theme is that all of these technologies and assays are built on a deep understanding of, and ability to manipulate, nucleic acid chemistry
- This is NOT an exhaustive list

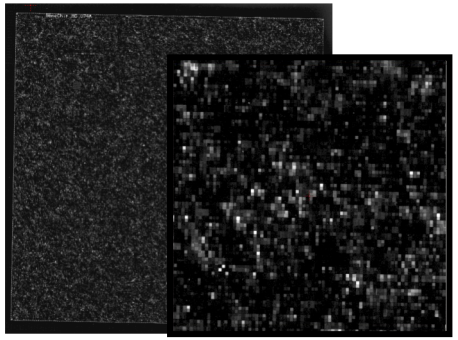
Microarray basics

Microarrays

Illumina Beadarray



Affymetrix Genechip



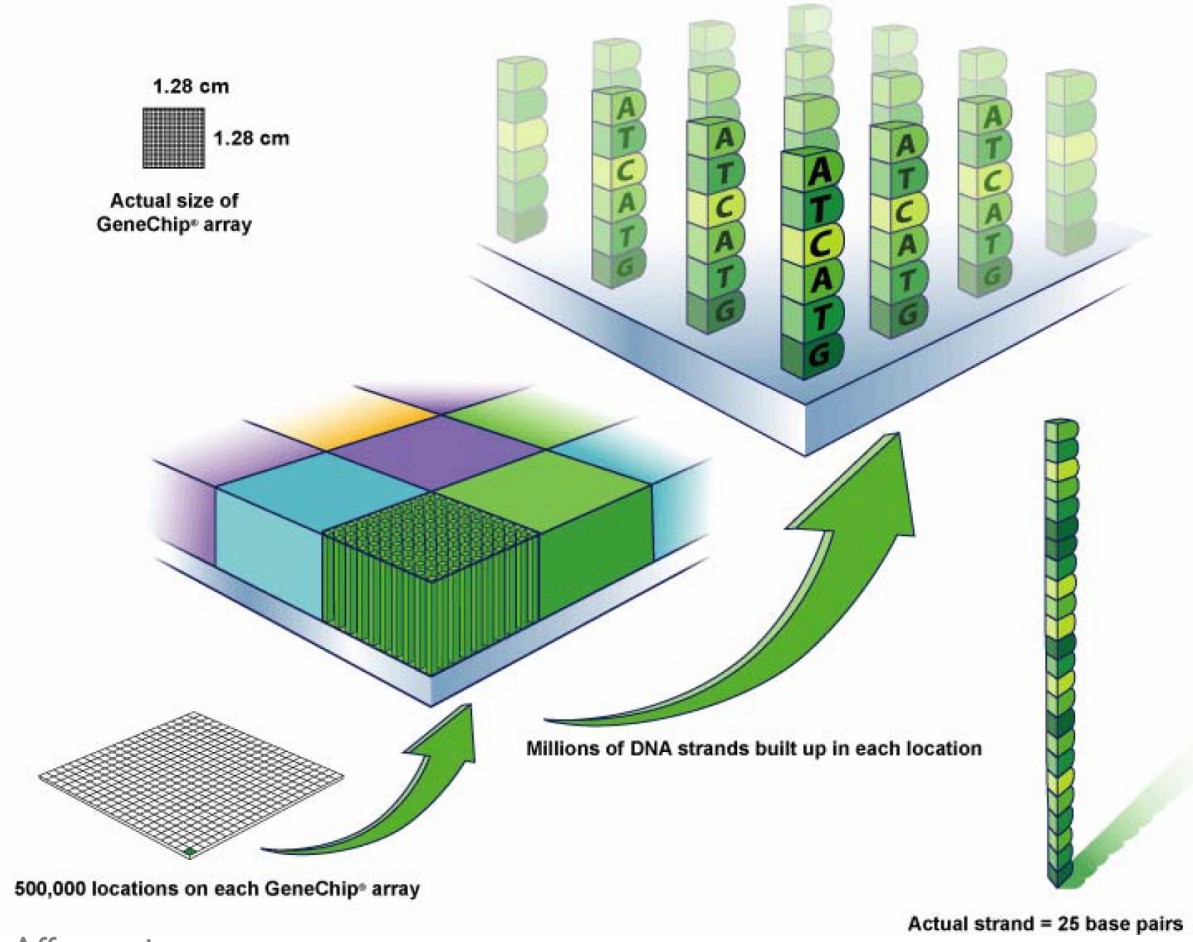
Based on fundamental principles of assay-by-hybridization developed since the 1970s

- Substrate glass slide contains specific short DNA "probes" that have been previously designed
- Each spot has only one probe sequence (and we know the mapping of location to probe)
- Start with a small amount of RNA or DNA purified from your sample (<< 10^{-6}g (or µg); one cell has few picograms: 10^{-12}g)
- **Hybridize** a labeled mixture of RNA or DNA from sample
- Readout is fluorescence of the spot: brighter = more of the labeled target in your sample

Design and construction of microarrays

Example: Affymetrix

- 25-bp oligonucleotides ("probes") synthesized in specific locations ("spots") on the array
- The sequence of the probes is designed to match sequences we expect to see in our sample
- Typically "off-the-shelf" but can be customized for cost (e.g. custom set of probes)

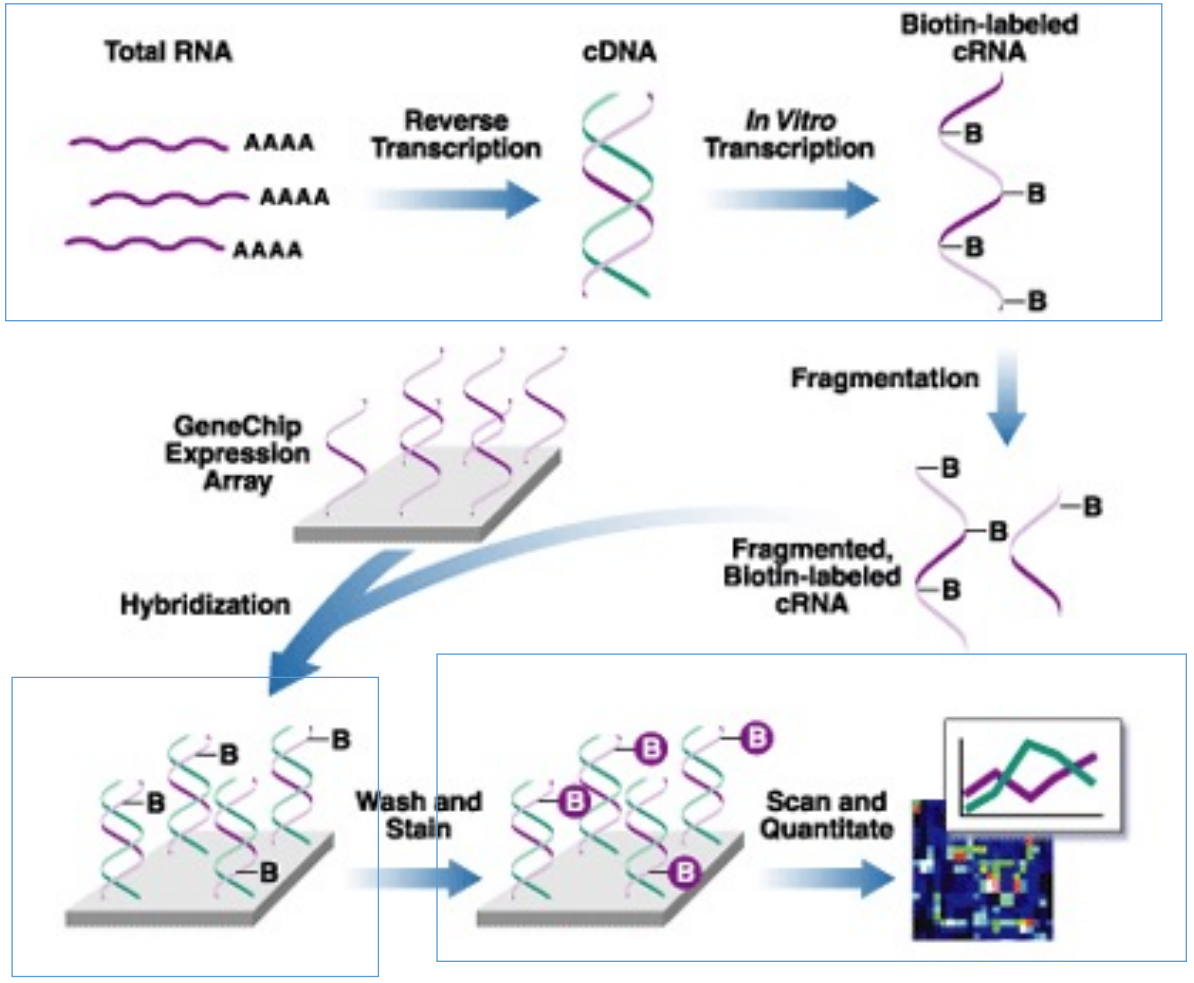


Affymetrix

Detection with Hybridization (Affymetrix)

Illustration is for RNA
Similar idea for DNA assays

Label RNA from cells of interest



Hybridize to chip

Detect where labeled RNA has bound

Pros and cons of hybridization-based assays

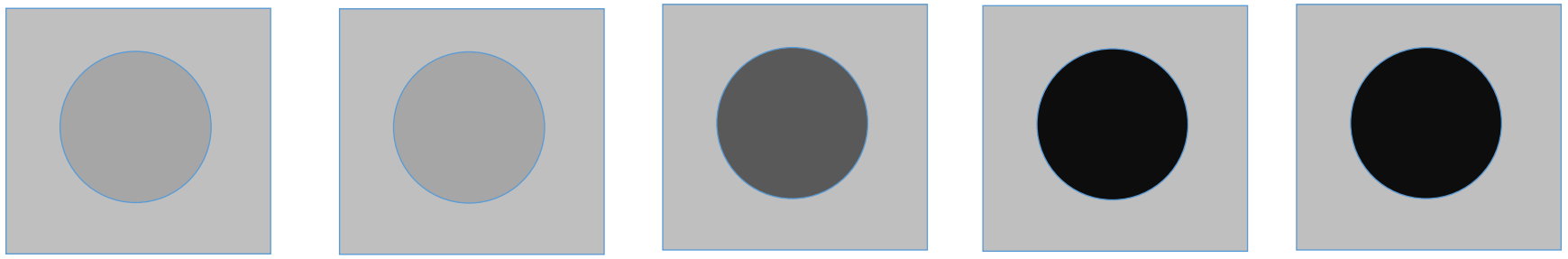
Pros: Mature, inexpensive, small data files

Cons:

- Only detect what you probe (acceptable for many applications)
- … so can't use on an "uncharacterized genome"

For *quantitative* uses:

- Cross/nonspecific hybridization → background + ambiguity
- Dynamic range further limited by saturation



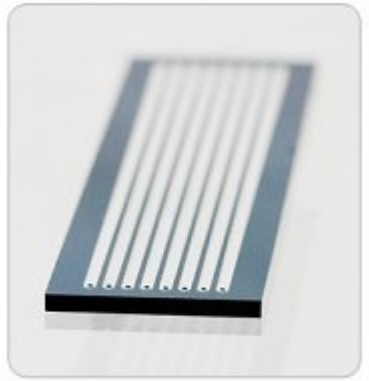
Sequencing basics

Sequencing-based assays

- Start with a small amount of RNA or DNA purified from your sample (<<1μg)
- Determine nucleotide sequences from the sample

Until recently (~2008) this was not feasible compared to hybridization approaches

- Generally "shotgun" – you *randomly* sample millions of short pieces of the input sample DNA/RNA and sequence them
- Then figure out what the pieces are/put them back together



Illumina device and sample flowcell

Sequencing coverage/depth

- Key statistic of interest in many settings is **coverage/depth** – how many times have you sequenced each base that you care about.
- Redundancy is important for confidently calling genotypes (DNA) or quantification (e.g. RNA)
- Additionally, without redundancy the set of bases you have sequenced will be limited because of the random sampling of which targets get sequenced
- Example "30x" – so if you want to sequence a 3Gb genome at 30x on average, you actually need to sequence 90Gb of DNA. With 100bp reads, that's 900 million reads
  - Lander-Waterman eq: Coverage = LN/G where L: read length; N reads; G haploid genome size; Lander and Waterman (1988) describe probability of getting desired depth at any given base based on Poisson statistics, etc.

Illumina reversible terminator sequencing

- Originally developed by Solexa; currently dominant platform
- It shares some technology with microarrays: glass slides, bound nucleotides, fluorescence detection
- Instead of hybridizing, do sequencing (base by base) on the chip

- Up to billions of reads per run on one machine
- Sequences up to ~150bp (x 2 for paired ends)*
- Per-base sequence error rate <1%

*MiSeq machine can do 300bp reads, but much lower throughput per machine. Also error rates get higher at longer read lengths, so common to use 100 or 125bp since they'd have to trim out most of the rest anyway.

See https://www.illumina.com/systems/sequencing-platforms.html

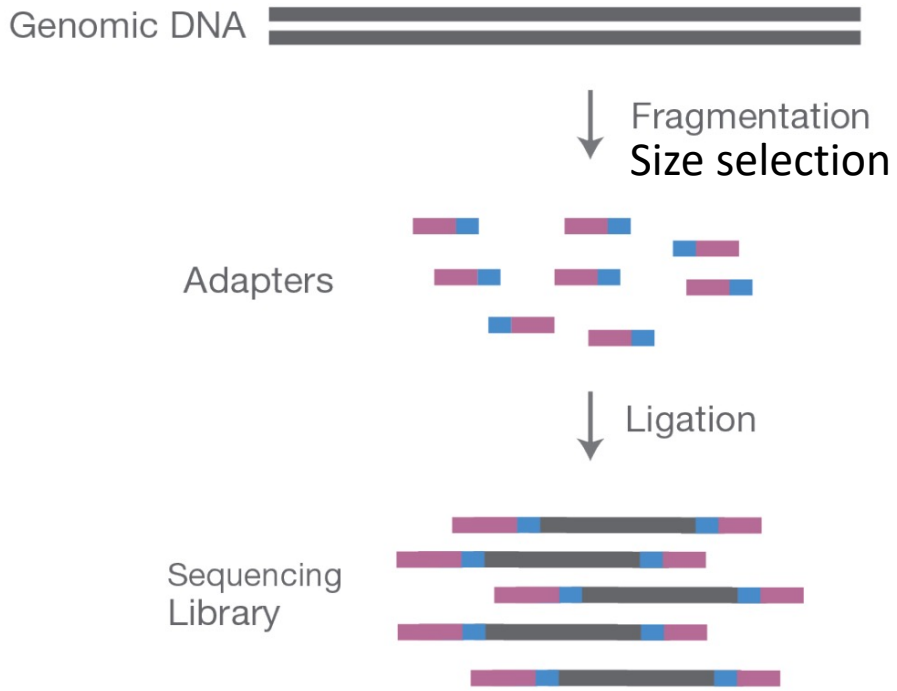
Throughput of some Illumina platforms

	NextSeq 550 Series	NextSeq 1000 & 2000	NovaSeq 6000
Run Time	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	360 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.2 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 250**

- For many applications samples are multiplexed (multiple samples run at once in a single lane)
- To see more details of how different scenarios would work on different Illumina devices see https://support.illumina.com/downloads/sequencing_coverage_calculator.html

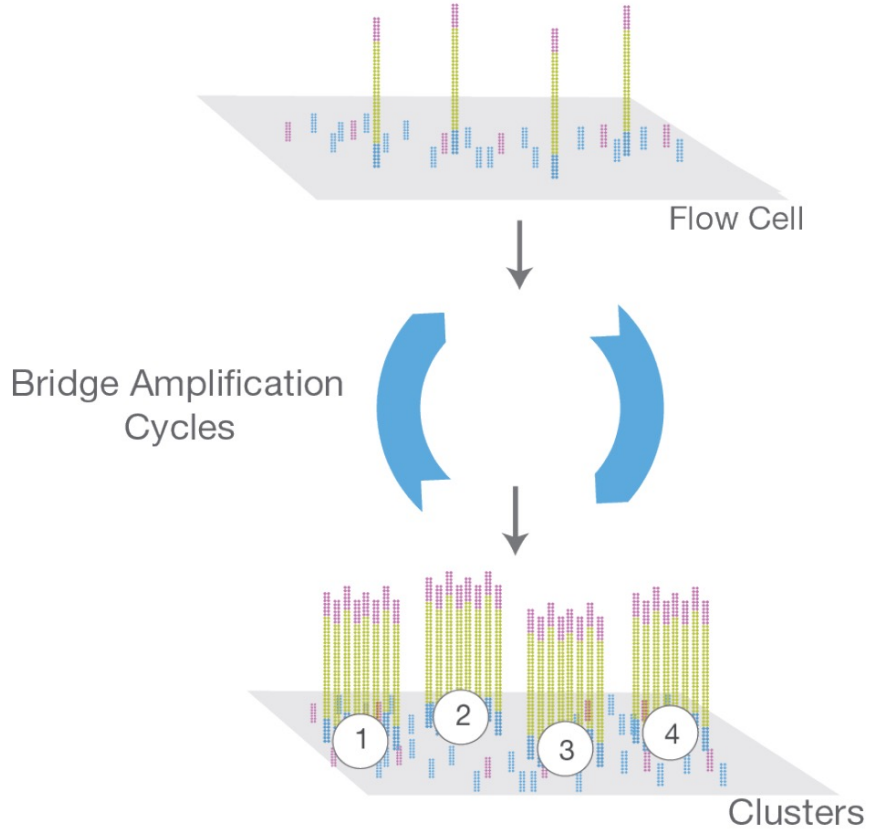
Source: https://www.illumina.com/systems/sequencing-platforms.html

A. Library Preparation



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

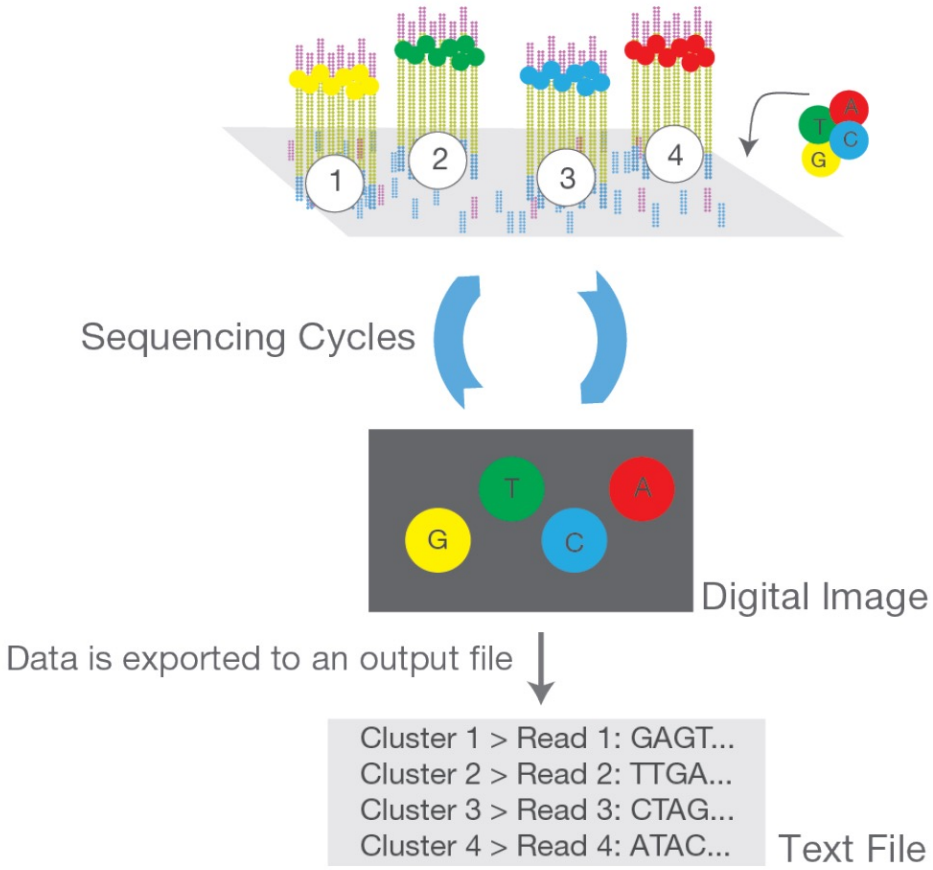
B. Cluster Amplification



Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

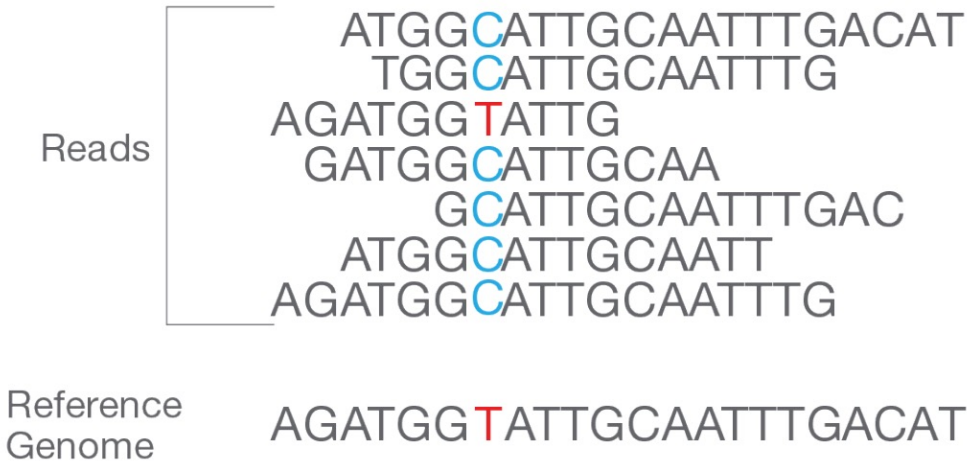
An introduction to Next-Generation Sequencing Technology
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

D. Alignment and Data Anaylsis



Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Sequencing modes/choices

- **Read length** (Pay more for longer reads, but usually worth it e.g. higher alignment accuracy)
- **Single end**: only read one end of a fragment vs **Paired end**: read both ends (takes twice as long, but also usually worth it e.g. alignment accuracy, detection of alternative splicing events)
- **Strand-specific** (important for RNA)
- **Multiplexing**: add a short identification sequence to each library (this is necessary to make it economical)
- **Single molecule indexing**: add an identification sequence to each input molecule; used in some single-cell methods (also called unique molecular identification / UMI)
- DNA: **exome** vs. **whole genome** (Human exome is ~50M bases, counting exon-flanks)
- RNA-seq: we'll come back to

Long-read sequencing

Biggest problem with many approaches: **reads are short**

e.g. Challenge accurately detecting and quantifying different transcripts from a gene; or isolating one genome from a microbial population.

What we want are long reads – entire mRNAs, long stretches of chromosomes – ideally without sacrificing throughput and accuracy

To the rescue (?) Single molecule sequencing

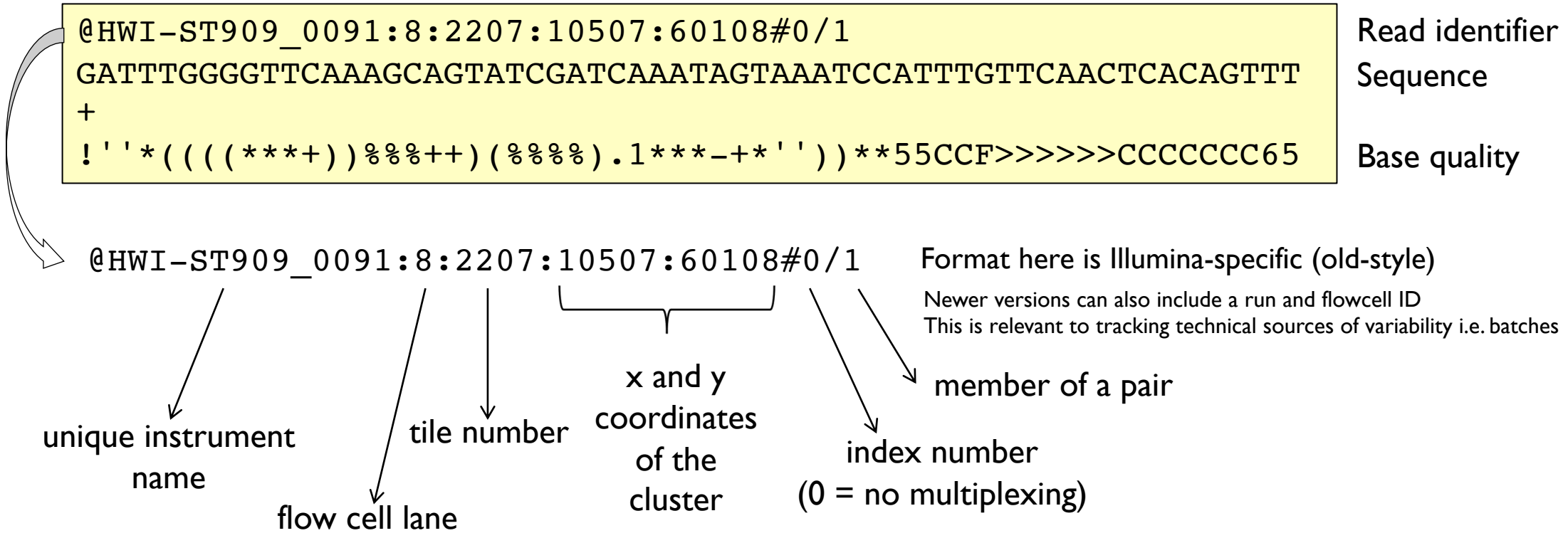
- e.g. Pac Bio SMRT, Oxford Nanopore
- Major benefit: Can produce *very long reads* (>>1kb)
- Downside: higher error rates (not as good for genotyping), lower throughput (not as good for quantification)

Often used as an adjunct to short-read seq. – vast majority of data still comes from short-read

Raw sequence data: FASTQ format

Text file; Sequences with quality information; ~60GB file for 500M paired-end reads, 100bp

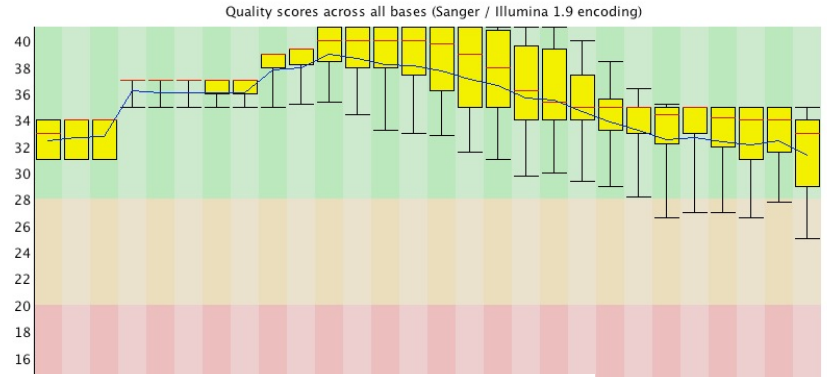
Data for one Illumina read:



Base qualities (Q) are interpreted as ASCII byte from 33 (new-style). Read as values from 0 so "!" = 0; "C" = 34

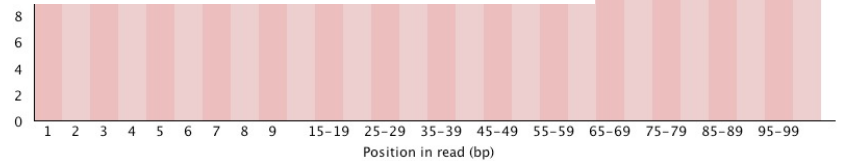
Interpretation: P(base call is wrong) = $10^{-Q/10}$ (AKA "Phred")

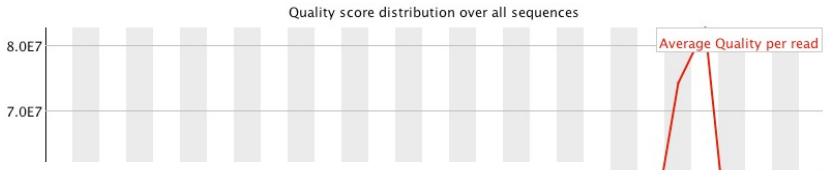
Sequence quality: FASTQC software



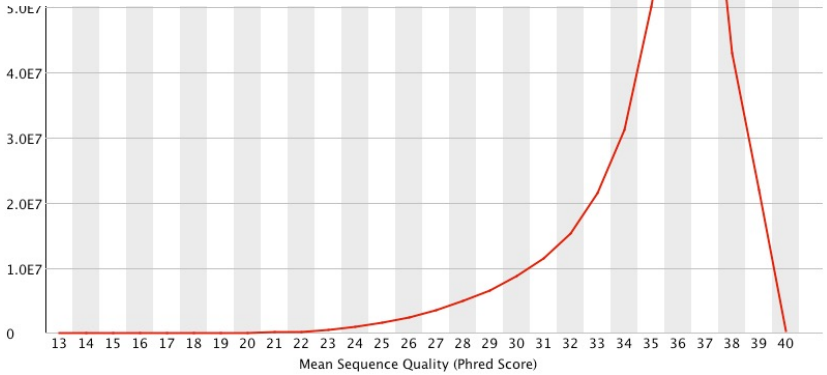
Phred quality Q20 = 1 in 100
prob of being incorrect

Quality score distribution per position





Mean quality score over all sequences



Also monitored for quality control:

- GC content
- sequence length distribution
- sequence duplication level

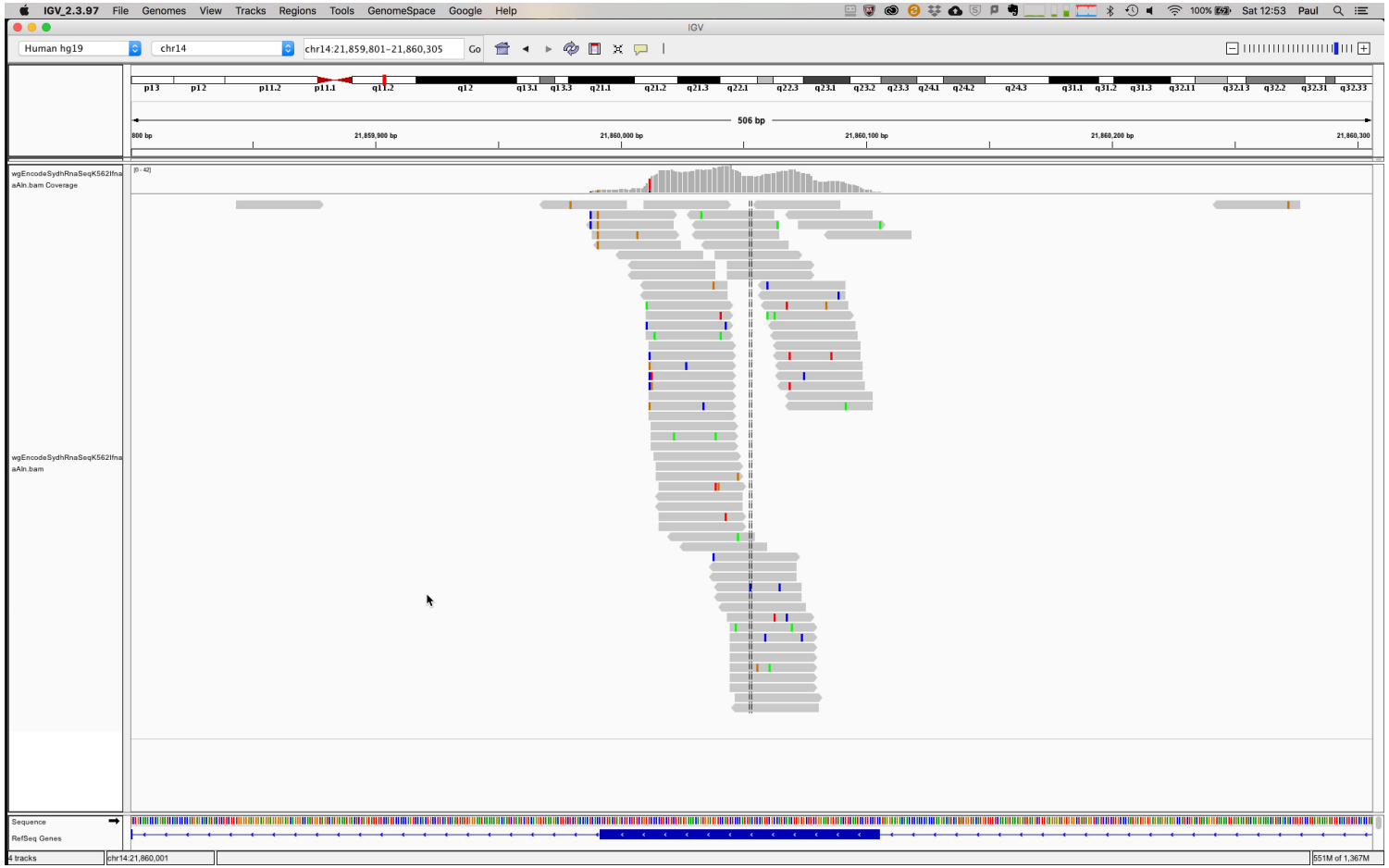
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aligned data – SAM/BAM format

- Contains the original sequence information *plus* the **alignment** information (where in the genome is the sequence located)
- SAM – tab-delimited text file with alignment info (human readable)
- BAM – binary version
  - Efficient storage (typically compressed), but not human readable
  - Indexing for quick data retrieval by genome coordinate; auxilliary .bai file
  - BAM files can be 30Gb or more per RNA-seq sample; WGS: 80Gb+ (depends on how many reads and their length).

Viewing alignments from a BAM file

- Somewhat like the genome browser I showed, but shows data from one or more BAM files
- This example is an RNA-seq sample (and low quality data by current standards)
- Observe
  - Read length and depth
  - Read direction – what does this tell you about the way the library was made?
  - Most reads align to exons as expected
  - Many reads have mismatches that don't look like SNPs



Software: IGV (Broad Institute)

Data from http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeSydhRnaSeq

Part II: Specific assays

- For many things we'd want to measure, it can be done multiple ways; often there are both microarray methods and sequencing methods

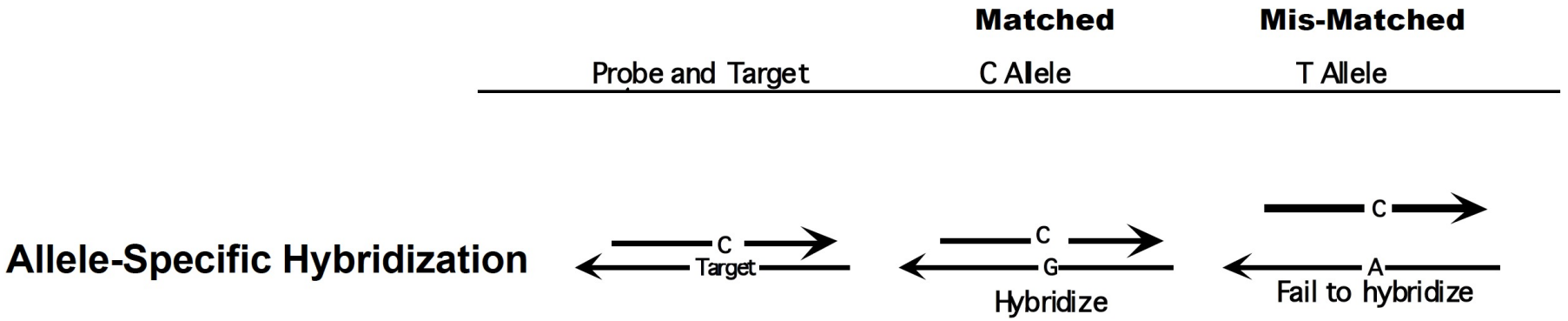
- Choice often comes down to tradeoffs in a given experiment
  - Can microarrays sufficiently answer the question? If so, they are generally cheaper
  - If higher resolution is desired/needed, sequencing is often the answer (if budget allows)

Pause for Q2

SNPs (Single Nucleotide Polymorphisms)

SNP detection with microarrays

- Recall: SNPs (Single Nucleotide Polymorphisms) are a class of common single base differences in DNA
- Because the polymorphic sites are (mostly) known in some well-studied organisms / populations, can use microarrays of pre-designed probes covering a set of SNP sites
- Arrays currently assay ~1 million SNPs (or more). *Imputation* used to infer additional sites
- One method: microarray is constructed with probes that distinguish the two alleles.
  - Affymetrix: multiple 25-base probes for both alleles, compare signal
- Output: The number of reference alleles (0,1 or 2) at each site

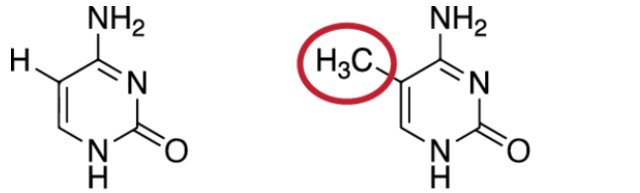


Debbie Nickerson via https://egp.gs.washington.edu/workshop/download/Day2_1045am_Nickerson.pdf

DNA Methylation

DNA methylation

- A chemical DNA modification associated with gene regulation; not all species have it
- Most commonly studied form is the methylation of cytosine in CG-dinucleotides (CpG)
- A site is either methylated or not – in a diploid cell, can have 0, 1 or 2 methylated.
- 60-80% of CpGs are generally methylated
- "CpG islands" – regions of high CpG density, often in promoters, mostly unmethylated
  - Methylation in promoters tends to be associated with low transcription



Wikipedia

http://www.mpipsykl.mpg.de

A seminar later in the term is focused on DNA methylation analysis

Role of DNA methylation

- DNAme is relatively stable, but dynamics are of interest
  - It is thought that relatively long-lasting "epigenetic states" might be partly reflected in DNA methylation
  - e.g. Different cell types have very different methylation patterns
- Methylation is classically associated with gene silencing – often via DNAme-binding proteins that result in changes in chromatin state
  - e.g. transposable elements, imprinting (only one inherited allele is expressed)
- Goal: survey genome-wide methylation states for populations of cells in different individuals/conditions/cell types

DNAme assay types

- Human genome contains about 28 million CpGs
- Whole Genome Bisulfite Sequencing (**WGBS**) can assay all of them
  - but inefficient; many reads (~65%) will not contain a CpG
- Reduced-representation bisulfite sequencing (**RRBS**) is a targeted approach
  - use a digestion and purification step to first enrich sample for CpG-rich DNA
  - typically ~2 million CpGs assayed – about 1% of the genome
- Microarray: Illumina beadarray platforms contain probes designed to distinguish me-C vs. C at selected CpGs
  - Older platform: "450k" ~485k sites
  - "EPIC" >860k sites

Assaying DNA Methylation

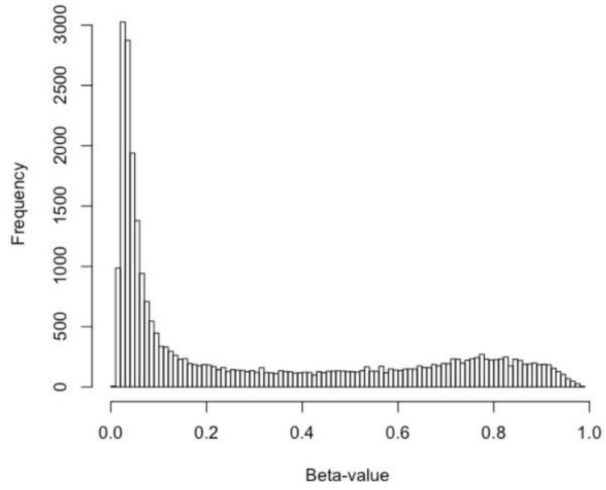
- Methods take advantage of distinct chemical properties of methyl-C
- Bisulfite conversion: chemical treatment turns Cs into Us *unless they are methylated*
- Then compare the sequence of a bisulfite-treated sample to the non-converted reference
  - **Microarray**: have probes that distinguish these cases for a particular set of pre-selected sites
  - **Sequencing**: all or part of genome (U will be read as T) – requires some modification to aligners
- Output for each site: "β"
  - **Arrays**: relative intensity of methylated vs unmethylated probes
  - **Sequencing**: the fraction of reads methylated
  - β in [0,1]; often analysis is done on $\log_2(\beta/(1-\beta))$ "M-value" (logit)
  - See Du et al. https://doi.org/10.1186/1471-2105-11-587
- Other approaches exist - e.g. use antibodies that are specific to methyl-C ("**MeDIP**") but don't give base-level resolution

Me
ATATGTTCGAAACGTAG

Bisulfite

Me
ATATGTTCGAAAUGTAG

Not shown: other strand of DNA

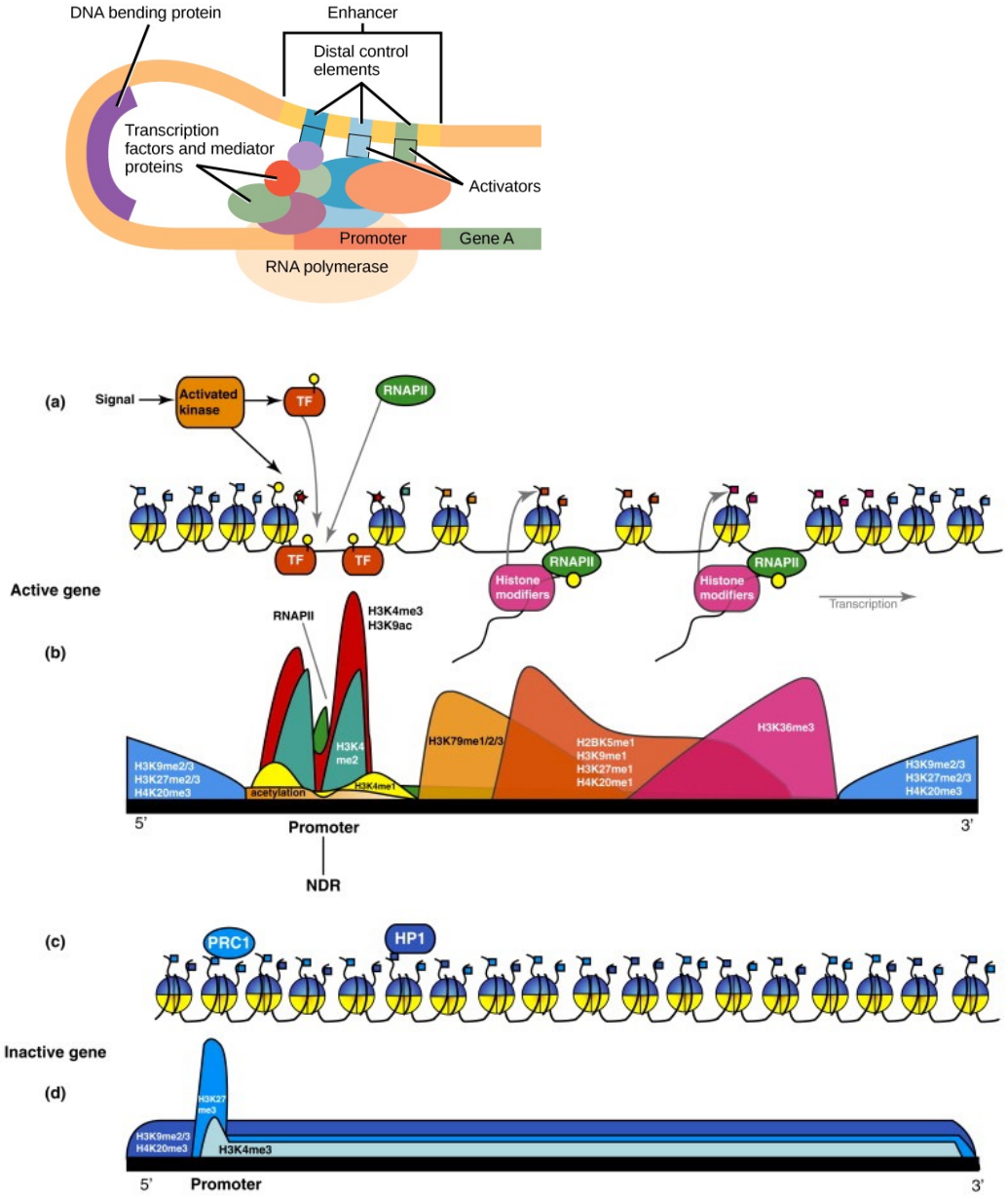


Typical beta distribution from Du et al.

DNA-binding proteins

ChIP-seq

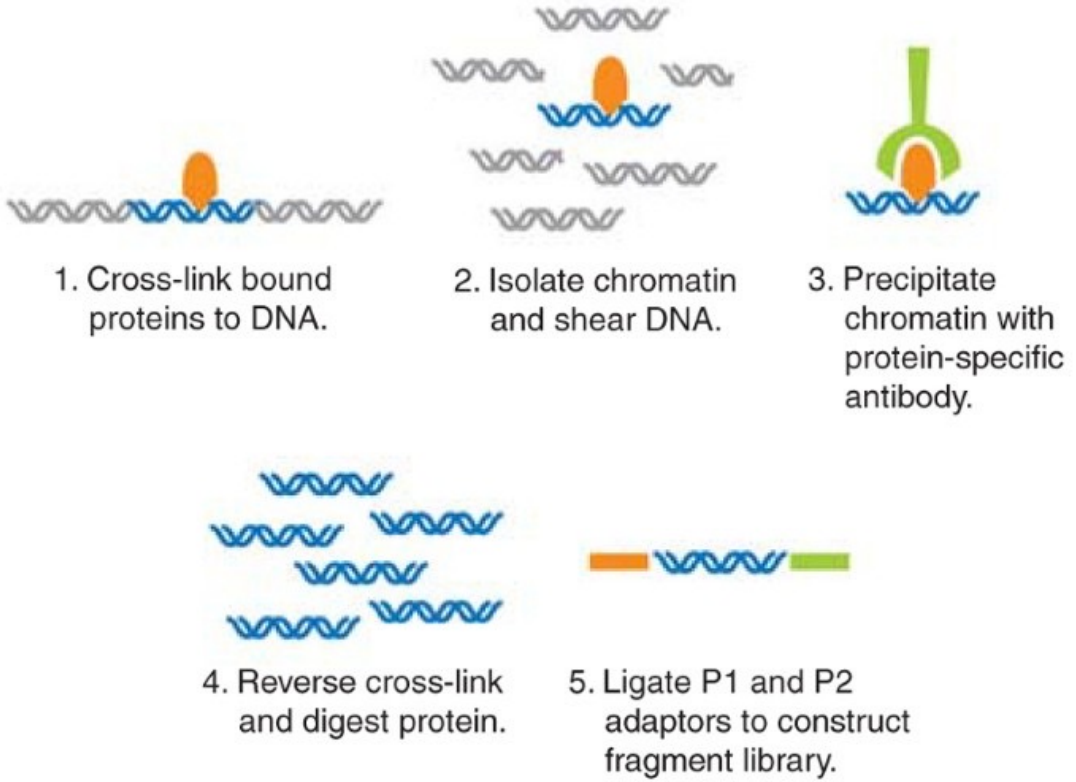
- Sequencing of genomic DNA fragments that are bound by **DNA-binding proteins**
  - Transcription factors
  - RNA polymerase
  - Chromatin-modifying enzymes
  - Histone states
- Two steps: chromatin immunoprecipitation (ChIP) and sequencing (seq)
  - Key reagent is a **specific antibody** against the protein you want to study
  - Some related methods use other ways to isolate fragments of interest e.g. DNAase-seq, ATAC-seq
  - Microarray methods preceded ChIP-seq (e.g. ChIP-chip)



Barth and Imhof, 2010 http://dx.doi.org/10.1016/j.tibs.2010.05.006

ChIP-seq protocol

1. Cross-linking protein to DNA – reversible formaldehyde fixation
2. Fragmentation – double-stranded DNA fragments <1kb
3. Immunoprecipitation – using specific antibodies to select bound fragments
4. Reversing crosslinking and amplification
5. Make sequencing libraray
6. Sequencing

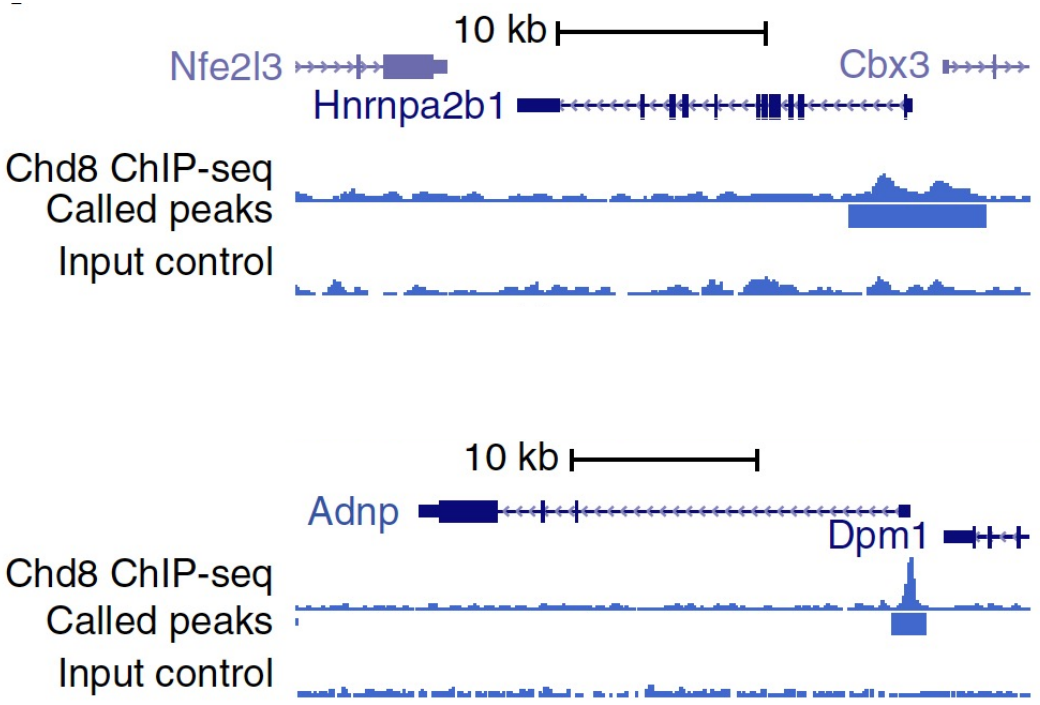


https://www.nature.com/articles/nmeth.f.247

Control sample: "input DNA" only without IP (less good control: "non-specific" antibody)
There is an extensive literature on various problems and artifacts including coverage biases

ChIP-seq analysis

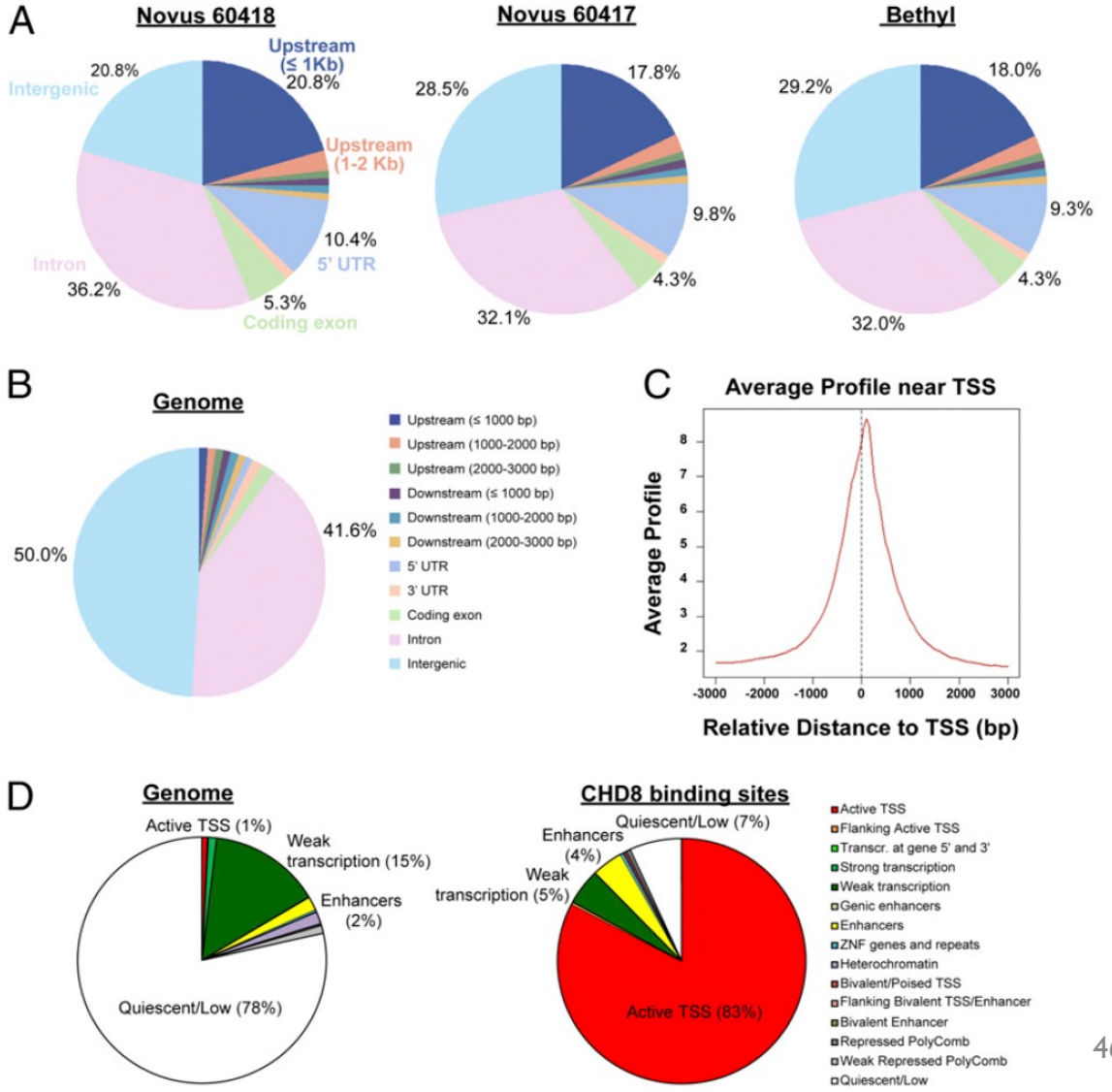
- Aligning sequence reads with short-read aligners
- Quantifying binding - peak finding
  - Goal: identify real peaks; estimate confidence
  - Basic idea: count the number of reads in windows and determine whether this number is above background – if so, define that region as bound
  - Many algorithms (>60)
  - Approaches: hard thresholding, HMMs, statistical tests, …



Gompers et al. 2017

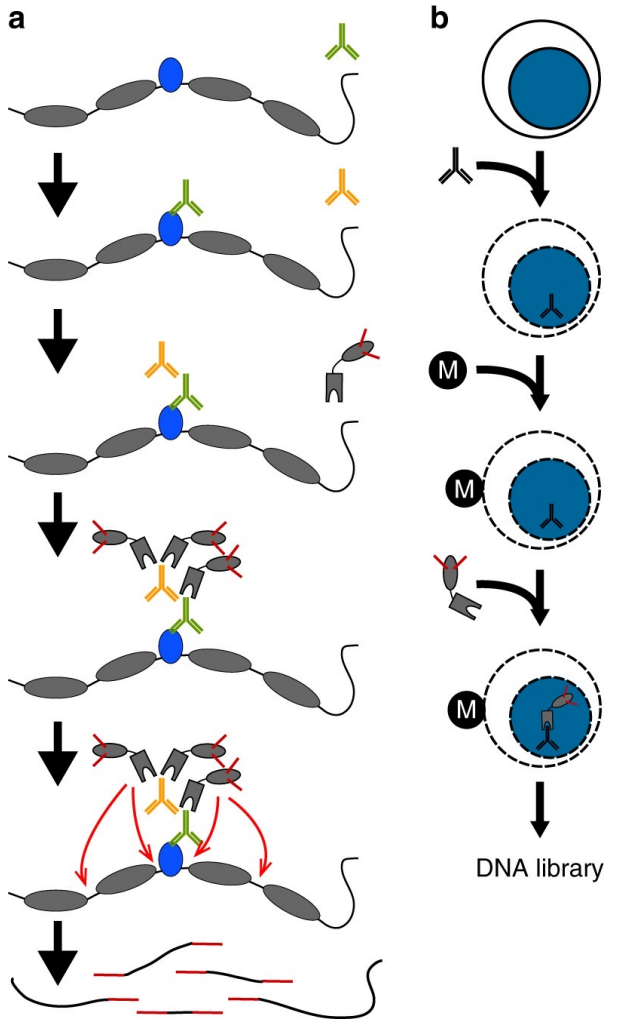
Example: CHD8 ChIP-seq study

- Tested 3 different antibodies
- Used a "neural precursor" cell line as DNA source
- Figure shows distribution of sites relative to various gene parts (or intergenic); There were about 15,000 sites found in each analysis.
- "… 7,324 sites that were replicated by all three antibodies at a Benjamini–Hochberg q value < 0.05 …" – tend to be near "active transcription start sites"



Sugathan et al. https://www.ncbi.nlm.nih.gov/pubmed/25294932

CUT&Tag-seq (Cleavage Under Targets & Tagmentation)



- Uses a secondary antibody that anchors a transposase enzyme to guide cleavage of DNA at the target binding site
- Much increased specificity (lower background); suitable for small numbers of cells / single-cell
- Doesn't require lysing of cells
- Central analysis task is still Peak calling, but need modification for high signal to noise ratio
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1802-4
  - https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-019-0287-4

Kaya-Okur et al., 2019: https://pubmed.ncbi.nlm.nih.gov/31036827/

Transcriptomes (Expression)

Assaying RNA expression (transcriptomics)

- Obtain a quantitative measure of the expression level of each gene (~20,000, though not all expressed at any given time)
- Ideally: tell different transcripts (isoforms) for the same gene apart

Why?

- It's relatively easy to measure (compared to proteins, in particular)
- **Gene expression is regulated:** changes in response to environment, disease, age, etc.
- The pattern of gene expression can be used as a '**fingerprint' of the state the sample** was in at the time of measurement
- Examining the details about which genes are relevant to the fingerprint ("differentially expressed") give **insight into the process/disease/condition** of interest or **biomarkers**
- The **"readout" of genetic variation** is partly in gene expression – it can help us understand the link between genetics and phenotypes

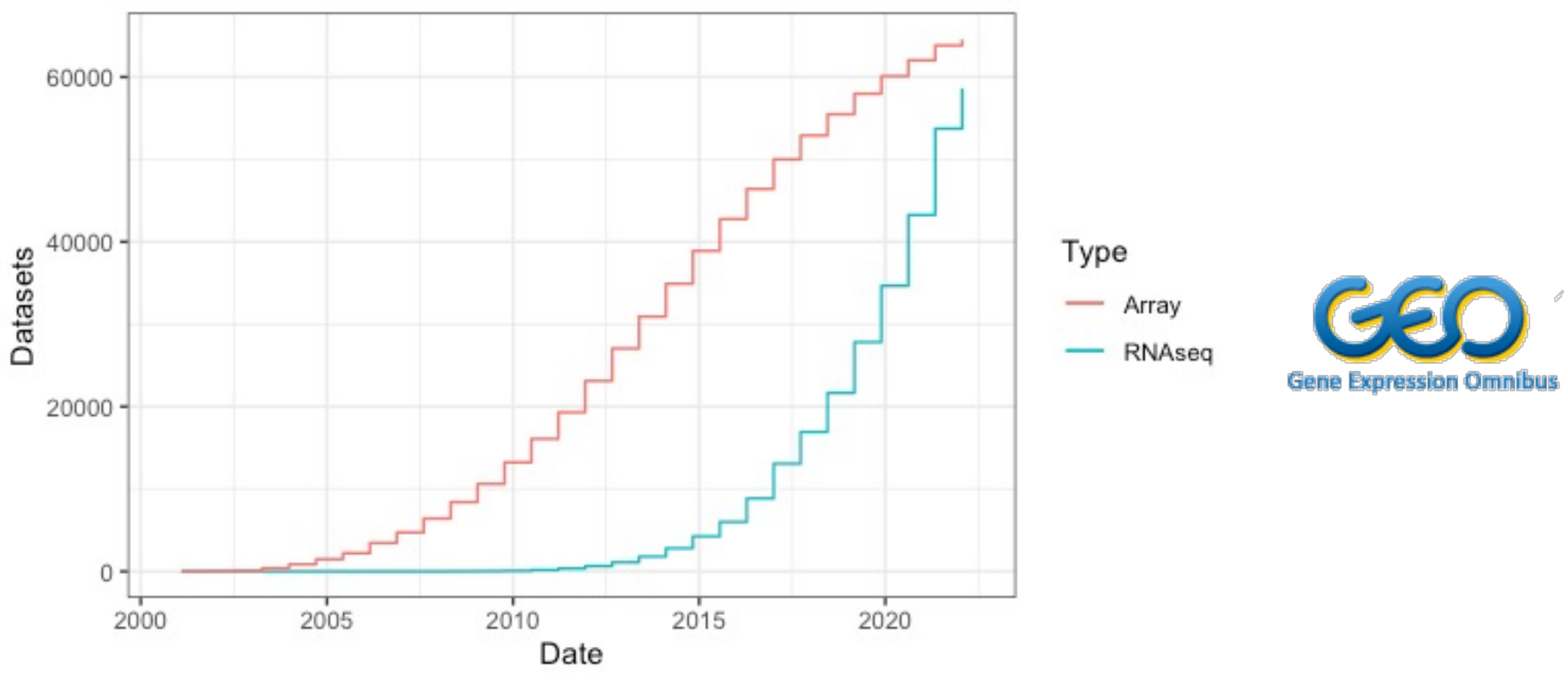
Choice of method: the move to RNA-seq

Microarrays were the dominant method until quite recently. Still used, but RNA-seq is now cheap enough to be widely used. Additional benefits of RNA-seq:

- Doesn't require prior genome annotation (e.g. probe design)
- Very little signal background (i.e. if a gene is not expressed, won't see any signal)
- Potential to detect splicing variation, gene fusions
- Allows other applications such as variant detection, allele-specific expression quantification, or RNA editing assessment
- Greater dynamic range → increase in sensitivity, precision with sufficient read depth. For some applications ("typing") only 50k reads might be acceptable; for quantitative need more like 20M.

Minor drawback: Much larger data files compared to microarrays

Trends in uses of RNA-microarrays vs RNA-seq



As of January 11 2022: Total of **58,638** RNA-seq and **64,571** microarray studies

Expression Profiling study counts from https://www.ncbi.nlm.nih.gov/geo/summary/

RNA-seq general approach

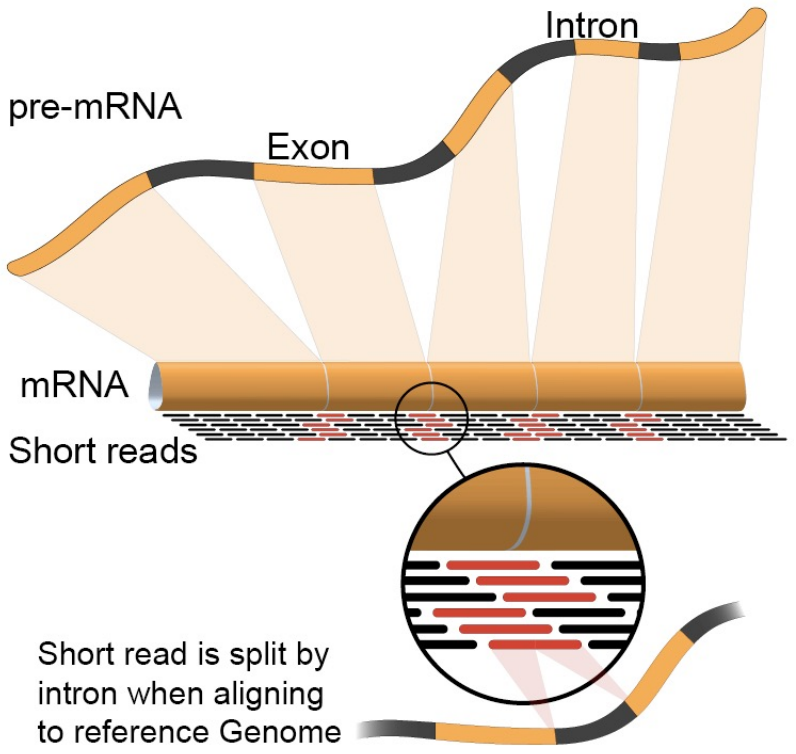
- Randomly sample molecules from your RNA sample
- Sequence their end(s)
  - Typically obtain ~20M paired-end reads for each sample for quantitation
  - Multiplexing allows dozens of samples to be run at once
- Figure out what transcript/gene the reads come from

RNA-seq choices

- Most cellular RNA is ribosomal RNA (rRNA; >90%) so generally use either polyA selection or ribosomal RNA depletion to isolate mRNA
  - Most mRNAs are polyadenylated; but some of interest might not be
- RNA size – for very short RNAs (i.e. miRNA) use different library prep methods
- Typically want strand-specific sequences
- RNA-seq can be used on very small samples - even single cells using special techniques (we revisit this topic later this term!)

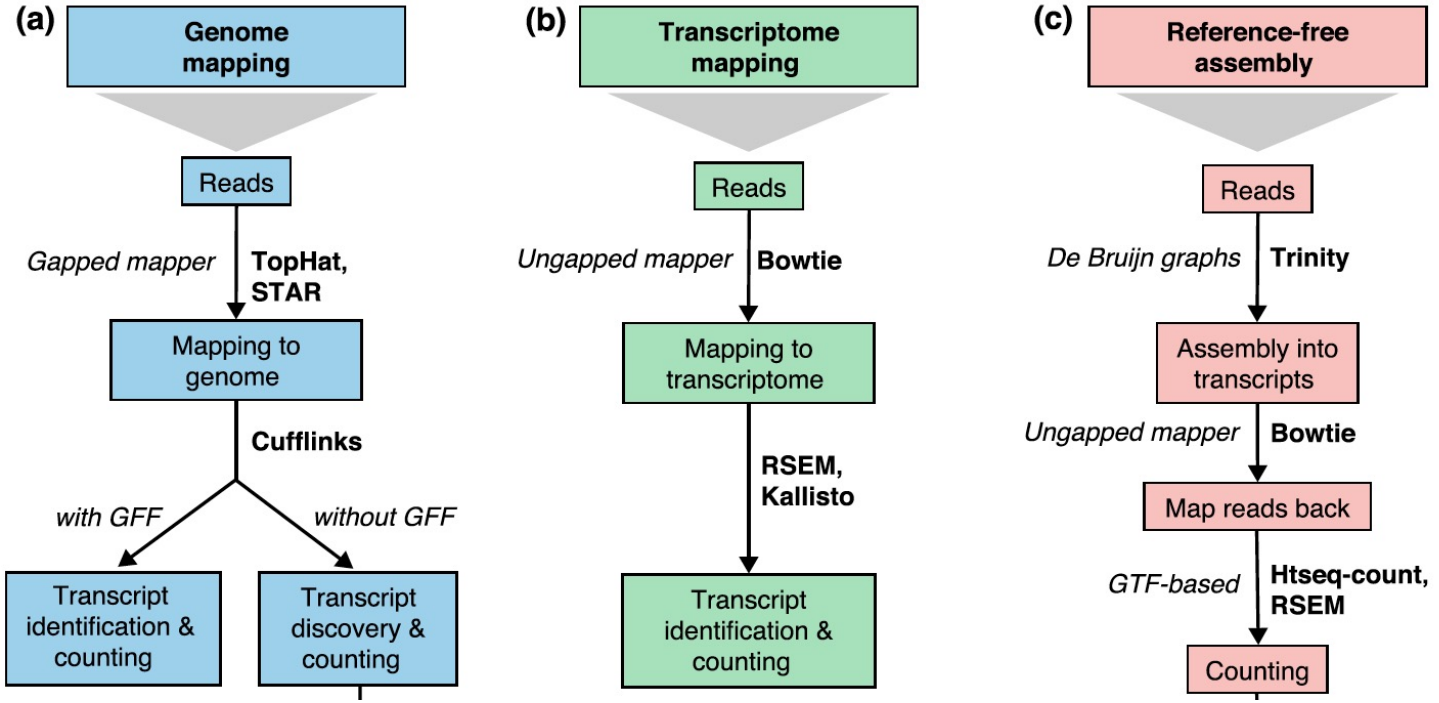
The challenge of exon junctions

- Recall that the RNA we're analyzing is already **spliced**. The junction sequences are not present in the genome.
- While many transcripts are annotated, we will potentially see 'unknown' transcript structures (unknown exon-exon junctions)
- We're only getting short reads
  - Typical human exons 50-200bp long; Mean of ~8 exons per gene; mRNA length can be thousands of bp. Compare to typical read lengths of 100bp.
- Many computational methods/approaches ...



Rodrigo Goya/Wikipedia

“Upstream” analysis of RNA-seq data



- By “upstream” I mean taking the raw RNA-seq data from **raw fastq files to counts per gene or transcript**
- Often have to think about a pipeline that is built from multiple tools
- Figure shows several types of pipelines, but it’s not this simple – methods can be mixed and matched, or not, to a varying extent
- Seminar 6 gives practice on this portion of the analysis; next few slides are an overview

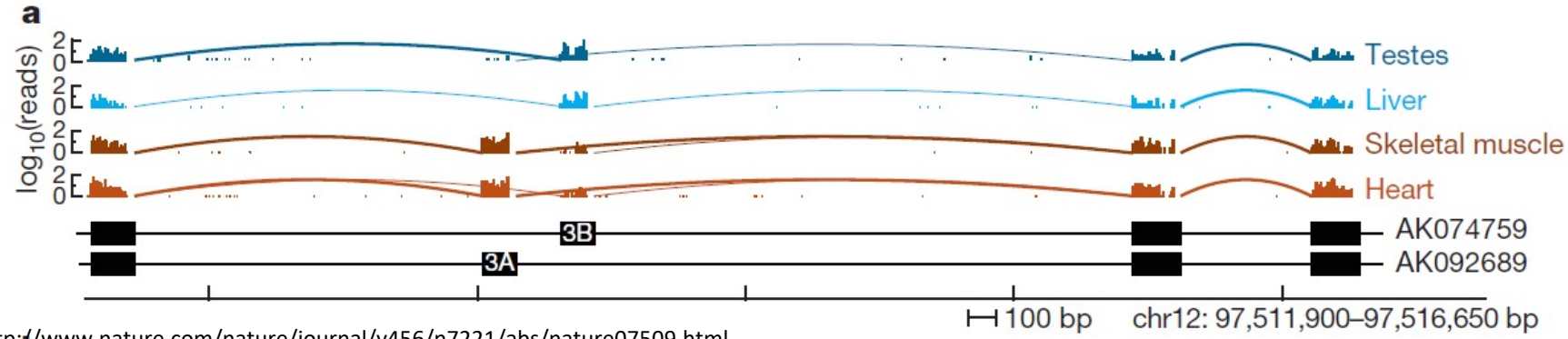
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8 - but already probably out of date

RNA-seq sequence alignment / mapping

- **Task**: find out where in the transcriptome each read came from
- Many tools: TopHat, Bowtie, STAR, GSNAP, GSTRUCT, HISAT…
- Different ways to deal with splice junctions
  - Align (or 'map') to the annotated transcriptome
  - Align (or 'map') directly to the genome
  - Reference-free *de novo* assembly
- Challenge: speed and memory usage
  - Some trade-off between speed and sensitivity
  - Parallelization important
- Still an active area (and will be so long as we have short reads)

Gene/transcript quantification

- **Task**: Given alignment result, count how many reads per gene or transcript
- As for alignment, multiple tools/methods: e.g. RSEM, HTSeq, Subread…
- *Gene* quantification is relatively simple
  - Basically, look at what gene aligned sequence belongs to. Assign read to that gene. Count.
  - **Complications**:
    - Comparing counts between larger and smaller genes
    - Comparing counts between samples with different depth (need normalization!)
- *Transcript* quantification is much harder
  - Have to guess which transcript a read came from – often ambiguous
  - Uncertainty of transcript structures present captured by probabilistic approaches (e.g. RSEM)



http://www.nature.com/nature/journal/v456/n7221/abs/nature07509.html

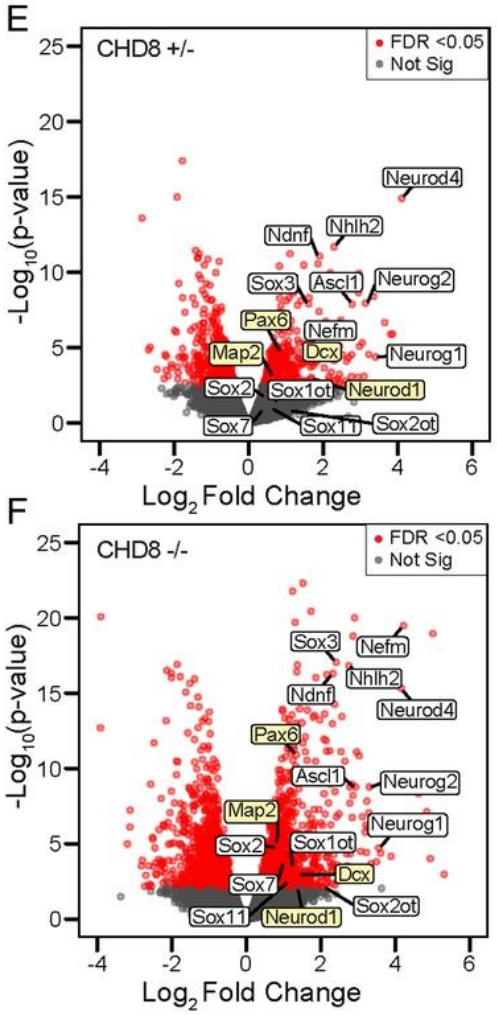
Quantification **without** alignment

- So far we discussed a two-step process of 'upstream' RNA-seq analysis:
  1. Align or map reads to transcriptome
  2. Quantify reads mapping to each gene or transcript
- **Pseudoalignment** methods (Salmon, Sailfish, Kallisto) go from raw reads to transcript quantification in *one step*
- Major advantage – speed
  - Alignment methods focus much computational effort in finding the best alignment
  - This is not so important when we just want to count/quantify how many reads are coming from each transcript
  - Pseudoalignment methods are optimized for this task

Pause for Q3

“Downstream” analysis of RNA-seq

- Counts in hand, we can focus on the “downstream” analysis – focus of this course
- Example study involving CHD8 gene: RNA-seq comparing wild type (CHD8 +/+) mice with two mutant lines
  - WT vs **CHD8 +/-** (~50% levels of CHD8 protein)
  - WT vs **CHD8 -/-** (undetectable levels of CHD8 protein)
- “Volcano” plots on the right show effect size (fold change) versus significance of each gene for both comparisons
  - Highlights neuronal development genes that are differentially expressed when CHD8 is knocked down / out
  - Possible hints to mechanisms of CHD8 involvement with Autism Spectrum Disorders



“Volcano plots of RNA-seq data for CHD8$^{+/-}$ and CHD8$^{-/-}$. DEGs are highlighted in red (FDR < 0.05), and genes involved with mature neuronal development and Sox TFs are labeled”.

Sood et al. 2020, PNAS https://doi.org/10.1073/pnas.1921963117