

Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

High-dimensional genomics assays & data

Keegan Korthauer

9 January 2025

with slide contributions from Paul Pavlidis

Learning Objectives

Part I: 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 (SNP arrays or seq)
 - DNA methylation (bisulfite conversion + seq or arrays)
 - DNA-protein interactions (ChIP-seq, CUT&Tag)
 - RNA quantification (RNA-seq or arrays)

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”

- Low-dimensional approach doesn't tell us how the parts work together
- Technology enabling increasingly detailed analyses – measure many features at once (“High-throughput”, “High-dimensional” or “High-content”)
- Hypotheses/questions with a “non-specific” flavour: e.g. “What genes does CHD8 regulate?”
- Criticisms of the approach include
 - The experiments rarely give easy/simple answers
 - Sometimes just generating “biomarkers” without gaining insight
 - “Fishing expeditions”

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

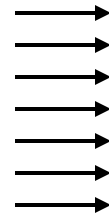
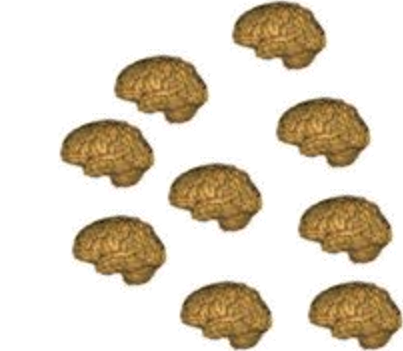
Questions:

- Can we distinguish between the subtypes molecularly?
- Can we find new targets (“biomarkers”) for drugs?
 - Drug targets are often proteins, encoded by genes
- Are there other subtypes that can be discovered? (“clusters”)

A prototype genomics experiment

Each value =
Measurement of one gene in one sample

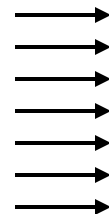
Samples
Subtype A



Assay
(e.g. RNAseq)



Subtype B



One column per sample

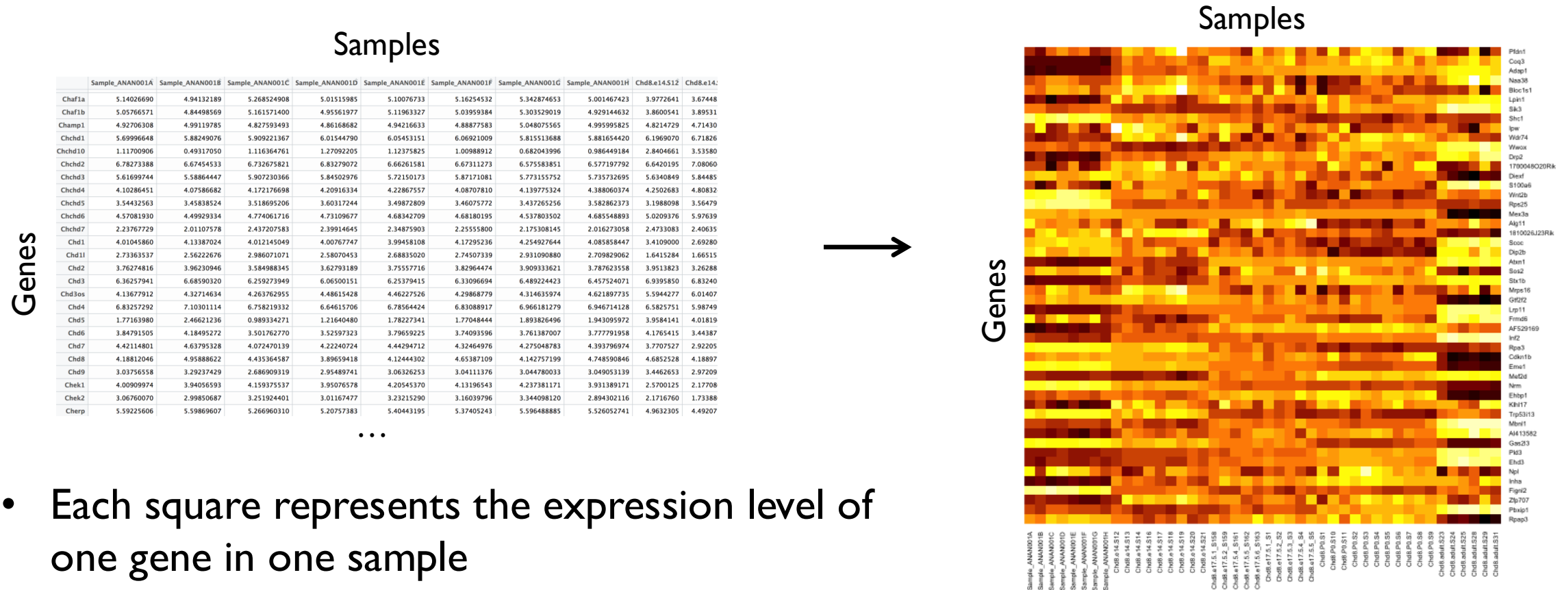
	Sample_ANAN0001A	Sample_ANAN0001B	Sample_ANAN0001C	Sample_ANAN0001D	Sample_ANAN0001E	Sample_ANAN0001F	Sample_ANAN0001G	Sample_ANAN0001H	Chd8.e14.S12	Chd8.e14.
Chaf1a	5.14026690	4.94132189	5.268524908	5.01515985	5.10076733	5.11254532	5.342874653	5.001467423	3.9772641	3.67448
Chaf1b	5.05766571	4.84498569	5.161571400	4.95561977	5.11963327	5.33959384	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

One row per gene (thousands)

Subtype A

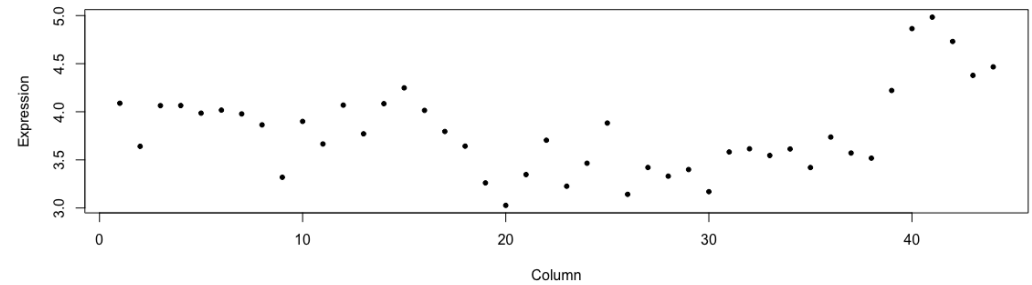
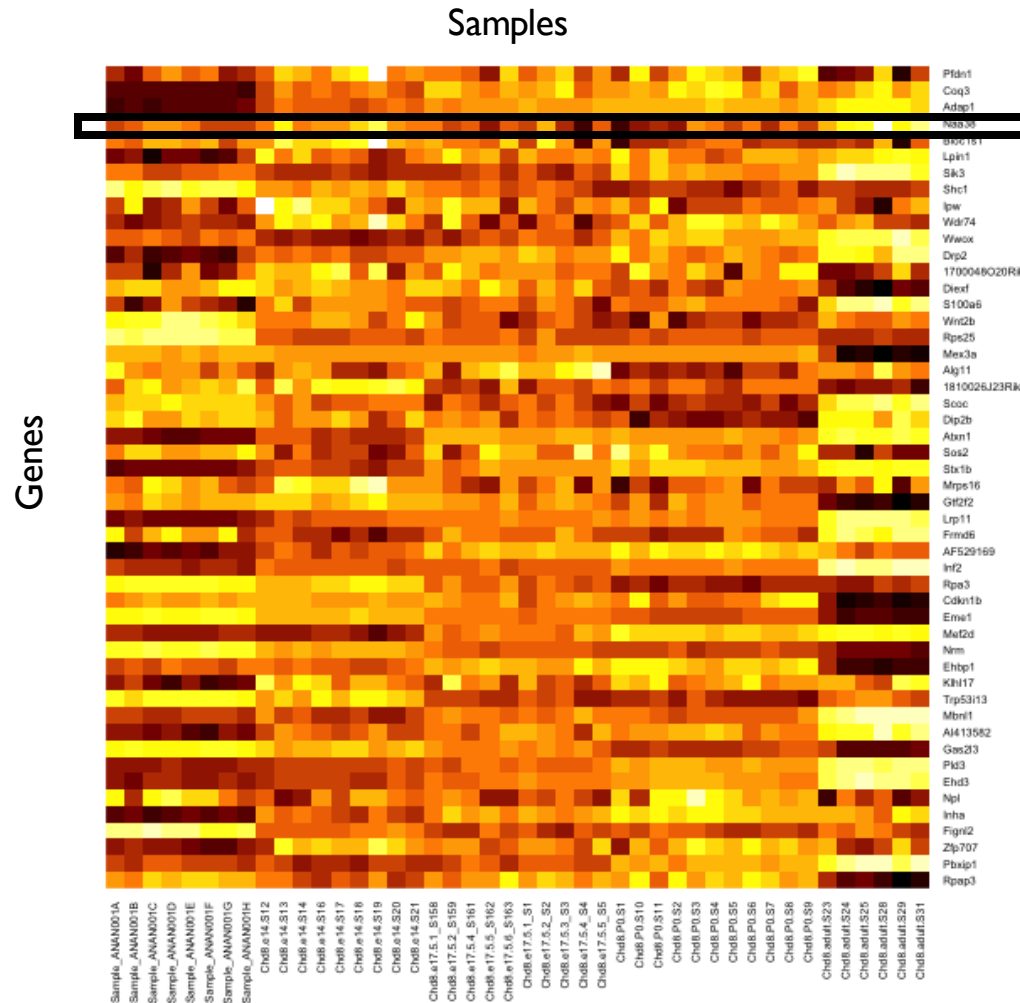
Subtype B

Alternative representation as colours



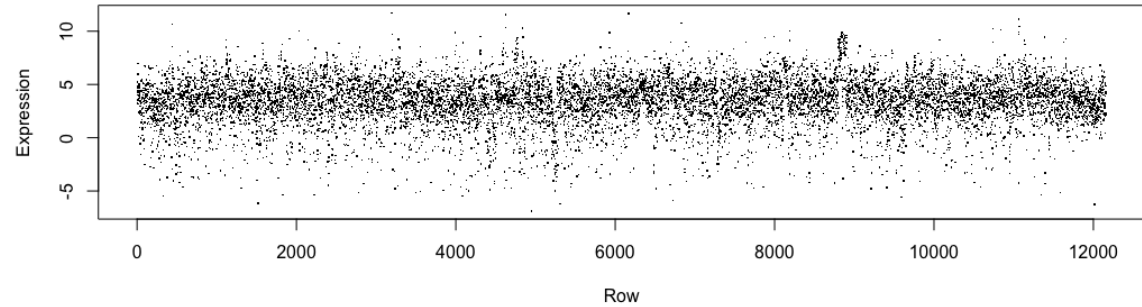
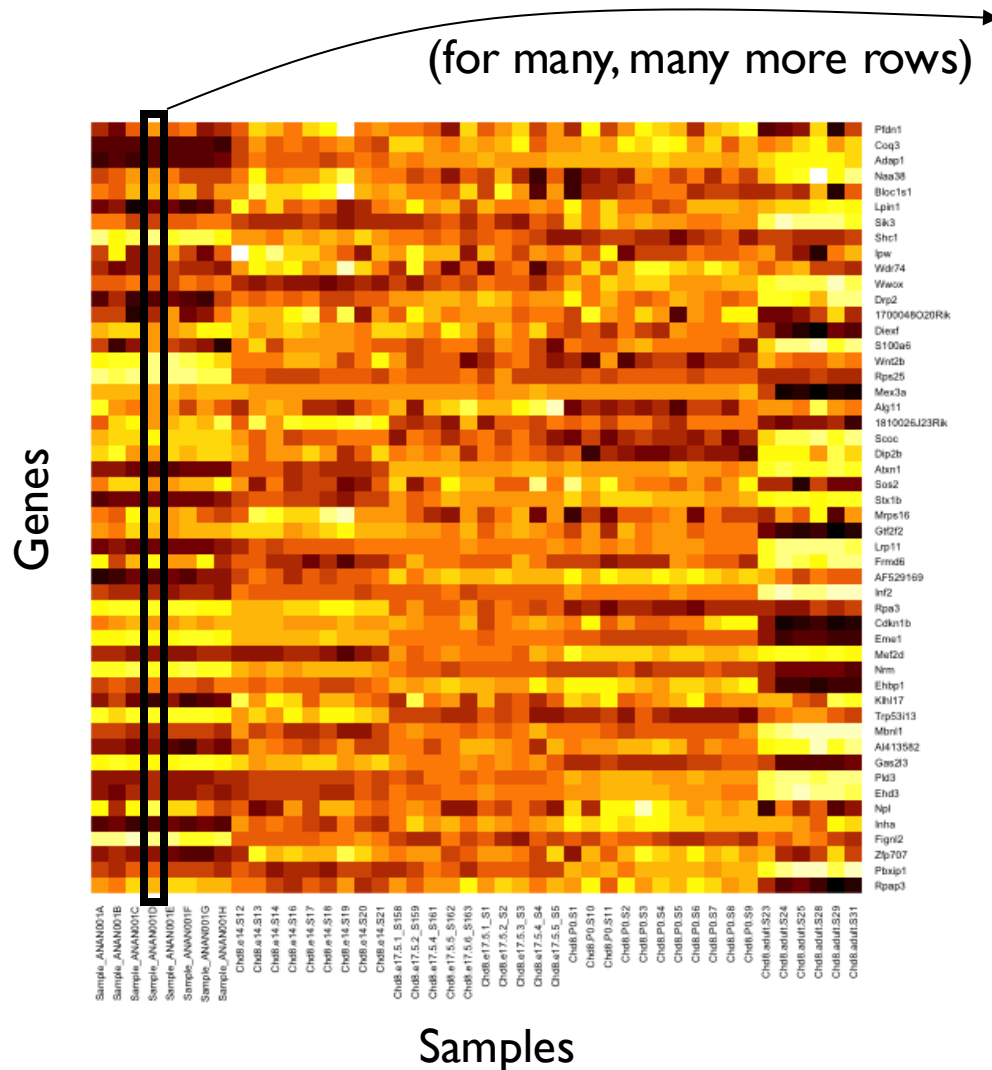
- Each square represents the expression level of one gene in one sample
- In this scheme, lighter colours mean higher levels of gene expression (“activity”)
- Only a portion of the data is shown

Profile for a single gene



- The order of columns (samples) is not necessarily important – we can rearrange without changing 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

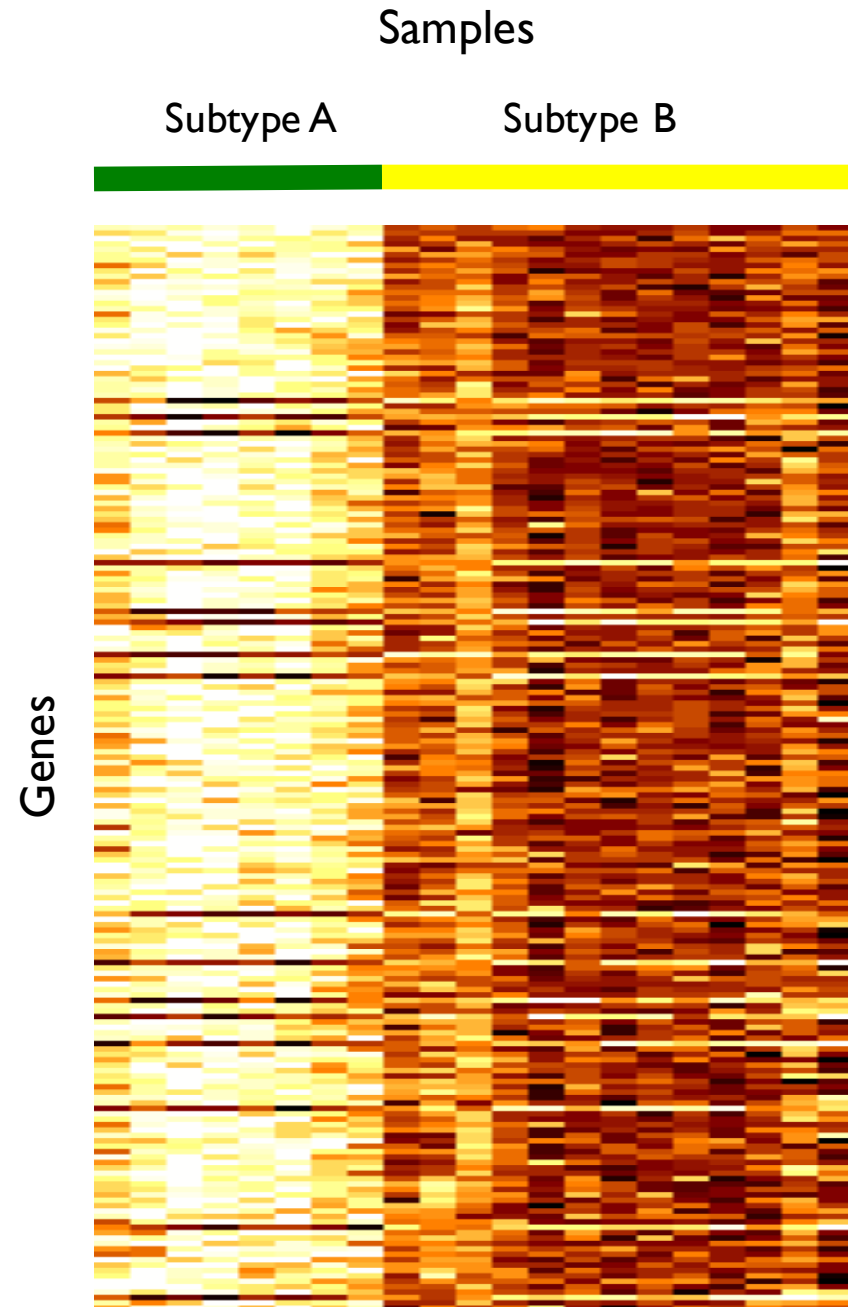


- (We usually wouldn't plot 12000 points like this)
- The “shape” (bumps, dips) of this isn't particularly meaningful – suggests the rows aren't in a completely random order in this example.
- Could rearrange the order of rows without changing the data (of course keeping track of which row is which)
 - Ordering by genomic position may be meaningful for other data types with spatially correlated 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

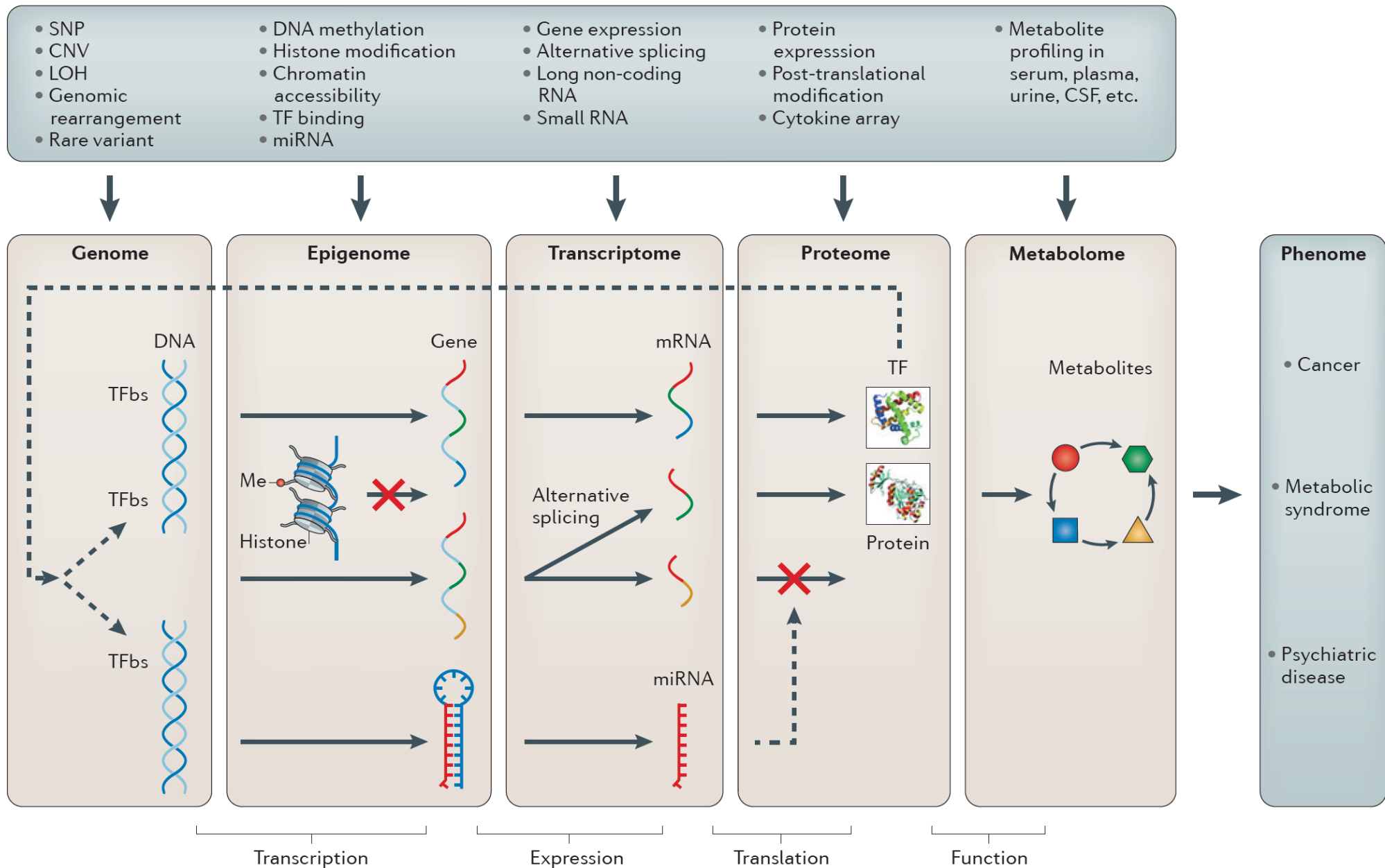
What's the big deal?



A few pitfalls and challenges

(why we need this course)

- Signals can be low: data are noisy with finite sensitivity
- False negatives are often a given, and false positives are a major danger
- Need to address confounders, 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 (multi-modal)



High-dimensional assays

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

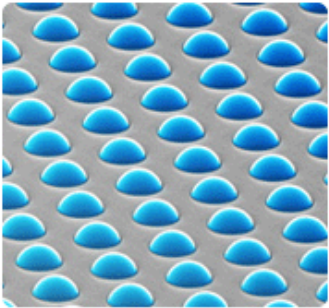
Target	Microarray/chip-based	Sequencing-based
DNA genotyping	SNP array	Targeted, exome, or whole genome sequencing
DNA methylation quantification	Methylation array	Bisulfite sequencing, methyl-binding enrichment sequencing
DNA copy number variation	Comparative genomic hybridization array	Targeted, exome, or whole genome sequencing
RNA quantification	Expression array	RNA sequencing, single-cell/nucleus RNA sequencing
DNA-protein binding	Chromatin Immunoprecipitation Chip (ChIP-chip)	Chromatin Immunoprecipitation sequencing (ChIP-seq), Cleavage Under Targets and Tagmentation sequencing (CUT&Tag)

⋮

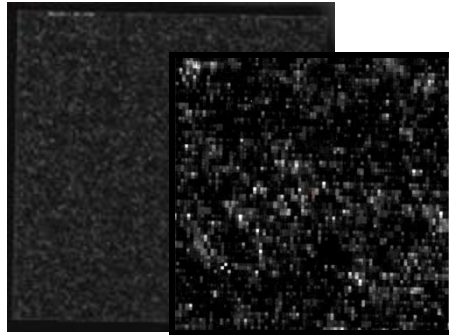
Microarray basics

Microarrays

Illumina Beadarray



Affymetrix Genechip



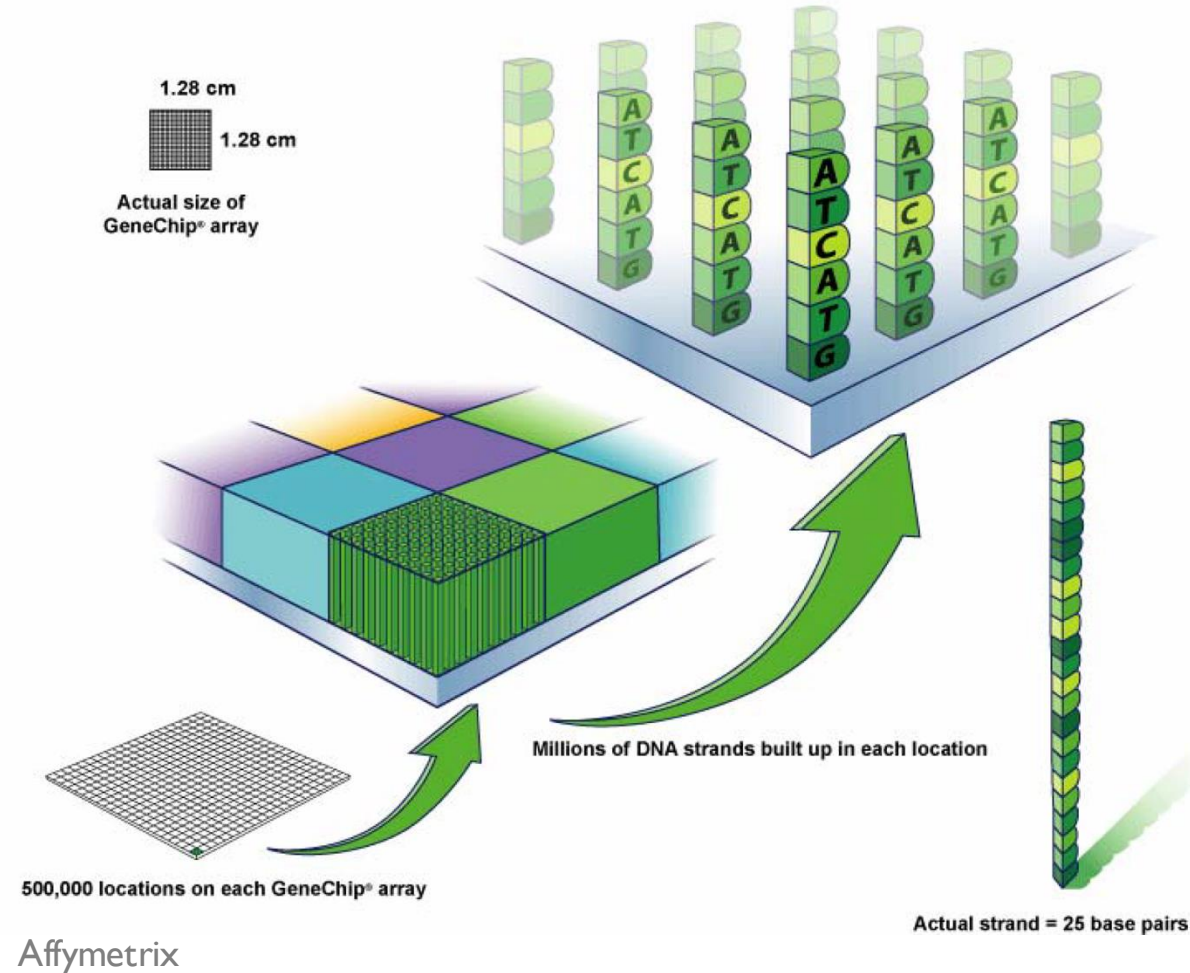
Based on fundamental principles of assay-by-hybridization developed ~1970s

- Substrate glass slide contains specific short DNA “probes”
- Each spot has only one probe sequence (mapping of location to probe is known)
- Start with a small amount of RNA or DNA purified from a sample (typically requiring many cells)
- 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

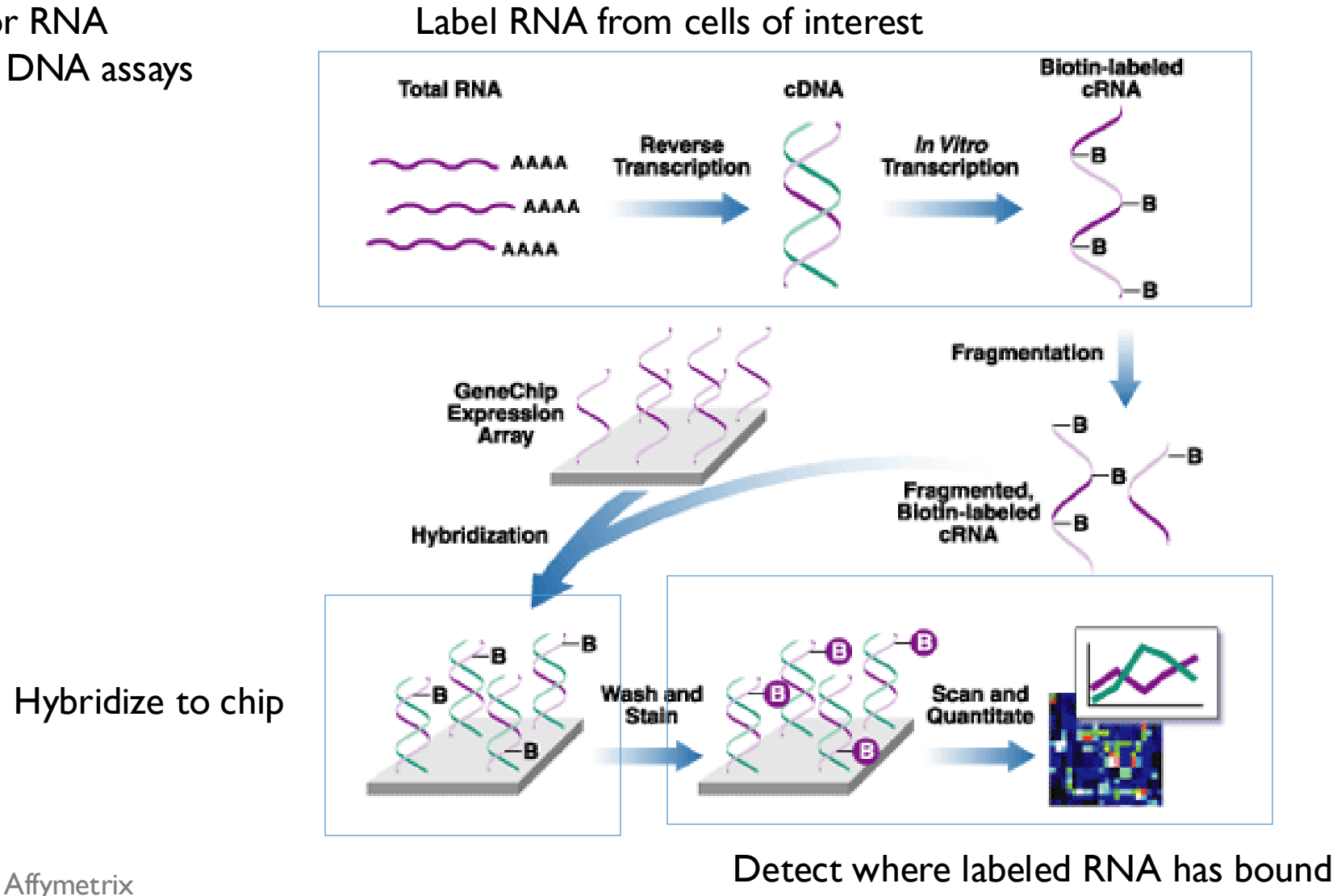
Example: Affymetrix

- ~Thousands of 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)



Detection with Hybridization (Affymetrix)

Illustration is for RNA
Similar idea for DNA assays

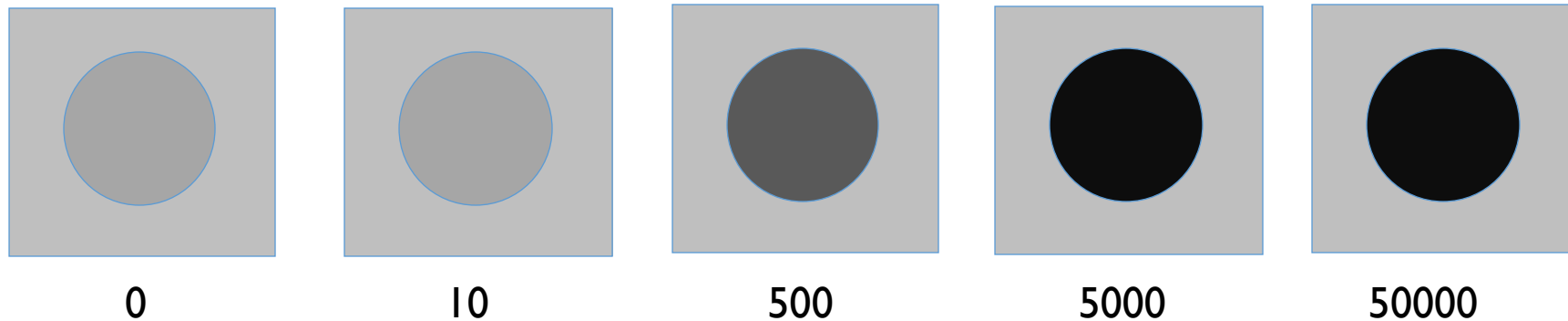


Pros and cons of hybridization-based assays

Pros: Mature, inexpensive, small data files

Cons:

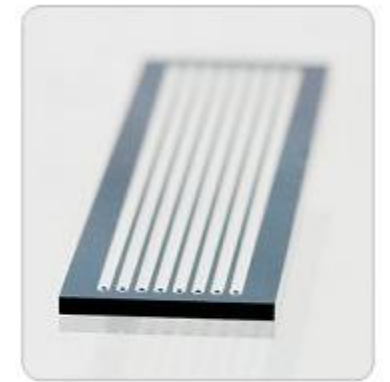
- Only detect what you probe, so can't use on an “uncharacterized genome”
- Cross-hybridization → background + ambiguity
- Dynamic range further limited by saturation



Sequencing basics

Sequencing-based assays

- Start with a small amount of purified RNA or DNA
- Amplify (e.g. create many copies via PCR) to get enough starting material
- Determine nucleotide sequences from the sample
- Until ~2008 this was not feasible compared to hybridization approaches
- Most commonly “shotgun” strategy (e.g. Illumina)
 - randomly sample millions of short pieces (100s bps) of the input DNA/RNA and sequence them
 - Then figure out what the pieces are/put them back together
- Long read approaches have also been developed (e.g. nanopore ~ 10s of thousands of bps)



Illumina device and sample flowcell

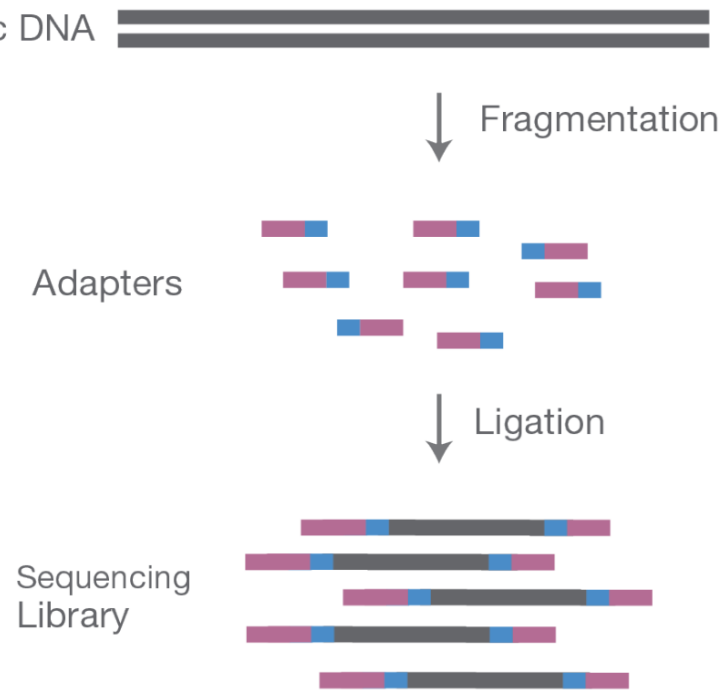
Sequencing coverage/depth

- 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)
 - due to the random sampling of which targets get sequenced
- Example – if you want to sequence a 3Gb genome at “30x”, you need to sequence 90Gb of DNA
 - With 100bp reads, that’s 900 million reads

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 a fragment, examine fragments base by base
- Up to billions of reads per run on one machine
- Sequences ~150-300bp (x 2 for paired ends)
- Per-base sequence error rate varies (typically <1%) [\[ref\]](#)

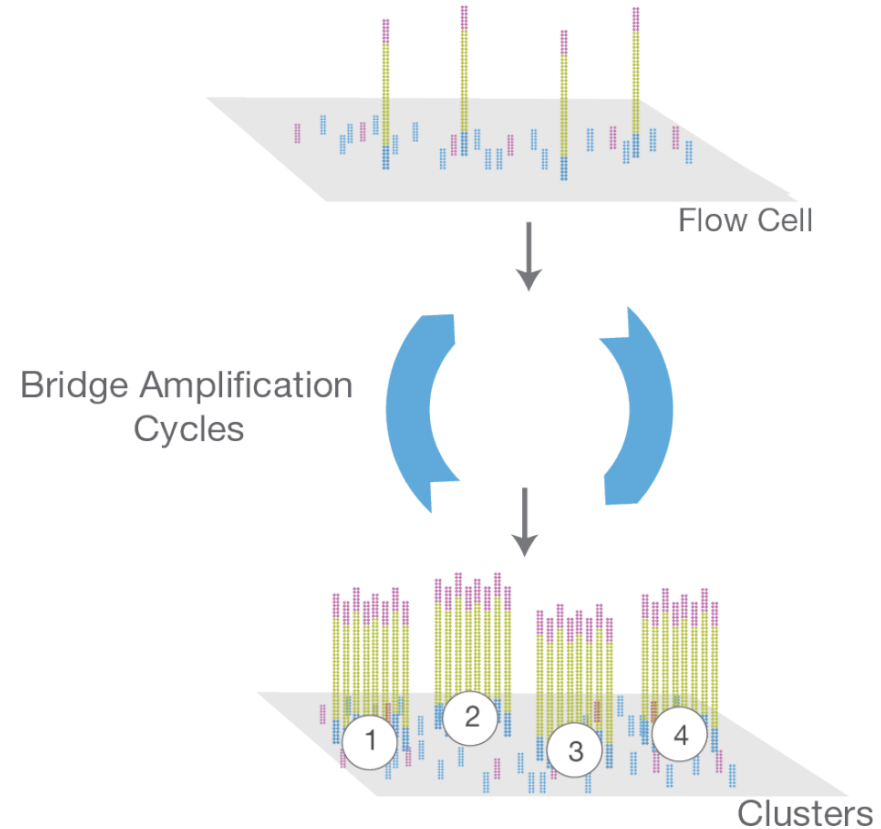
A. Library Preparation



Optionally: PCR amplification
Size Selection

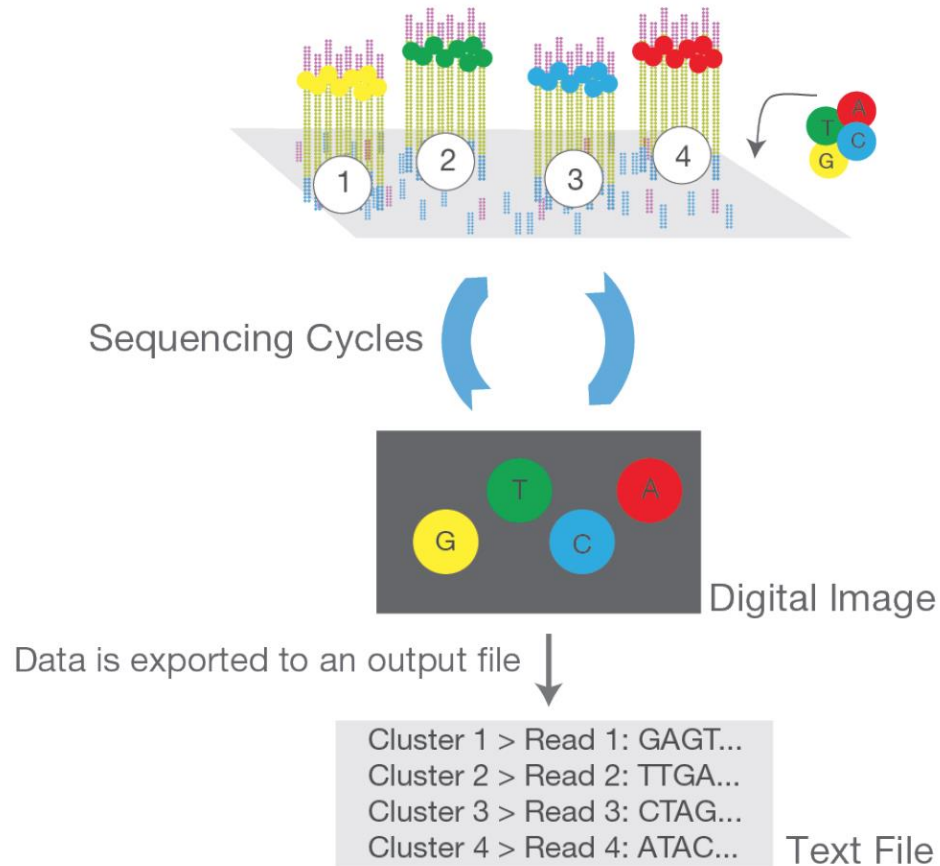
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification



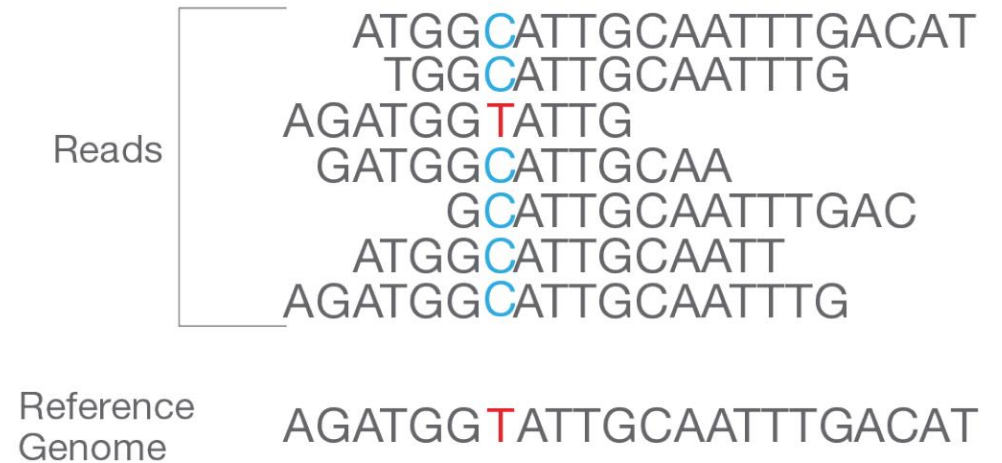
Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

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 Analysis



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.

Throughput of some Illumina platforms

	     					
Key specifications	iSeq 100 System	MiniSeq System	MiSeq System	MiSeq i100 Series^a	NextSeq 550 System	NextSeq 1000 and 2000 Systems
Max output per flow cell	1.2 Gb ^b	7.5 Gb ^c	15 Gb ^d	30 Gb ^a	120 Gb ^c	540 Gb ^e
Run time (range) ^e	~9.5–19 hr	~5–24 hr	~5.5–56 hr	~4–15.5 hr	~11–29 hr	~8–44 hr
Max reads per run (single reads)	4M ^{ab}	25M ^c	25M ^d	100M ^a	400M ^c	1.8B ^e
Max read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp

- 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

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. difficult to detect/quantify different transcripts from a gene
- e.g. difficult to isolate 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 ($\gg 1$ kb)
- Downside: higher error rates (not as good for genotyping), lower throughput (not as good for quantification)

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

Raw sequence data: FASTQ format

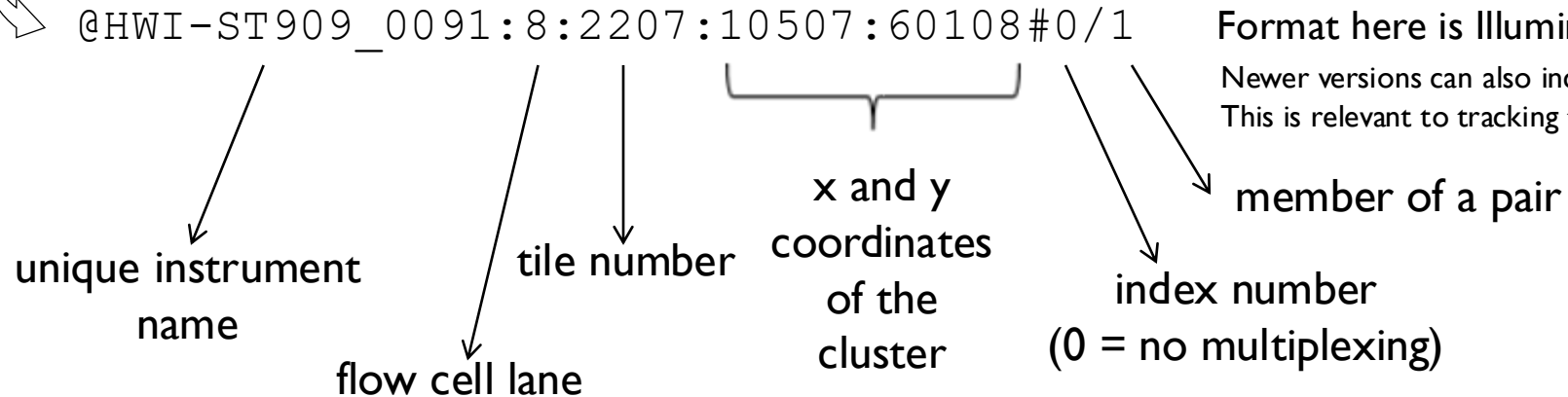
Text file; Sequences with quality information; ~60GB file for 500M paired-end reads, 100bp

Data for one Illumina read:

```
@HWI-ST909_0091:8:2207:10507:60108#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (**+)) %%%++) (%%%) .1***-+*'') ) **55CCF>>>>>CCCCCCC65
```

Read identifier
Sequence

Base quality



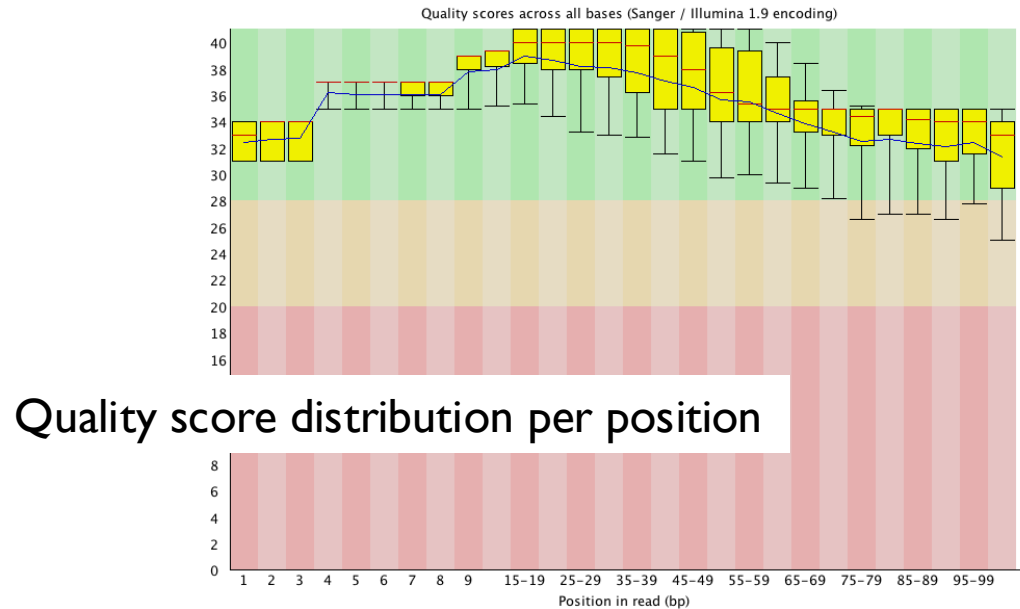
Format here is Illumina-specific (old-style)

Newer versions can also include a run and flowcell ID

This is relevant to tracking technical sources of variability i.e. batches

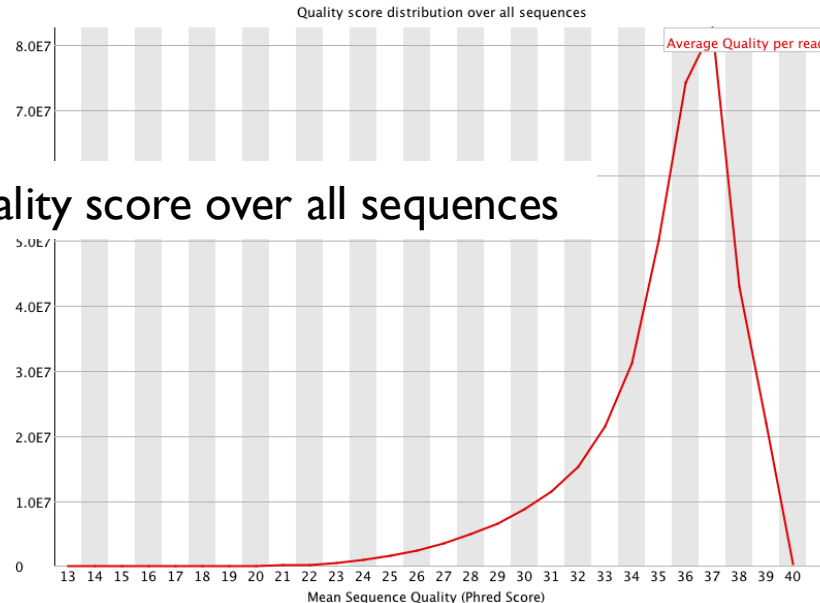
Base qualities (Q) are interpreted as ASCII byte from 33 (new-style). Read as values from 0 so “!” = 0; “C” = 34
Interpretation: $P(\text{base call is wrong}) = 10^{-Q/10}$ (AKA “Phred”)

Sequence quality: FASTQC software



Phred quality Q20 = 1 in 100
prob of being incorrect

Mean quality score over all sequences



Also monitored for quality control:

- GC content
- sequence length distribution
- sequence duplication level

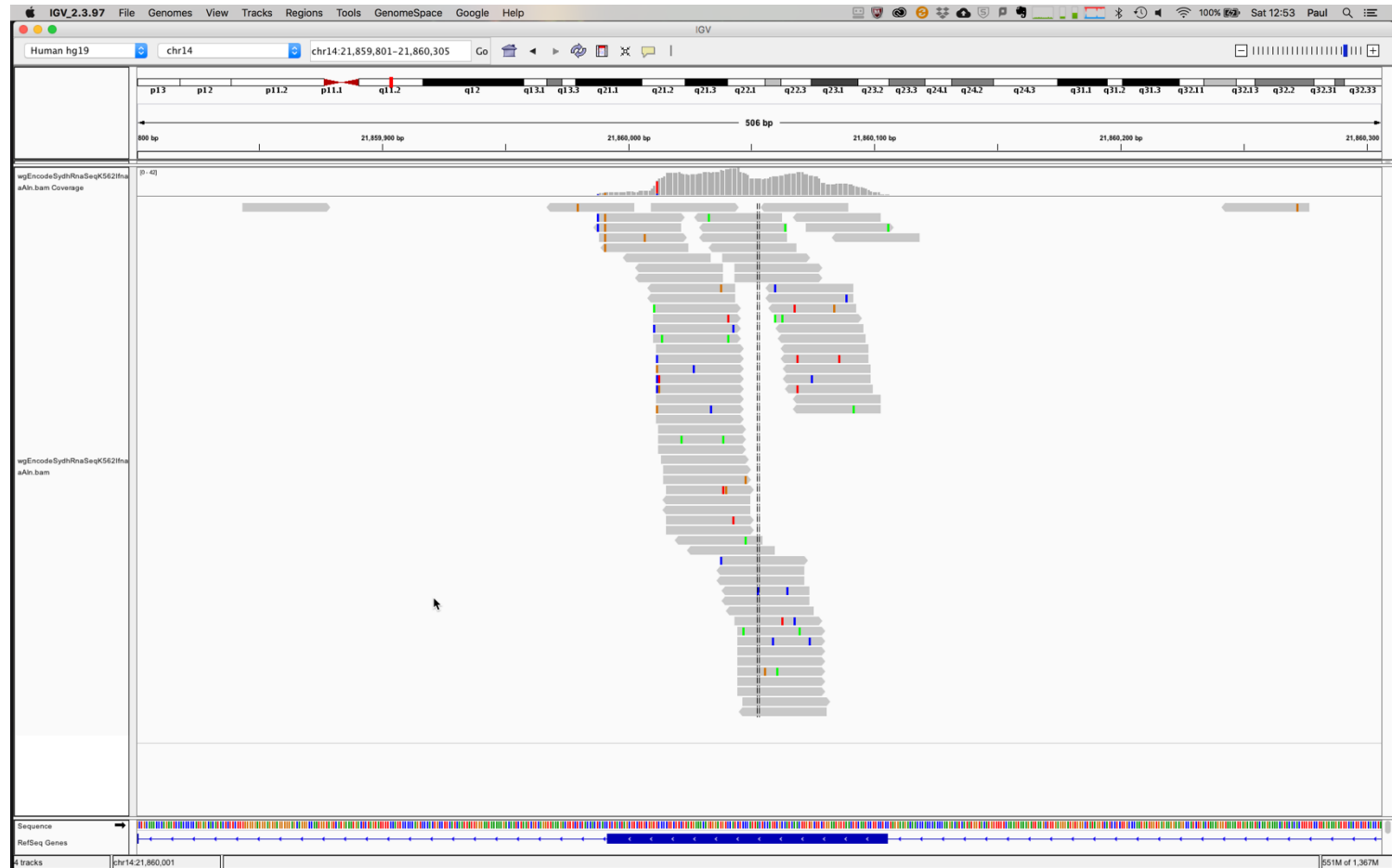
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

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)

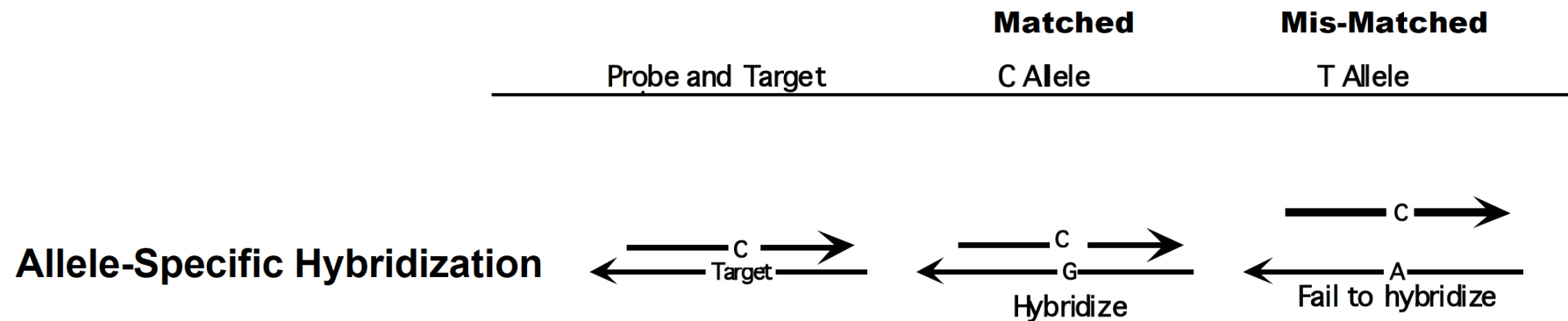
Part II: Specific assays

- 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 often cheaper
 - If higher resolution is desired/needed, sequencing is often the answer (if budget allows)

SNPs (Single Nucleotide Polymorphisms)

SNP detection with microarrays

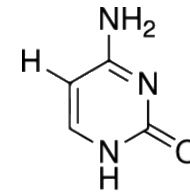
- Recall: SNPs (Single Nucleotide Polymorphisms) are a class of common single base differences in DNA
- Because the polymorphic sites are (mostly) known in some well-studied organisms / populations, can use microarrays of pre-designed probes covering a set of SNP sites
- Arrays currently assay ~1+ million SNPs; Imputation used to infer millions more
- 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



DNA Methylation

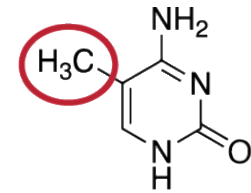
DNA methylation

- 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 CpG is either methylated or not – in a diploid cell, can have 0, 1, or 2 methylated*
- 60-80% of mammalian CpGs are generally methylated

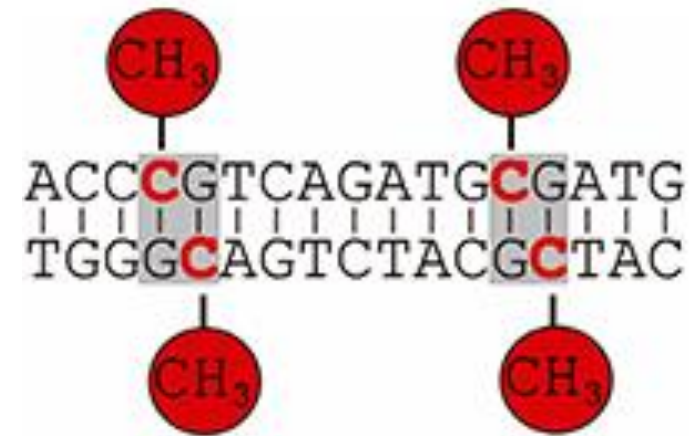


Cytosine

Wikipedia



methylated Cytosine



<http://www.mpipsykl.mpg.de>

Role of DNA methylation

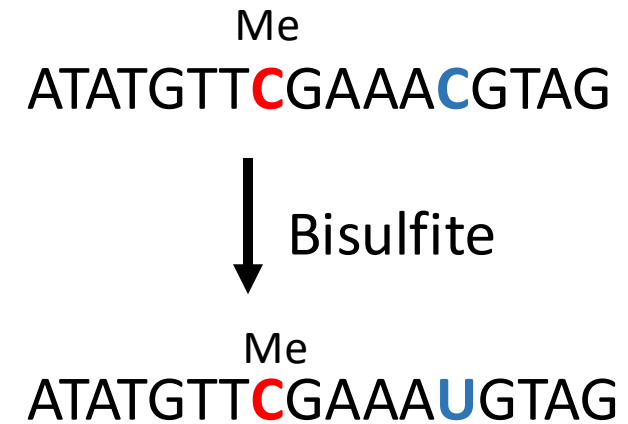
- Methylation is classically associated with gene silencing
 - often via DNAm-binding proteins that result in changes in chromatin state
- DNAm is relatively stable, but dynamics are of interest
 - Long-lasting “epigenetic states” partly reflected in DNA methylation (e.g. cell type-specific patterns, aging patterns)
 - Can also be influenced by environment
- Goal: survey genome-wide methylation states for populations of cells in different individuals/conditions/cell types

DNAme assay types

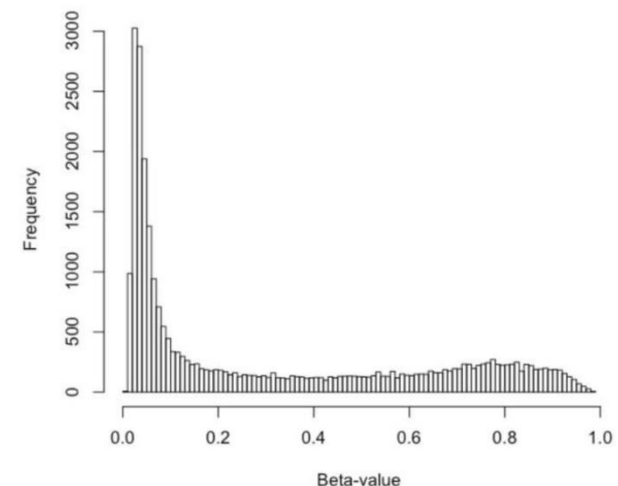
- Human genome contains ~28 million CpGs
- Whole Genome Bisulfite Sequencing (WGBS) can assay all of them
 - but inefficient; many reads (~65%) will not contain a CpG
- Reduced-representation bisulfite seq (RRBS) is more efficient
 - use a digestion/purification step to enrich sample for CpG-dense DNA
 - ~2 million CpGs assayed (~1% of genome)
- Microarray: Illumina beadarray platforms contain probes designed to distinguish mC vs. C at selected CpGs
 - Older platform: “450k” ~485k sites; newer “EPIC” >860k sites
- Other approaches: affinity enrichment (e.g. MeDIP)
 - enrich for methylated DNA; hybridize on array or sequence

Assaying DNA Methylation

- Take advantage of distinct chemical properties of mC
- Bisulfite conversion: chemical treatment turns Cs into Us unless they are methylated
- Compare bisulfite-converted & standard sequence
 - Microarray: probes distinguish these cases for a particular set of pre-selected sites
 - Sequencing: U will be read as T – requires some modification to aligners
- Output for each site: “ β ” (beta value)
 - Arrays: relative intensity of methylated vs unmethylated probes
 - Sequencing: the fraction of reads methylated



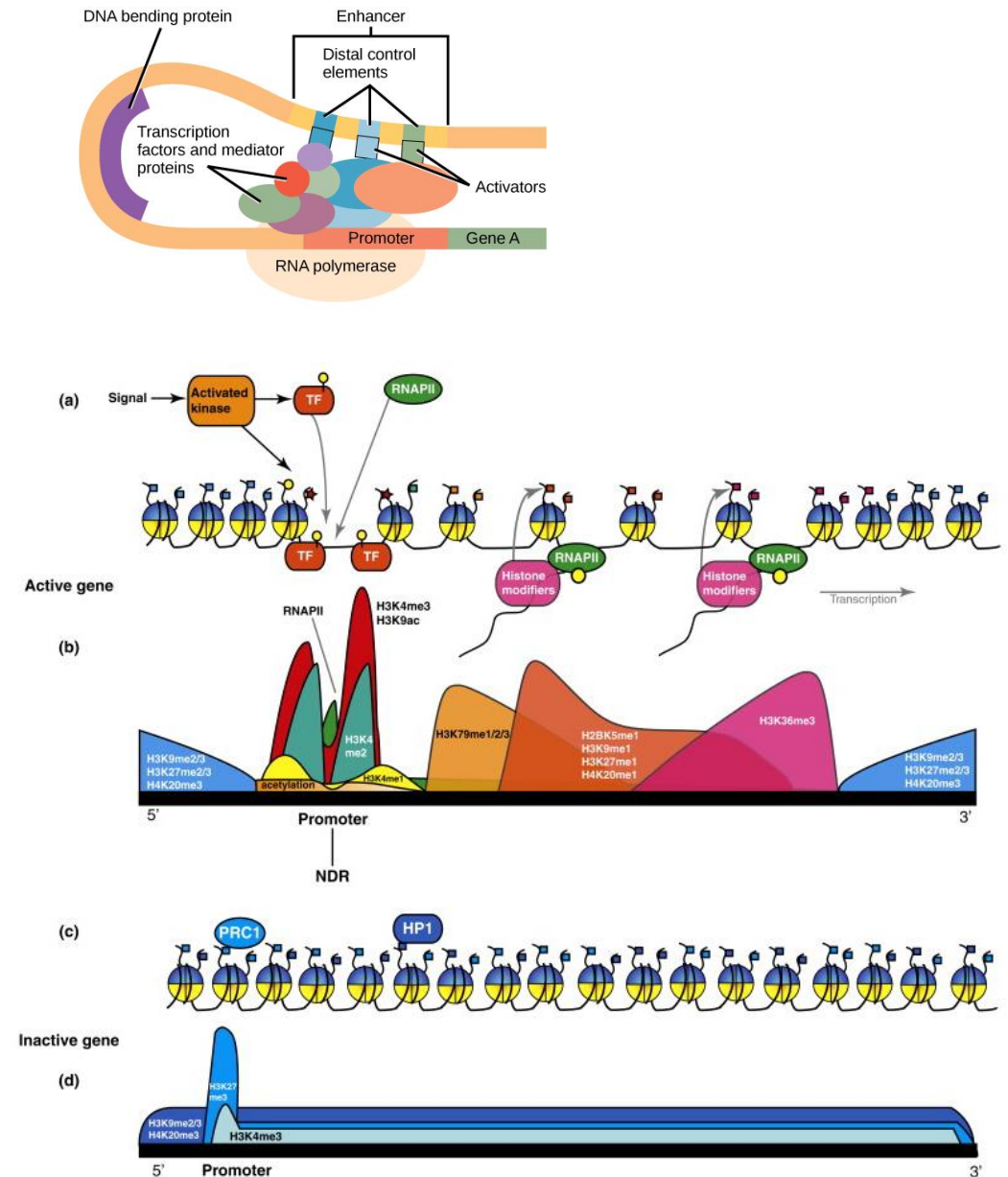
Not shown: other strand of DNA



DNA-binding proteins

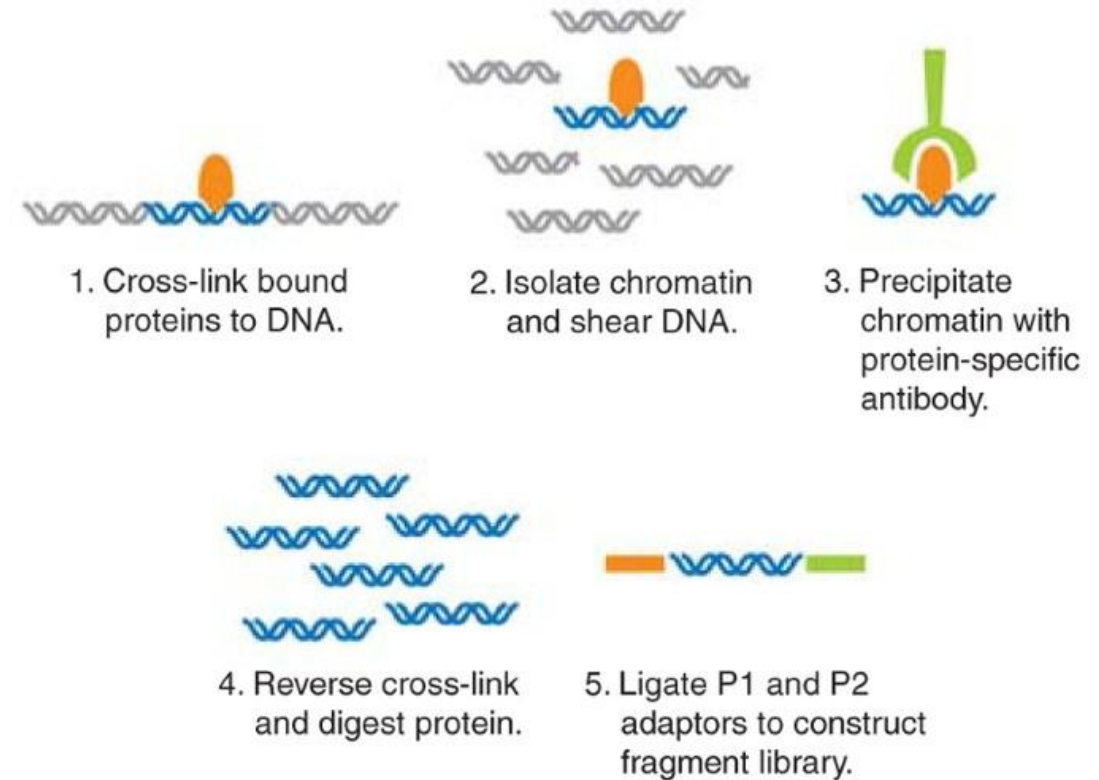
ChIP-seq

- Sequencing of DNA fragments that are bound by DNA-binding proteins
 - Transcription factors
 - RNA polymerase
 - Histone modifications / enzymes
- Two steps: chromatin immunoprecipitation (ChIP) and sequencing (seq)
 - Key reagent is a specific antibody against the protein you want to study (next slide)
- 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)



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 library
6. Sequencing

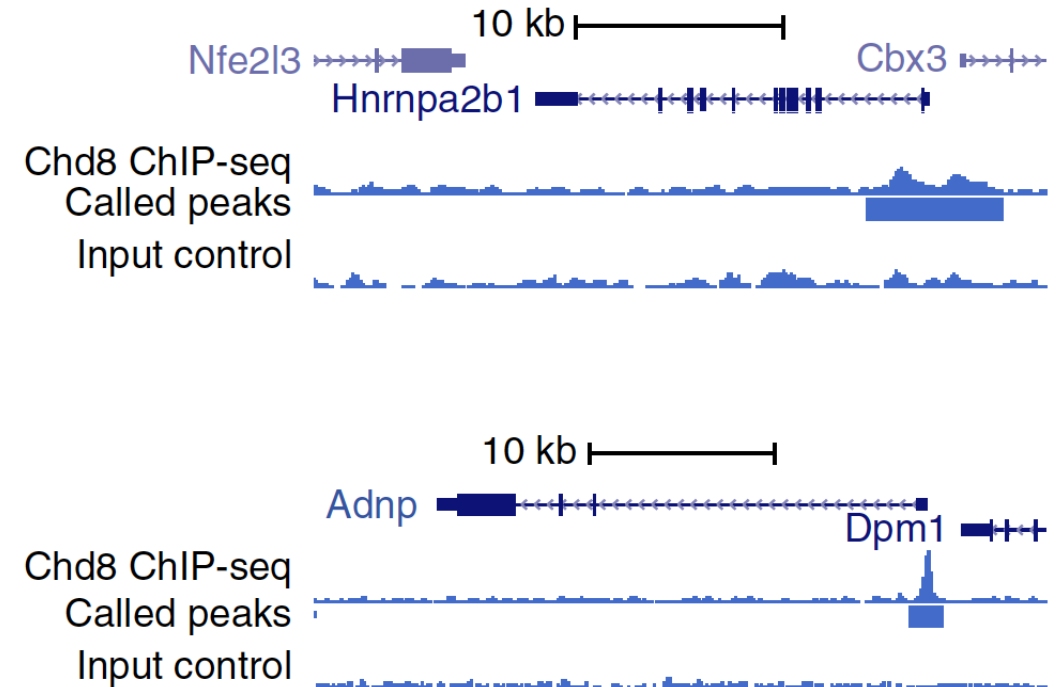


<https://www.nature.com/articles/nmeth.f.247>

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

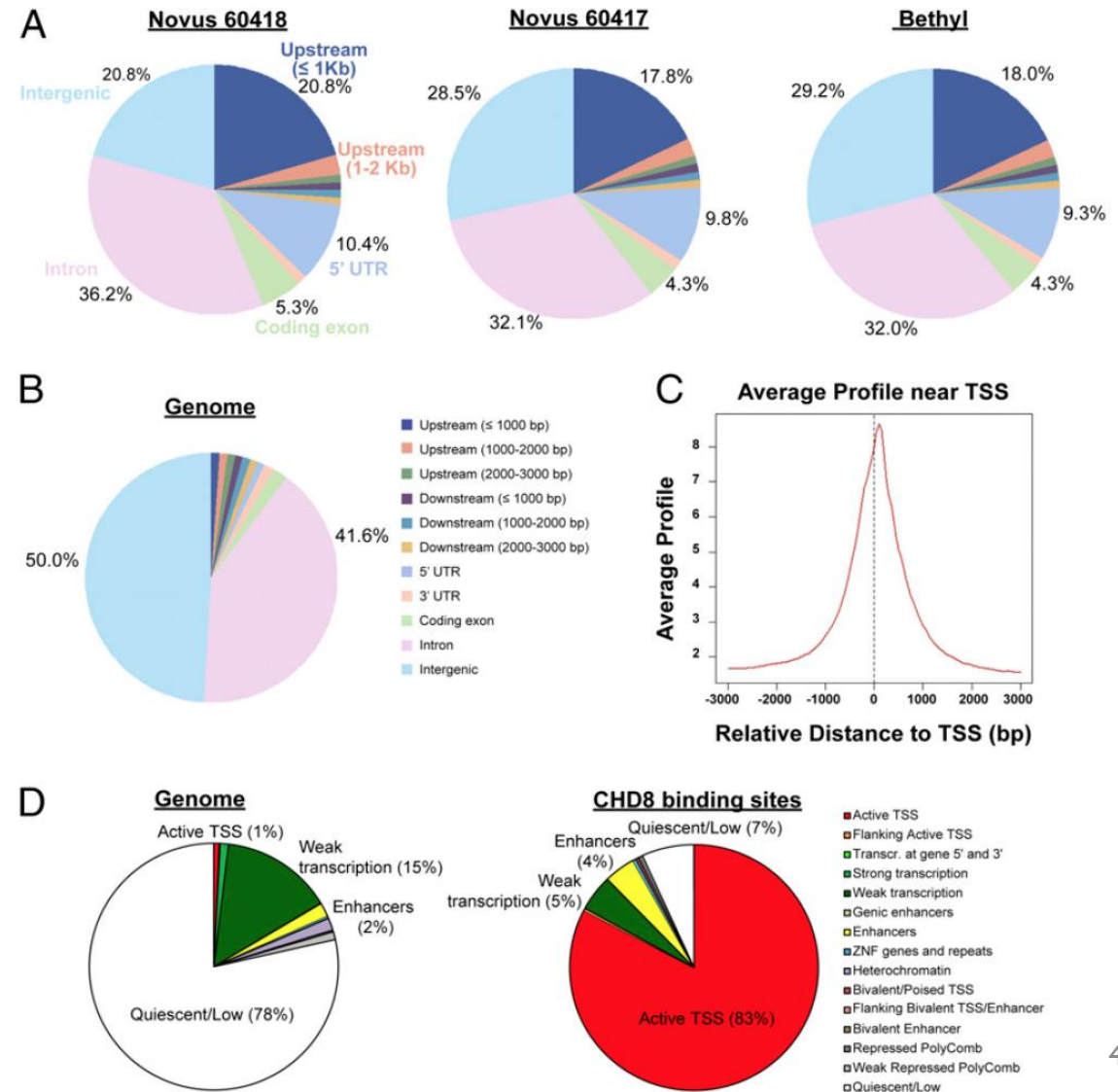
- Aligning sequence reads with short-read aligners
- Quantifying binding - peak finding
 - Goal: identify real peaks; estimate confidence
 - Basic idea: count the number of reads in windows and determine whether this number is above background – if so, define that region as bound
 - Many algorithms (>60)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



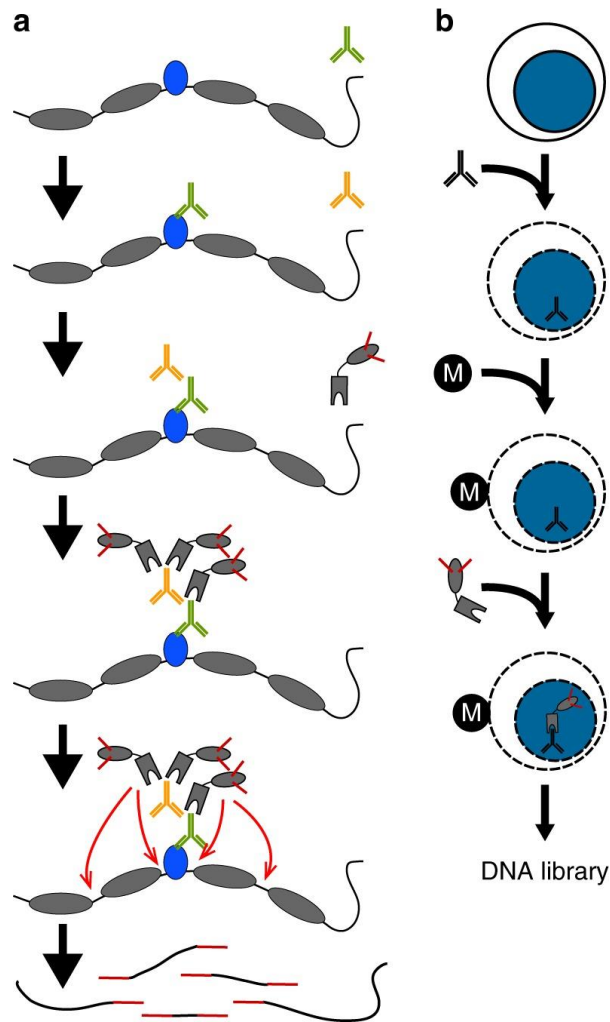
Gompers et al. 2017

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”



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
- 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

Transcriptomes (Expression)

Assaying RNA expression (transcriptomics)

- Quantitative measure of expression level of each gene
 - 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") gives insight into the process/disease/condition of interest
- The "readout" of genetic variation is partly in gene expression – it can help us understand the link between genetics and phenotypes

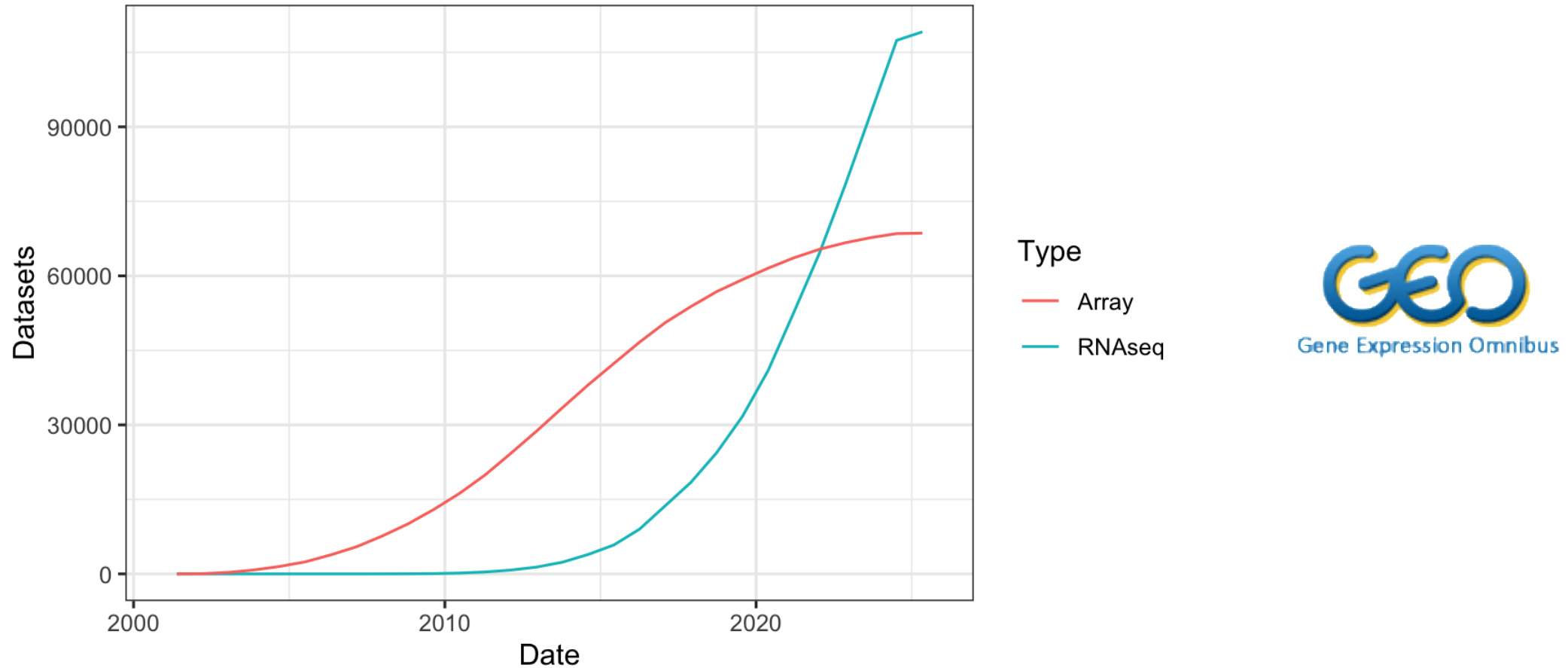
Choice of method: the move to RNA-seq

Microarrays are still used, but RNA-seq is now cheap enough to be the dominant method. 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 2025: Total of **109,121** RNA-seq and **68,597** microarray studies

RNA-seq general approach

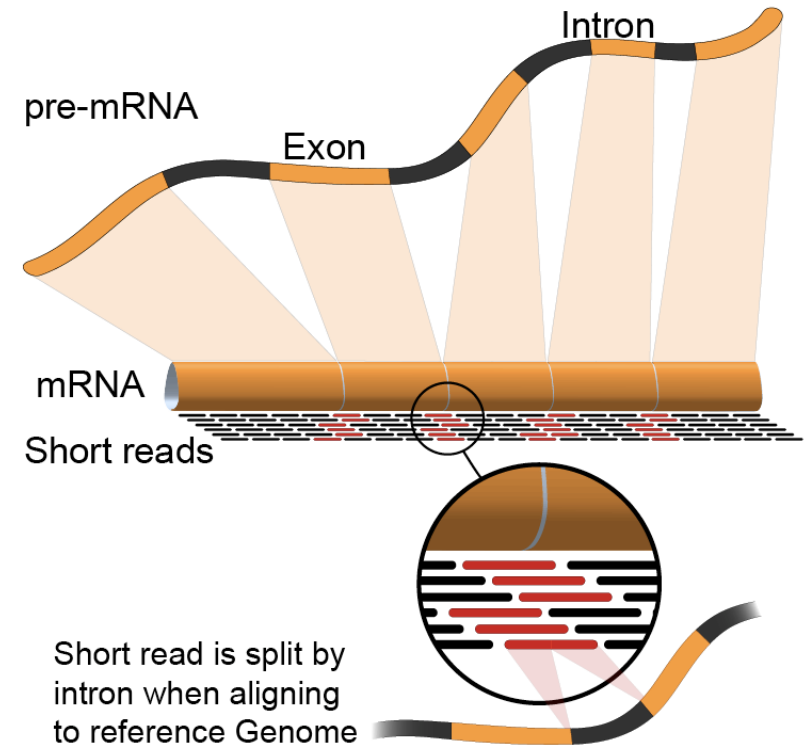
- Randomly sample molecules from your RNA sample
- Sequence their end(s)
 - Typically obtain ~20M paired-end reads for each sample for quantitation
 - Multiplexing allows dozens of samples to be run at once
- Figure out what transcript/gene the reads come from

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!)

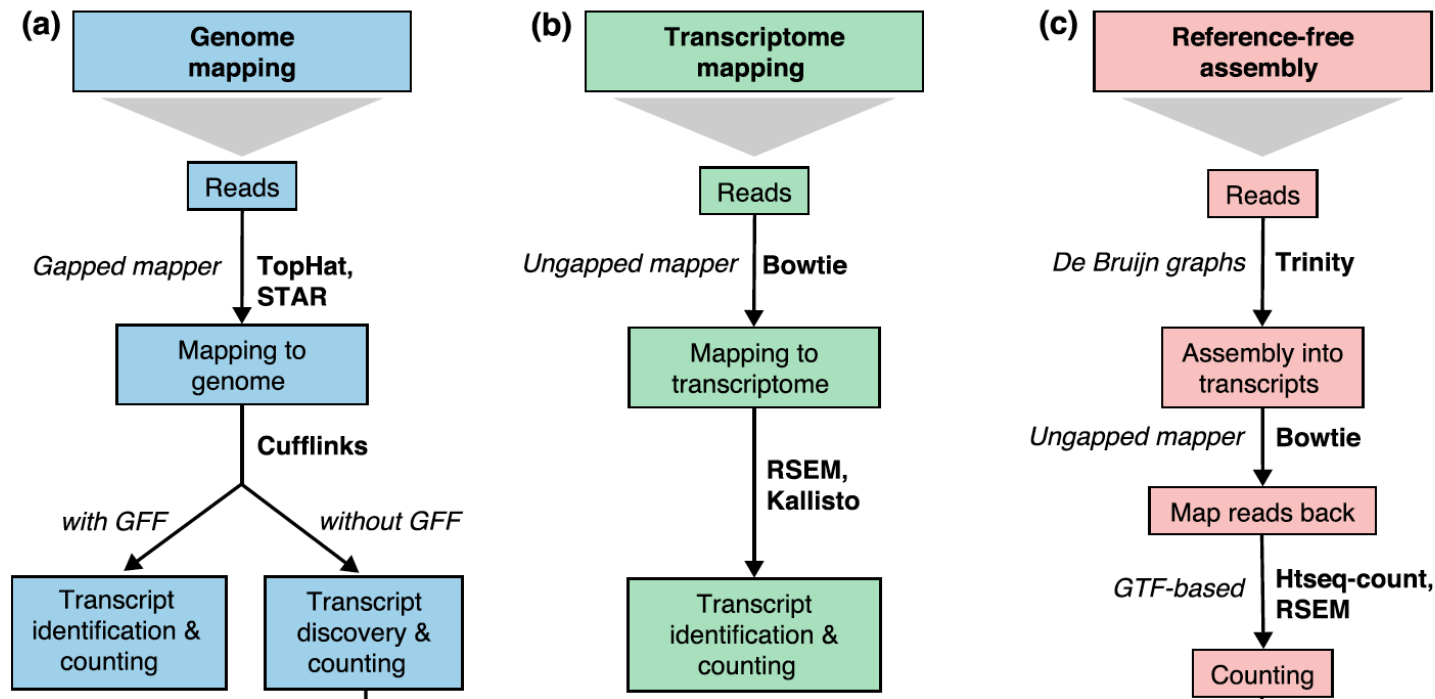
The challenge of exon junctions

- Recall that the RNA we're analyzing is already spliced
- 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 150bp.
- Many computational methods, plus alternative (long-read) sequencing approaches ...



Rodrigo Goya/Wikipedia

“Upstream” analysis of RNA-seq data



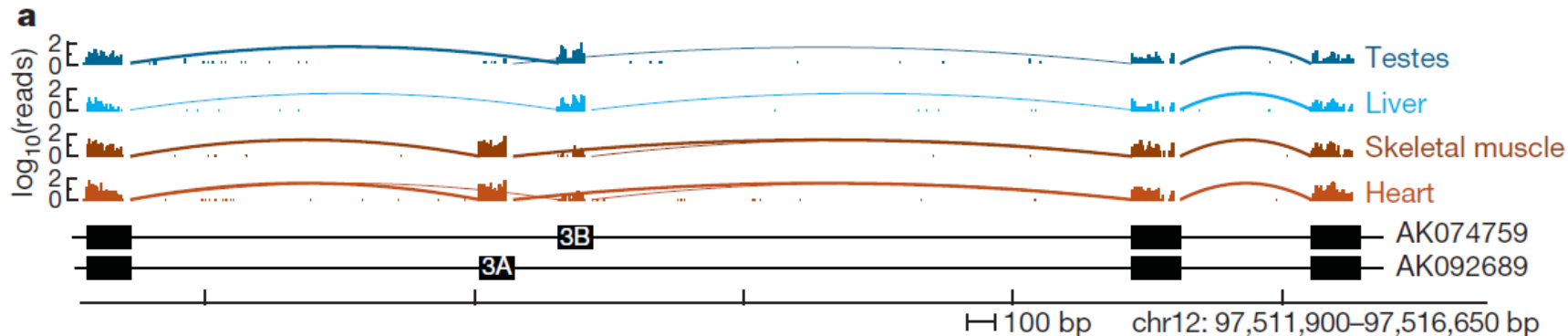
- By “upstream” I mean taking the raw RNA-seq data from raw fastq files to counts per gene or transcript
- Often have to think about a pipeline that is built from multiple tools
- Figure shows several types of pipelines, but it’s not this simple – methods can be mixed and matched, or not, to a varying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

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)

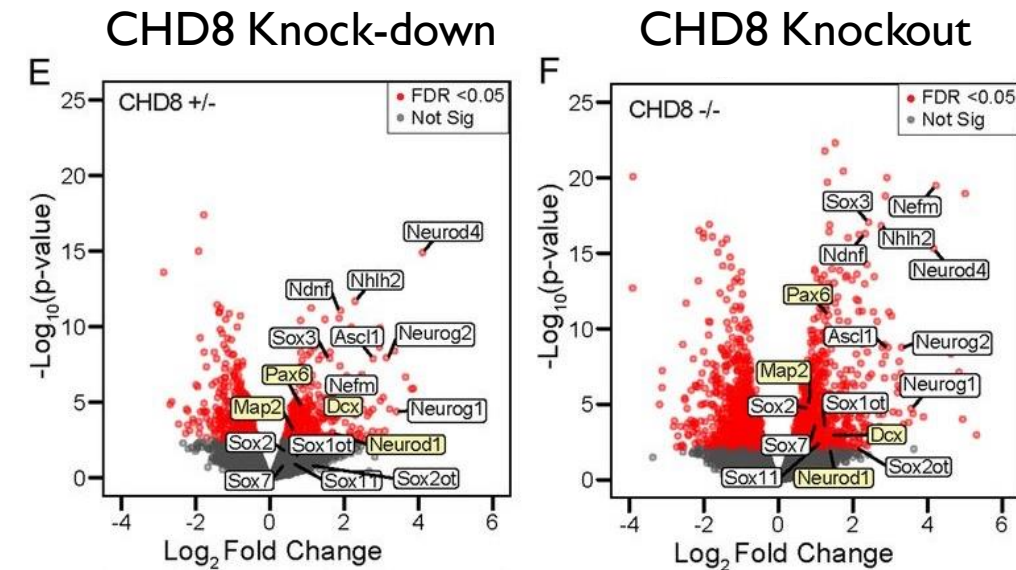


Quantification without alignment

- 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 – determine which transcripts each read is compatible with

“Downstream” analysis of RNA-seq

- Gene or transcript counts in hand, we move to “downstream” analysis: focus of this course
- Example study involving CHD8: RNA-seq comparing wild type (CHD8 +/+) mice with two mutant lines
 - WT vs CHD8 +/- (knock down; ~50% levels of CHD8)
 - WT vs CHD8 -/- (knock out; CHD8 undetectable)
- “Volcano” plots show effect size (fold change) vs 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”