Genomic technologies

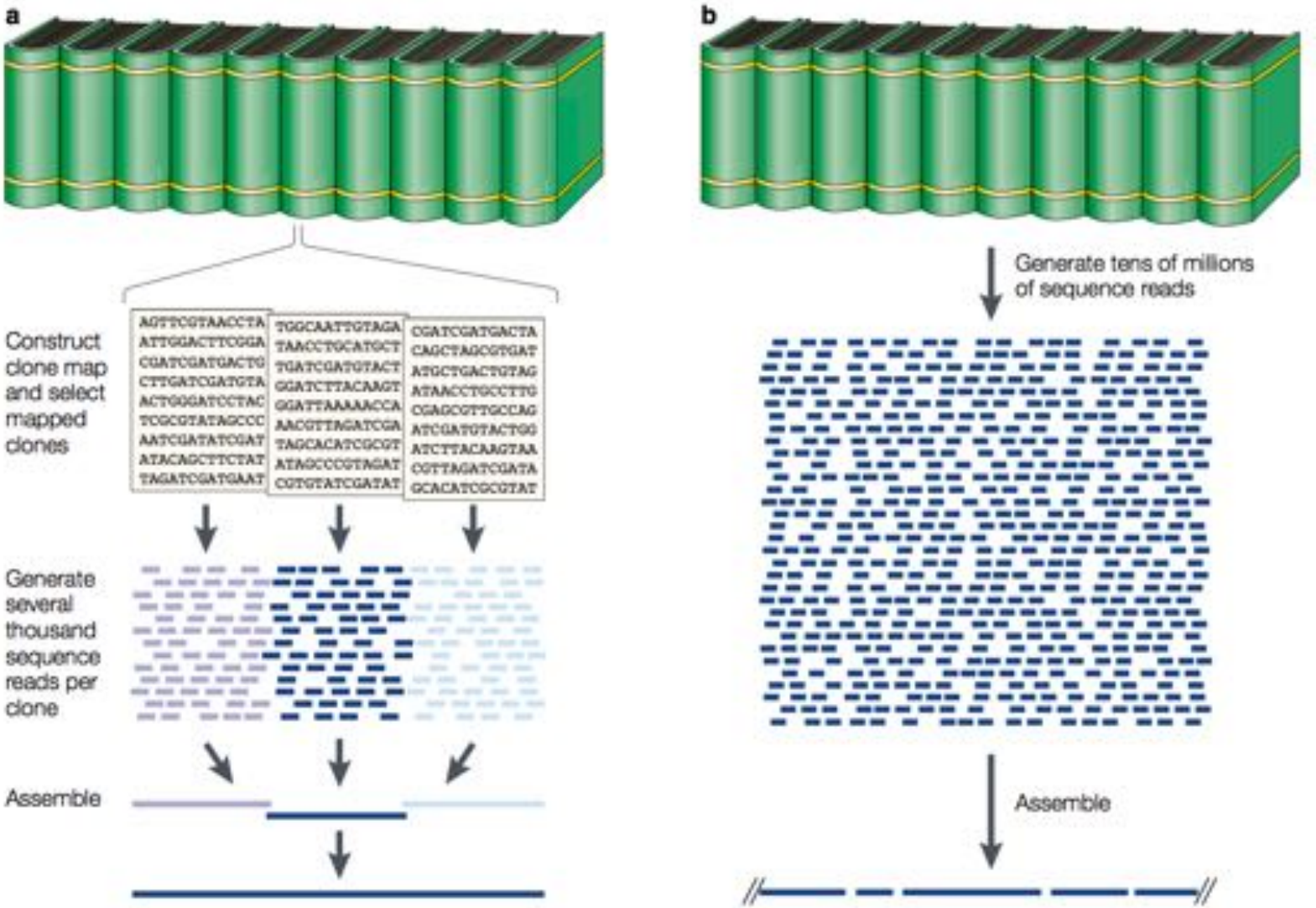
Mikhail Dozmorov mdozmorov@vcu.edu

May 15, 2019

Sequencing in a nutshell

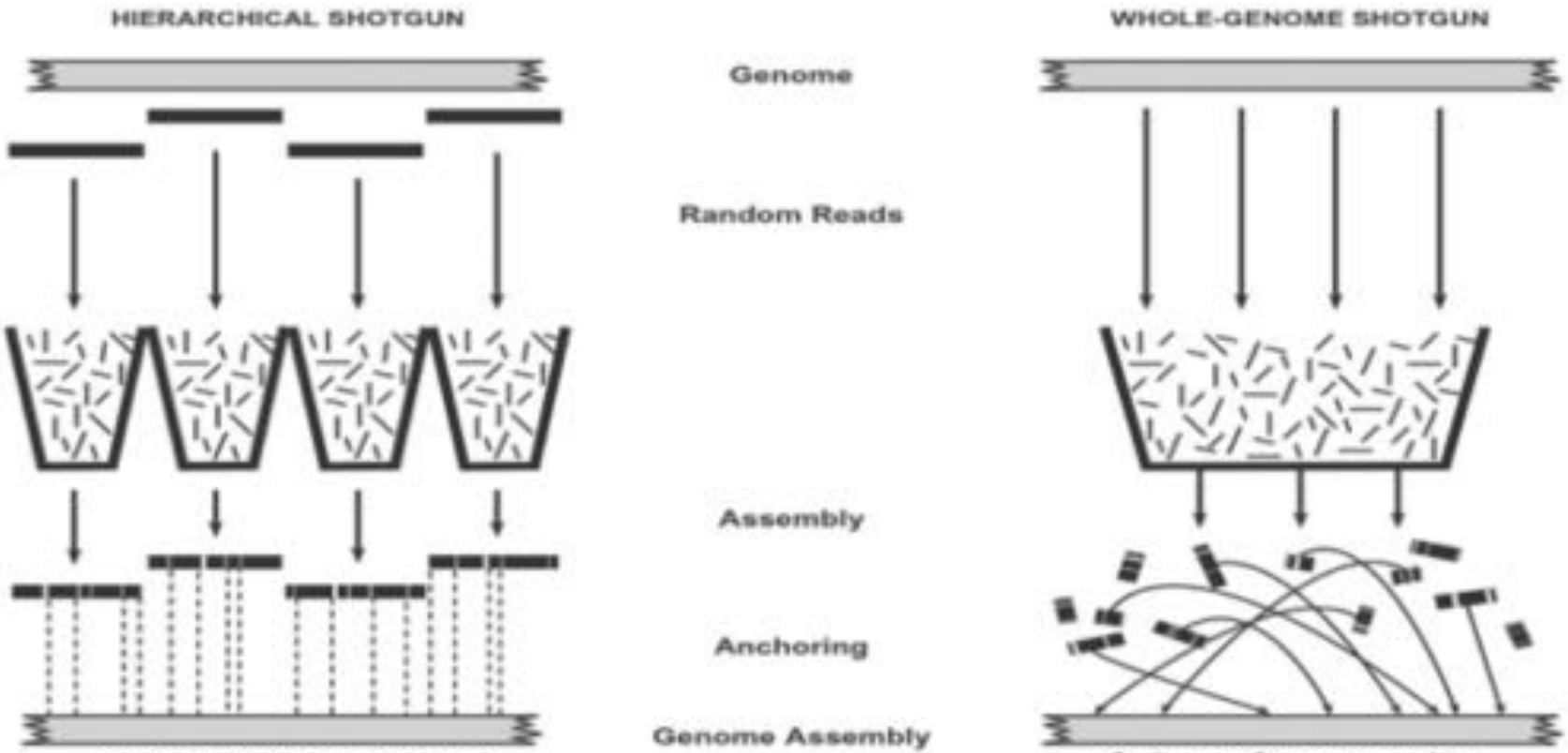
- Cut the long DNA into smaller segments (several hundreds to several thousand bases).
- Sequence each segment: start from one end and sequence along the chain, base by base.
- The process stops after a while because the noise level is too high.
- Results from sequencing are many sequence pieces. The lengths vary, usually a few thousands from Sanger, and several hundreds from NGS.
- The sequence pieces are called “reads” for NGS data.

Two shotgun-sequencing strategies



https://www.nature.com/nrg/journal/v2/n8/full/nrg0801_573a.html

The competing human genome projects



Evolution of sequencing technologies

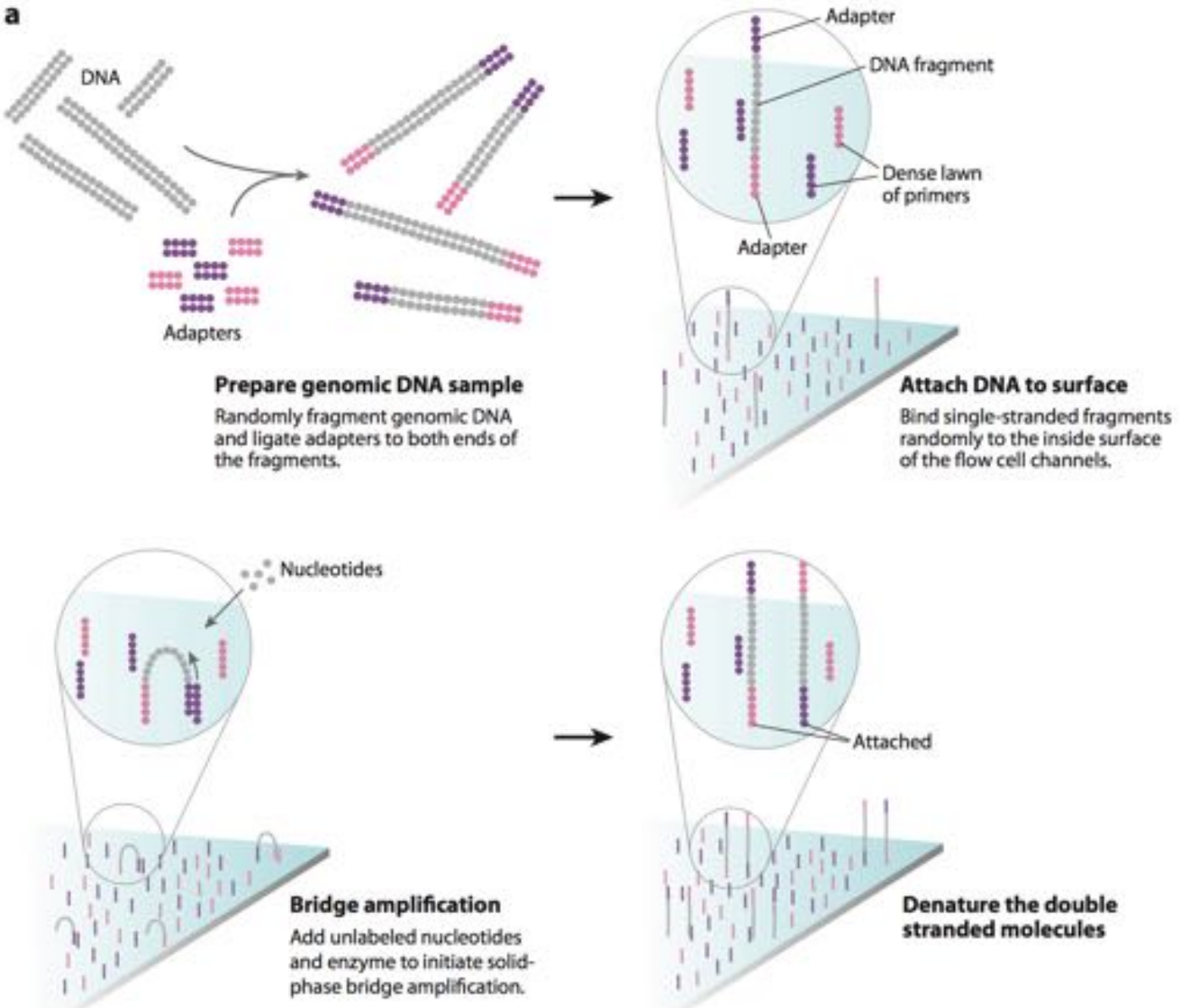
- 2006: Solexa (Illumina)
- 2010: Ion Torrent (Life Technologies)
- 2011: Pacific Biosciences
- 2015: Oxford Nanopore Technologies

Solexa (Illumina) sequencing (2006)

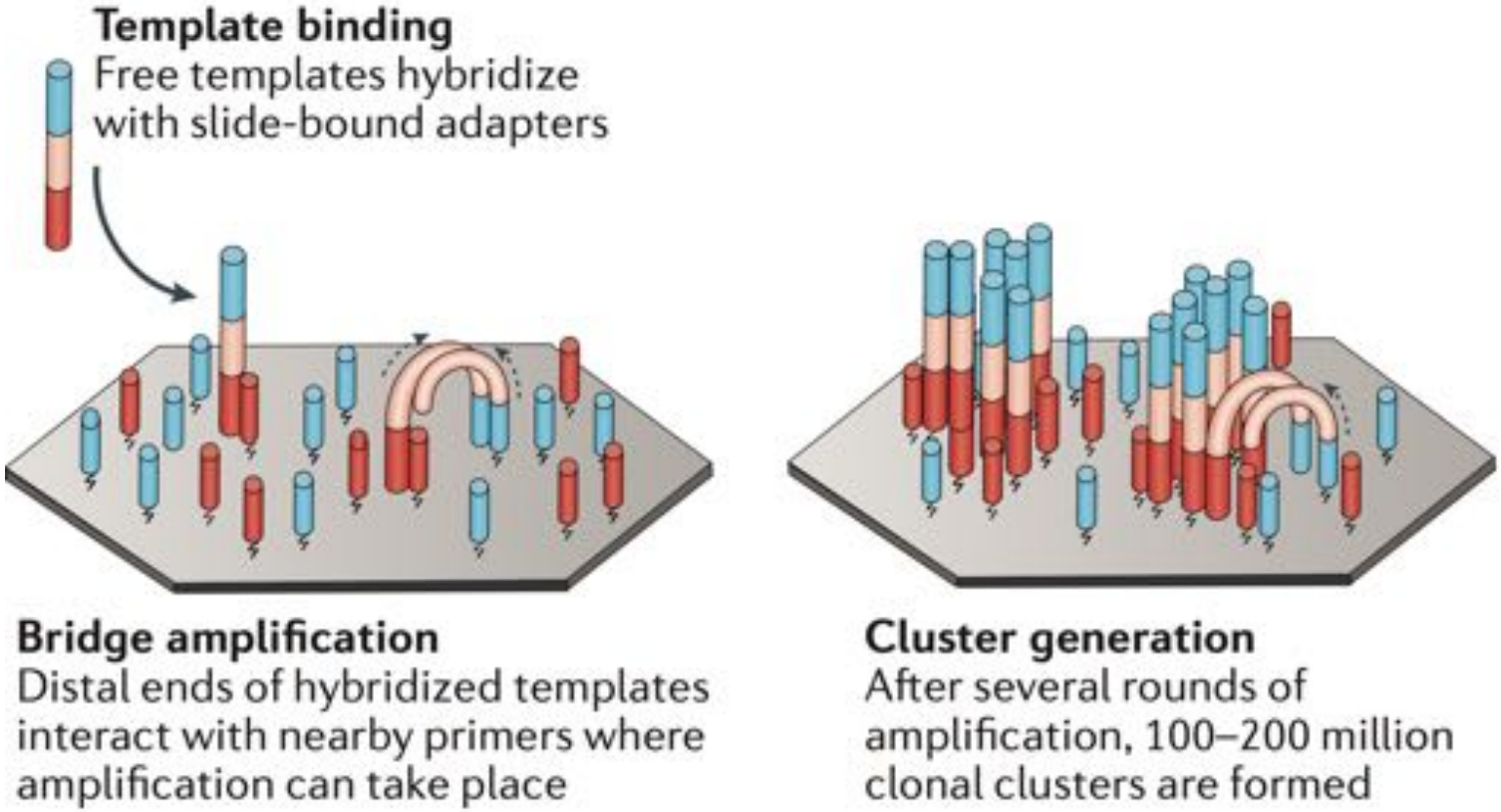
- Polymerase Chain Reaction (PCR) amplify DNA fragments
- Immobilize fragments on a solid surface, amplify
- Reversible terminator sequencing with 4 color dye-labelled nucleotides

Video of Illumina sequencing, http://www.youtube.com/watch?v=77r5p8IBwJk (1.5m),
https://www.youtube.com/watch?v=fCd6B5HRaZ8 (5m)

Solexa (Illumina) sequencing (2006)

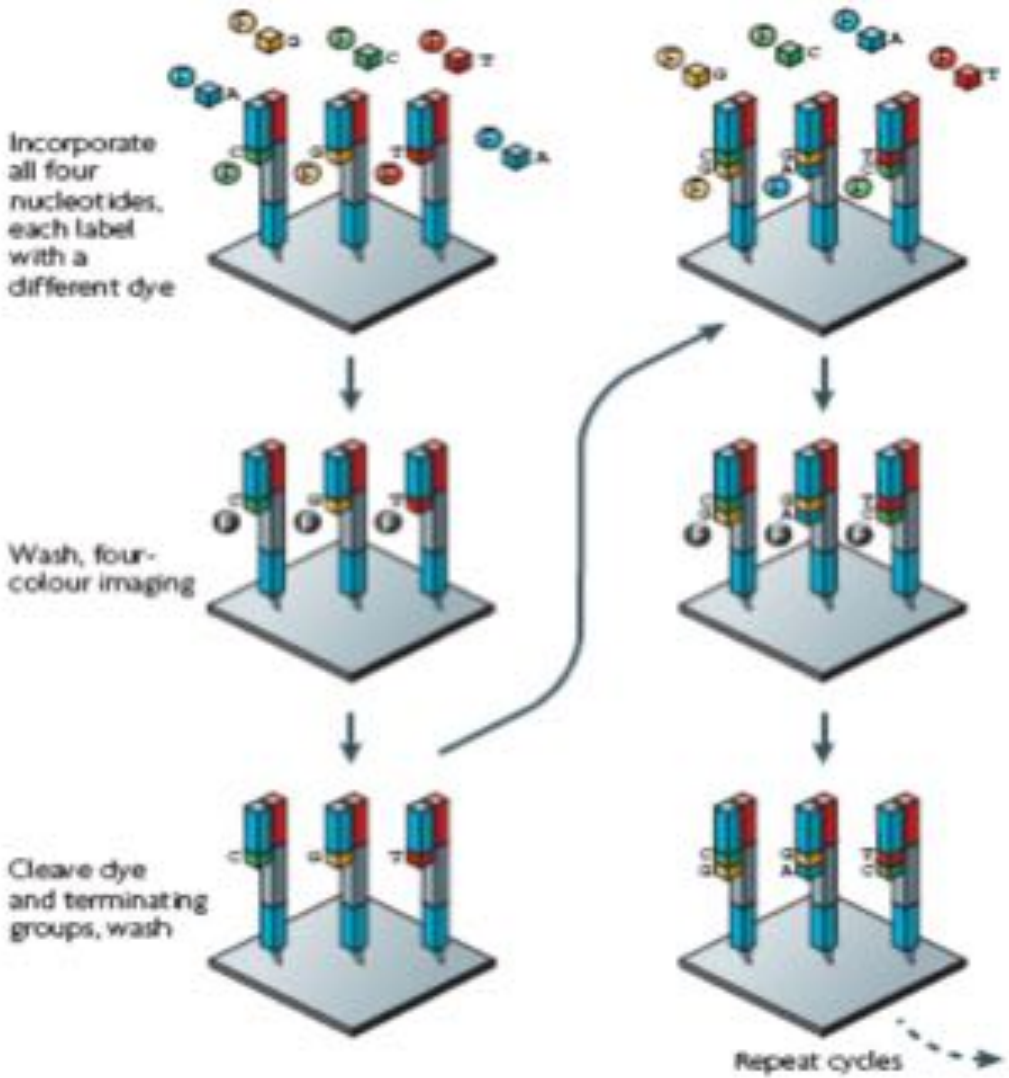


Cluster amplification by “bridge” PCR



https://binf.snipcademy.com/lessons/ngs-techniques/bridge-pcr

Clonal amplification



Base calling

- 6 cycles with base-calling

"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: TACAC, corresponding to the template sequence

TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

https://www.youtube.com/watch?v=IzXQVwWYFv4

https://www.youtube.com/watch?time_continue=65&v=tuD-ST5B3QA

Illumina sequencers

- **Illumina HiSeq**: ~3 billion paired 100bp reads, ~600Gb, $10K, 8 days (or "rapid run" ~90Gb in 1-2 days)
- **Illumina X Ten**: ~6 billion paired 150bp reads, 1.8Tb, <3 days, ~1000 / genome($$), (or "rapid run" ~90Gb in 1-2 days)
- **Illumina NextSeq**: One human genome in <30 hours

http://www.businesswire.com/news/home/20150112006333/en/Illumina-Expands-World%E2%80%99s-Comprehensive-Next-Generation-Sequencing-Portfolio

Illumina sequencers

NovaSeq™ 6000 Sequencing System

Scalable throughput and flexibility for virtually any genome, sequencing method, and scale of project.

Highlights

- **Scalable platform**
  Match data output, time to results, and price per sample to study needs
- **Flexible performance**
  Configure sequencing method, flow cell type, and read length to support a broad range of applications
- **Streamlined operation**
  Increase lab efficiency with a simplified workflow and reduced hands-on time

- Massive improvement of the cluster density - higher output
- Less expensive than the previous sequencers
- Faster runs

https://blog.genohub.com/2017/01/10/illumina-unveils-novaseq-5000-and-6000/

http://www.mrdnalab.com/illumina-novaseq.html

Solexa (Illumina) sequencing: summary

Advantages:

- Best throughput, accuracy and read length
- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)
- Some runs are error prone

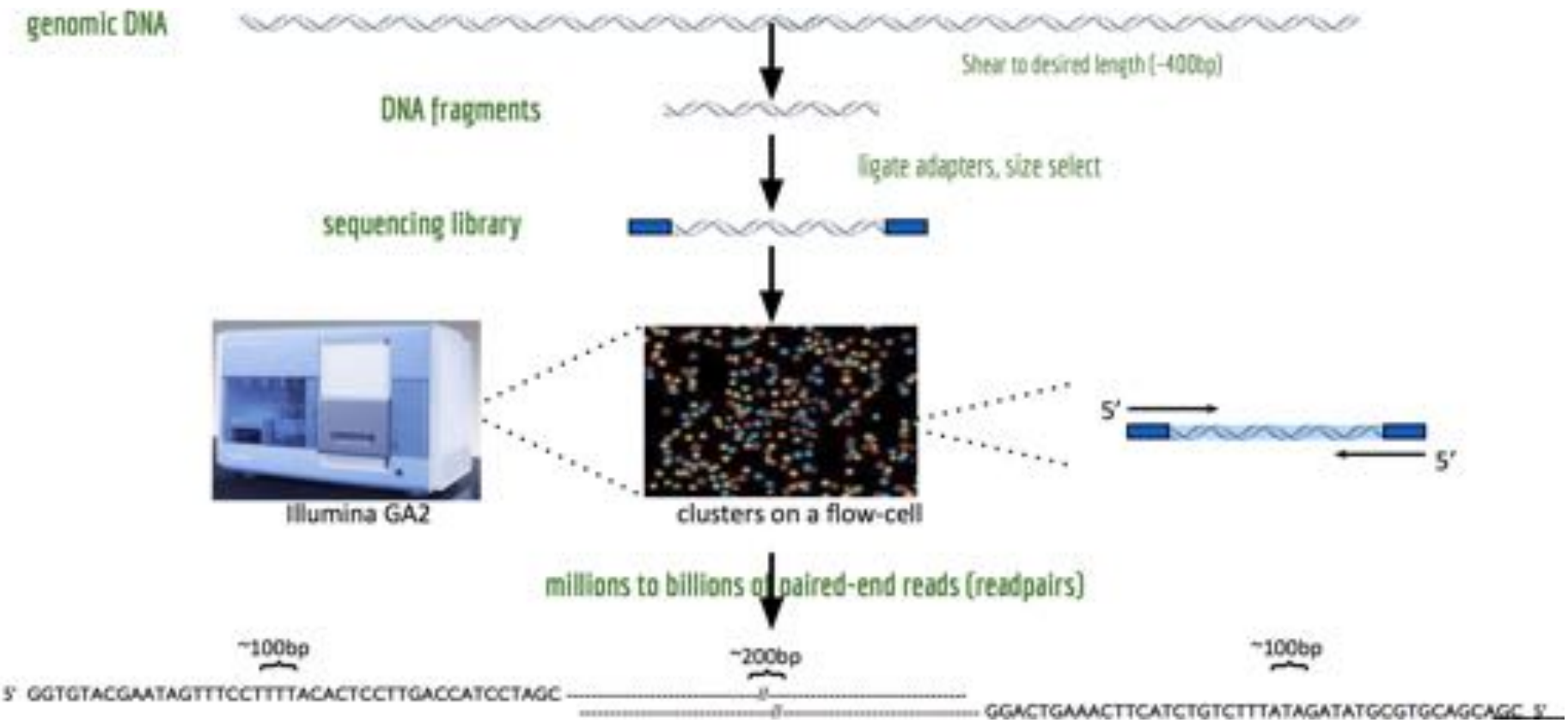
Video of Illumina sequencing https://www.youtube.com/watch?v=womKfikWlxM (5m)

Single-end vs. paired-end sequencing

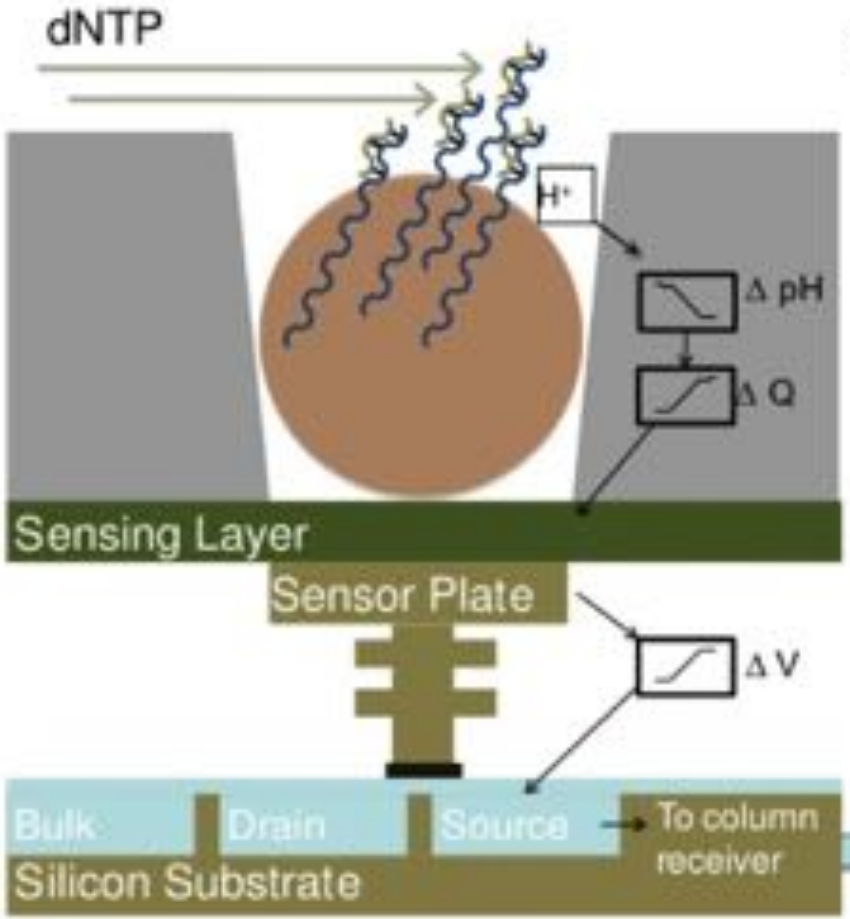
- Single-end sequencing: sequence one end of the DNA segment.
- Paired-end sequencing: sequence both ends of a DNA segments.
  - Result reads are "paired", separated by certain length (the length of the DNA segments, usually a few hundred bps).
  - Paired-end data can be used as single-end, but contain extra information which is useful in some cases, e.g., detecting structural variations in the genome.
  - Modeling technique is more complicated.

Paired-end sequencing - a workaround to sequence longer fragments

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



ION Torrent-pH Sensing of Base Incorporation



- DNA → Ions → Sequence
  - Nucleotides flow sequentially over Ion semiconductor chip
  - One sensor per well per sequencing reaction
  - Direct detection of natural DNA extension
  - Millions of sequencing reactions per chip
  - Fast cycle time, real time detection

Platforms: Ion Torrent

PGM

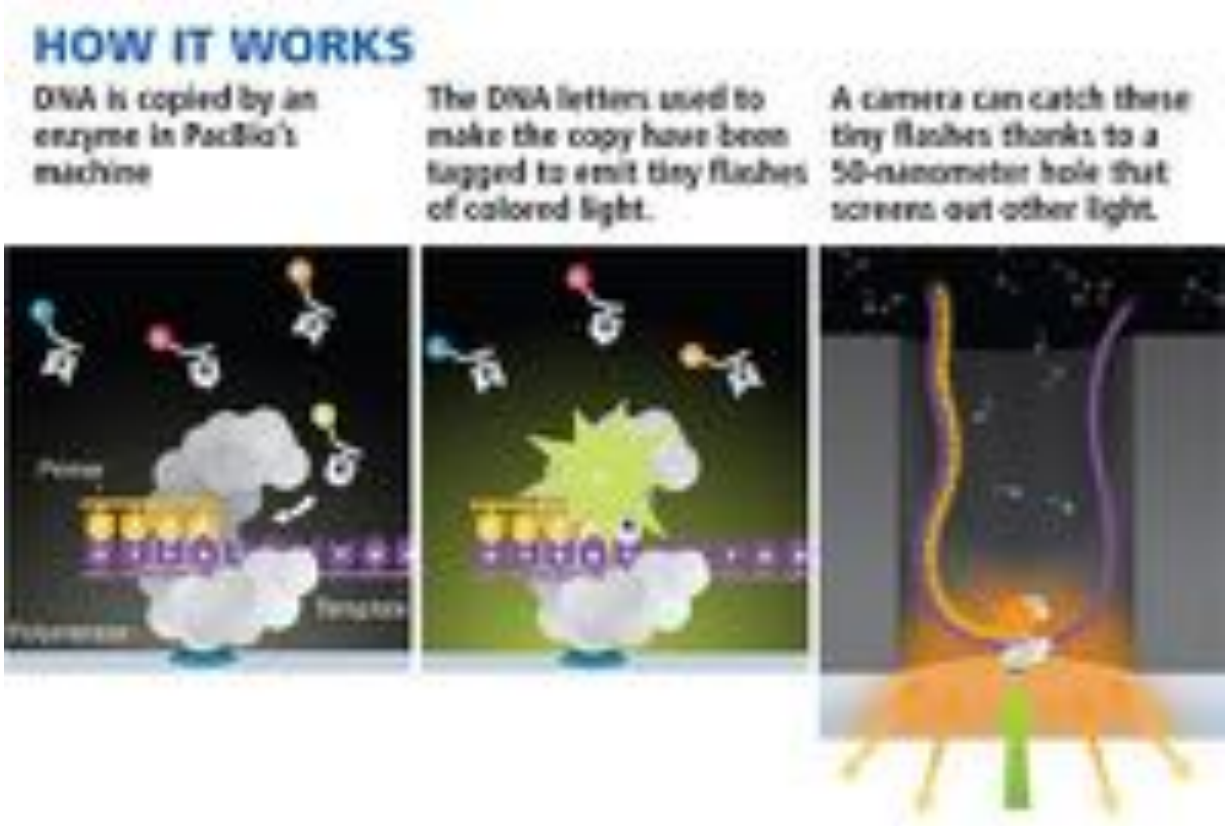
- Three sequencing chips available:
  - 314 = up to 100 Mb
  - 316 = up to 1 Gb
  - 318 = up to 2 Gb
- 2-7 hour/run
- up to 400 bp read length
- 400kreads up to 5 Mreads

Proton

- Two human exomes (Proton 1 chip) or one genome (@20X-Proton 2 chip) per run
- Ion One Touch or Ion Chef preparatory modules
- 2-4 hour/run
- ~200 bp average read length
- Proton 1 produces 60-80 Mreads $\geq$50 bp

- Low substitution error rate, in/dels problematic, no paired end reads
- Inexpensive and fast turn-around for data production
- Improved computational workflows for analysis

Pacific Biosciences



- Long reads
  - Structural variant discovery
  - *De novo* genome assembly

https://www.forbes.com/forbes/2009/1005/revolutionaries-science-genomics-gene-machine.html

Pacific Biosciences: summary

Key Points:

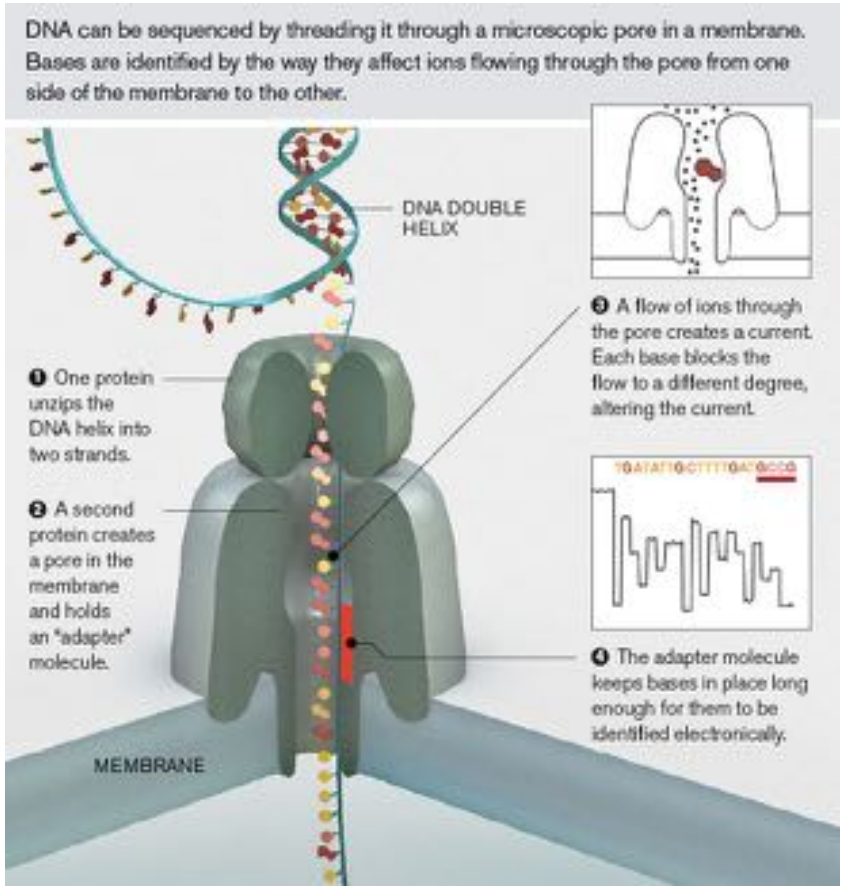
- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide)
- 4 colors flash in real time as polymerase acts
- Methylated cytosine has distinct pattern
- No *theoretical* limit to DNA fragment length

Caveats:

- Higher error rate (11-15%), but they are random
- Lower throughput, roughly 5 gigabases per run

Nanopore sequencing

- Nearly 30-years old technology



http://www2.technologyreview.com/news/427677/nanopore-sequencing/

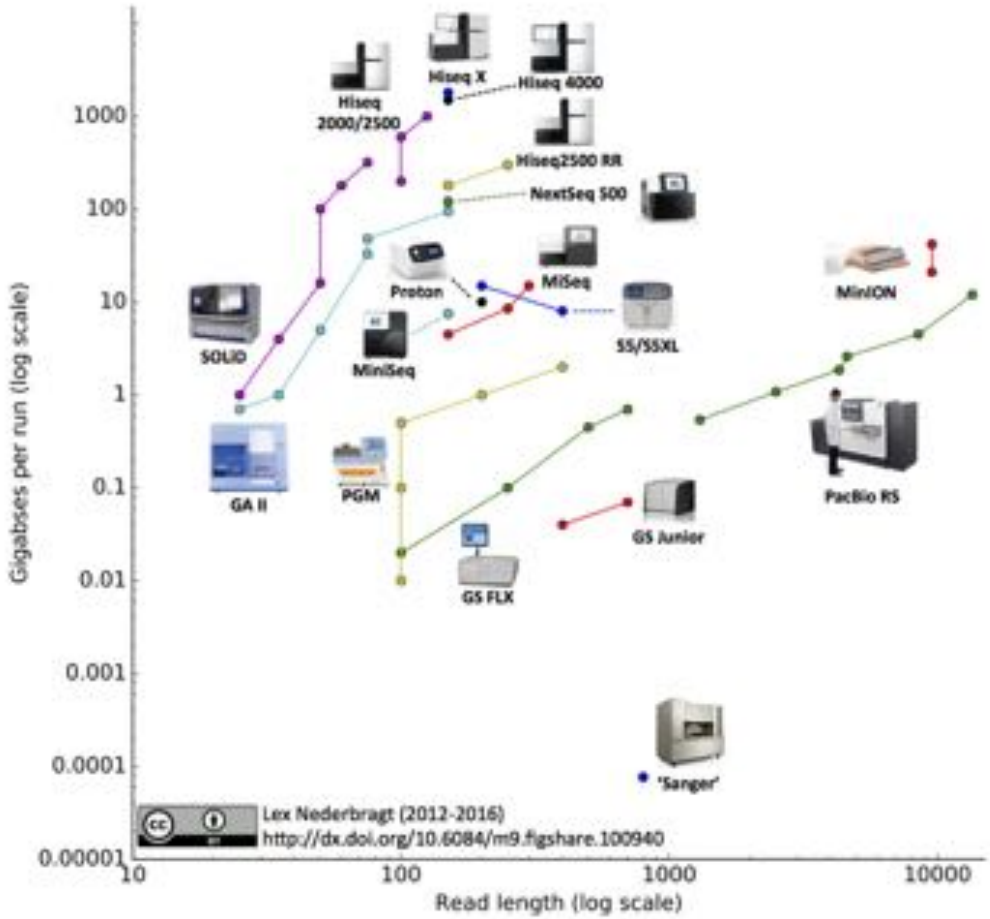
Nanopore sequencing

- Nanopore sequencing with ONT is accurate and relatively reliable
- Current yield per run ("R9.4" chemistry): ~5 Gbp, 5-15% error rate

https://www.technologyreview.com/s/600887/with-patent-suit-illumina-looks-to-tame-emerging-british-rival-oxford-nanopore/

Video of Ion Torrent chemistry, http://www.youtube.com/watch?v=yVf2295JqUg (2.5m)

Developments in next generation sequencing: instruments, read lengths, throughput.



Section 1

Sequencing applications

Applications

- NGS has a wide range of applications.
  - DNA-seq: sequence genomic DNA.
  - RNA-seq: sequence RNA products.
  - ChIP-seq: detect protein-DNA interaction sites.
  - Bisulfite sequencing (BS-seq): measure DNA methylation strengths.
  - A lot of others.
- Basically replaced microarrays with better data: greater dynamic range and higher signal-to-noise ratios.

What matters is what you feed into the sequencing machine

https://liorpachter.wordpress.com/seq/

Evolution of sequencing technologies

Technology	Brief description
ChIP-seq	Locate protein-DNA interaction or histone modification sites.
CLIP-seq	Map protein-RNA binding sites
RNA-seq	Quantify expression
SAGE-seq	Quantify expression
RIP-seq	capture TF-bound transcripts
GRO-seq	evaluate promoter-proximal pausing
BS-seq	Profile DNA methylation patterns
MeDIP-seq	Profile DNA methylation patterns
TAB-seq	Profile DNA hydroxyl-methylation patterns
MIRA-seq	Profile DNA methylation patterns
ChiRP-seq	Map lncRNA occupancy
DNase-seq	Identify regulatory regions
FAIRE-seq	Identify regulatory regions
FRT-seq	Quantify expression
Repli-seq	Assess DNA replication timing
MNase-seq	Identify nucleosome position
Hi-C	Infer 3D genome organization
ChIA-PET	Detect long distance chromosome interactions
4C-seq	Detect long distance chromosome interaction
Sono-seq	Map open-chromatin sites
NET-seq	determine *in vivo* position of all active RNAP complexes.
NA-seq	Map Nuclease-Accessible Sites

Section 2

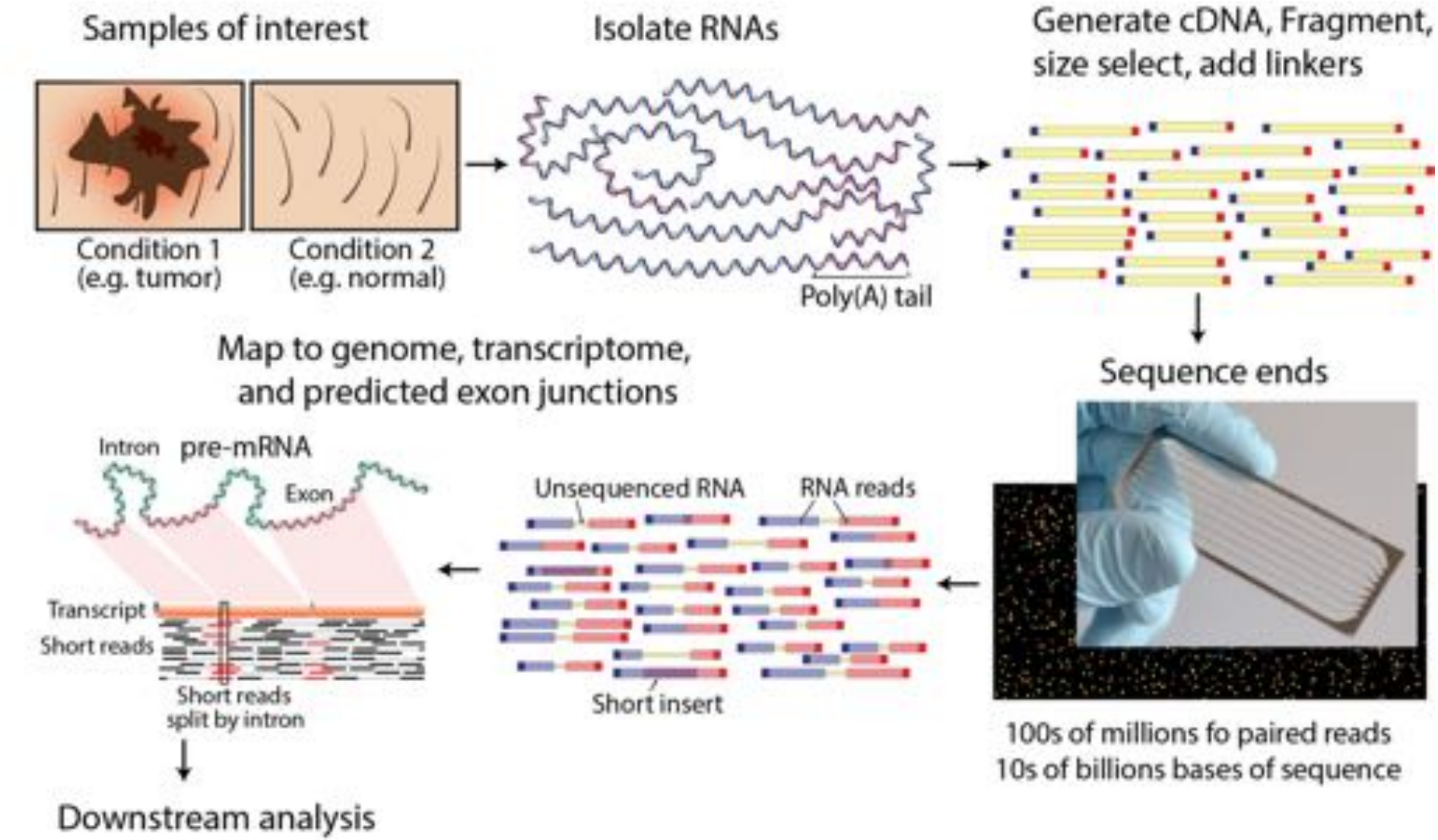
RNA-seq

What is RNA sequencing?

- Massive parallel sequencing to **characterize and quantify transcriptomes** (all actively transcribed genes)
- Detection of **differential gene expression**
- **Transcriptome reconstruction**, identification of **new transcripts**
- Detection of **alternative splicing events**
- Detection of **structural variants**, e.g., fusion transcripts
- **Allele-specific** gene expression measurements
- **Mutation analysis** – presence of genomic mutations and their effect on gene expression

http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s003

Overview of RNA sequencing technology



Source: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Library preparation steps

- **RNA isolation and QC**, to extract RNA relevant to the experimental question
- **Fragmentation**, to recover short reads across full length of long genes
- **Size selection**, suitable for RNA sequencing. 300-500bp - mRNA, 20-150bp - small/miRNA
- **Amplification**, typically by PCR. Up to $0.5 - 10ng$ of RNA
- **Library normalization/Exome capture**
- **Barcoding** and **multiplexing**
- Optionally, add **External RNA Control Consortium (ERCC) spike-in controls**
- **Single** or **paired end** sequencing. The latter is preferrable for the *de novo* transcript discovery or isoform expression analysis

Sample prepatation and library construction strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s005

RNA isolation

- **Ribosomal RNA (rRNA) depletion**
  - $0.1 - 1\mu g$ original total RNA (One cell contains ~10 picogram of total RNA)
  - rRNAs constitute over 90 % of total RNA in the cell, leaving the 1–2 % comprising messenger RNA (mRNA) that we are normally interested in (One cell contains ~0.1 picogram mRNA)
  - Enriches for mRNA + long noncoding RNA.
  - Hybridization to bead-bound rRNA probes

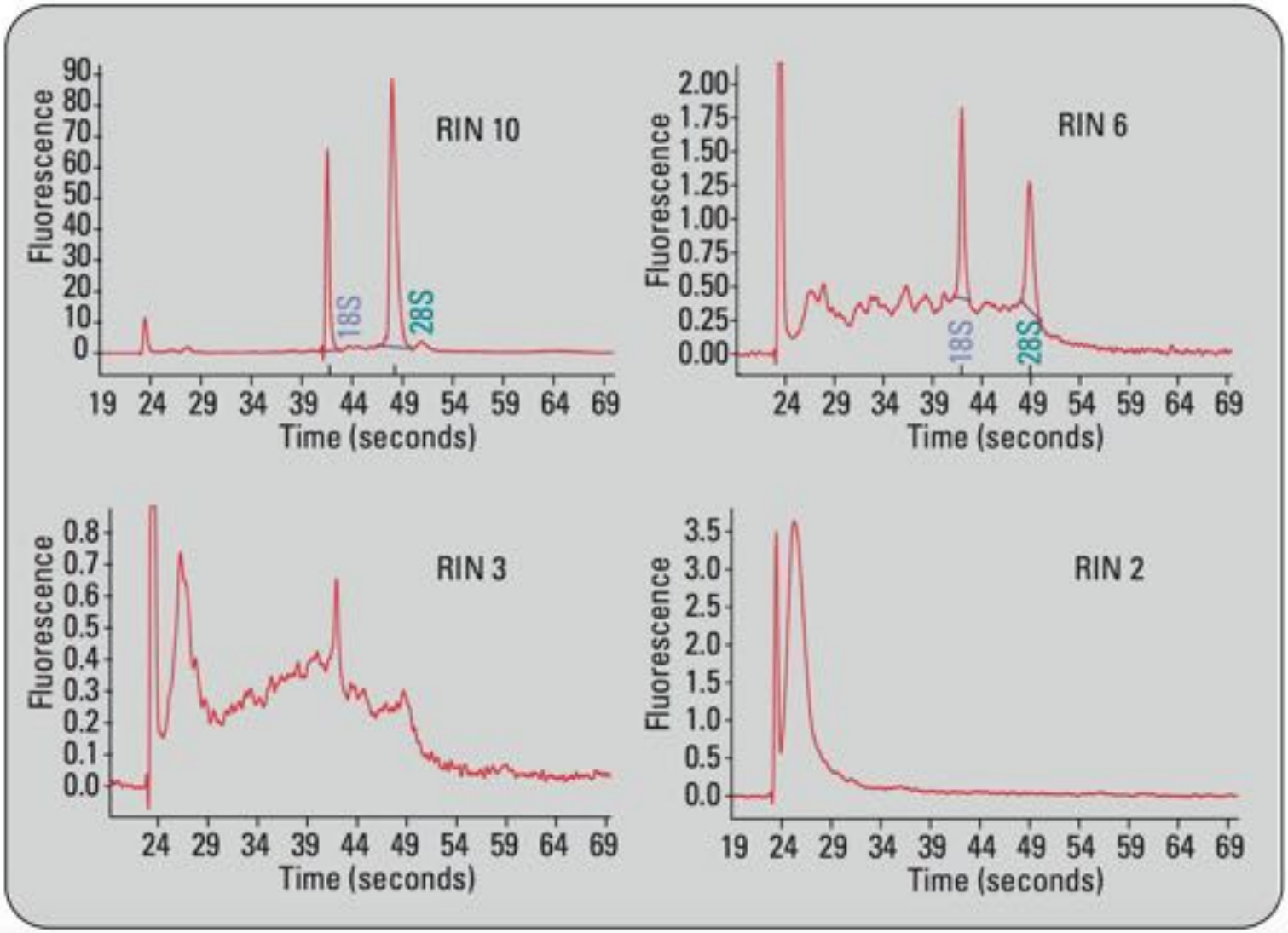
RNA isolation

- **Poly(A) selection (for eukaryotes only)**
  - Enrich for mRNA.
  - Hybridization to oligo-dT beads
- **Small RNA extraction**
  - Specific kits required to retain small RNAs
  - Optionally, size-selection by gel

Description of RNA-seq library enrichment strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s006

RNA quality

Agilent 2100 bioanalyzer. RIN - RNA integrity number (should be >7)



Unstranded vs. Strand-specific library

- **Unstranded**: Random hexamer priming to reverse-transcribe mRNA
- **Stranded**: dUTP method - incorporating UTP nucleotides during the second cDNA synthesis, followed by digestion of the strand containing dUTP

Strand-related settings for RNA-seq tools:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s007}

Sequencing length/depth

- Longer reads improve mappability and transcript quantification
- More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level
- Up to 100 million reads is needed to precisely quantify low expressed transcripts.
- In reality, 20-30 million reads is OK for human genome.

Summarization of read counts

- From RNA-seq, the alignment result gives the chromosome/position of each aligned read.
- For a gene, there are reads aligned to the gene body. How to summarized them into a number for the expression?

Counts of reads

- Easiest: The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it ~ number of aligned reads.
- Disadvantages: longer gene produce more reads, library depth (total counts) influence counts of individual transcripts
- **htseq-count** - command line tool for counting reads in features
- **featureCounts** - summarize multiple datasets at the same time

https://htseq.readthedocs.io/en/release_0.9.1/count.html

http://bioinf.wehi.edu.au/featureCounts/

Expression estimation for known genes and transcripts

- **Counts per million**: counts scaled by the library depth in million units. $CPM = C * 10^6 / N$
- **RPKM**: Reads Per Kilobase of transcript per Million mapped reads. Introduced by Mortazavi, 2008
- **FPKM**: Fragments Per Kilobase of transcript per Million mapped reads. Introduced by Salzberg, Pachter, 2010

Expression estimation for known genes and transcripts

- **FPKM** (or **RPKM**) attempt to normalize for gene size and library depth

$$RPKM\ (or\ FPKM_i) = (10^9 * C_i)/(N * L_i)$$

- C_i - number of mappable reads/fragments for a i gene/transcript/exon/etc.
- N - total number of mappable reads/fragments in the library
- L_i - number of base pairs in the i gene/transcript/exon/etc.

https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

TPM: Transcript per Kilobase Million

- **TPM**: Transcripts per million. Introduced by Li, 2011. Normalized by total transcript count instead of read count in addition to average read length.

If you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen for transcript i.

$$TPM_i = 10^6 * Z * \frac{C_i}{N * L_i}$$

- Z - sum of all length normalized transcript counts

TPM: Transcript per Kilobase Million

FPKM is calculated as

1. Sum sample/library fragments per million
2. Divide gene/transcript fragment counts by #1 – fragments per million, FPM
3. Divide FPM by length of gene in kilobases (FPKM)

TPM reverses the order - length first, library size second

1. Divide fragment count by length of transcript – fragments per kilobase, FPK
2. Sum all FPK for sample/library per million
3. Divide #1 by #2 (TPM)

https://youtu.be/TTUrtCY2k-w?t=23

https://www.ncbi.nlm.nih.gov/pubmed/22872506

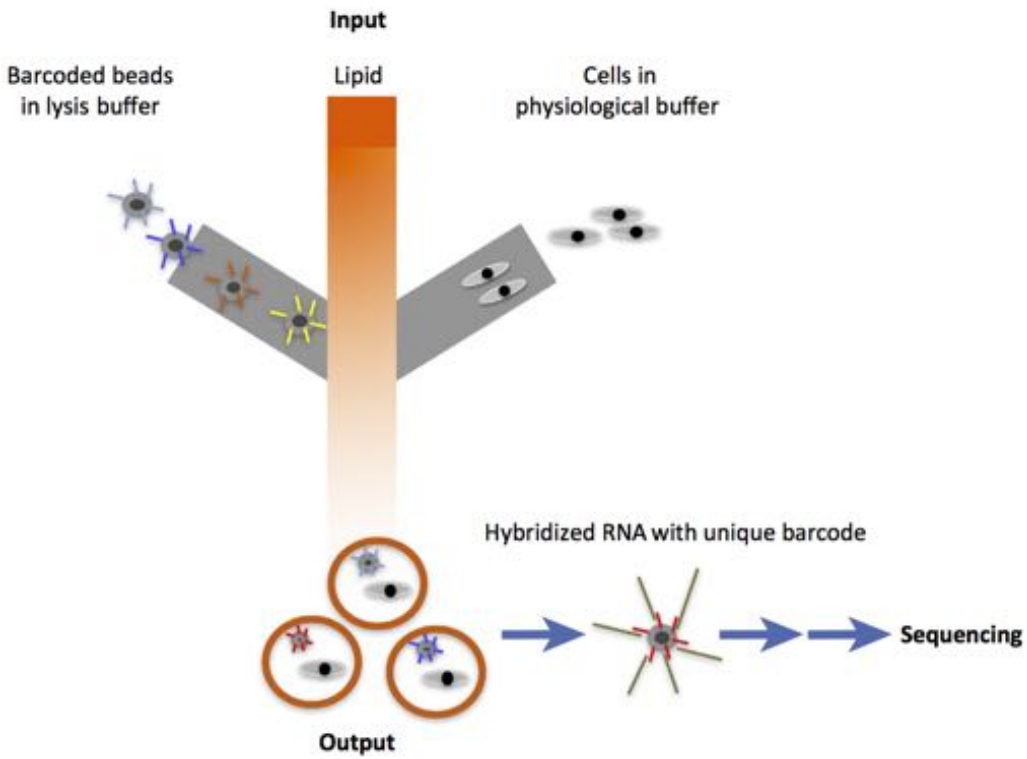
Section 3

Single cell sequencing

Single cell sequencing applications

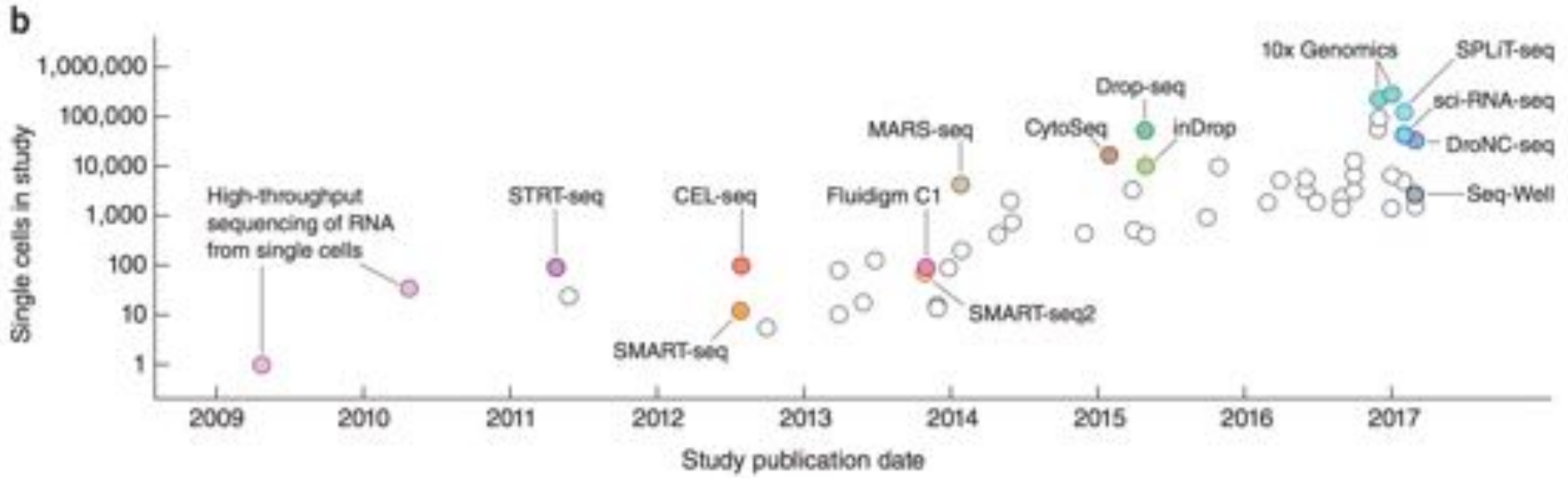
- Infer cell lineages
- Identify subpopulations
- Outline temporal evolution
- Define cell-specific biological characteristics, e.g., differentially expressed genes

Single-cell Sequencing Technology



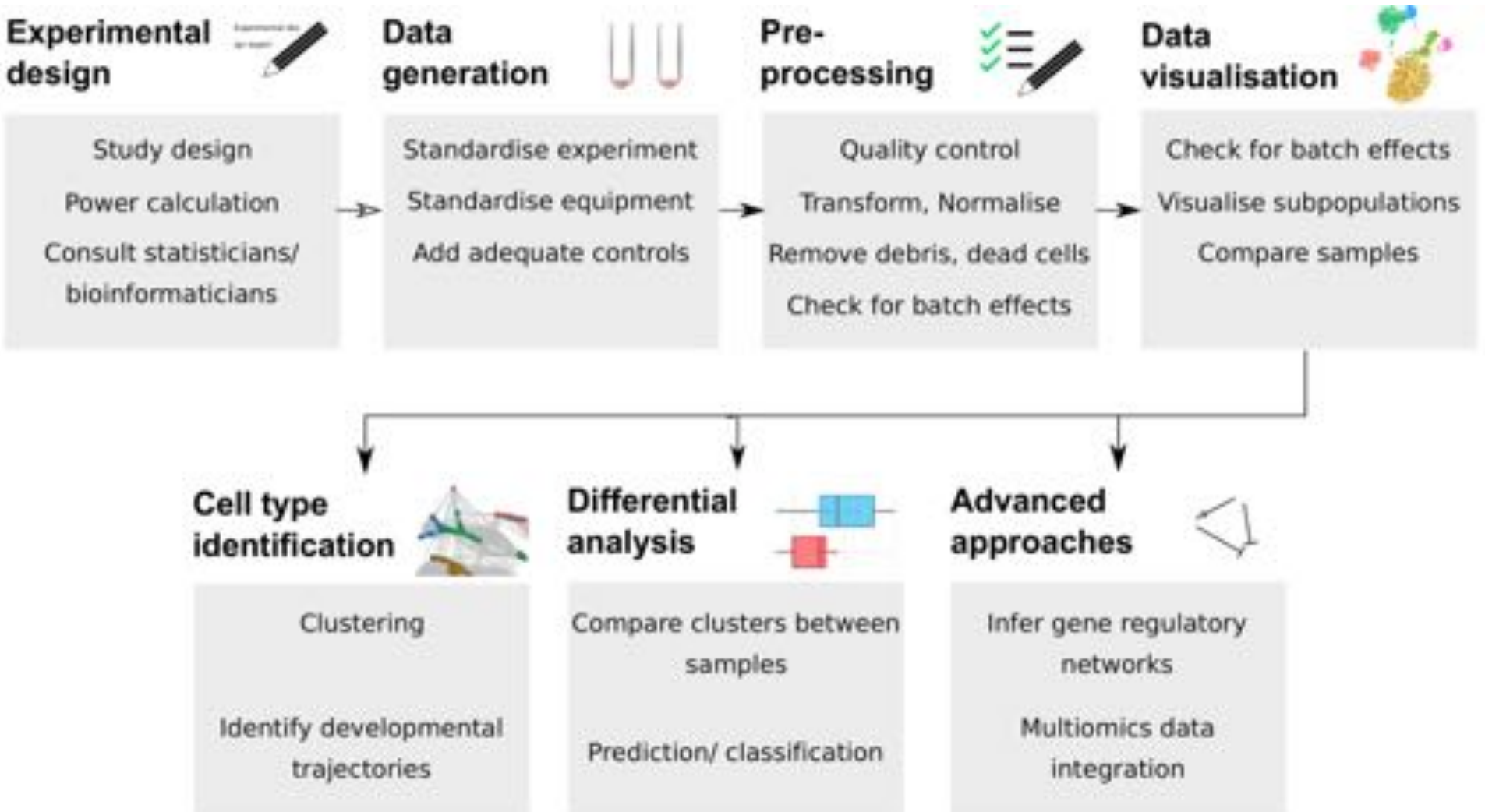
A single device has three input ports (oil, barcoded beads in lysis buffer, and cells of interest) and a single output port used for collecting bead–cell-containing lipid droplets. Then each cell (or RNA in the cell) is marked by the unique barcode and processed on the bead for sequencing.

Timeline of single-cell technologies



Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann. "Exponential Scaling of Single-Cell RNA-Seq in the Past Decade." Nature Protocols 13, no. 4 (April 2018): 599–604. https://doi.org/10.1038/nprot.2017.149.

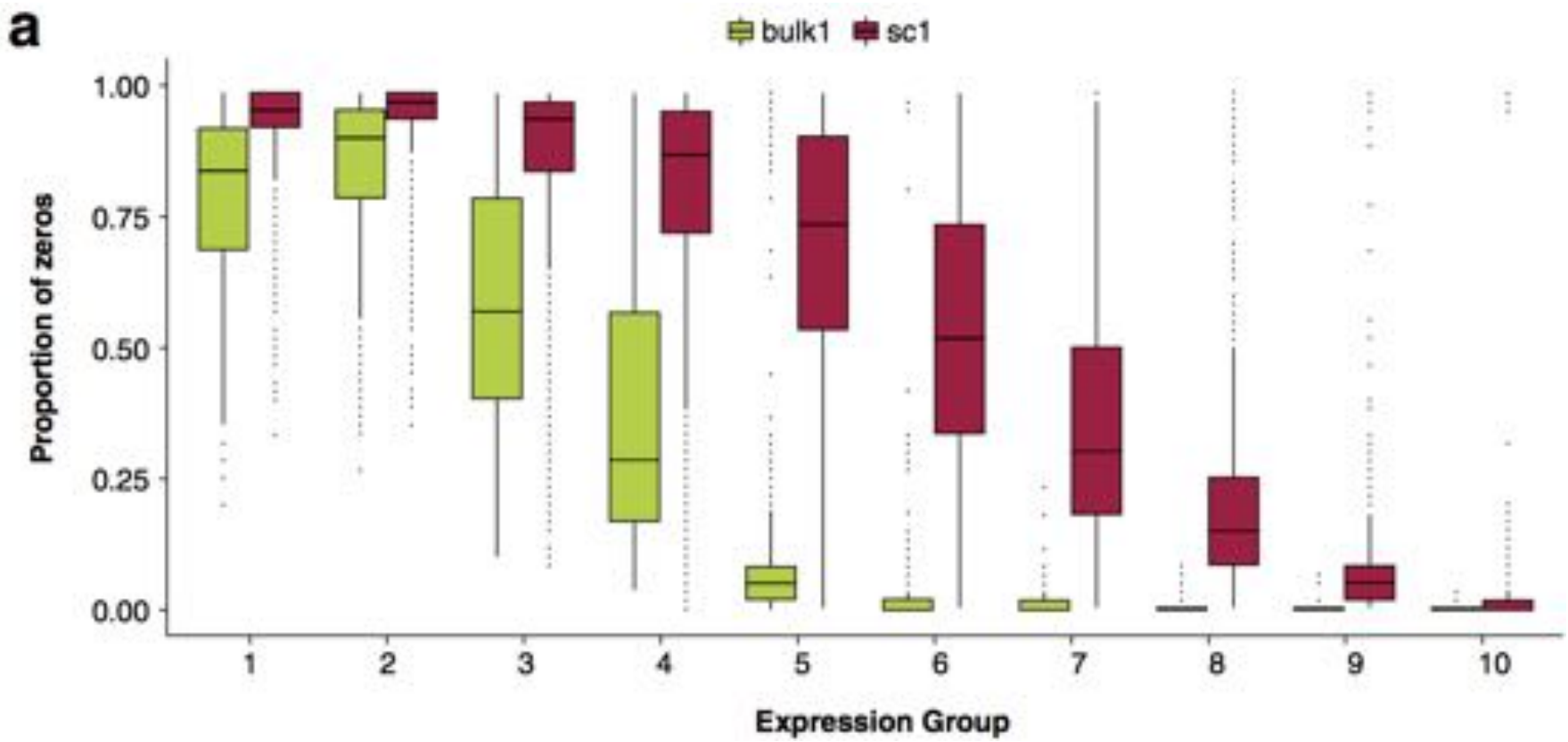
The computational workflow for single cell experiments detailed in steps



How does single-cell data differ from bulk RNA-seq

- In theory, up to 600 cells could be sequenced on a single lane (assuming 120 million reads per lane)
- Even with the most sensitive platforms, the data are relatively sparse owing to a high frequency of dropout events (lack of detection of specific transcripts)
- The numbers of expressed genes detected from single cells are typically lower compared with population-level ensemble measurements
- The commonly used 'reads per kilobase per million' (RPKM) transcript quantification is biased on a single-cell level, at the very least the 'transcripts per million' (TPM) should be used

Abundance of zeros



Bacher, Rhonda, and Christina Kendziorski. "Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments." Genome Biology 17 (April 7, 2016): 63. https://doi.org/10.1186/s13059-016-0927-y.

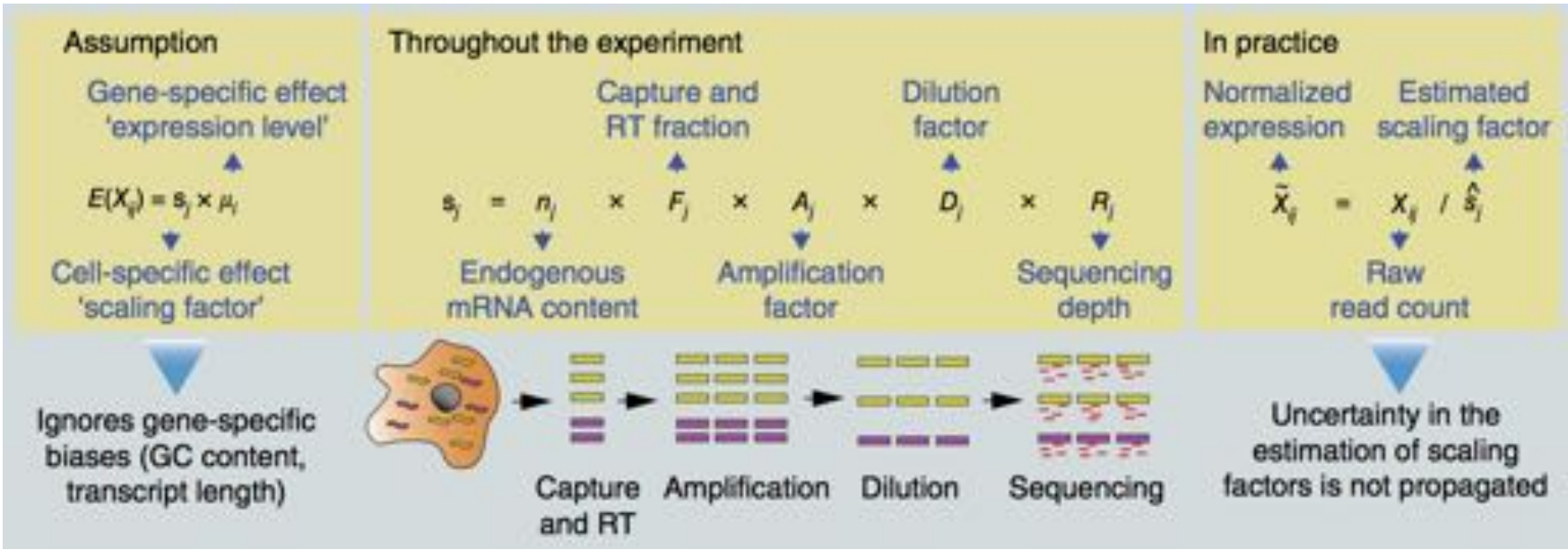
How does single-cell data differ from bulk RNA-seq

- scRNA-seq data, in general, are much more variable than bulk data
- Distributions of transcript quantities are often more complex in single-cell datasets than in bulk RNA-seq - negative binomial or multimodal distributions

Filtering

- Filter cells and/or genes
- No single consensus, frequently used criteria include:
  - relative library size
  - number of detected genes ($<$ 10,000)
  - fraction of reads mapping to mitochondria-encoded genes or synthetic spike-in RNAs ($<$ 500)

Global-scaling normalization



Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. "Normalizing Single-Cell RNA Sequencing Data: Challenges and Opportunities." Nature Methods 14, no. 6 (May 15, 2017): 565–71. https://doi.org/10.1038/nmeth.4292.

Sub-population identification

Standard methods used in RNA-Seq

- **Hierarchical clustering**, **PCA**, **tSNE** of highly variable, or differentially expressed, genes. Zeros can be a problem
- **ZIFA** - Zero-inflated dimensionality reduction algorithm for single-cell data
- **SNN-Cliq** - A clustering method for high dimensional dataset. Rank-based (not expression) similarity

https://github.com/epierson9/ZIFA

http://bioinfo.uncc.edu/SNNCliq/

Many more at https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0927-y

Cell clustering

- Perhaps the most active topic in scRNA-seq
- The goals include:
  - Cluster cells into subgroups
  - Model temporal transcriptomic dynamics: reconstruct "pseudo-time" for cells. This is useful for understanding development or disease progression
- Traditional method like k-means or hierarchical clustering need to be used with caution due to dropout events

t-SNE: a useful visualization tool

- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map
- When project high-dimensional data into lower dimensional space, preserve the distances among data points
  - This alleviate the problem that many clusters overlap on low dimensional space
- Try to make the pairwise distances of points similar in high and low dimension
- This is used in almost all scRNA-seq data visualization
- Has "tsne" package on CRAN

Pseudotemporal ordering

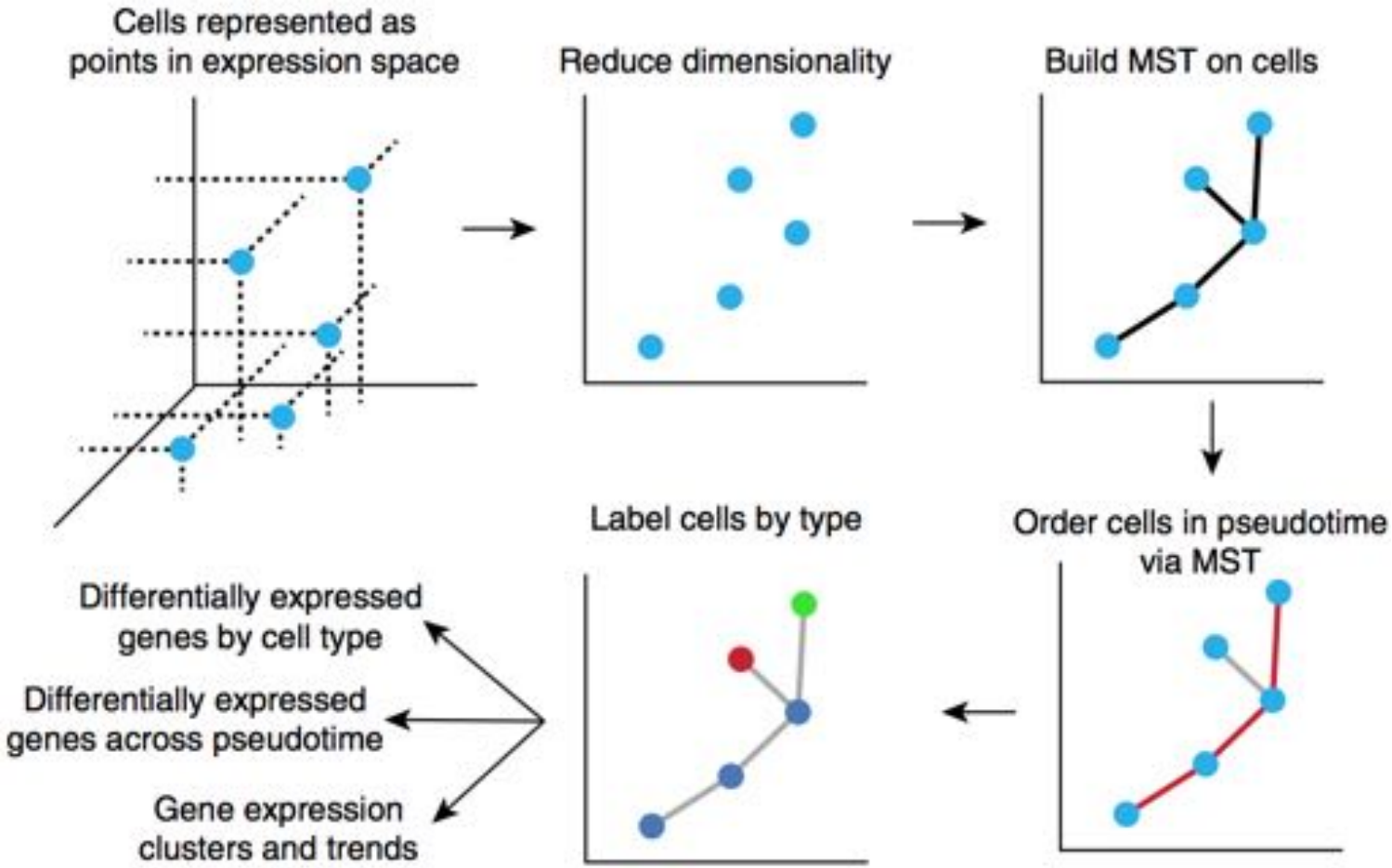
- Idea - cells at different differentiation (or other biological process) stage are presented with different expression profiles
- Dynamics of cellular processes can be reconstructed from expression profiles
- Key assumption: genes do not change direction very often, thus samples with similar transcriptional profiles should be close in order
- Most approaches are dimensionality reduction-based, and apply graph theory designed to traverse nodes in a graph efficiently
- **Monocle** - Independent component analysis, then a minimum spanning three through the dimension-reduced data

https://cole-trapnell-lab.github.io/monocle-release/

Many more at https://github.com/agitter/single-cell-pseudotime

Monocle, An analysis toolkit for single-cell RNA-seq

Single-cell trajectories, clustering, visualization, differential expression



https://cole-trapnell-lab.github.io/monocle-release/

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Pranti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J.

Differentially expressed genes

- Need to accomodate unobserved dropouts, bimodality in expression levels due to abundance of zero or low values (**MAST**, **SCDE**)
- **scDD** - Distinguishes four types of differential expression changes to increase power:
  - shifts in unimodal distribution
  - differences in the number of modes
  - differences in the proportion of cells within modes
  - combination of the previous two

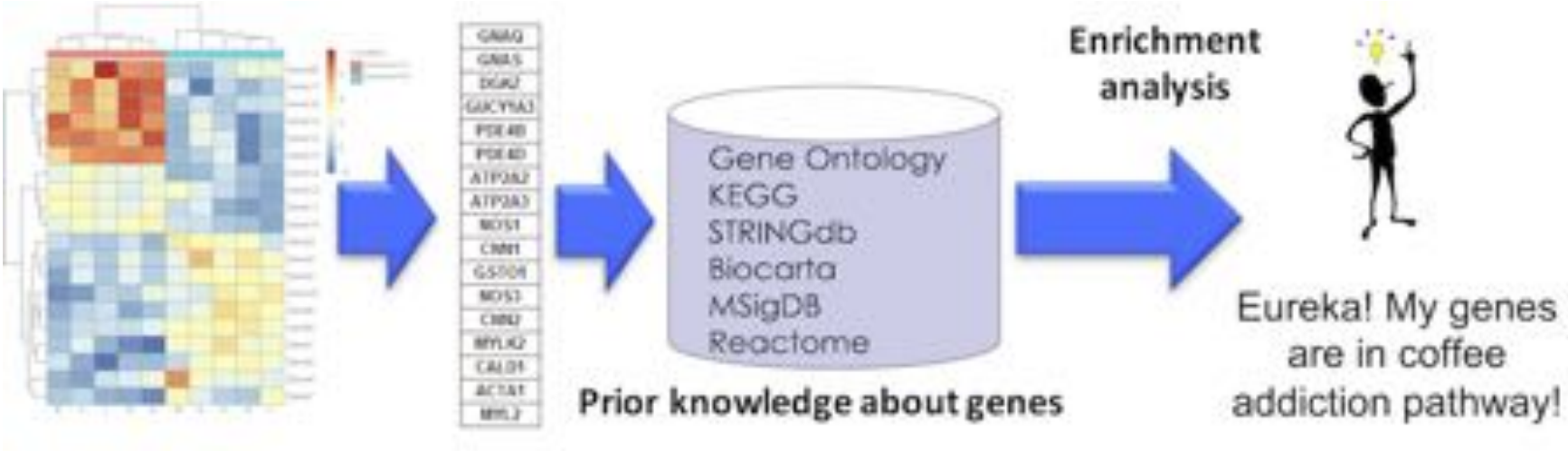
https://github.com/kdkorthauer/scDD

Section 4

Functional enrichment analysis

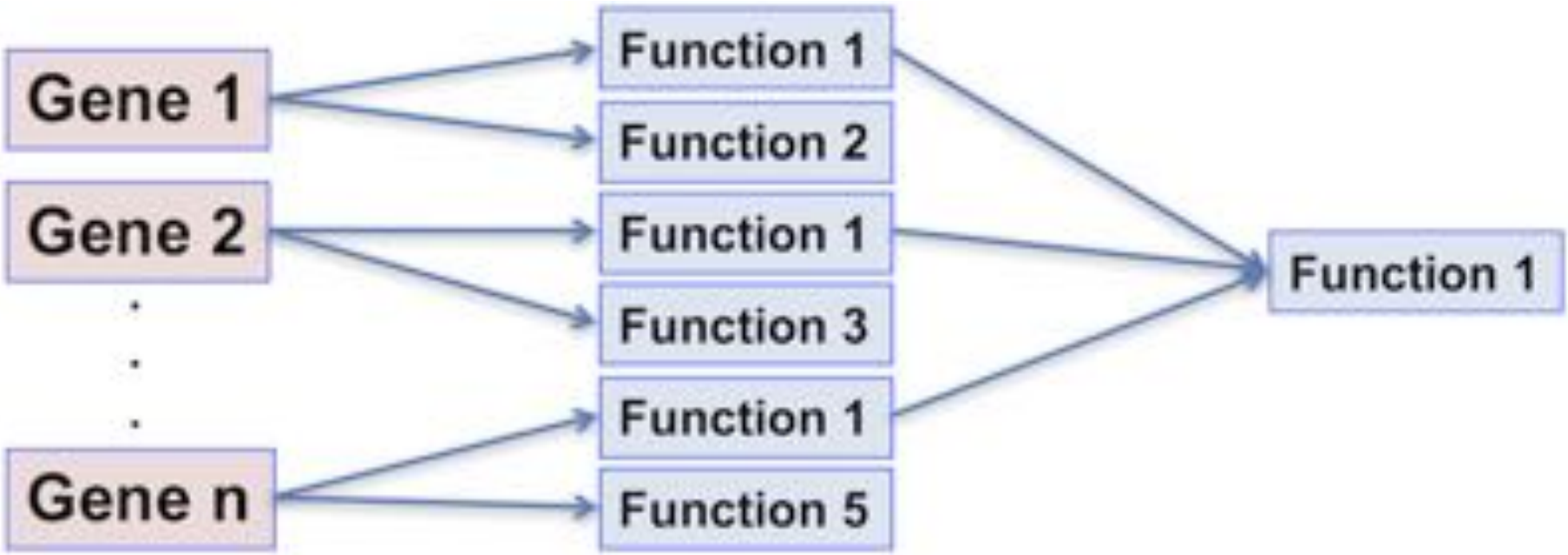
Why enrichment analysis?

- Translating changes of **hundreds/thousands of differentially expressed genes** into a few biological processes (reducing dimensionality)
- High level understanding of the biology behind gene expression – **Interpretation!**



What is enrichment analysis

- **Enrichment analysis** - summarizing common functions associated with a group of objects



Annotation databases and ontologies

- An annotation database annotates genes with functions or properties - sets of genes with shared functions
- Structured prior knowledge about genes

Gene ontology structure

Gene ontology describes multiple levels of detail of gene function.

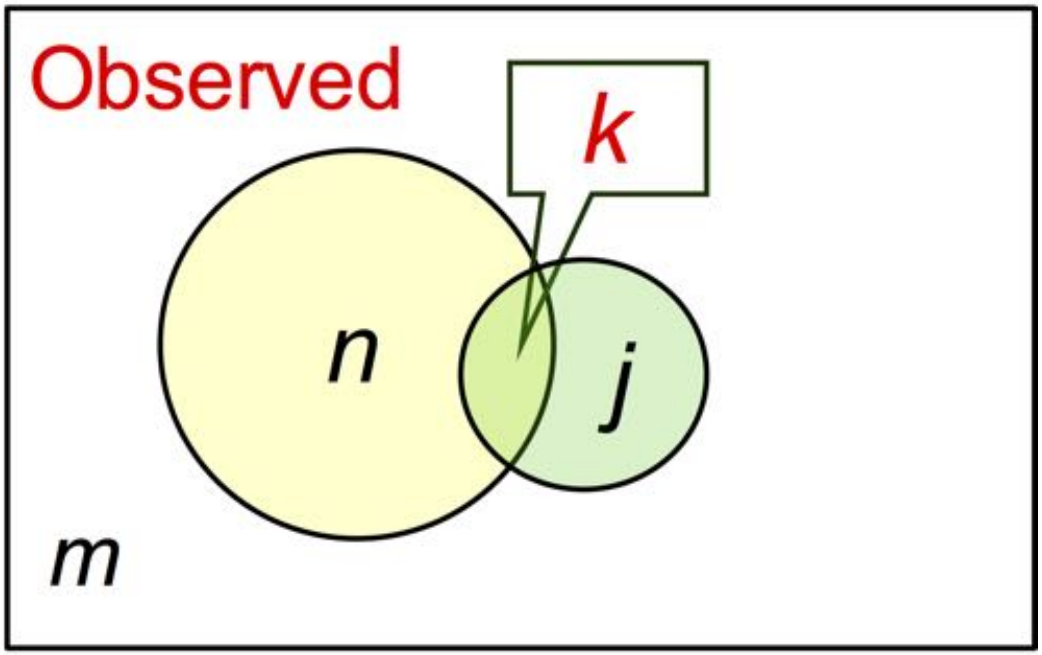
- **Molecular Function** - the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*
- **Biological Process** - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions
- **Cellular Component** - subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

Other gene annotation databases

- **KEGG: Kyoto Encyclopedia of Genes and Genomes** - a collection of biological information compiled from published material, http://www.genome.jp/kegg/
- **Reactome** - Curated human pathways encompassing metabolism, signaling, and other biological processes, http://www.reactome.org/
- **MSigDb** - Molecular Signatures Database, http://software.broadinstitute.org/gsea/msigdb/
- **PathwayCommons**, version 8 has over 42,000 pathways from 22 data sources, http://www.pathwaycommons.org/
- **PathGuide**, lists ~550 pathway related databases, http://www.pathguide.org/
- **WikiPathways**, community-curated pathways, http://wikipathways.org/
- **BioCarta**, pathway genes and diagrams, https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
- **Consensus-PathDB**, pathway interactions, enrichment, data, http://www.consensuspathdb.org/

Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category



Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

What is the probability of having k or more genes from the category in the selected n genes?

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Overrepresentation analysis, Hypergeometric test

1. Find a set of differentially expressed genes (DEGs)
2. Are *DEGs in a set* more common than *DEGs not in a set*?

- Fisher test `stats::fisher.test()`
- Conditional hypergeometric test, to account for directed hierachy of GO `GOstats::hyperGTest()`

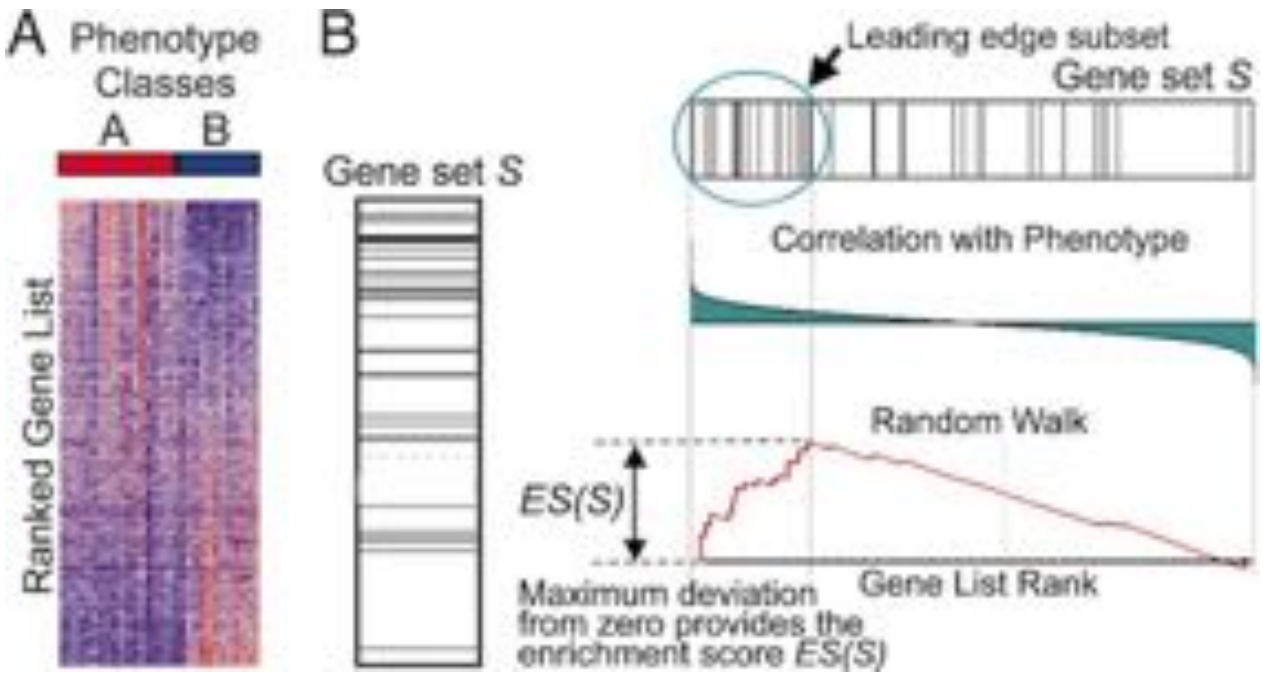
Example: https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

GSEA: Gene set enrichment analysis

- **Gene set analysis (GSA)**. Mootha et al., 2003; modified by Subramanian, et al. "**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.**" PNAS 2005 http://www.pnas.org/content/102/43/15545.abstract
- Main rationale – functionally related genes often display a coordinated expression to accomplish their roles in the cells
- Aims to identify gene sets with "subtle but coordinated" expression changes that would be missed by DEGs threshold selection

GSEA: Gene set enrichment analysis

1. Sort genes by log fold change
2. Calculate running sum - increment when gene in a set, decrement when not
3. Maximum of the runnig sum is the enrichment score - larger means genes in a set are toward top of the sorted list
4. Permute subject labels to calculate significance p-value



Gene set enrichment analysis

- **GSEA** (https://www.broadinstitute.org/gsea/index.jsp) - Better way of doing enrichment analysis
- **g:Profiler** (http://biit.cs.ut.ee/gprofiler/) - gene ID converter, GO and pathway enrichment, and more
- **ToppGene** (https://toppgene.cchmc.org) - Quick gene enrichment analysis in multiple categories
- **Metascape** (http://metascape.org/) - Enrichment analysis of multiple gene sets
- **DAVID** (https://david.ncifcrf.gov/) - Newly updated gene enrichment analysis
- **FRY** (http://shiny.bioinf.wehi.edu.au/giner.g/FRY_GeneSetExplorerApp/) - Fast Interactive Biological Pathway Miner, from WEHI group

R packages: enrichment analysis

- **clusterProfiler** (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) - statistical analysis and visualization of functional profiles for genes and gene clusters
- **limma** (https://bioconductor.org/packages/release/bioc/html/limma.html) - Linear Models for Microarray Data, includes functional enrichment functions `goana`, `camera`, `roast`, `romer`
- **GOstats** (https://www.bioconductor.org/packages/2.8/bioc/html/GOstats.html) - tools for manimpuating GO and pathway enrichment analyses. https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

R packages: Gene annotation databases

- **annotables** (https://github.com/stephenturner/annotables) - R data package for annotating/converting Gene IDs
- **msigdf** (https://github.com/stephenturner/msigdf) - Molecular Signatures Database (MSigDB) in a data frame
- **pathview** (https://www.bioconductor.org/packages/devel/bioc/html/pathview.html) - a tool set for pathway based data integration and visualization

Section 5

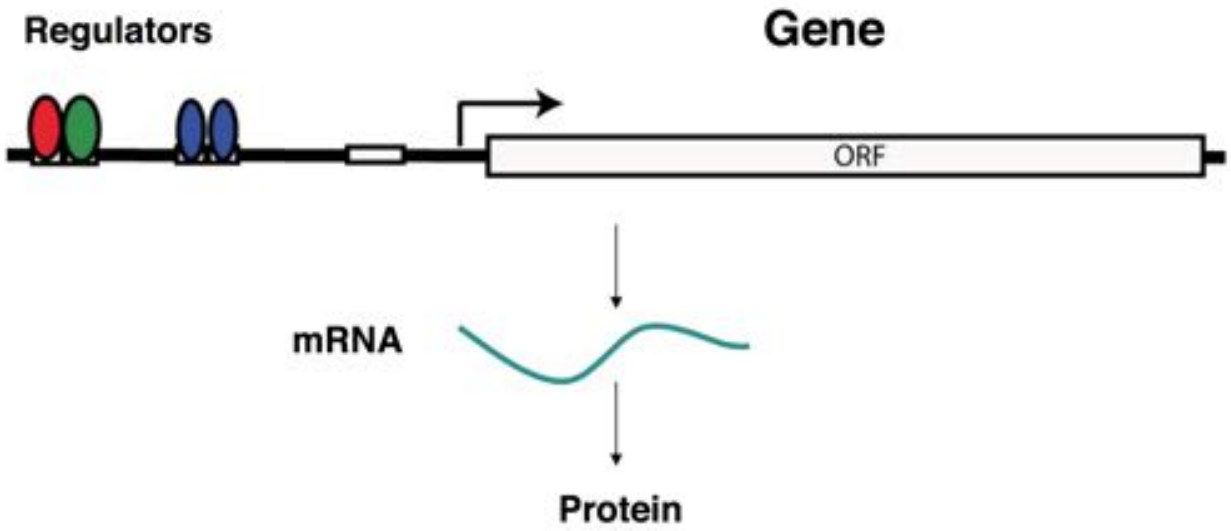
ChIP (Chromatin Immunoprecipitation) sequencing

ChIP-seq

- Chromatin-Immunoprecipitation (ChIP) followed by sequencing (seq): sequencing version of ChIP-chip.
- Used to detect locations of certain "events" on the genome:
  - Transcription factor binding.
  - DNA methylations and histone modifications.
- A type of "captured" sequencing. ChIP step is to capture genomic regions of interest.

Gene Regulation: DNA -> RNA -> Protein

- What are the transcription factors (TFs) that control gene expression?
- At what genes do these TFs operate?
- Understanding transcriptional regulatory network will
  - Reveal how cellular processes are connected and coordinated
  - Suggest new strategies to manipulate phenotypes and combat disease



ChIP-seq big picture

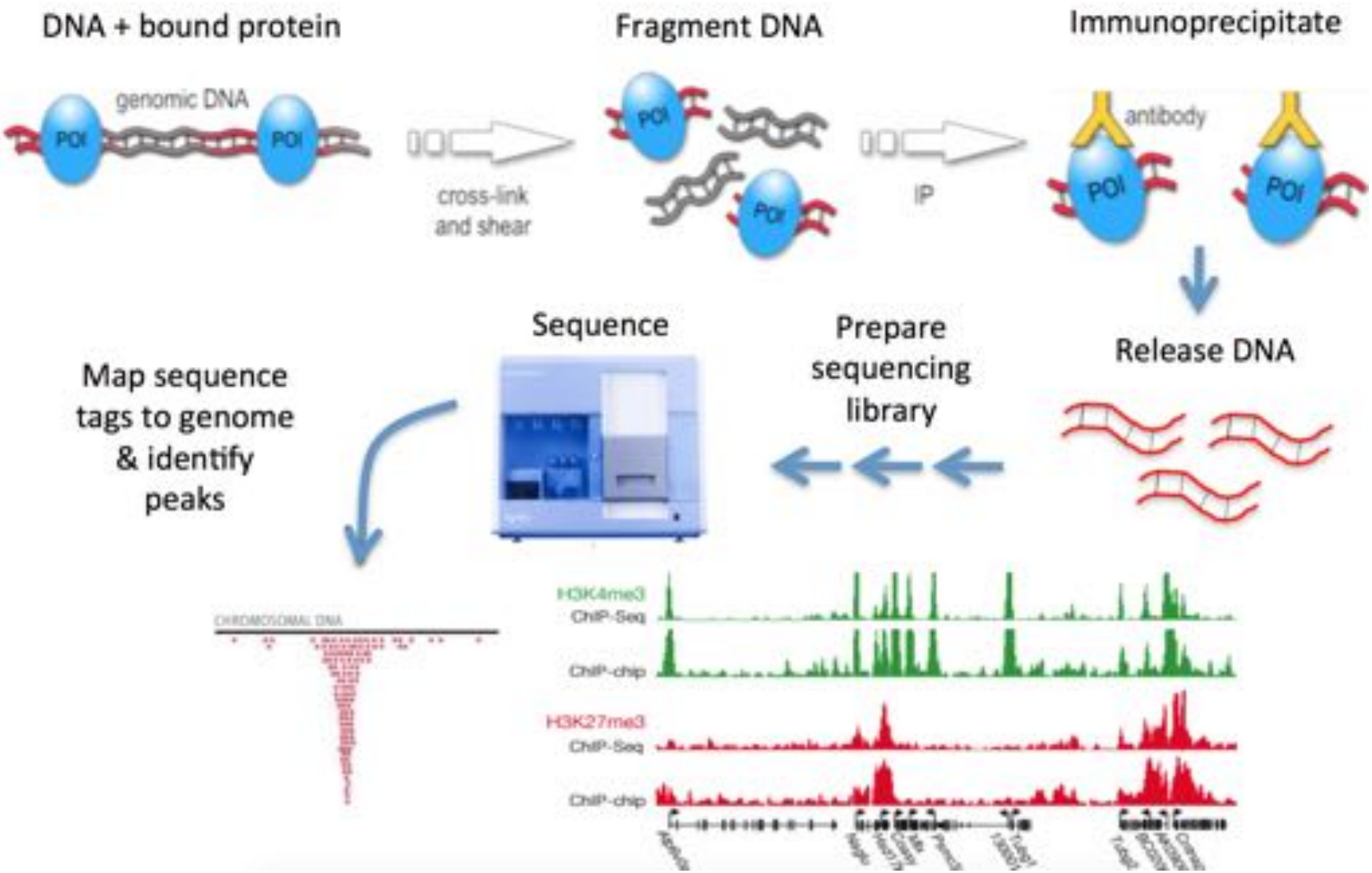
Combine high-throughput sequencing with Chromatin Immunoprecipitation to identify specific protein-DNA interactions genome-wide, including those of:

- Transcription factors
- Histones (various types and modifications)
- RNA Polymerase (survey of transcription)
- DNA polymerase (investigate DNA replication)
- DNA repair enzymes
- Fragments of DNA that are modified (e.g. methylated)

Experimental procedures

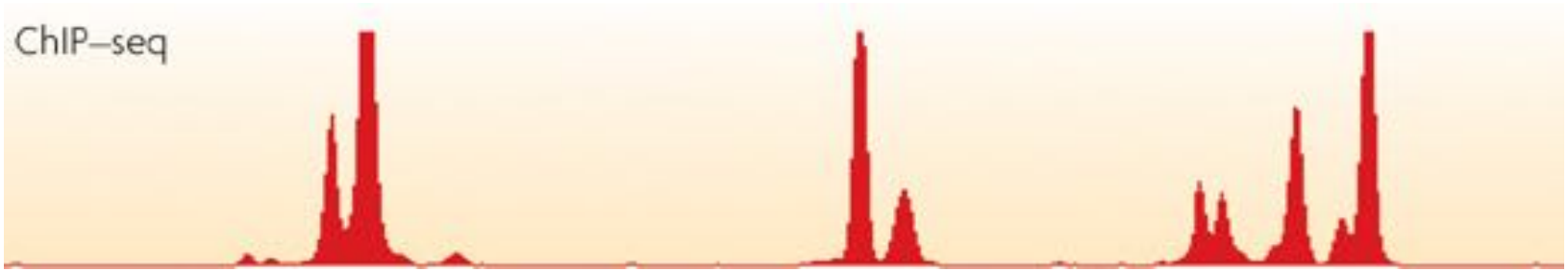
1. **Crosslink**: fix proteins on Isolate genomic DNA.
2. **Sonicate**: cut DNA in small pieces of ~200bp.
3. **Immunoprecipitate (IP)**: use antibody to capture DNA segments with specific proteins.
4. **Reverse crosslink**: remove protein from DNA.
5. **Sequence** the DNA segments.

ChIP-seq overview



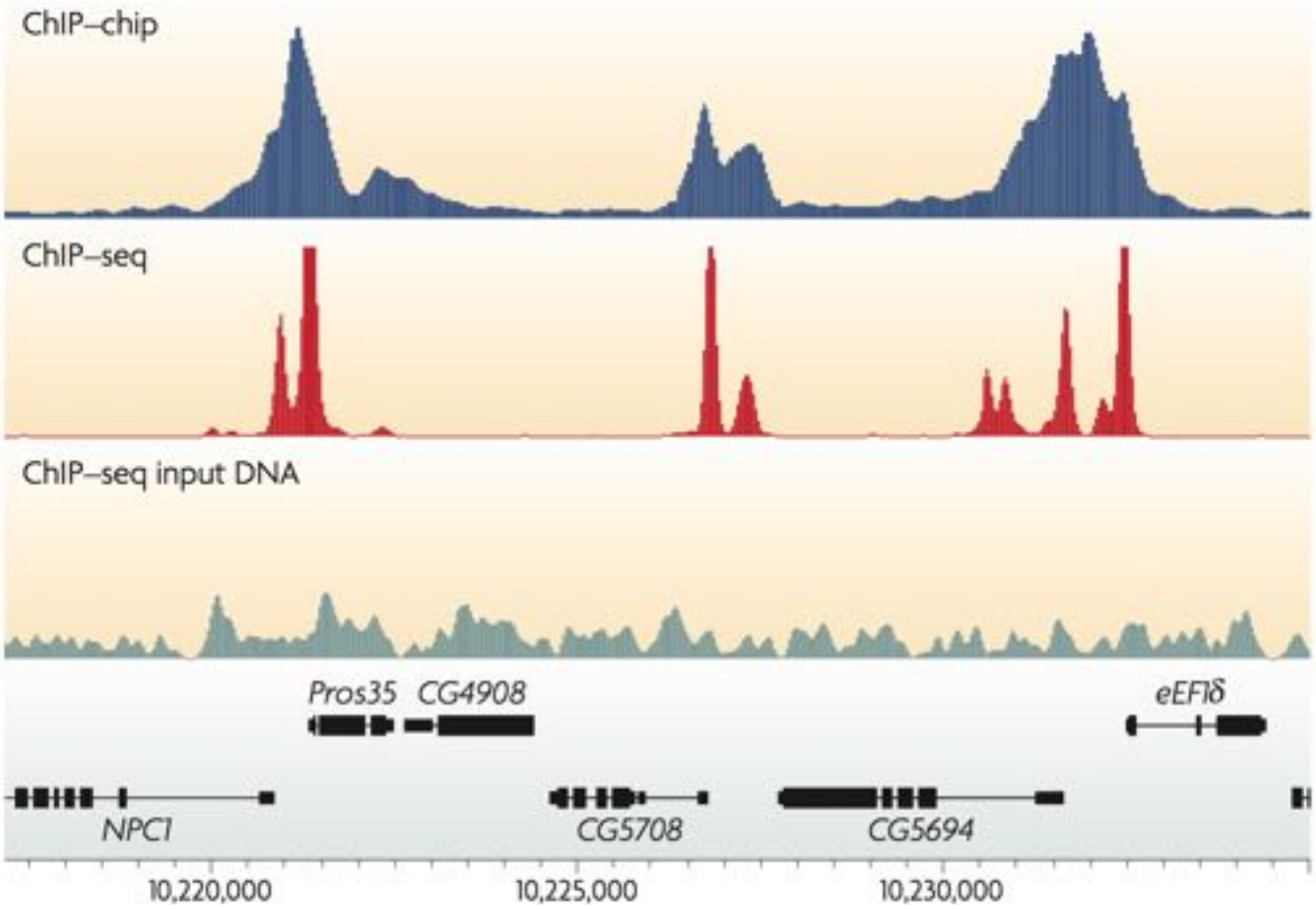
ChIP-seq "peak" detection

- When plot the read counts against genome coordinates, the binding sites show a tall and pointy peak. So "peaks" are used to refer to protein binding or histone modification sites



- Peak detection is the most fundamental problem in ChIP-seq data analysis

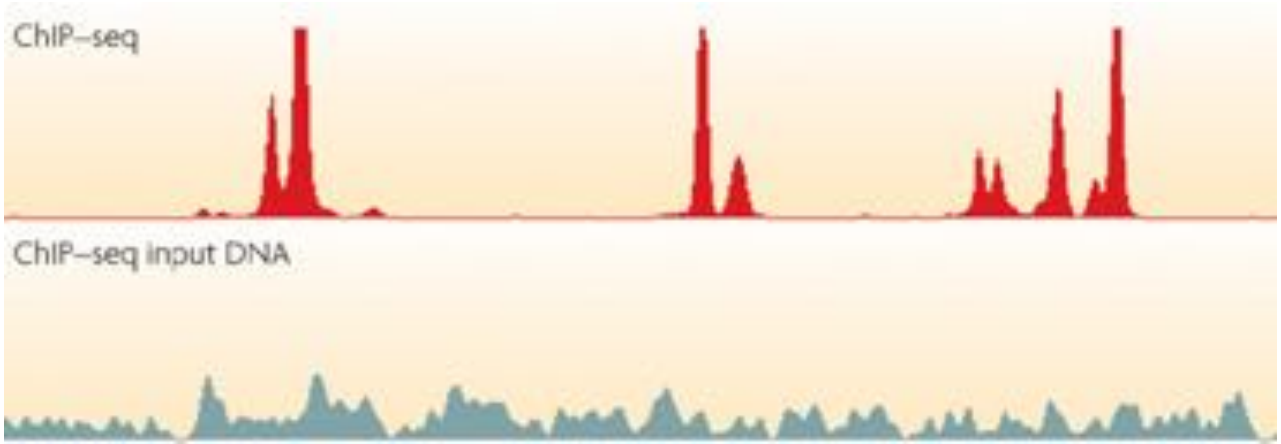
Peak calling: a classic signal versus noise problem



Park, Peter J. "ChIP–seq: Advantages and Challenges of a Maturing Technology." Nature Reviews Genetics 10, no. 10 (October 2009): 669–80. https://doi.org/10.1038/nrg2641.

Control sample is important

- A control sample is necessary for correcting many artifacts: DNA sequence dependent artifacts, chromatin structure, repetitive regions, etc.
- Importantly, control samples should be sequenced significantly deeper than the ChIP ones in a TF experiment and in experiments involving diffused broad-domain chromatin data. This is to ensure sufficient coverage of a substantial portion of the genome and non-repetitive autosomal DNA regions

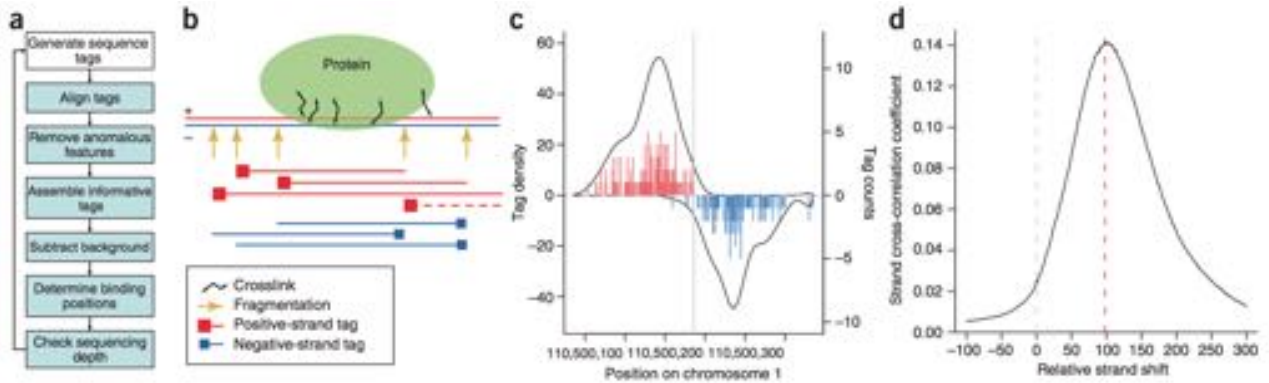


Control sample is important

- There are three commonly used types of control sample:
  1) **Input DNA** (a portion of the DNA sample removed prior to immunoprecipitation (IP))
  2) **Mock IP DNA** (DNA obtained from IP without antibodies)
  3) **DNA from nonspecific IP** (IP performed using an antibody, such as immunoglobulin G, against a protein that is not known to be involved in DNA binding or chromatin modification)

How do peak-finders map binding sites?

- Fragments contain the TF binding site at a (mostly) random position within them
- Reads are randomly generated from left or right edges (sense or antisense) of fragments
- Binding site position = mid-way between sense tag peak and antisense tag peak
- To get binding site peak, shift sense downstream by 1/2 fragsize & antisense upstream by 1/2 fragsize

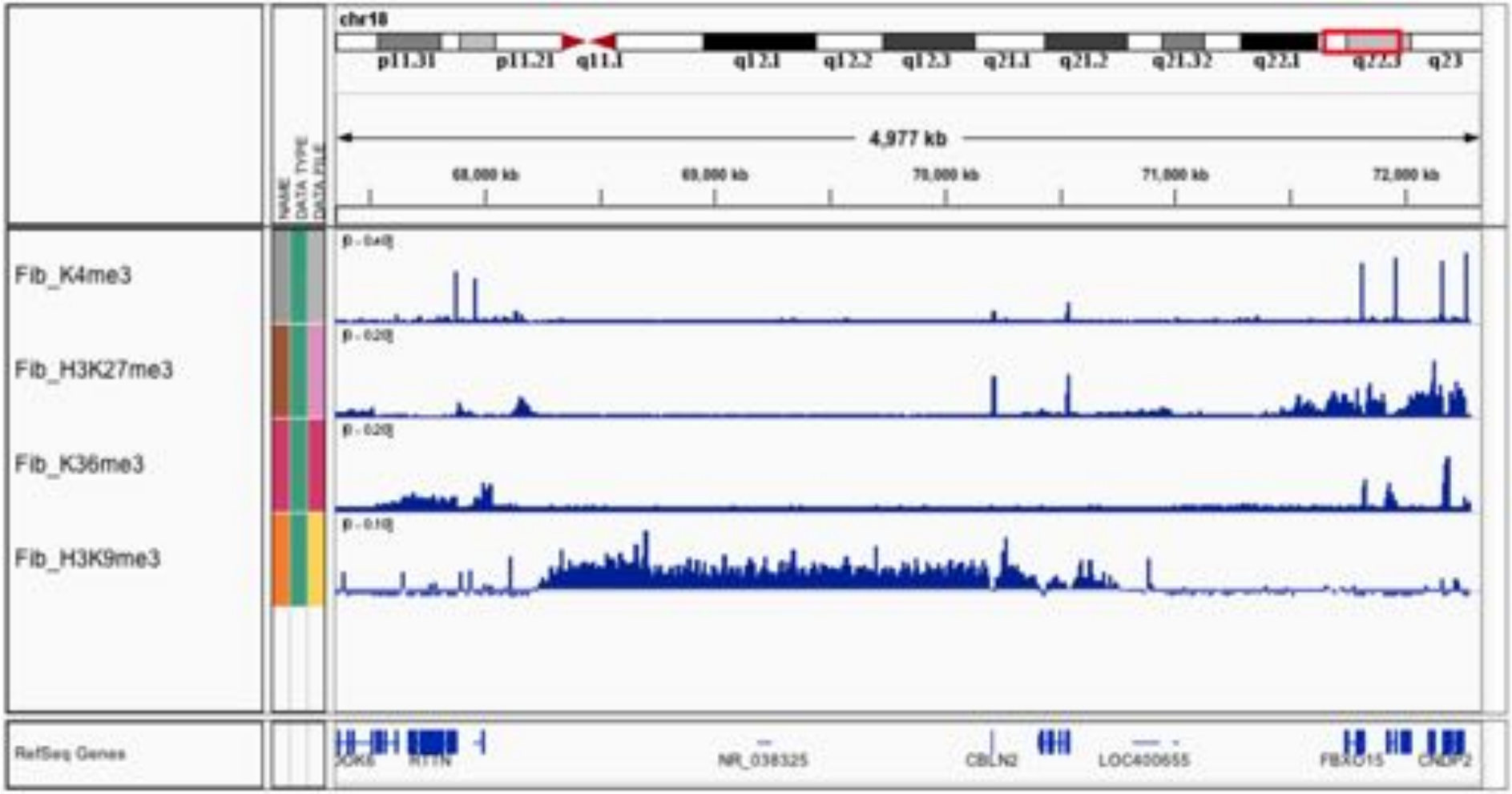


Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." Nature Biotechnology 26, no. 12 (December 2008): 1351–59. https://doi.org/10.1038/nbt.1508.

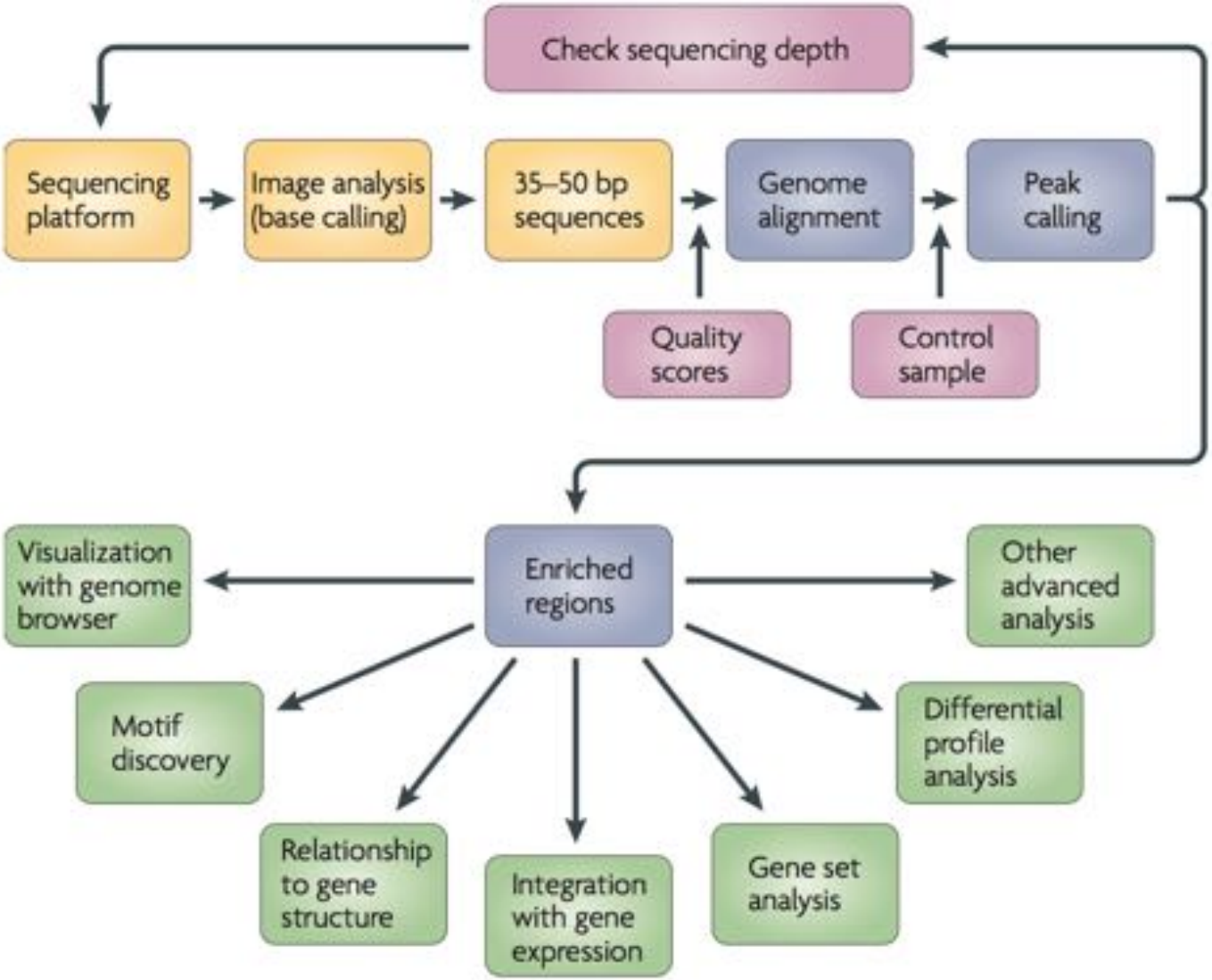
ChIP-seq for histone modification

- Histone modifications have various patterns
- Some are similar to protein binding data, e.g., with tall, sharp peaks: H3K4
- Some have wide (mega-bp) "blocks": H3K9
- Some are variable, with both peaks and blocks: H3K27me3, H3K36me3. Also, RNA Pol II - stalled binding shows as peaks, moving along with transcription shows as broad stretches

Histone modification ChIP-seq data



After peak/block calling

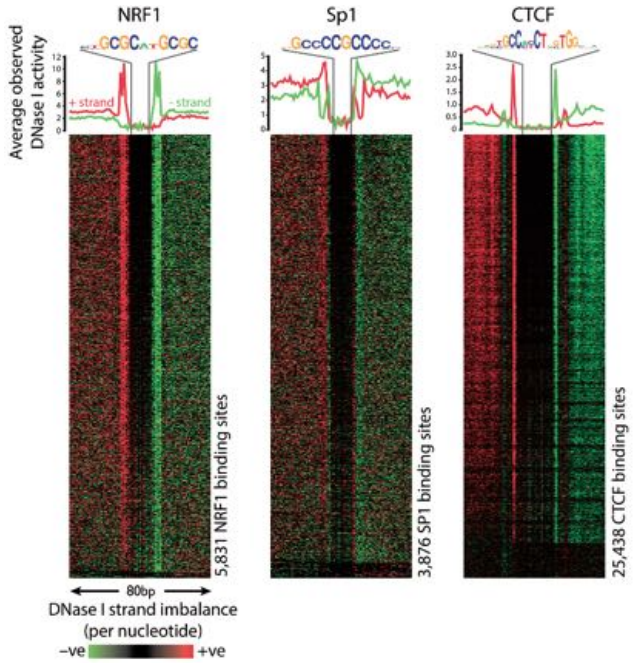


Other "captured/targeted" sequencing technologies

- Enrich and then sequence selected genomic regions.
  - **MeDIP-seq**: measure methylated DNA.
  - **DNase-seq**: detect DNase I hypersensitive sites.
  - **FAIRE-seq**: detect open chromatin sites.
  - **Hi-C**: study 3D structure of chromatin conformation.
  - **GRO-seq**: map the position, amount and orientation of transcriptionally engaged RNA polymerases.
  - **Ribo-seq**: detect ribosome occupancy on mRNA. This is captured RNA- seq.

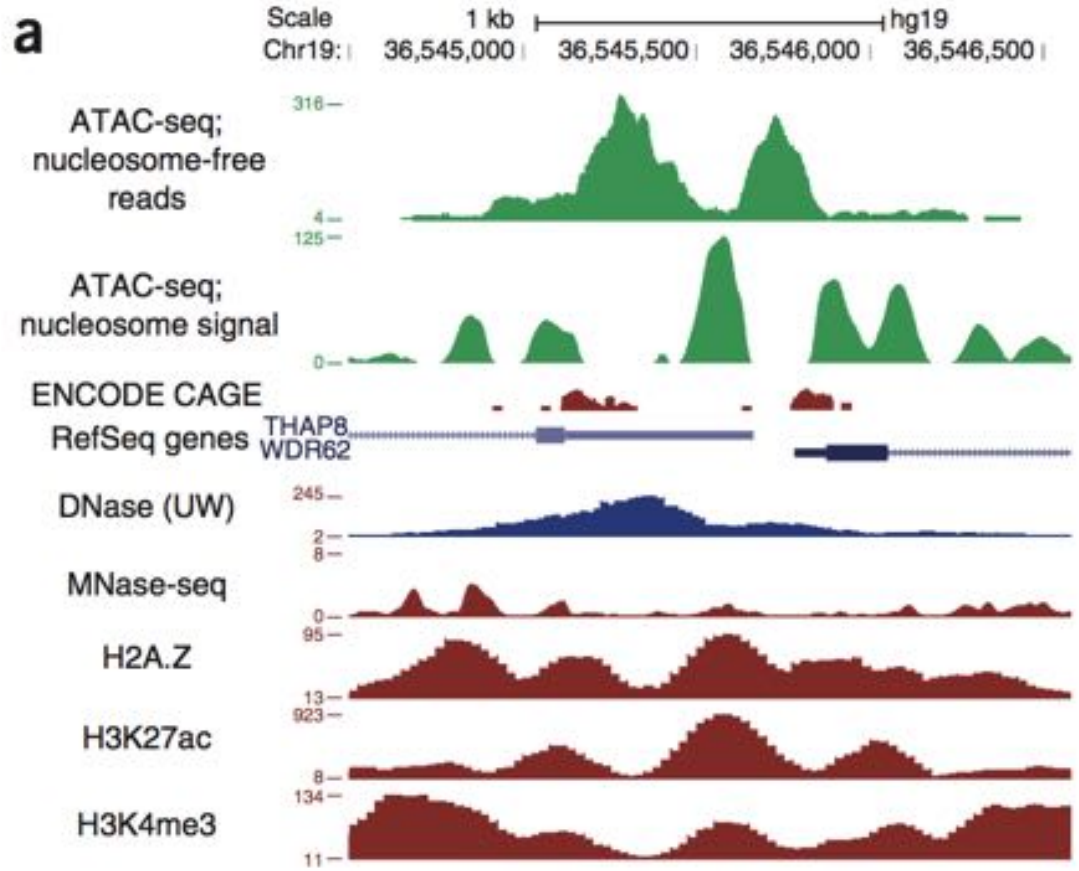
DNAse-seq

- A widely used approach in gene regulation studies uses DNase I as a tool to identify DNase I Hypersensitive Sites (DHSs) within chromatin
- DHSs represent open chromatin regions that are normally only accessible at sites of active regulatory elements such as transcriptional enhancers



Cockerill,P.N. (2011) Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J., 278, 2182–2210.

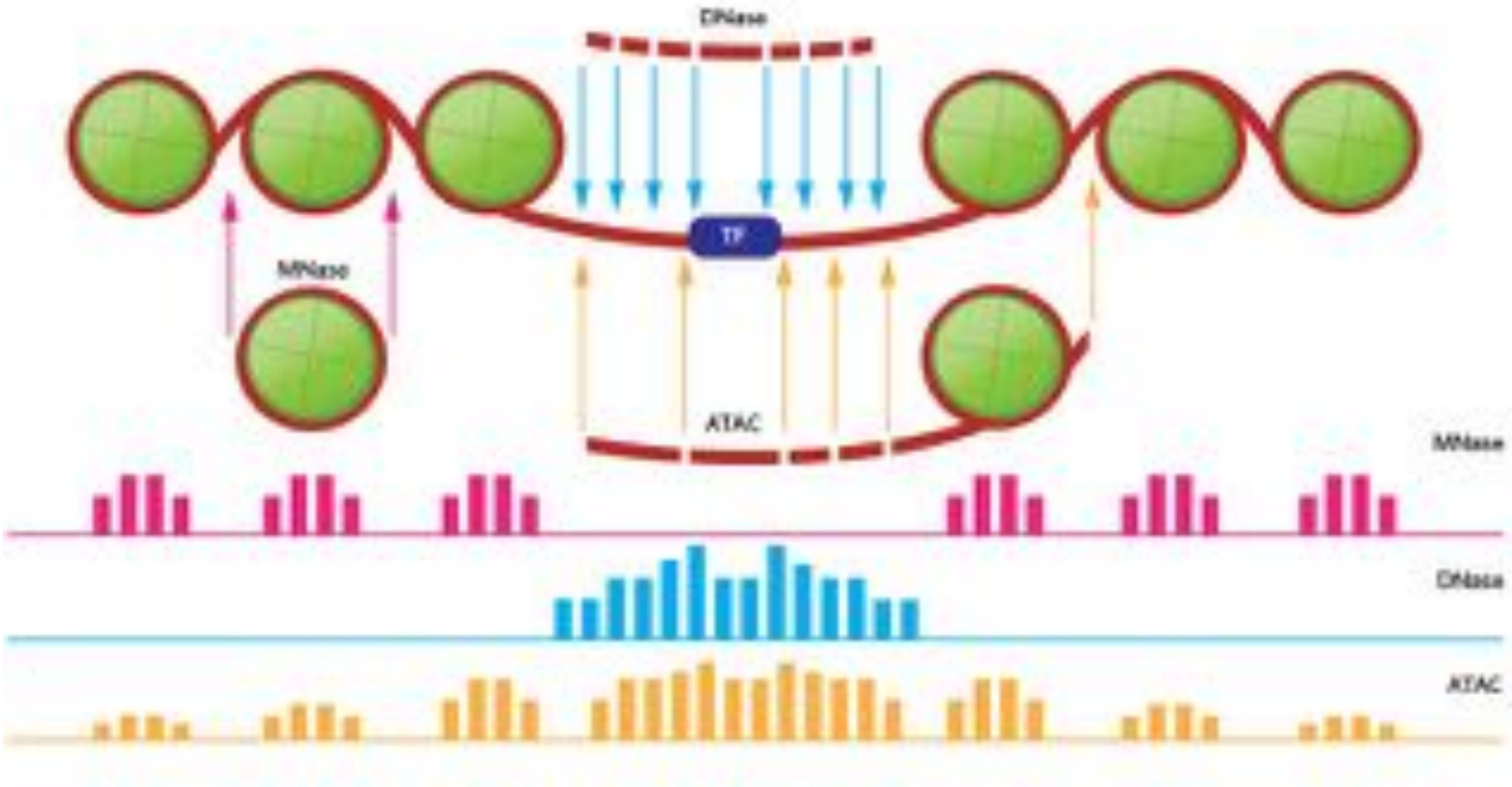
ATAC-seq: finding open chromatin regions



Jason D Buenrostro et al., "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position," Nature Methods 10, no. 12 (December 2013): 1213–18, https://doi.org/10.1038/nmeth.2688.

Technology-specific data

Peaks produced by different technologies are different - calling peaks should be adjusted



Learn more (https://github.com/mdozmorov/Talk_Genomics)

Name	Description	URL	Stars	Watchers	Forks
GENERAL BIOINFORMATICS COLLECTIONS					
Deeplearning-biology	A list of deep learning implementations in biology	https://github.com/hussius/deeplearning-biology	775	148	198
Deep-review	A collaboratively written review paper on deep learning genomics and precision medicine	https://github.com/greenelab/deep-review	742	120	188
Awesome-bioinformatics	A curated list of awesome Bioinformatics libraries and software	https://github.com/danielecook/Awesome-Bioinformatics	583	80	158
Awesome	Awesome resources on Bioinformatics data science machine learning programming language Python Golang R Perl and miscellaneous stuff	https://github.com/shenwei356/awesome	304	21	115
Genomicspapers	The Leek group guide to genomics papers	https://github.com/jtleek/genomicspapers	299	54	134
Biotools	A list of useful bioinformatics resources	https://github.com/jdidion/biotools	205	24	60
Getting-started-with-genomics-tools-and-resources	Unix R and python tools for genomics	https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources	157	27	69
FIELD-SPECIFIC BIOINFORMATICS COLLECTIONS					
Awesome-single-cell	List of software packages for single-cell data analysis including RNA-seq ATAC-seq etc.	https://github.com/seandavi/awesome-single-cell	712	154	303
RNA-seq-analysis	RNAseq analysis notes from Ming Tang	https://github.com/crazyhottommy/RNA-seq-analysis	260	44	104
ChIP-seq-analysis	ChIP-seq analysis notes from Ming Tang	https://github.com/crazyhottommy/ChIP-seq-analysis	252	41	136
Awesome-cancer-variant-databases	A community-maintained repository of cancer clinical knowledge bases and databases focused on cancer variants	https://github.com/seandavi/awesome-cancer-variant-databases	109	23	25
Awesome-10x-genomics	List of tools and resources related to the 10x Genomics GEMCode/Chromium system	https://github.com/johandahlberg/awesome-10x-genomics	63	8	12
DNA-seq-analysis	DNA sequencing analysis notes from Ming Tang	https://github.com/crazyhottommy/DNA-seq-analysis	53	7	34
Awesome-microbes	List of computational resources for analyzing microbial sequencing data	https://github.com/stevetsa/awesome-microbes	33	5	16
DNA-methylation-analysis	DNA methylation analysis notes from Ming Tang	https://github.com/crazyhottommy/DNA-methylation-analysis	25	4	22

Statistical Methods for High-throughput Genomic Data I, II: https://mdozmorov.github.io/BIOS567.2017/syllabus/, https://mdozmorov.github.io/BIOS668.2018/syllabus/

Collections of bioinformatics resources: https://github.com/mdozmorov/blogs/tree/master/Bioinformatics

https://www.frontiersin.org/articles/10.3389/fbioe.2018.00198/full