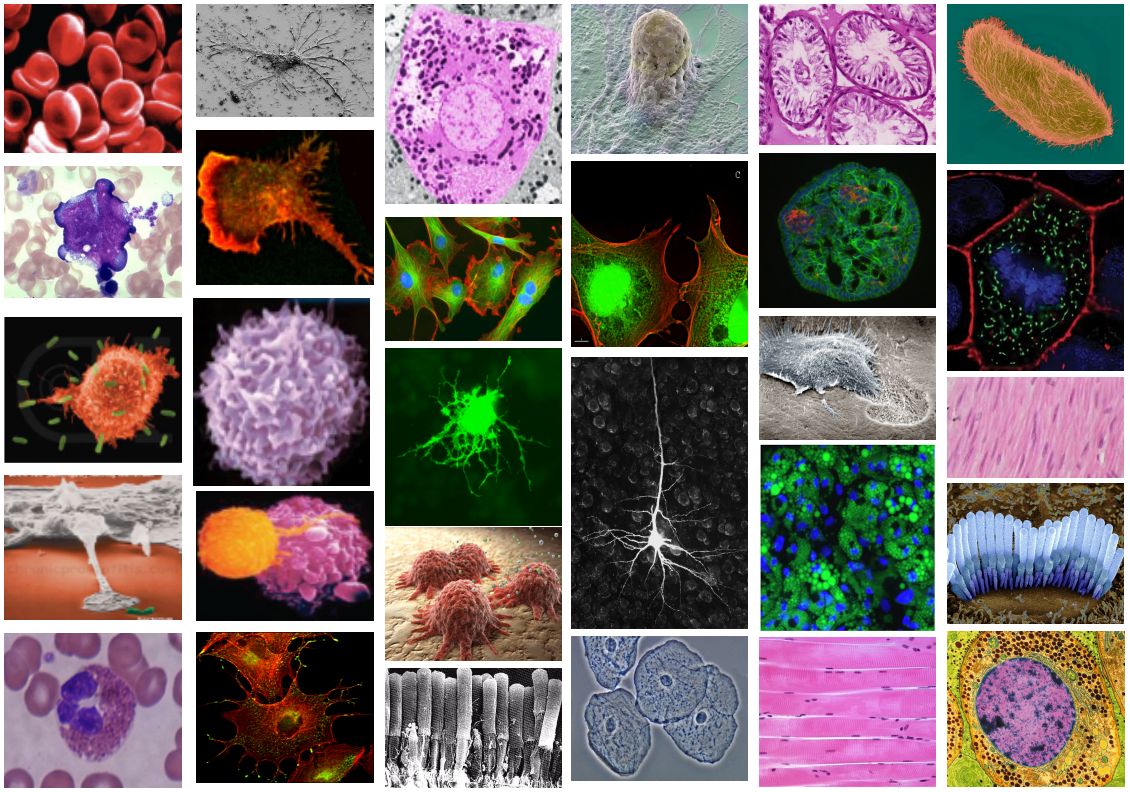
Computational Methods for RNA Sequencing Analysis

ASCB Conference Workshop

Dec. 13, 2015

What controls and determines cell fate and function



Sequencing: applications

Counting applications

- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
  - 3D genome conformation
- Nucleic acid Interactions

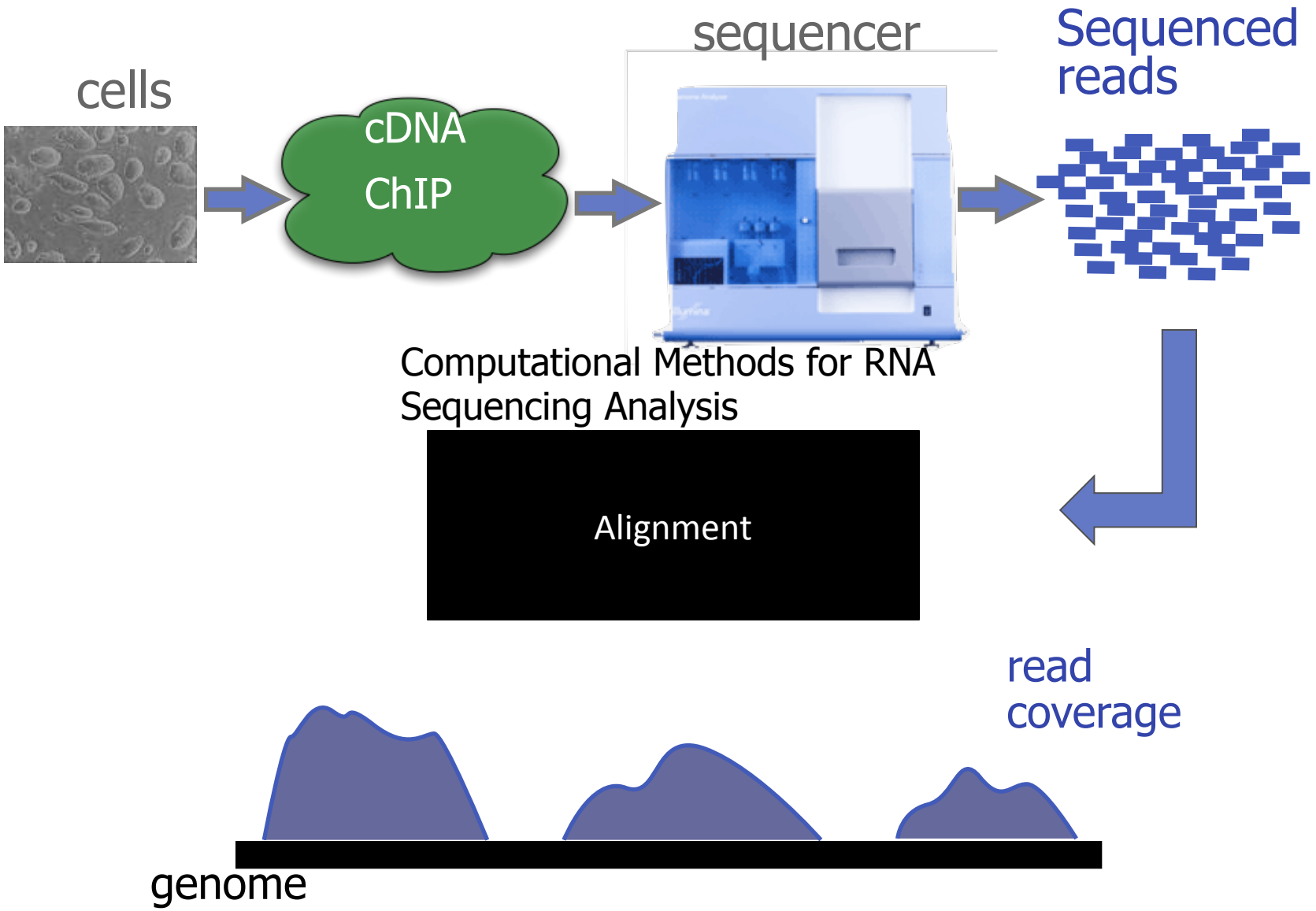
Polymorphism/mutation discovery

  - Bacteria
  - Genome dynamics
  - Exon (and other target) sequencing
  - Disease gene sequencing
- Variation and association studies
- Genetics and gene discovery

Other Applications

- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics

Once sequenced the problem becomes computational



Overview of the session

- RNA Sequencing.
- Processing
- Quantification

Overview of the session

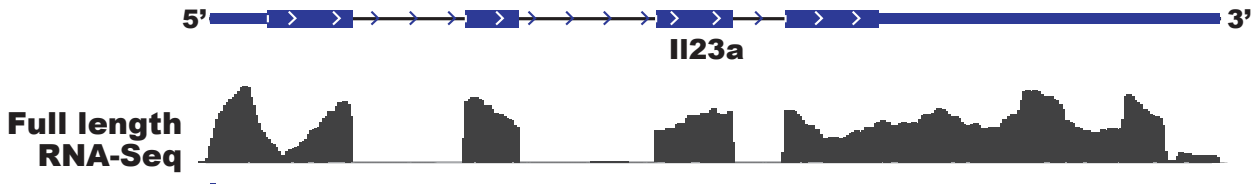
- **RNA Sequencing**
  - **Survey of RNA-Seq libraries.**
- Processing
- Quantification

Sequencing libraries to probe the genome

- **RNA-Seq**
  - **Transcriptional output**
  - **Annotation**
  - **miRNA**
  - **Ribosomal profiling**
- ChIP-Seq
  - Nucleosome positioning
  - Open/closed chromatin
  - Transcription factor binding
- CLIP-Seq
  - Protein-RNA interactions
- Hi-C
  - 3D genome conformation

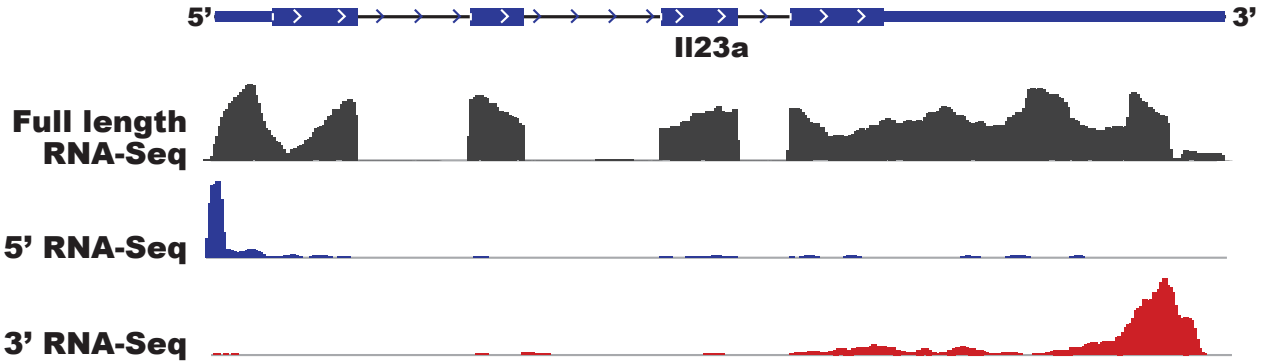
RNA-Seq libraries I: "Standard" full-length

- Source: intact, **high qual** RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
  - Differential Gene expression. Requires ~ 5-10 mill, at least 3 replicates, single end



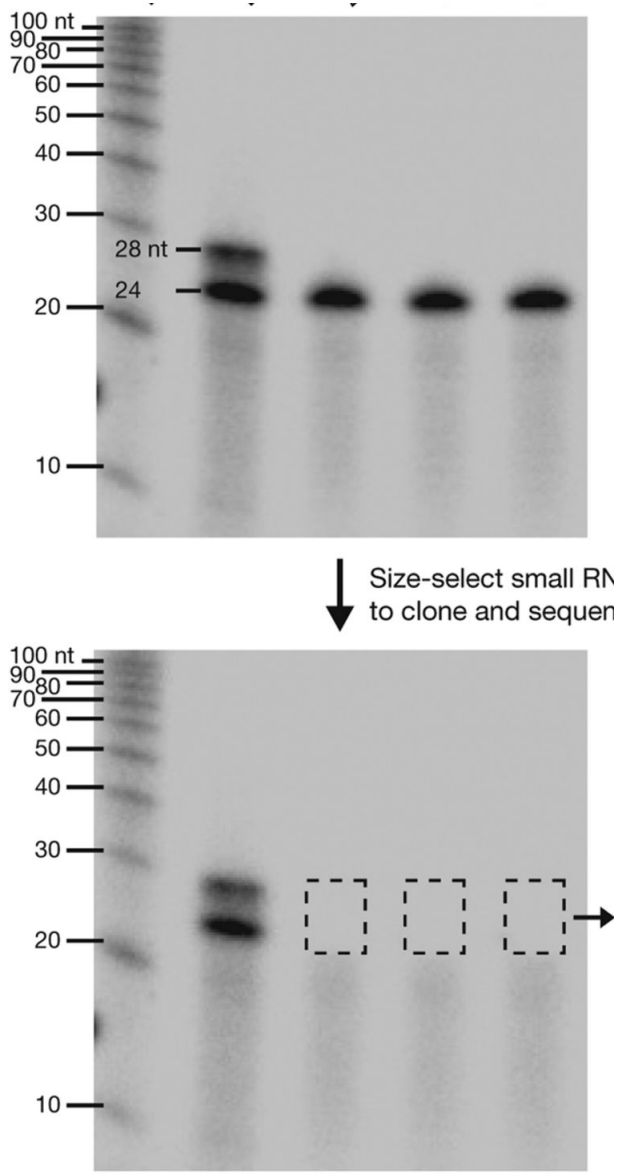
RNA-Seq libraries II: End-sequence libraries

- Target the start or end of transcripts.
- Source: End-enriched RNA
  - Fragmented then selected
  - Fragmented then enzymatically purified
- Uses:
  - Annotation of transcriptional start sites
  - Annotation of 3’ UTRs
  - Quantification and gene expression
  - Depth required 3-8 mill reads
  - Low quality/quantity (single cell) RNA samples



RNA-Seq libraries III: Small RNA libraries

- Source: size selected RNA
- Uses: miRNA, piRNA annotation and quantification
  - Short single end 30-50 bp reads
  - Depth: 3-5 mill reads



Malonne et al. CSHL protocols, 2011

When you need both annotation and quantification

- Attempt three replicates per condition
- Pool libraries to obtain ~15 mill reads per replicate
- Sequence using paired ends
- Analysis:
  - Merge replicate alignments for annotation
  - Split alignments for differential expression analysis

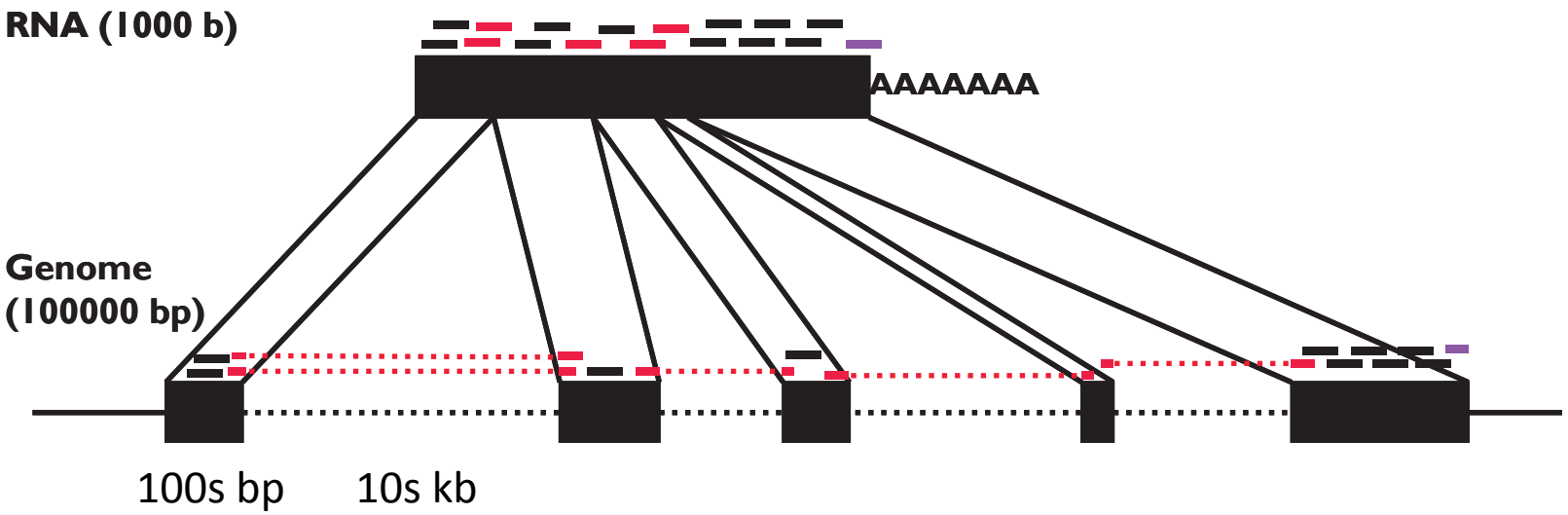
Overview of the session

- RNA Sequencing
  - Different -Seq libraries.
- Processing
  - Read mapping (alignment): Placing short reads in the genome
- Quantification

The short read alignment problem

- Finding 100,000s of small (30-500 bp) sequence in a 10 - 10000 million bp genome.
- Sequences are error prone (~1% error rate)
- Reference and sequence may not be the same haplotype
- **Many techniques are great at finding perfect matches**

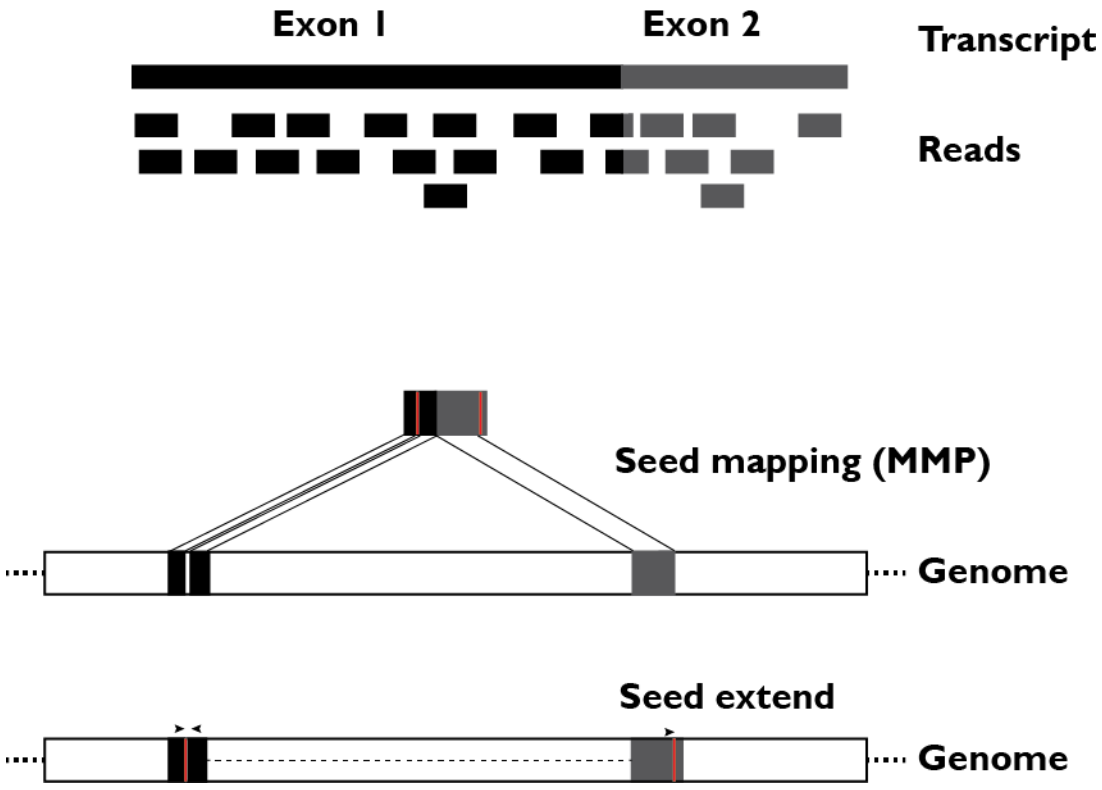
The RNA-Seq alignment problem



Challenges:

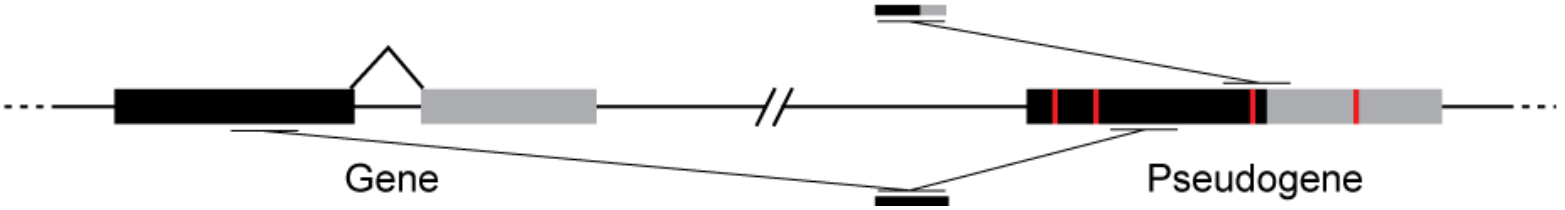
- Genes exist at many different expression levels, spanning several orders of magnitude.
- Reads originate from both mature mRNA (exons) and immature mRNA (introns) and it can be problematic to distinguish between them.
- Reads are short and genes can have many isoforms making it challenging to determine which isoform produced each read.

Mapping RNA-Seq reads

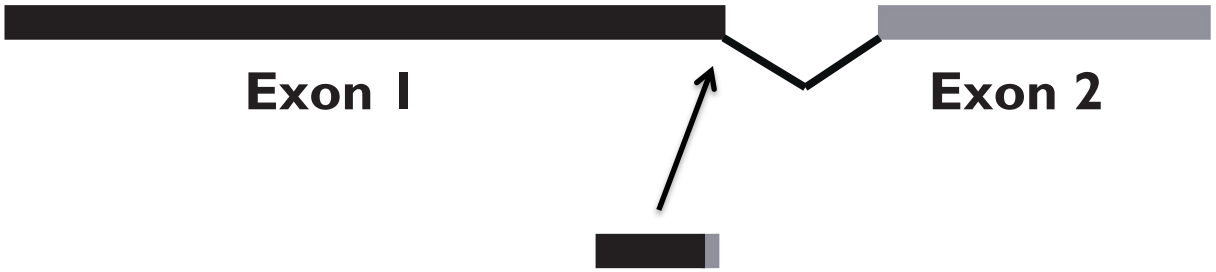


RNA-Seq specific problems

Pseudo gene attraction problem



Intron invasion



Current aligners deal directly with these problems

Short read alignment strategies

Breaks reads into "seeds" that can be perfectly matched

- Create an easily searchable genome (*index*)
  - Hash table: address map of small words (*k-mers*)
  - Suffix Arrays: Efficient way to look up words
  - FA indices (i.e. Burrows Wheelers)
- Seed search using the index:
  - Matching of smaller portions (seeds) of the read
  - Grouping and prioritizing seeds
- Extending seed alignments as read length increases
  - Use algorithms that handle mismatches and gaps
- Align directly to the transcriptome and not the genome
  - Specially when quantifying a good transcriptome

Evolution of the RNA-Seq aligner

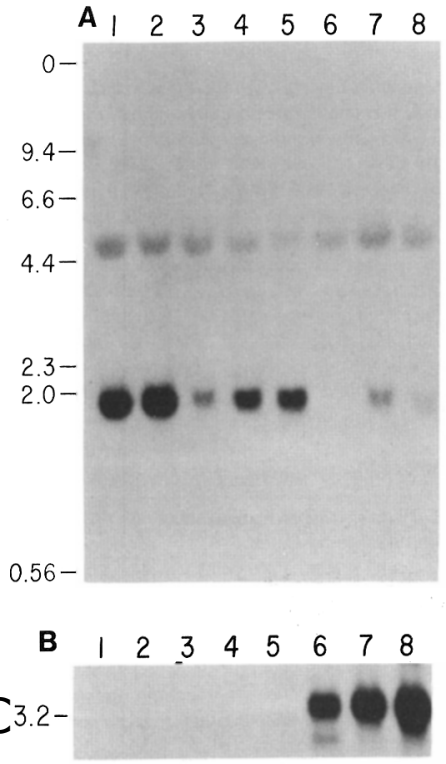
- Tophat (2009)
  - Designed for short reads
  - Requires all reads to be the same length
  - Intron invasion
  - Pseudo-gene attraction
- Tophat2 (2013)
  - Designed to handle long (> 150 bp) reads
  - Specifically handles both Intron invasion & pseudo-gene attraction
    - Automatic re-mapping
- Hisat (2015)
  - Multiple indexing strategy
  - An order of magnitude faster

Overview of the session

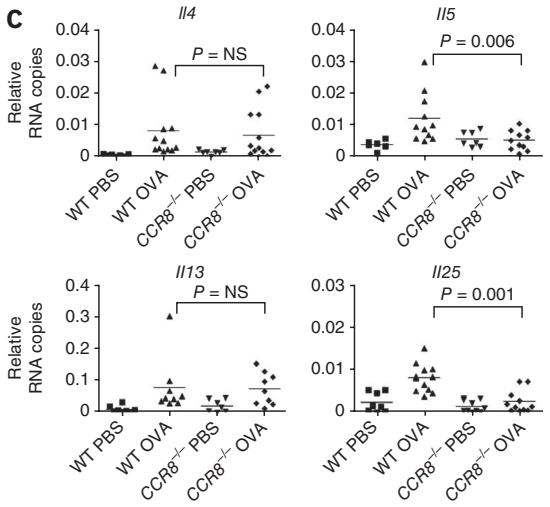
- RNA Sequencing
  - Different -Seq libraries.
- Processing
  - Read mapping (alignment): Placing short reads in the genome
- Quantification
  - Assigning scores to genes/transcripts
  - Determining whether a gene is expressed
  - Normalization
  - Finding genes/transcripts that are differentially represented between two or more samples.

Where is different about sequence data?

80s

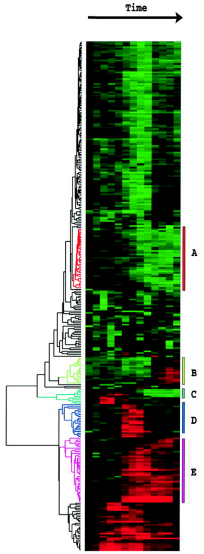


90s



qPCR

Islan et al Nat. Imm. 2011



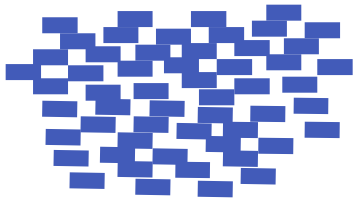
microarrays

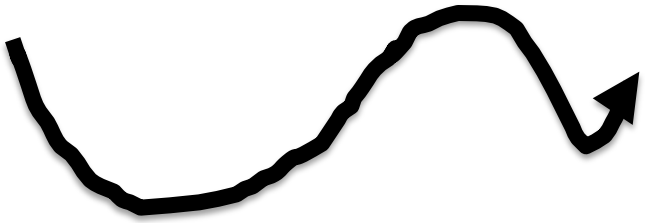
Michael B. Eisen et al. PNAS 1998

Biology slowly becoming a “big data” science

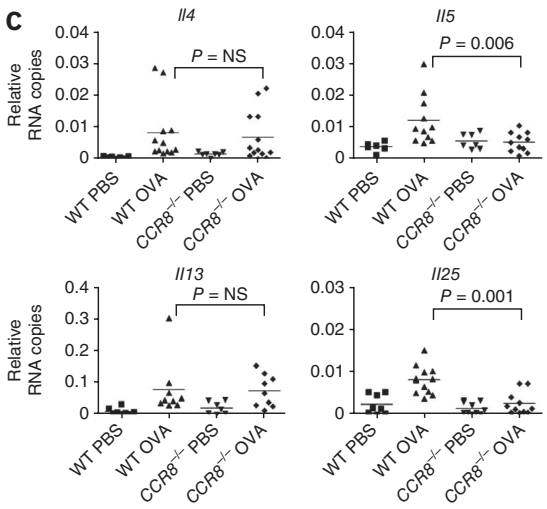
2010s

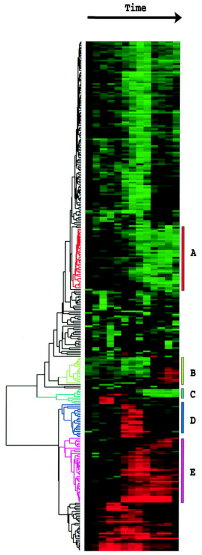
Sequenced reads





Millions-billions

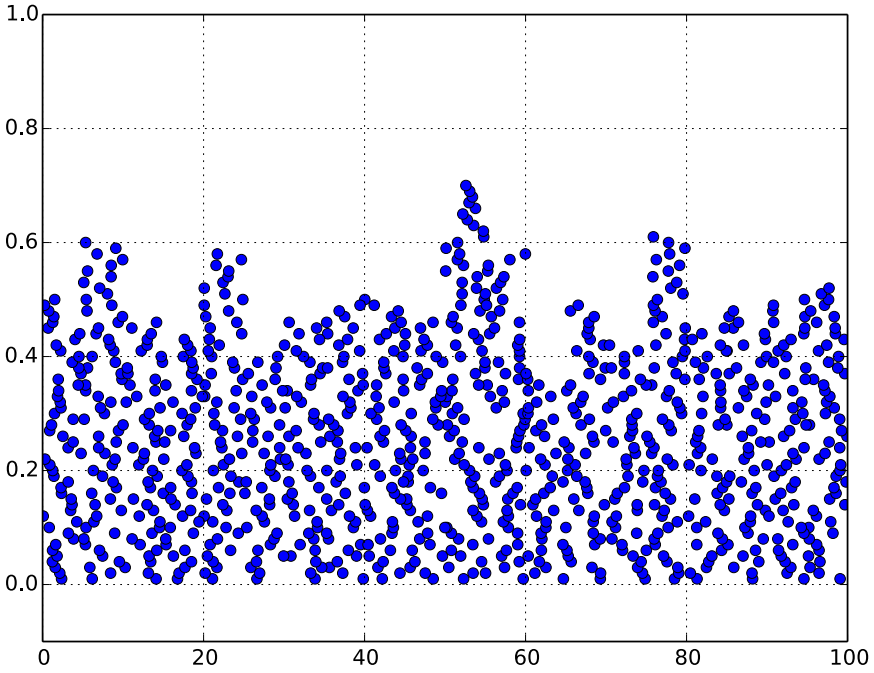


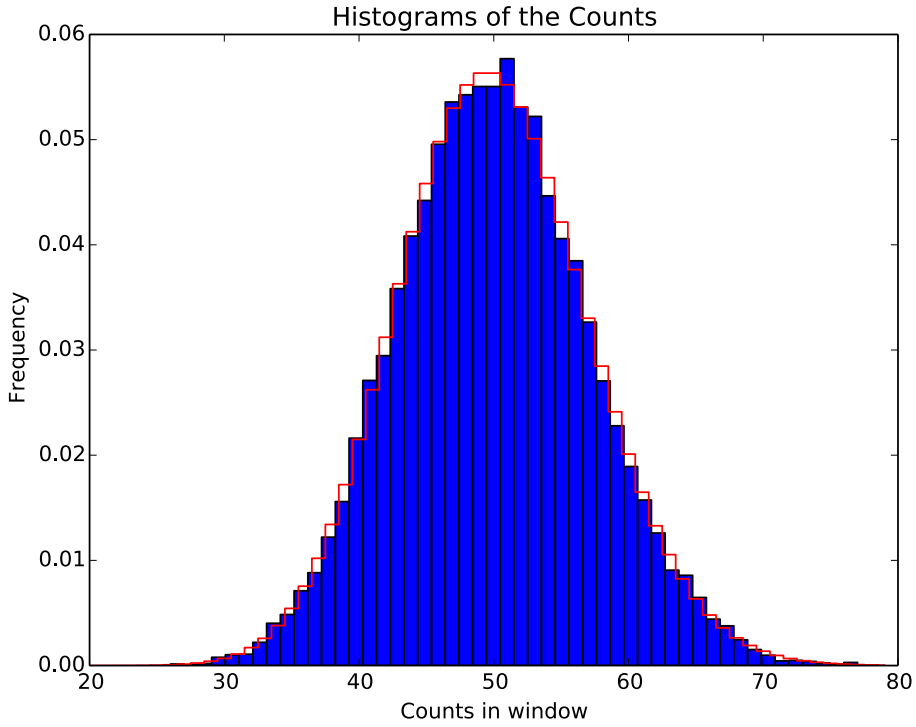


More interesting example: Modeling sequencing data

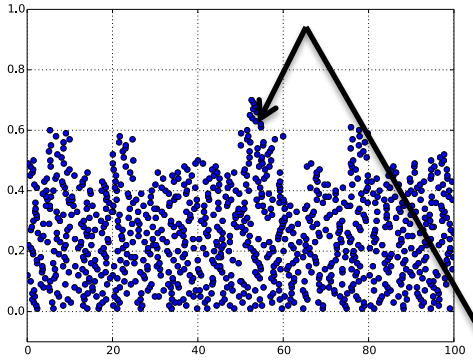
Even randomly generated data using the underlying assumption may look like signal.

Distribution of coverage under the assumption that every position is equally likely to be sequenced.



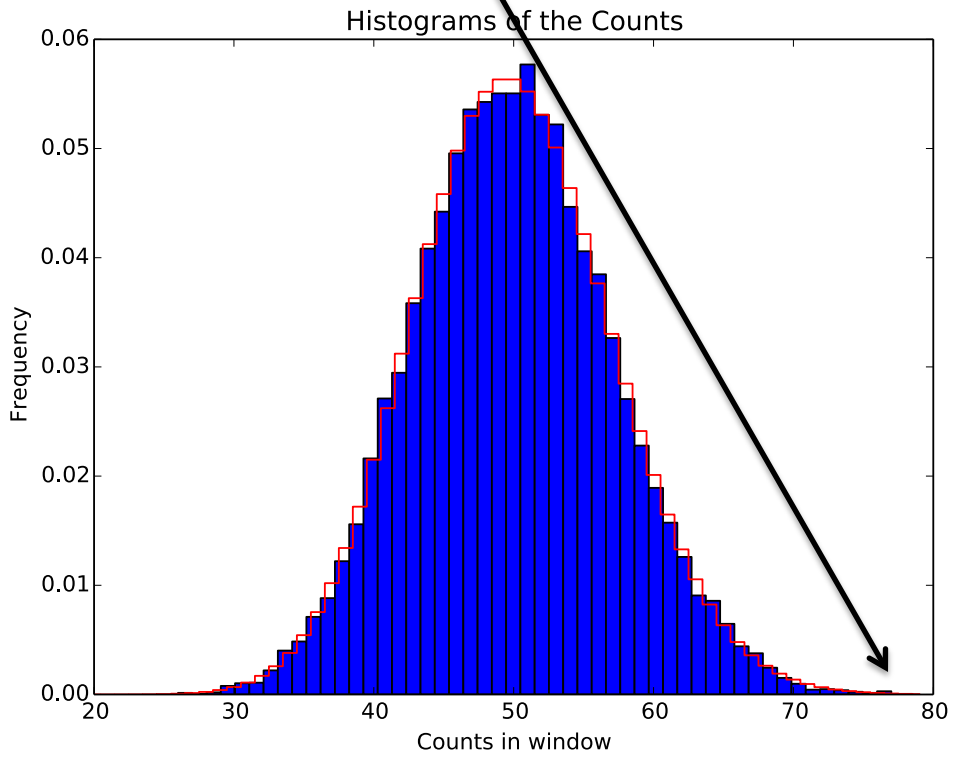


Two very different questions



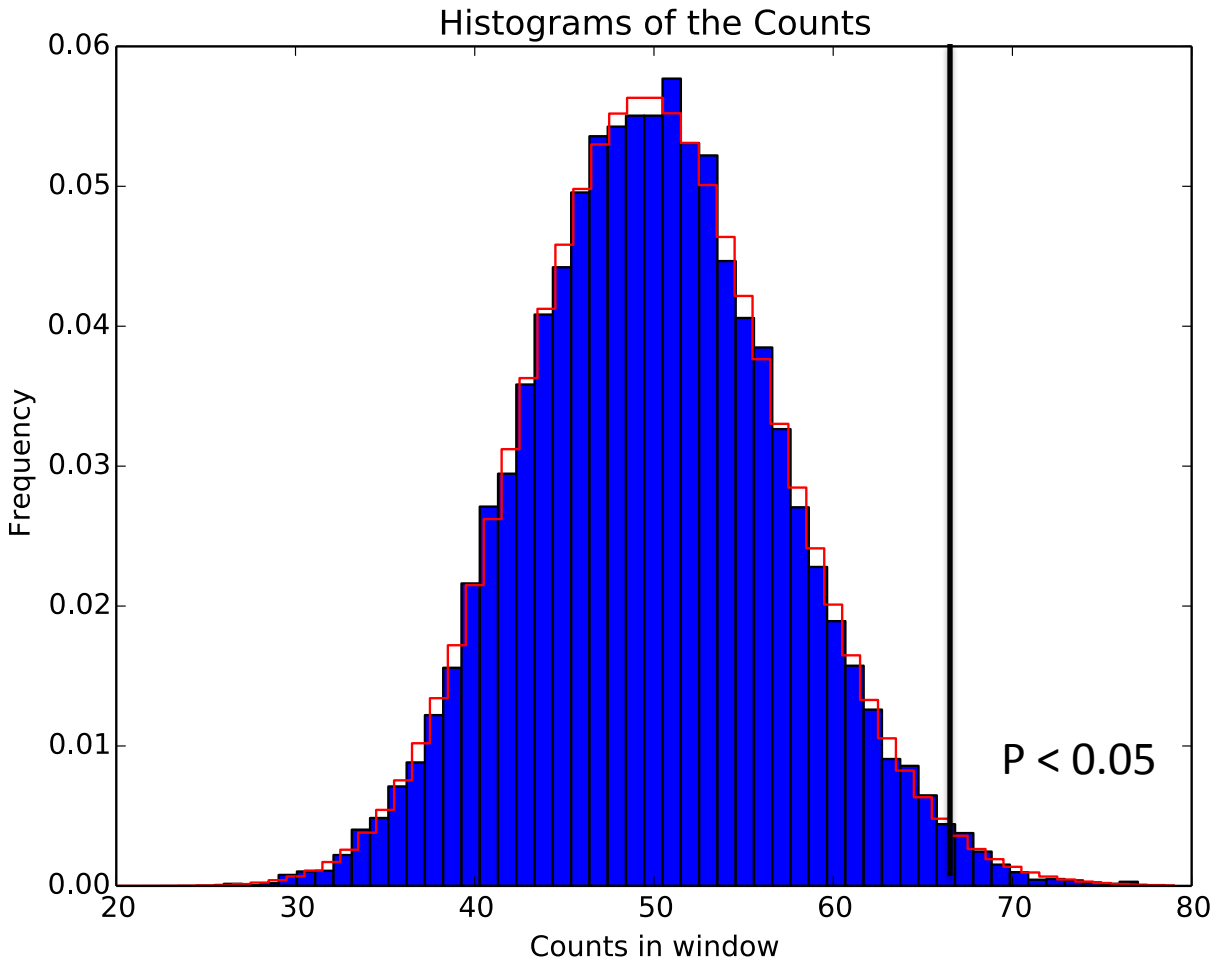
When analyzing a single loci or gene:
Null hypothesis: **coverage of the gene is average**
If I were to choose any other window how likely it is that I get this value?

P (count >gene) = 0



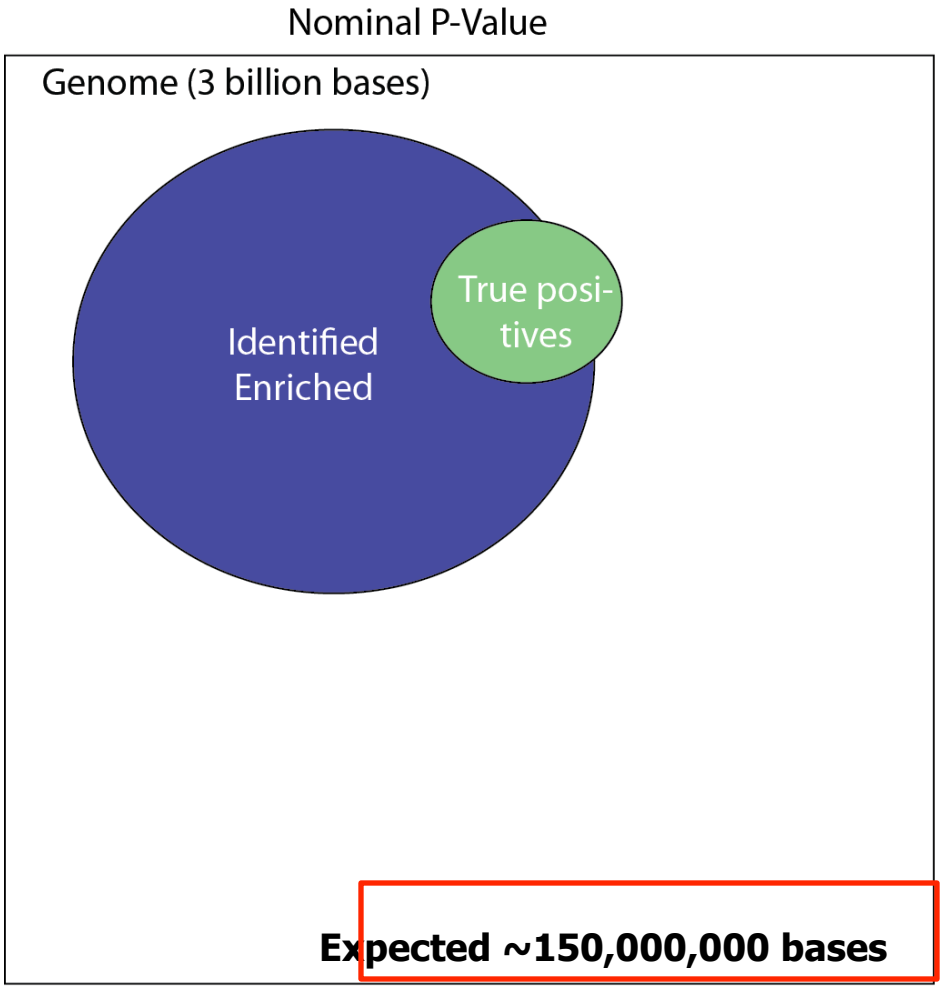
When we sequence the genome we want to **discover** regions or genes that have greater coverage than expected under the null hypothesis. That is we want to identify **outliers**.

We can’t use a nominal p-value any longer



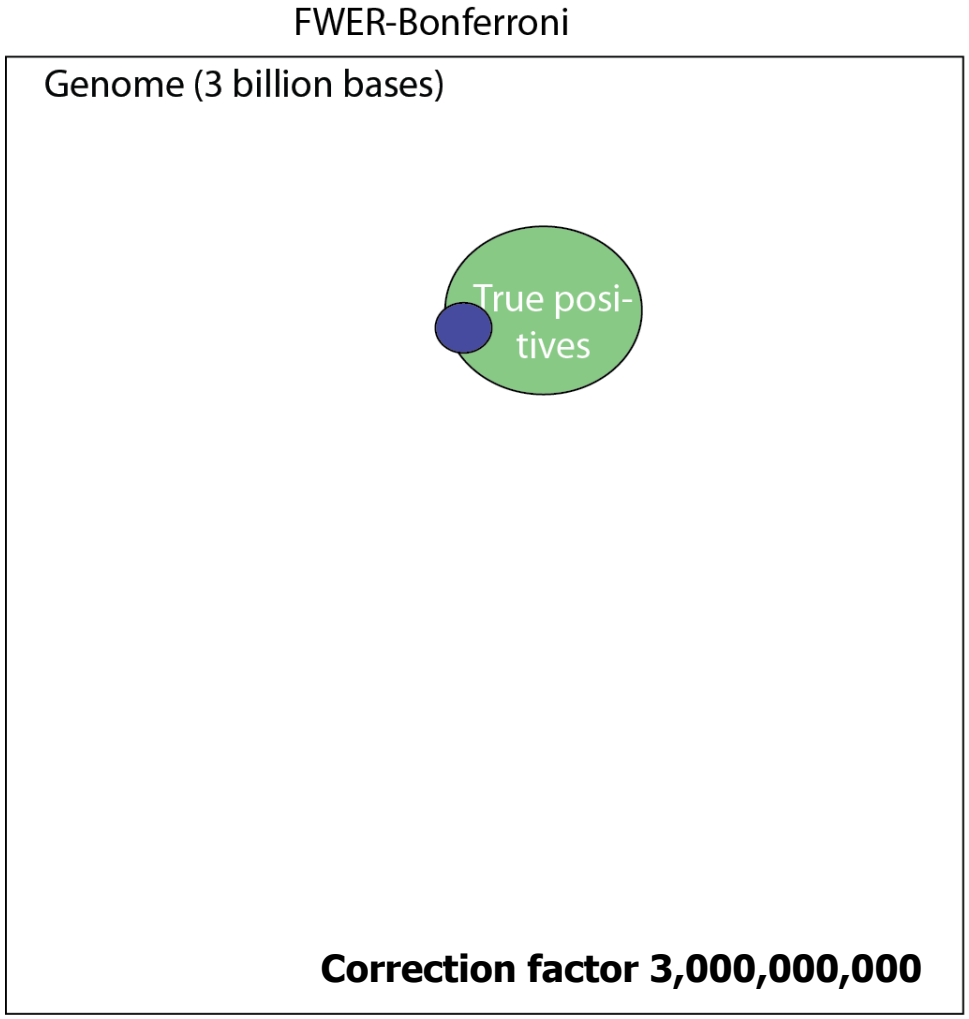
All will be noise!

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative



Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

A better approach: FDR

We test m hypothesis (e.g. gene i is differentially expressed m =20,000).
We wish to detect ($m-m_0$) genes that change between conditions

	Do not pass significance	pass significance	Total
Random noise (Null hypothesis is true)	**U**	**V**	**m_0 = U+V**
True signal (Null hypothesis is false)	**T**	**S**	**$m-m_0$**
Total	**m-R**	**R**	**m**

V = **# Type I errors** (False Positives)
T = **# Type II errors** (False Negatives)

FDR = **E(V | significance)** estimated by the Benjamini-Hochberg procedure

The FDR can be easily computed from nominal p-values!

What does significance means?

- RNA-Seq: The gene is expressed
- ChIP-Seq: Factor binds the region
- CLIP-Seq: Protein binds RNA region
- Ribosomal footprinting:
  - Transcript is translated
  - Ribosomes stalling at region

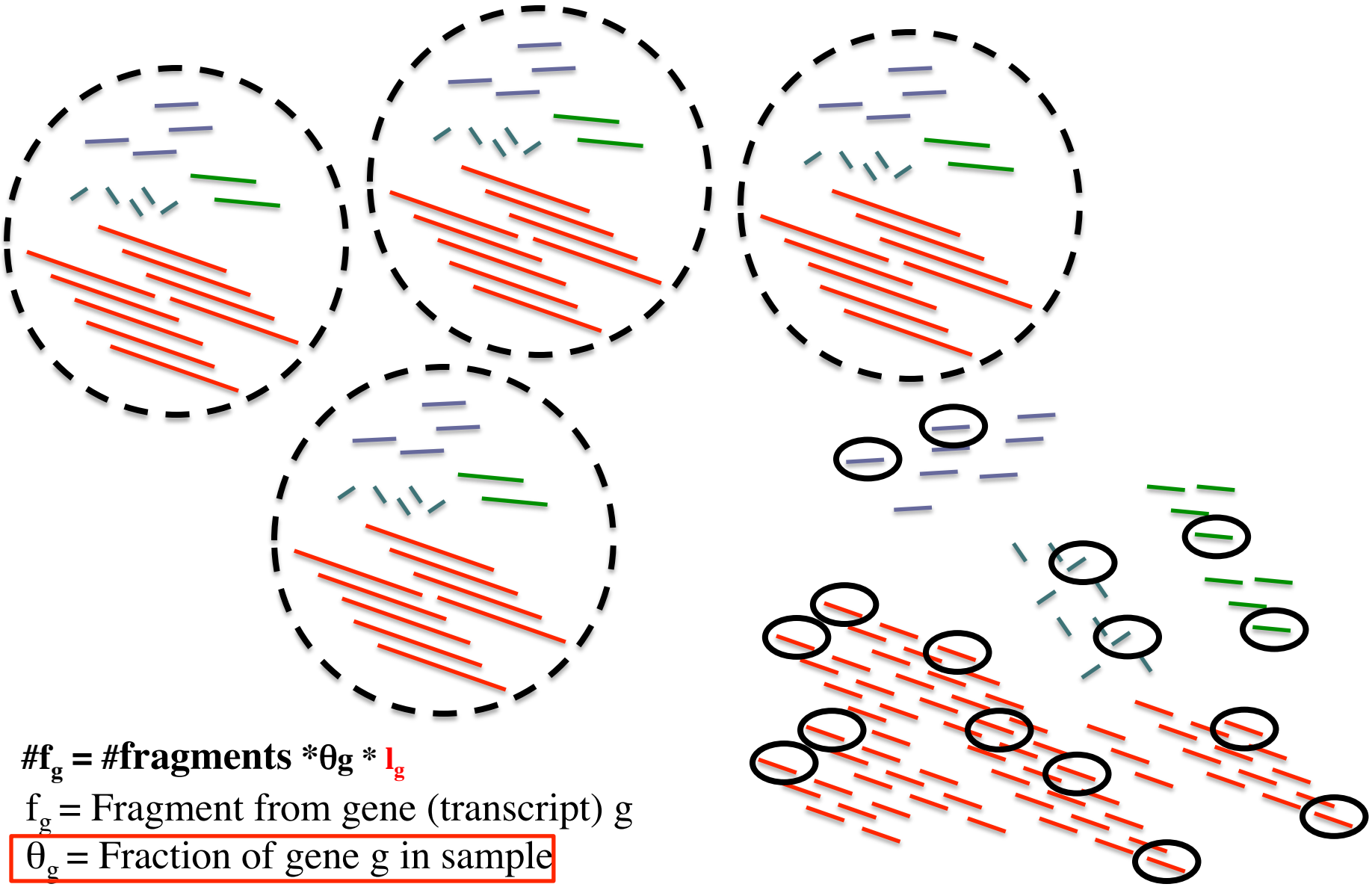
Considerations and assumptions in RNA-Seq

- High library complexity
  - #molecules in library >> #sequenced molecules
- Short reads
  - Read length << sequenced molecule length

Not all applications satisfy this:

- miRNA sequencing
- Small input sequencing (e.g. single cell sequencing)

Quantification given assumptions 1 & 2



#f_g = #fragments *θ_g * l_g

f_g = Fragment from gene (transcript) g

θ_g = Fraction of gene g in sample

l_g = (effective) length of gene g

Corollaries

- Libraries satisfying assumptions 1 & 2 only measure relative abundance
- Key quantity: # fragments sequenced for each transcript.
  - **Key: Which transcript generated the observed read?**
- Isn't this easy?
  - Reads do not uniquely map
  - Transcripts or genes have different isoforms
  - Sequencing has a ~ 1% error rate
  - Transcripts are not uniformly sequenced

The RNA-Seq quantification problem (simple case)

- Start with a set of previous gene/transcript annotations
- Assume only one isoform per gene
- Assume 1-1 read to transcript correspondence.

$$\text{Let } \Theta = \{\theta_g\} \text{ the relative abundance of each gene}$$
$$\text{let } n_g \text{ the number of reads aligned to gene } g$$
$$N = \sum n_g$$

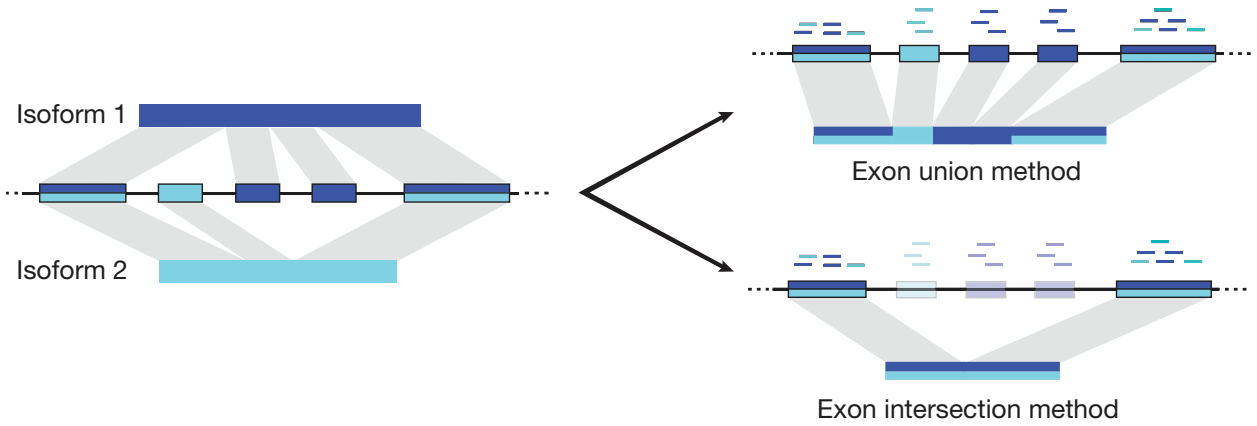
Read counts are a direct estimator of abundance and indeed:

$$\theta_g = \frac{n_g}{\sum n_g}$$

The abundance of gene g is exactly the abundance of the reads that map to it

Application: small RNA sequencing

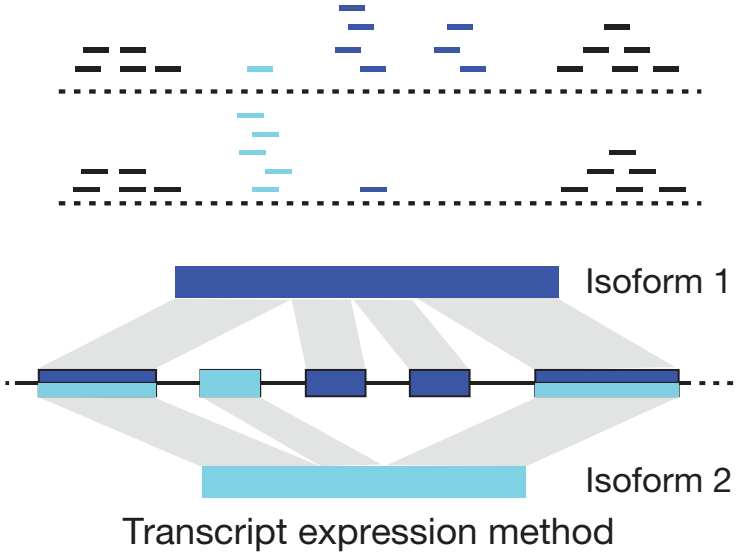
The RNA-Seq quantification problem – With isoforms



- Start with a set of previous gene/transcript annotations
- ~~Assume~~ Define only one isoform per gene
- ~~Assume 1-1 read to transcript correspondence.~~ Reads (fragments) are now short, one transcript generates many fragments.
- **Now, reads depend on both abundance and the transcript length**

$\tilde{l}_g$ Transcript effective length $\theta_g = \frac{n_g}{\tilde{l}_g N}$ The FPKM

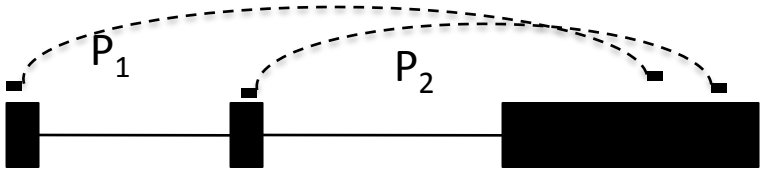
The RNA-Seq quantification problem - Isoform deconvolution



Main difference: quantification involves read assignment. Our model must capture read assignment uncertainty.

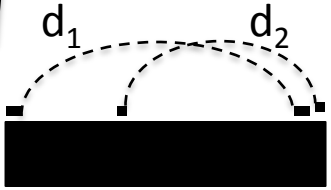
Parameters: Transcript relative abundance
Latent variables: Fragment alignment source
Observed variables: N fragment alignments, transcripts, *fragment length distribution*

We can estimate the insert size distribution

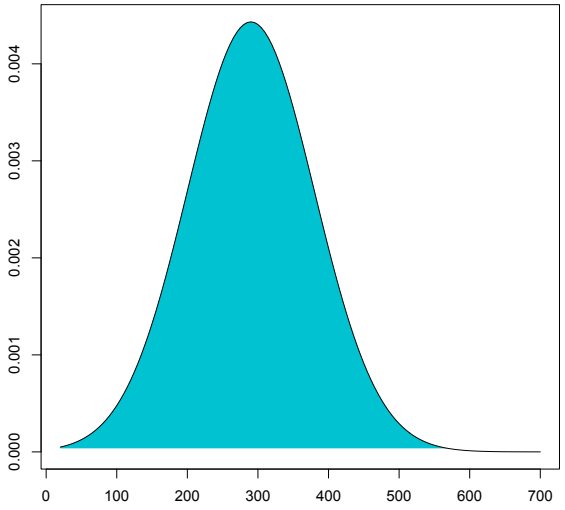


Get all single isoform reconstructions

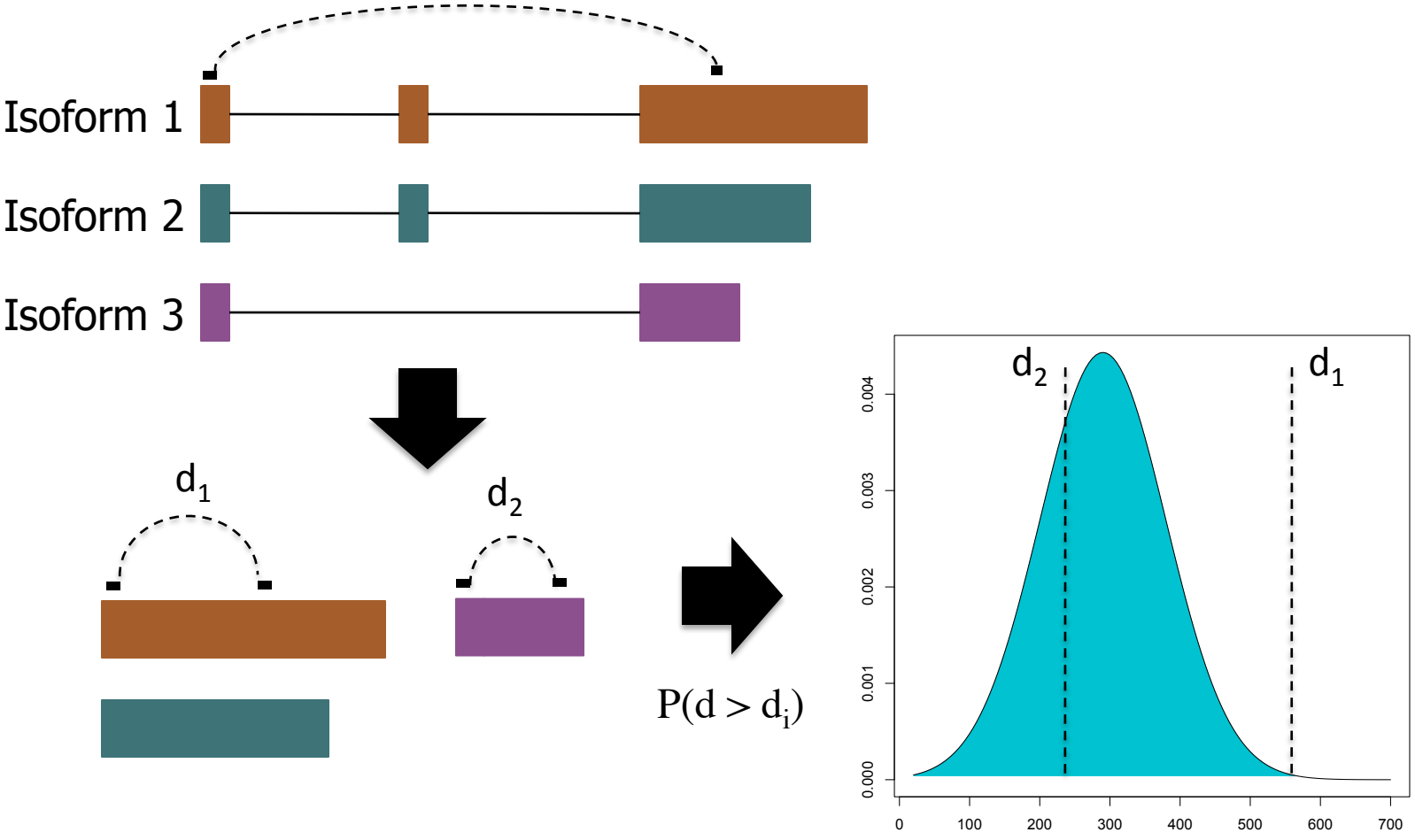
Splice and compute insert distance



Estimate insert size empirical distribution

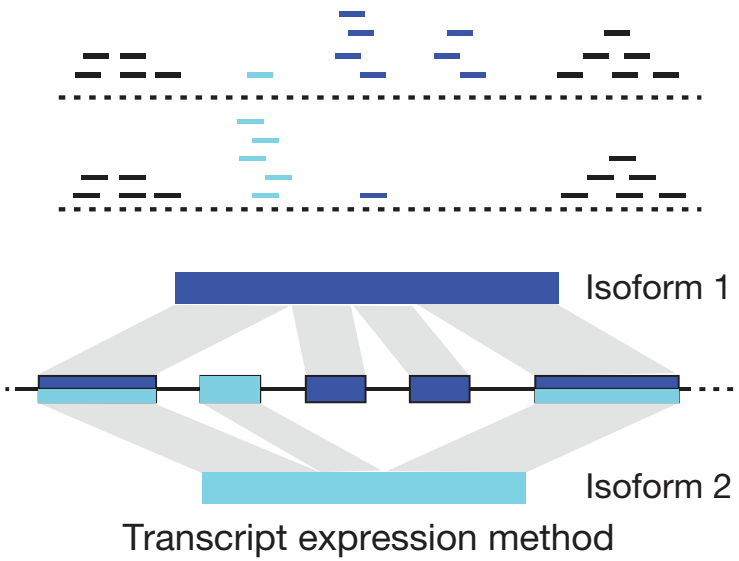


… and use it for probabilistic read assignment

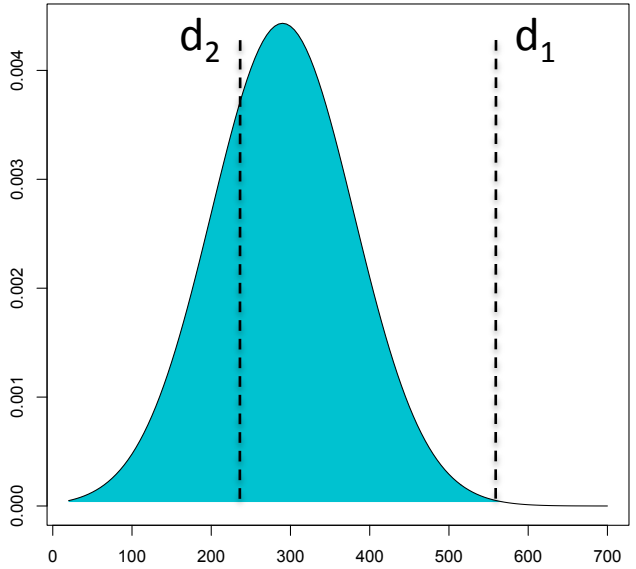


In methods such as MISO, Cufflinks and RSEM, paired-end data is critical

The RNA-Seq quantification problem. Isoform deconvolution



Transcript expression method



Parameters: Transcript relative abundance
Latent variables: Fragment alignment source
Observed variables: N fragment alignments, transcripts, **fragment length distribution**

$$P(a \in t|D,\theta_t) = \frac{\theta_t \tilde{l}_t}{\sum_{a \in s} \theta_s \tilde{l}_s} P(l(a)|t,D)$$

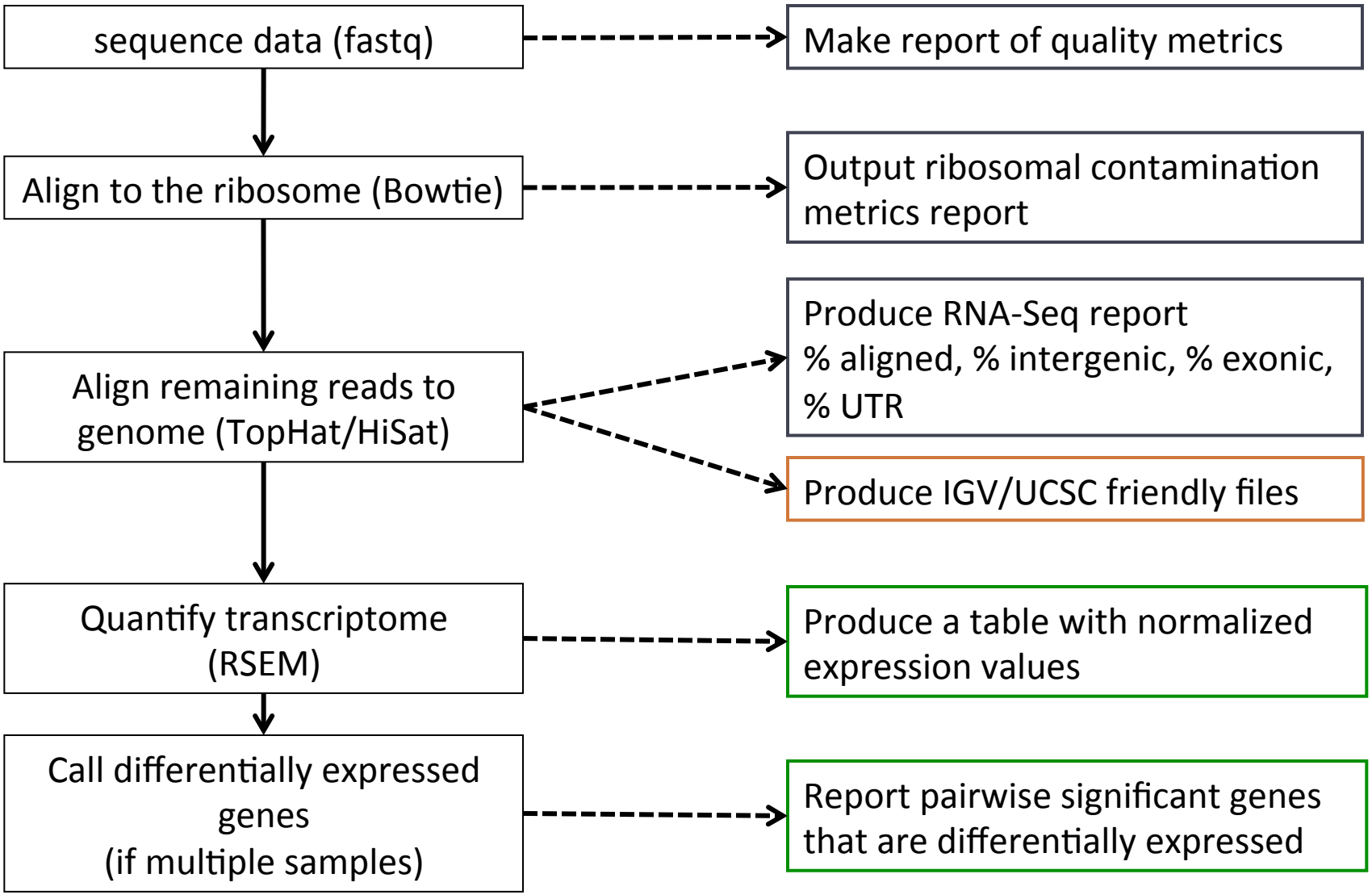
Probability of the fragment alignment originating from t

$$\mathcal{L}(\Theta \mid D, A, G) = \prod_{t \in G} \prod_{a \in t} P(a \in t|D,\theta_t)$$

RNA-Seq quantification summary

- Counts must be estimated from ambiguous gene/ transcript assignment.
  - Using simplified gene models (intersection)
  - Probabilistic read assignment
- Counts must be normalized
  - RPKM/FPKM/TPM are designed for intra-library comparisons:
    - Is gene A more highly expressed than gene B

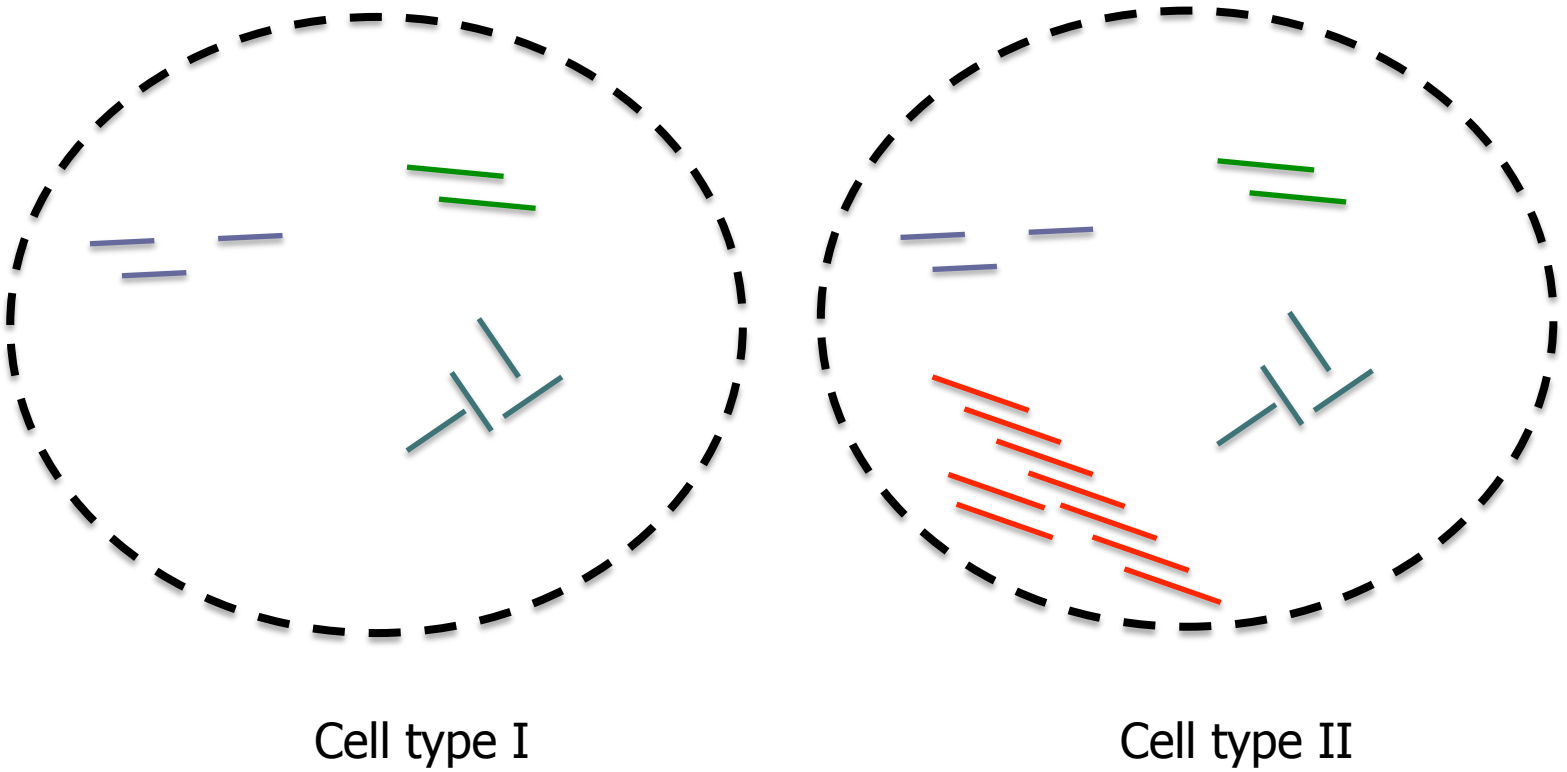
Our typical pipeline (e.g. RNA-Seq)



Overview of the session

- RNA Sequencing
- Processing
- Quantification
  - Assigning scores to genes/transcripts
  - Determining whether a gene is expressed
  - Normalization
  - Finding genes/transcripts that are differentially represented between two or more samples.

Sample composition impacts transcript ***relative*** abundance



Cell type I

Cell type II

Normalizing by total reads does not work well for samples with very different RNA composition

Example normalization techniques

Counts for gene i in experiment j

$$s_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}.$$

Geometric mean for that gene over ALL experiments

	c_1	c_2	c_3	c_4
g_1	k_{11}	k_{12}	k_{13}	k_{14}
g_2	k_{21}	k_{22}	k_{23}	k_{24}
g_3	k_{31}	k_{32}	k_{33}	k_{34}
g_4	k_{41}	k_{42}	k_{43}	k_{44}
g_5	k_{51}	k_{52}	k_{53}	k_{54}

i runs through all *n* genes
j through all *m* samples
k_{ij} is the observed counts for gene *i* in sample *j*
s_j Is the normalization constant

Alders and Huber, 2010

Lets do an experiment (and do a short R practice)

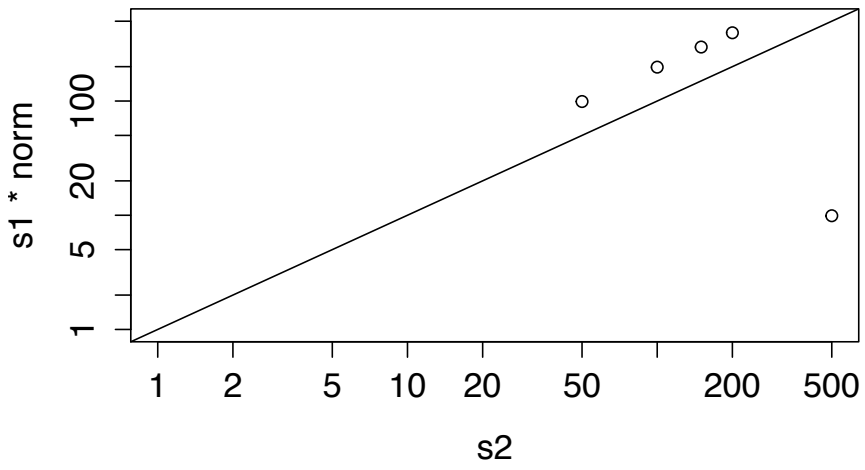
$> s1 = c(100, 200, 300, 400, 10)$
$> s2 = c(50, 100, 150, 200, 500)$

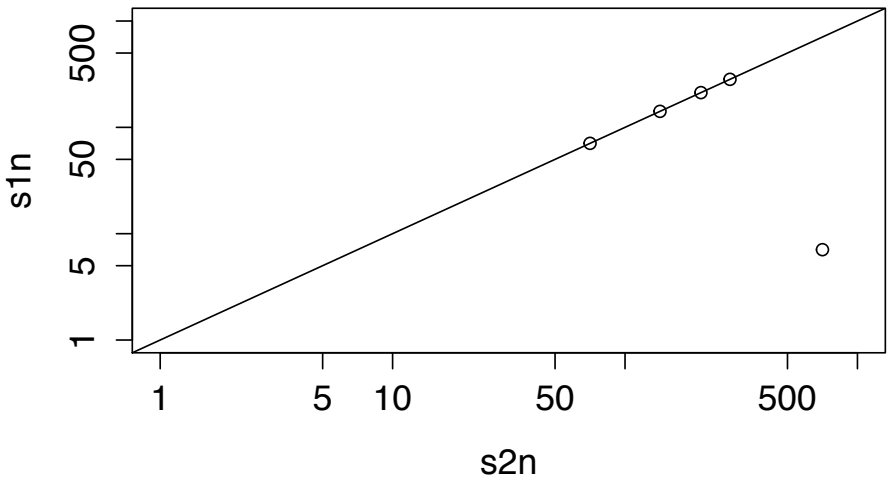
Similar read number,
one transcript many fold changed

$>\text{norm}=\text{sum}(s2)/\text{sum}(s1)$
$>\text{plot}(s2, s1*\text{norm},\text{log}=\text{"xy"})$
$>\text{abline}(a = 0, b = 1)$

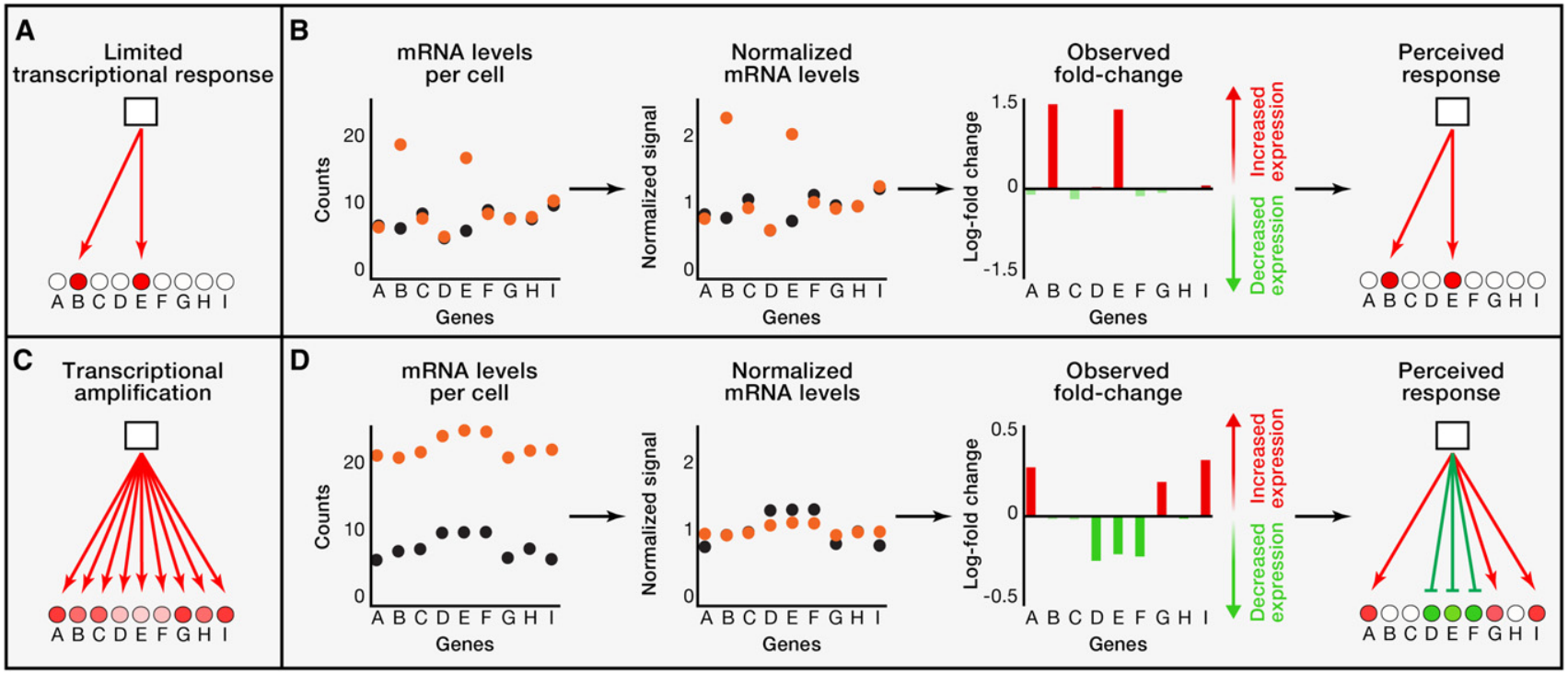
Size normalization results in 2-fold changes in *all* transcripts

$>\text{g} = \text{sqrt}(s1 * s2t)$
$>\text{s1n} = s1/\text{median}(s1/g);\ \ \text{s2n} = s2/\text{median}(s2/g)$
$>\text{plot}(s2n, s1n,\text{log}=\text{"xy"})$
$>\text{abline}(a = 0, b = 1)$





When everything changes: Spike-ins



Lovén et al, Cell 2012

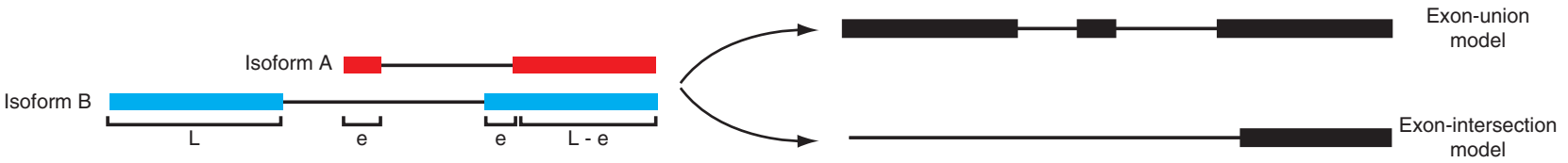
Differential Gene Expression Questions

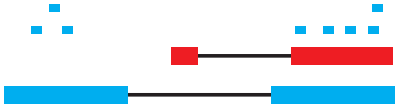
- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3’ UTRs

General strategy for differential gene expression

- Normalize *count* data
  - Key: We only compare each gene across samples NOT one gene to another.
- Estimate normalized mean gene counts
- Estimate *gene variance*
  - Assume variance is similar for similarly expressed transcripts
  - Model variance as a function of expression
- Define a test
  - DESeq: Generalization of a fisher exact test
  - Cufflinks: Log transformed of FPKMs divided by its variance (~ normally distributed).
    - Null hypothesis: log ratio = 0

Why not just simple models?



Condition A	Condition B	Log fold-change (union count)	Log fold-change (intersect count)	Log fold-change (true expression)
		$\log_2\left(\frac{10}{10}\right) = 0$	$\log_2\left(\frac{8}{7}\right) = 0.19$	$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
		$\log_2\left(\frac{6}{8}\right) = -0.41$	$\log_2\left(\frac{5}{5}\right) = 0$	$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
		$\log_2\left(\frac{5}{10}\right) = -1$	$\log\left(\frac{4}{5}\right) = -0.1$	$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Trapnell et al. Nat. Biotech 2012

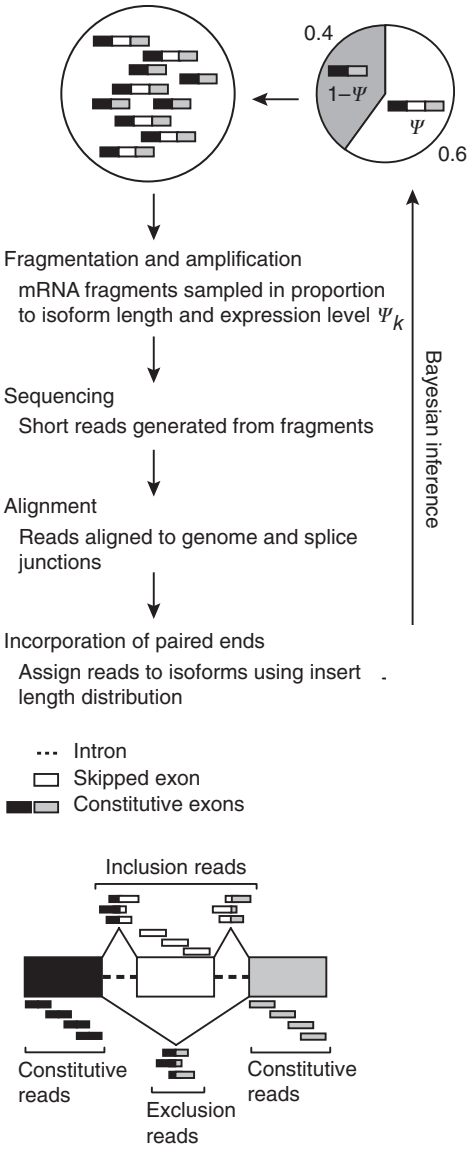
Differential analysis strategies

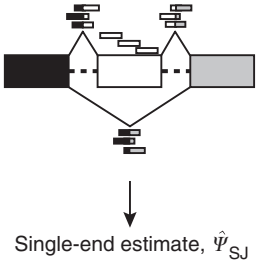
Use read counts and Standard Fisher exact test

	Condition A	Condition B
Gene A reads	n_a	n_b
Rest of reads	N_a	N_b

– Not naturally extendable to experiments with replicates

MISO: Specifically testing exon inclusion





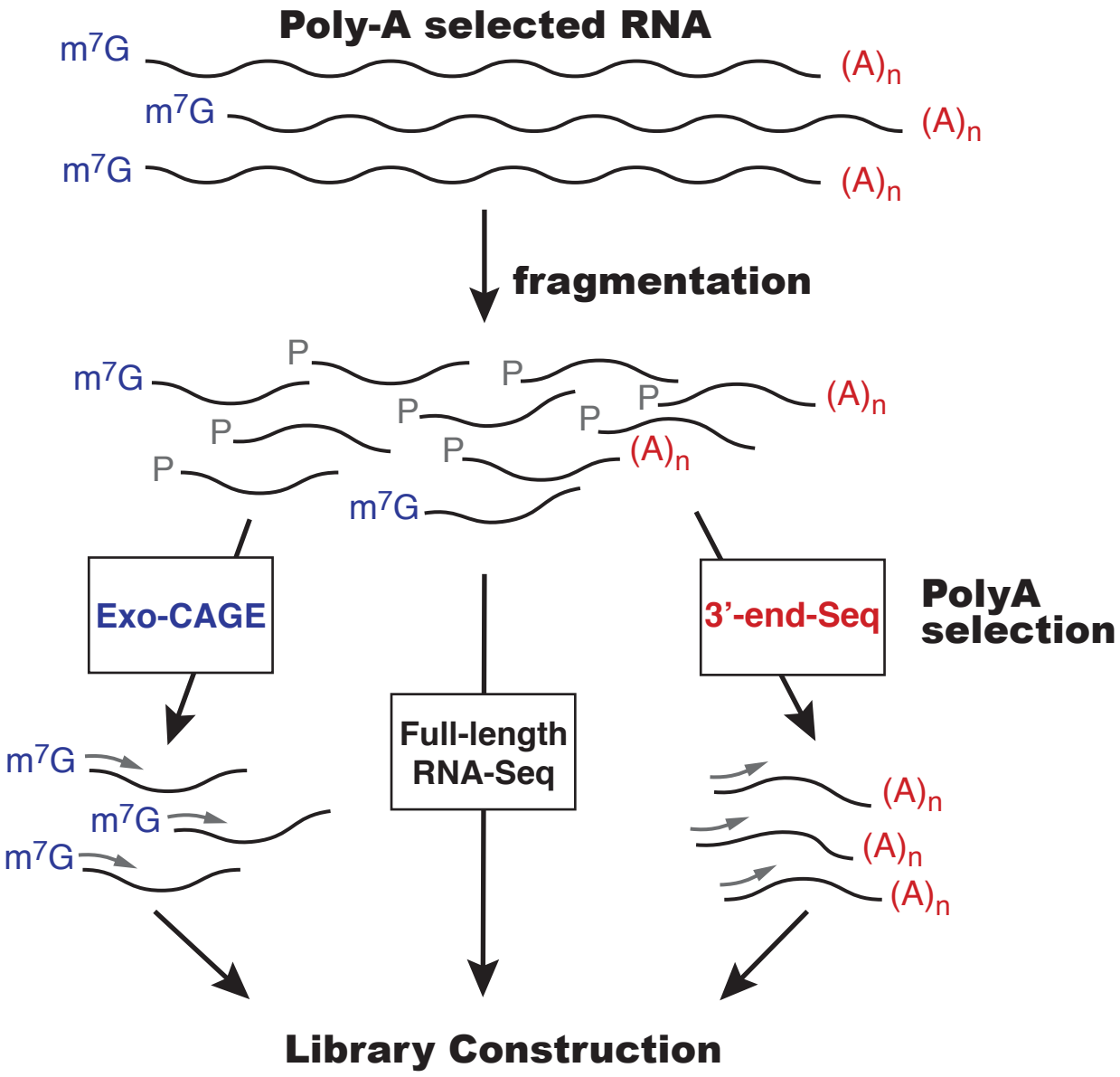
Katz et al Nat. Methods 2010

Digital Gene Expression (DGE)

The quest for inexpensive expression assays

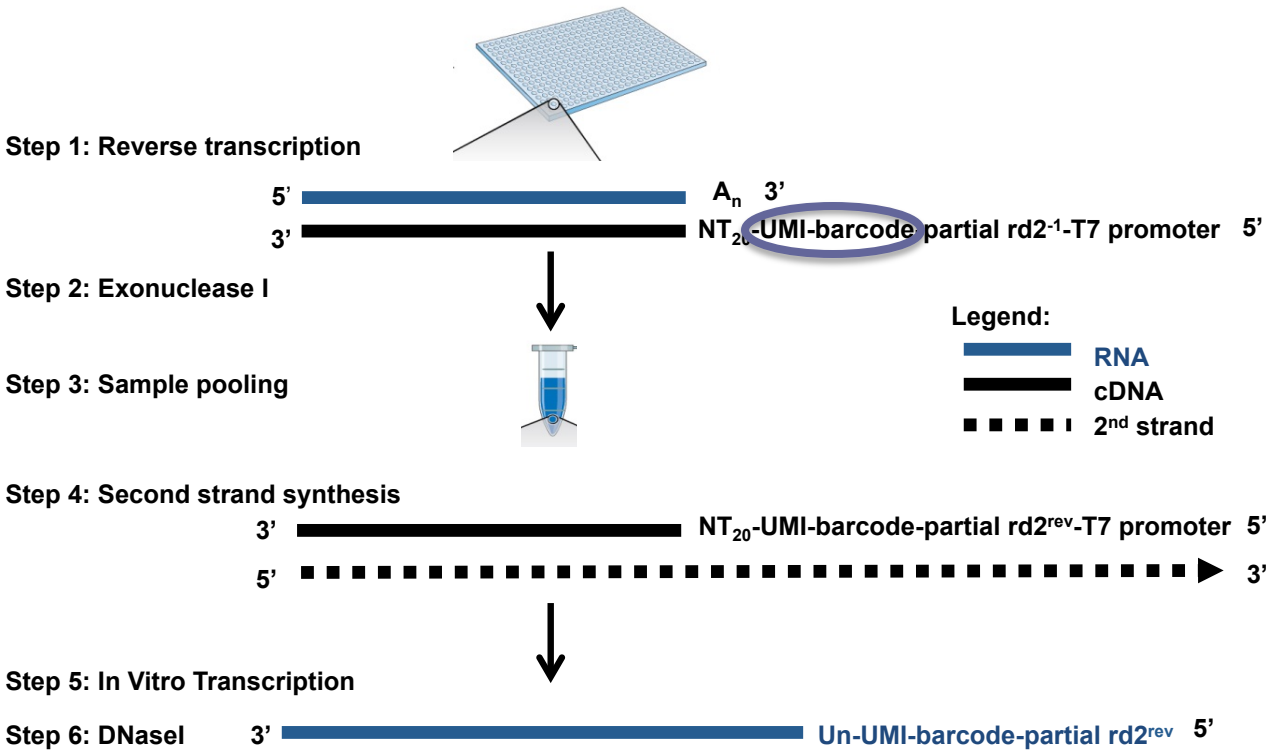
- Goal: Routinely profile hundreds of samples
- Why?
  - Human variability in health and disease
  - Perturbation studies
  - Clinical applications of expression profiling
  - Single cell sequencing
- Current costs
  - Afffy ~$300-$400/sample
  - Illumina bead arrays $150/sample
  - RNA-Seq (20 mill reads) ~$400-$500/sample ($350 in sequencing)
- RNA-Seq disadvantages
  - Complex analysis
  - Length bias

Reading molecules: end-sequencing and molecular barcodes



Maxim Artyomov

Molecule counting – Unique Molecular Identifiers (UMI)



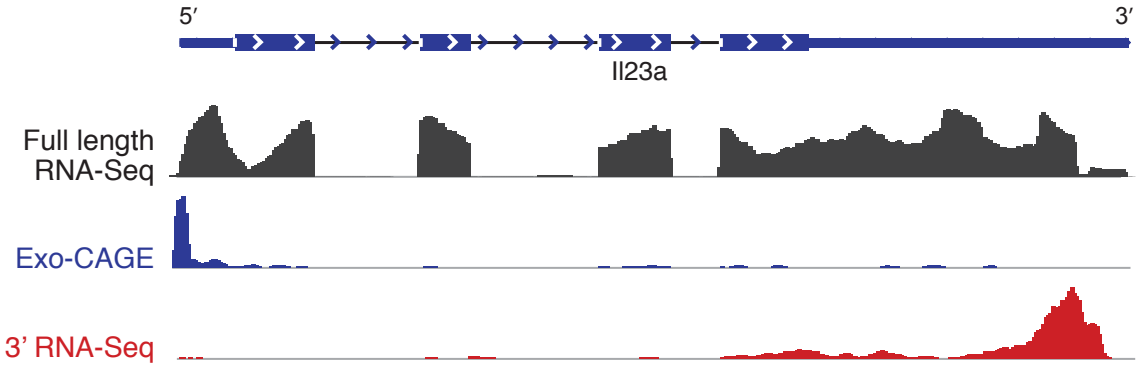
NT_{20}**NNNNNNNN-SSSSSS**-adapter

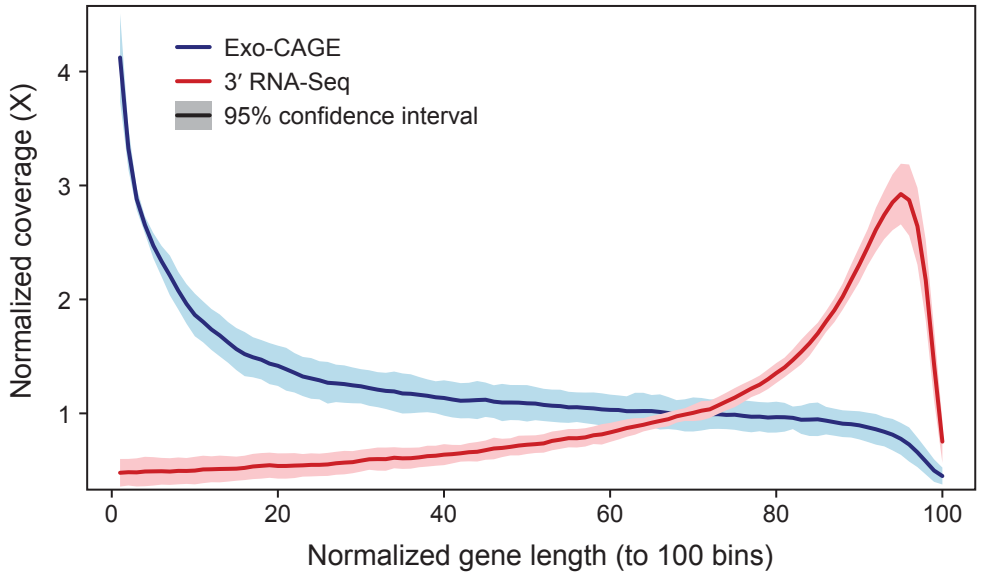
NNNNNNNN: UMI

SSSSSS: Sample Barcode

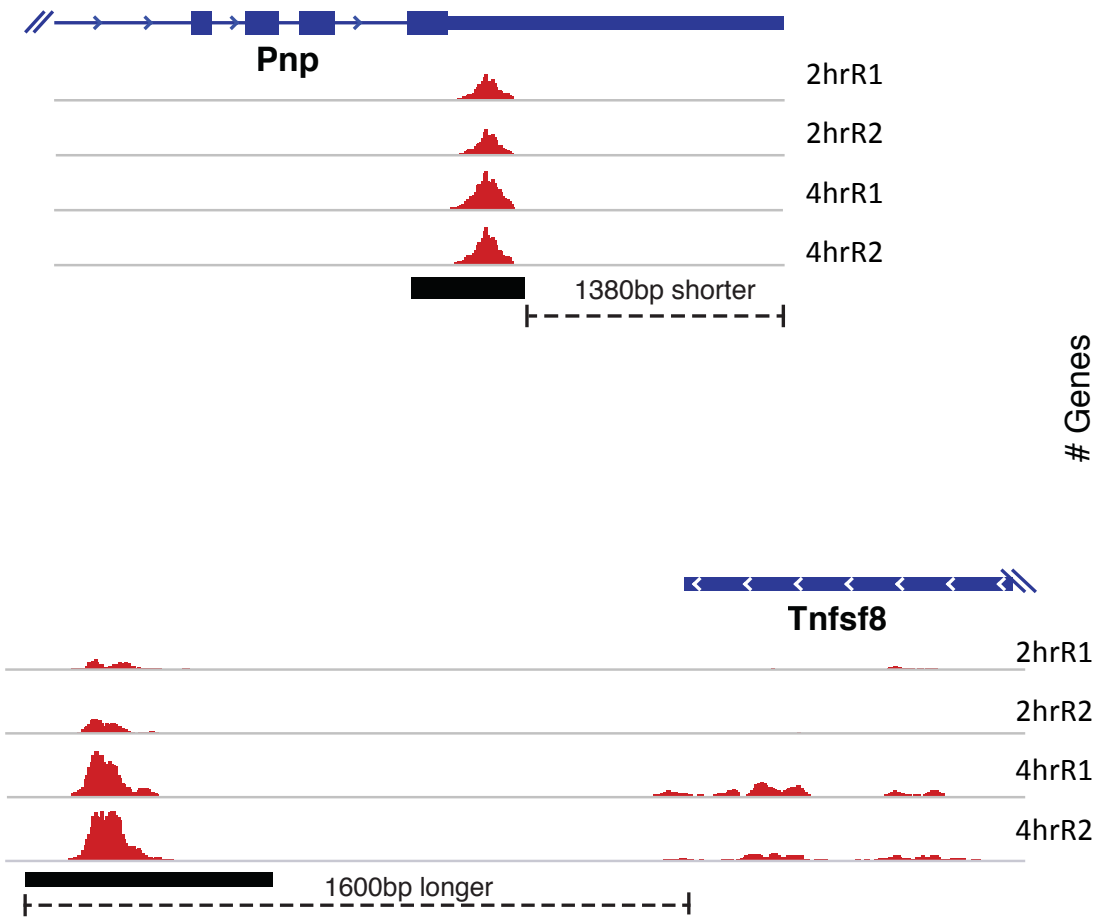
Jaitin et al. Science 2014

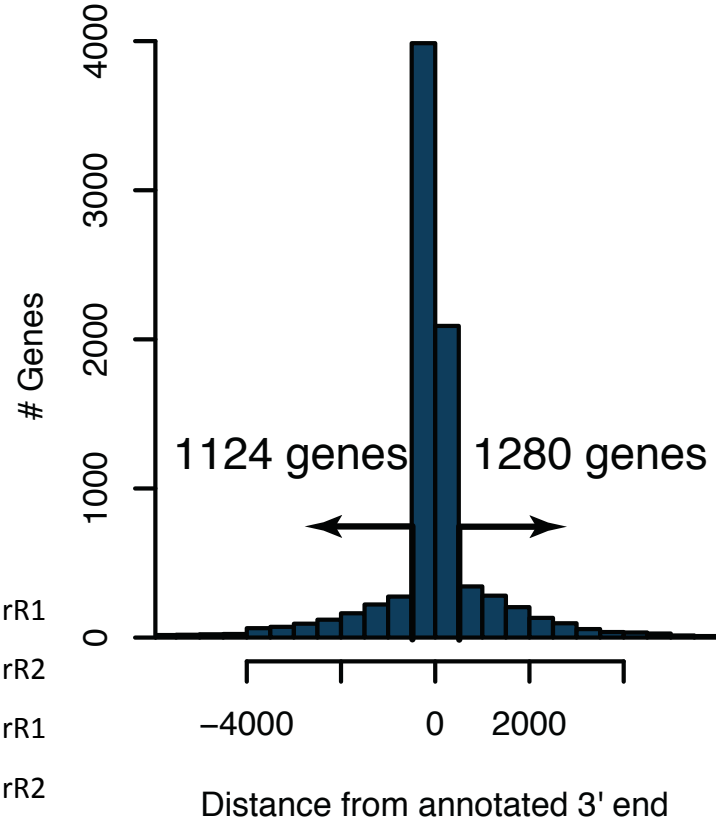
End-sequencing really sequences transcript ends



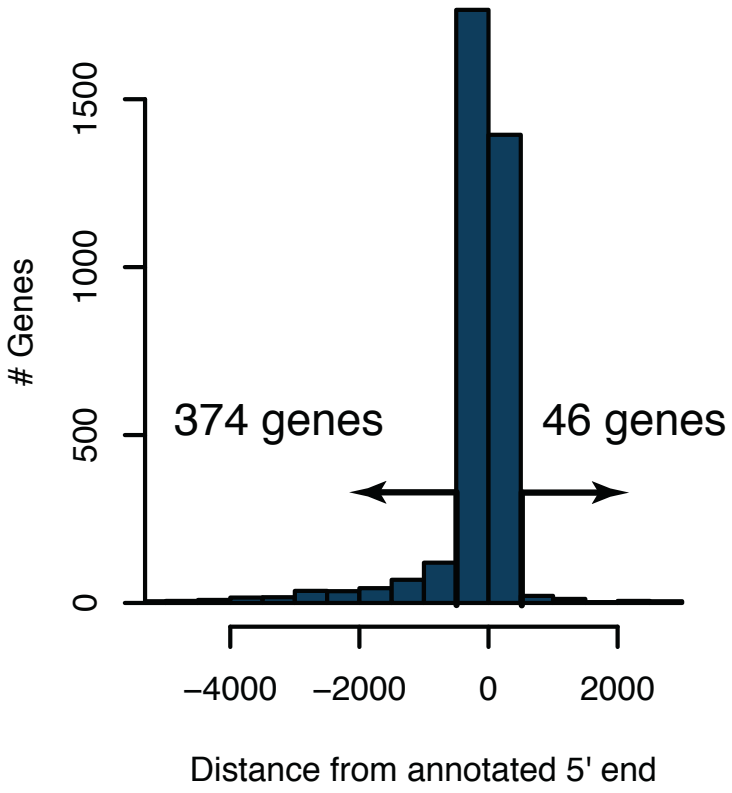


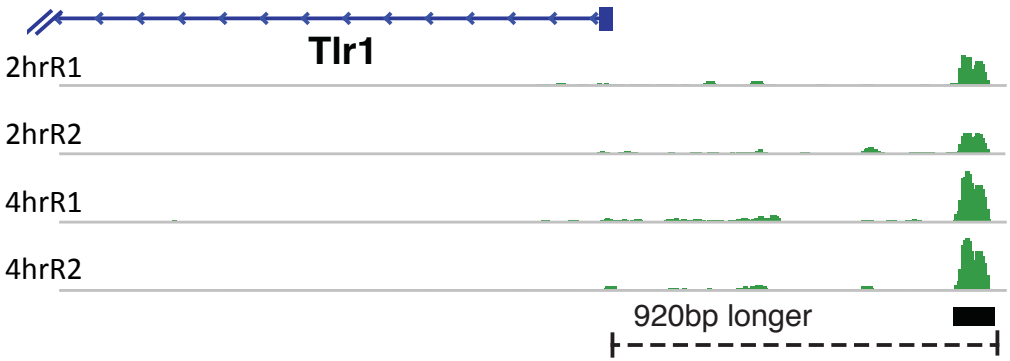
Challenge: annotated ends far from perfect

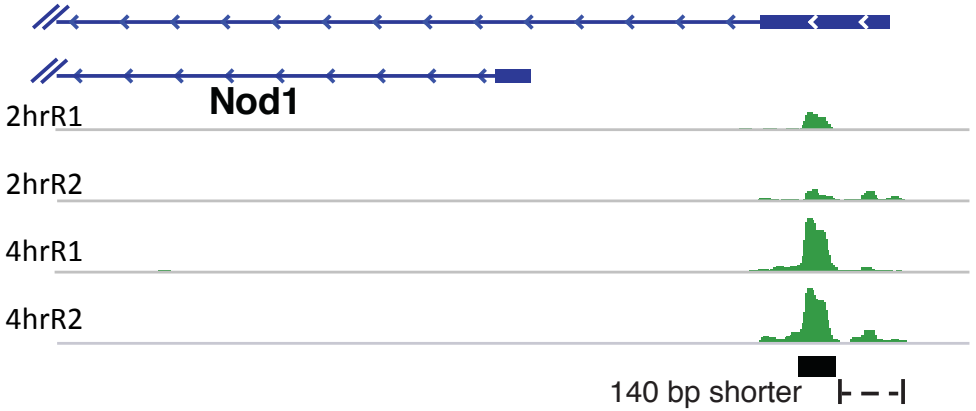




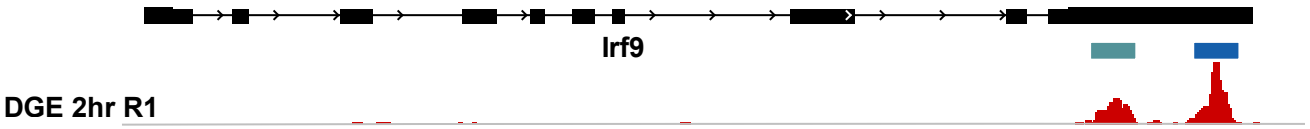
While annotated starts are much more conserved







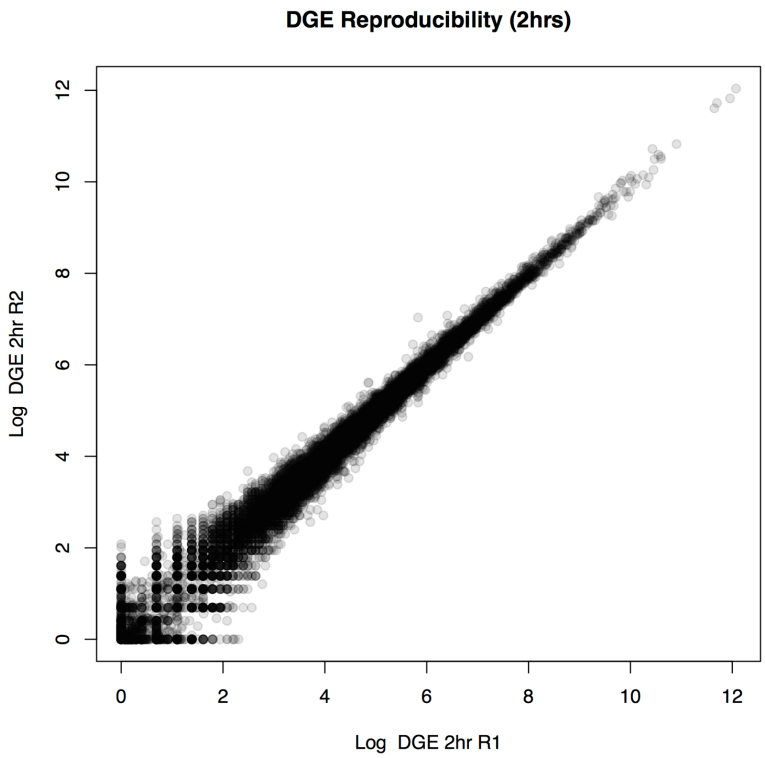
End sequencing give alt. UTR ussage

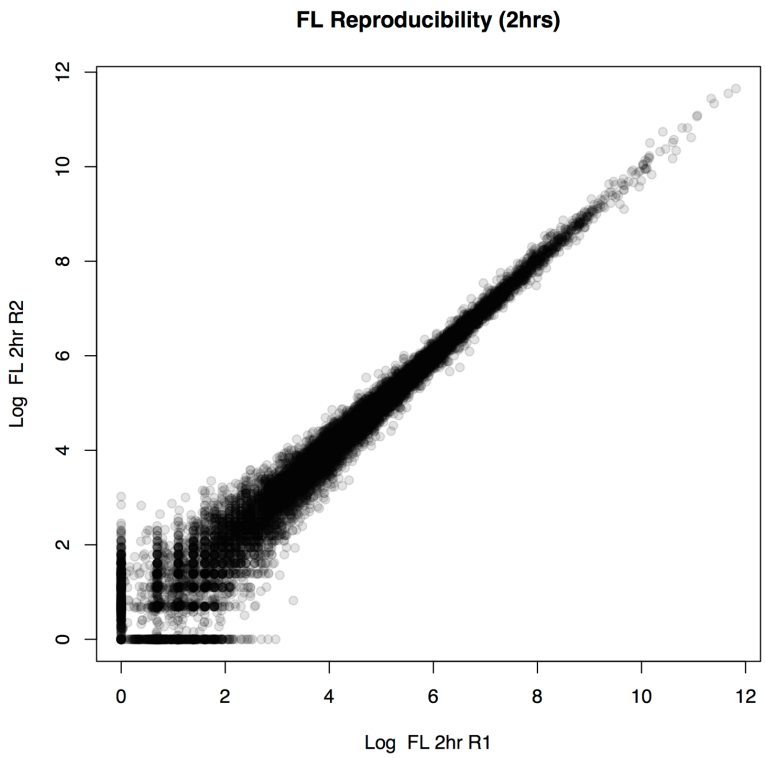


1. Slide a window and identify major 3’ end
2. Identify all other significant windows (using a local background)
3. Repeat for each sample
4. Take all significant windows across samples

5.1 Report gene level counts: Sum across all sig. windows
5.2 Report isoform level counts: Each sig. window

Reproducibility is as good as with full length





DGE has been the bases of single cell sequencing

End sequencing is ideal for single cell sequencing:

- Each mRNA is represented by one read
- Amenable to molecular indexing
- Works well with low quality RNA
- Uses a universal primer: oligo-dT

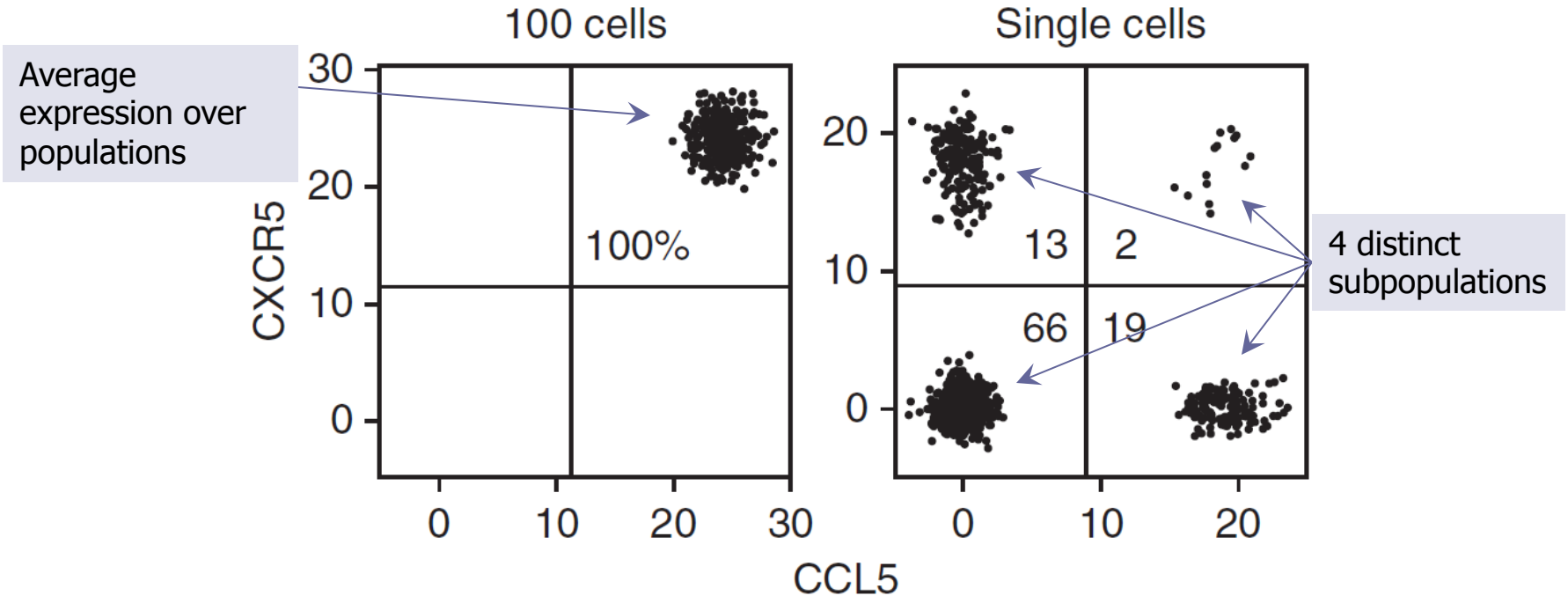
As a result most existing single cell protocols use end sequence:

- Cell-Seq (*Hamshimshony, et al Cell Reports 2012*)
- MARS-Seq (Jaitin, *et al. Science 2014)*
- SCRB-Seq (*Soumillion, et al, bioRxiv 2014*) Cell-Seq

- Which have been adapted to microfluidic approaches
  - In-Drop (*Klein et al. Cell 2015*)
  - Drop-Seq (*Macosko et al., Cell, 2015*)

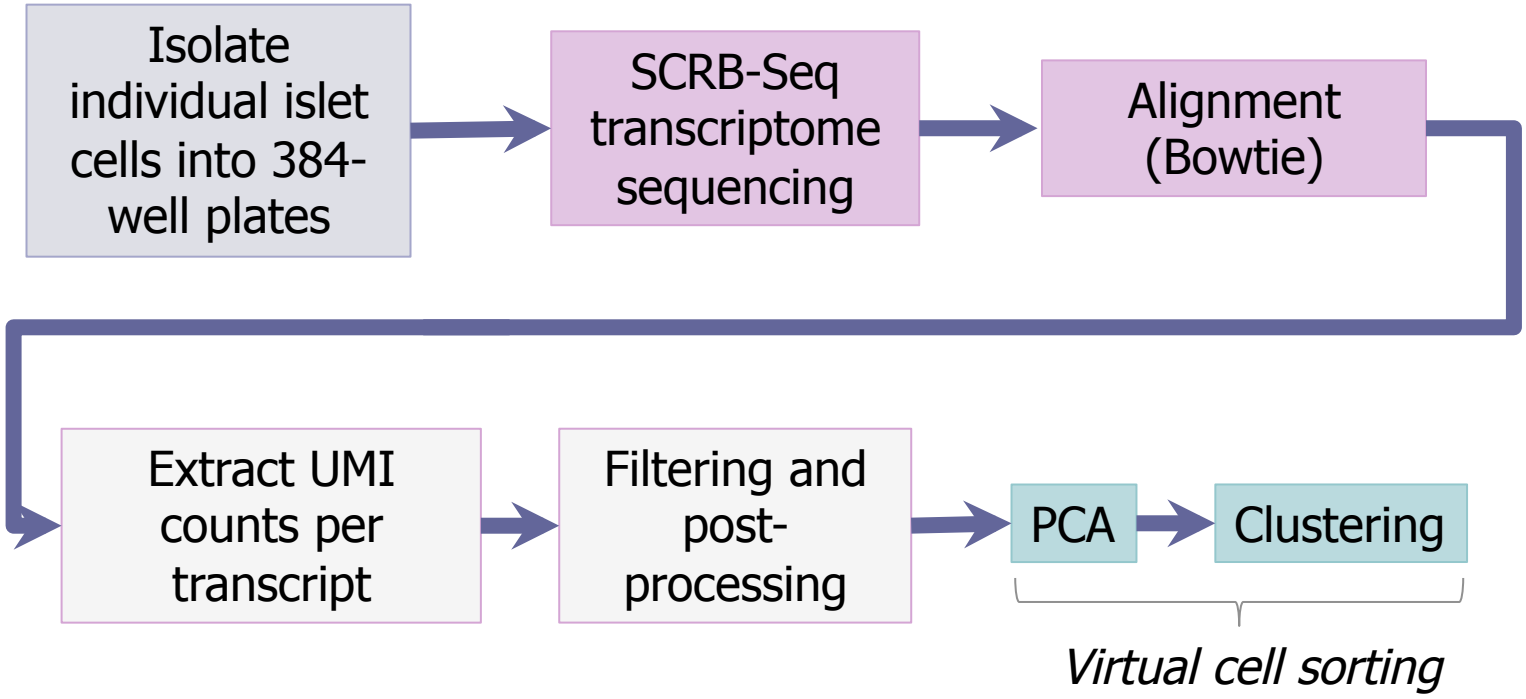
Why Single-cell analysis?

qPCR analysis of *CXCR5* vs *CCL5* expression in 'bulk' 100-cell T cell populations and single T cells:



From: ***Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells***, Newell E, Davis M; Nature Biotechnology, Feb. 2014

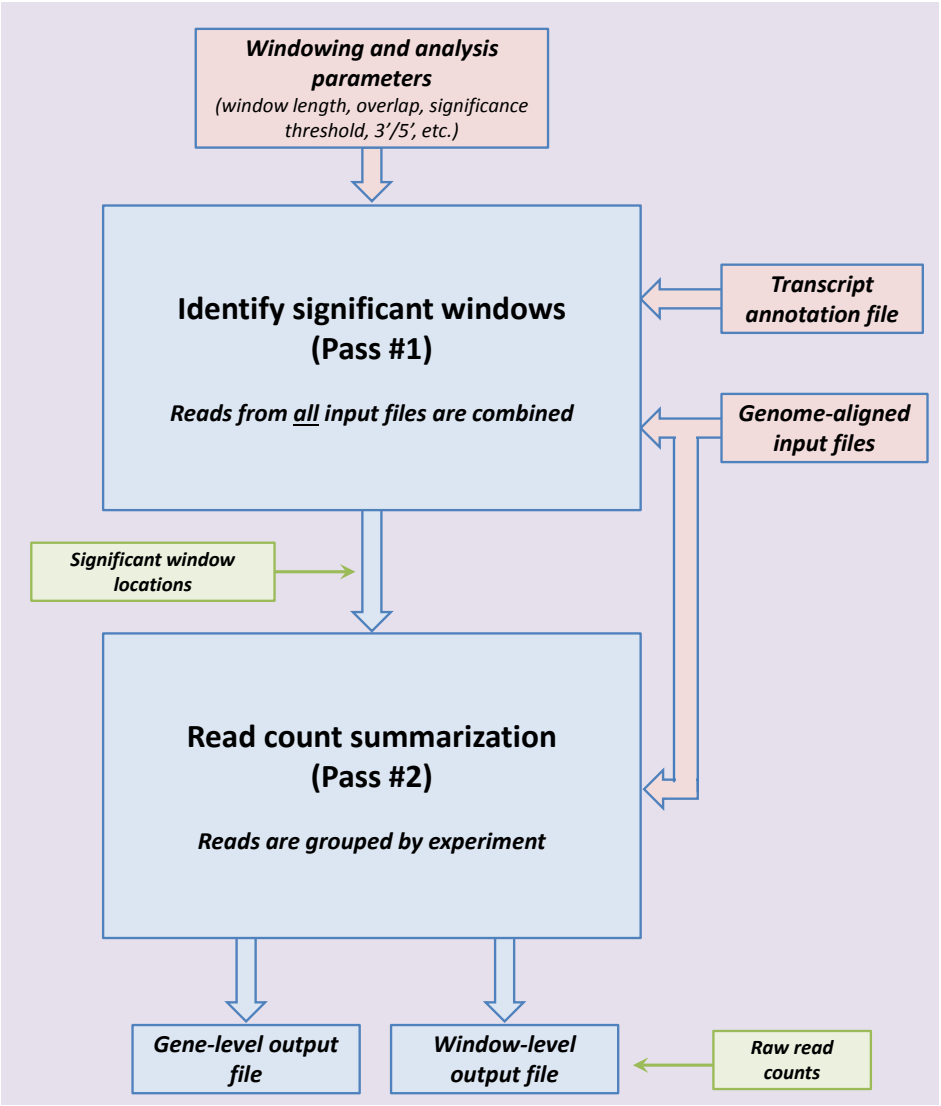
Typical study

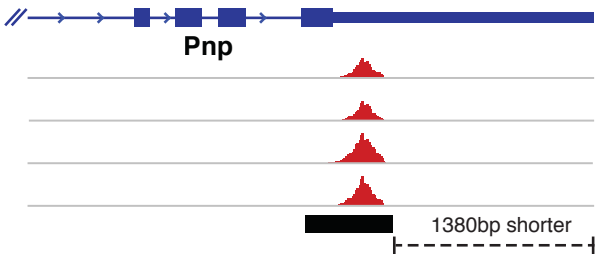


Post-processing

- Major difference from bulk RNA-Seq:
  - Transcript expression level is proportional to the number of **unique UMIs** aligning to the transcript, **not** the number of reads.
  - Expression estimates the # of molecules not the proportion anymore.
- Number of reads per UMI is an indication of quality:
  - If average reads/UMI is low, indicates poor-quality cell material (e.g., dead cells).
  - Low reads per individual UMI for a transcript indicates questionable alignment.

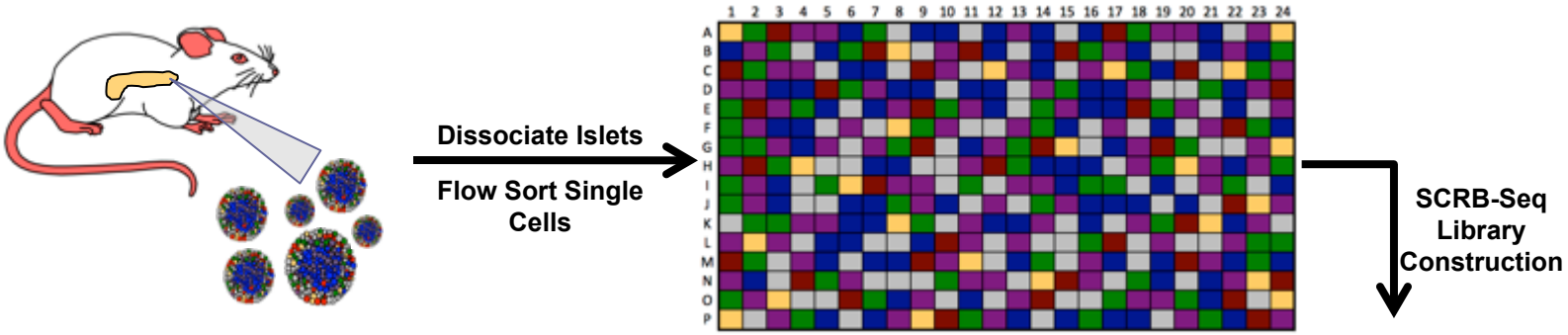
End-Sequencing Analysis Toolkit (ESAT)





http://garberlab.umassmed.edu/software/esat/

Islet single cell sequencing (SCRB-Seq)



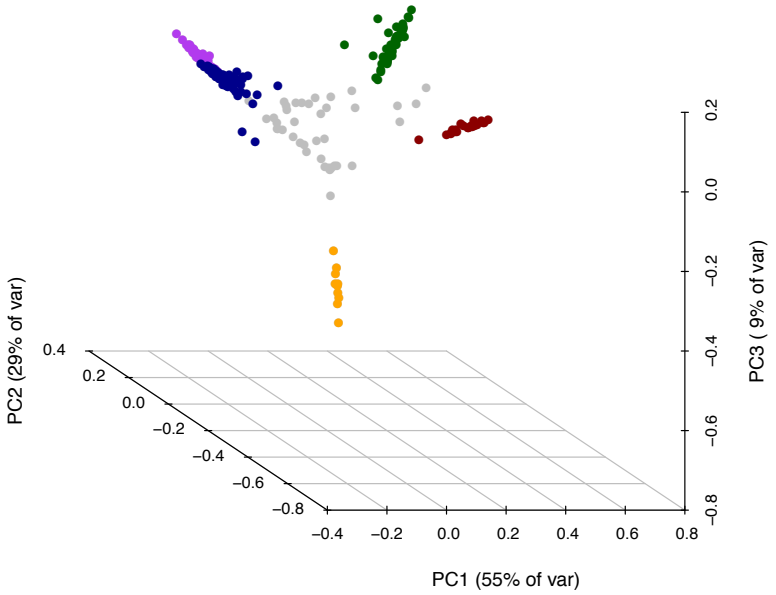
	no extension	5kb extension	% increase
Total unique genes with >0 UMI for any cell:	8574	9648	12.5%

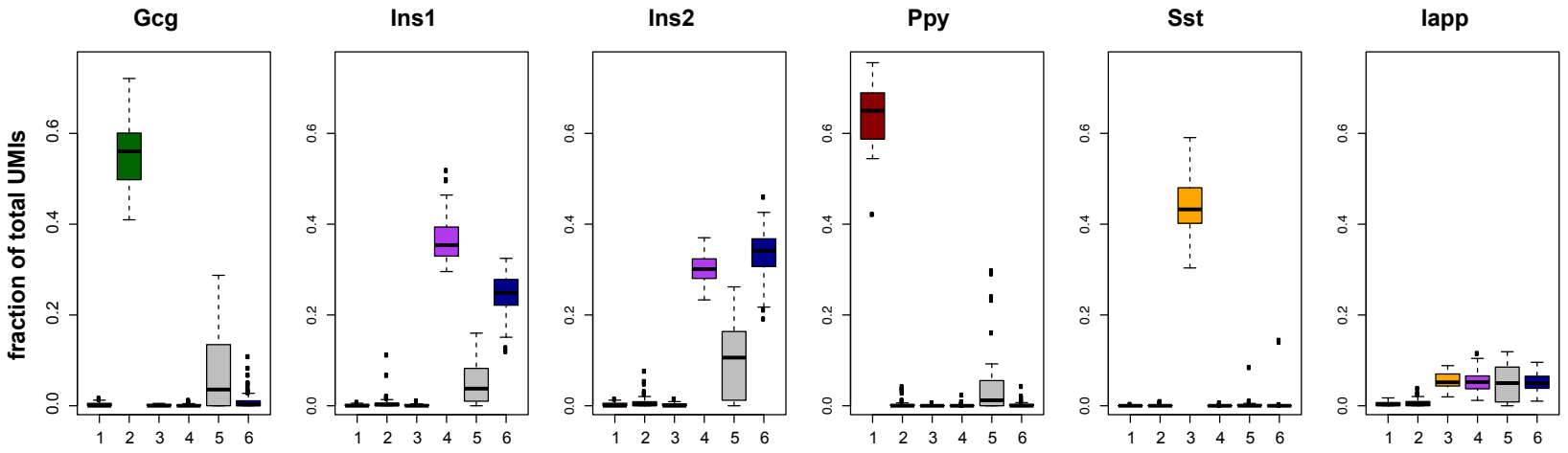
After filtering:

cells with >200 total UMIs:	263	283	7.6%
genes with >50 total UMIs:	296	367	24.0%

Alan Derr, Magali Soumillion, Tarjei Mikkelsen

Which allow us to recover the known islet composition



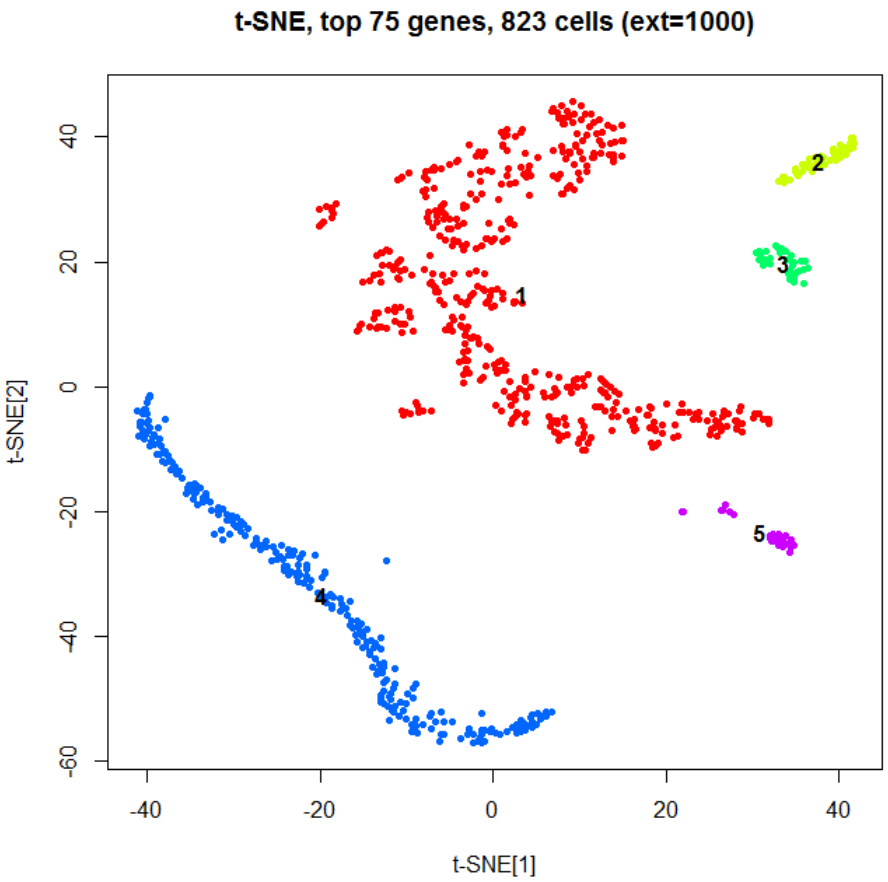


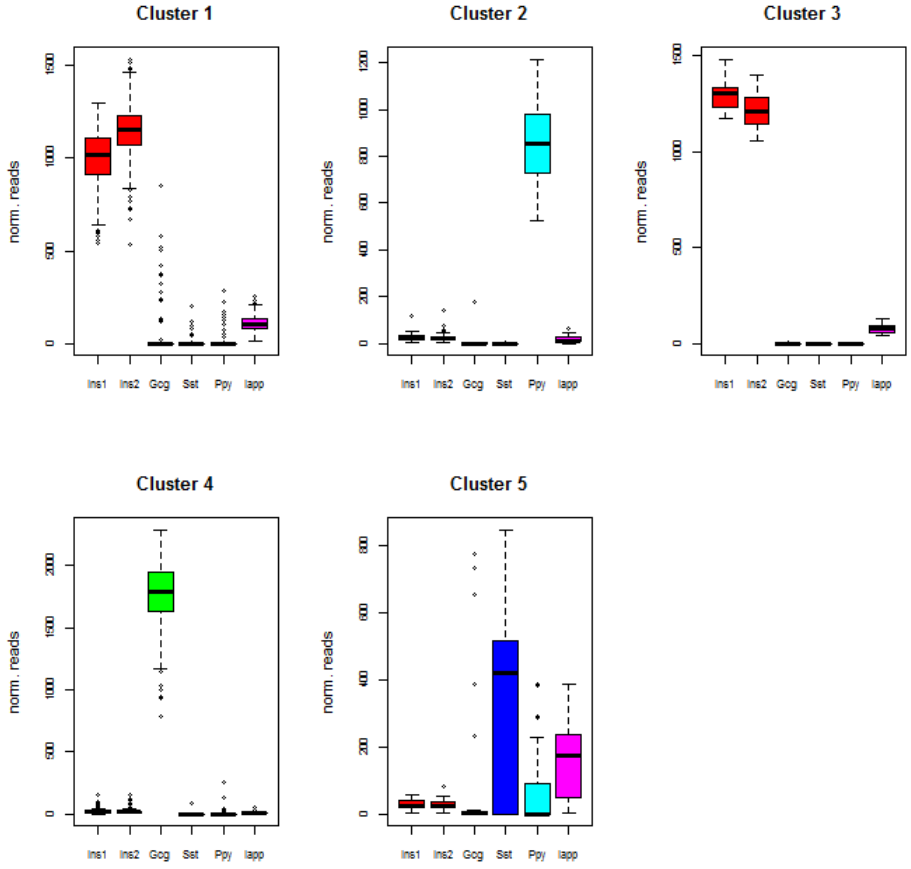
Islet single cell sequencing (In-Drop)

- Sample of healthy rat pancreatic islet cells
- Single-cell RNA-Seq performed using inDrop methodology with ~1000 cells per sequencing library
- Sequenced with Illumina MiSeq: 18.7M reads
- Reads with valid BC and UMI (after BC correction): 10.3M
- Mapped reads (Tophat genome alignment): 7.2M
- After removal of low-count BCs (<2K reads/BC): 6.1M
- Transcript-mapped reads (ESAT): 2.6M
- Total cells analyzed: 955
- PCA performed with remaining **980** genes with **823** cells
- Approximate PCR duplication rate: 2.3X

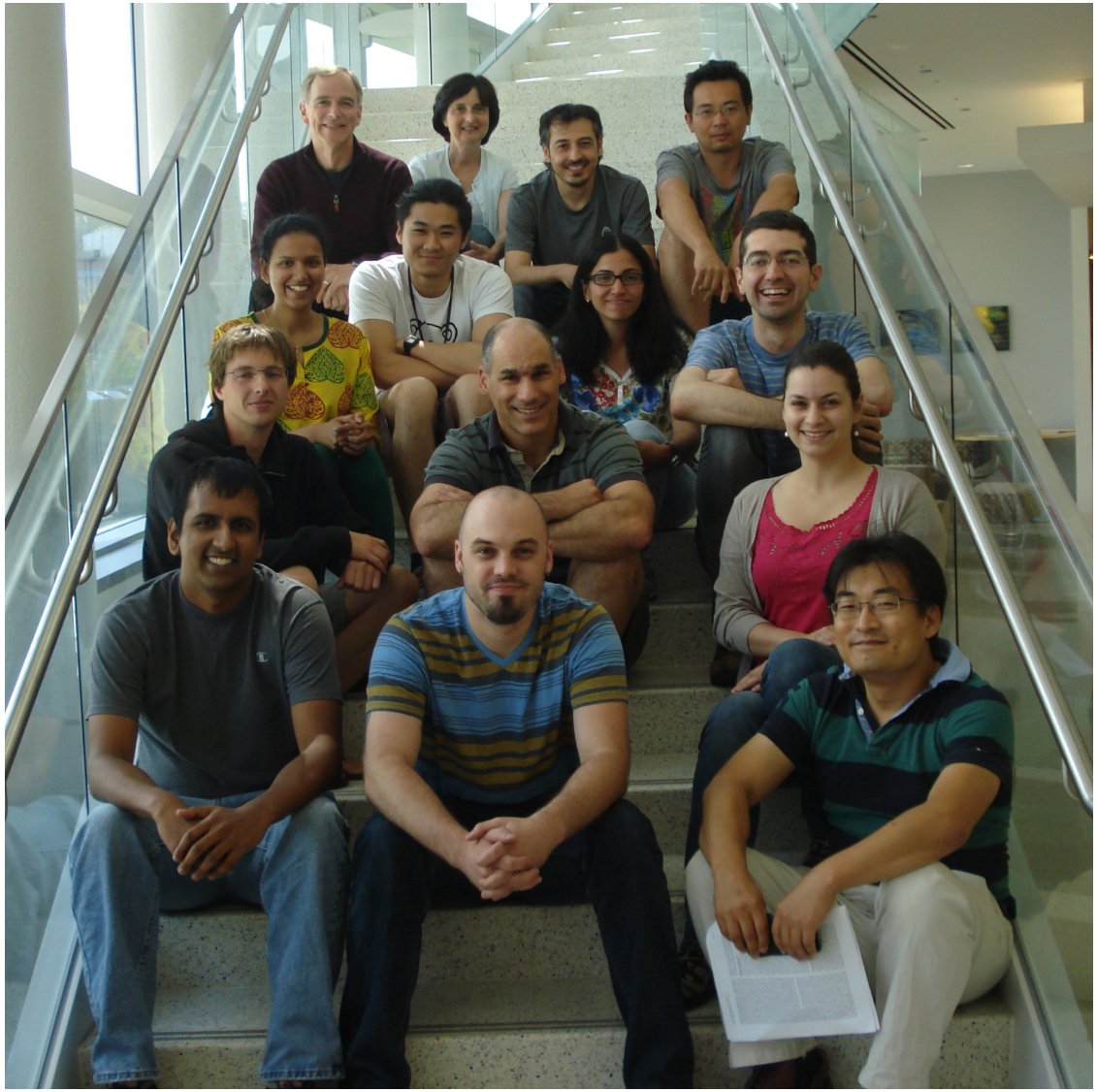
Alan Derr, Alon Klein, Sambra Redrik

t-SNE mapping of cells





Expression distributions of marker genes for each cluster



Diabetes Center
Sambra Redrik
David Blodget
Chaoxing Yang
Rita Bortel
Dale Greiner
David Harlan

Broad Technology Labs
Tarjei Mikelsen
Magali Soumillon

Harvard System Biology
Alon Klein

Garber Lab

DARPA

genome.gov
National Human Genome Research Institute
National Institutes of Health