UMASS University of Massachusetts Medical School

Short read sequence analysis

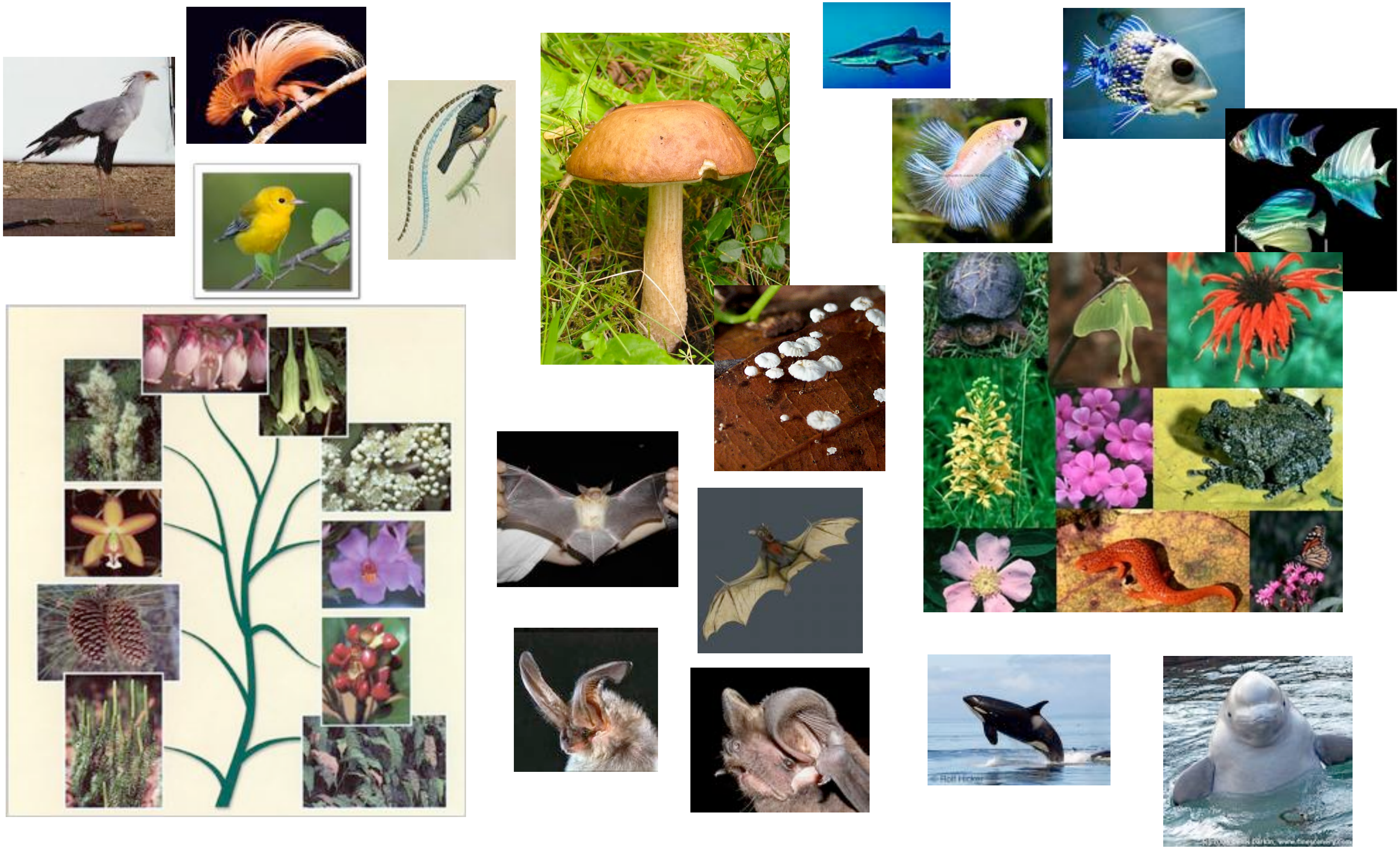
Manuel Garber

Krumlov 2015

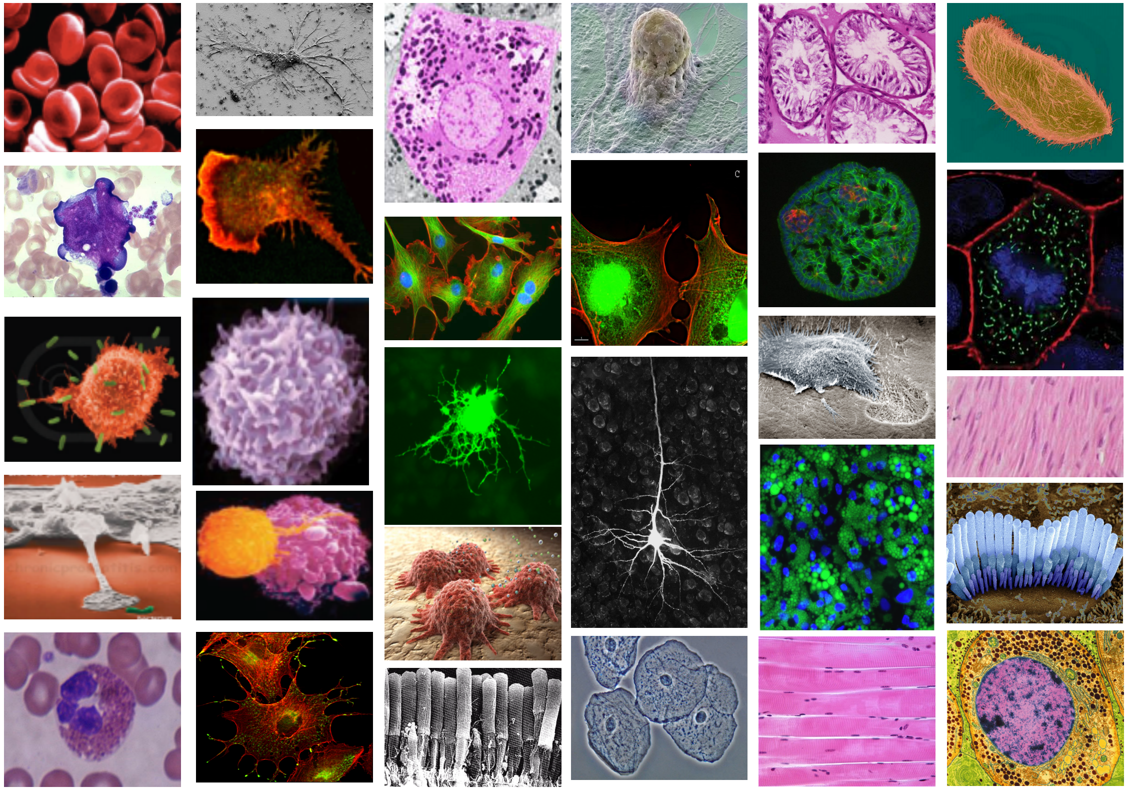
Overview of the session

- Explaining diversity: Transcriptional regulation
  - A short story from our recent work
- RNA Sequencing
  - The different BLA-Seq libraries. A common theme
  - Read mapping (alignment): Placing short reads in the genome
  - Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.
    - How much depth?
  - ~~Reconstruction: Finding the regions that originated the reads~~
- RNA-Seq Vignette: non-coding RNA evolution

Why do organisms look the way that they do?

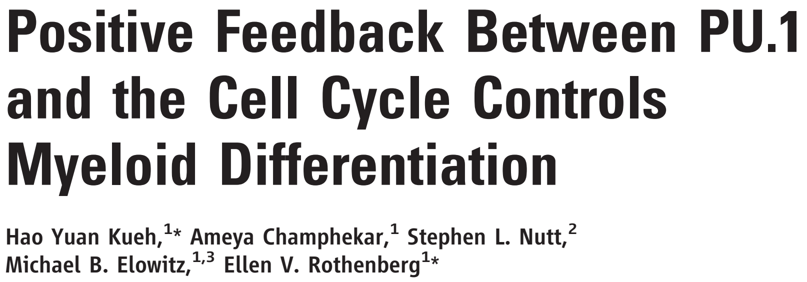


Why do different cell types do what they do!



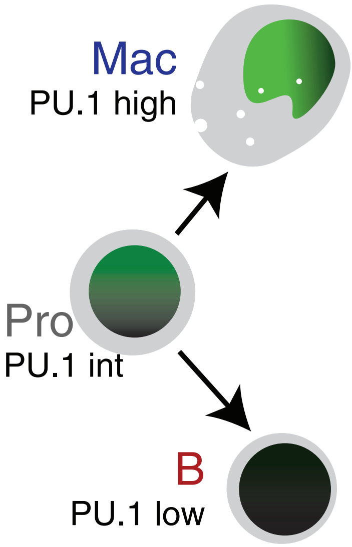
However, all this diversity arises from the same genome sequence!
Proteins are very conserved across vertebrates, what is the driving force of variability?

Cell identity is determined by gene regulation

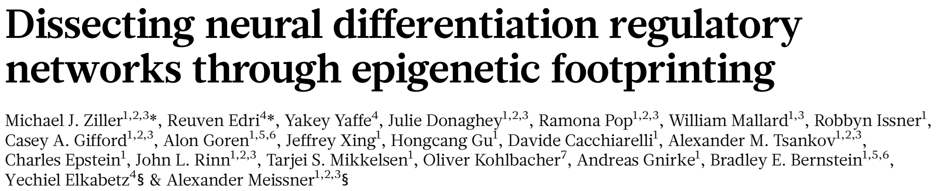


Positive Feedback Between PU.1 and the Cell Cycle Controls Myeloid Differentiation

Hao Yuan Kueh,[1]* Ameya Champhekar,[1] Stephen L. Nutt,[2]
Michael B. Elowitz,[1,3] Ellen V. Rothenberg[1]*

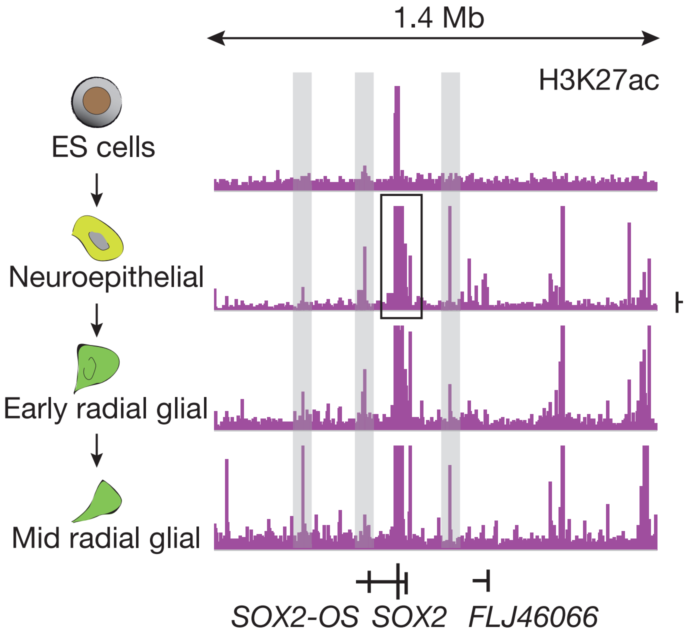


And their epigenetic state



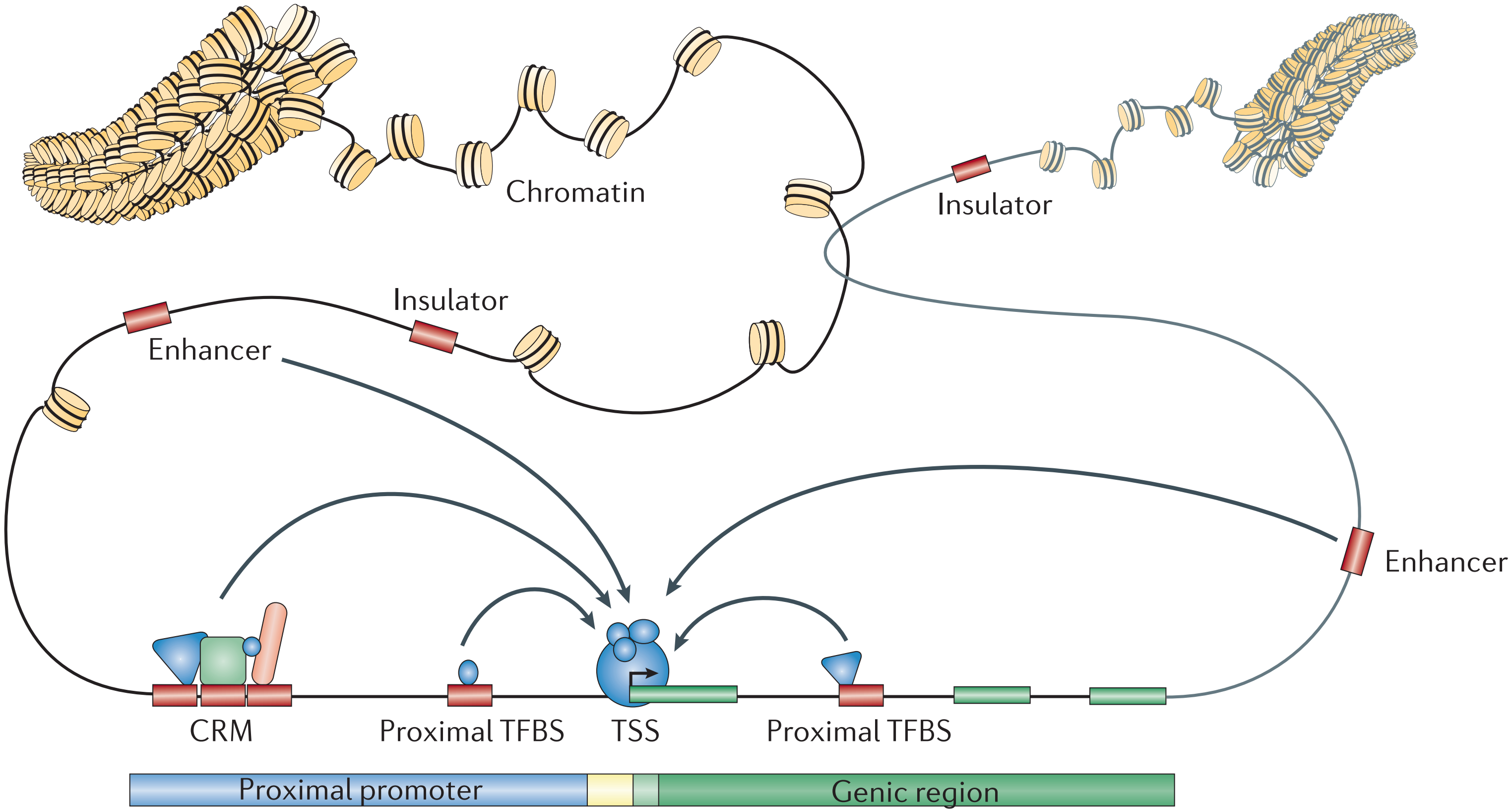
Dissecting neural differentiation regulatory networks through epigenetic footprinting

Michael J. Ziller[1,2,3]*, Reuven Edri[4]*, Yakey Yaffe[4], Julie Donaghey[1,2,3], Ramona Pop[1,2,3], William Mallard[1,3], Robbyn Issner[1], Casey A. Gifford[1,2,3], Alon Goren[1,5,6], Jeffrey Xing[1], Hongcang Gu[1], Davide Cacchiarelli[1], Alexander M. Tsankov[1,2,3], Charles Epstein[1], John L. Rinn[1,2,3], Tarjei S. Mikkelsen[1], Oliver Kohlbacher[7], Andreas Gnirke[1], Bradley E. Bernstein[1,5,6], Yechiel Elkabetz[4]§ & Alexander Meissner[1,2,3]§



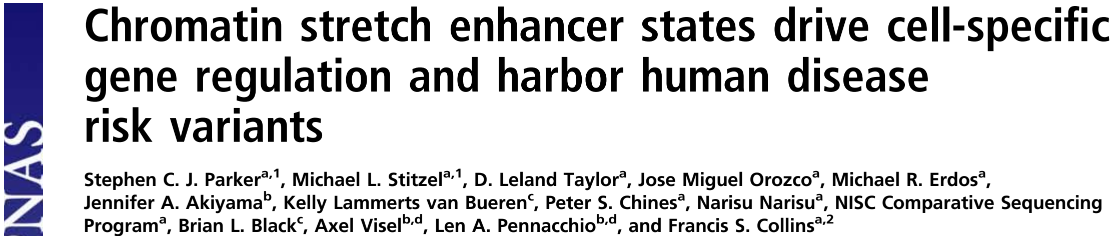
Transcription factors regulate gene programs. Epigenome informs (determines?) potential for expression

Multicellular development requires complex regulation



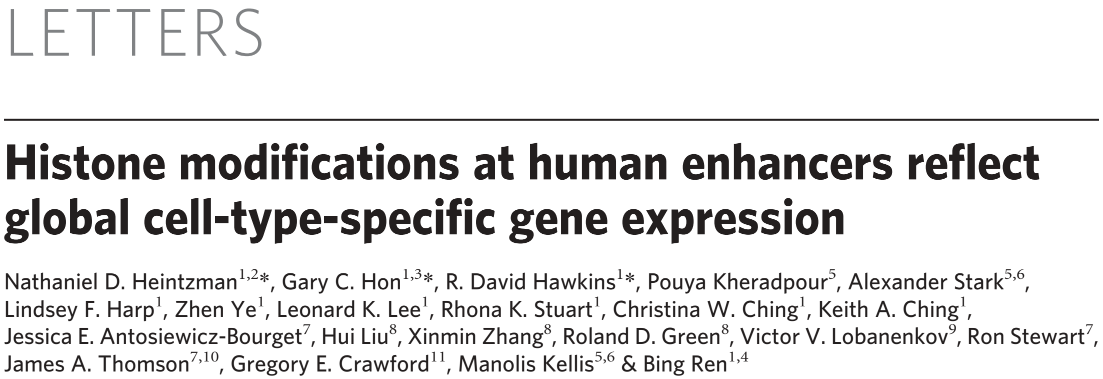
Lenhard et al. Nat Review Genetics 2012

Indeed Enhancers are both species and cell type specific



Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants

Stephen C. J. Parker[a,1], Michael L. Stitzel[a,1], D. Leland Taylor[a], Jose Miguel Orozco[a], Michael R. Erdos[a], Jennifer A. Akiyama[b], Kelly Lammerts van Bueren[c], Peter S. Chines[a], Narisu Narisu[a], NISC Comparative Sequencing Program[a], Brian L. Black[c], Axel Visel[b,d], Len A. Pennacchio[b,d], and Francis S. Collins[a,2]



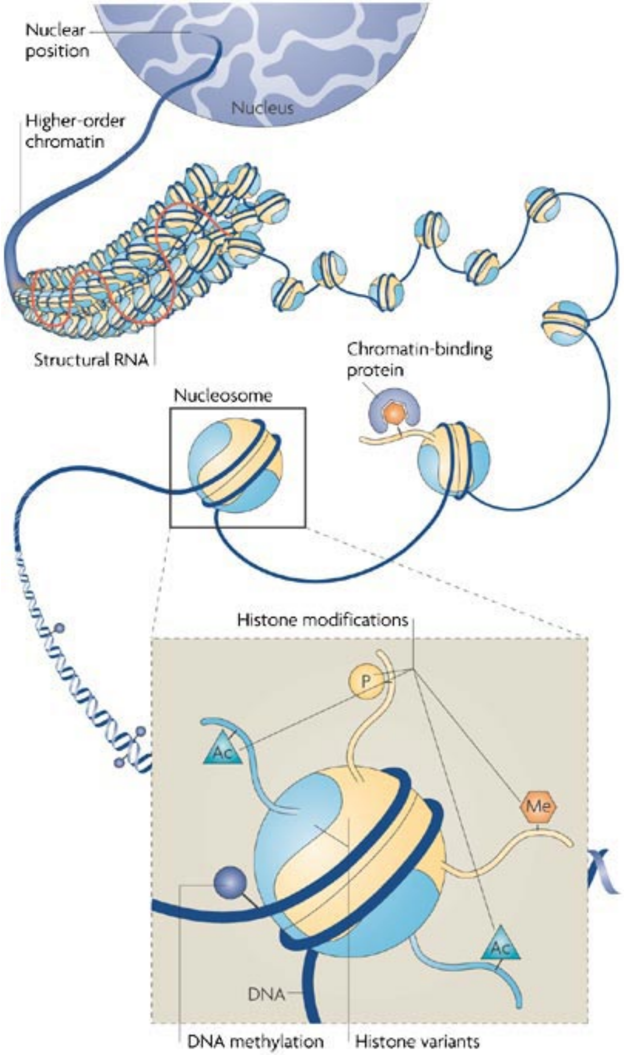
LETTERS

Histone modifications at human enhancers reflect global cell-type-specific gene expression

Nathaniel D. Heintzman[1,2]*, Gary C. Hon[1,3]*, R. David Hawkins[1]*, Pouya Kheradpour[5], Alexander Stark[5,6], Lindsey F. Harp[1], Zhen Ye[1], Leonard K. Lee[1], Rhona K. Stuart[1], Christina W. Ching[1], Keith A. Ching[1], Jessica E. Antosiewicz-Bourget[7], Hui Liu[8], Xinmin Zhang[8], Roland D. Green[8], Victor V. Lobanenkov[9], Ron Stewart[7], James A. Thomson[7,10], Gregory E. Crawford[11], Manolis Kellis[5,6] & Bing Ren[1,4]

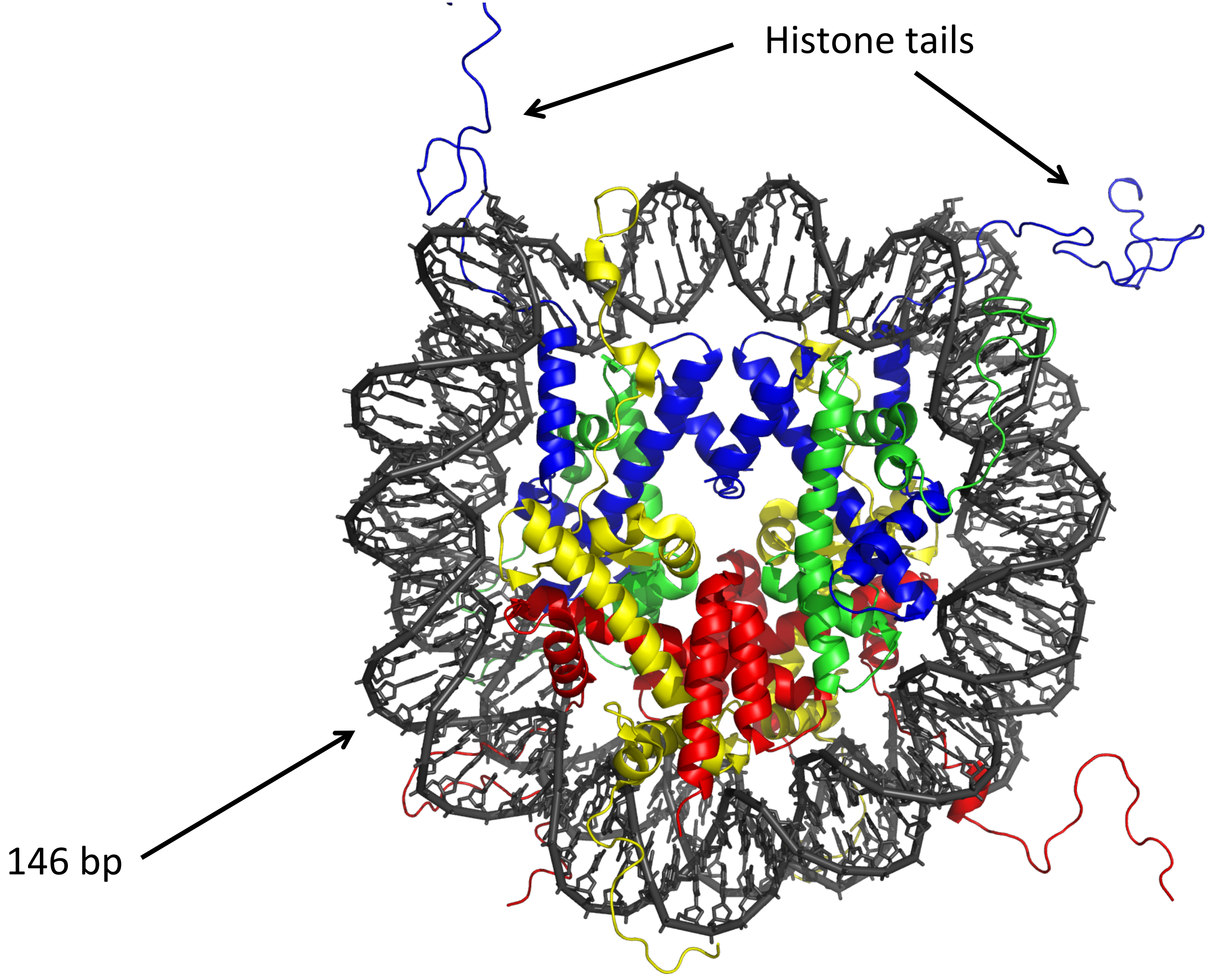
Enhancer elements are poorly conserved, are cell type specific, How do we find them?

DNA is not naked



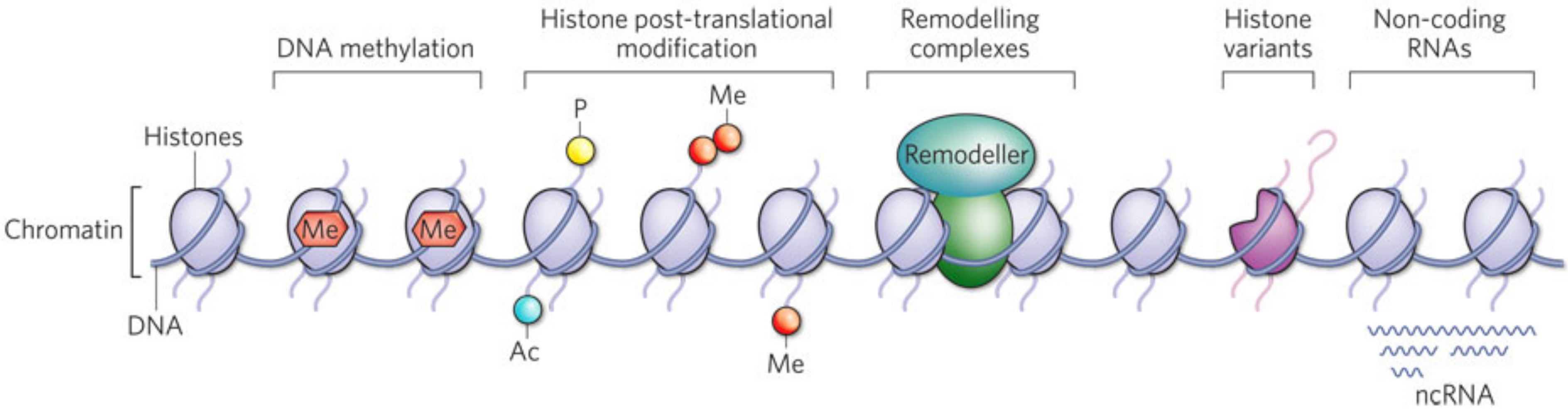
Nature Reviews | Molecular Cell Biology

Nucleosomes interact with nuclear factors through tails



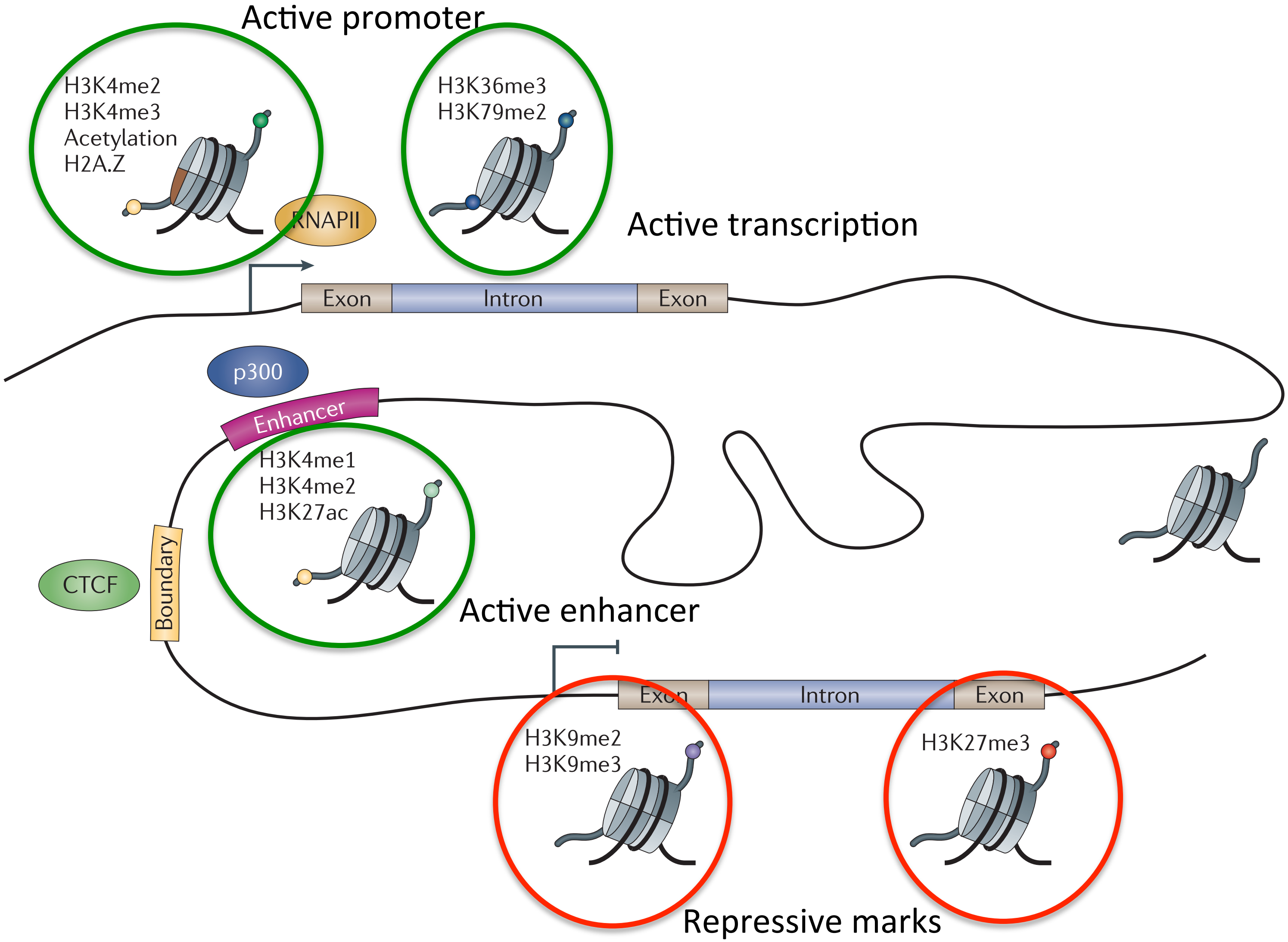
Wikipedia

Cell identity is determined by its epigenetic state



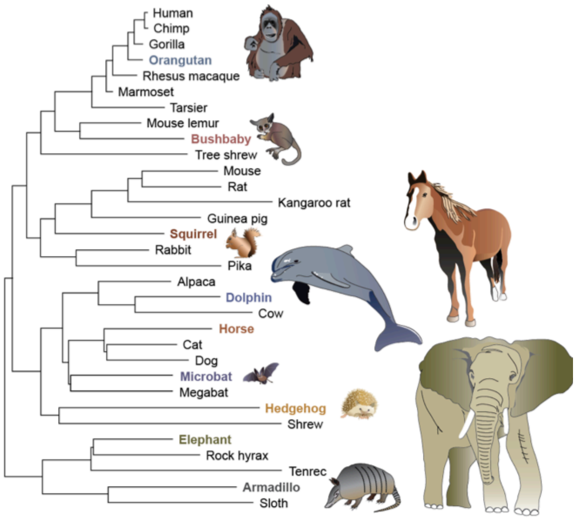
Catherine Dulac, Nature 2010

Which controls the genome functional elements

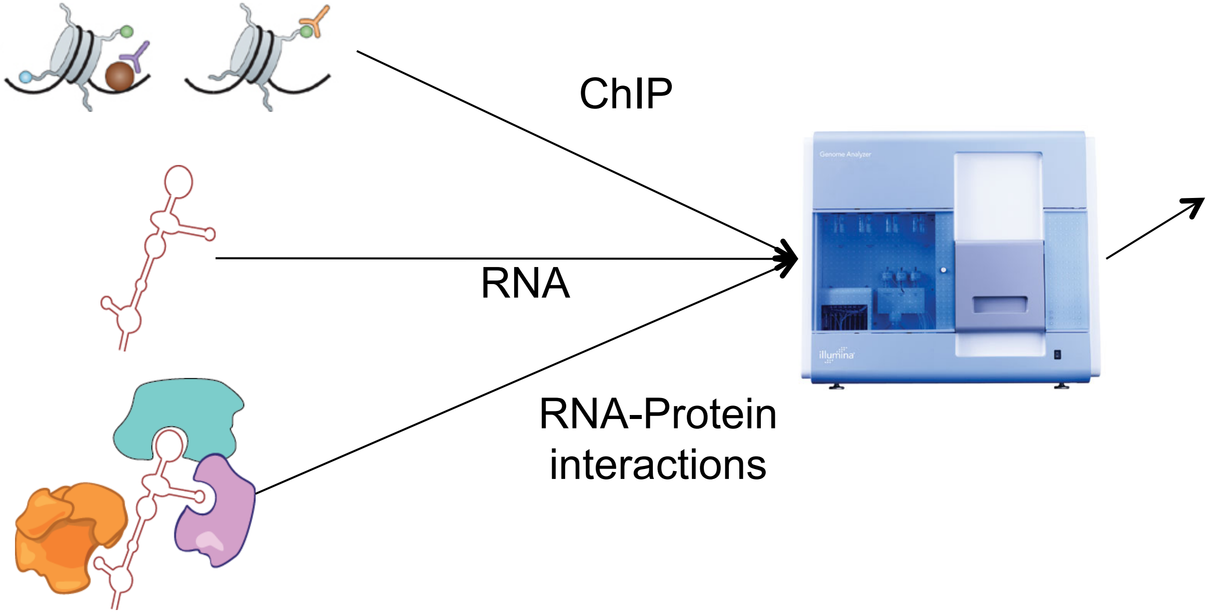


Zhou, Goren Berenstein, Nature Rev. Genetics 2011

Dissecting a gene regulatory network



Comparative genomics – measure constraint



ChIP

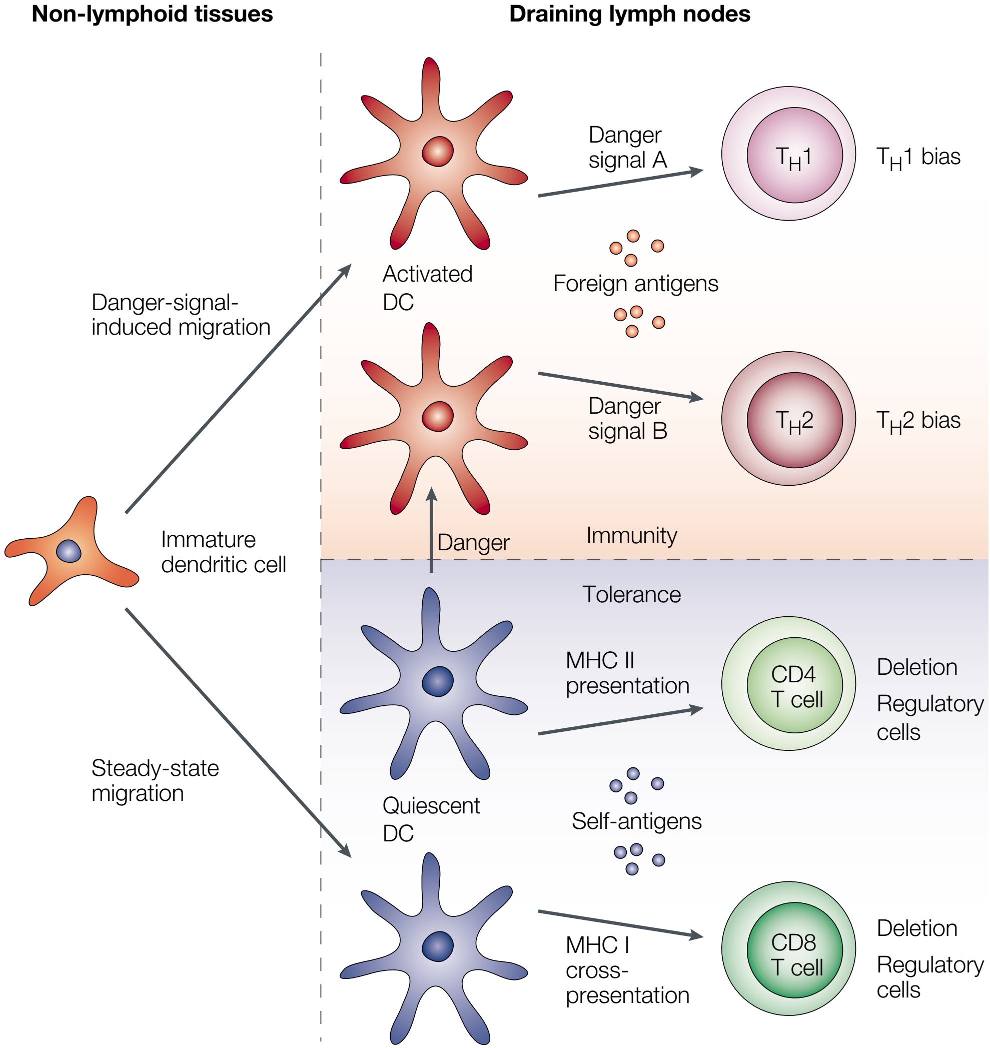
RNA

RNA-Protein interactions

- new methods
- models of transcriptional regulation
- models of epigenetic interactions
- perturbations

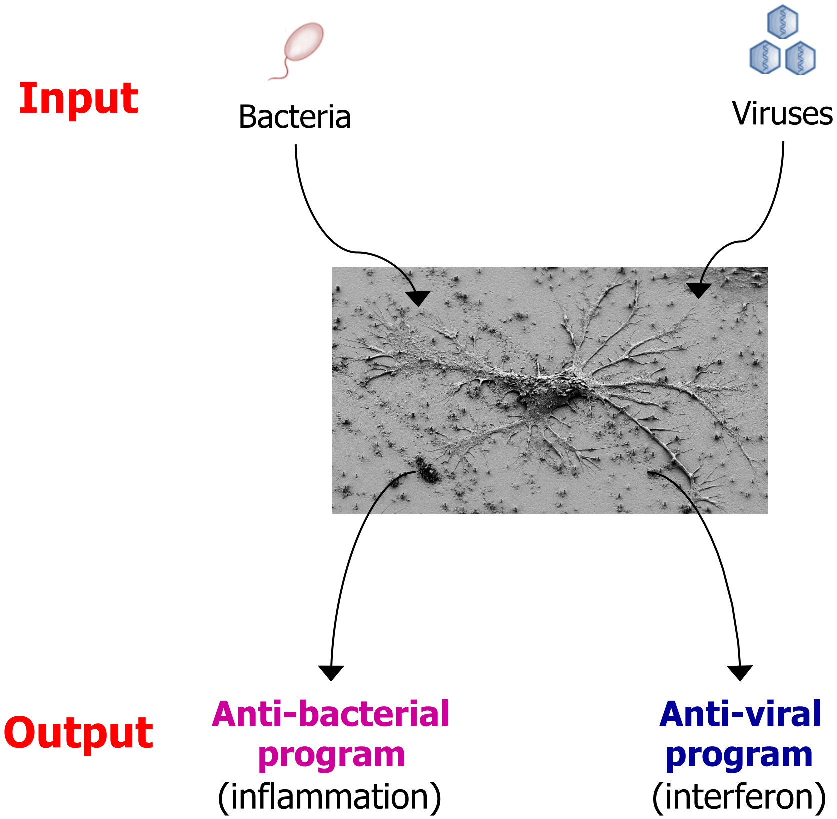
We want to ultimately understand the cell circuits of the cell

Understanding innate immunity



Shortman and Liu, Nat. Reviews Immunology 2002

Gene expression programs in response to LPS



Input

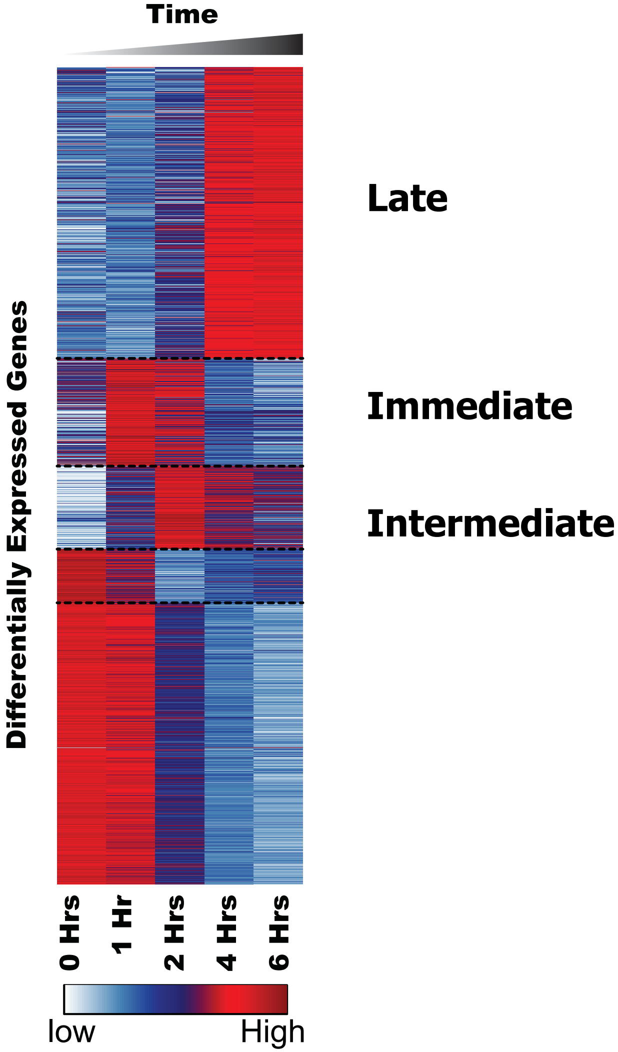
Bacteria

Viruses

Output

Anti-bacterial program (inflammation)

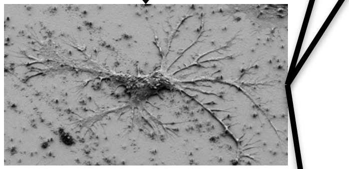
Anti-viral program (interferon)



LPS (TLR4 receptor) stimulation as it elicits the most broad gene expression response.

Chip-Seq + RNA-Seq to map and relate components

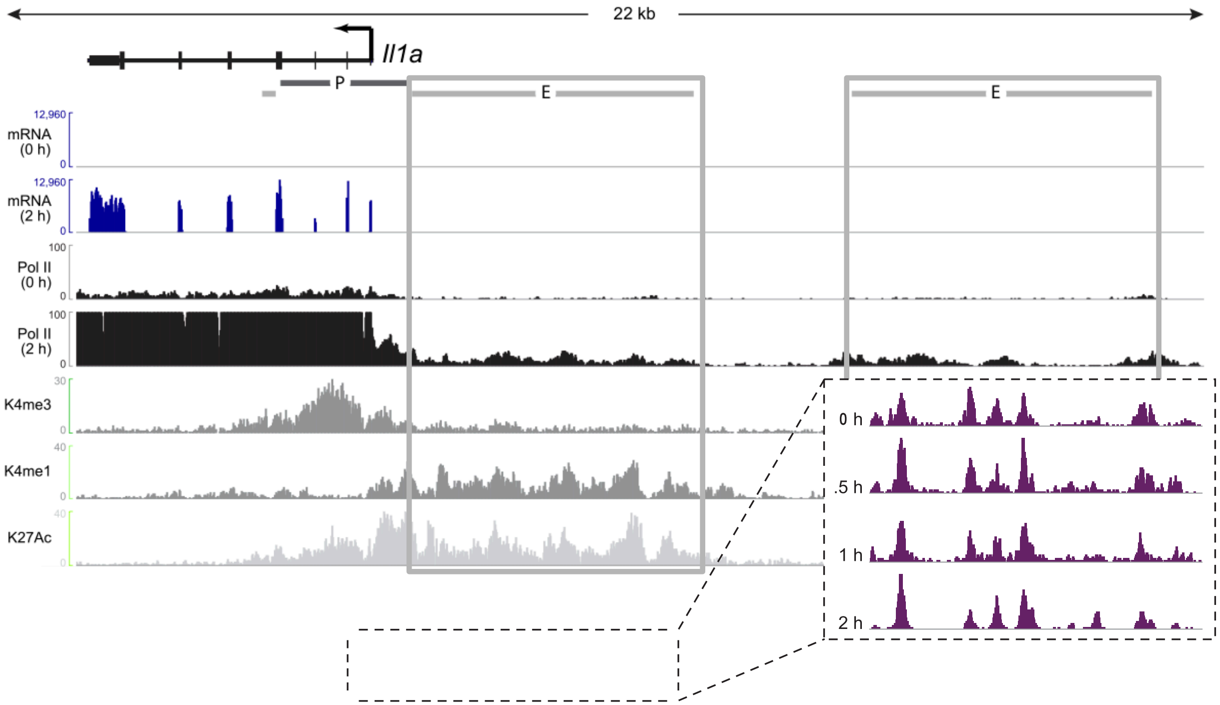
LPS



RNA-Seq
Expression

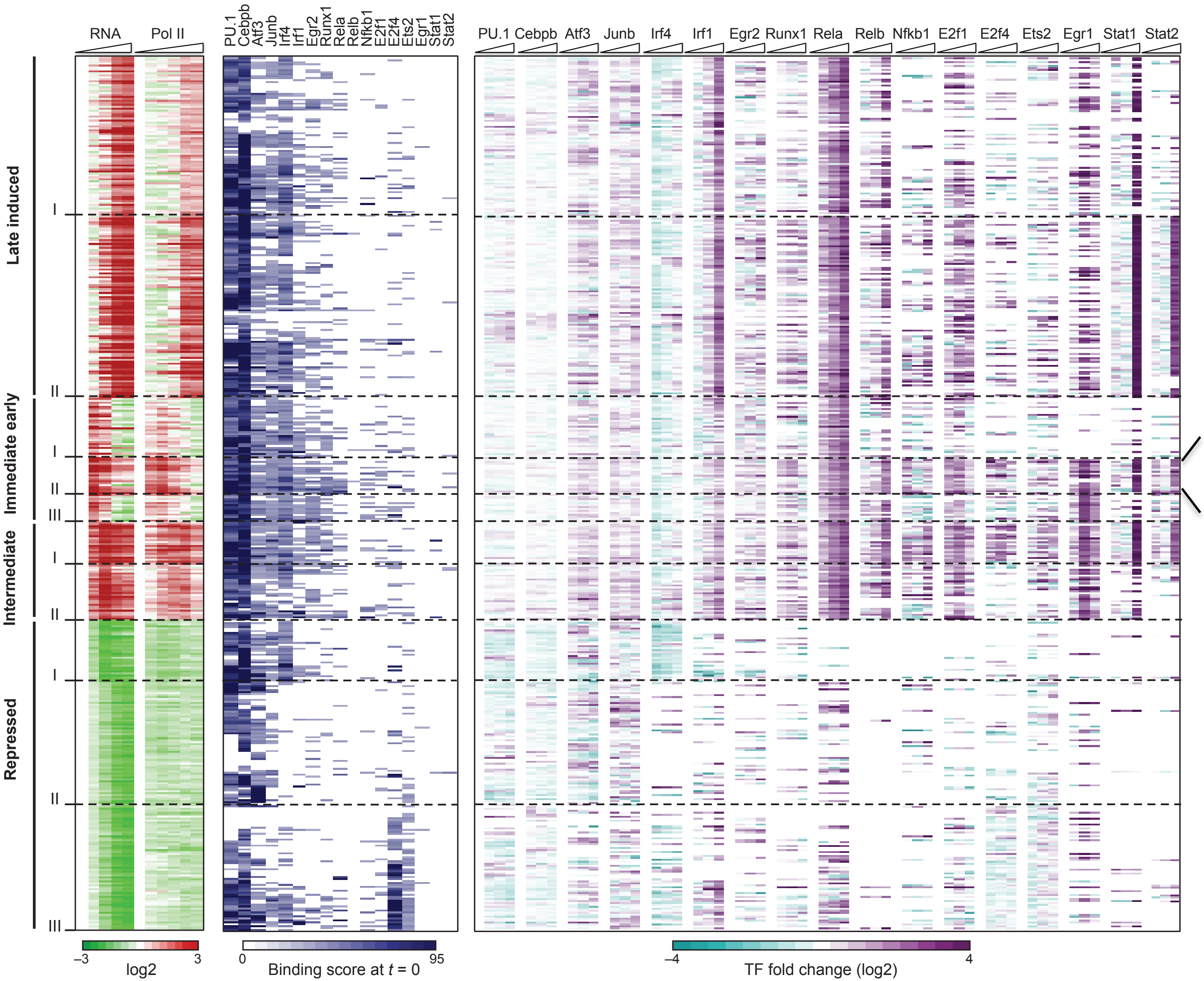
Chromatin:
Promoters
Enhancers

Transcription
Factors:
Regulation



Sequencing libraries allow us to map output, state and the circuit of the cell

Transcription factors control specific pathways



Cd274
Cd38
Tnfrsf14

Inflammation:
B, T activation

Stat1
Irf7
Mx2

Anti-viral

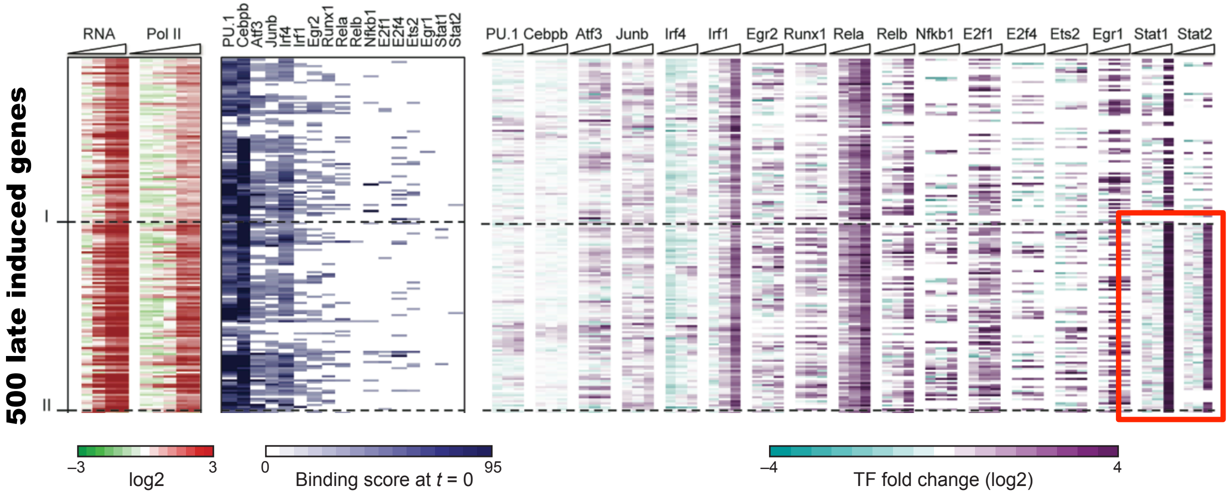
TNF
Cxcl2
Nfkbia

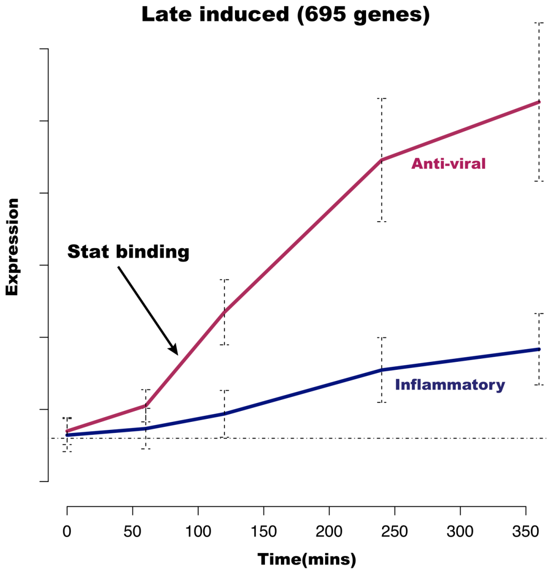
Inflammation

Anti-apoptotic

Cell cycle

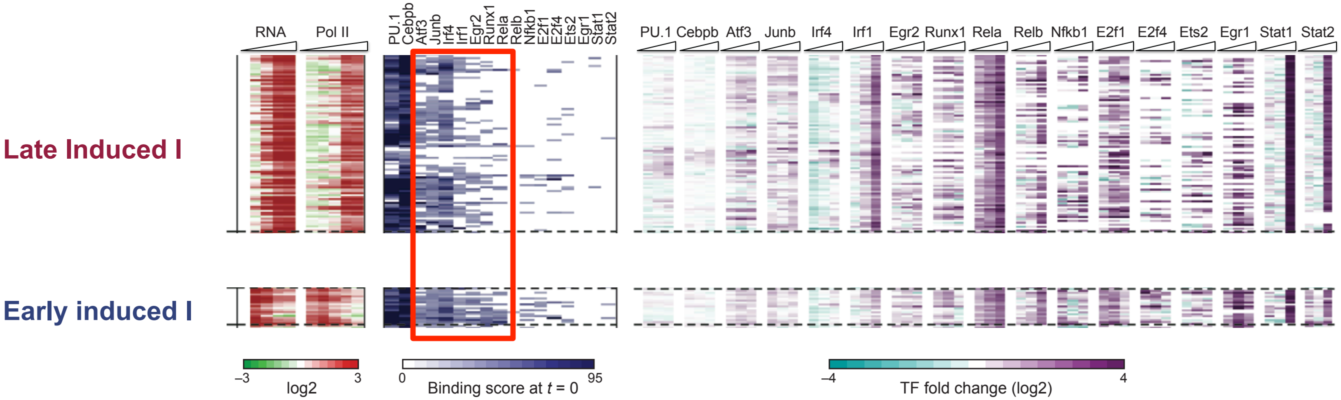
Specific factors control amplitude of expression

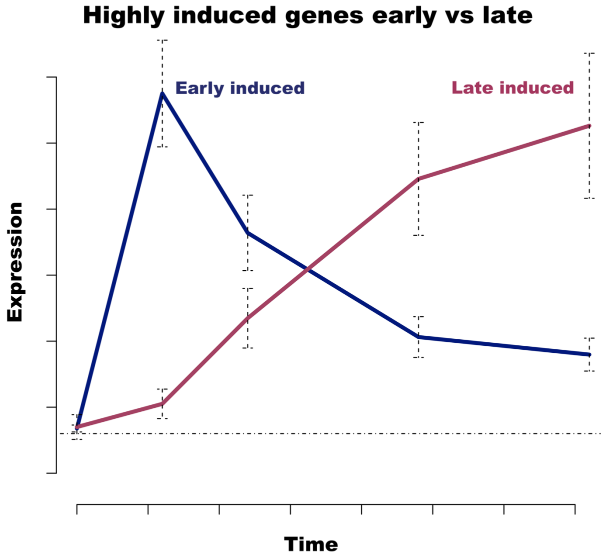




How different is the regulation of different expression patterns?

Different control of early vs. late induced genes





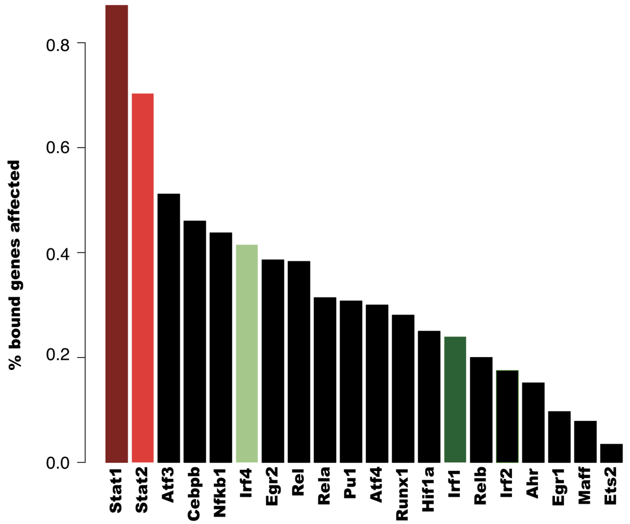
Late Induced I

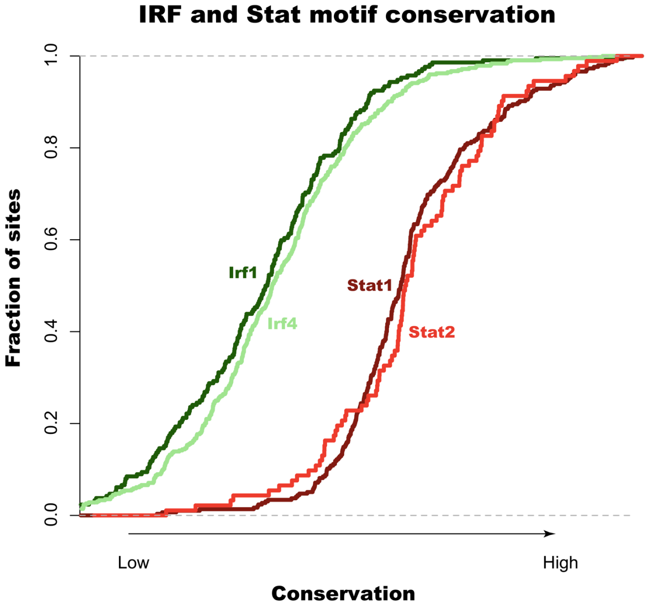
Regulated by few factors

Early induced

Highly pre-bound

Factors that control early induced genes are more redundant





Conclusions and considerations

- A large fraction of binding exist prior to stimulus
- Immediate vs. late regulation is drastically different:
  - Early induced genes regulators are more redundant
  - Late induced regulators are less redundant
  - *Are the early inflammation pathways evolutionary more malleable?*
- Factors act in layers, consistent with previous reports
- Genomic approaches like this are applicable to many systems
  - Protocols can handle smaller input material (Alon Goren, Oren Ram, Amit)
- *Test models using a genome wide genetic screens*
- *Map TFs with no available antibodies*

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

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

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

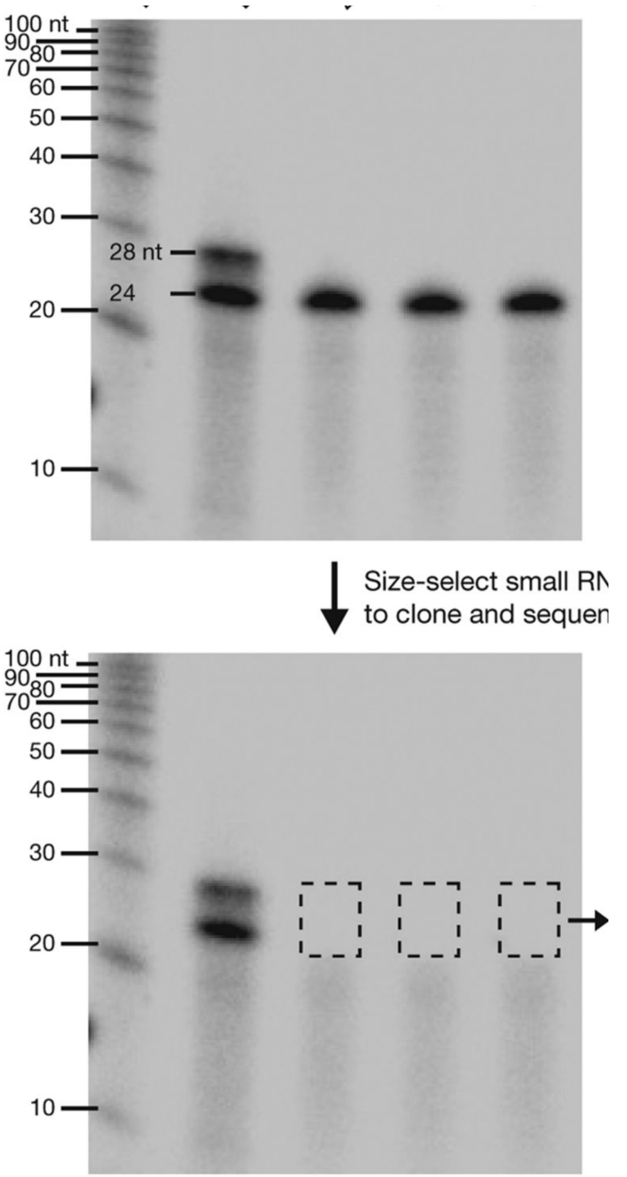
- "Source: intact, **high qual.** RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
  - Differential Gene expression. Requires ~ 5-10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

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: 5-10 mill reads

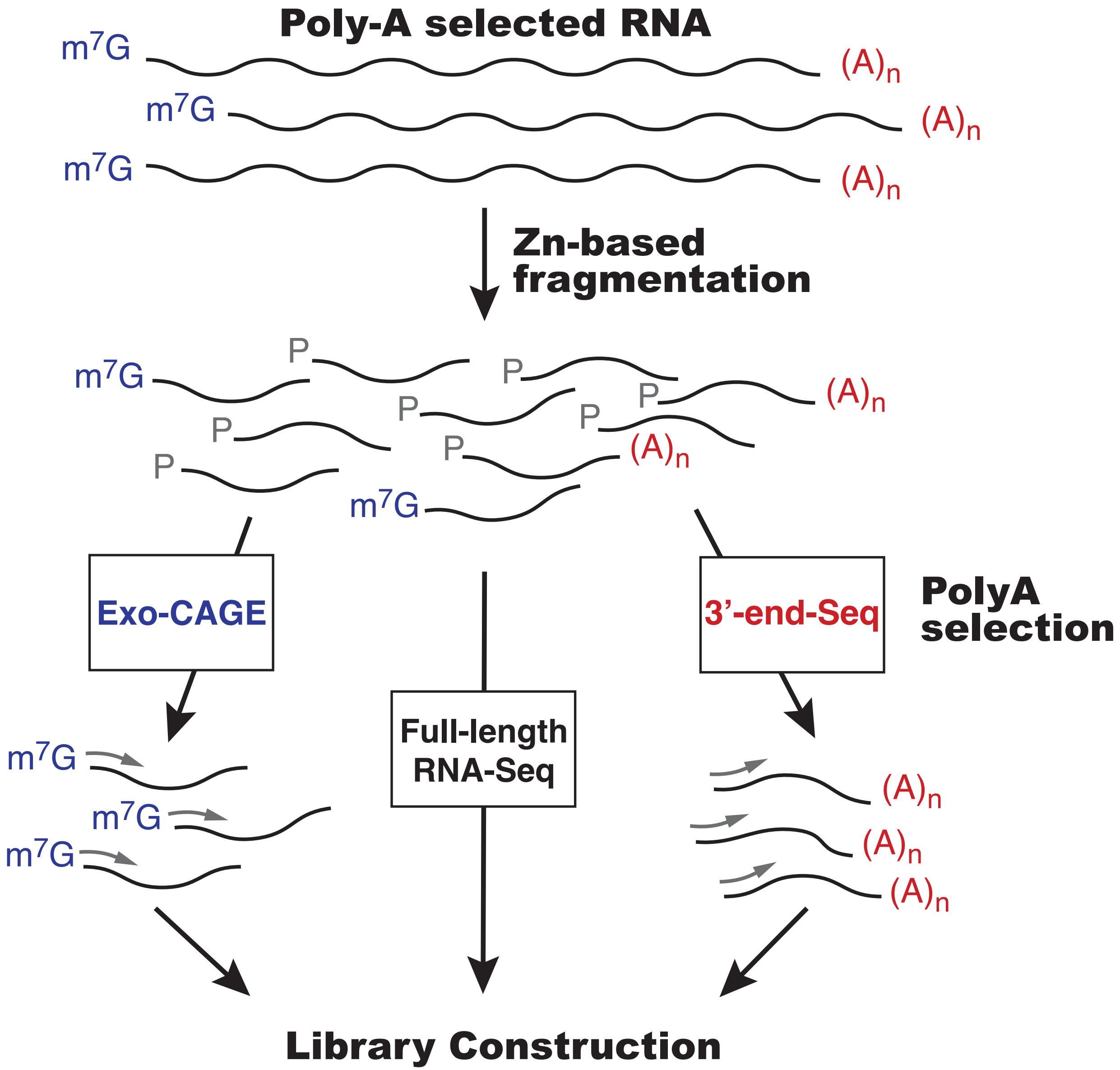


Malonne et al. CSHL protocols, 2011

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

RNA-Seq libraries: Summary



Poly-A selected RNA

Zn-based fragmentation

Exo-CAGE

Full-length RNA-Seq

3'-end-Seq

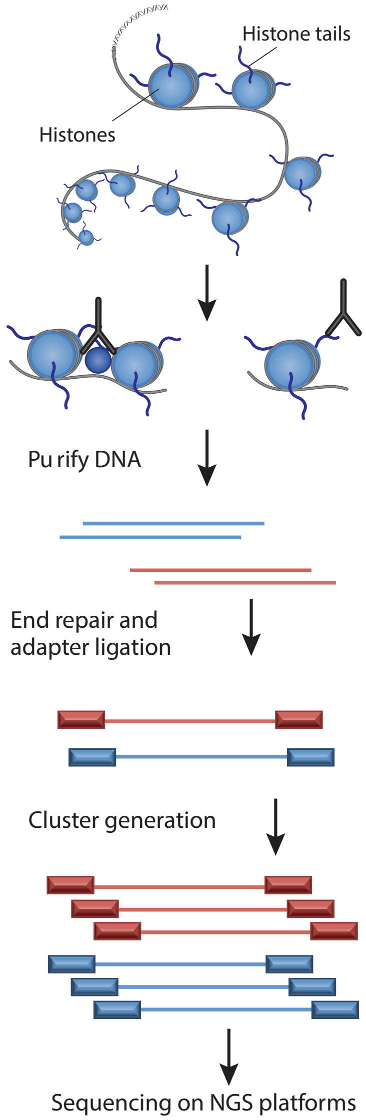
PolyA selection

Library Construction

ChIP-Seq libraries:

- Crosslinked, immunoprecipitated DNA
- DNA → sequence
- Uses:
  - Mapping nucleosomes (huge depth required)
  - Mapping histones with specific tails
  - Mapping transcription factor sites
  - Requires ~ 5-10 mill, at least 2-3 replicates, single end

ChIP-Seq protocol



Kidder et al. Nature Immunology, 2011

CLIP-Seq libraries and ribosome footprinting:

- Crosslinked, immunoprecipitated ***RNA***
- RNA→ cDNA →sequence
- Uses:
  - Mapping RNA/protein interactions
  - Find miRNA regulated transcripts
  - Mapping translation rates
  - Annotate ORFs

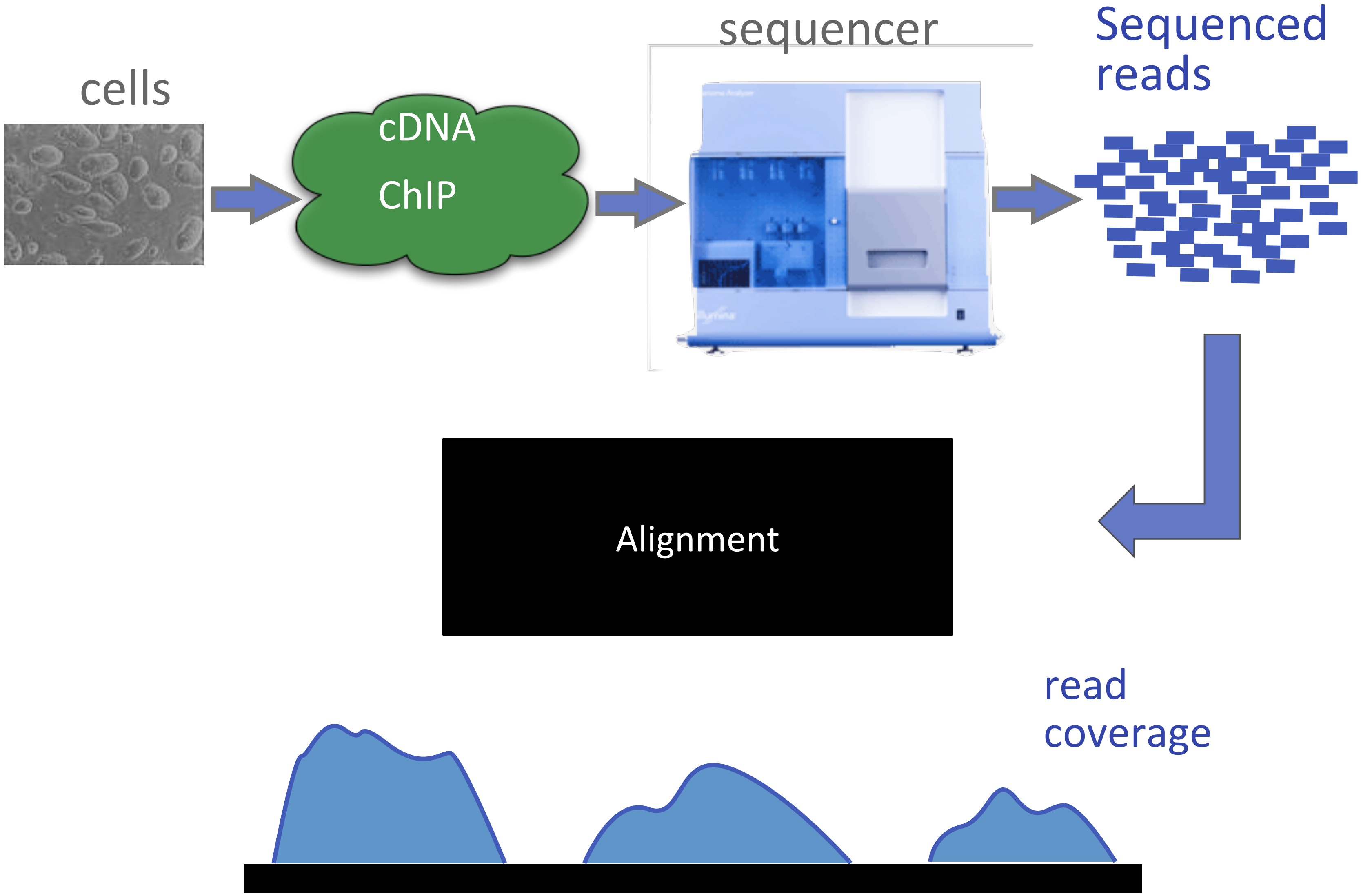
CLIP-Seq

Ribosomal profiling

Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

Once sequenced the problem becomes computational



cells

cDNA

ChIP

sequencer

Sequenced reads

Alignment

read coverage

genome

Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

The sequencing era alignment problem

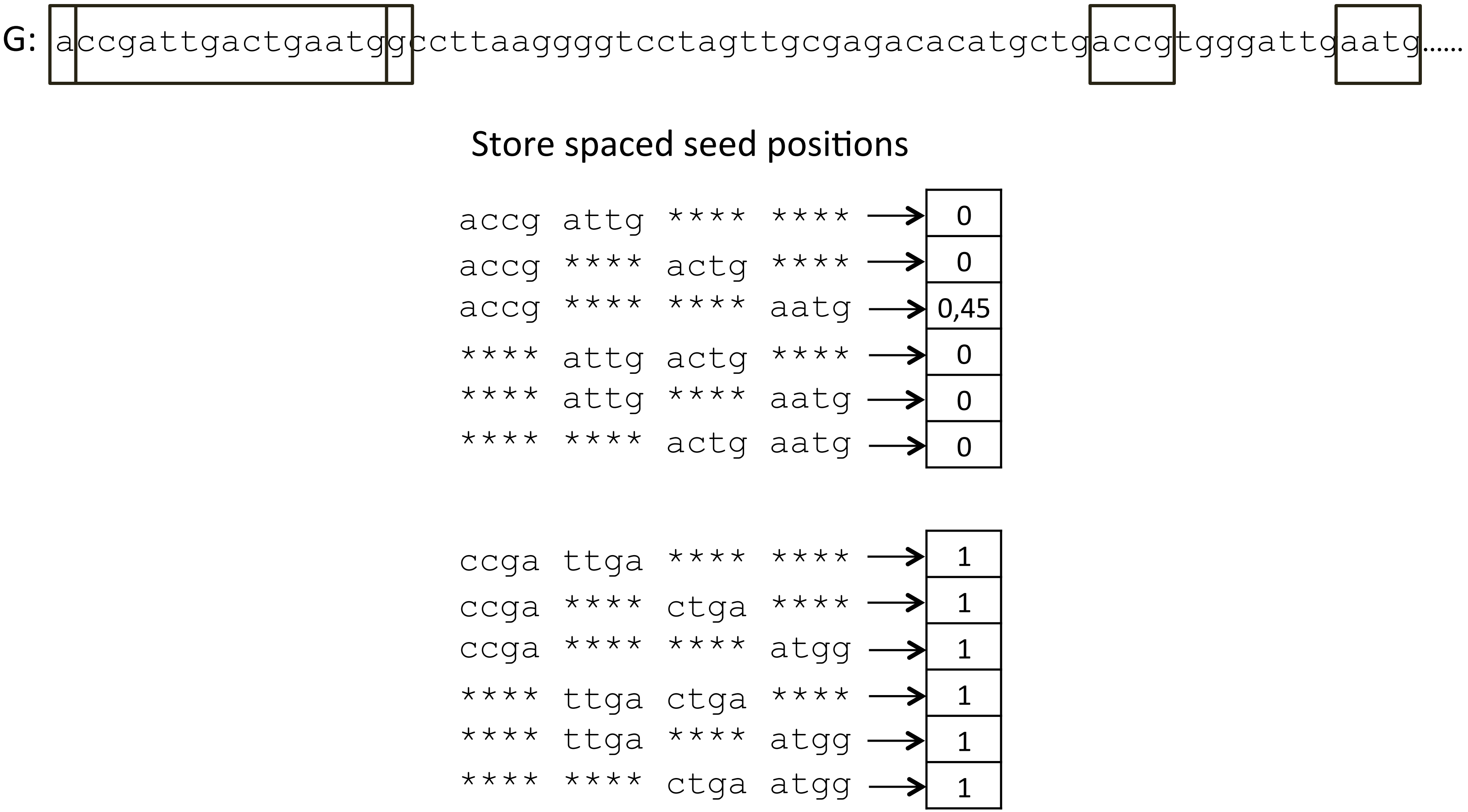
- Finding 100,000s of small (30-500 bp) sequence in a 10 - 10000 million bp genome.
- Sequences are error prone (~1% error rate)
- Reference and sequence may not be the same haplotype
- **Many techniques are great at finding perfect matches**

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
  - Use algorithms that handle mismatches and gaps

Spaced seed alignment – Hashing the genome



G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg......

Store spaced seed positions

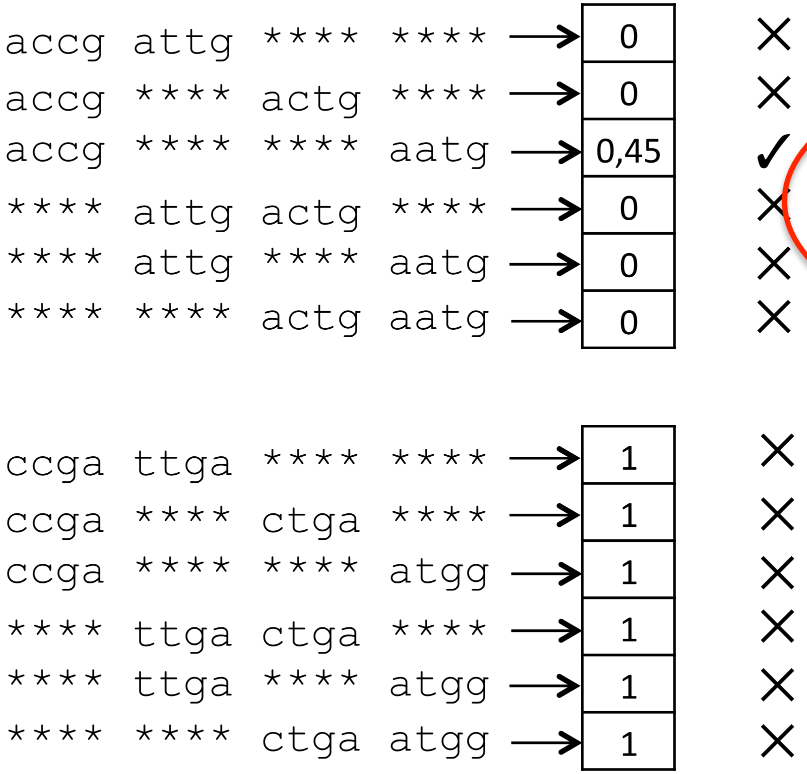
accg	attg	****	****	0
accg	****	actg	****	0
accg	****	****	aatg	0,45
****	attg	actg	****	0
****	attg	****	aatg	0
****	****	actg	aatg	0

ccga	ttga	****	****	1
ccga	****	ctga	****	1
ccga	****	****	atgg	1
****	ttga	ctga	****	1
****	ttga	****	atgg	1
****	****	ctga	atgg	1

Spaced seed alignment – Mapping reads

G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg.....

q: accg atag accg aatg



accg attg **** ****	0	✗
accg **** actg ****	0	✗
accg **** **** aatg	0,45	✓
**** attg actg ****	0	✗
**** attg **** aatg	0	✗
**** **** actg aatg	0	✗
ccga ttga **** ****	1	✗
ccga **** ctga ****	1	✗
ccga **** **** atgg	1	✗
**** ttga ctga ****	1	✗
**** ttga **** atgg	1	✗
**** **** ctga atgg	1	✗

accgattgactgaatg — 2 missmatches

accgtgggattgaatg — 5 missmatches

Report position 0

But, how confidence are we in the placement?

$$q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$$

Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

Lets compute the probability the read originated at genome position i

q: `accg atag accg aatg`

q_s: `30 40 25 30  30 20 10 20  40 30 20 30  40 40 30 25`

$q_s[k] = -10\log_{10} P(\text{sequencing error at base k})$, the PHRED score. Equivalently:

$$P(\text{sequencing error at base k}) = 10^{-\frac{q_s[k]}{10}}$$

So the probability that a read originates from a given genome position i is:

$$P(q \mid G,i) = \prod_{j \text{ match}} P(q_j \text{good call}) \prod_{j \text{ missmatch}} P(q_j \text{bad call}) \approx \prod_{j \text{ missmatch}} P(q_j \text{bad call})$$

In our example

$$P(q \mid G,0) = \left[(1-10^{-3})^6(1-10^{-4})^4(1-10^{-2.5})^2(1-10^{-2})^2\right]\left[10^{-1}10^{-2}\right] = [0.97]*[0.001] \approx 0.001$$

Mapping quality

What we want to estimate is $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$

That is, the posterior probability, the probability that the region starting at i was sequenced *given* that we observed the read *q*:

$$P(i \mid q) = \frac{P(q \mid i)P(i)}{P(q)} = \frac{P(q \mid i)P(i)}{\sum_j P(q \mid j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H *genome Research* 2008, for example)

$$q_{MS} = -10\log_{10}(1 - P(i \mid q))$$

Considerations

- Trade-off between sensitivity, speed and memory
  - Smaller seeds allow for greater mismatches at the cost of more tries
  - Smaller seeds result in a smaller tables (table size is at most 4^k), larger seeds increase speed (less tries, but more seeds)

Considerations

- BWT-based algorithms rely on perfect matches for speed
- When dealing with mismatches, algorithms “backtrack” when the alignment extension fails.
- Backtracking is expensive
- As read length increases novel algorithms are required

Short read mapping software for ChIP-Seq

Seed-extend

	Short indels	Use base qual
Maq	**No**	**YES**
RMAP	Yes	**YES**
SeqMap	Yes	NO
SHRiMP	Yes	NO

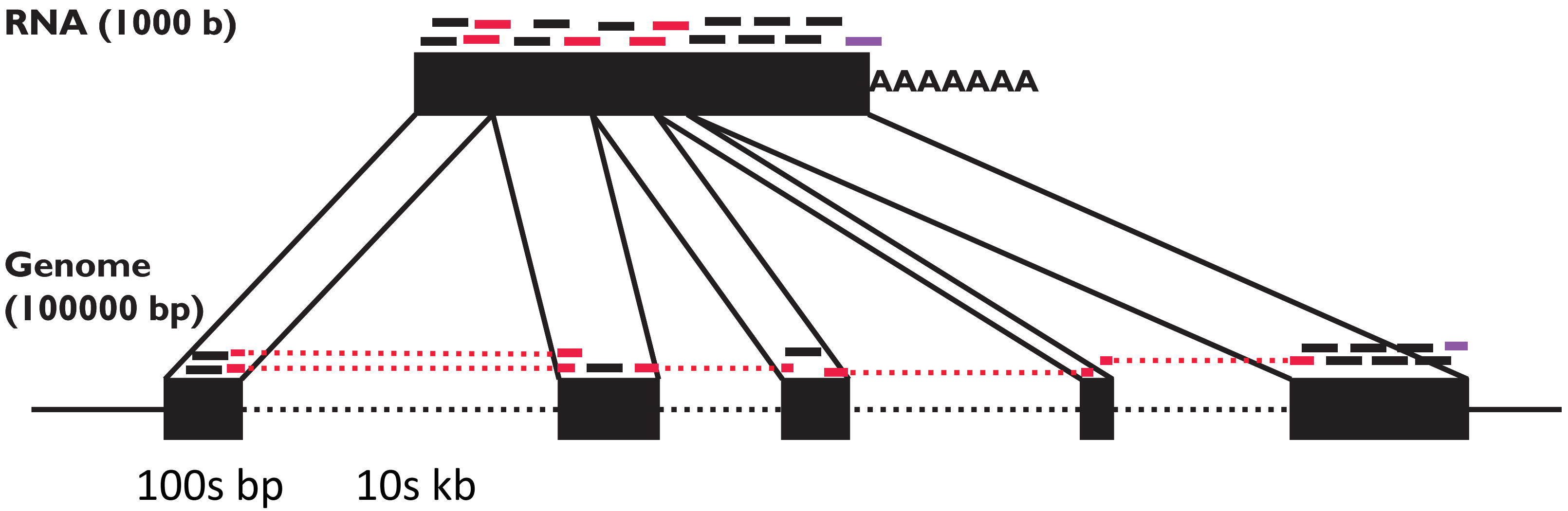
BWT

	Use Base qual
BWA	**YES**
Bowtie	NO
Stampy*	YES
Bowtie2*	(NO)

***Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA**

***Bowtie2 breaks reads into smaller pieces and maps these "seeds" using a BWT genome.**

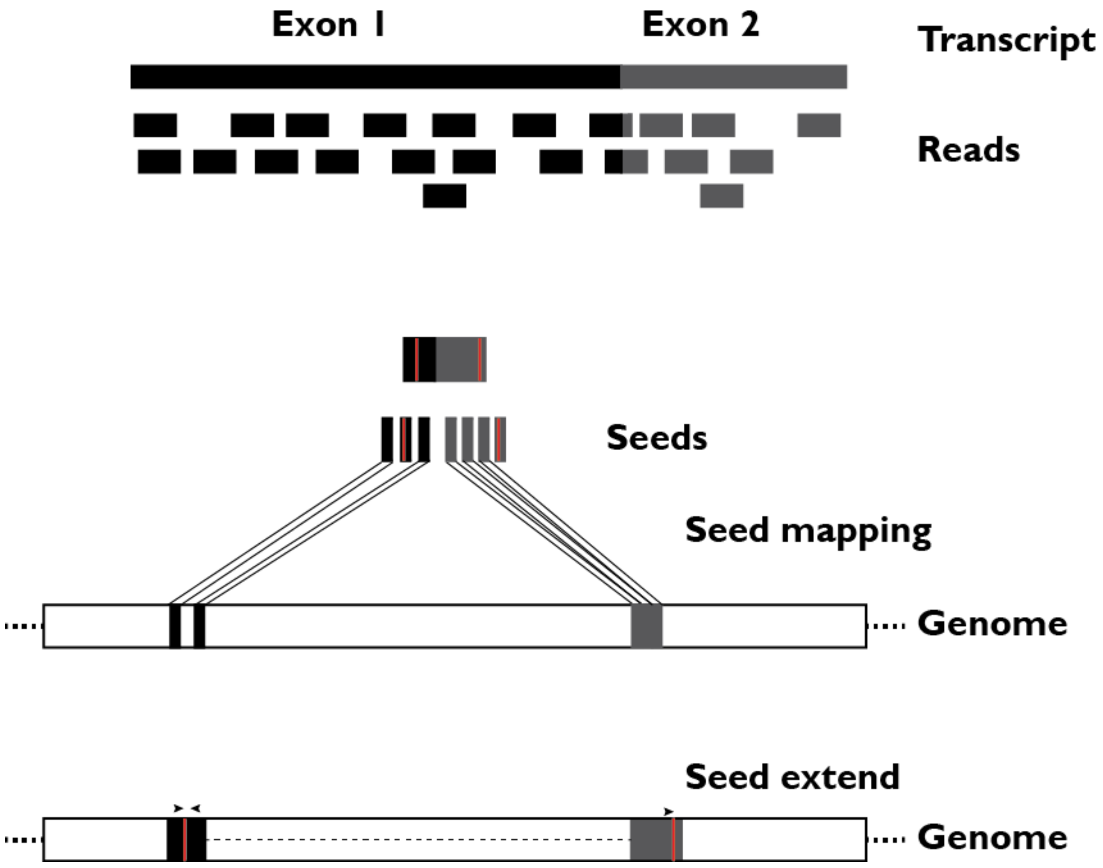
The RNA-Seq alignment problem



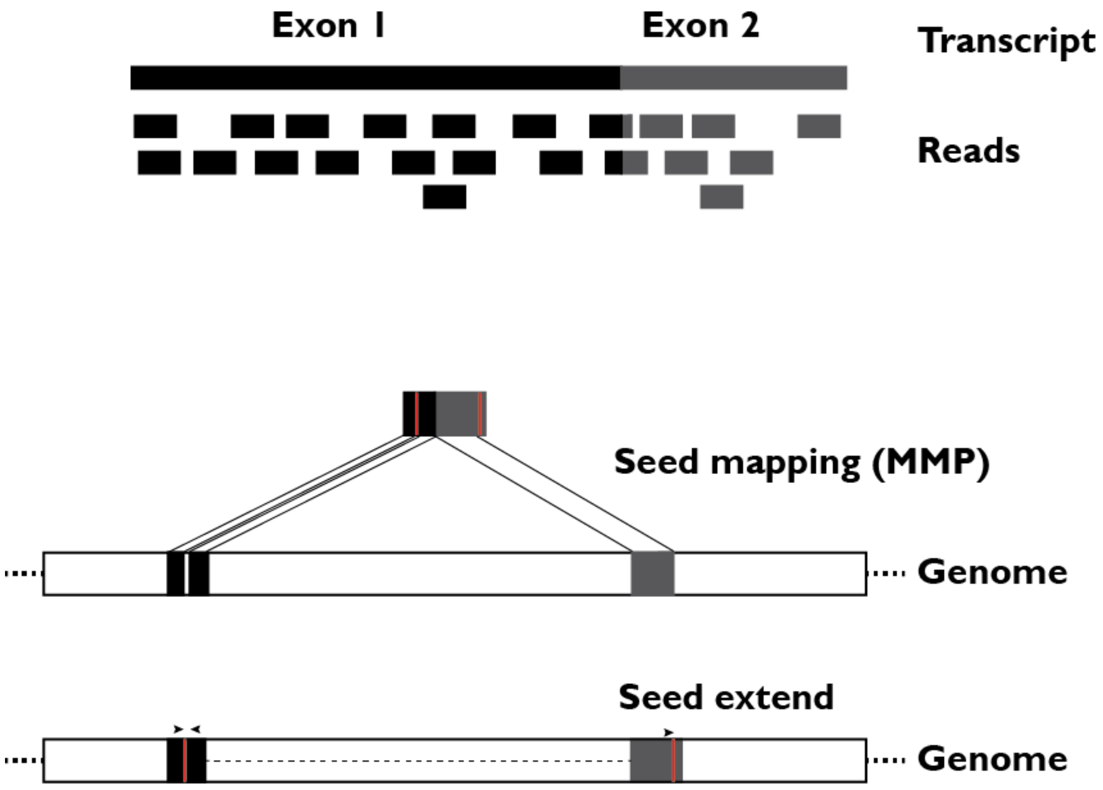
Challenges:

- 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: Exon-first spliced alignment (e.g. TopHat2)

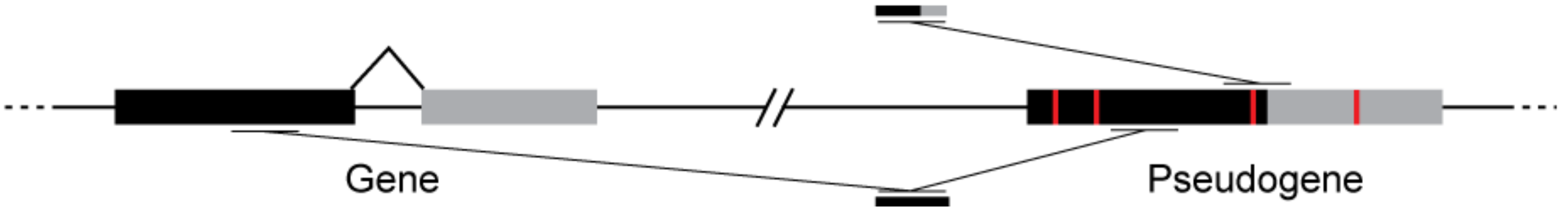


Mapping RNA-Seq reads: Maximal Mapping Prefix (STAR)

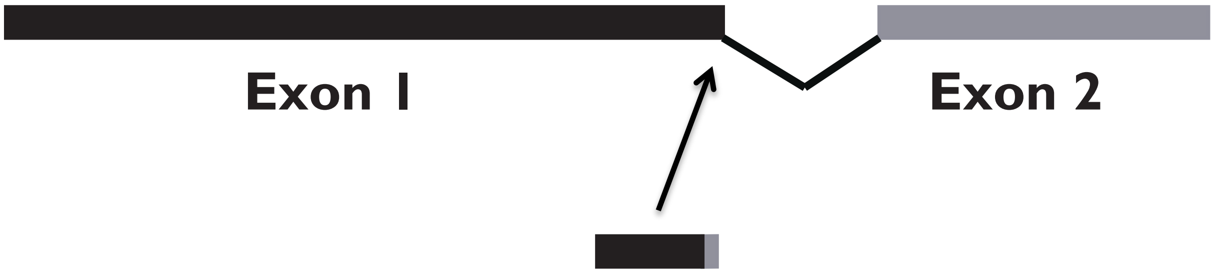


RNA-Seq specific problems

Pseudo gene attraction problem



Intron invasion



Current aligners deal directly with these problems

Short read mapping software for RNA-Seq

Seed-extend

	Short indels	Use base qual
STAR	**Yes**	?
QPALMA	Yes	NO
BLAT	Yes	NO

Exon-first

	Use base qual
TopHat2	NO

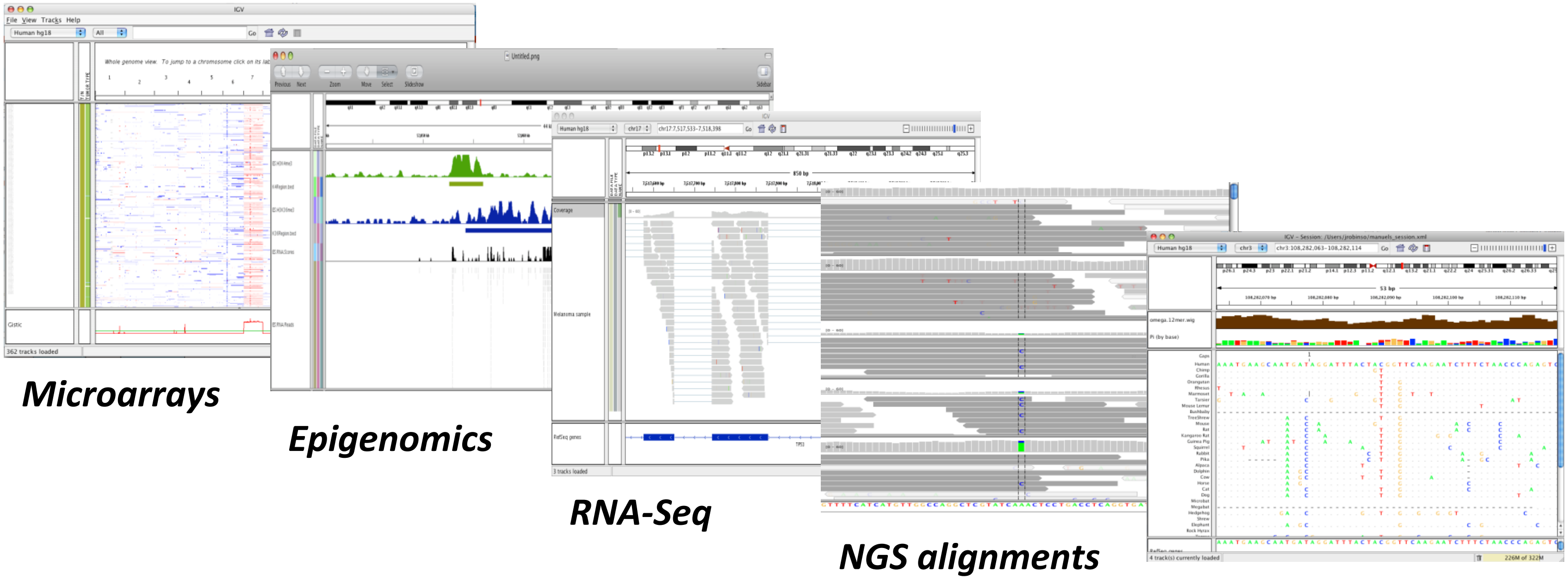
Exon-first alignments will map contiguous first at the expense of spliced hits

IGV: Integrative Genomics Viewer

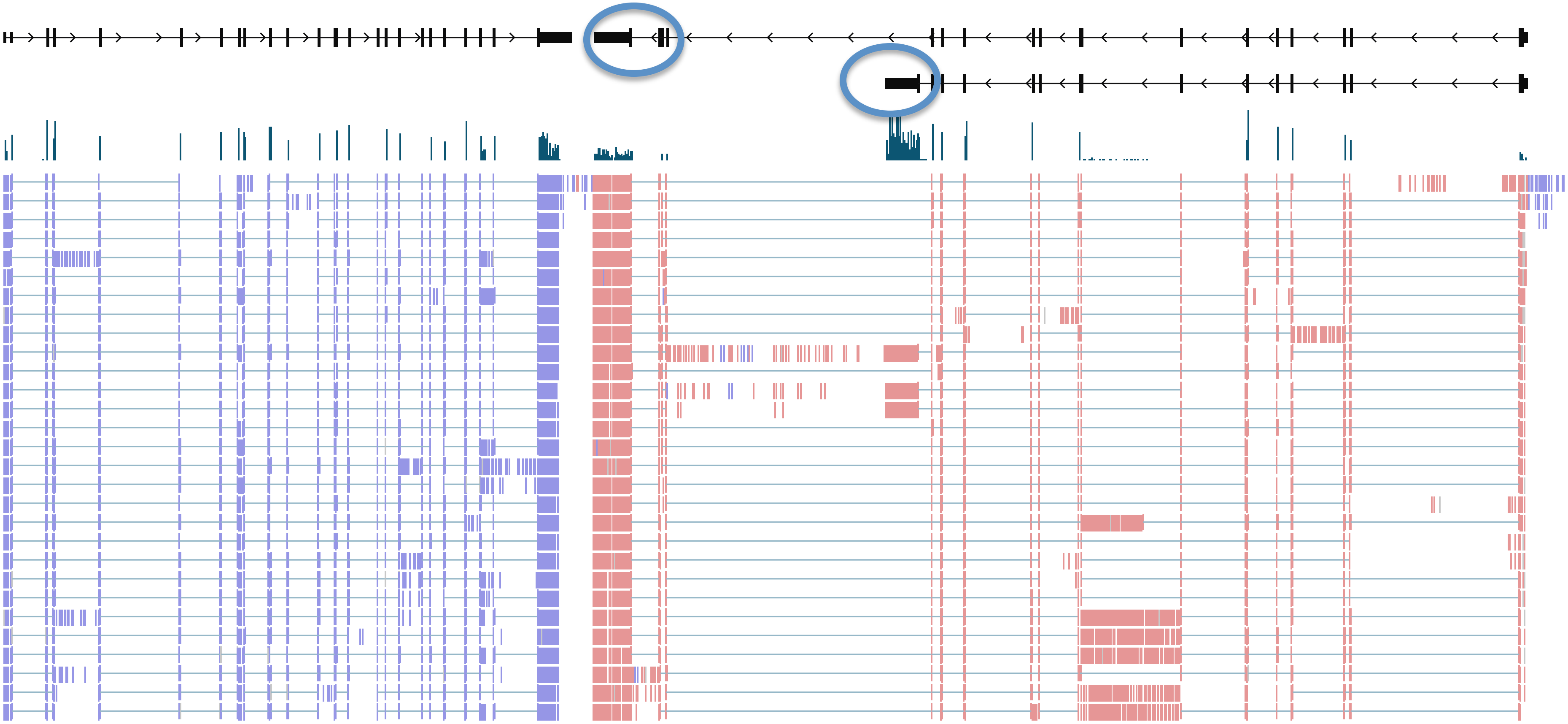
A desktop application

for the visualization and interactive exploration

of genomic data



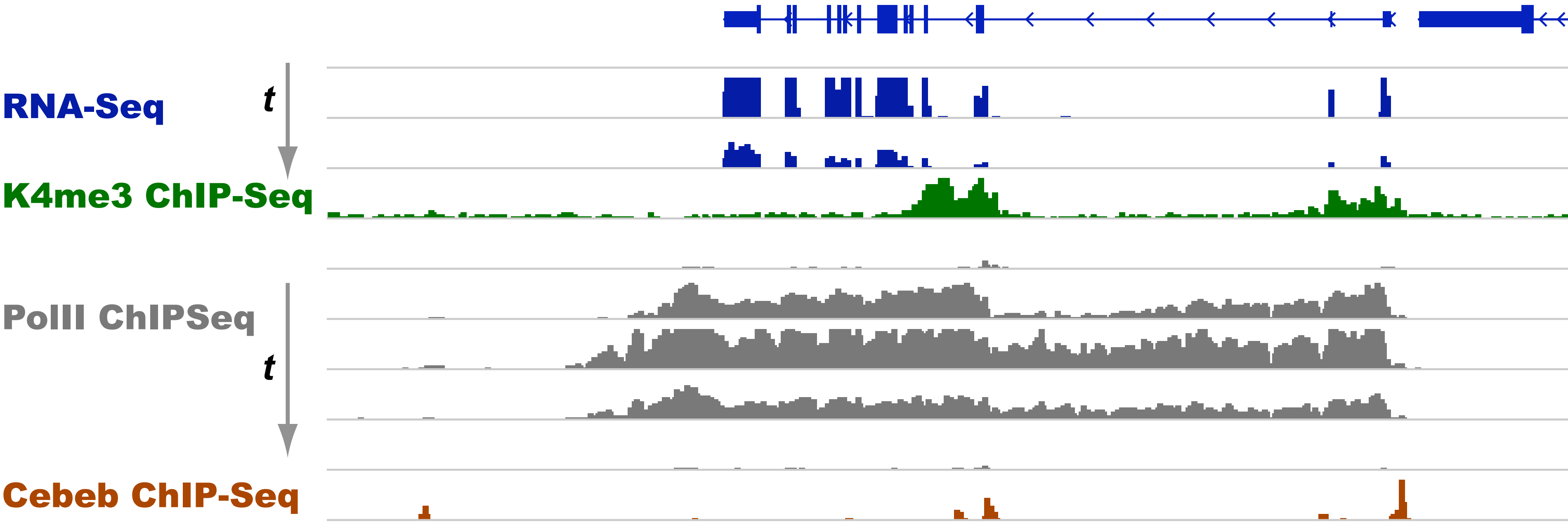
Visualizing read alignments with IGV — RNASeq



Strand specific library!

Gap between reads spanning exons

Visualizing read alignments with IGV — zooming out



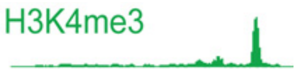
Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

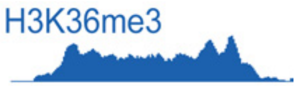
What does significance means?

- 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

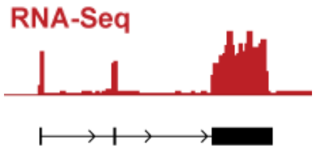
How do we find peaks?



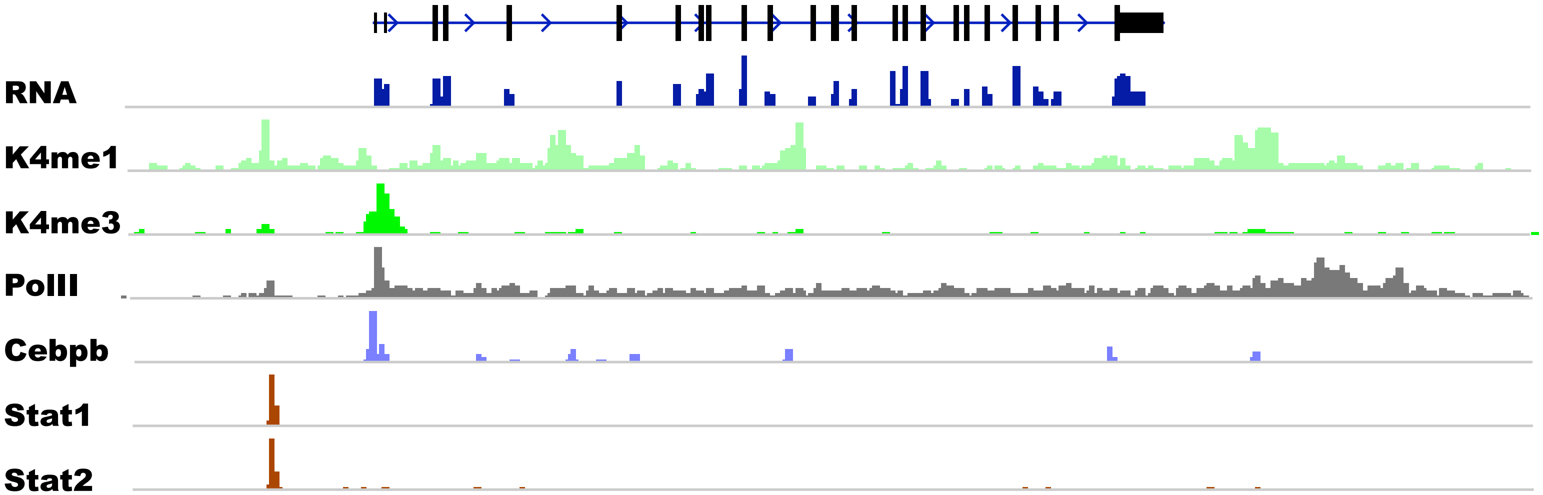
Short modification



Long modification

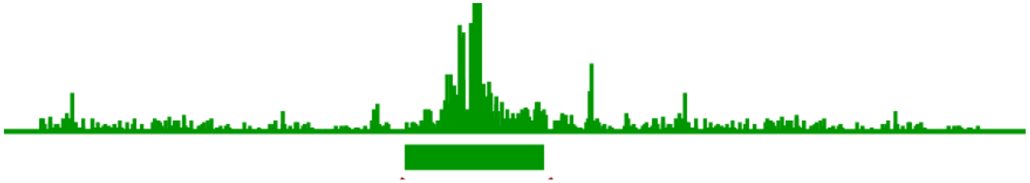


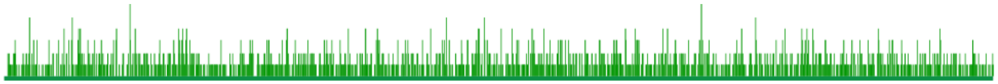
Discontinuous data

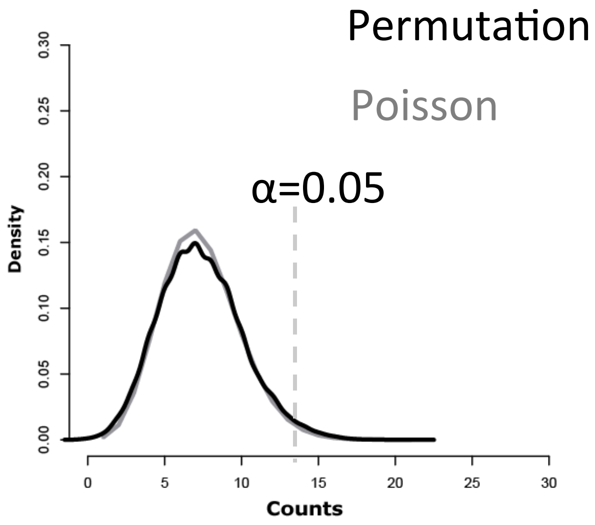


Scripture is a method to solve this general question

Our approach



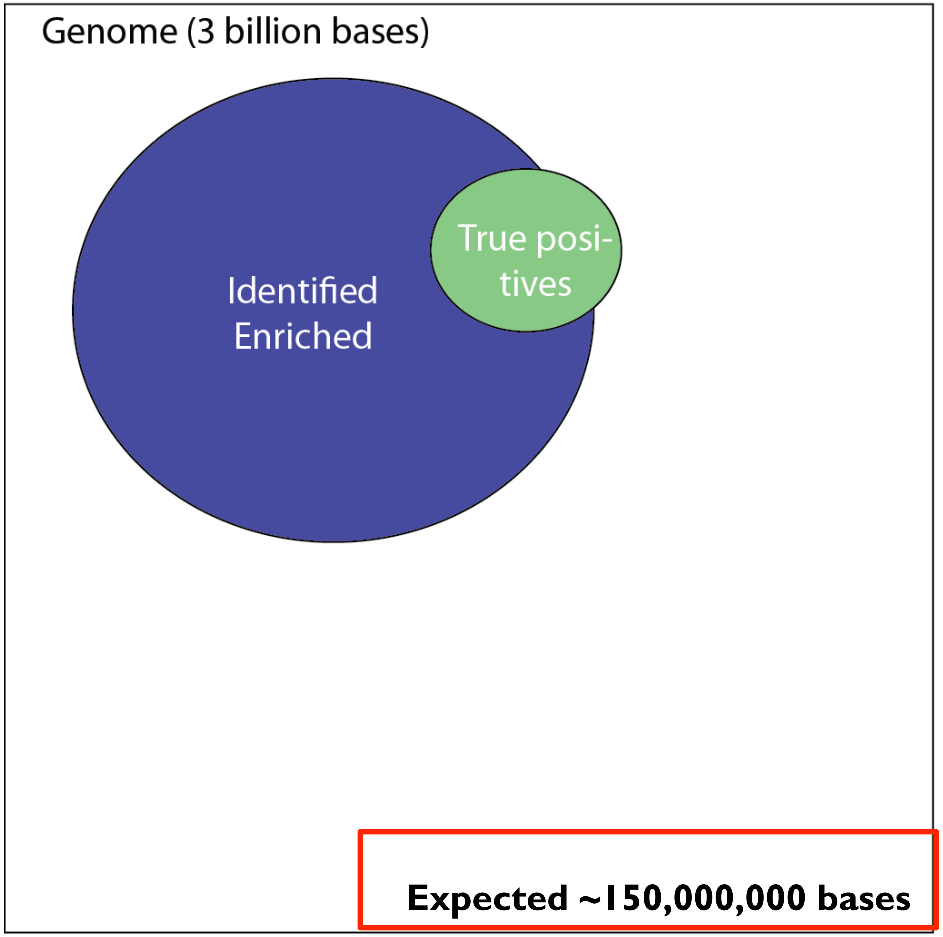




We have an efficient way to compute read count p-values …

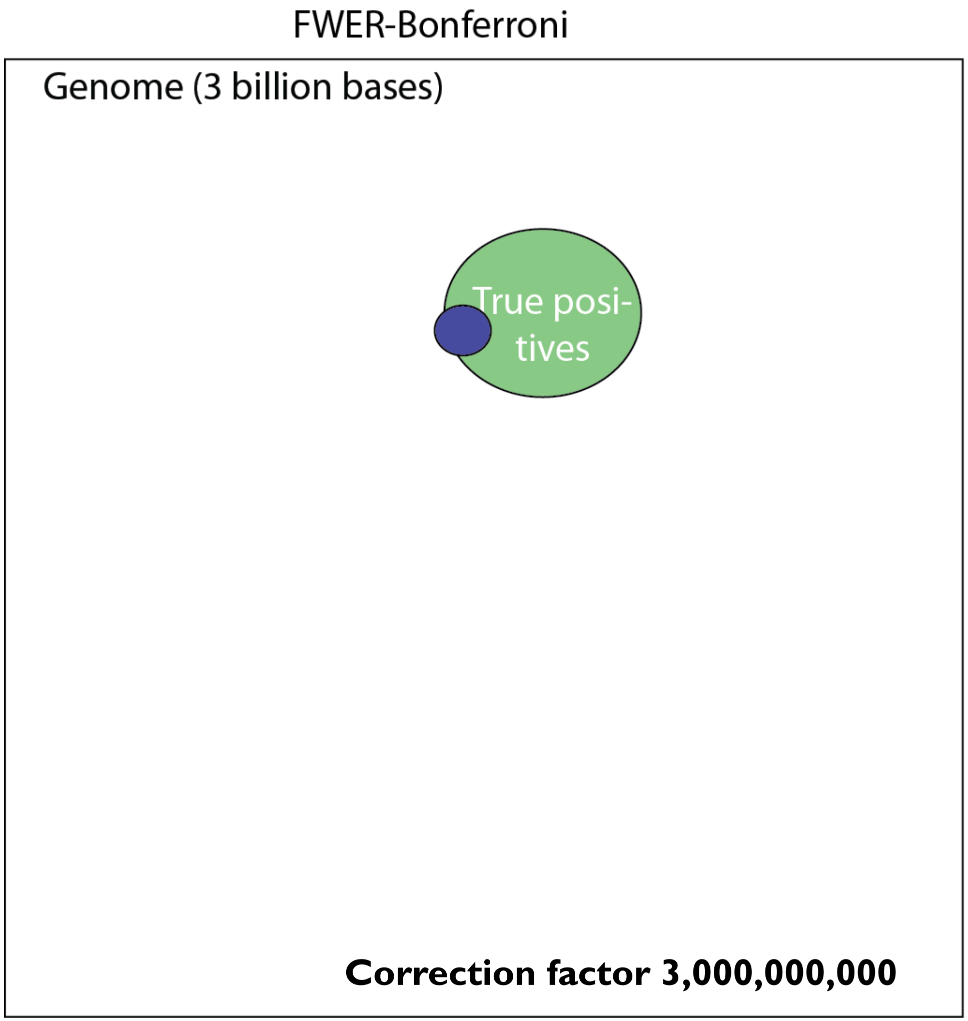
The genome is large, many things happen by chance

Nominal P-Value



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative

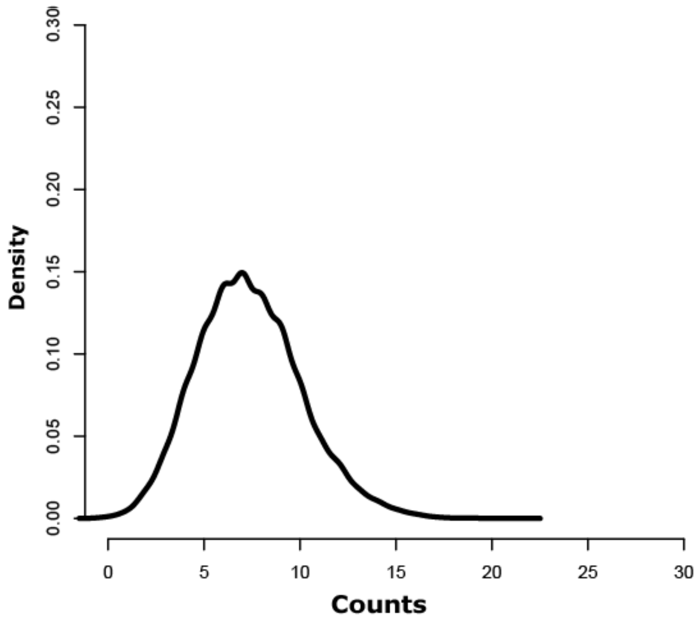


Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

Controlling FWER

Max Count distribution

α=0.05 α_{FWER}=0.05



Count distribution (Poisson)

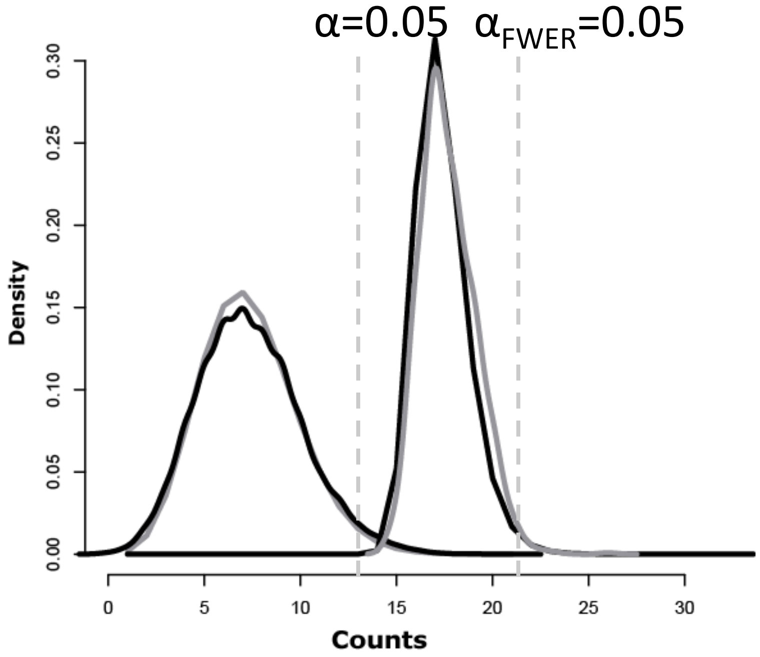
Given a region of size w and an observed read count n. What is the probability that one or more of the 3x10^9 regions of size w has read count >= n under the null distribution?

We could go back to our permutations and compute an FWER: **max of the genome-wide distributions of same sized region**)→
but really really really slow!!!

Scan distribution, an old problem

- Is the observed number of read counts over our region of interest high?
- Given a set of Geiger counts across a region find clusters of high radioactivity
- Are there time intervals where assembly line errors are high?

Scan distribution

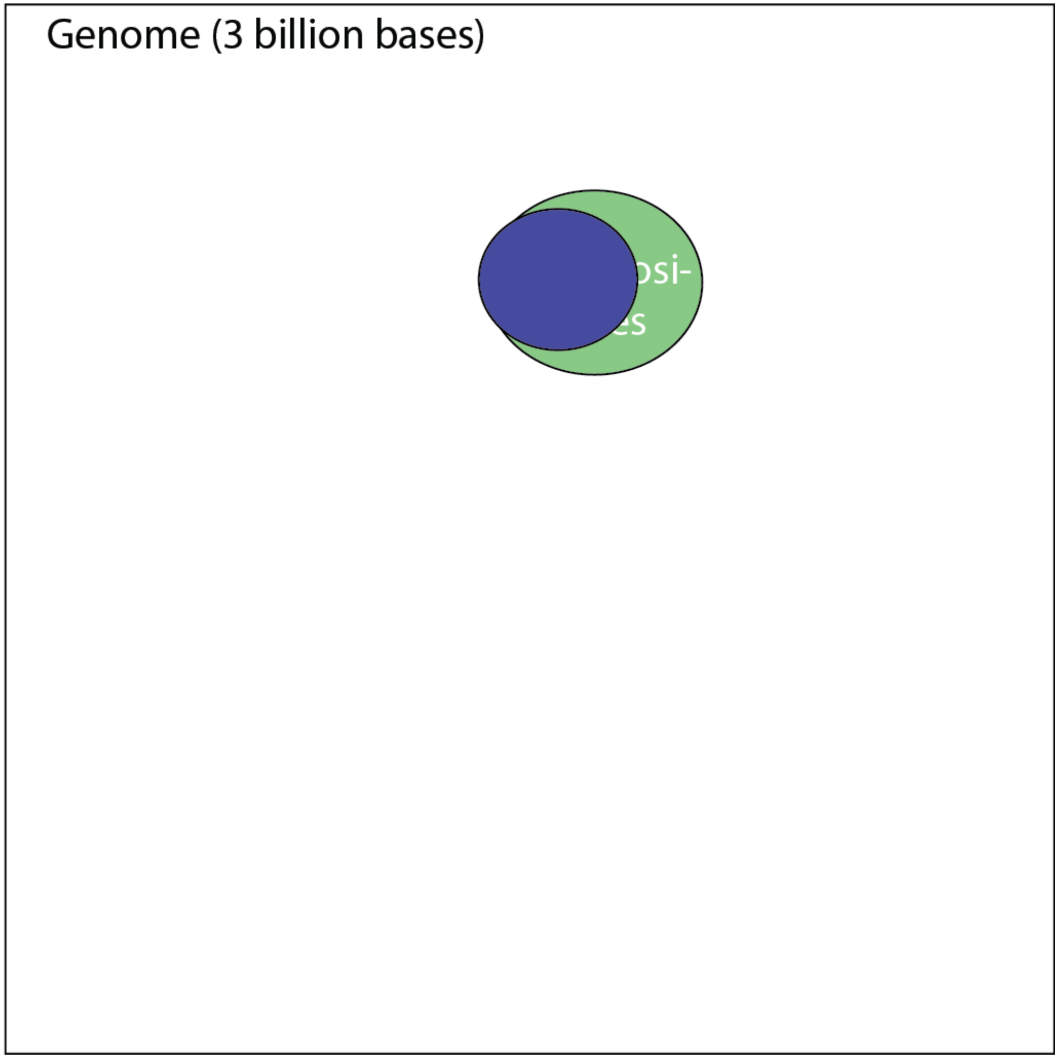


Poisson distribution

Thankfully, the ***Scan Distribution*** computes a closed form for this distribution.

ACCOUNTS for dependency of overlapping windows thus more powerful!

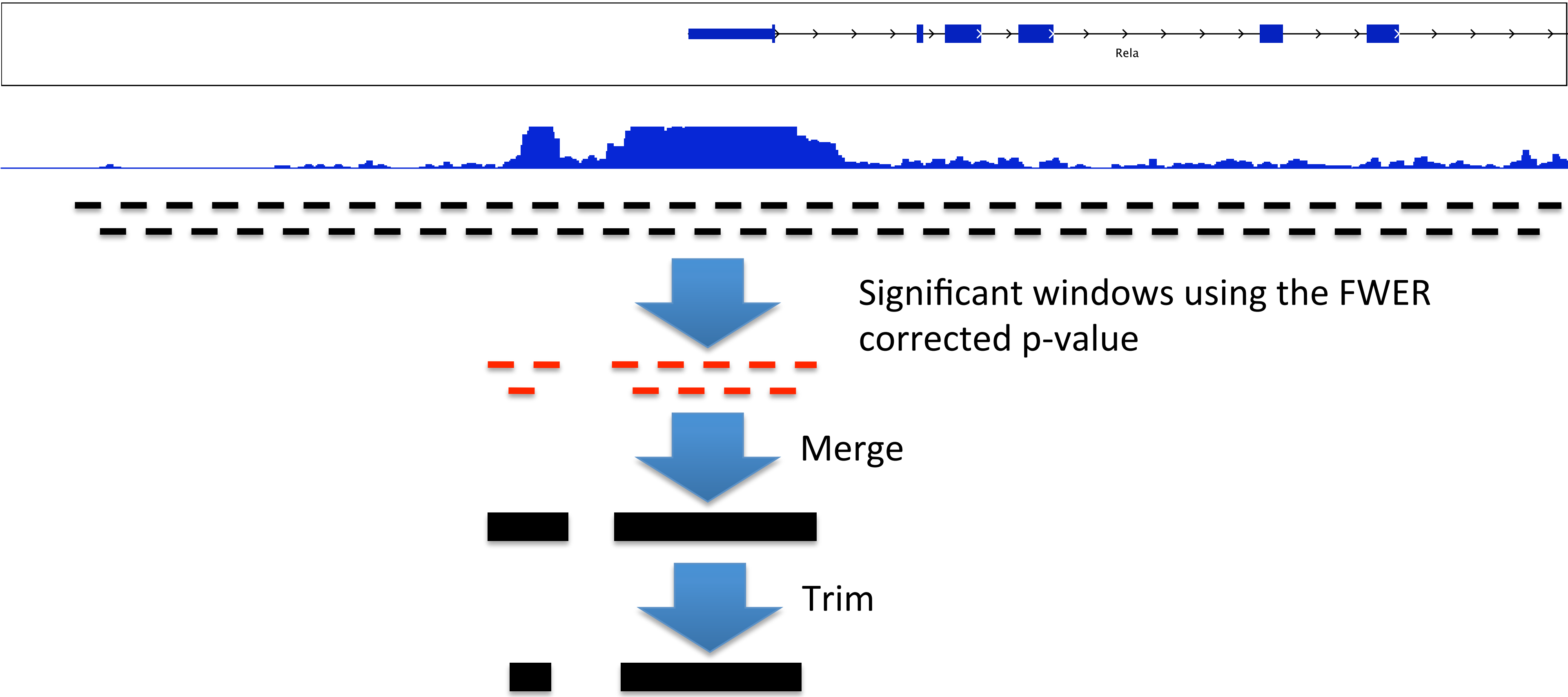
FWER-Scan Statistics



By utilizing the dependency of overlapping windows we have greater power, while still controlling the same genome-wide false positive rate.

Segmentation method for contiguous regions

Example : PolII ChIP



Significant windows using the FWER corrected p-value

Merge

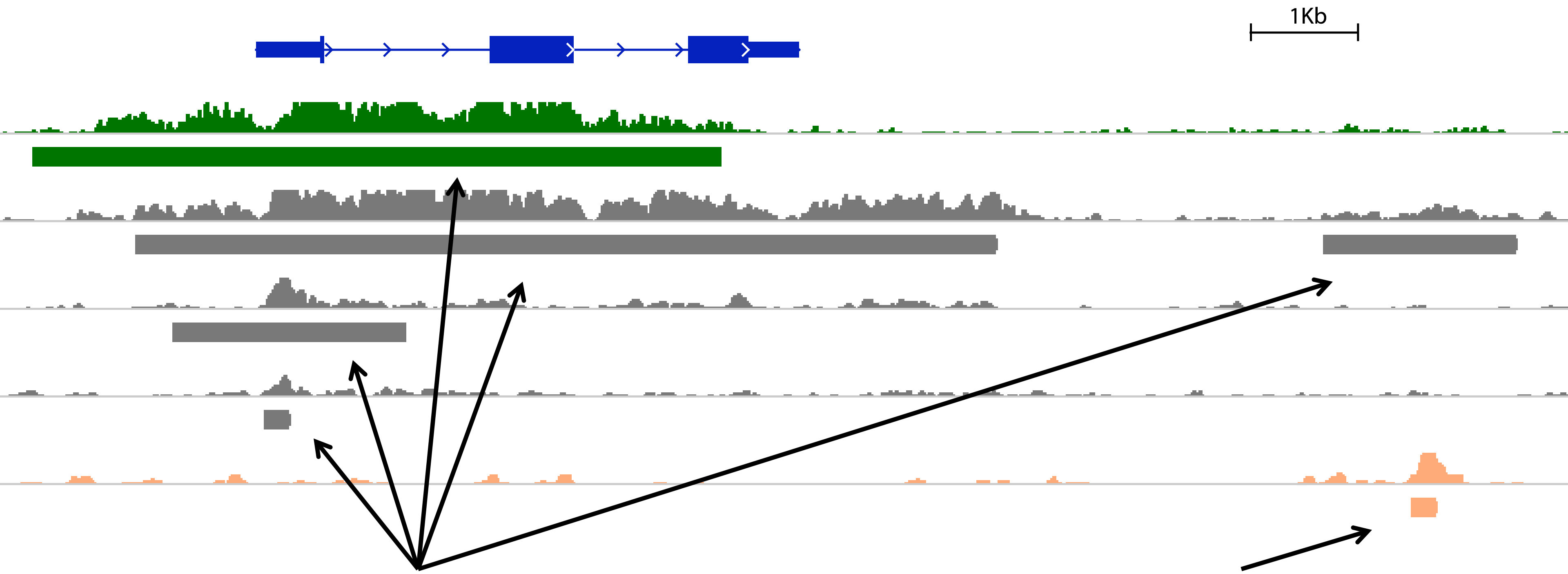
Trim

But, which window?

We use multiple windows

- Small windows detect small punctuate regions.
- Longer windows can detect regions of moderate enrichment over long spans.
- In practice we scan different windows, finding significant ones in each scan.
- In practice, it helps to use some prior information in picking the windows although globally it might be ok.

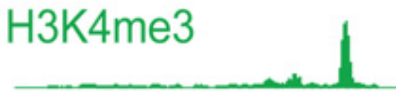
Applying Scripture to a variety of ChIP-Seq data



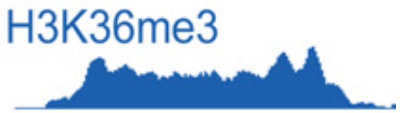
200, 500 & 1000 bp windows

100 bp windows

Can we identify enriched regions across different libraries?

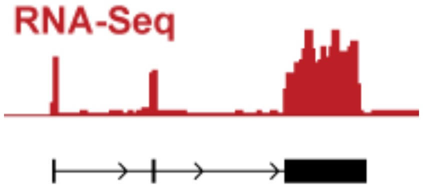


Short modification ✓



Long modification ✓

Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?



Discontinuous data: RNA-Seq to find gene structures for this gene-like regions

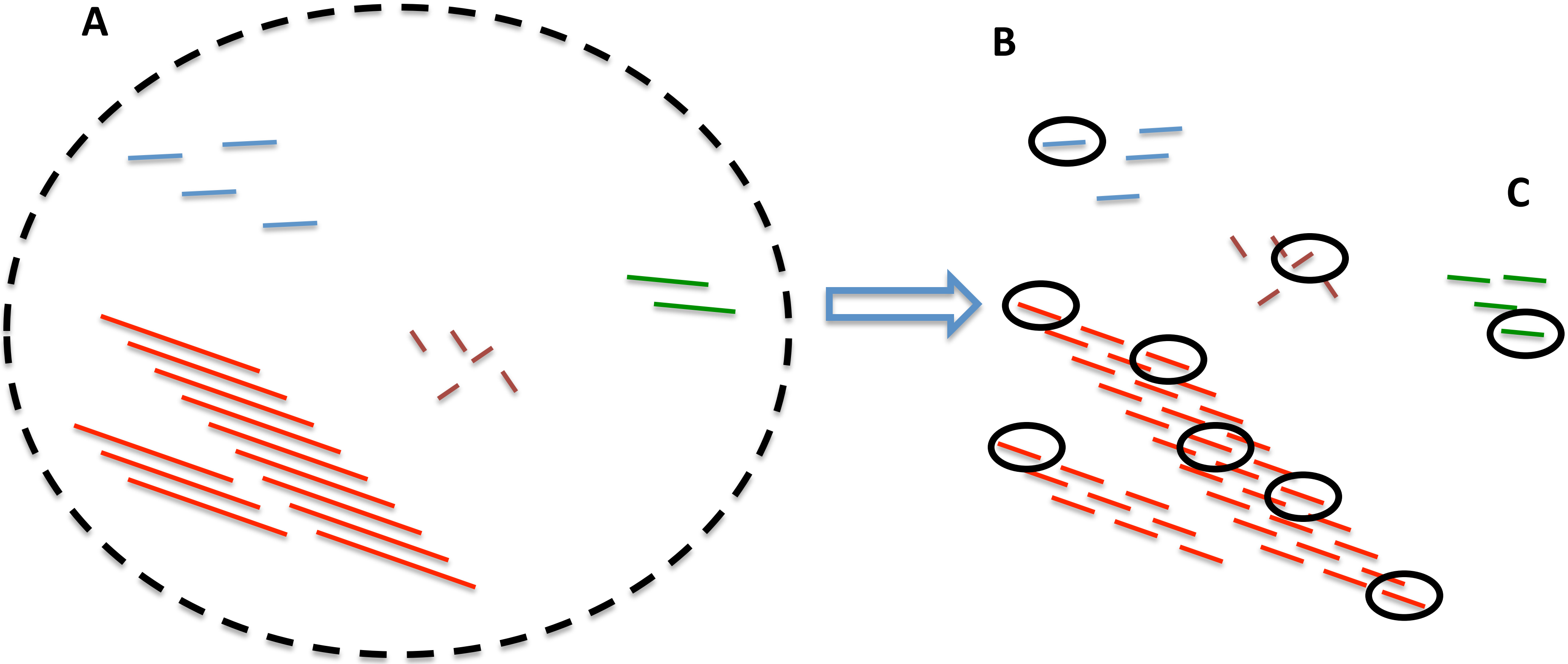
Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

RNA-Seq quantification

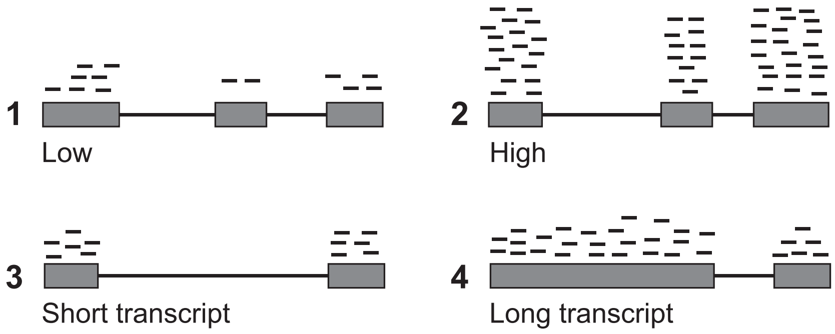
- Is a given gene (or isoform) expressed?
- Is expression gene A > gene B?
- Is expression of gene A isoform a_1 > gene A isoform a_2?
- Given two samples is
  - expression of gene A in sample 1 different from gene A in sample 2?
  - Is the expression of one isoform changing?

RNA-Seq measures relative abundance



RNA-Seq quantification: Infer fraction of molecules in sample

RNA-Seq quantification units



$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

Reads per kilobase of exonic sequence per million mapped reads
(*Mortazavi* et al Nature methods 2008)

- •Fragmentation of transcripts results in length bias: longer transcripts have higher counts
- •Different experiments have different yields. Normalization is key for cross lane comparisons

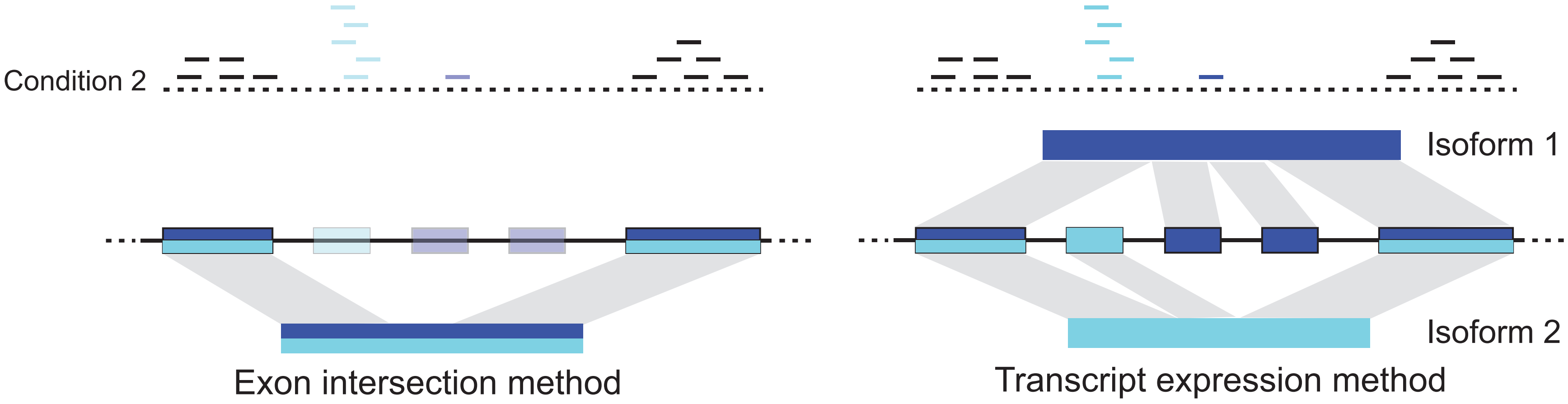
Garber et al. Nat. Methods 2011

RNA-Seq quantification "units"

- To compare within a sequence run (lane), RPKM accounts for length bias.
- RPKM (Mortazavi et al 2008) is not optimal for cross experiment comparisons.
  - Different samples may have different compositions.
- FPKM (Trapnell et al. 2011) superseded RPKM to deal with paired end data
  - Paired end reads originate from the same Fragment
- And later TPM = 10^6 x Fraction of transcript in sample (Li et al 2009)
  - More robust to changes in sample composition

Complexity increases when multiple isoforms exist

But, how to compute counts for complex gene structures?



Three popular options:

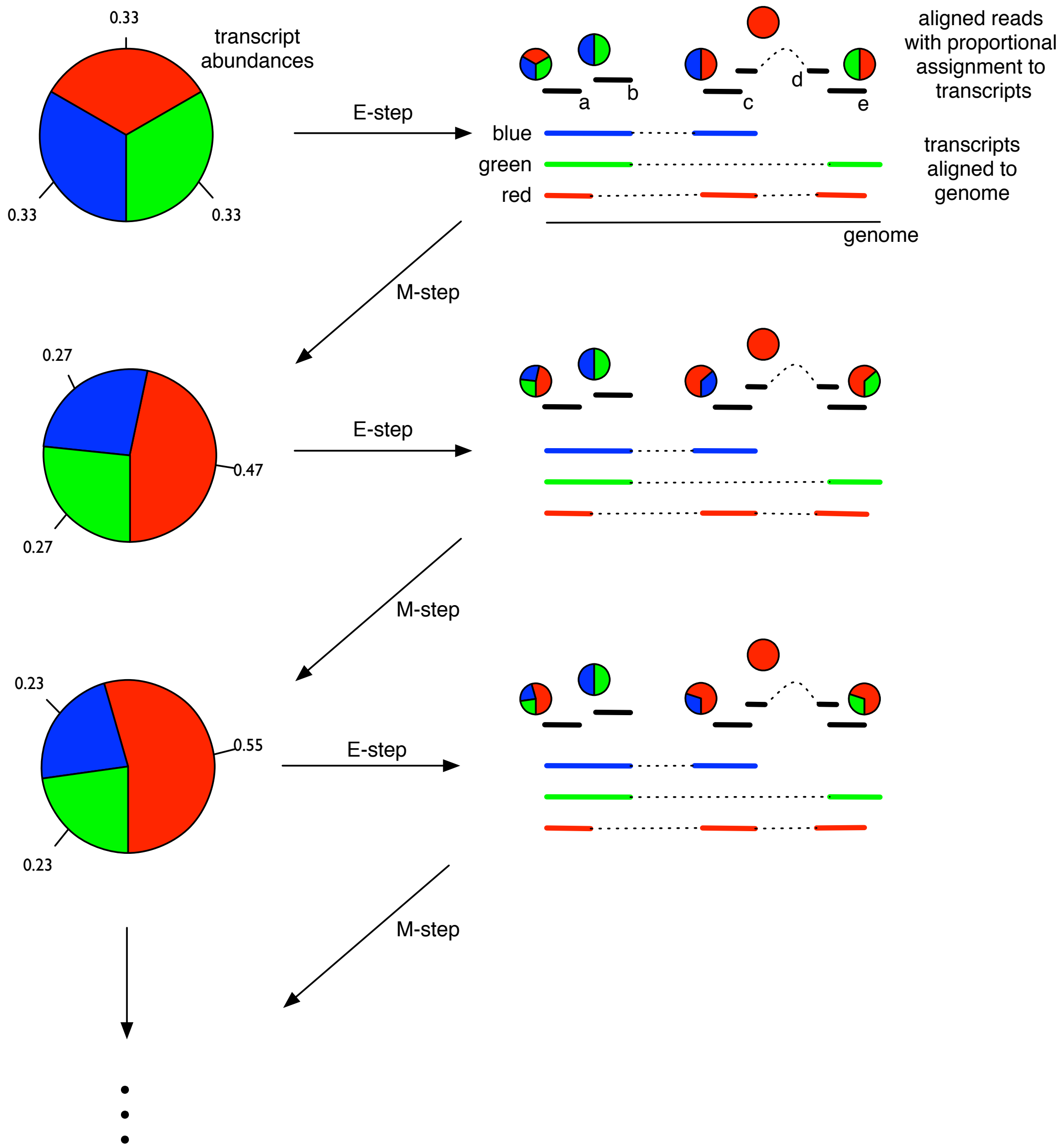
Exon *intersection* model: Score constituent exons

Exon *union* model: Score the the “merged” transcript

Transcript expression model: Assign reads uniquely to different isoforms. *Not a trivial problem!*

Garber et al. Nat. Methods 2011

Read assignment involves probabilistic assignment

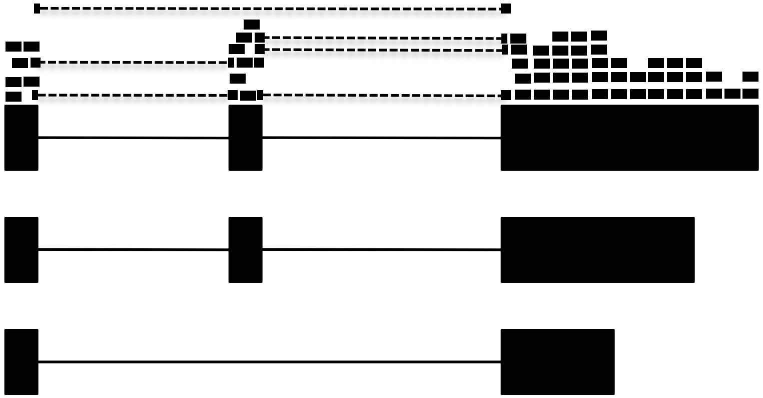


Pachter, L. (2011), (arXiv:1104.3889)

Current quantification models are complex

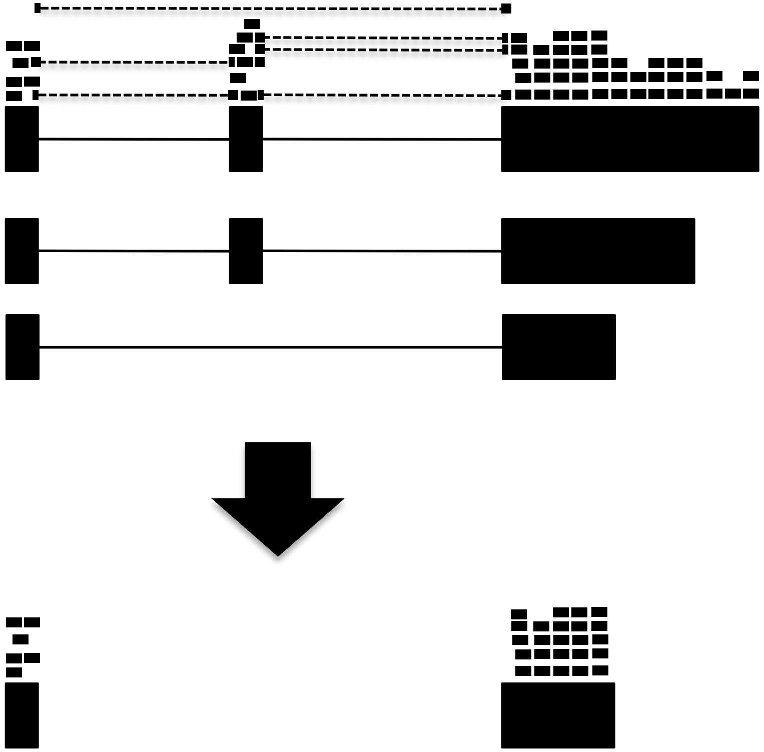
- In its simplest form we assume that reads can be unequivocally mapped. This allows:
  - Read counts distribute multinomial with rate estimated from the observed counts
- When this assumption breaks, multinomial is no longer appropriate.
- In general models use:
  - Fragments as inferred from paired-end data
  - Base quality scores
  - Sequence mapability
  - Protocol biases (e.g. 3' bias)
- Handling each of these involves a more complex model where reads are assigned probabilistically not only to an isoform but to a *different loci*

Why paired end matters for isoform quantification?



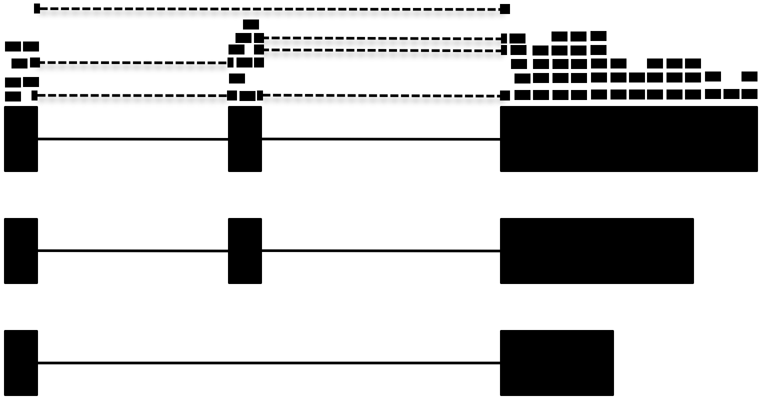
How do we define the gene expression?
How do we compute the expression of each isoform?

Computing gene expression

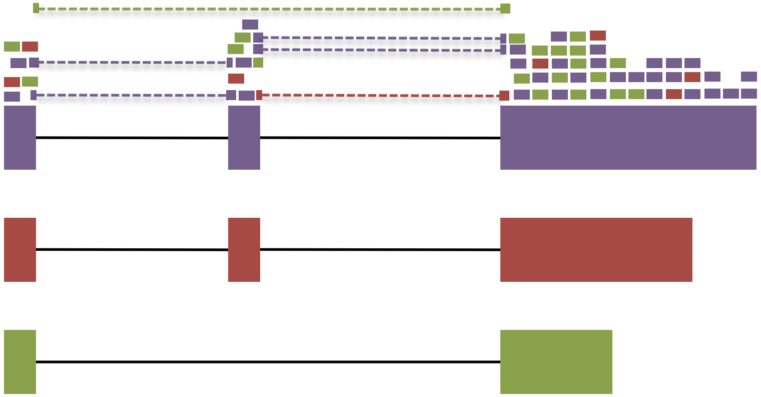


Idea1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)

Computing gene expression — isoform deconvolution



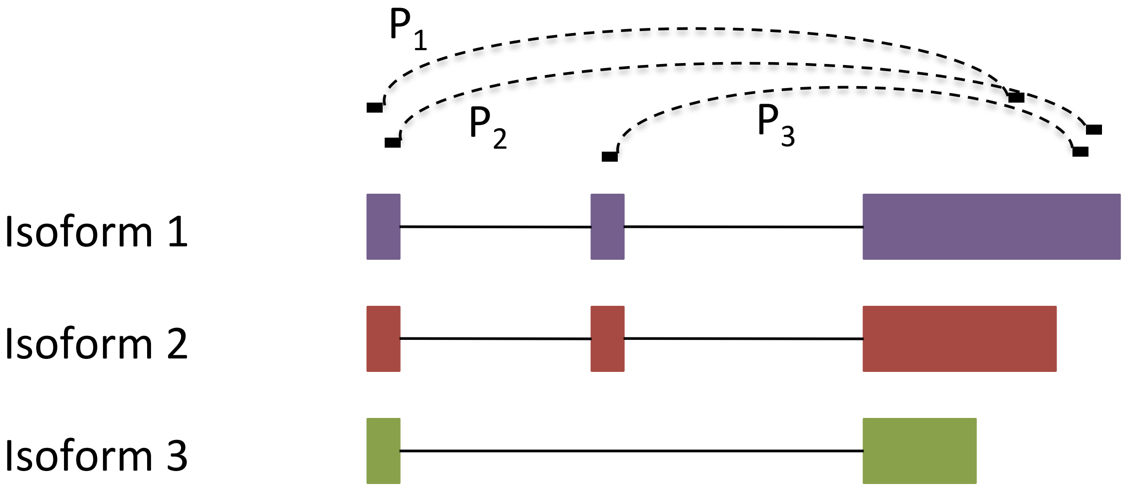
Computing gene expression — isoform deconvolution



If we knew the origin of the reads we could compute each isoform's expression.
The gene's expression would be the sum of the expression of all its isoforms.

$$E = \mathrm{RPKM}_1 + \mathrm{RPKM}_2 + \mathrm{RPKM}_3$$

Paired-end reads are easier to associate to isoforms

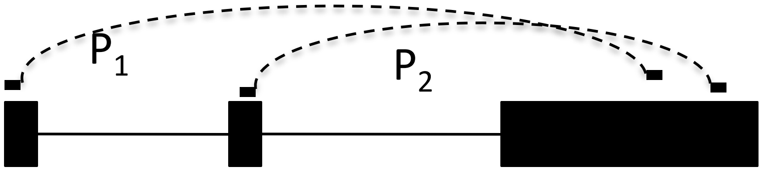


Paired ends increase isoform deconvolution confidence

- P_1 originates from isoform 1 or 2 but not 3.
- P_2 and P_3 originate from isoform 1

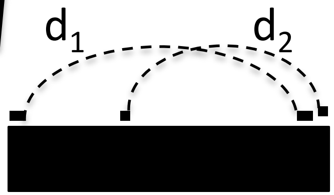
Do paired-end reads also help identifying reads originating in isoform 3?

We can estimate the insert size distribution

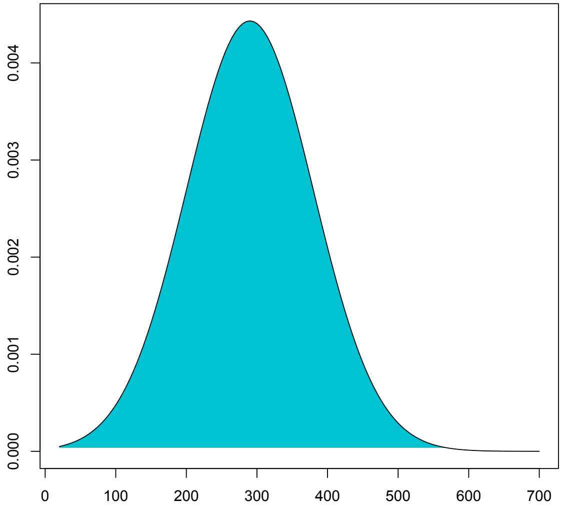


Get all single isoform reconstructions

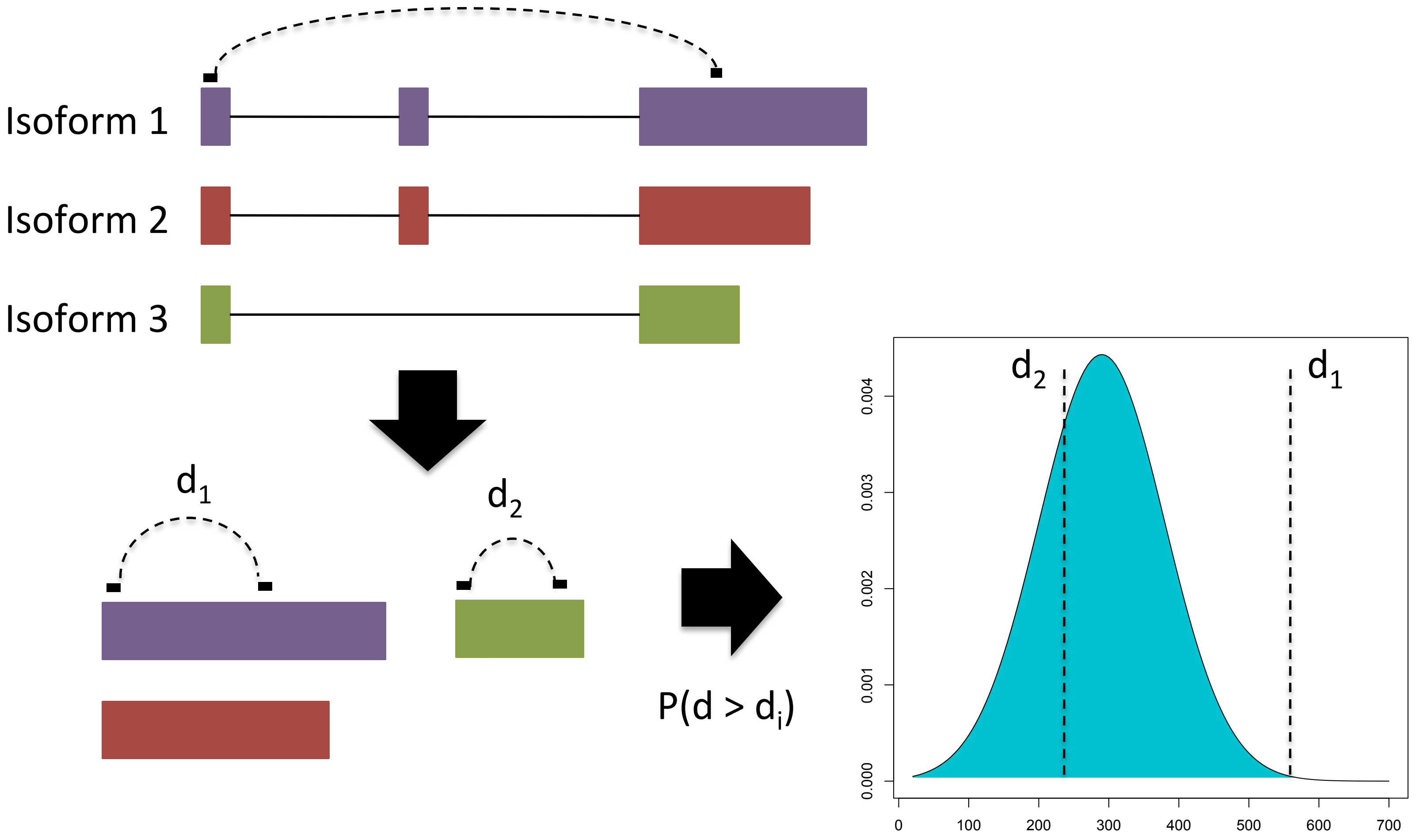
Splice and compute insert distance



Estimate insert size empirical distribution



… and use it for probabilistic read assignment



For methods such as MISO, Cufflinks and RSEM, it is critical to have paired-end data

Other considerations

- Duplicates – What to do with PCR artifacts
- Multimapper reads – What to do with reads that map to multiple places in the genome

RNA-Seq quantification summary

- Counts must be estimated from ambiguous read/ 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
- How do we normalize More sophisticated normalization to account for differences in library composition for inter-library comparisons.

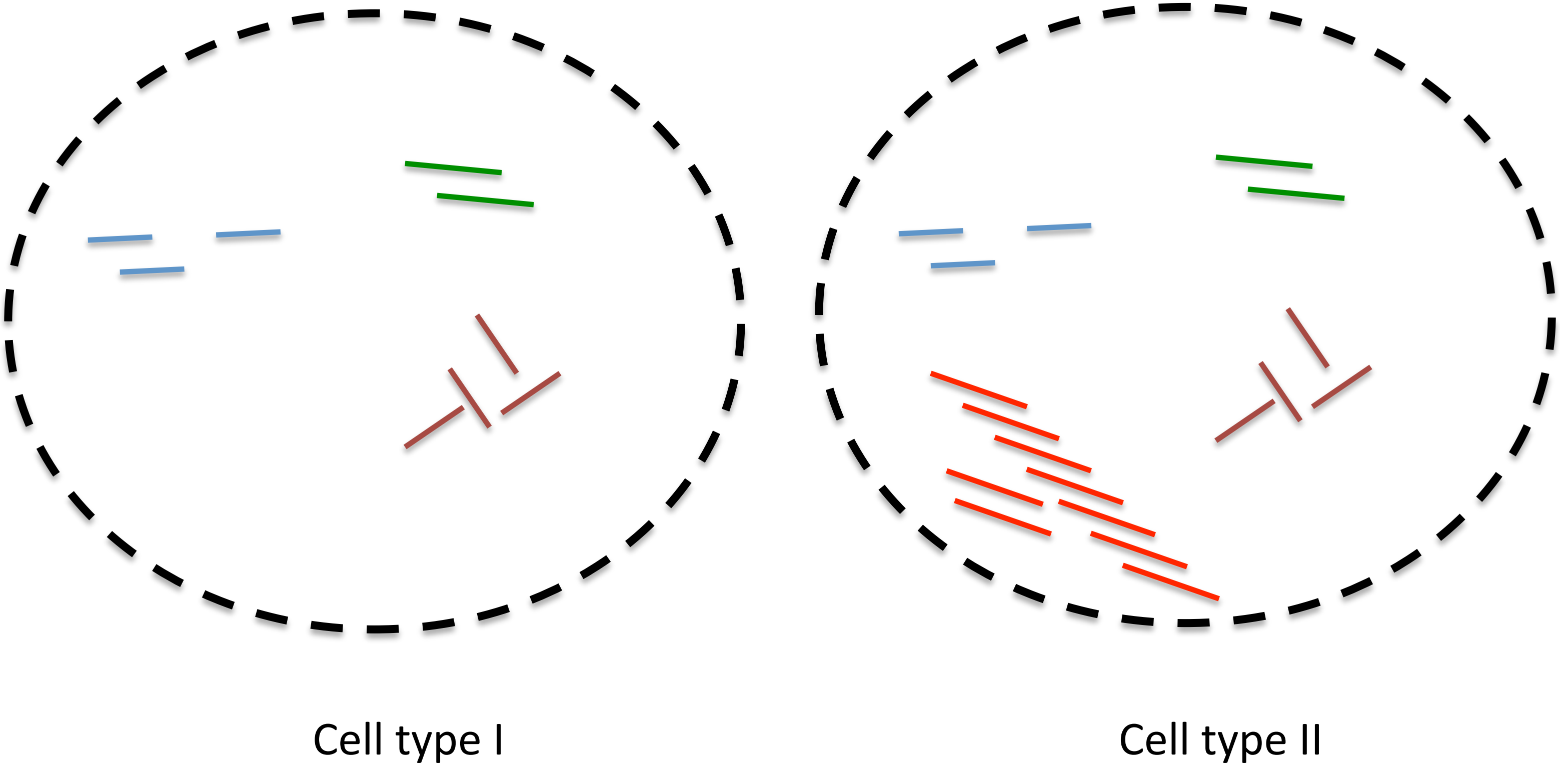
Programs to measure transcript expression

	Implemented method
Cufflinks2	Transcript deconvolution by solving the maximum likelihood problem
RSEM	Transcript deconvolution by solving the maximum likelihood problem
eXpress	Incorporated biases into model

Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

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



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

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

Geometric mean for that gene over ALL experiments

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

$>\text{norm}=\text{sum}(s2)/\text{sum}(s1)$

$>\text{plot}(s2, s1{*}\text{norm},\text{log}=\text{"xy"})$

$>\text{abline}(a = 0, b = 1)$

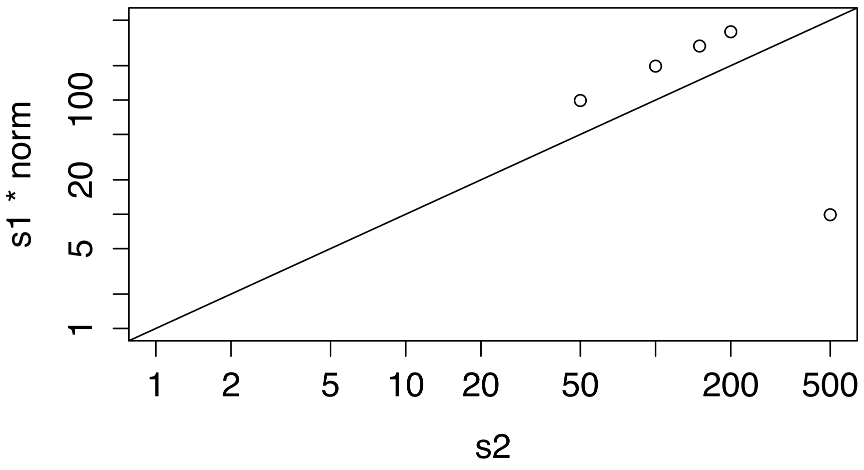
Size normalization results in 2-fold changes in *all* transcripts

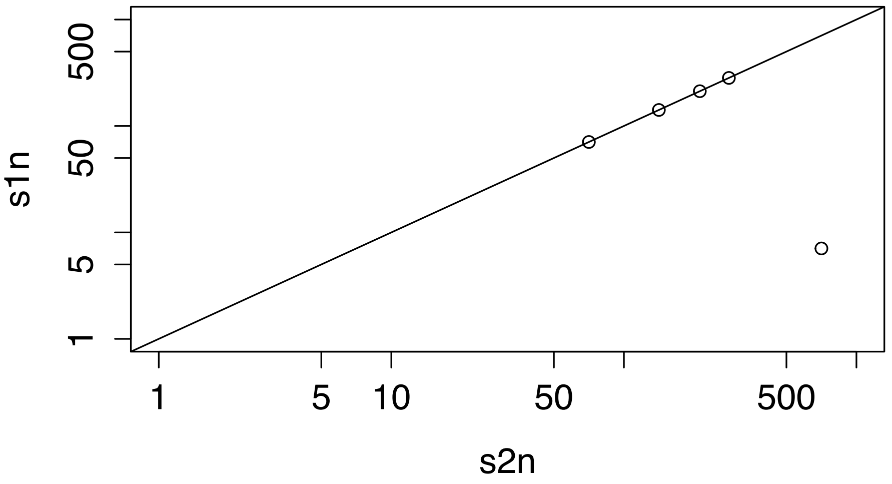
$>\text{g} = \text{sqrt}(s1 * s2t)$

$>\text{s1n} = s1/\text{median}(s1/g);\;\; \text{s2n} = s2/\text{median}(s2/g)$

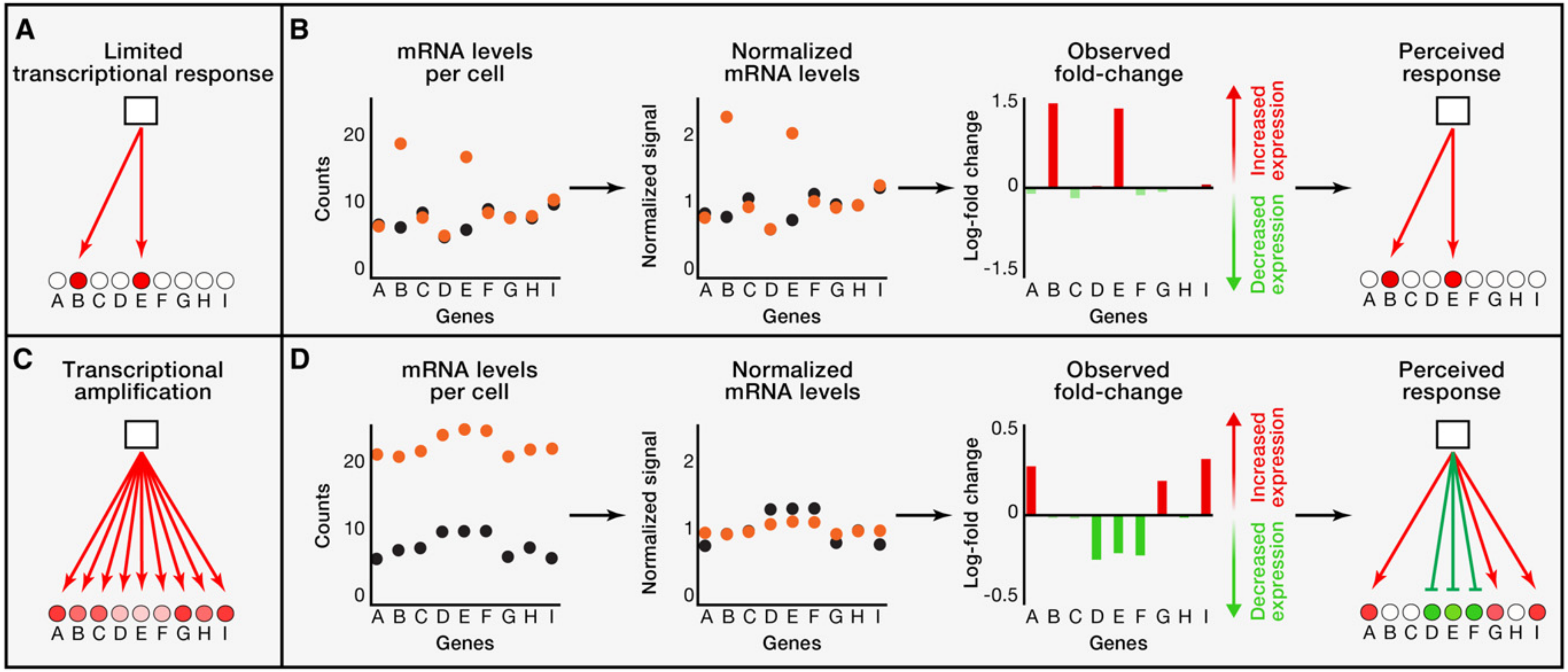
$>\text{plot}(s2n, s1n,\text{log}=\text{"xy"})$

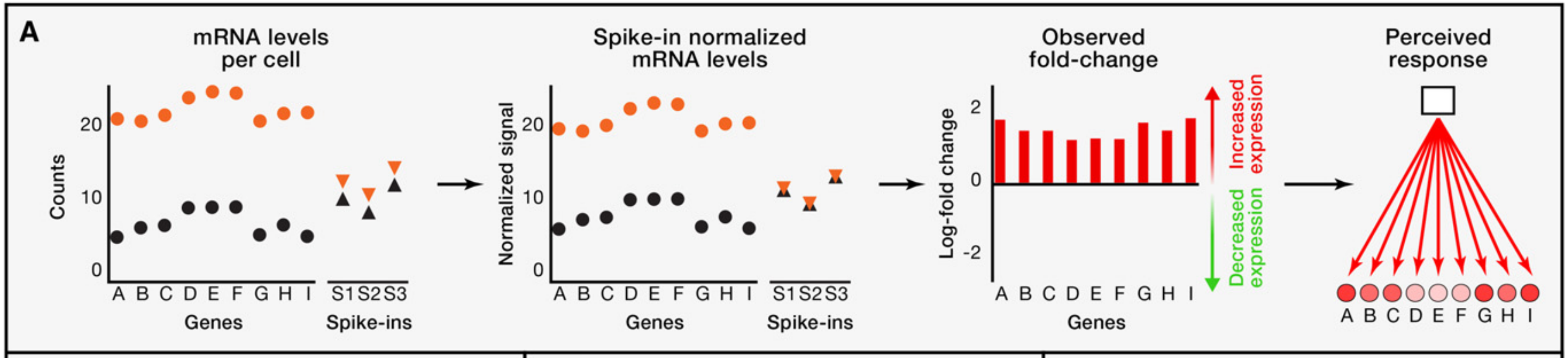
$>\text{abline}(a = 0, b = 1)$





When everything changes: Spike-ins





Lovén et al, Cell 2012

Analysis of counting data requires 3 broad tasks

- 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.
- Reconstruction: Finding the regions that originated the reads

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
  - Use model to estimate variance for a transcript given its mean count
- Define a test
  - DESeq: Generalization of a fisher exact test
  - Cufflinks: Log transformed of counts divided by its variance (~ normally distribute).
    - Null hypothesis: log ratio = 0

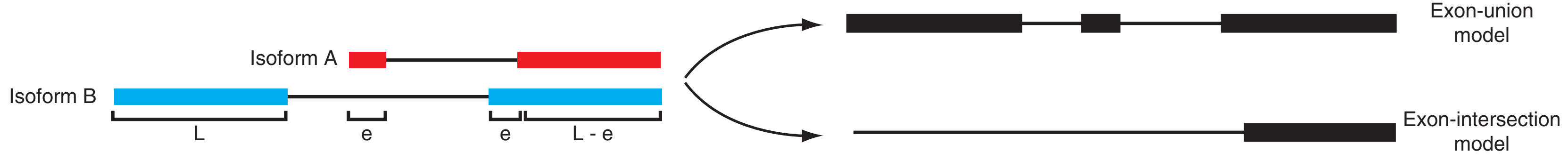
Differential analysis strategies

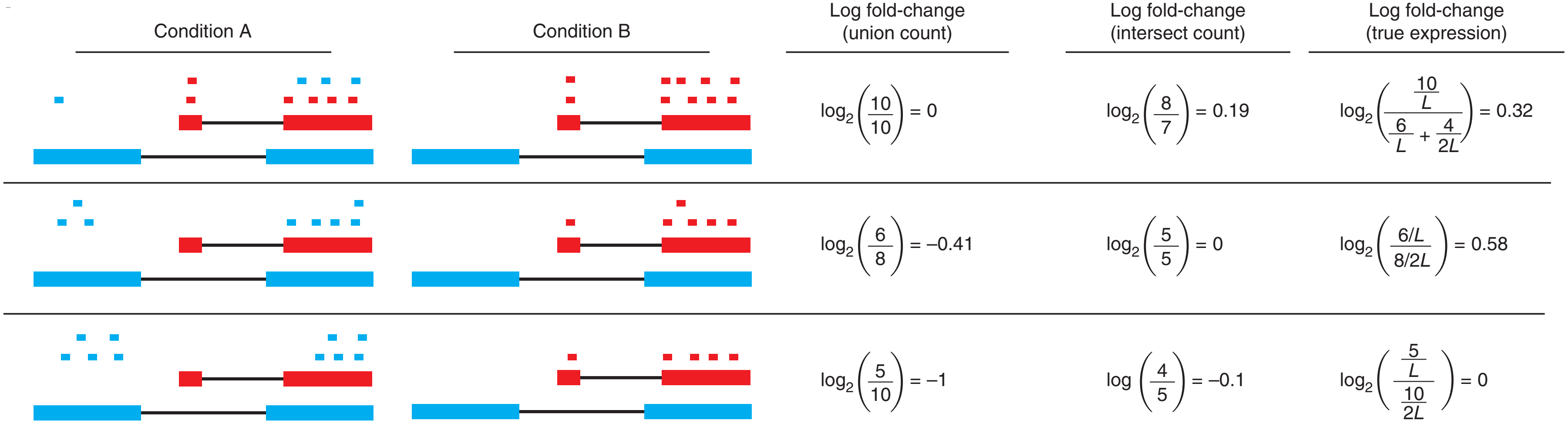
- Use read counts and Standard Fisher exact test

	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

Why not just simple models?





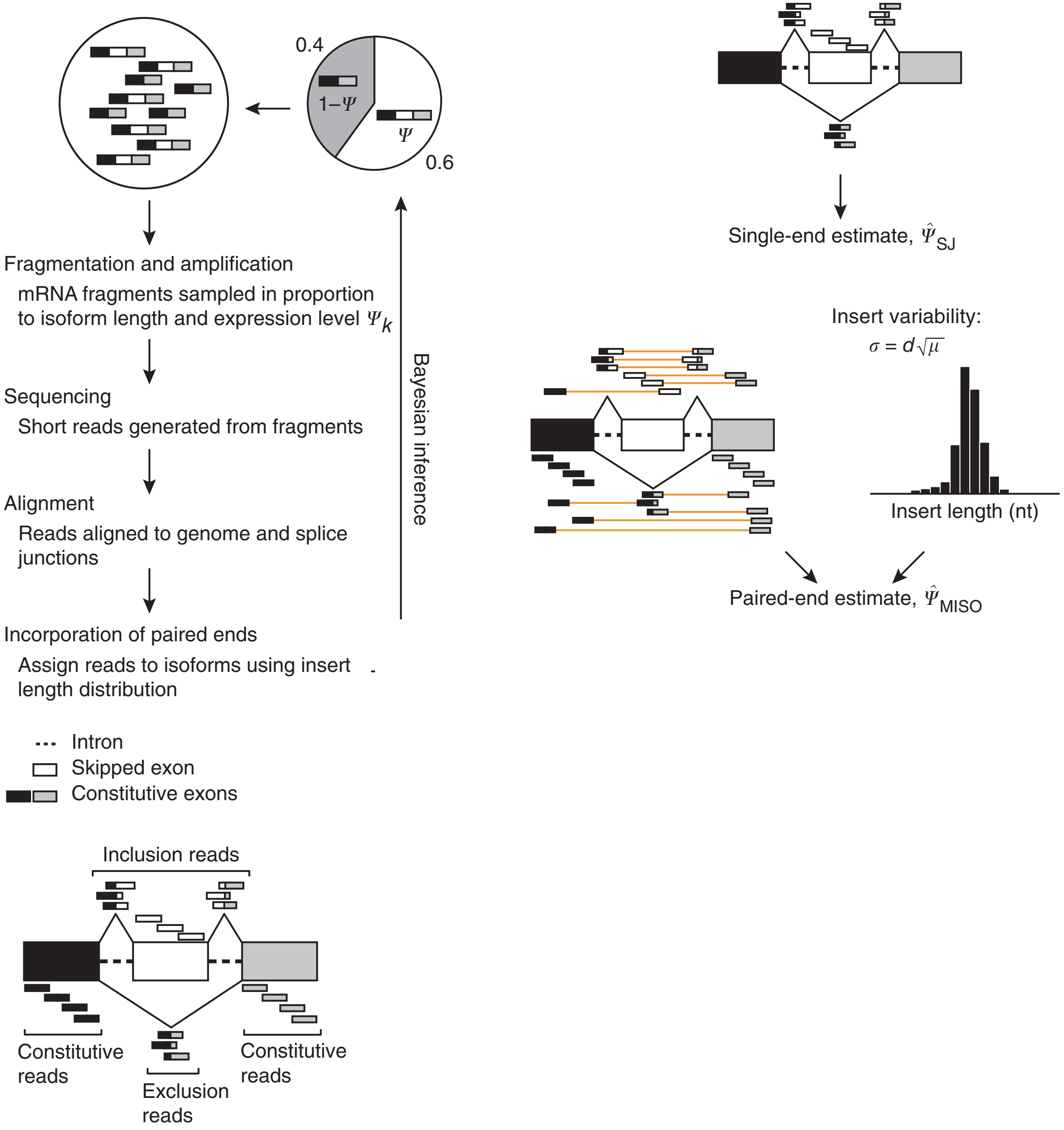
Condition A	Condition B	Log fold-change (union count)	Log fold-change (intersect count)	Log fold-change (true expression)
		$\log_2\left(\frac{10}{10}\right) = 0$	$\log_2\left(\frac{8}{7}\right) = 0.19$	$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L} + \frac{4}{2L}}\right) = 0.32$
		$\log_2\left(\frac{6}{8}\right)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

Trapnell et al. Nat. Biotech 2012

RNA-Seq differential expression software

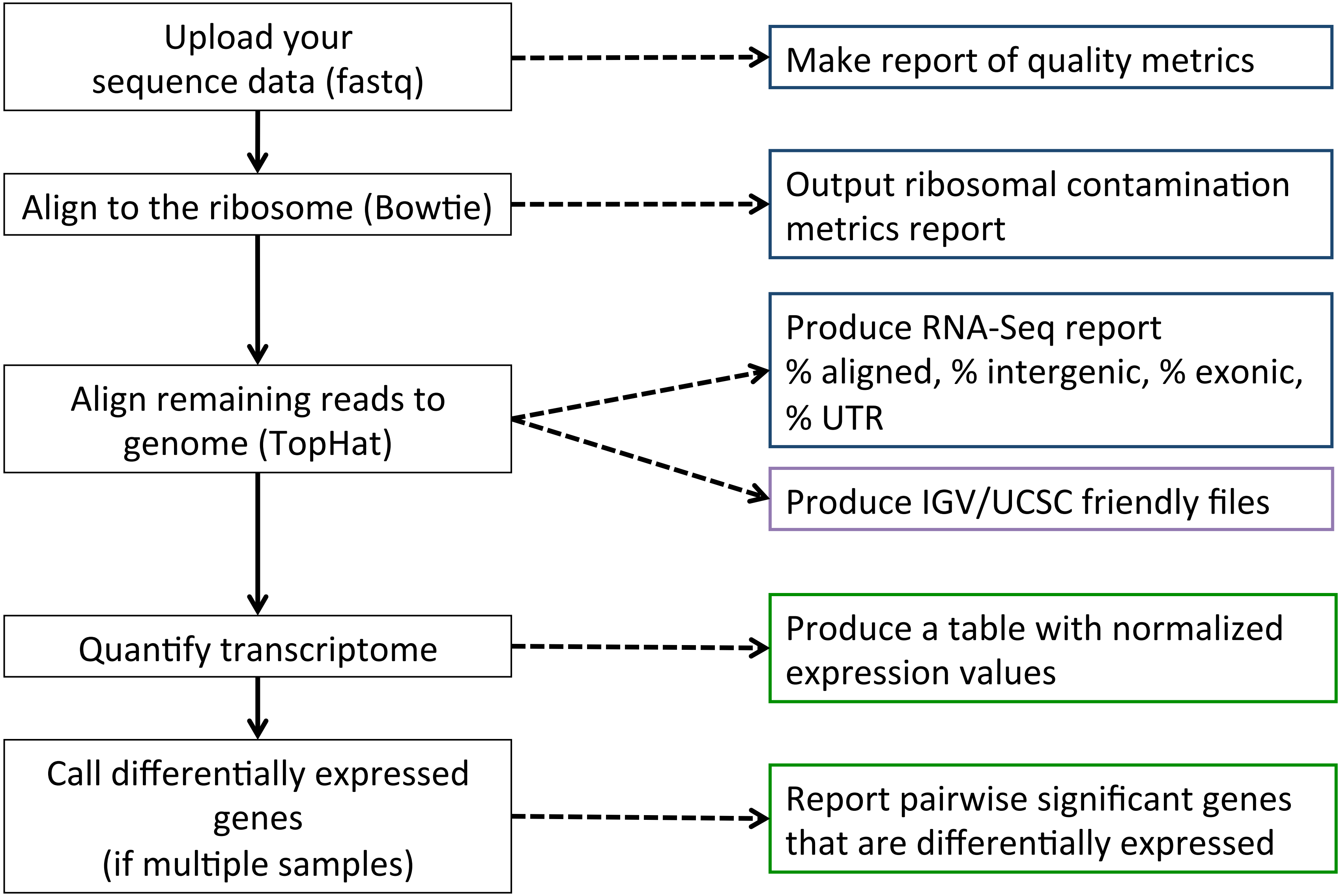
	Underlying model	Notes
EdgeR	Negative Bionomial	Gene read counts table
DESeq2	Negative Bionomial	Gene read counts table
Cufflinks2	~~Poisson~~ Negative Bionomial	Works directly from the alignments
Myrna	Empirical	Sequence reads and reference transcriptome
Miso	Multinomial	Specifically to test exon cassette inclusion/ exclusion.

MISO: Specifically testing exon inclusion


0.4
1−Ψ
Ψ
0.6
Fragmentation and amplification
mRNA fragments sampled in proportion to isoform length and expression level Ψ_k
Sequencing
Short reads generated from fragments
Alignment
Reads aligned to genome and splice junctions
Incorporation of paired ends
Assign reads to isoforms using insert length distribution
Bayesian inference
Intron
Skipped exon
Constitutive exons
Inclusion reads
Constitutive reads
Exclusion reads
Constitutive reads
Single-end estimate, $\hat{\Psi}_{SJ}$
Insert variability:
$\sigma = d\sqrt{\mu}$
Insert length (nt)
Paired-end estimate, $\hat{\Psi}_{MISO}$

Katz et al Nat. Methods 2010

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



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat)

Produce RNA-Seq report
% aligned, % intergenic, % exonic, % UTR

Produce IGV/UCSC friendly files

Quantify transcriptome

Produce a table with normalized expression values

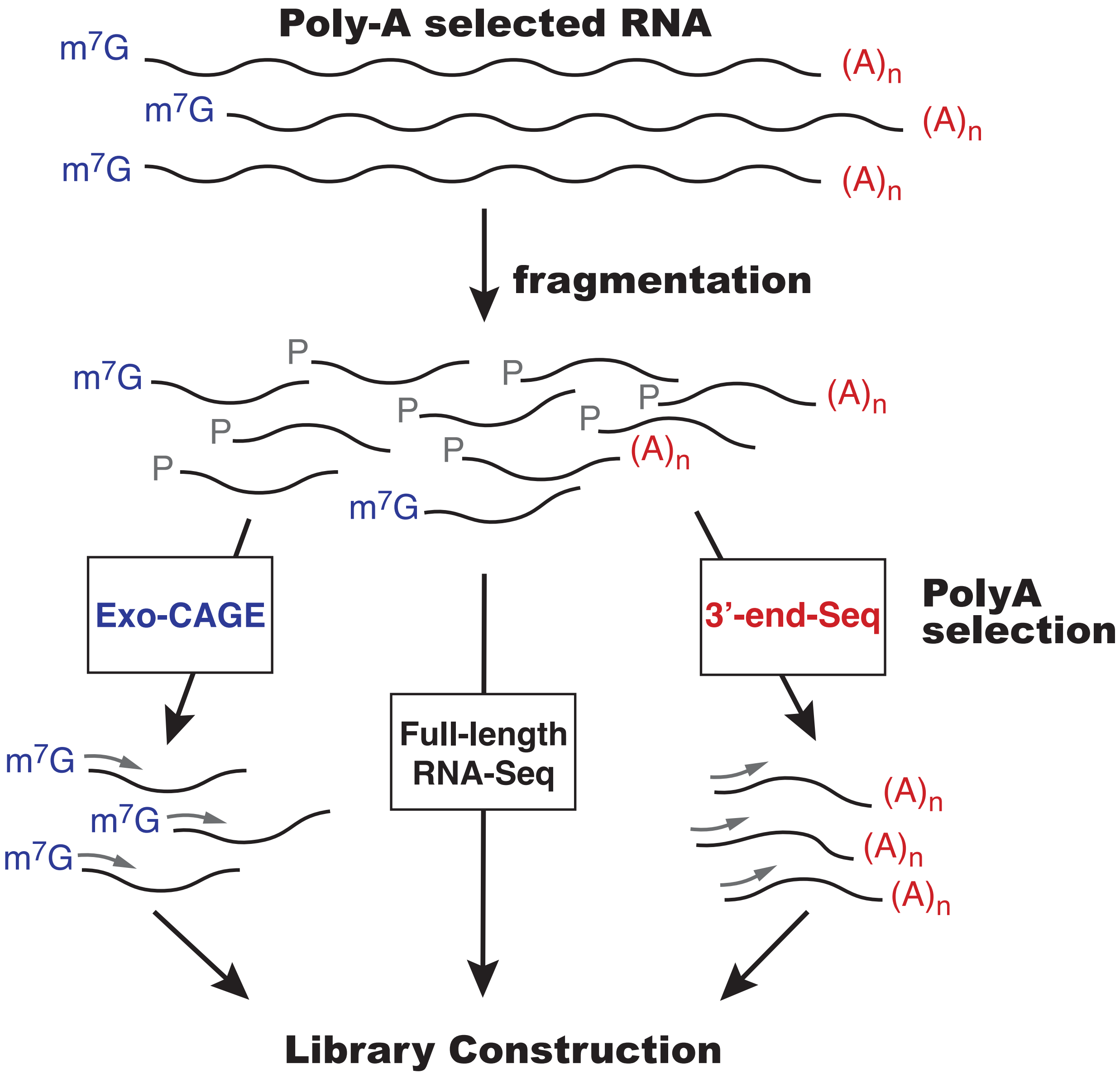
Call differentially expressed genes
(if multiple samples)

Report pairwise significant genes that are differentially expressed

The quest for inexpensive expression assays

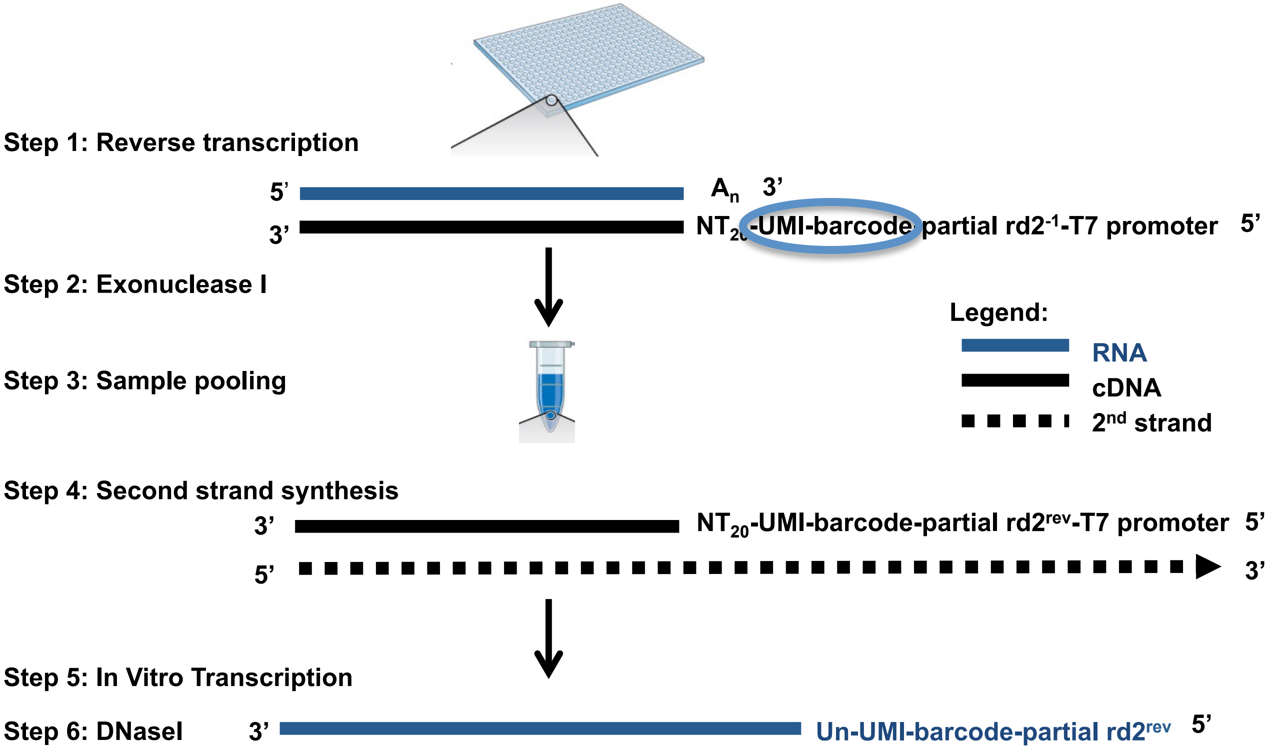
- Goal: Routinely profile hundreds of samples
- Why?
  - Human variability in health and disease
  - Perturbation studies
  - Clinical applications of expression profiling
  - Single cell sequencing
- Current costs
  - Afffy ~$300-$400/sample
  - Illumina bead arrays $150/sample
  - RNA-Seq (20 mill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



Step 1: Reverse transcription

5' A$_n$ 3'

3' NT$_{20}$-UMI-barcode-partial rd2^{-1}-T7 promoter 5'

Step 2: Exonuclease I

Step 3: Sample pooling

Legend:

- ▬ RNA
- ▬ cDNA
- ▪▪▪▪ 2nd strand

Step 4: Second strand synthesis

3' NT$_{20}$-UMI-barcode-partial rd2^{rev}-T7 promoter 5'

5' 3'

Step 5: In Vitro Transcription

Step 6: DNaseI 3' Un-UMI-barcode-partial rd2^{rev} 5'

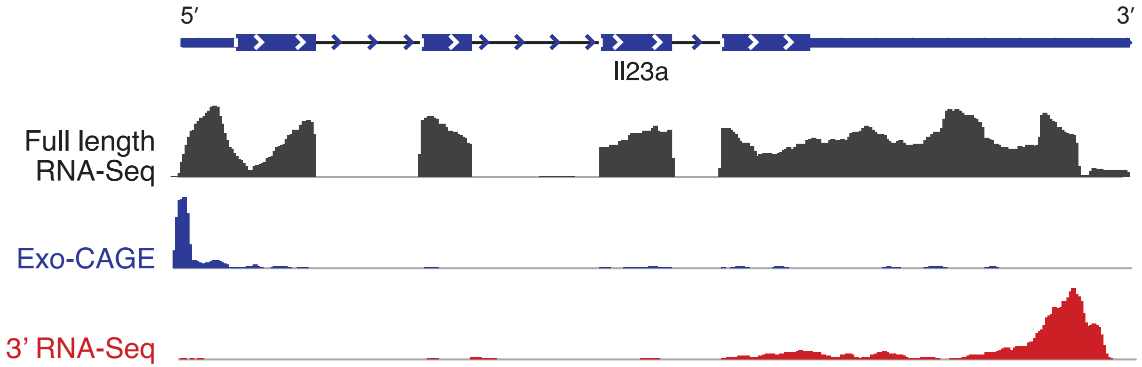
NT$_{20}$XXXXXXXXX-SSSSSS-adapter

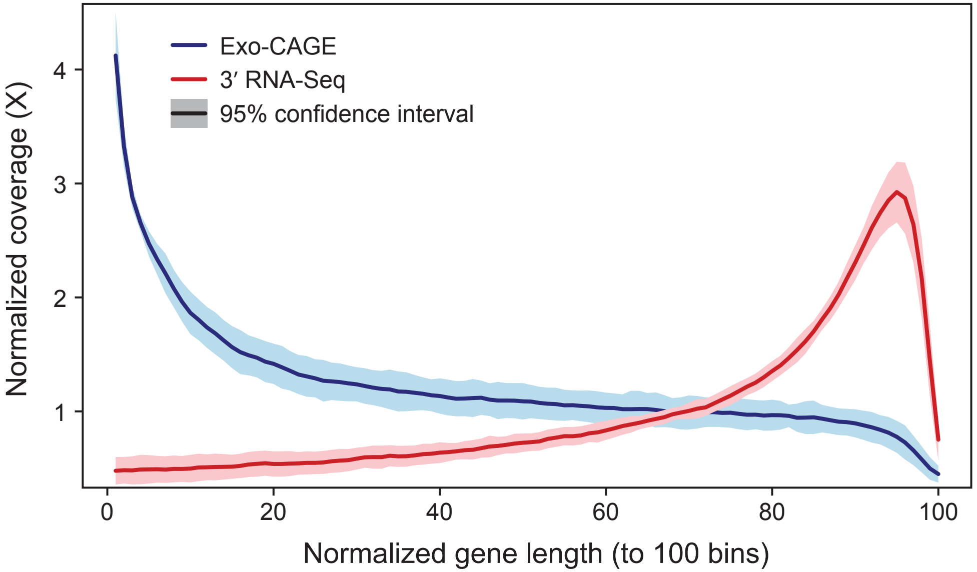
XXXXXXXXX: UMI

SSSSSS: Sample Barcode

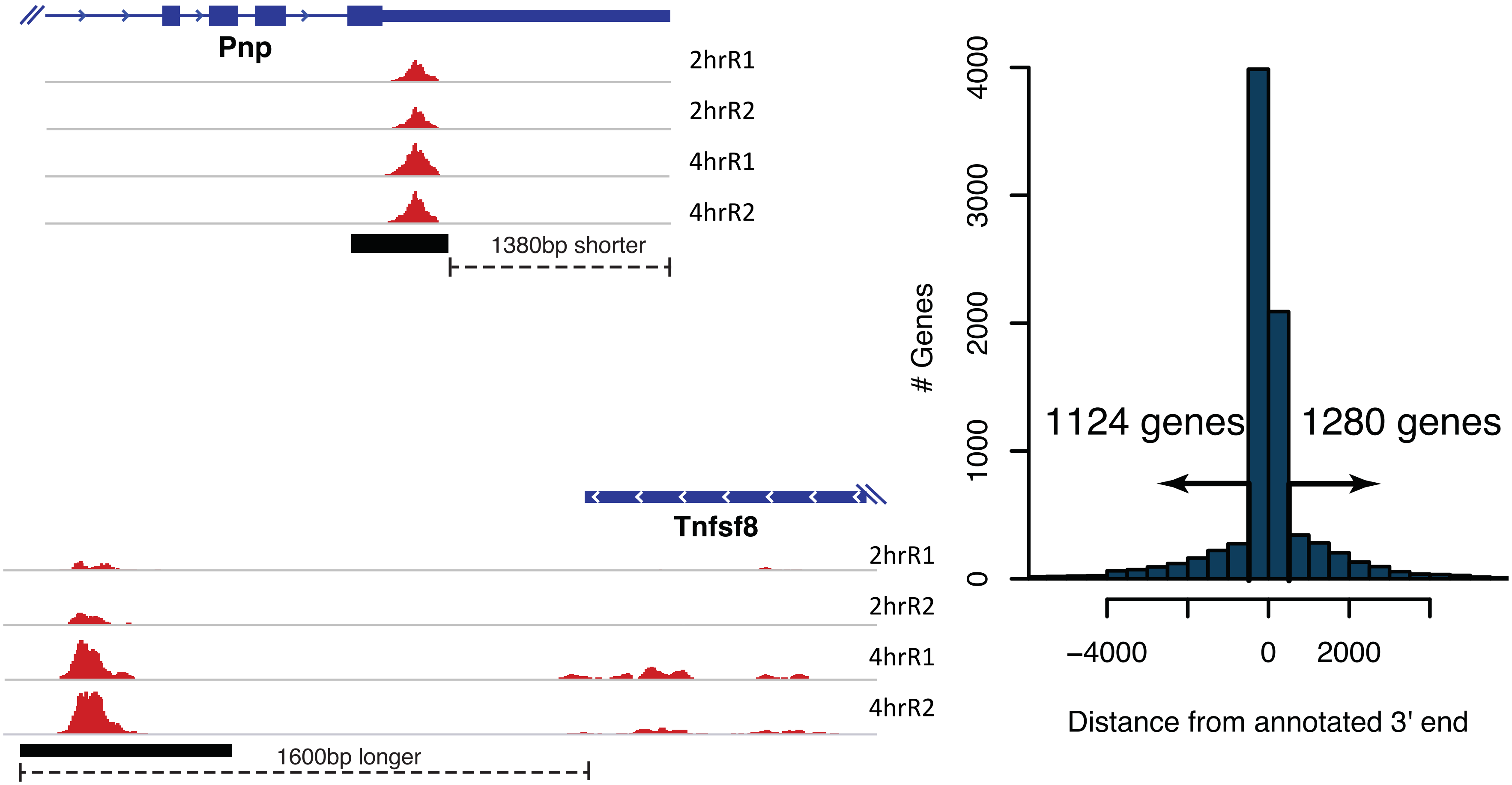
Jaitin et al. Science 2014

End-sequencing solution

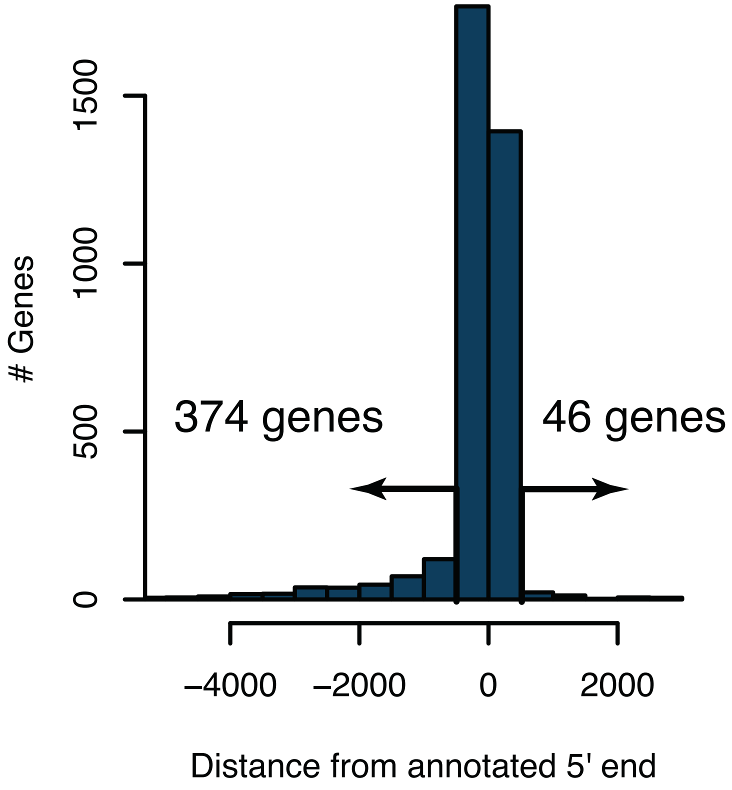


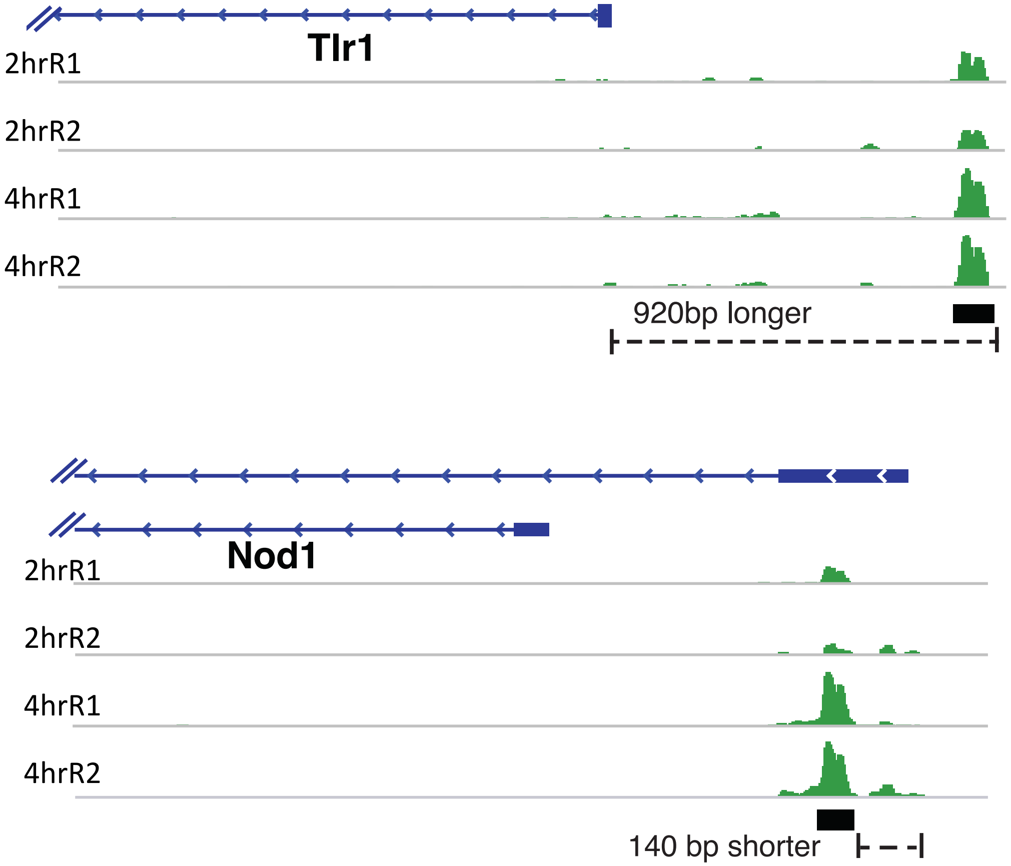


Although annotated ends far from perfect

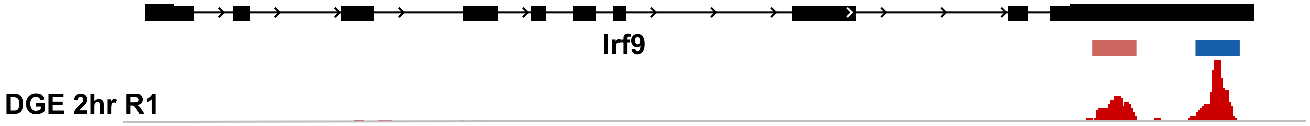


While annotated starts are much more conserved





We take full advantage of the data



1. Slide a window and identify major 3’ end

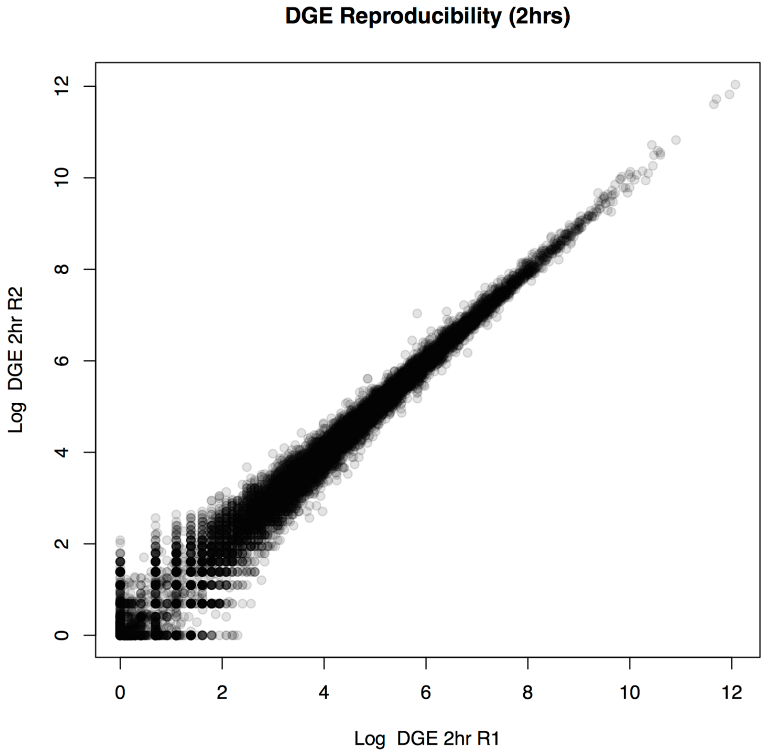
2. Identify all other significant windows (using a local background)

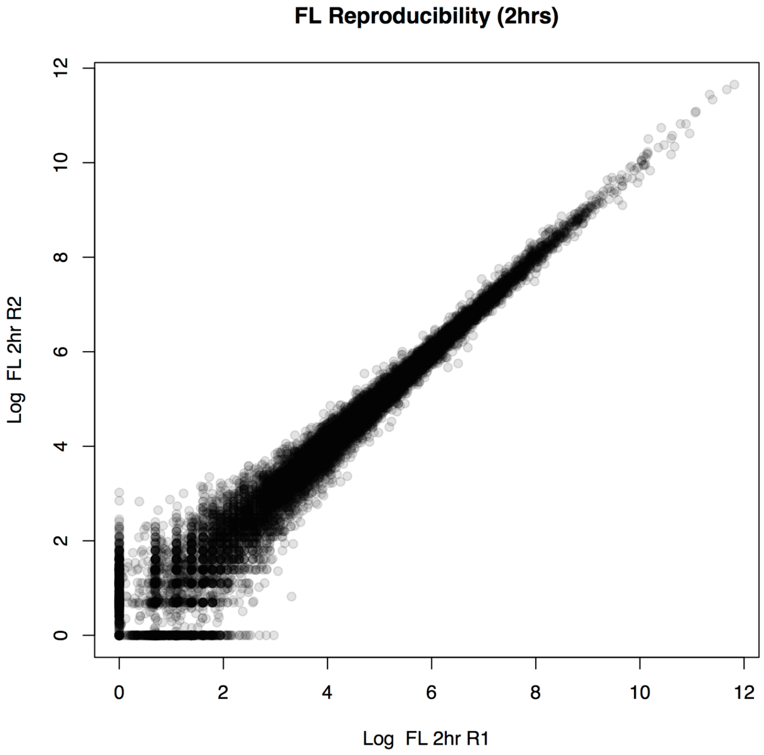
3. Repeat for each sample

4. Take all significant windows across samples

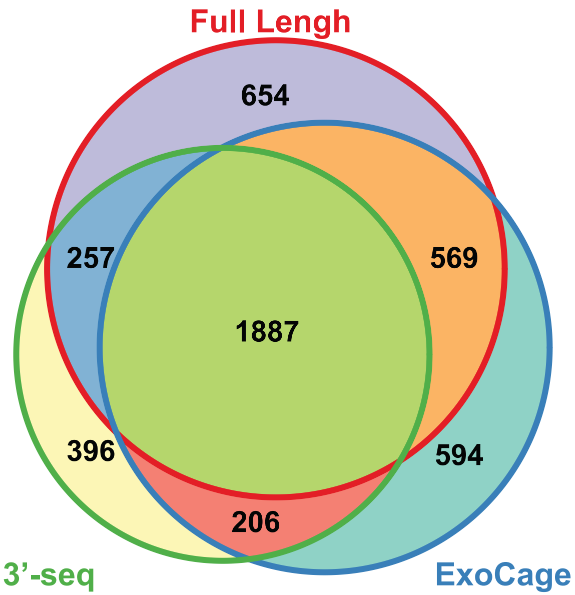
5.1 Report gene level counts: Sum across all sig. windows
5.2 Report isoform level counts: Each sig. window

Reproducibility is as good as with full length


DGE Reproducibility (2hrs)


FL Reproducibility (2hrs)

With 8.5 Million reads similar yet somewhat reduced power

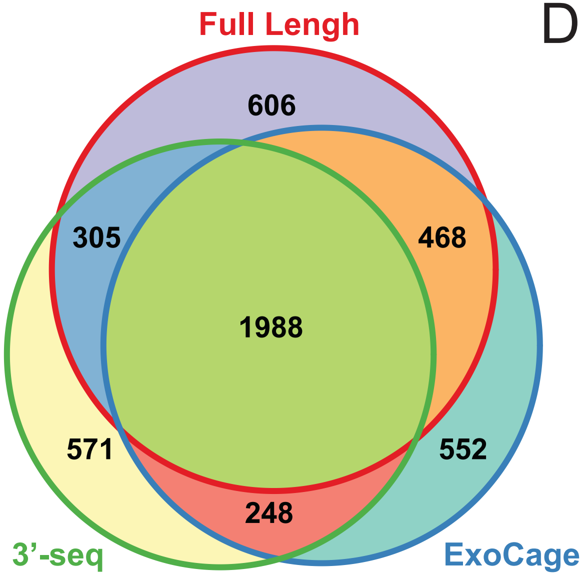


+

UTR correction

&

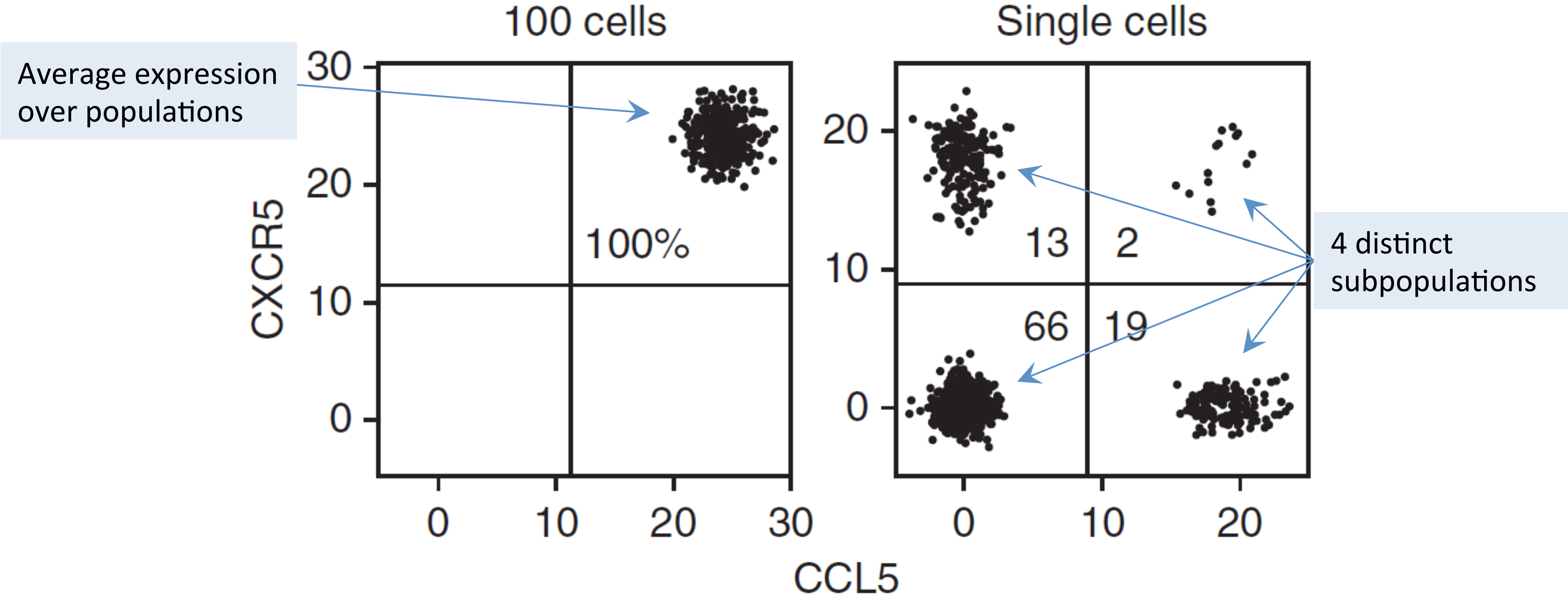
multimapper handling



Having established a robust analysis pipeline => Single cell RNA-Seq

Why Single-cell analysis?

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



Average expression over populations

4 distinct subpopulations

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

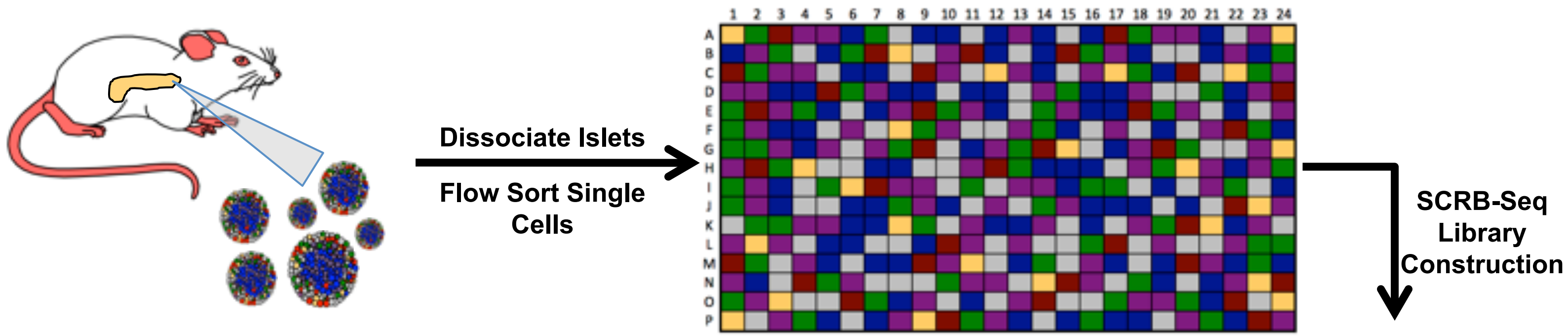
Type 1 Diabetes study

- It is unclear what triggers T1D
- The mechanism(s) of β-cell death are not well understood.
- Rat model with inducible T1D within 10 (± 1) days.
- Bulk RNA-Seq can’t reveal tissue composition

Cell sorting

- Pancreatic islets are composed of:
  - α-cells: primarily produce glucagon
  - β-cells: primarily produce insulin
  - δ-cells, PPY producing cells, and others
- Issues with sorting cells by FACS:
  - Only *known* cell types can be selected
  - Preprocessing may affect the observed cell state
  - Islet cells are very difficult to isolate, and FACS discards "other" cells in the sorting process (wasteful for rare cells)
- In addition, "bulk" RNA-Seq can mask underlying heterogeneity of even a sorted cell population…

Islet single cell sequencing



Dissociate Islets

Flow Sort Single Cells

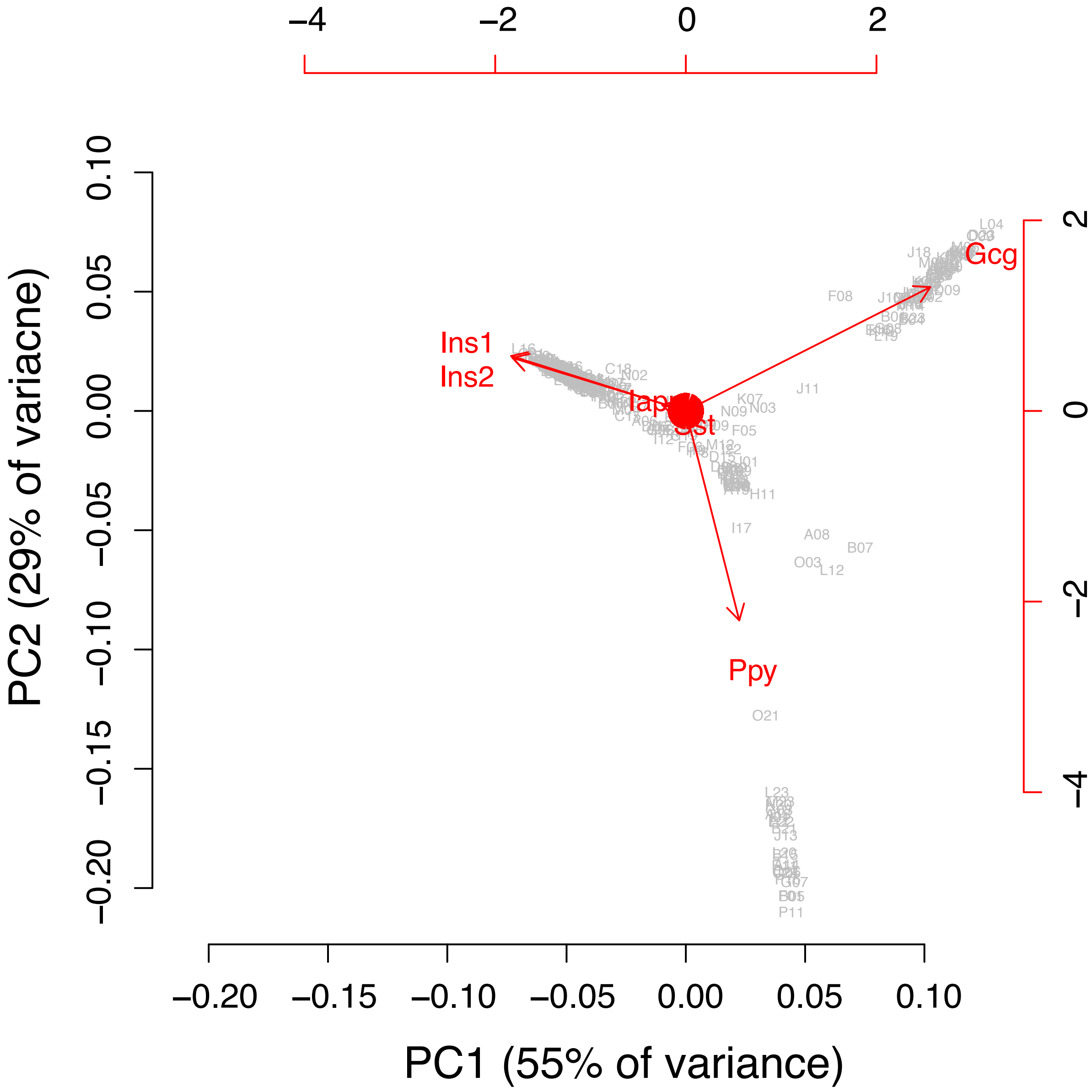
SCRB-Seq Library Construction

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

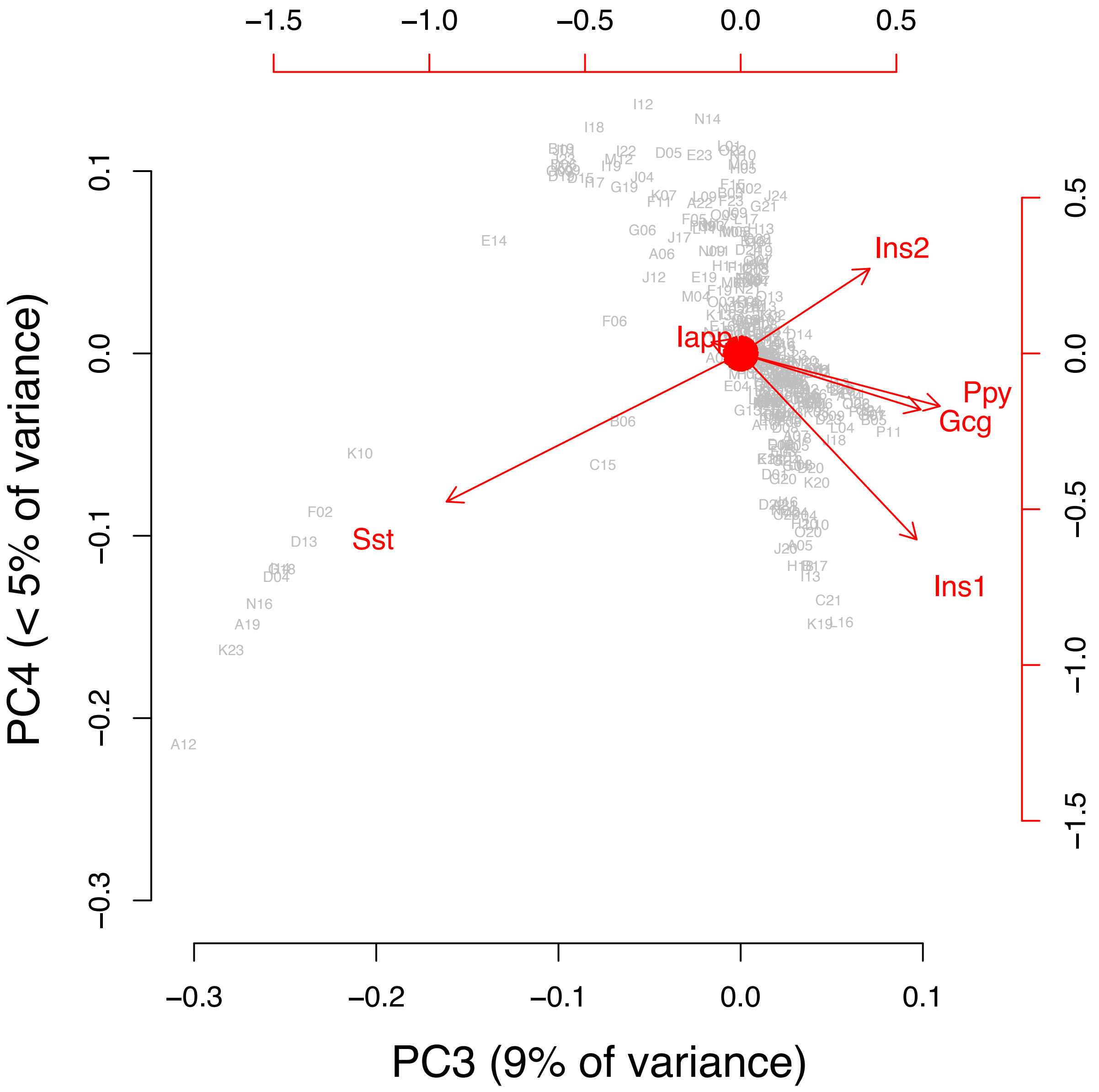
After filtering:

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

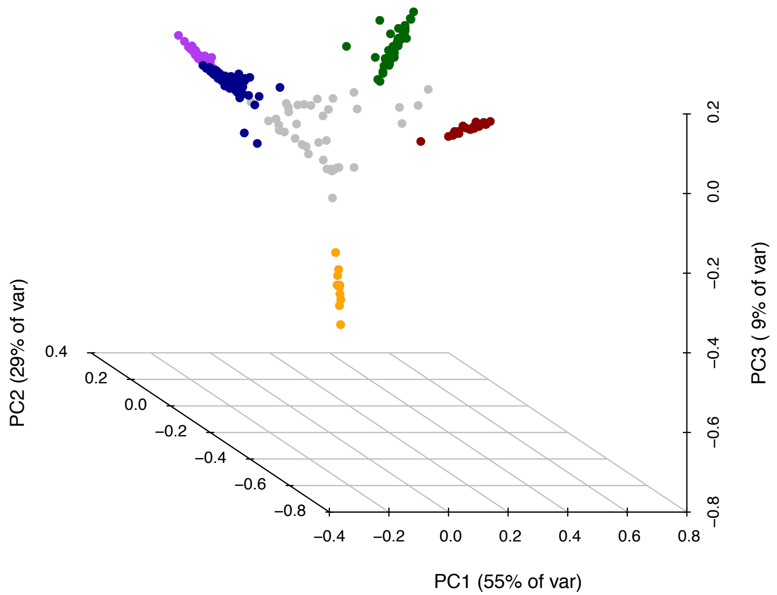
Single cell RNA-Seq cell sorting

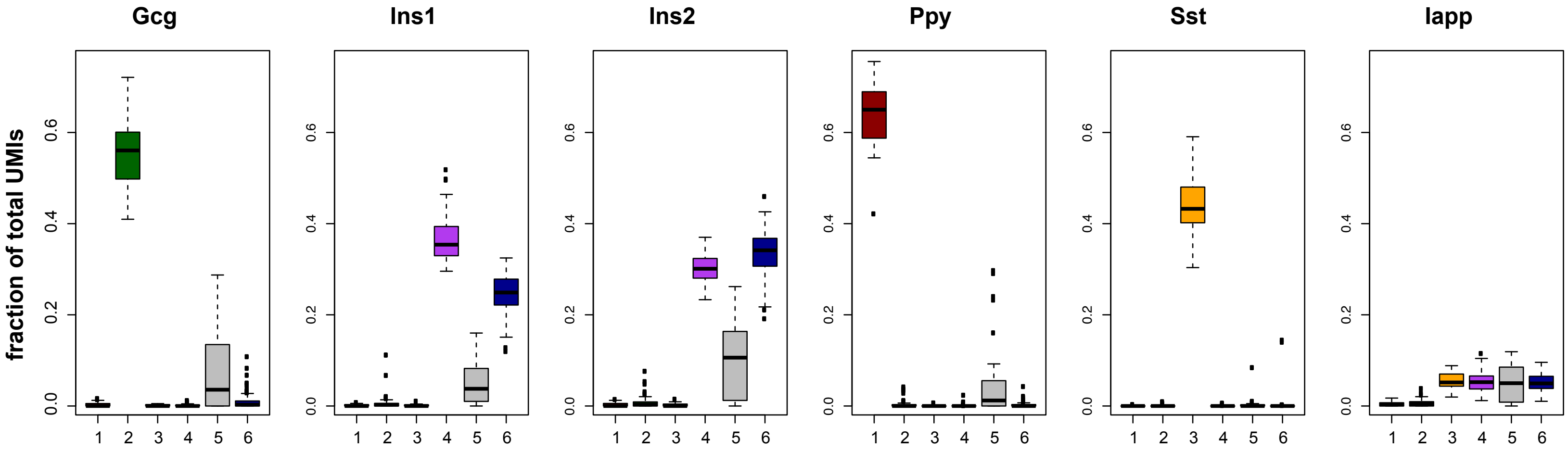


Single cell RNA-Seq cell sorting



Which allow us to recover the known islet composition





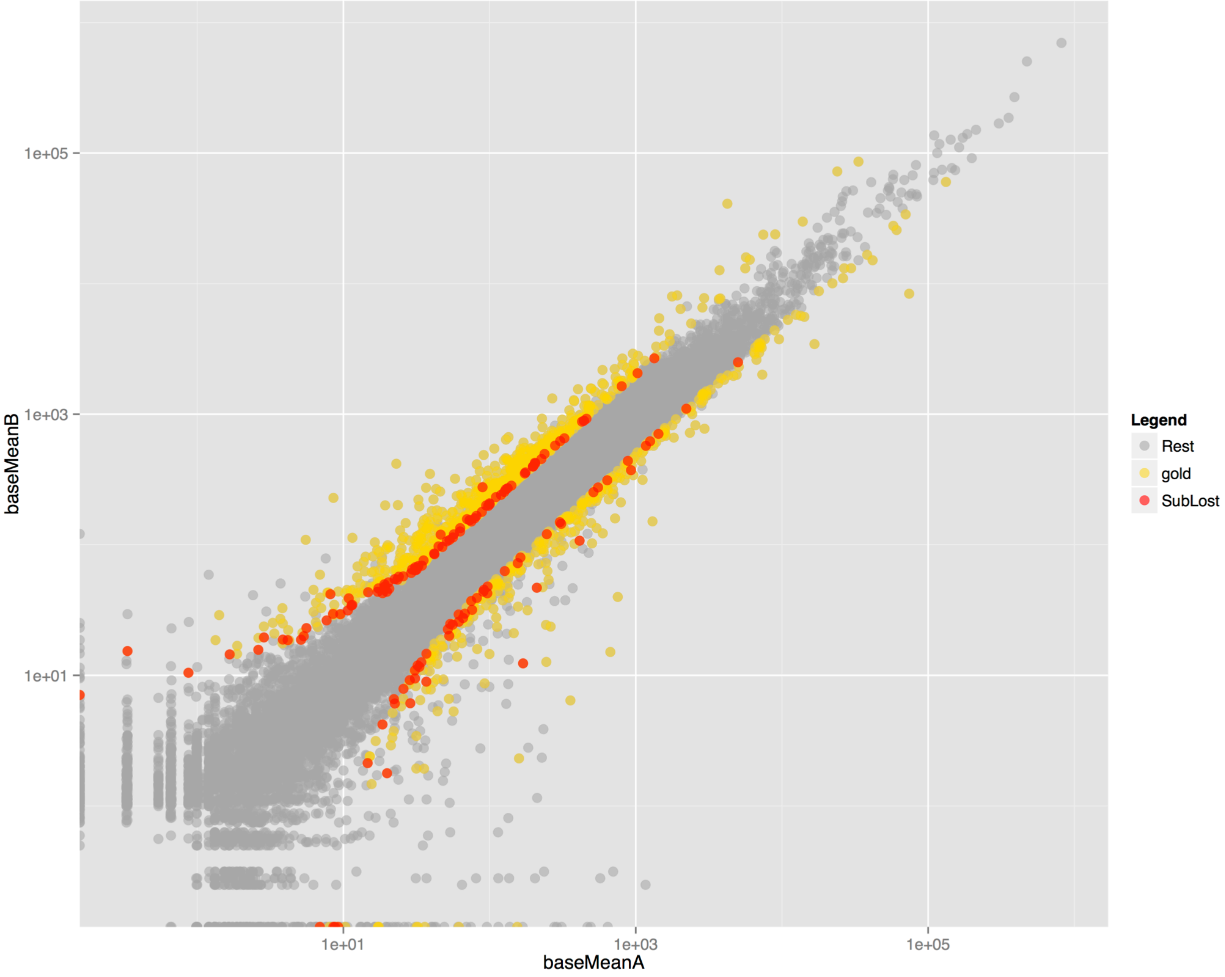
More in depth exploration of depth

- Very deep (30 million reads) dataset with triplicates.
  - Mouse WT vs double Jnk1/2 KO (Roger Davis)
  - Worm diet changes (Marian Walhout)
- Call DE with full dataset, then *in-silico* downsample data

Alper Kucukural

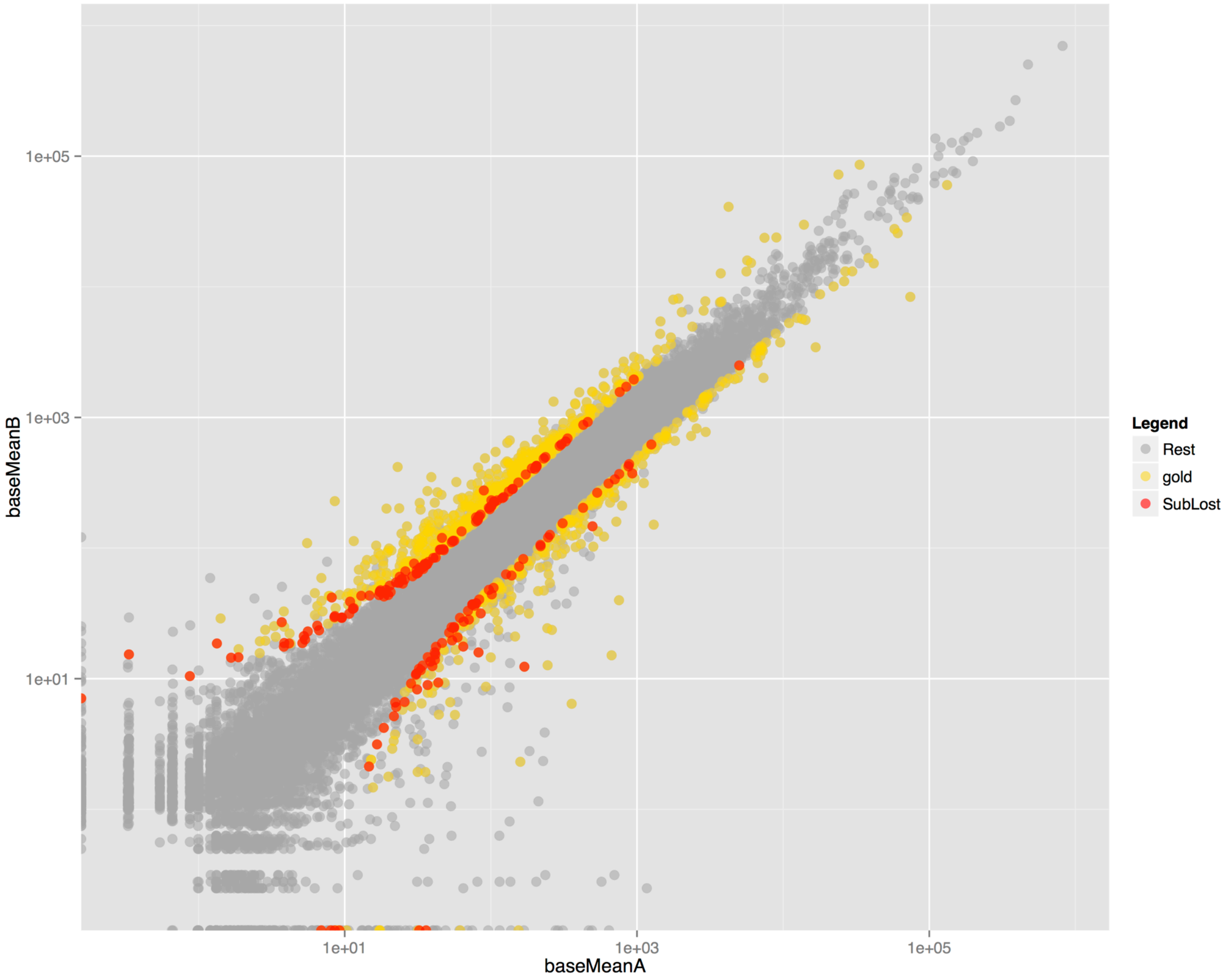
Is the loss qualitatively significant?

15 Million reads



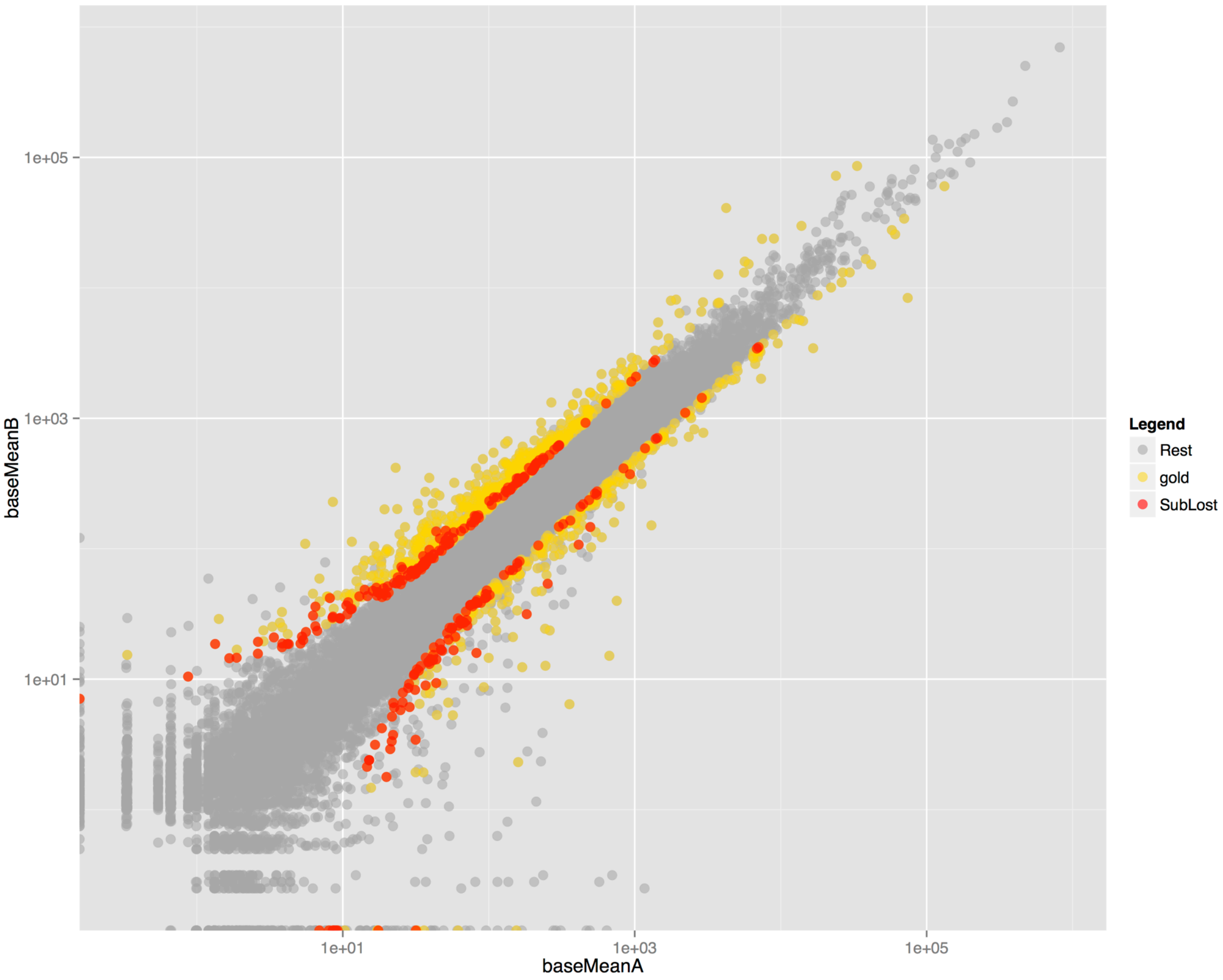
Is the loss qualitatively significant?

12.5 Million reads



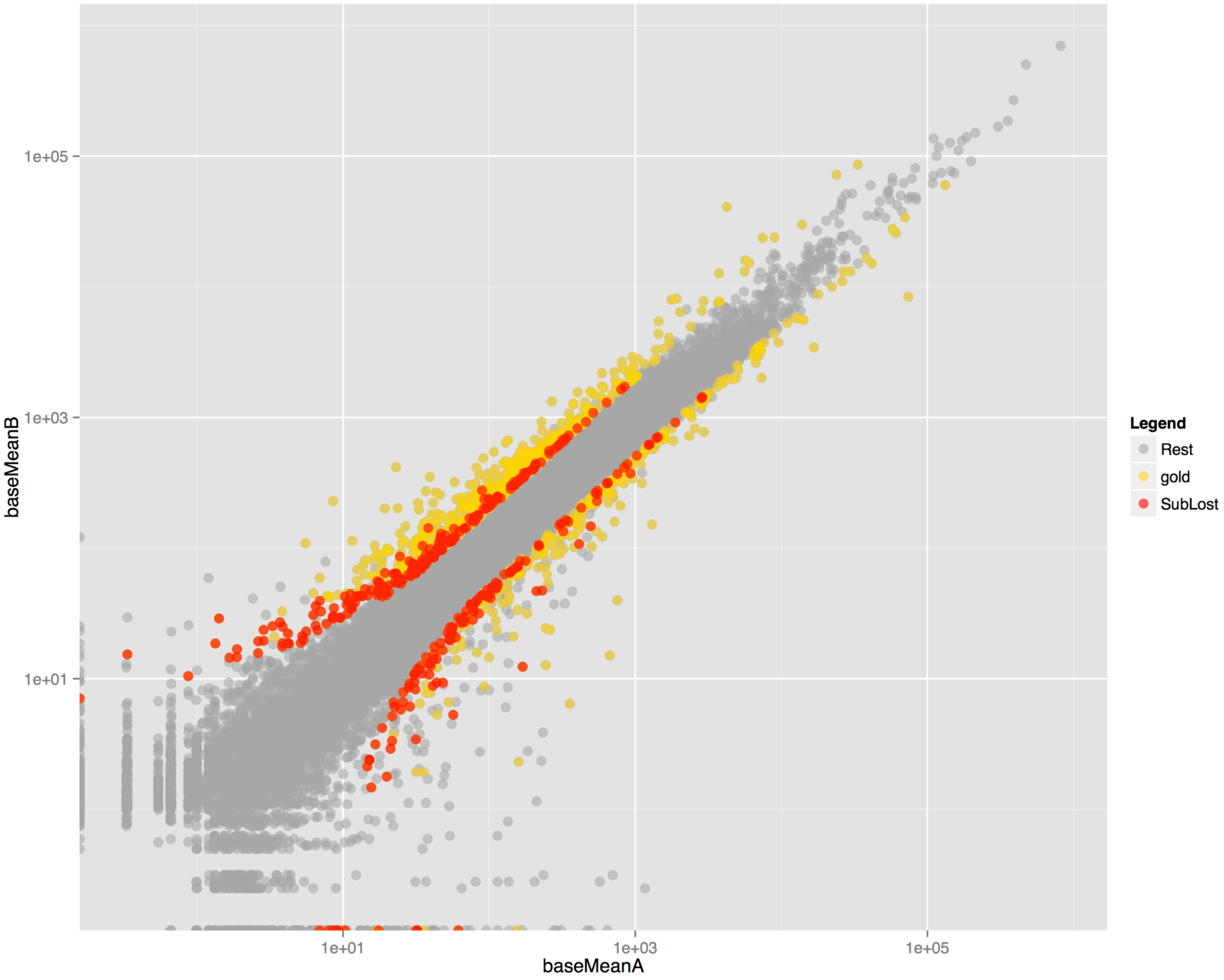
Is the loss qualitatively significant?

10 Million reads



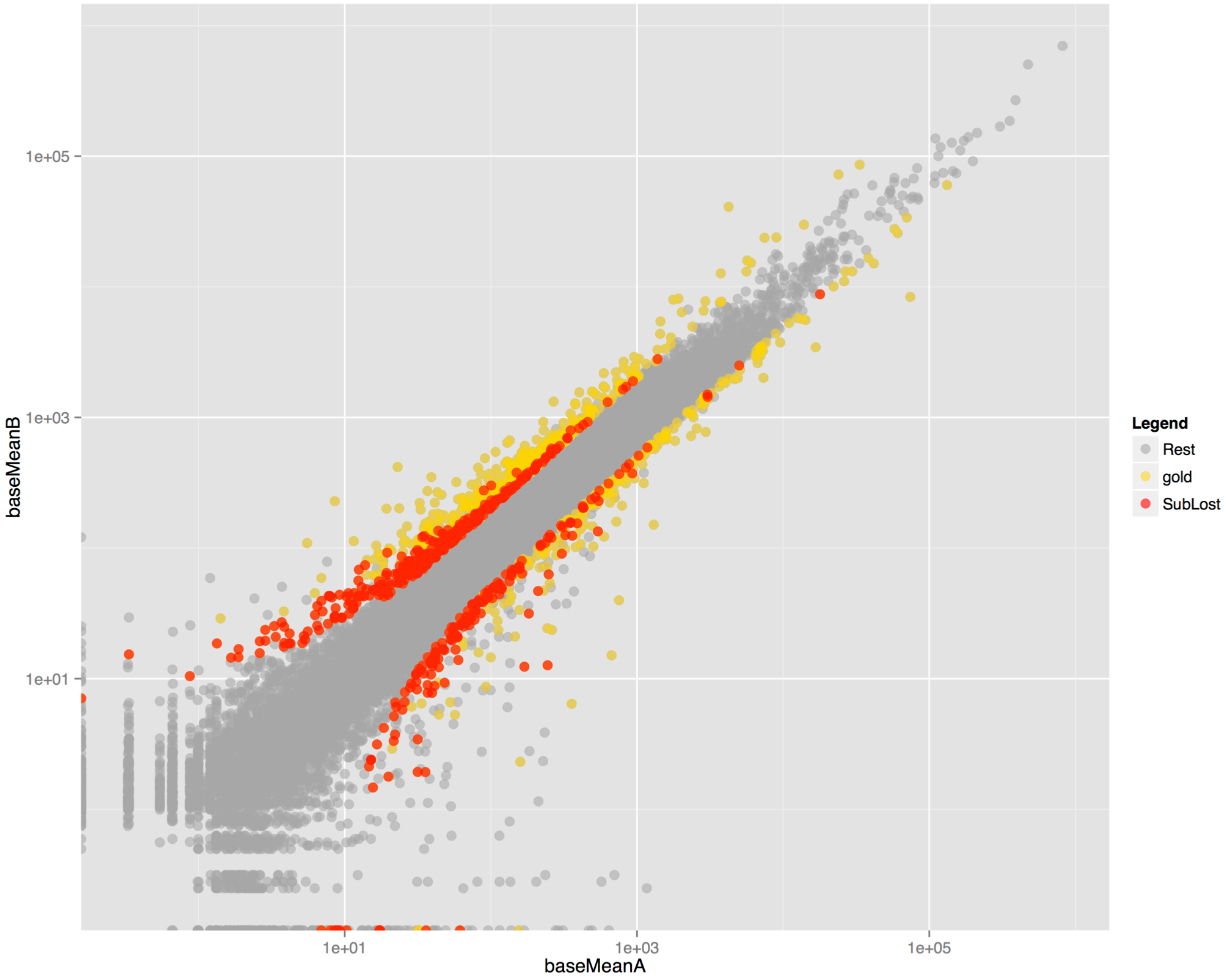
Is the loss qualitatively significant?

7.5 Million reads



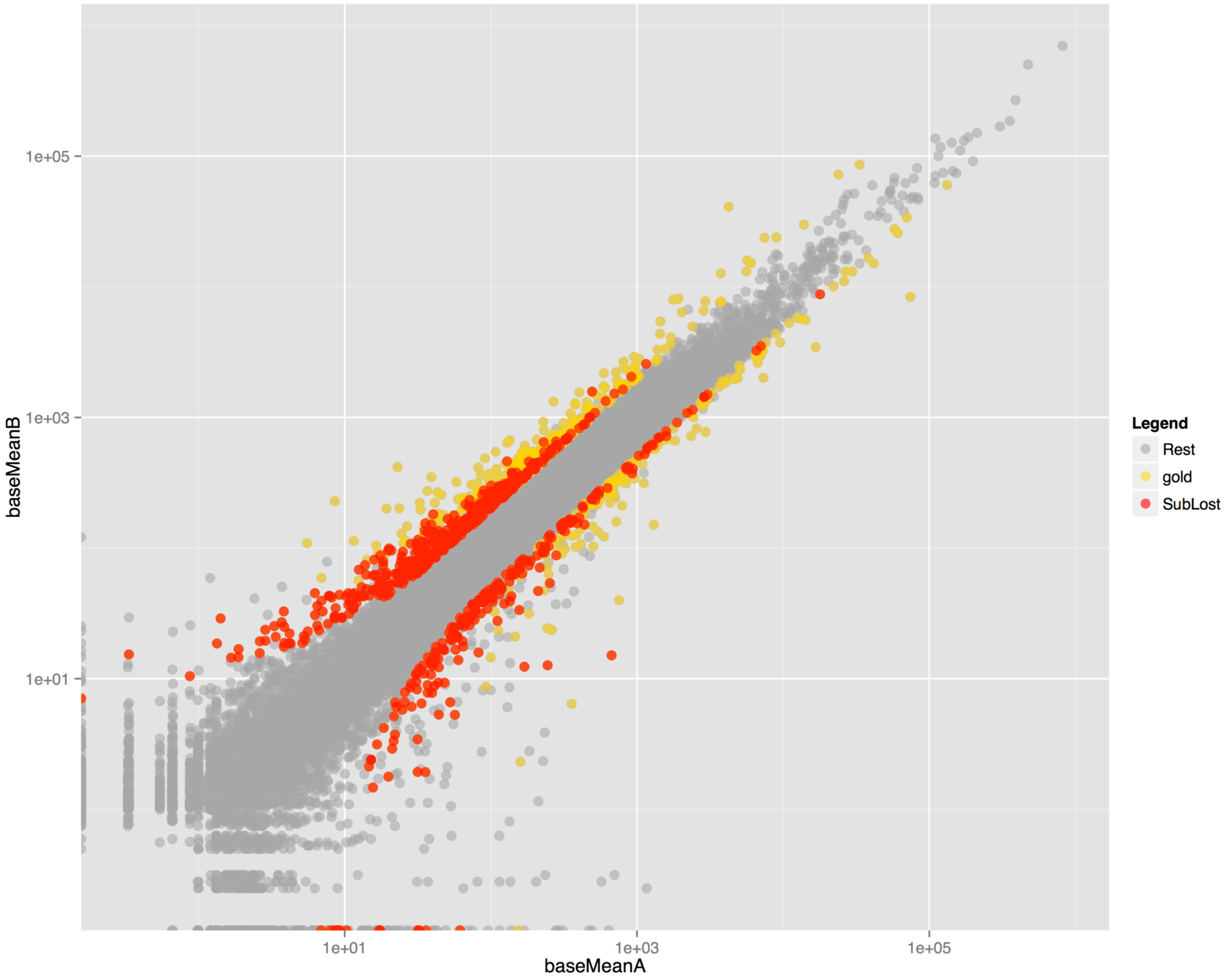
Is the loss qualitatively significant?

5 Million reads



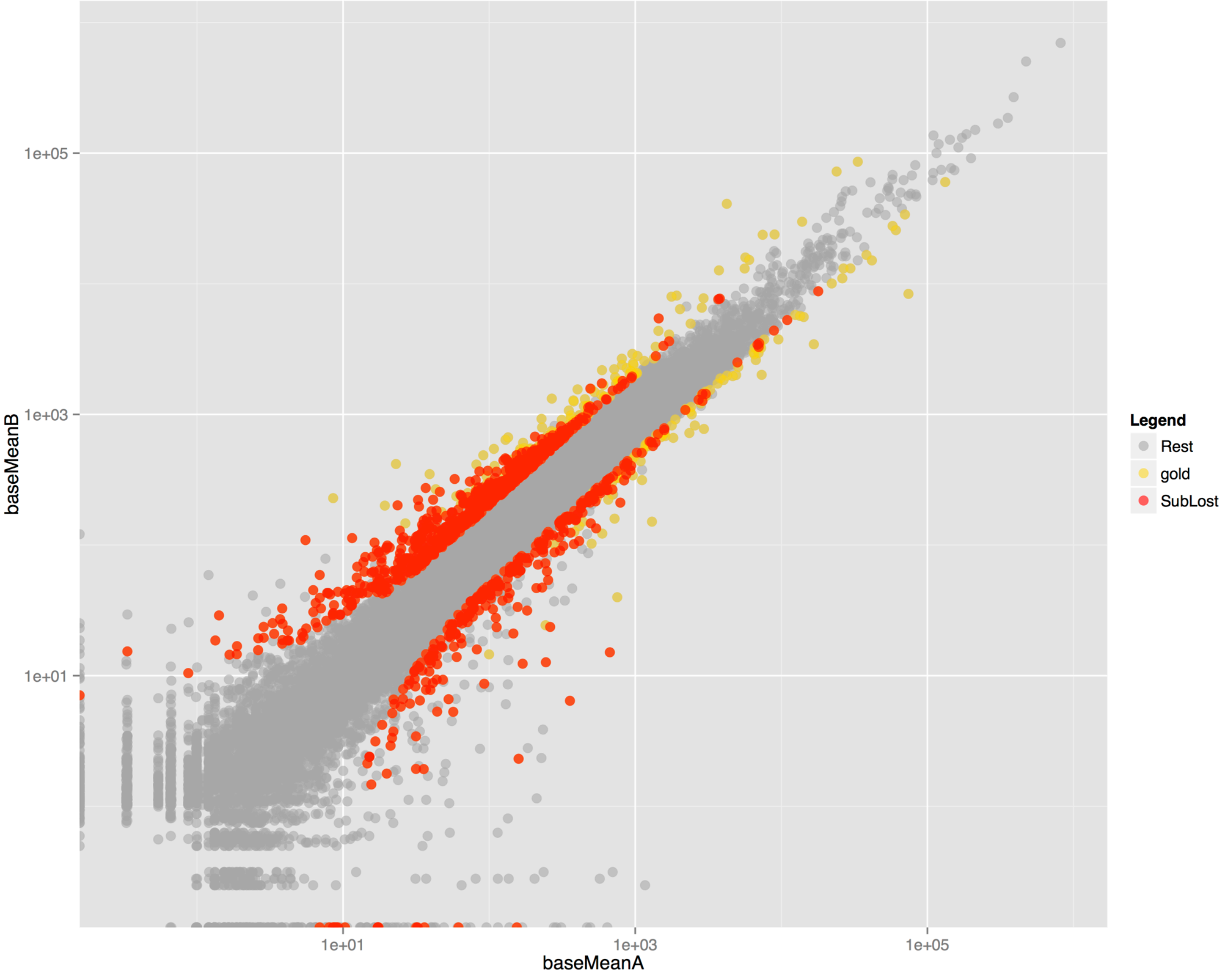
Is the loss qualitatively significant?

2.5 Million reads

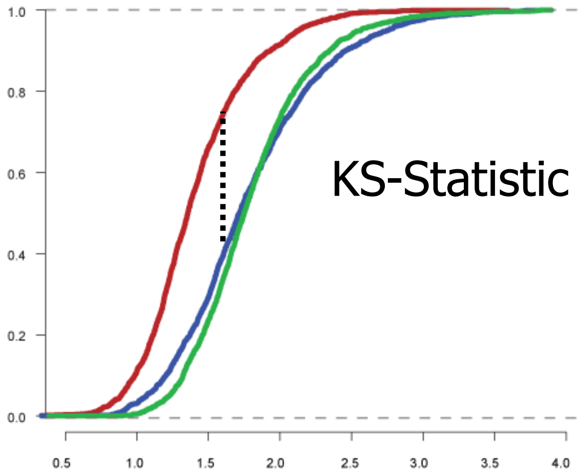


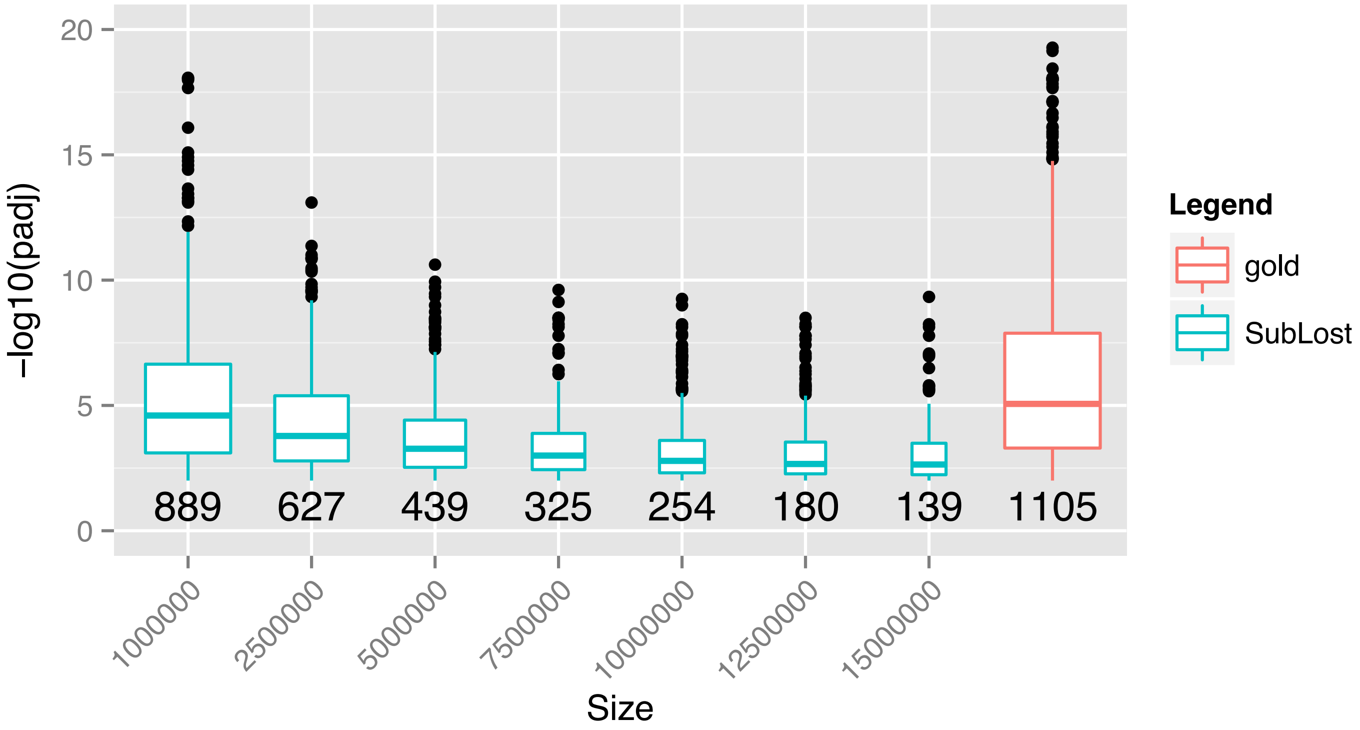
Is the loss qualitatively significant?

1 Million reads

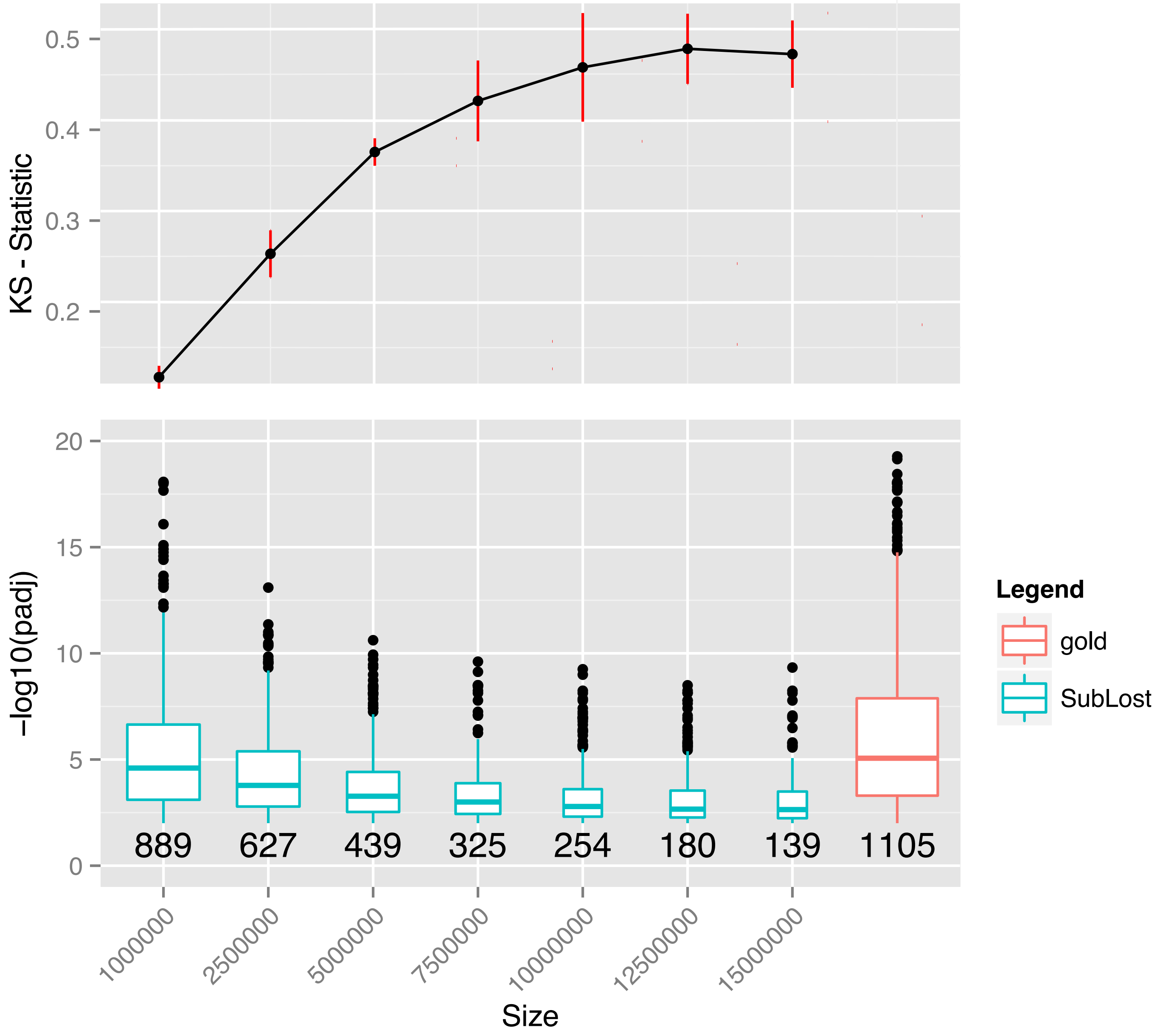


The loss is qualitatively small





The loss is qualitatively small

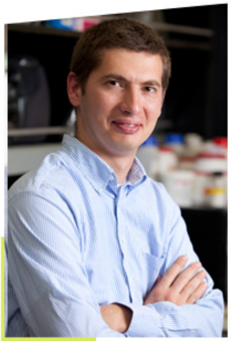


Final considerations: The steps of Sequencing analysis

- Filter reads (fastq file) by removing adapter, splitting barcodes.
  - Evaluate overall quality, look for drop in quality at ends. Trim reads if ends are of low quality
- Alignment to the genome
  - Use transcriptome if available
  - Filter out likely PCR duplicates (reads that align to the same place in the genome)
  - Evaluate ribosomal contamination
  - What percent of reads aligned
- Reconstruct(?)
- Quantify
  - Normalize according to application

Thanks

Sebastian Kadener
(HUJ)


Maxim Artyomov
(WashU)

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

Broad Technology Labs
Tarjei Mikelsen
Magali Soumillon

Garber lab & Bioinformatics Core

Alper Kucukural

Sabah Kadri
(Broad)

Jenny Chen
(MIT)

Alan Derr

DARPA

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

http://garberlab.umassmed.edu/