University of Massachusetts Medical School

Short read sequence analysis

Manuel Garber

HU-CFAR 2014

Overview of the session

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

Why do organisms look the way that they do?

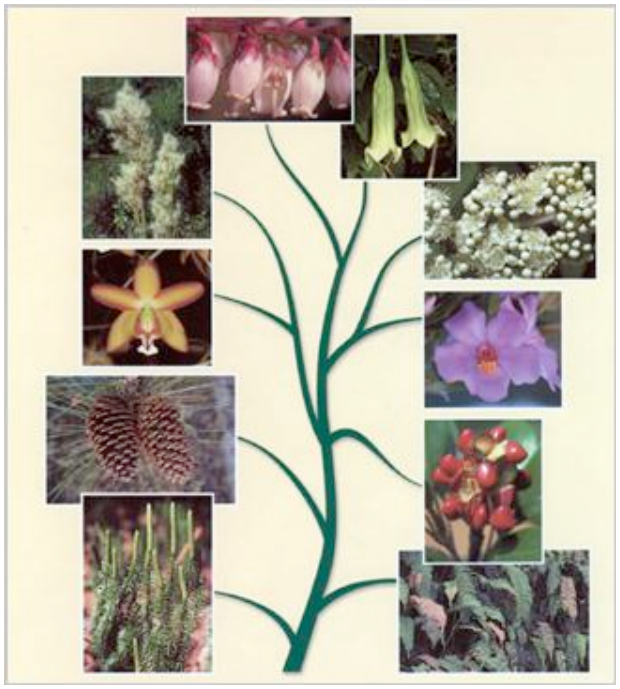








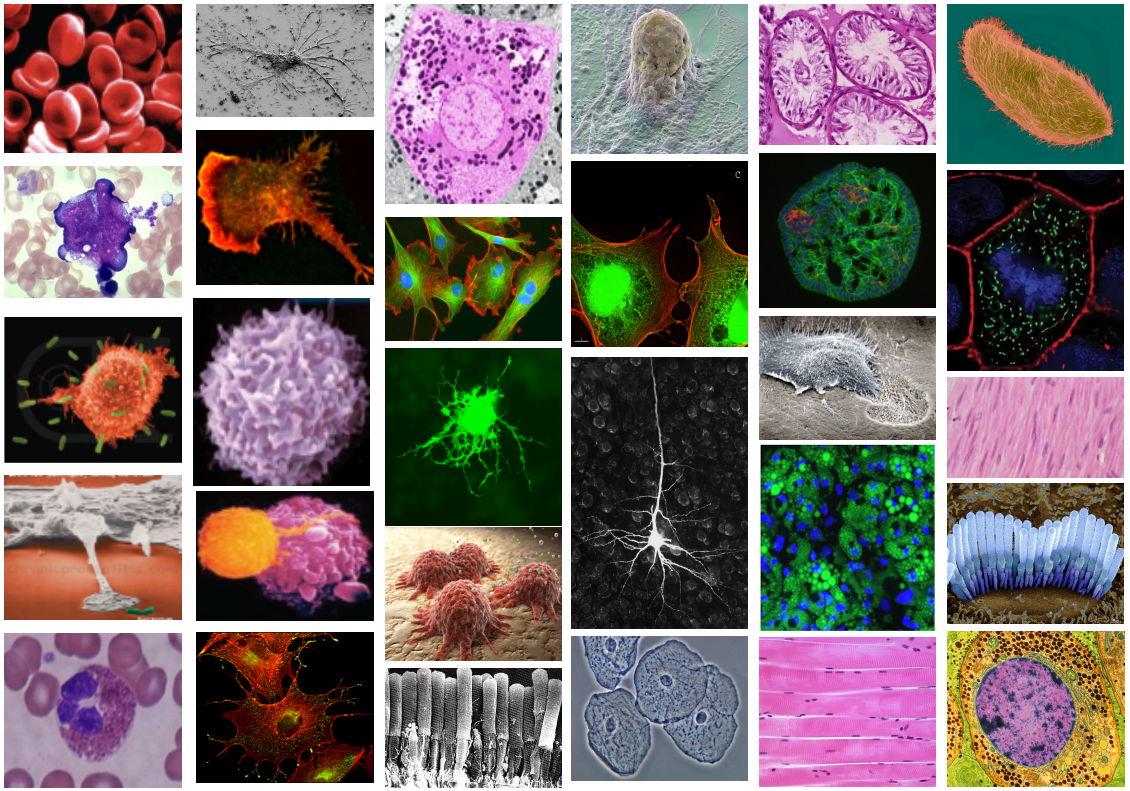






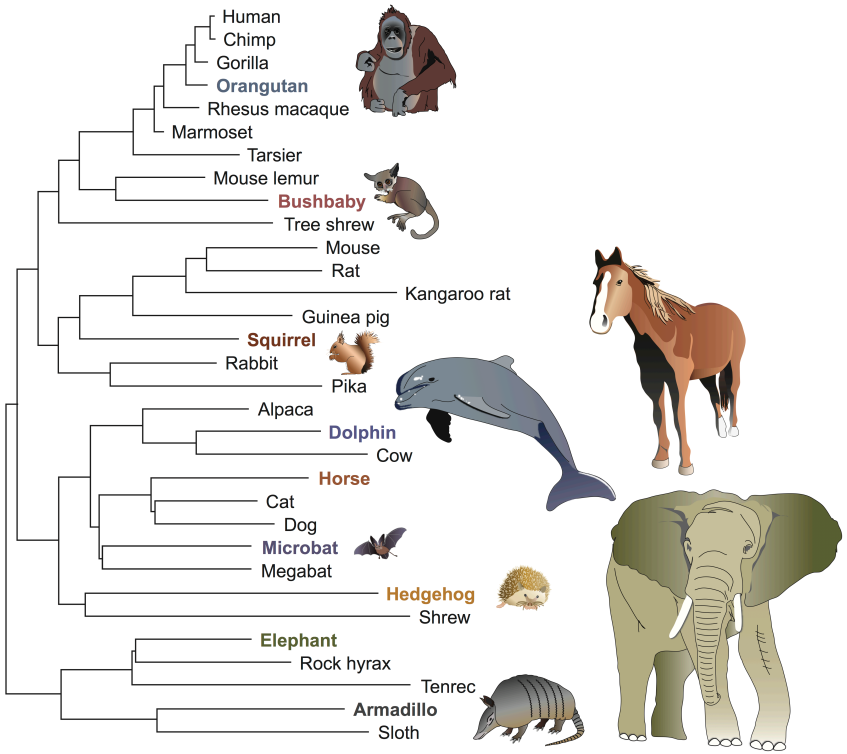
Why do different cell types do what they do!

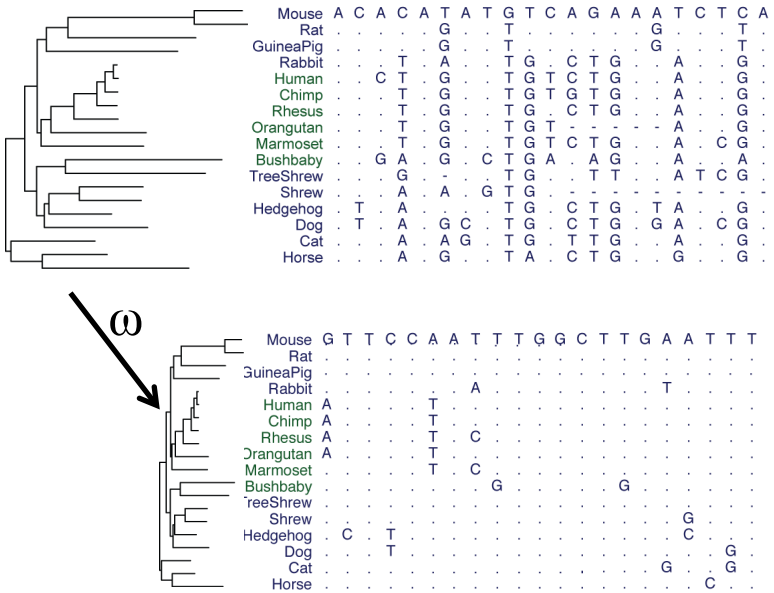


However, all this diversity arises from the same genome sequence!
Proteins are very conserved across vertebrates, what is the driving force of variability?

Sequence-based strategy: Comparative genomics

Technique: Identify regions undergoing selection





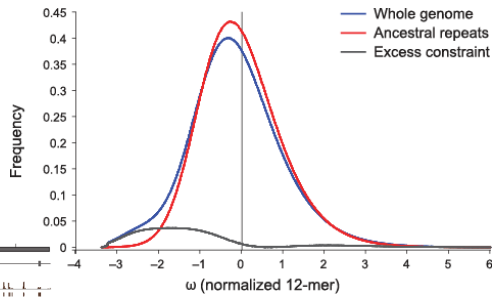
Neutral

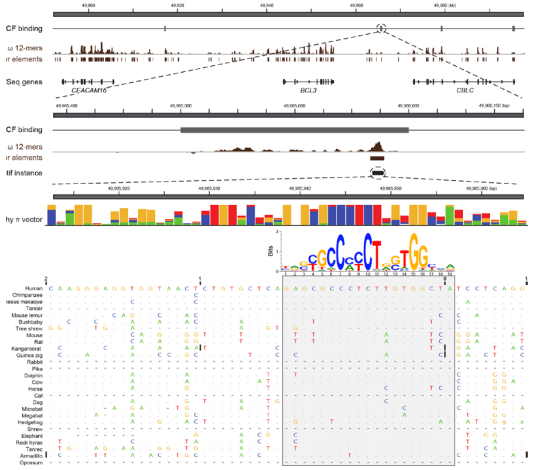
Under Selection

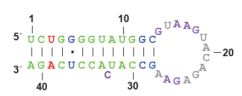
Implementation: Siphy (http://www.broadinstitute.org/genome_bio/siphy/)

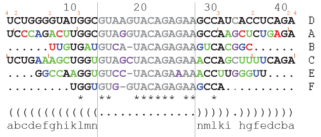
Comparative genomics genome annotation

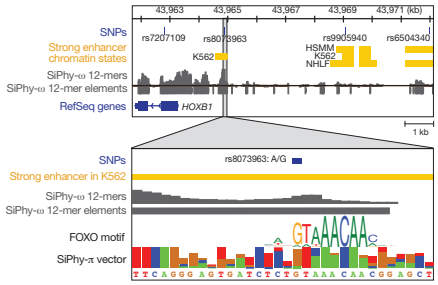
- ~7% under selection. 4.5% can be *pinpointed* at 5% FDR
- Thousands of conserved binding sites
- Hundreds of RNA structures
- Narrows associated SNPs candidates





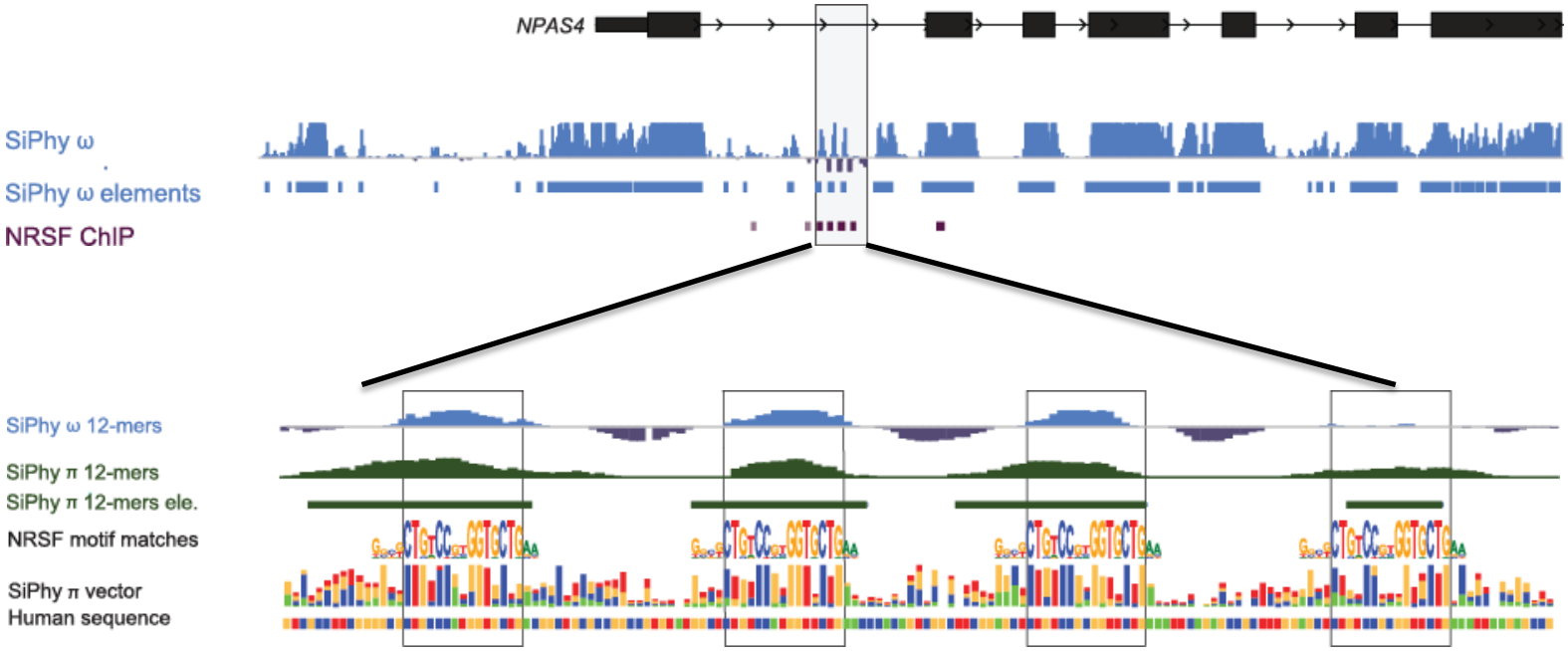






Lindblad-Toh, Garber et al. Nature 2011

In some cases resolution is astonishing

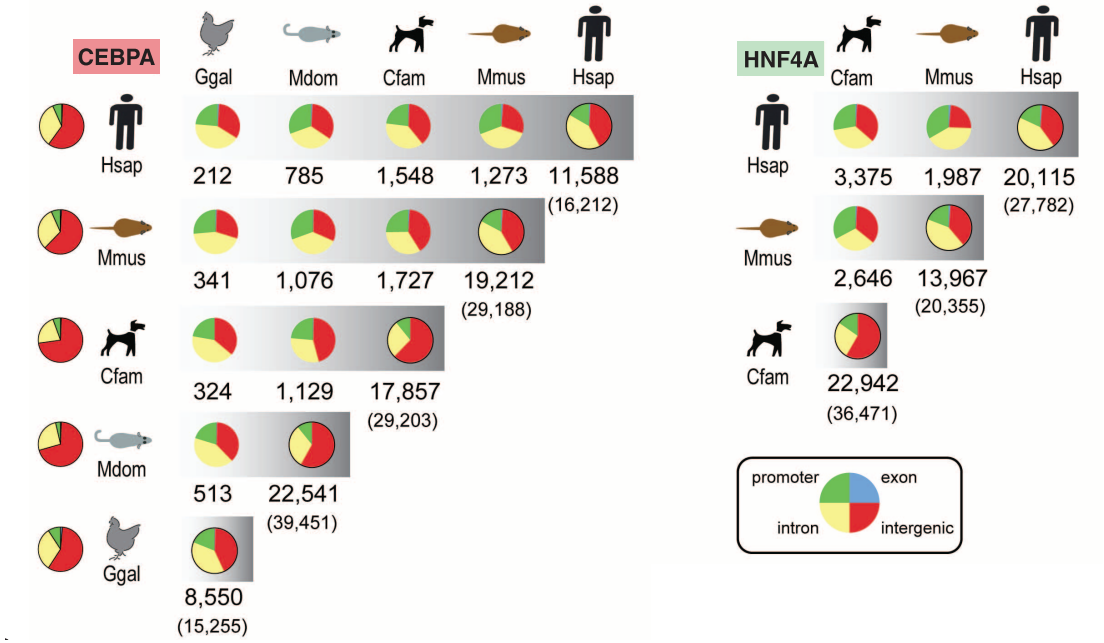


However, most binding is not conserved

REPORTS

Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding

Dominic Schmidt,[1,2]* Michael D. Wilson,[1,2]* Benoit Ballester,[3]* Petra C. Schwalie,[3] Gordon D. Brown,[1] Aileen Marshall,[1,4] Claudia Kutter,[1] Stephen Watt,[1] Celia P. Martinez-Jimenez,[5] Sarah Mackay,[6] Iannis Talianidis,[5] Paul Flicek,[3,7]† Duncan T. Odom[1,2]†



Transcriptional regulation may be a key driver of diversity and definitively of cell type diversity

Enhancers poorly conserved, 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

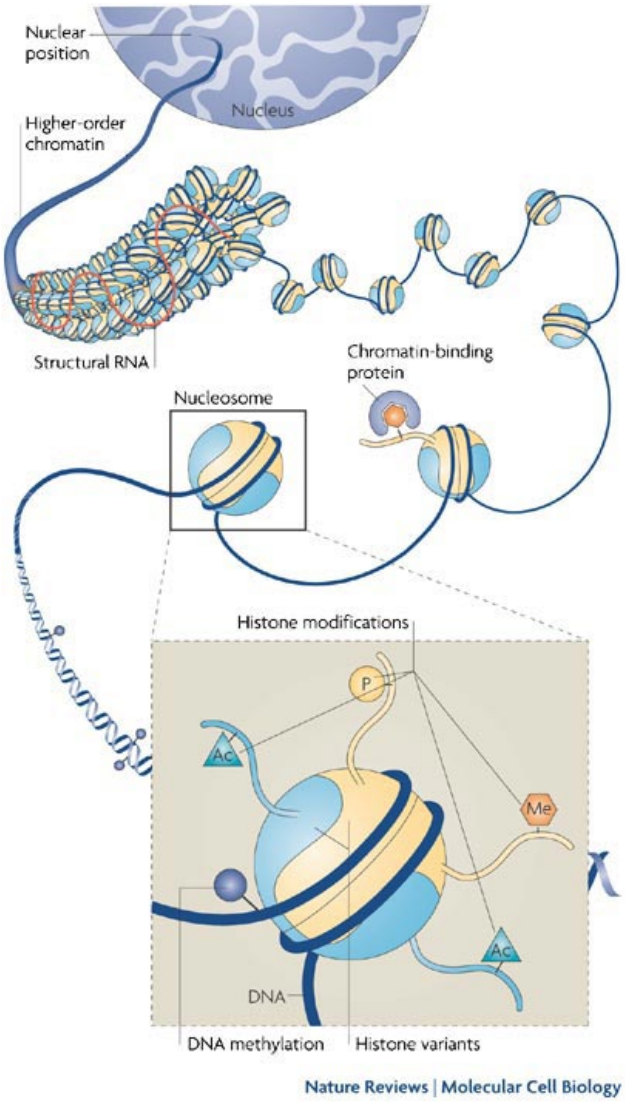
Nathaniel D. Heintzman[1,2]*, Gary C. Hon[1,3]*, R. David Hawkins[1]*, Pouya Kheradpour[5], Alexander Stark[5,6], Lindsey F. Harp[1], Zhen Ye[1], Leonard K. Lee[1], Rhona K. Stuart[1], Christina W. Ching[1], Keith A. Ching[1], Jessica E. Antosiewicz-Bourget[7], Hui Liu[8], Xinmin Zhang[8], Roland D. 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?

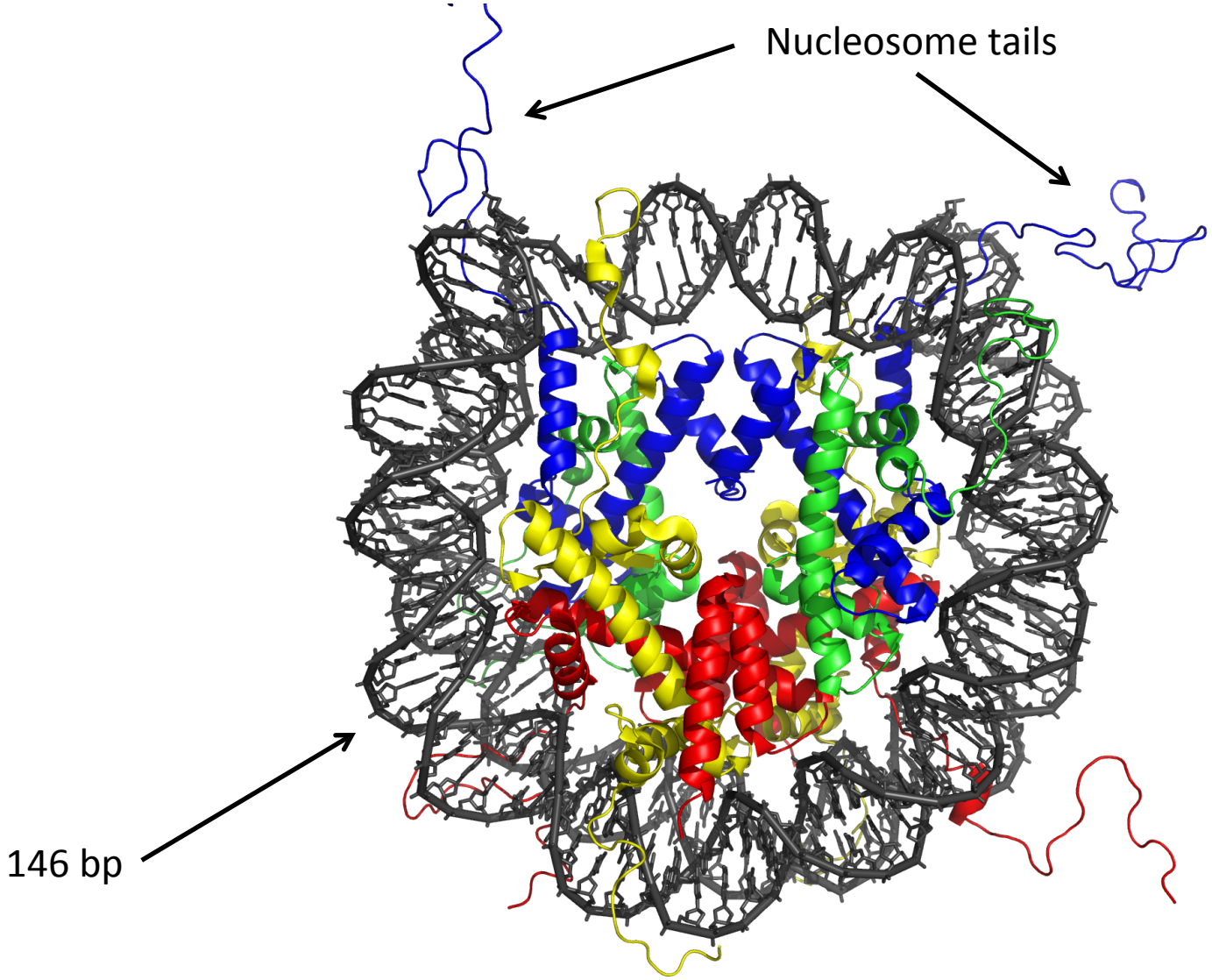
Transcription factor regulation

see:
https://www.youtube.com/watch?v=MkUgkDLp2iE

DNA is not naked

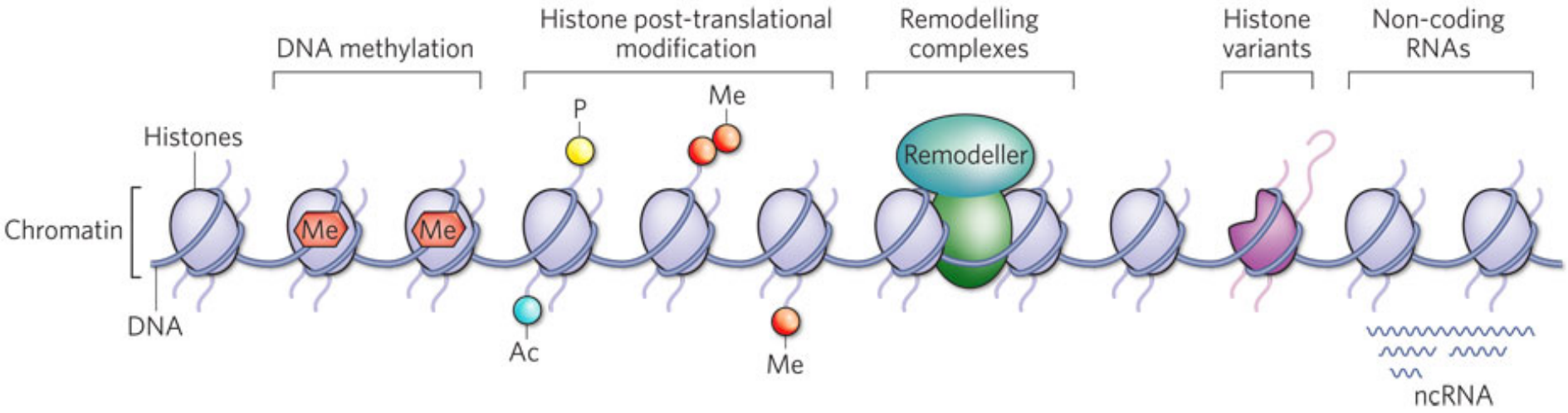


Nucleosomes interact with nuclear factors through tails



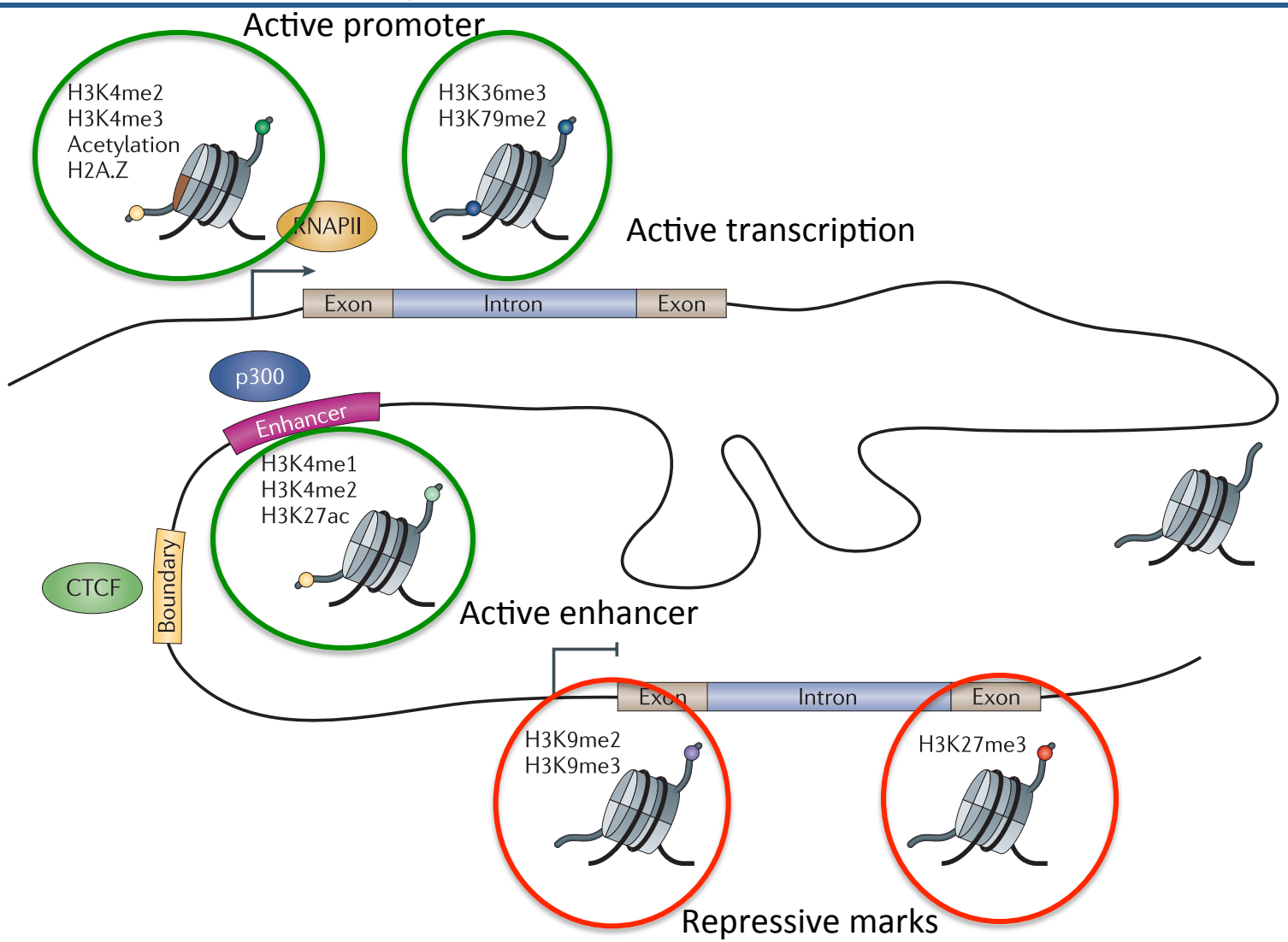
Wikipedia

Cell identity is determined by its epigenetic state



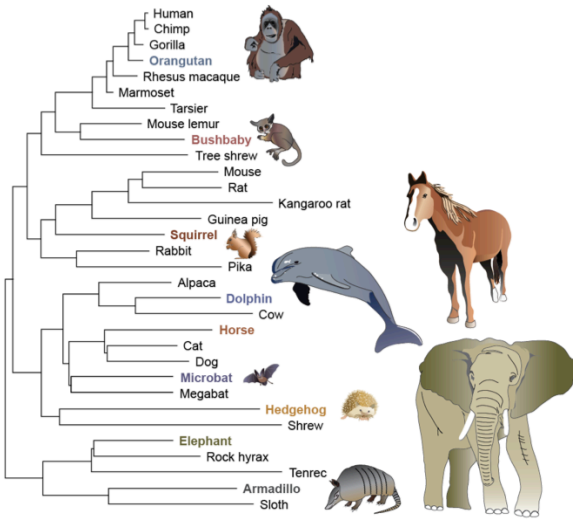
Catherine Dulac, Nature 2010

Which controls the genome functional elements

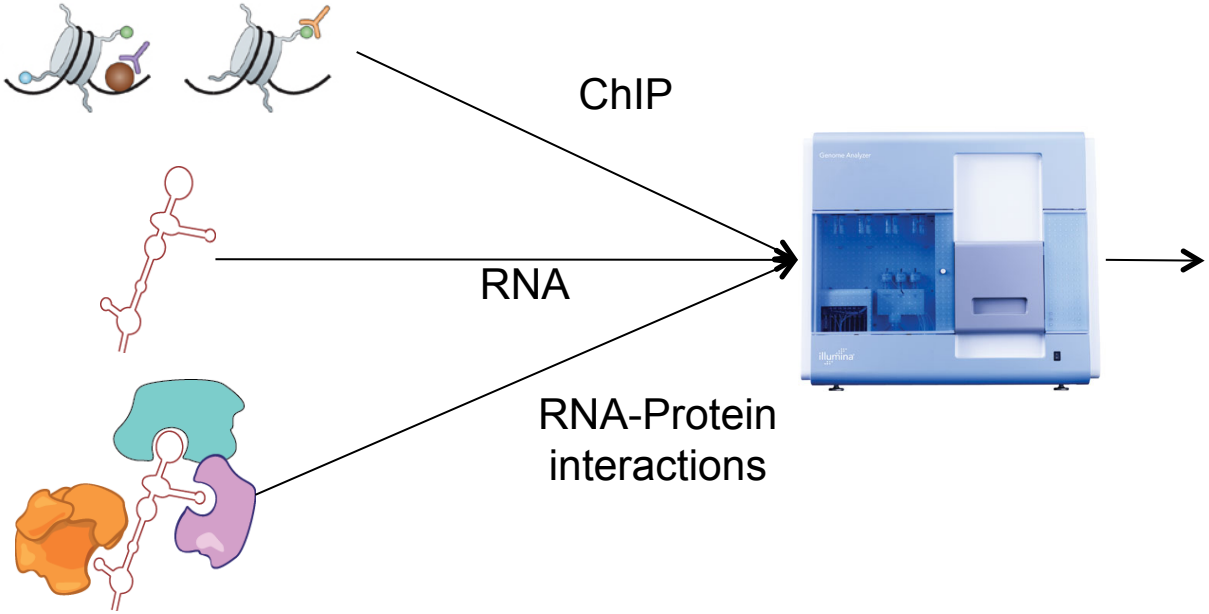


Zhou, Goren Berenstein, Nature Rev. Genetics 2011

We seek to map and functionally characterize elements



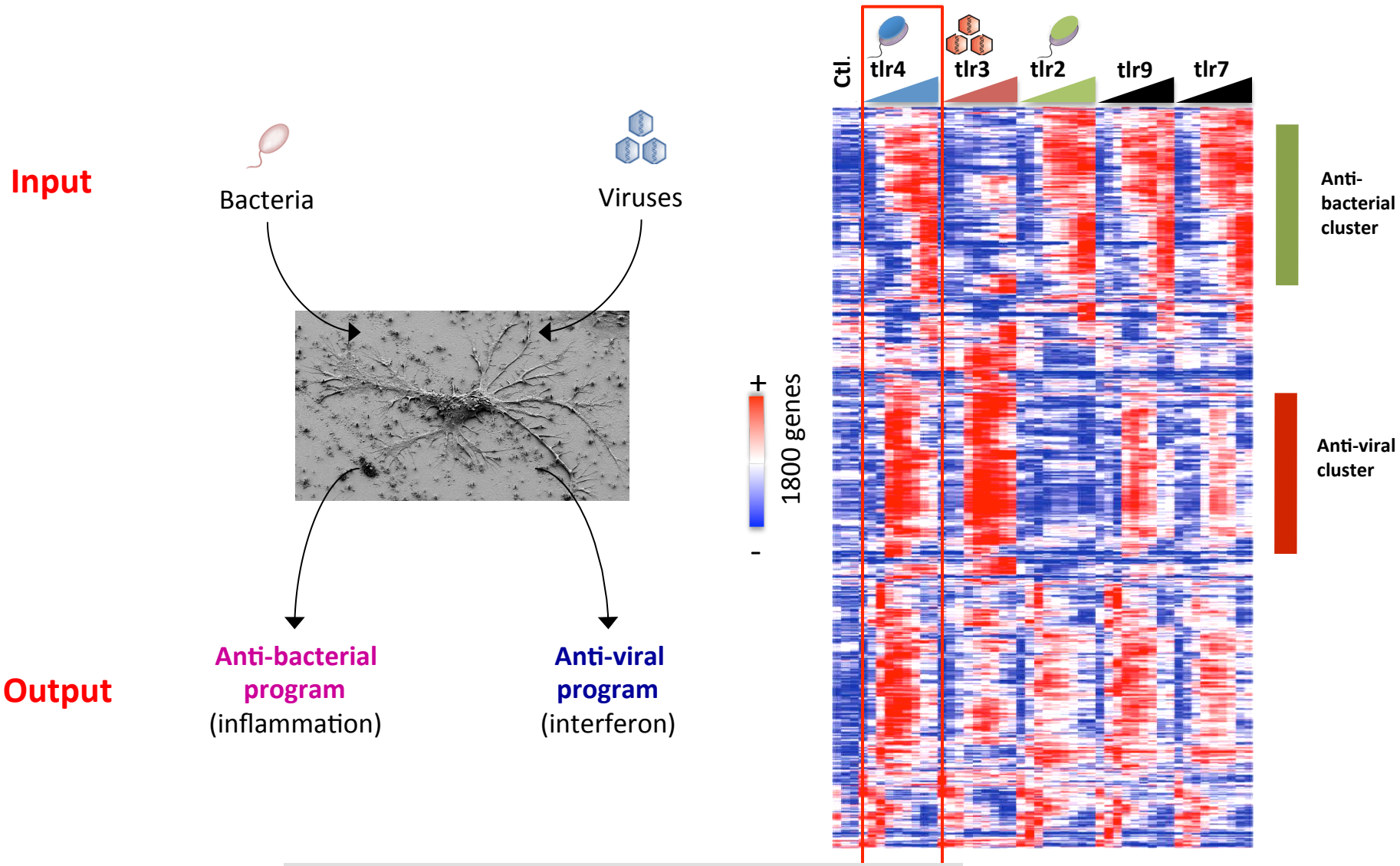
Estimate the "functional genome" by finding what is under selection



- Develop informatics tools for new methods
- Develop models of transcriptional regulation
- Develop models of epigenetic interactions
- Evolution of large non-coding RNAs

We want to ultimately understand the cell circuits of the cell

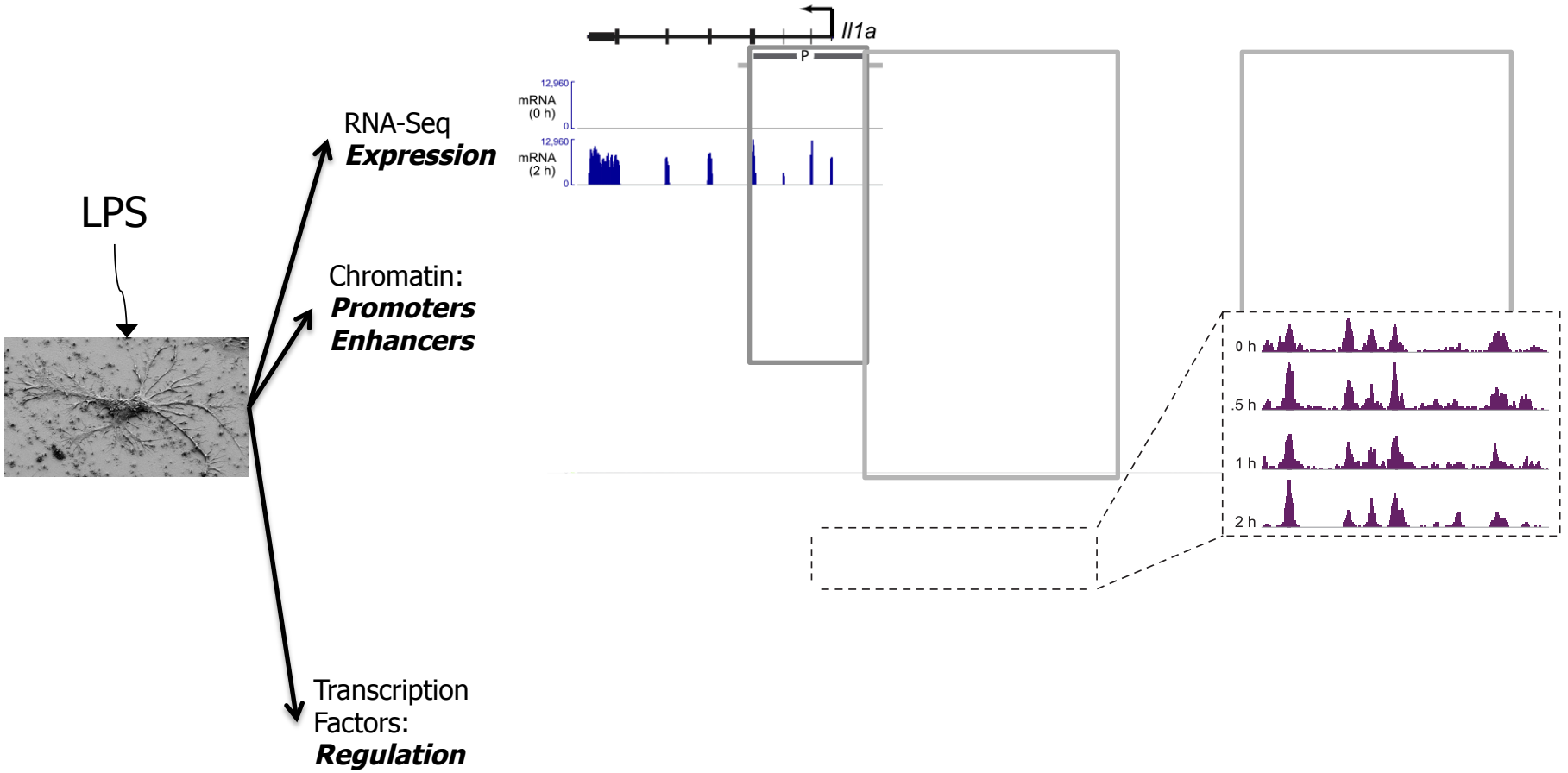
For example: wiring of innate immune cells



How is this response controlled?

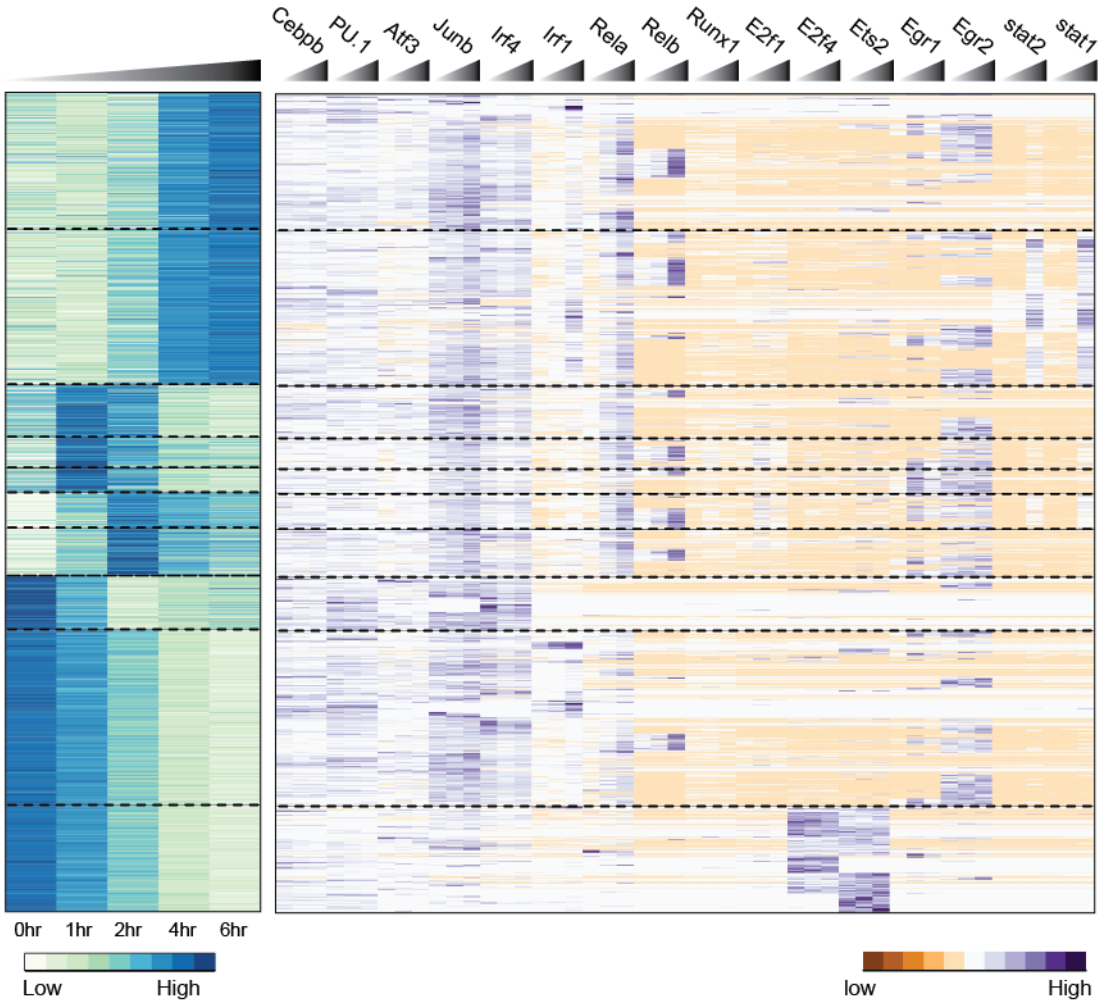
Amit, Garber et al, *Science* 2010

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



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

Into specific functional sets



Late inflammation

Anti-viral genes

No enrichment
Early inflammation

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

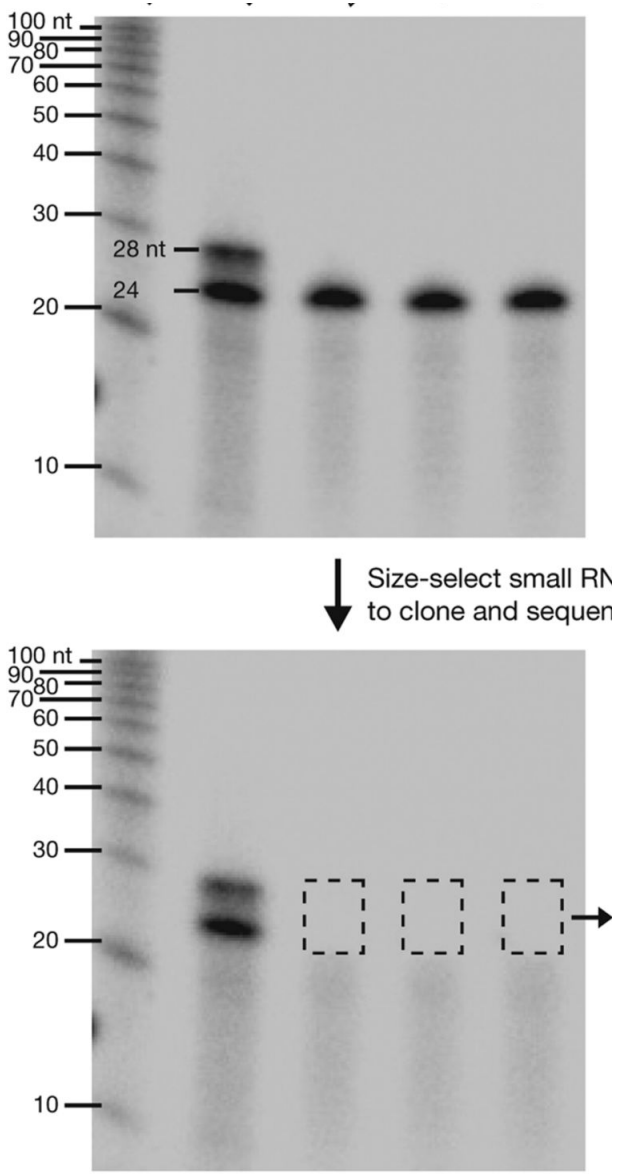
- "Source: intact, **high qual**. RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
  - Differential Gene expression. Requires ~ 5-10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

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

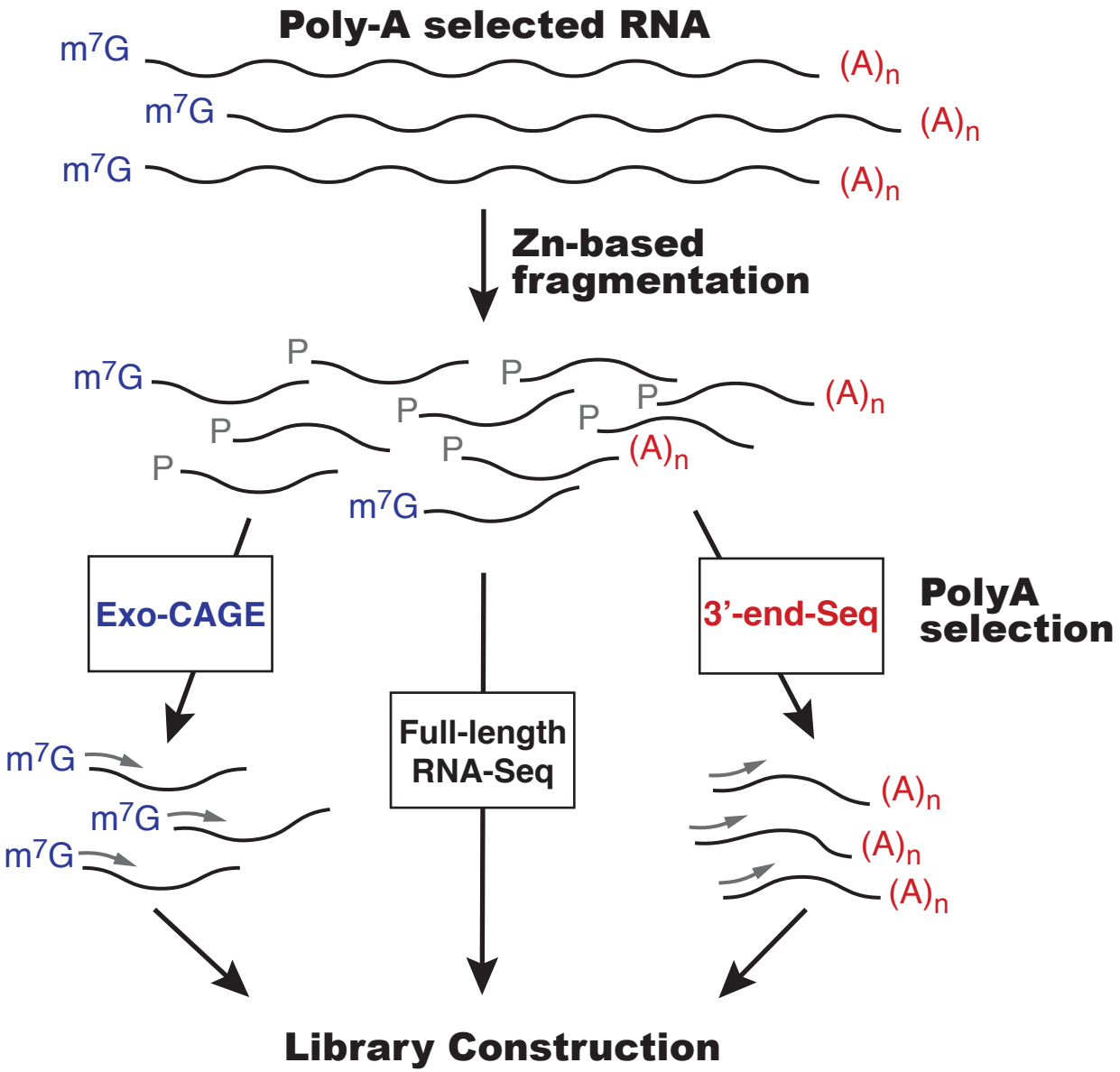


Malonne et al. CSHL protocols, 2011

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

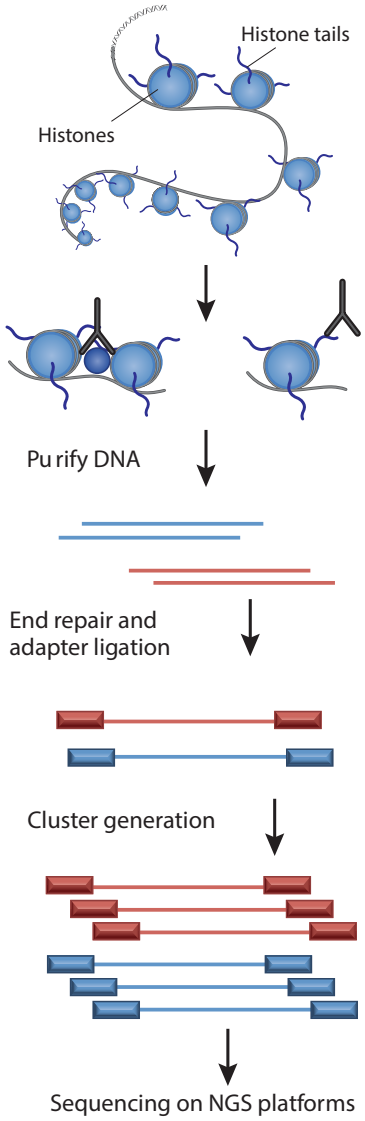
RNA-Seq libraries: Summary



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

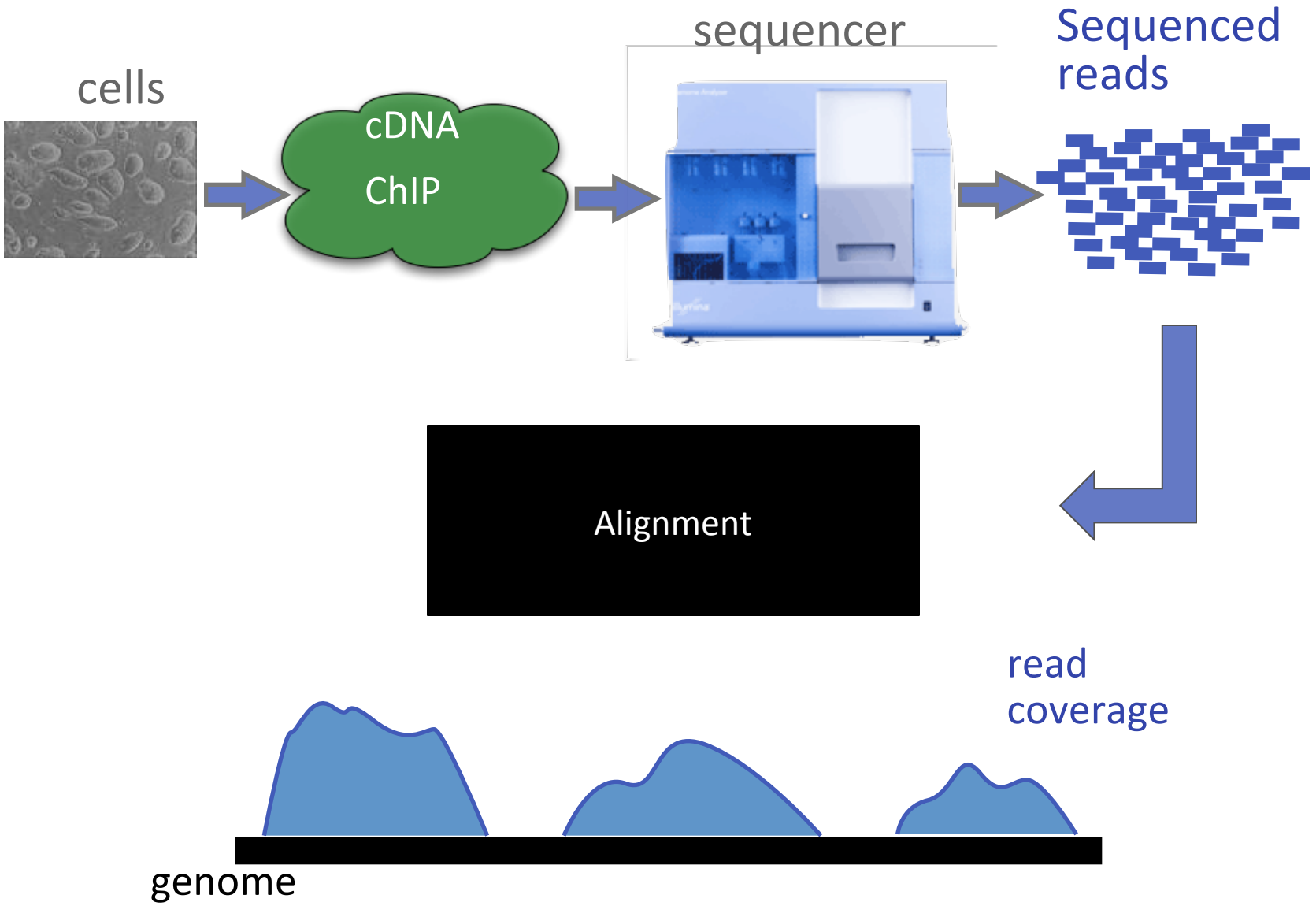
CLIP-Seq

Ribosomal profiling

Analysis of counting data requires 3 broad tasks

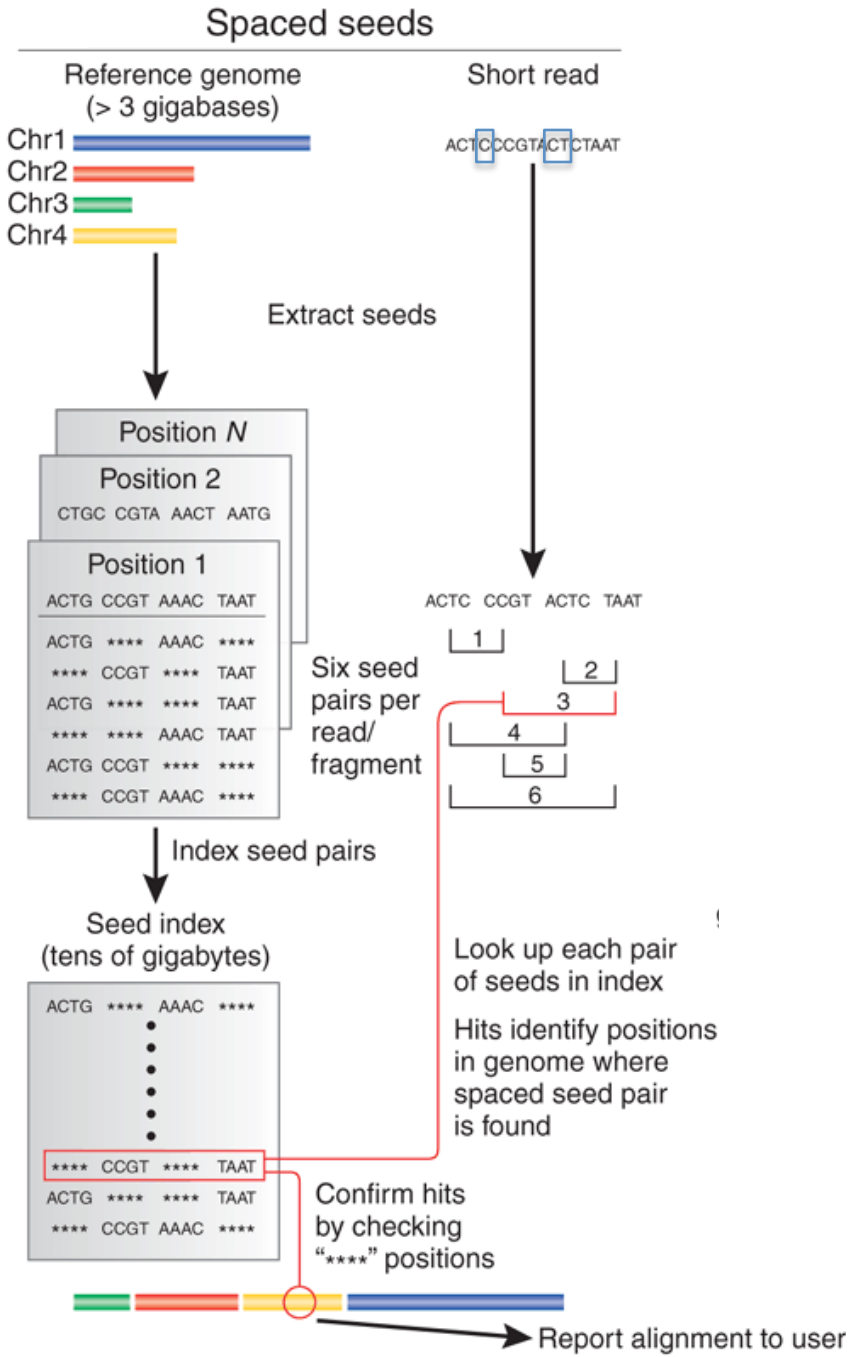
- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

Once sequenced the problem becomes computational



Analysis of counting data requires 3 broad tasks

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.



Trapnell, Salzberg, Nature Biotechnology 2009

Spaced seed alignment – Hashing the genome

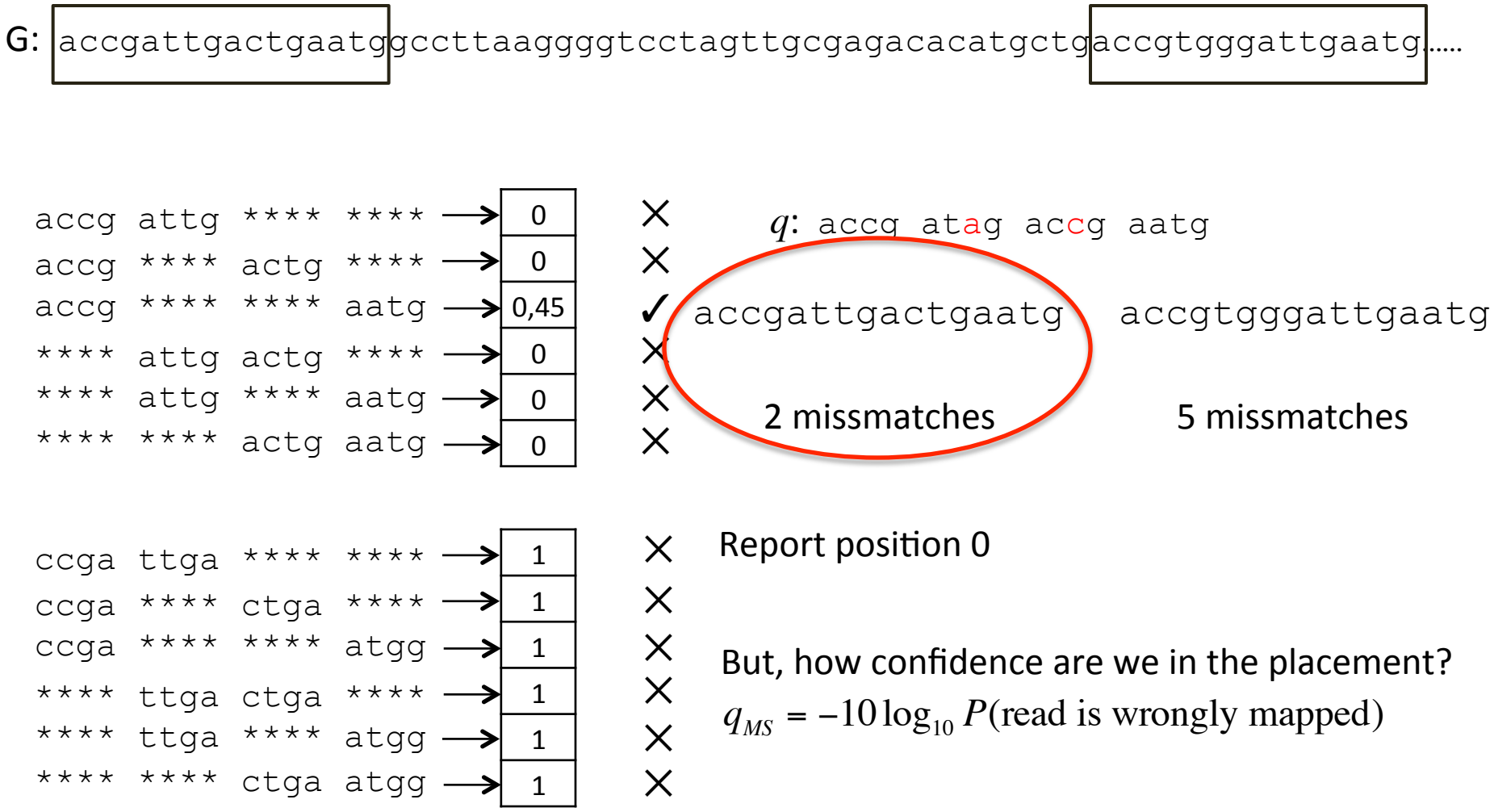
G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg……

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



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*:

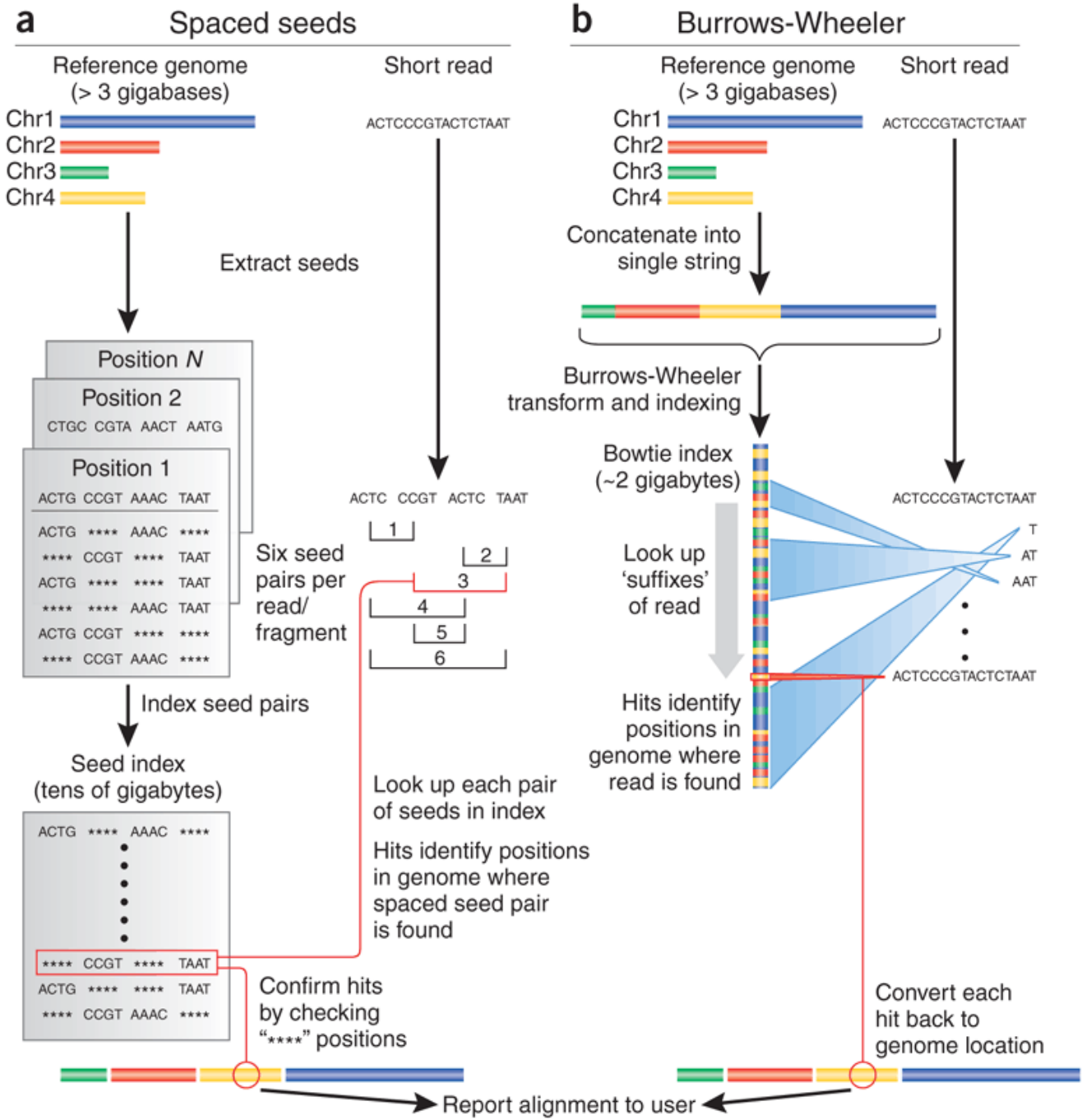
$$P(i \mid G,q) = \frac{P(q \mid G,i)P(i \mid G)}{P(q \mid G)} = \frac{P(q \mid G,i)P(i \mid G)}{\sum_j P(q \mid G,j)}$$

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

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

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)

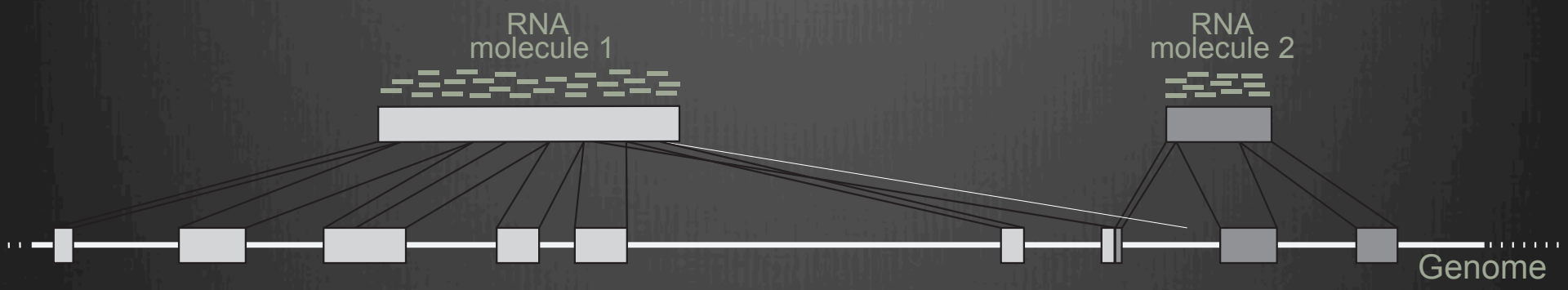


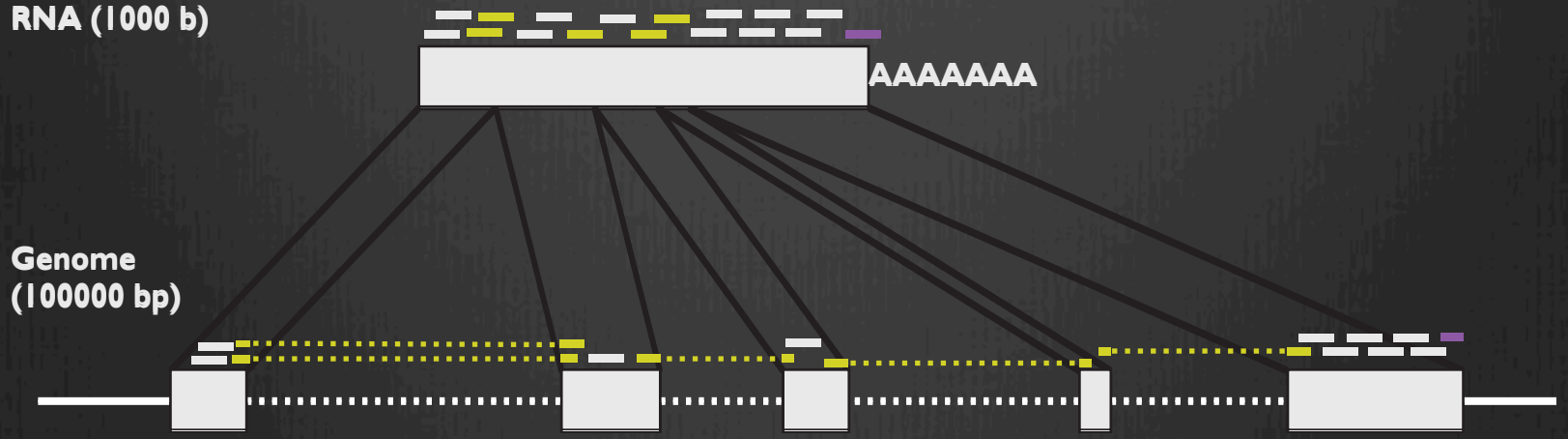
Trapnell, Salzberg, Nature Biotechnology 2009

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

RNA-Seq Read mapping





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

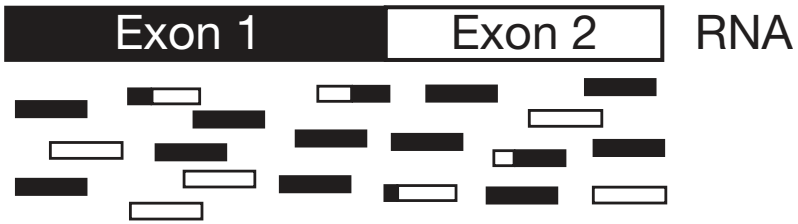
BWT

	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**

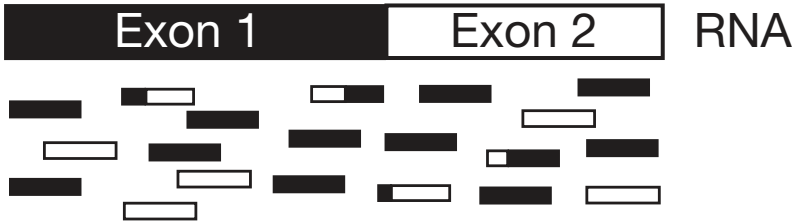
***Bowtie2 breaks reads into smaller pieces and maps these "seeds" using a BWT genome.**

Seed-extend spliced alignment (e.g. GSNAP)



Exon-first spliced alignment (TopHat)

Exon-first approach



Short read mapping software for RNA-Seq

Seed-extend

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

Exon-first

	Use base qual
STAR	NO
TopHat	NO

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

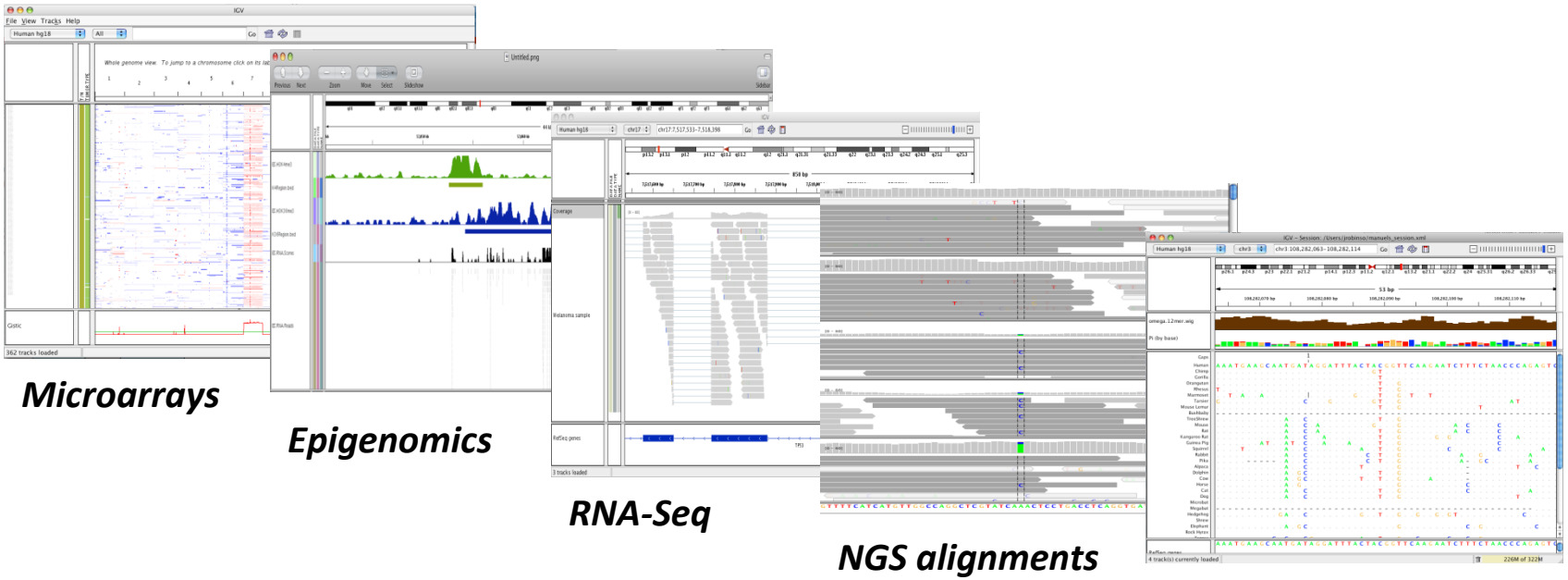
IGV: Integrative Genomics Viewer

Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



Microarrays

Epigenomics

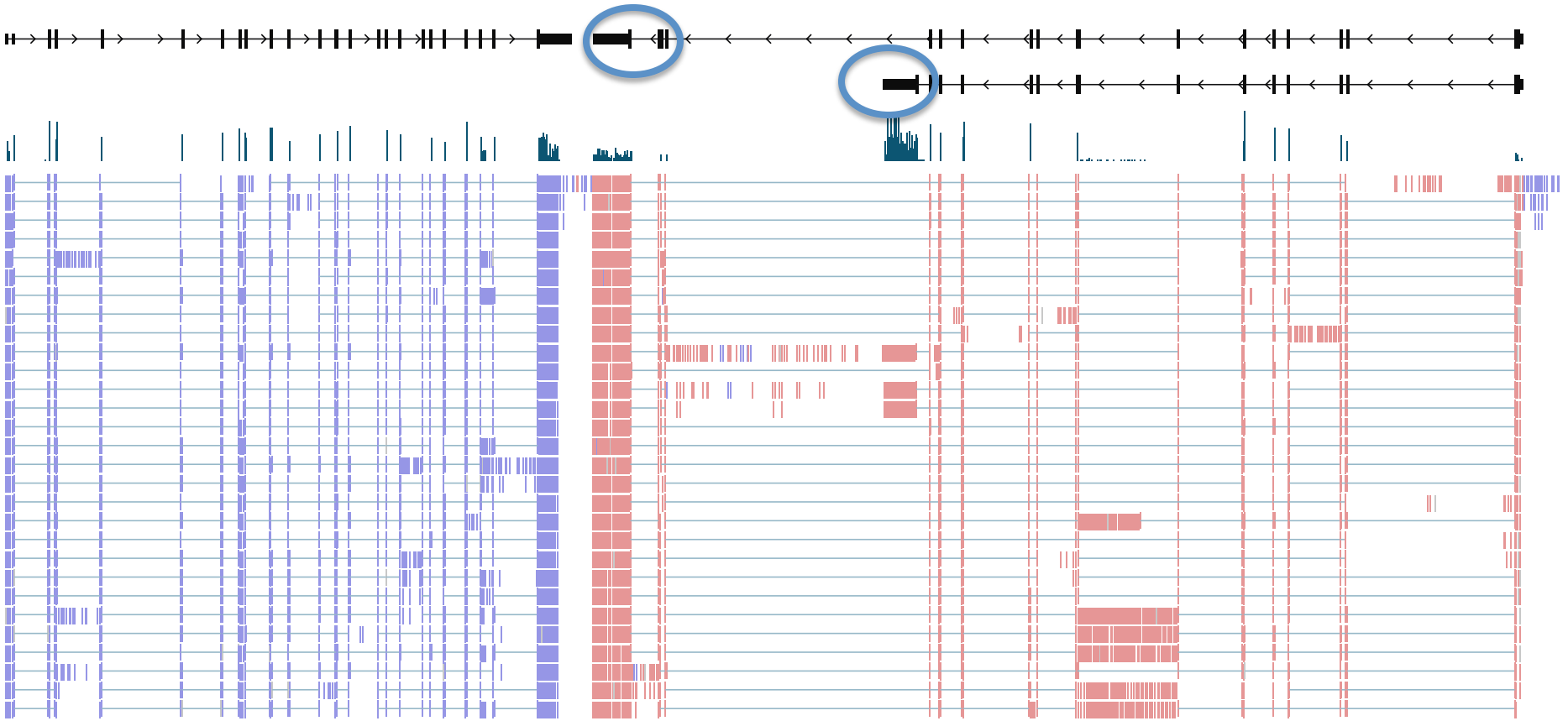
RNA-Seq

NGS alignments

Comparative genomics

BROAD INSTITUTE

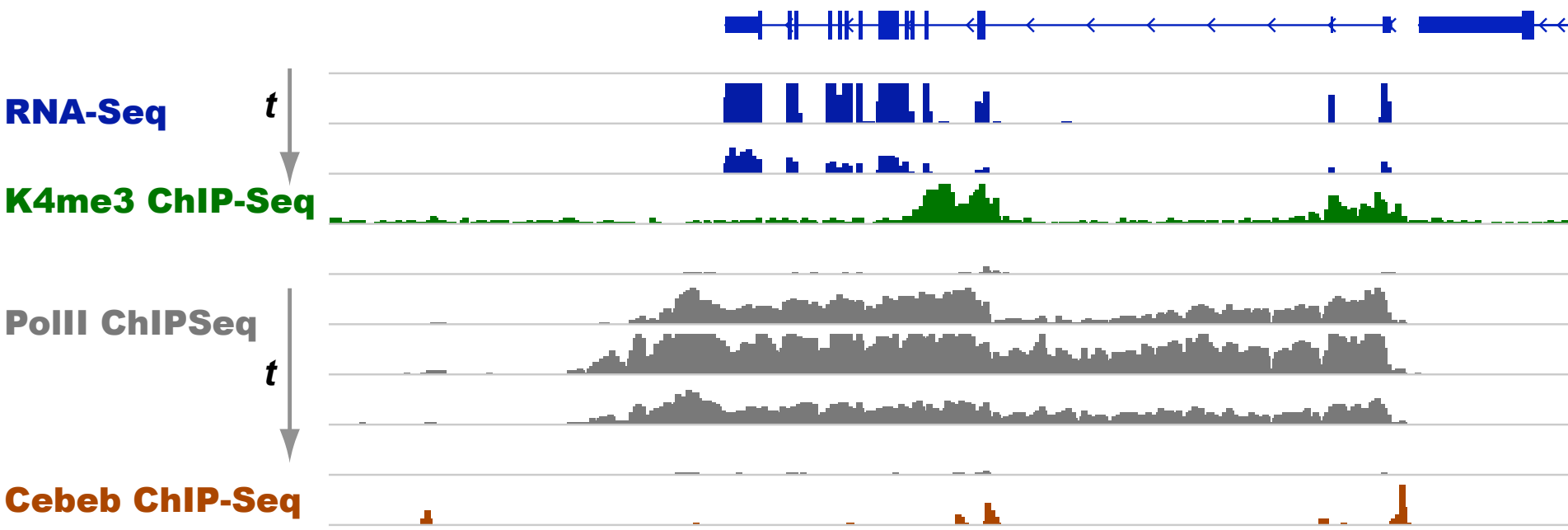
Visualizing read alignments with IGV — RNASeq



Strand specific library!

Gap between reads spanning exons

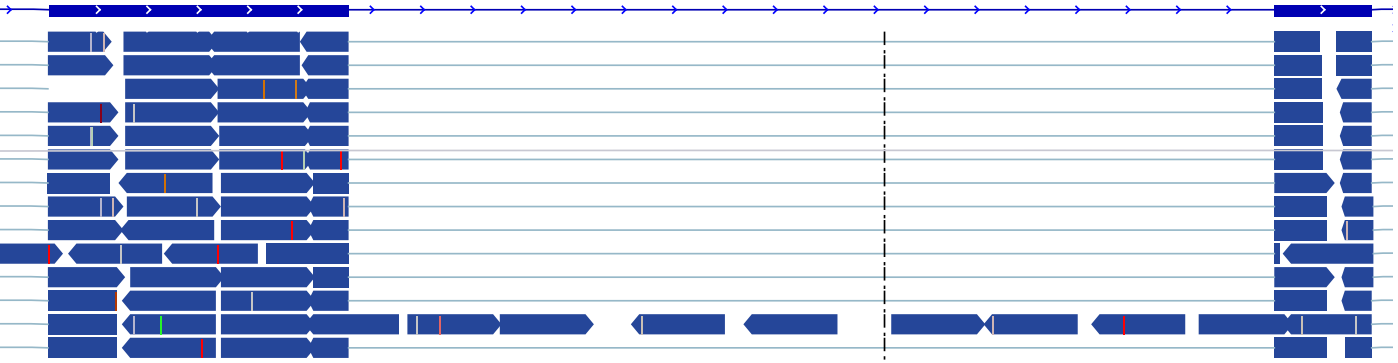
Visualizing read alignments with IGV — zooming out



Mapping longer reads

MiSeq “Bench” sequencer
~15 Million 2x250 base reads.
Ideal for **deep annotation of Targeted RNA**

Large number of expected mismatches
Given sequencing errors (>1.5%) + SNPs
expect many reads with >4 missmatches



Short (76b) reads

Long (250b) reads

Longer, reads mapping cannot be done with standard BWT based aligners

How do "short" read aligners responded to read increase?

- Break reads into seeds (e.g. 16nt every 10nt)
- Use BWT or HashTable to find candidate positions
- Prioritize candidates
- Extend top candidates using classical alignment techniques.

Aligner	Technique
TopHat2 (Bowtie2)	BWT
GSNAP	Hash Table
STAR	Suffix (similar to BWT)

Analysis of counting data requires 3 broad tasks

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

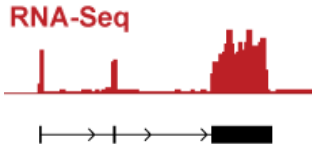
What does significance means?

- 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

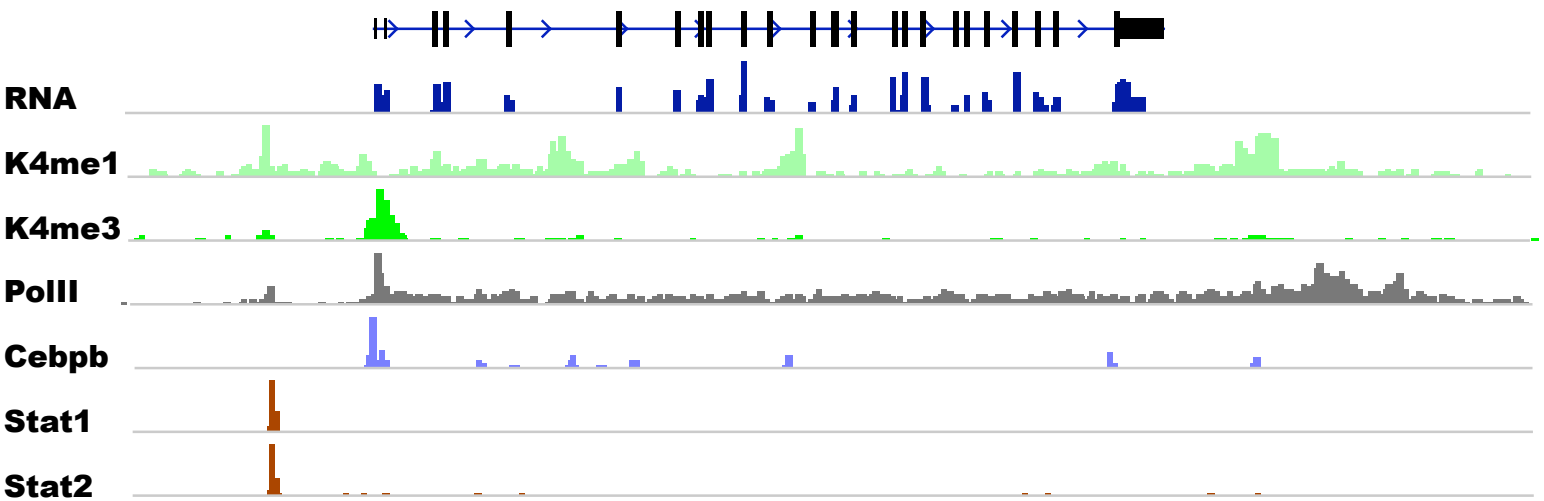
How do we find peaks?

Short modification

Long modification

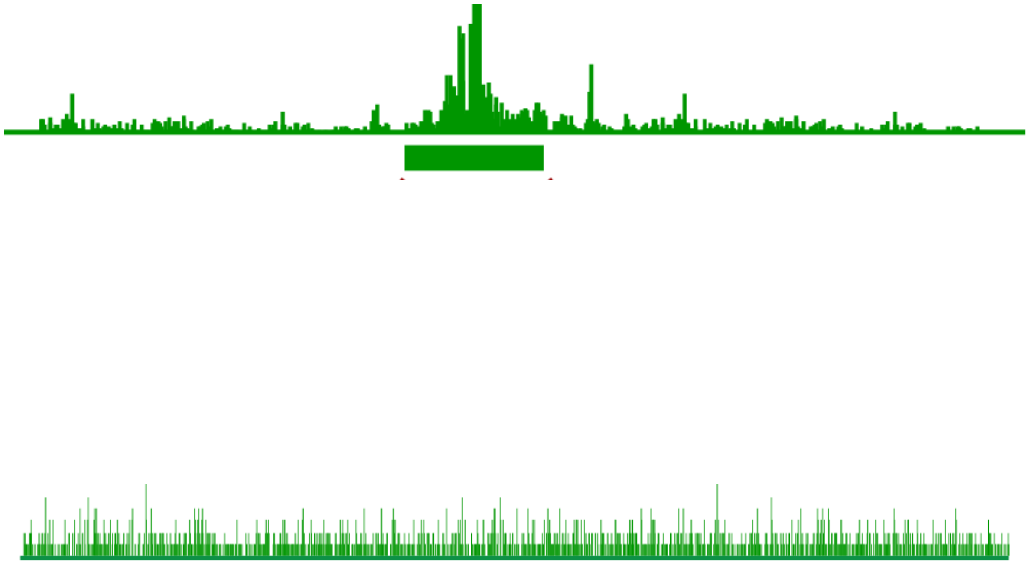


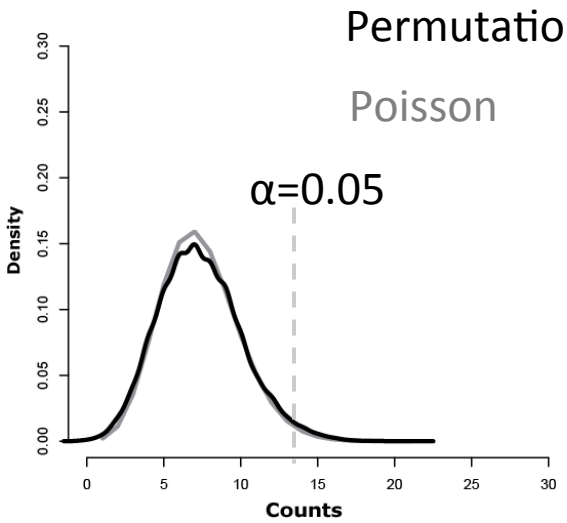
Discontinuous data



Scripture is a method to solve this general question

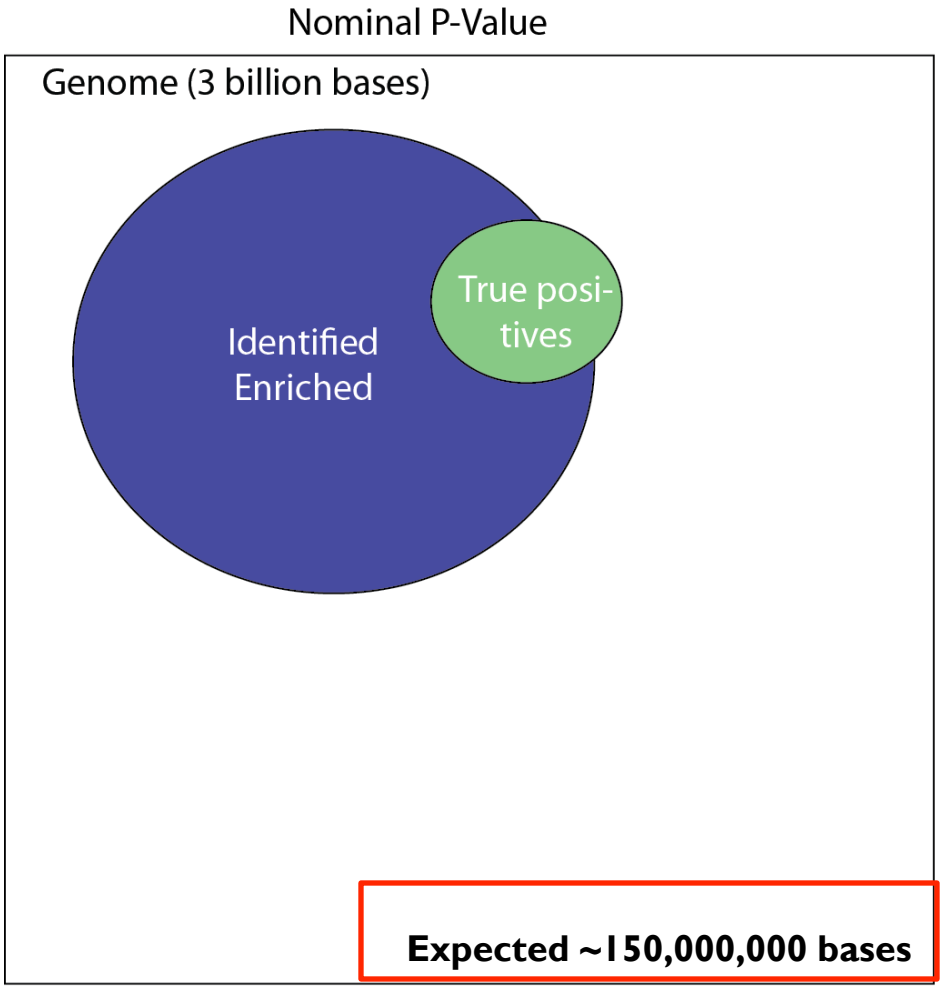
Our approach





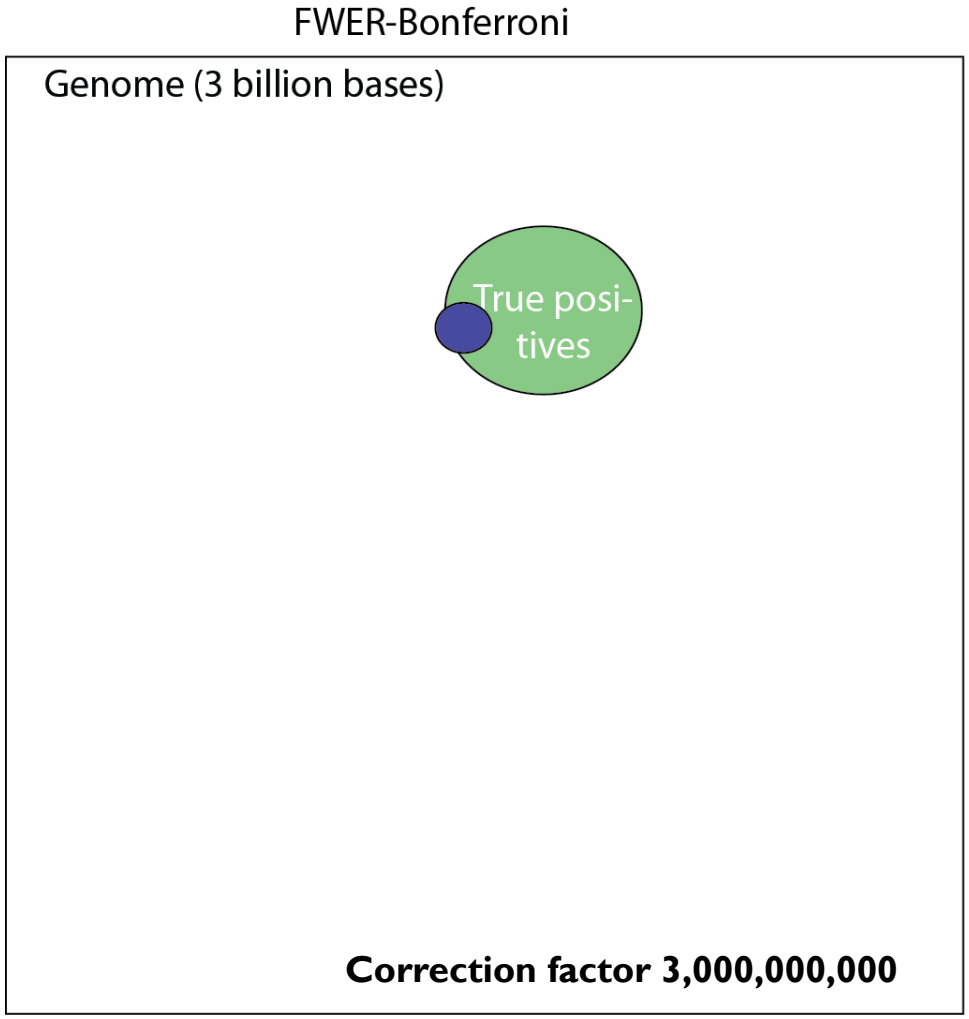
We have an efficient way to compute read count p-values …

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative

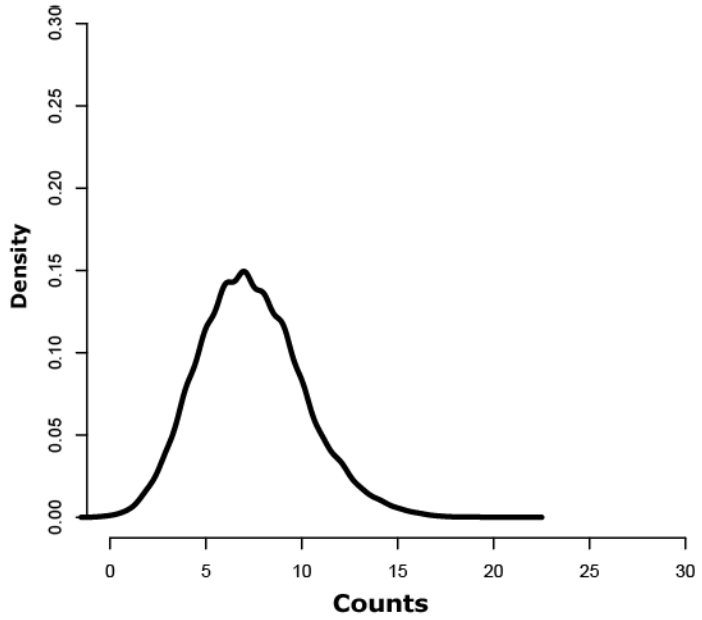


Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

Controlling FWER

Max Count distribution

α=0.05 α_{FWER}=0.05



Count distribution (Poisson)

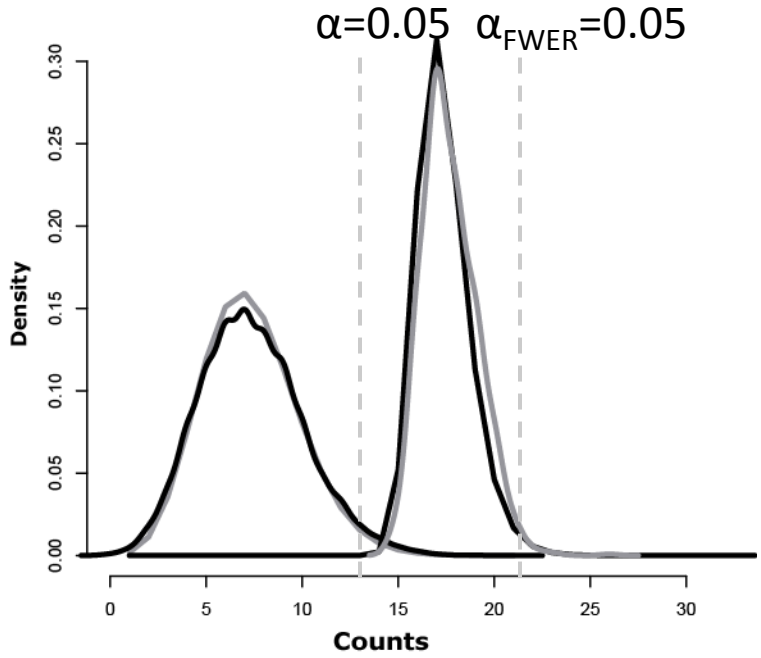
Given a region of size w and an observed read count n. What is the probability that one or more of the $3x10^9$ regions of size w has read count >= n under the null distribution?

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!

Scan distribution for a Poisson process

The probability of observing k reads on a window of size w in a genome of size L given a total of N reads can be approximated by (Alm 1983):

$$P(k|\lambda w, N, L) \approx 1 - F_p(k-1|\lambda w)e^{-\frac{k-w\lambda}{k}\lambda(T-w)P(k-1|\lambda w)}$$

where

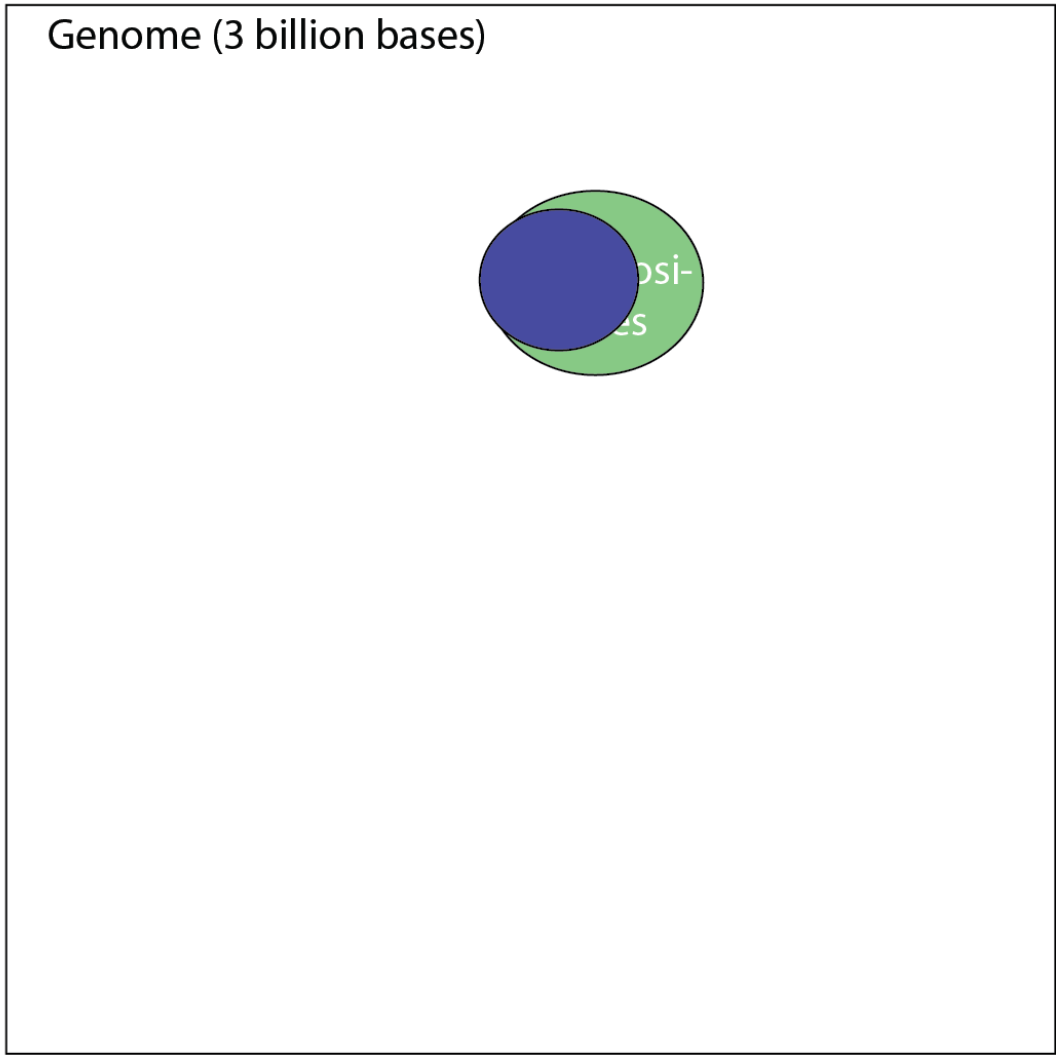
$P(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ counts given an expected count of λw

and

$F_p(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ or fewer counts given an expectation of λw reads

The scan distribution gives a computationally very efficient way to estimate the FWER

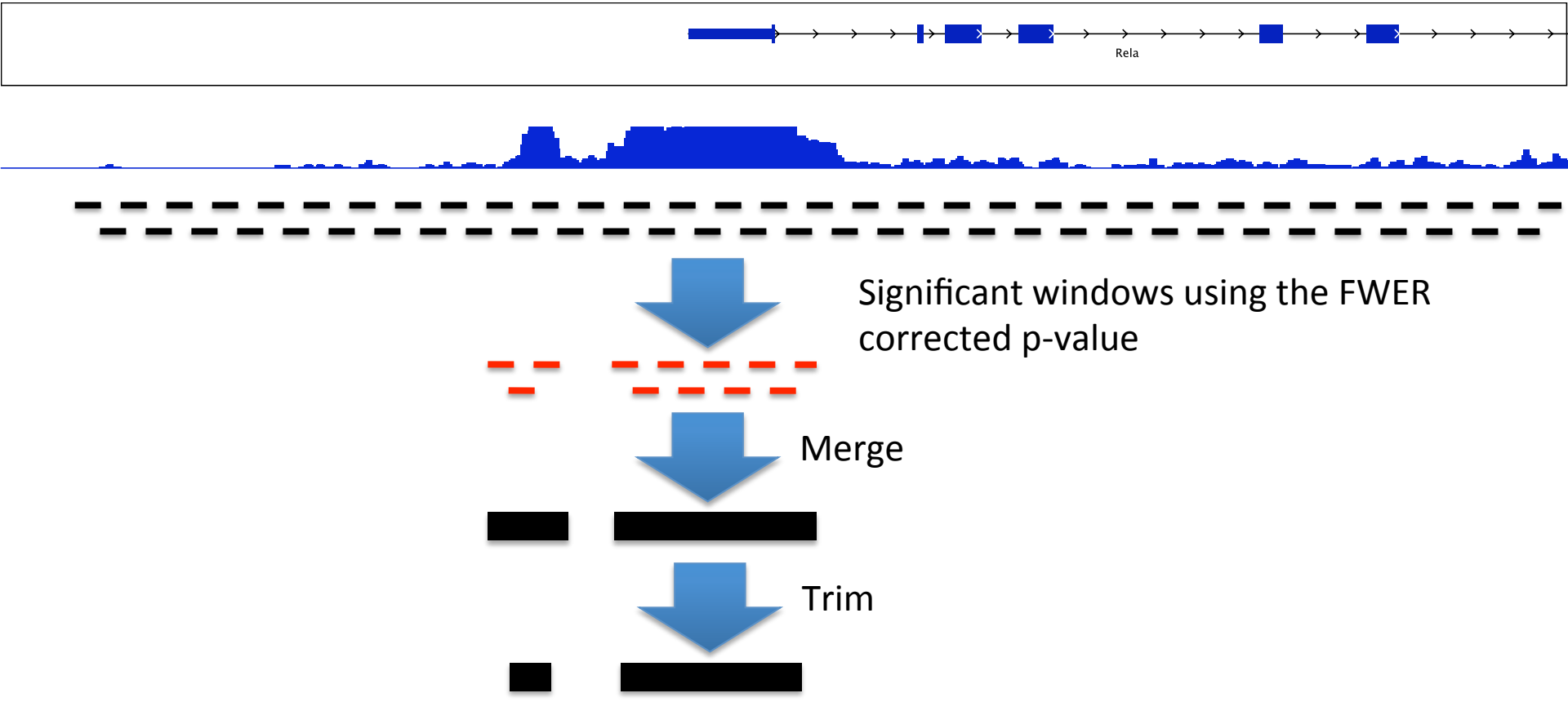
FWER-Scan Statistics



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

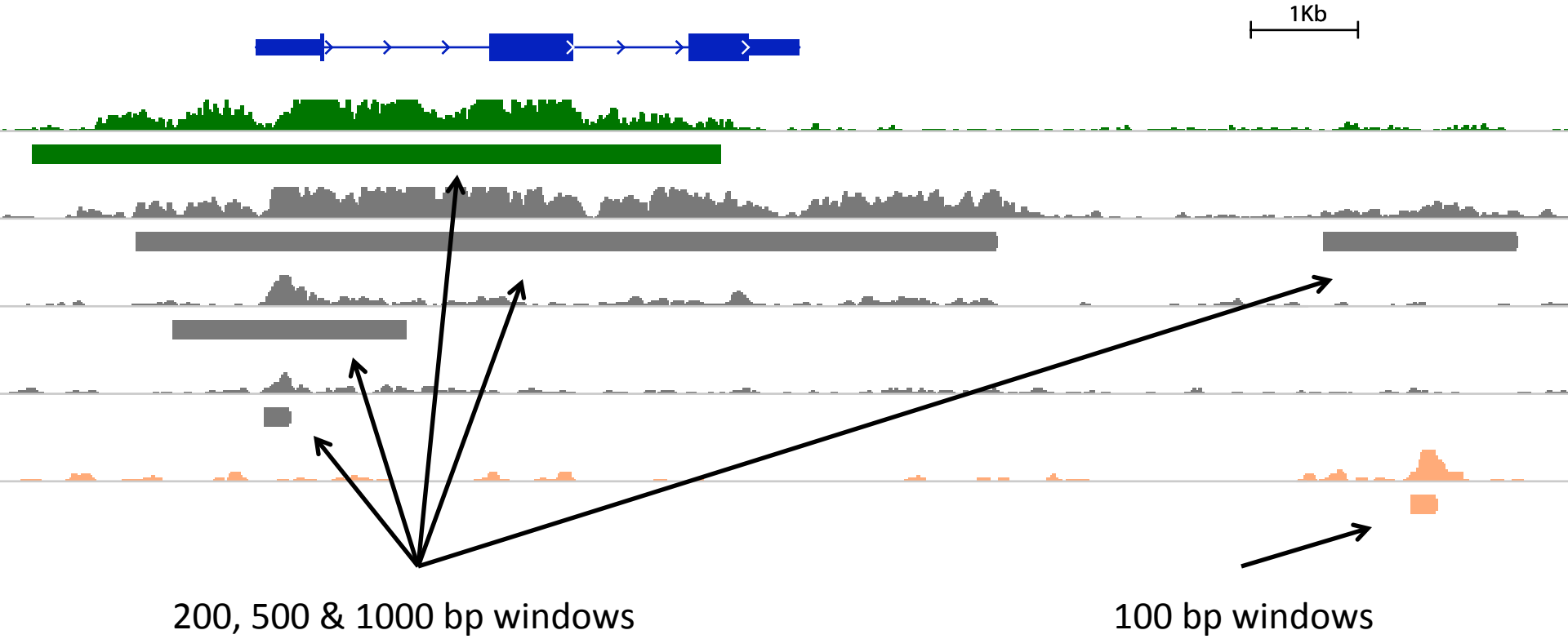


But, which window?

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

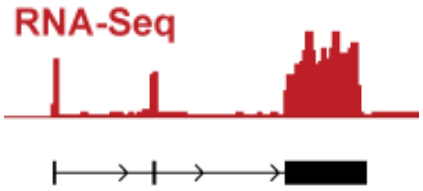


Can we identify enriched regions across different libraries?

Short modification ✓

Long modification ✓

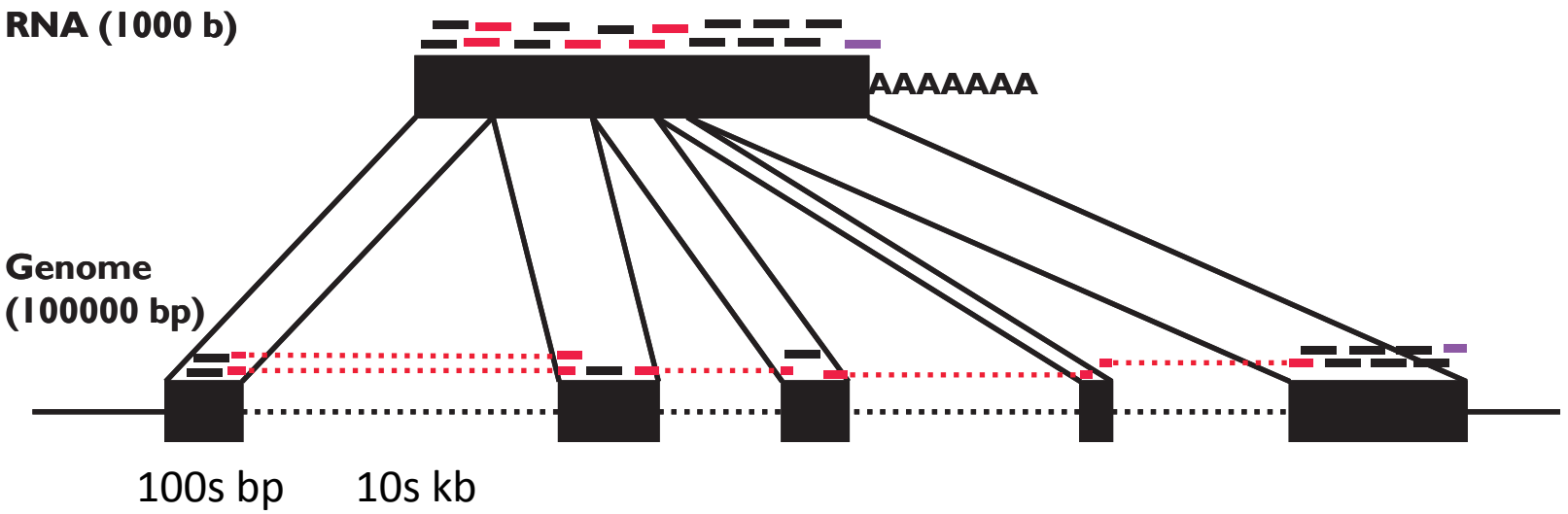
Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?



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

Scripture for RNA-Seq:
Extending segmentation to discontiguous regions

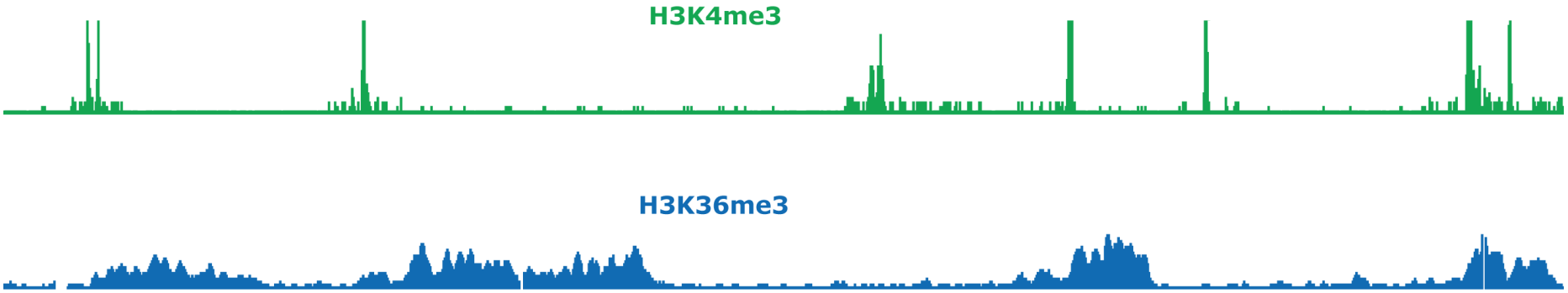
Transcript reconstruction problem as a segmentation problem



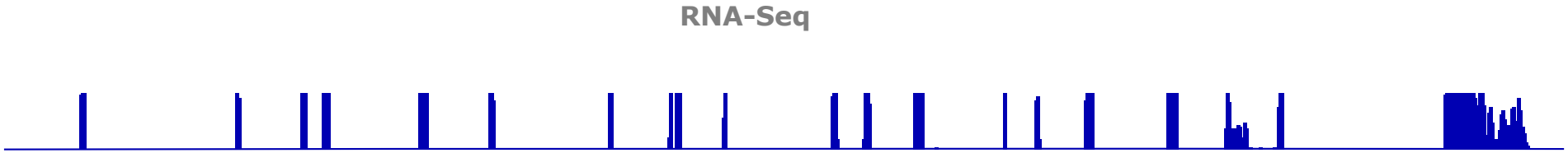
Challenges:

- 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.

Scripture: Genome-guided transcriptome reconstruction

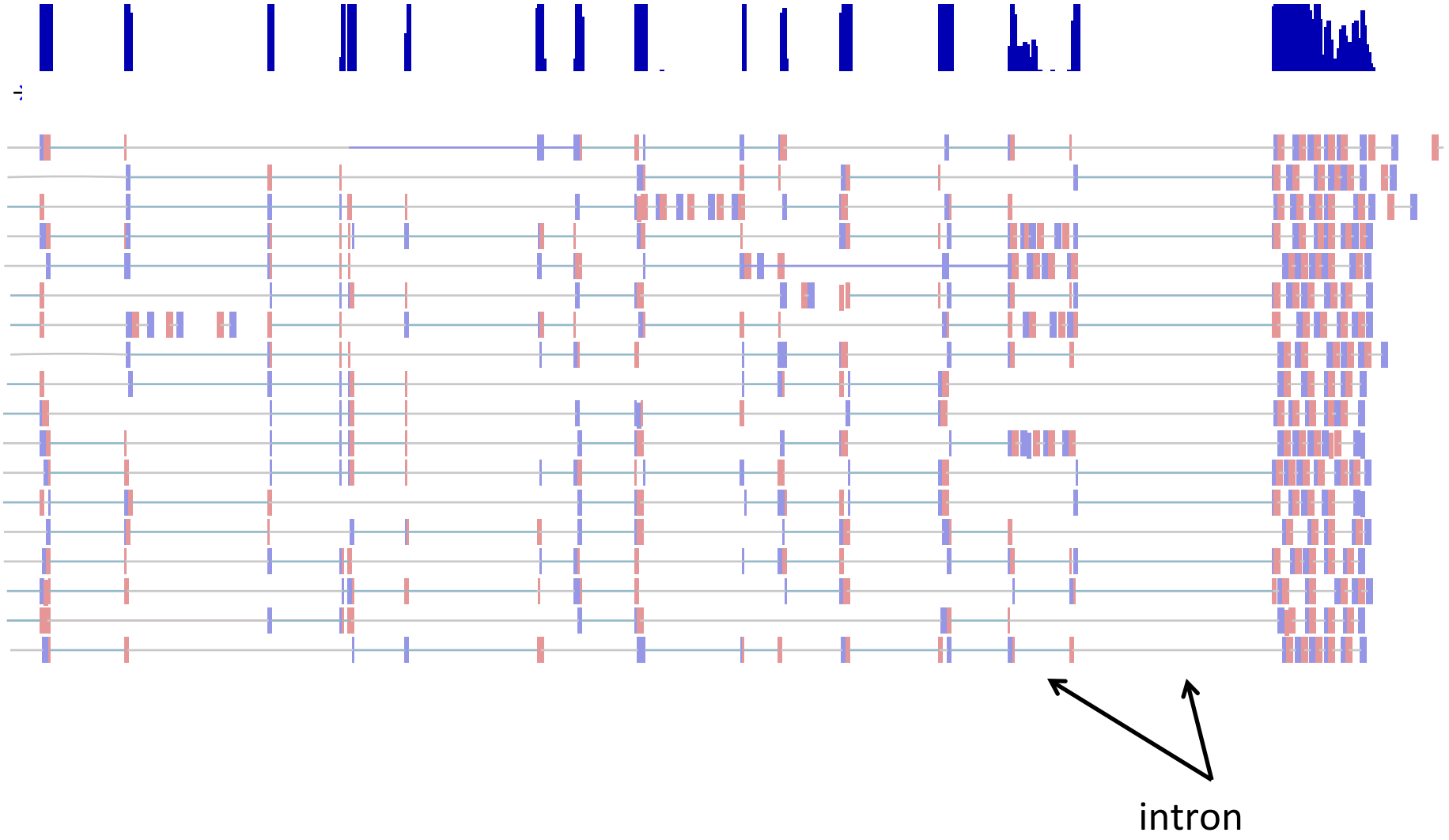


Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



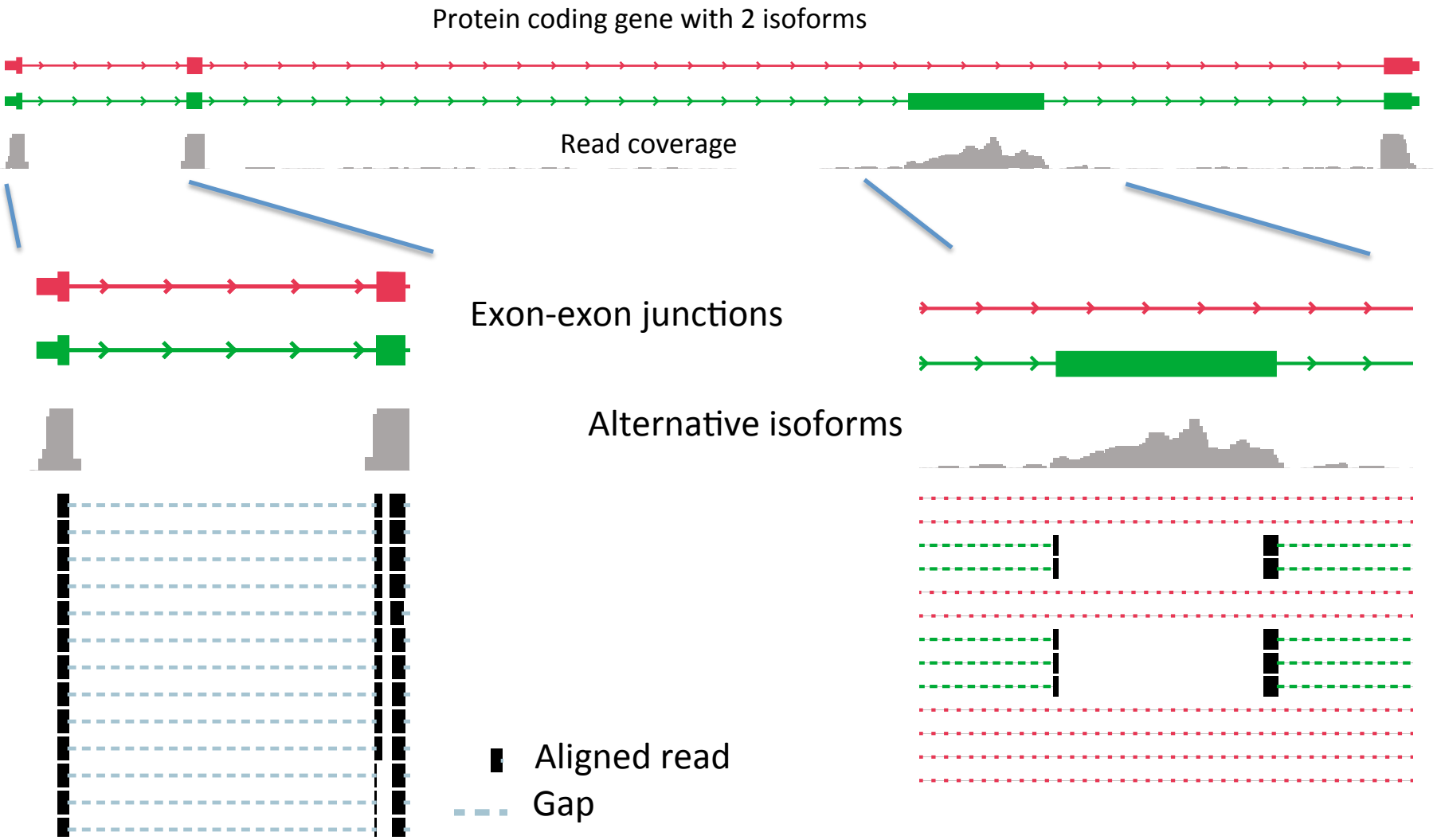
If we know the connectivity of fragments, we can increase our power to detect transcripts

Longer (76) reads increased number of junction reads



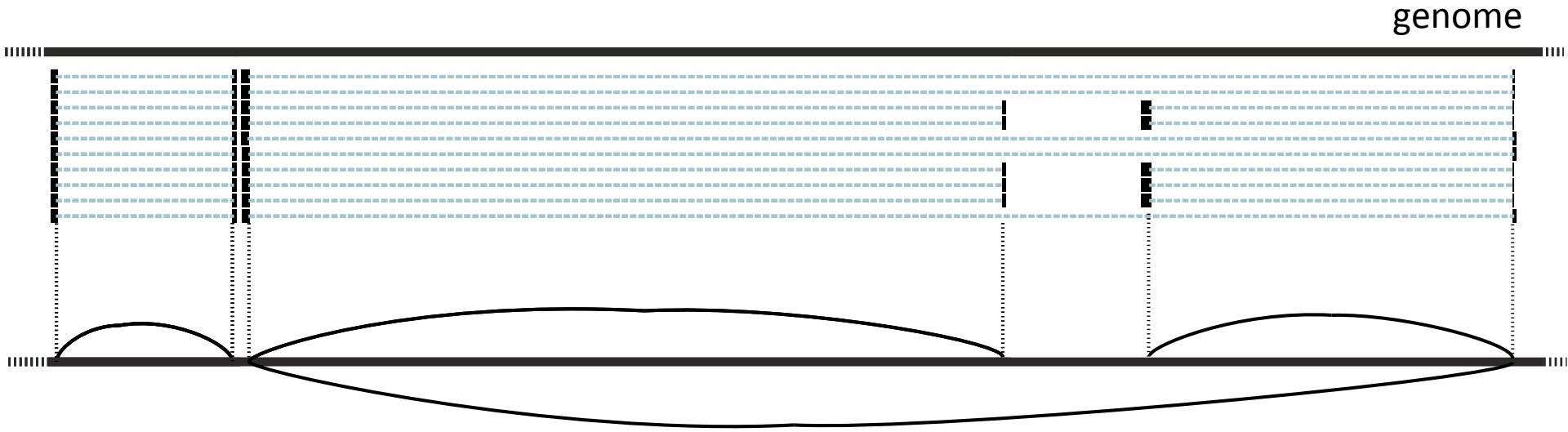
Exon junction spanning reads provide the connectivity information.

The power of spliced alignments



Statistical reconstruction of the transcriptome

Step 1: Align Reads to the genome allowing gaps flanked by splice sites

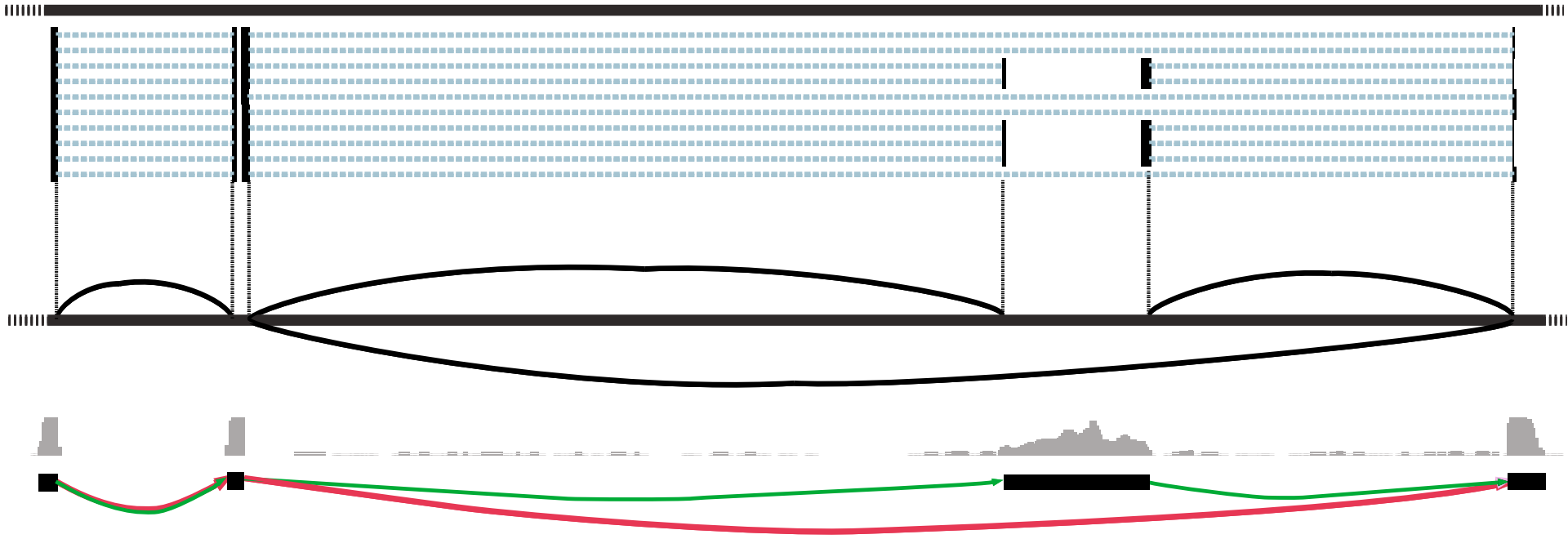


Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

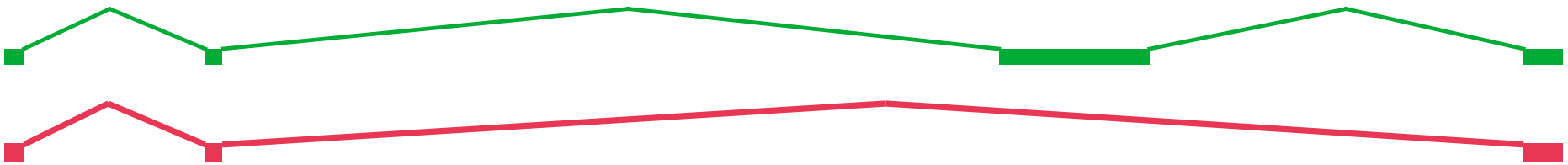
The "connectivity graph" connects all bases that are directly connected within the transcriptome

Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



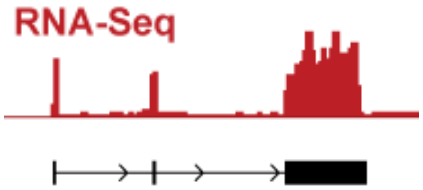
Step 4: Find significant segments



Can we identify enriched regions across different data types?

Short modification ✓

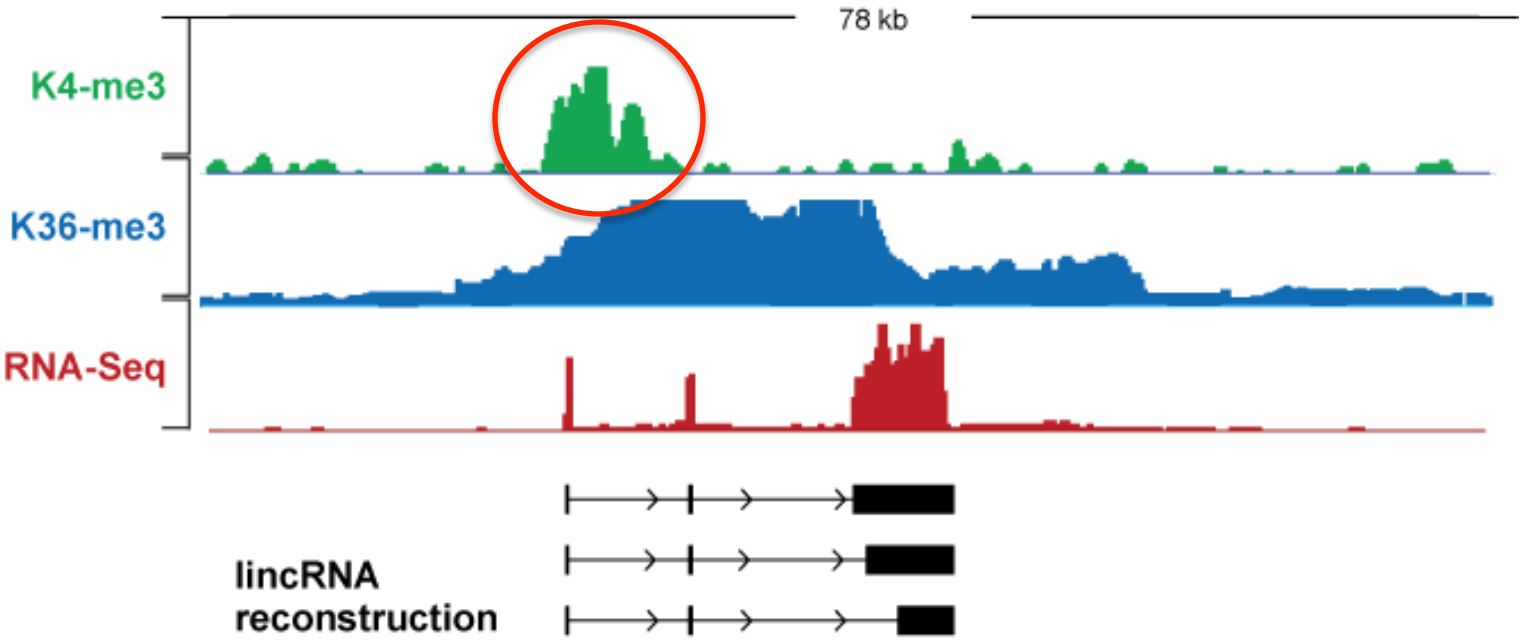
Long modification ✓



Discontinuous data ✓

Are we really sure reconstructions are complete?

RNA-Seq data is incomplete for comprehensive annotation

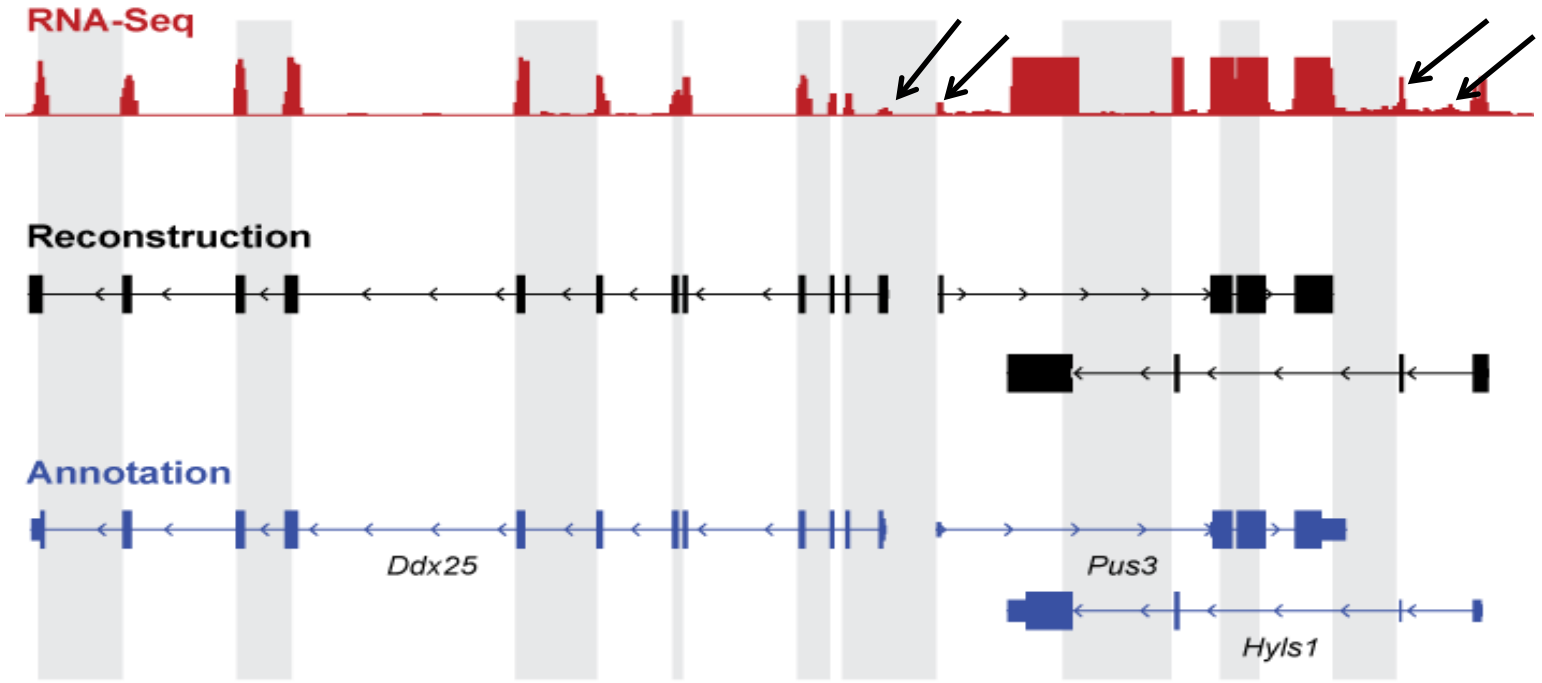


Library construction can help provide more information. More on this later

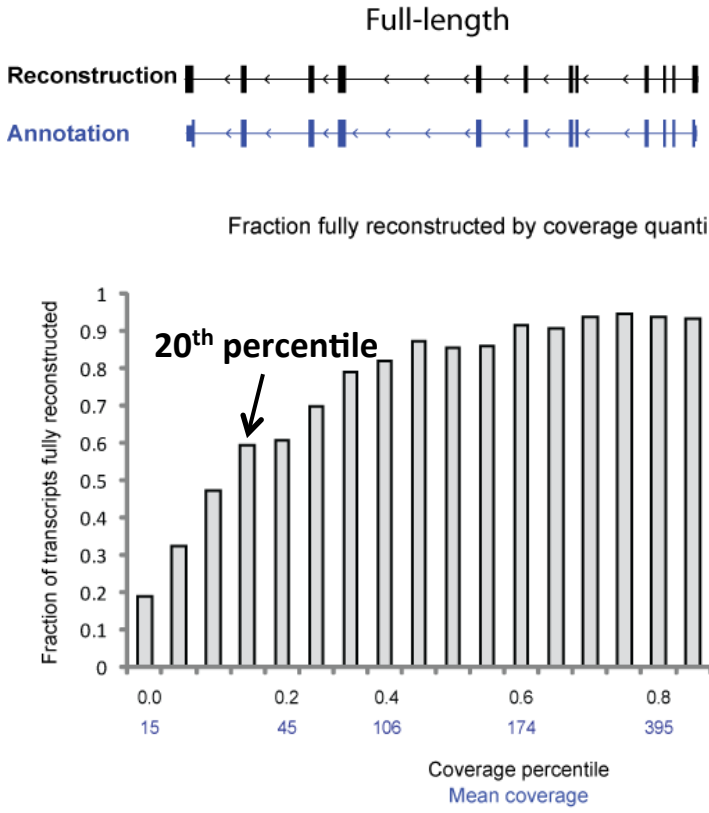
Applying scripture: Annotating the mouse transcriptome

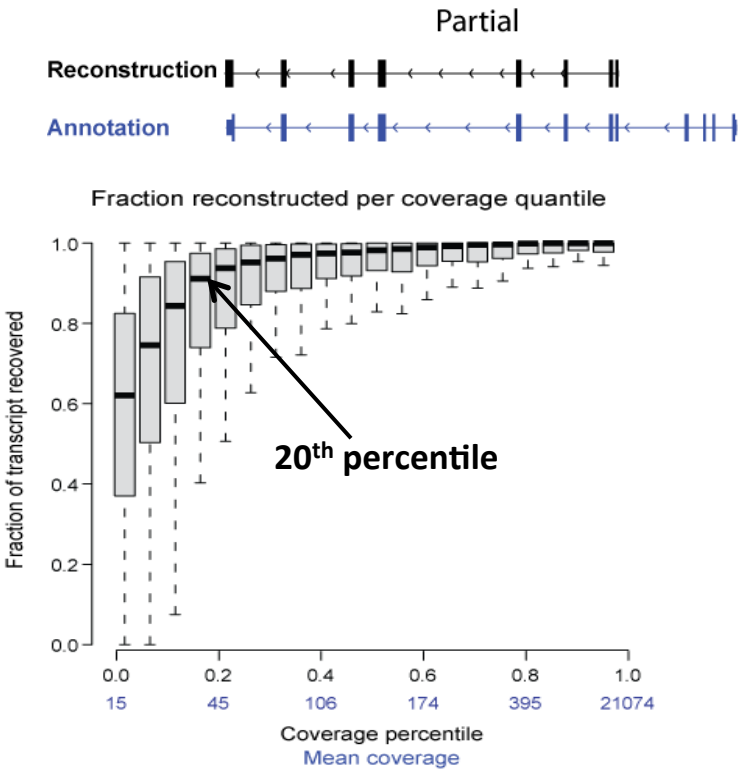
Reconstructing the mouse transcriptome (45M paired reads)

Mouse Cell Types → **Sequence** → **Reconstruct**



Sensitivity across expression levels

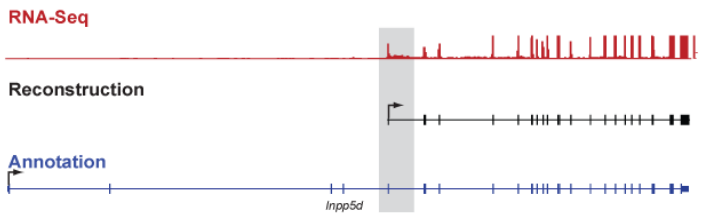




Even at low expression (20th percentile), we have:
average coverage of transcript is ~95% and 60% have full coverage

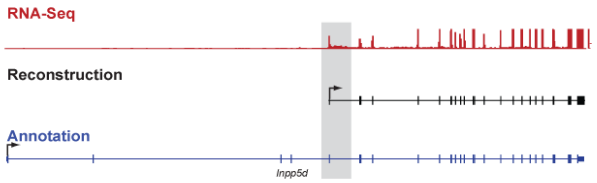
Novel variation in protein-coding genes

	ES cells	3 cell types
Novel 5’ Start Sites	**1,804**	**3,137**
	1,310	**2,477**
	588	**903**

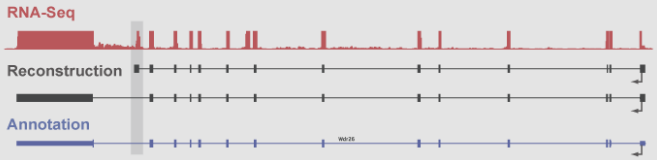


Novel variation in protein-coding genes

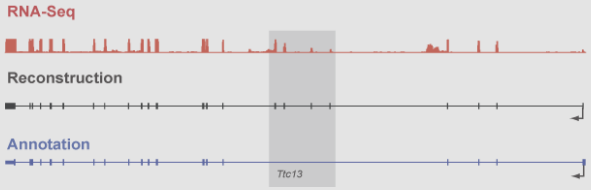
Novel 5’ Start Sites

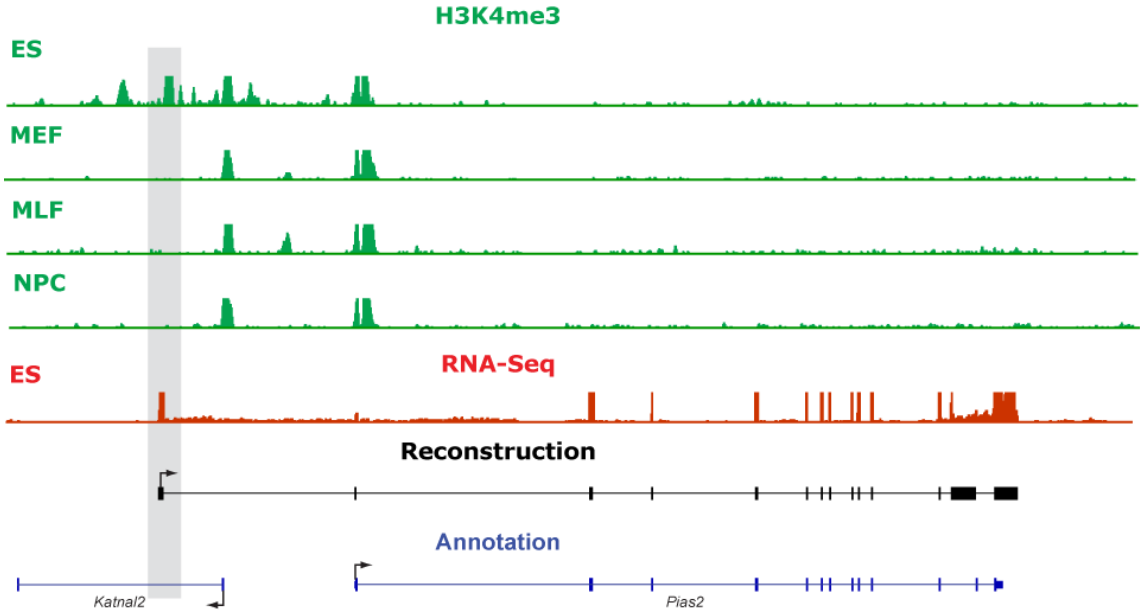


Novel 3’ End



Novel Coding Exons

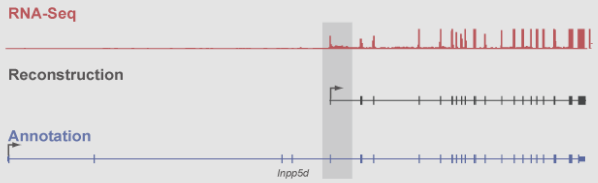




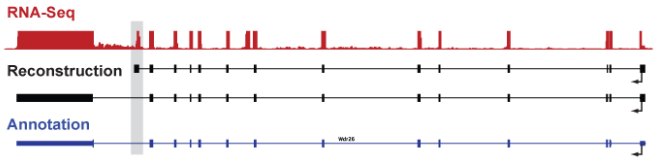
~85% overlap K4me3

Novel variation in protein-coding genes

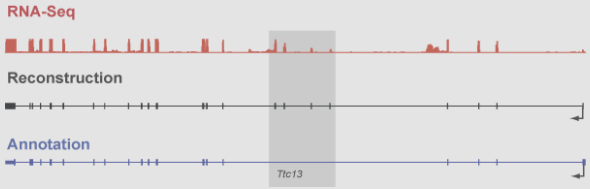
Novel 5’ Start Sites



Novel 3’ End

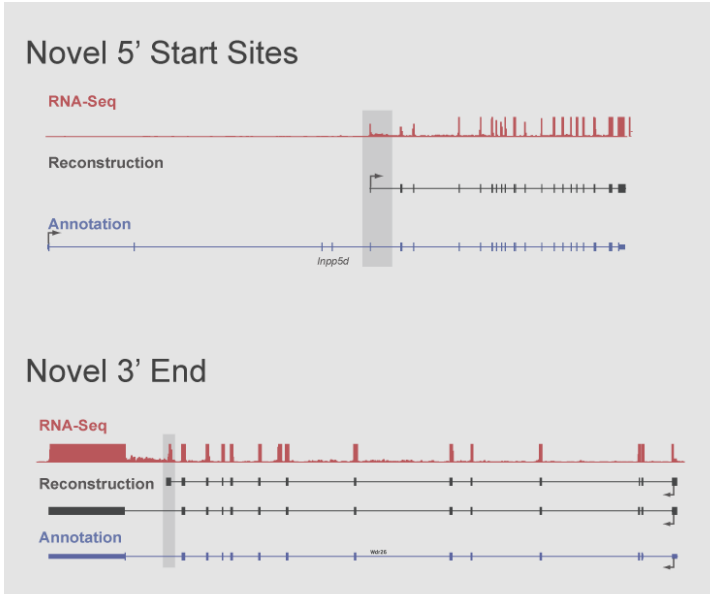


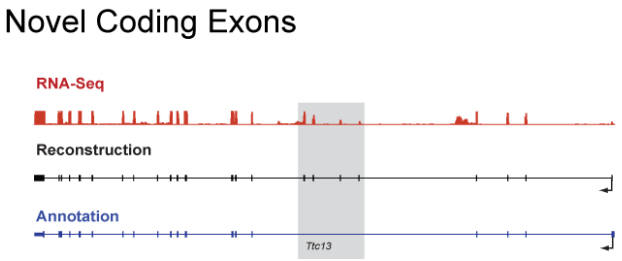
Novel Coding Exons

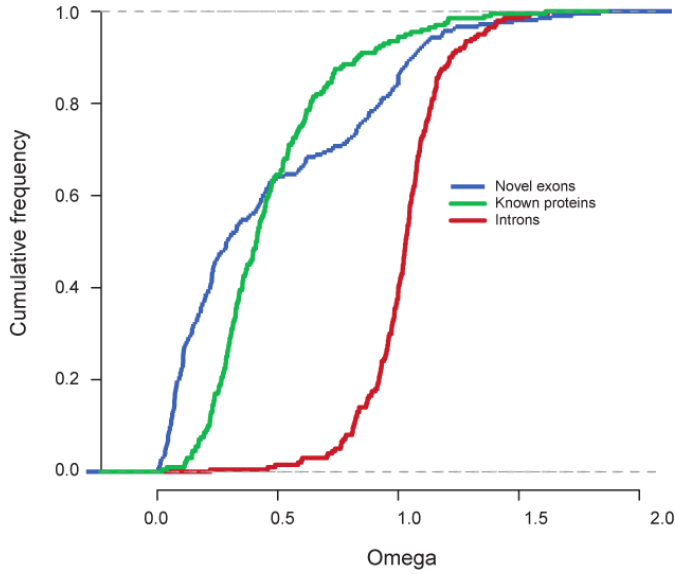


~50% contain polyA motif
Compared to ~6% for random

Novel variation in protein-coding genes







~80% retain ORF

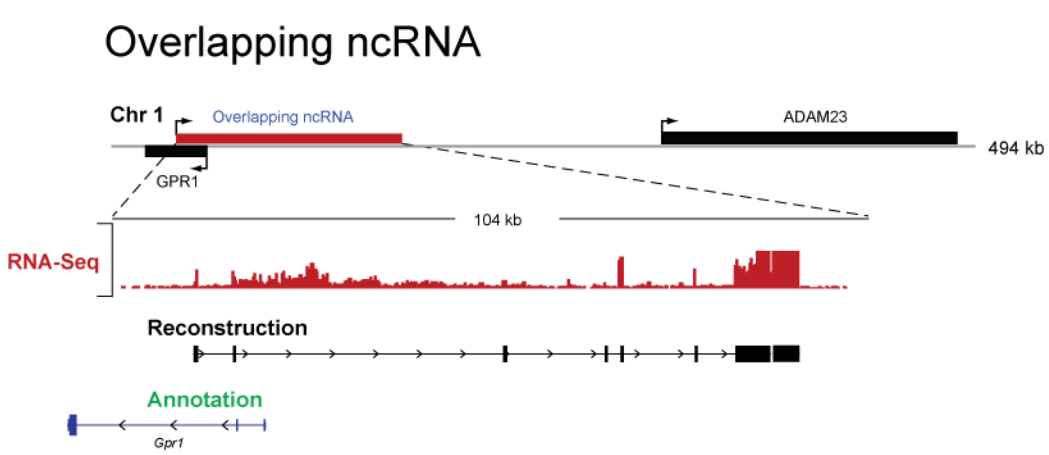
What about novel genes?

Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

Class 1: Overlapping ncRNA



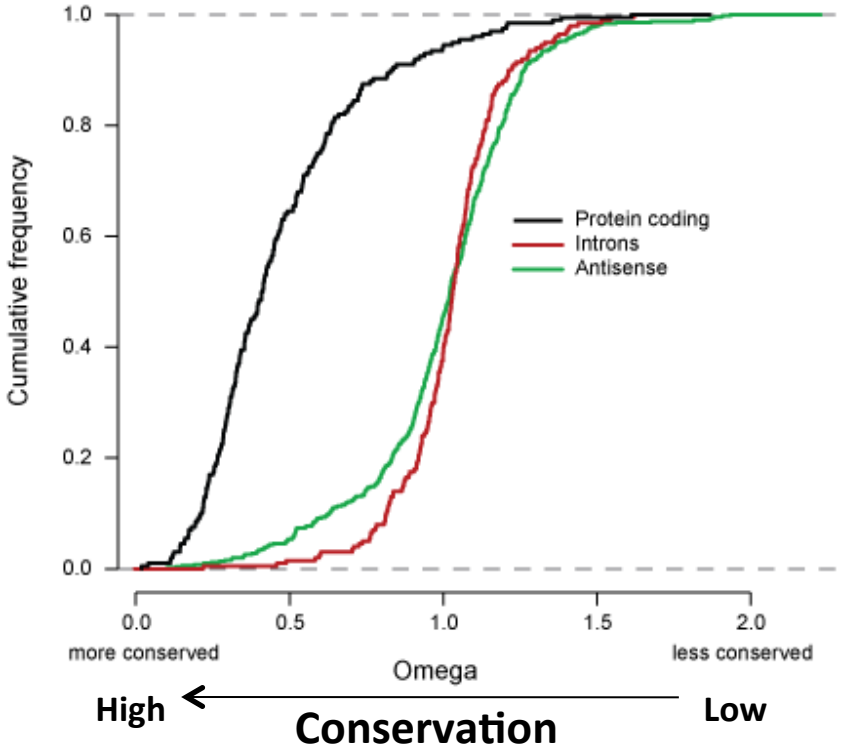
ES cells	3 cell types
201	**446**

Overlapping ncRNAs: low evolutionary conservation

Conserved

Neutral

SiPhy – (Garber et al. *Bioinformatics*, 2009)



Overlapping ncRNAs show little evolutionary conservation

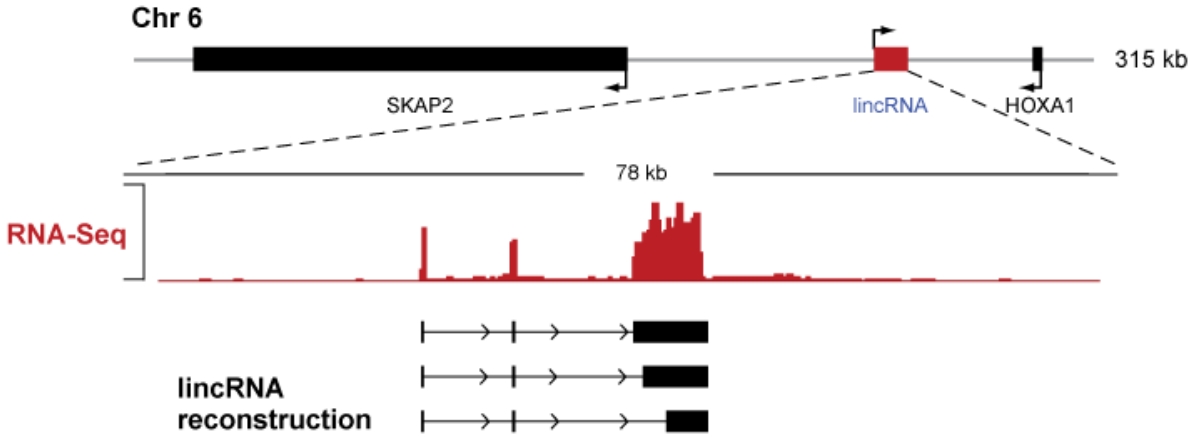
RNA (non-protein coding) Genes

Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

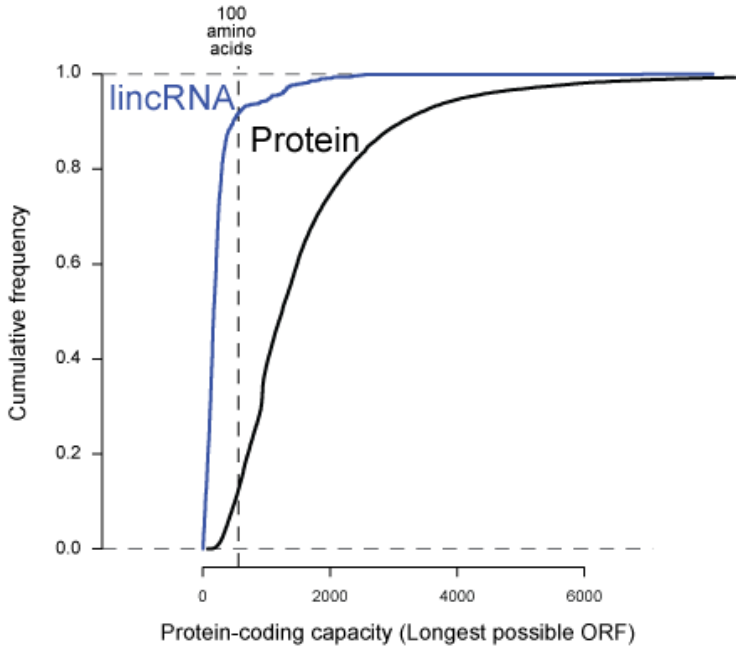
Class 2: Intergenic ncRNA (lincRNA)



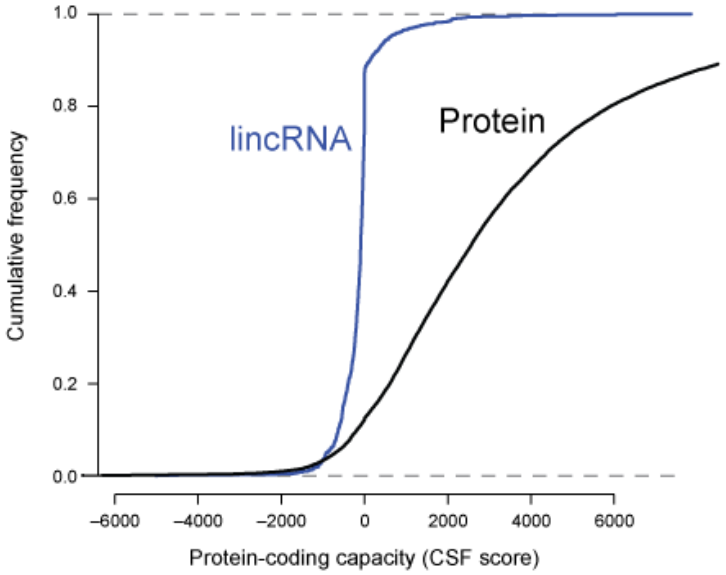
ES cells	3 cell types
~500	**~1500**

lincRNAs: How do we know they are non-coding?

ORF Length

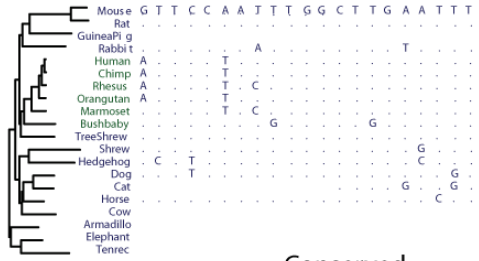


CSF (ORF Conservation)

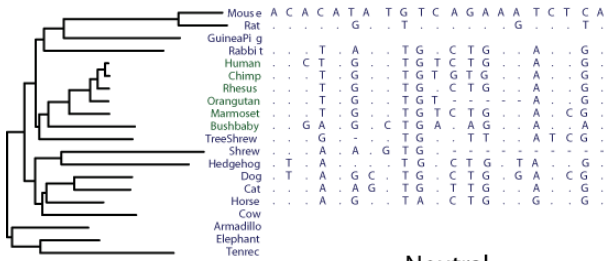


>95% do not encode proteins

lincRNAs: under slight constraint



Conserved



Neutral

High ← **Conservation** **Low**

What about novel coding genes?

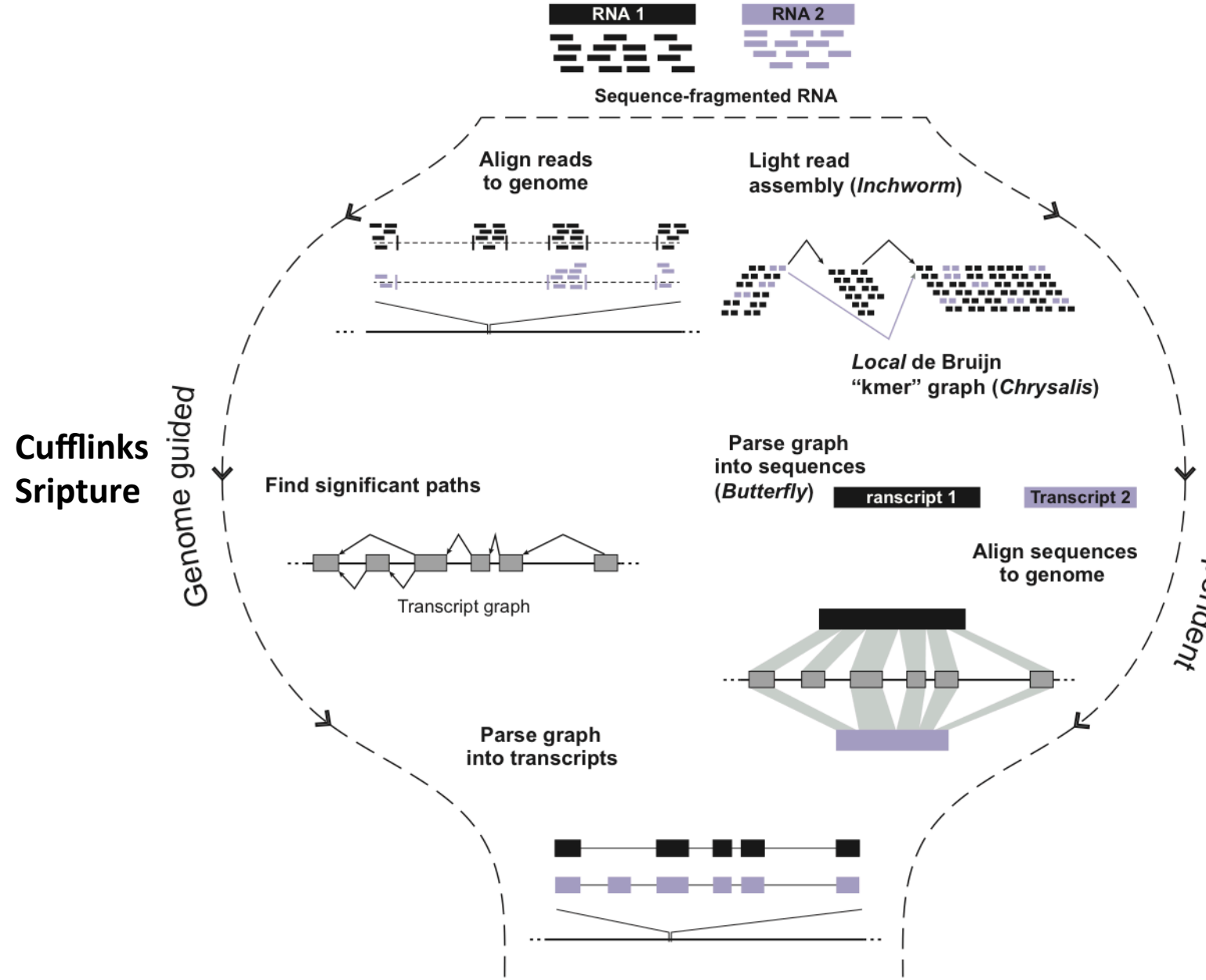
Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

~40 novel protein-coding genes

If there is no reference genome!
Genome independent methods



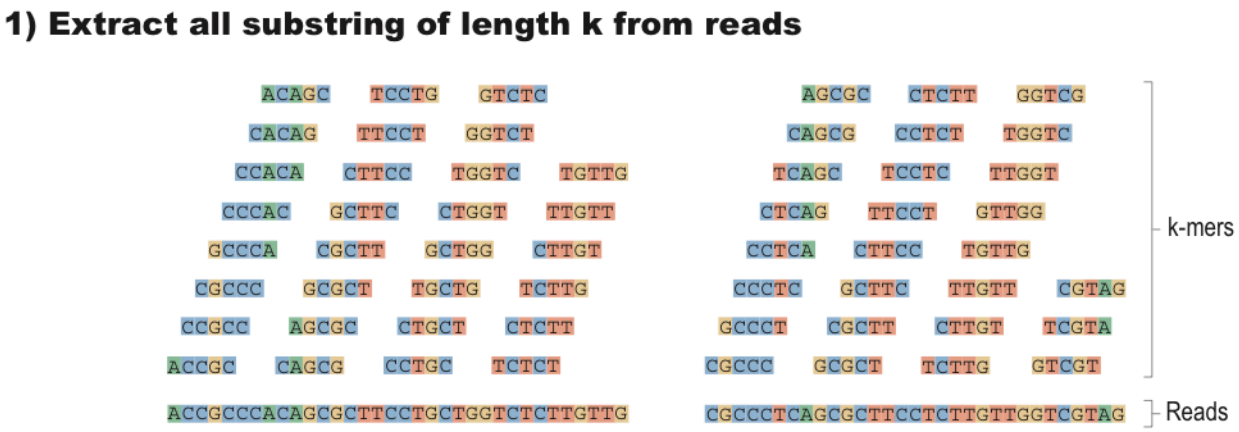
Rayan Chikhi lecture

Cufflinks
Sripture

Abyss
Trinity
Velvet

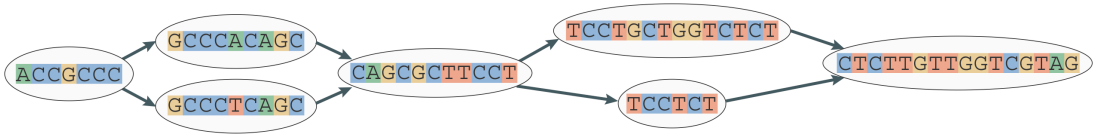
Garber et al, Nature Methods 2011

Assembly approach



Assembly approach

3) Collapse graph



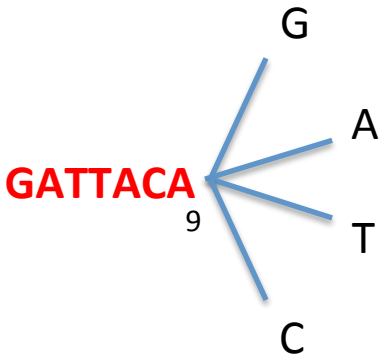
But this challenging already with DNA and RNA has many different challenges

The Trinity approach: Localize

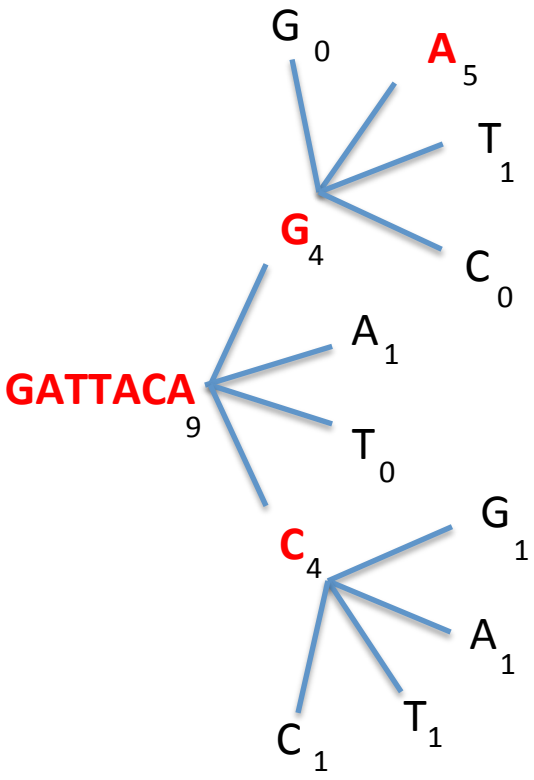
Decompose all reads into overlapping Kmers (25-mers)

Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3' end, guided by coverage.

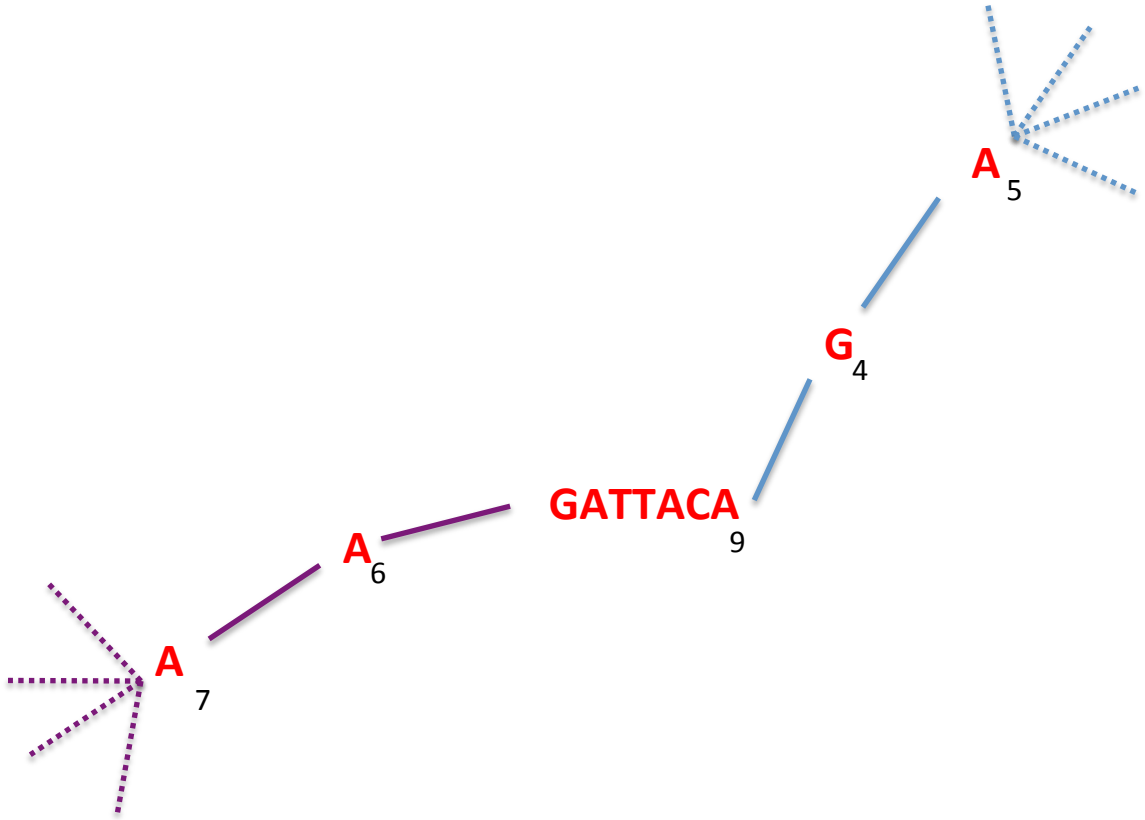


The Trinity approach: Localize



Briah Haas

The Trinity approach: Localize



Report contig: ....AAGATTACAGA....

Remove assembled kmers from catalog, then repeat the entire process.

Briah Haas

Trinity approach: Assemble

RNA-Seq reads

Group similar contigs

key: localize the assembly problem

Pros and cons of each approach

- Transcript assembly methods are the obvious choice for organisms without a reference sequence.
- Genome-guided approaches are ideal for annotating high-quality genomes and expanding the catalog of expressed transcripts and comparing transcriptomes of different cell types or conditions.
- Hybrid approaches for lesser quality or transcriptomes that underwent major rearrangements, such as in cancer cell.
- More than 1000 fold variability in expression leves makes assembly a harder problem for transcriptome assembly compared with regular genome assembly.
- Genome guided methods are very sensitive to alignment artifacts.

RNA-Seq transcript reconstruction software

Assembly	Genome Guided
Oasis (velvet)	Cufflinks
Trans-ABySS	Scripture
Trinity	

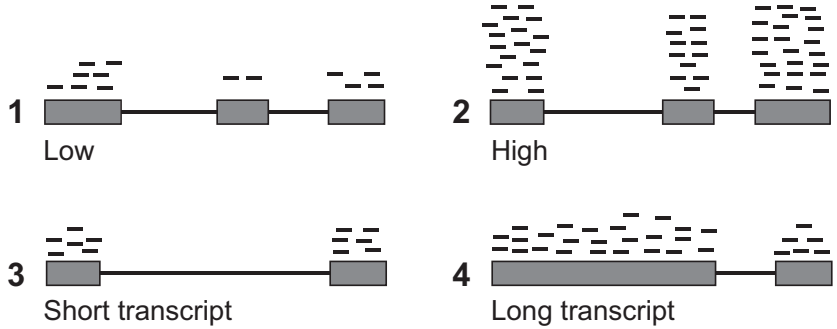
Analysis of counting data requires 3 broad tasks

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

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 > gene A in sample 2?

Quantification: only one isoform



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

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

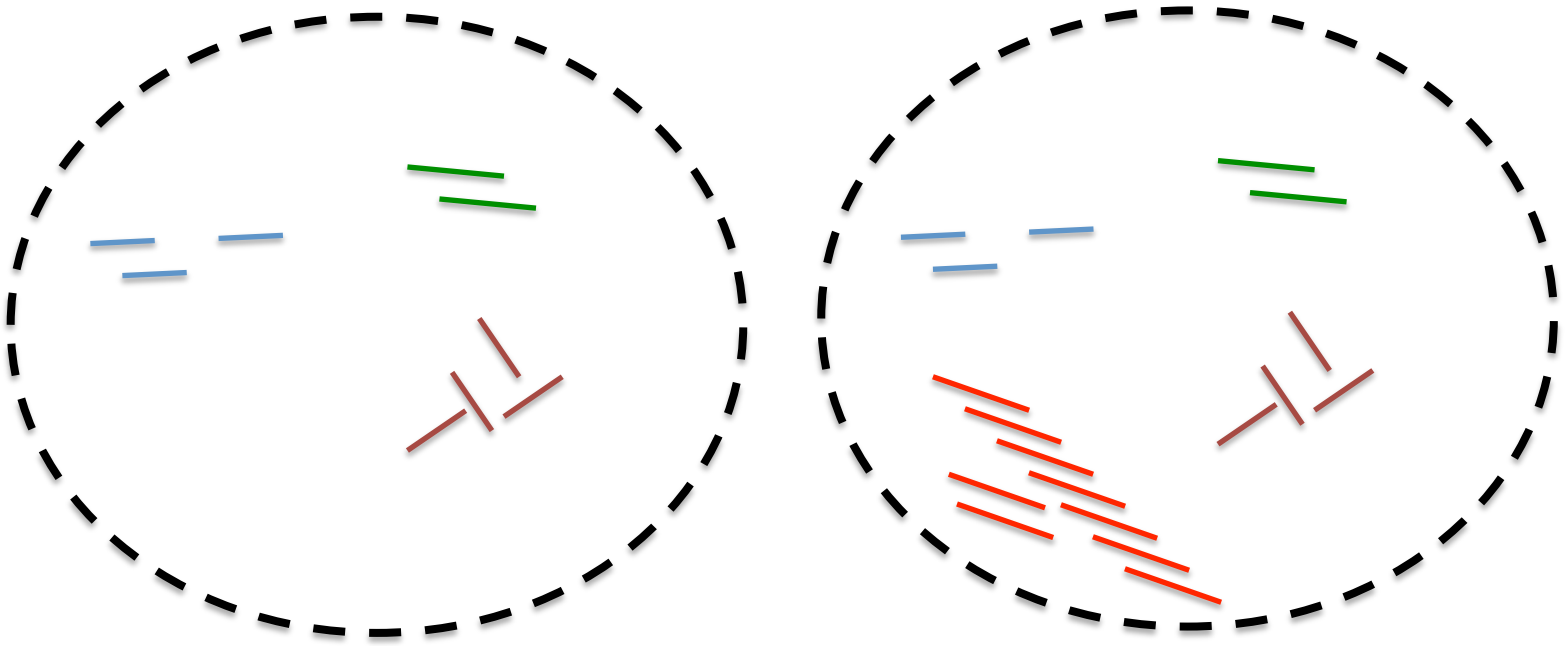
- Fragmentation of transcripts results in length bias: longer transcripts have higher counts
- Different experiments have different yields. Normalization is key for cross lane comparisons

Complexity increases when multiple isoforms exist

Normalization depends on the application

- To compare within a sequence run (lane), RPKM accounts for length bias.
- RPKM is not optimal for cross experiment comparisons.
  - Different samples may have different compositions.

Step 2: Different RNA compositions

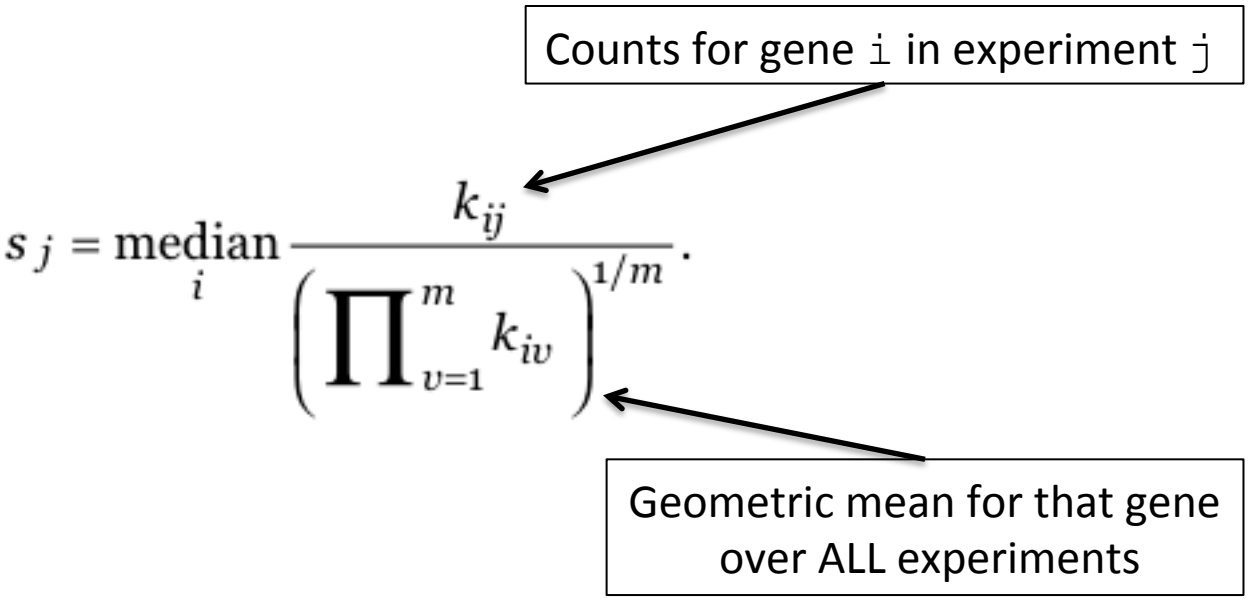


Cell type I

Cell type II

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

Step2: More robust normalization



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)

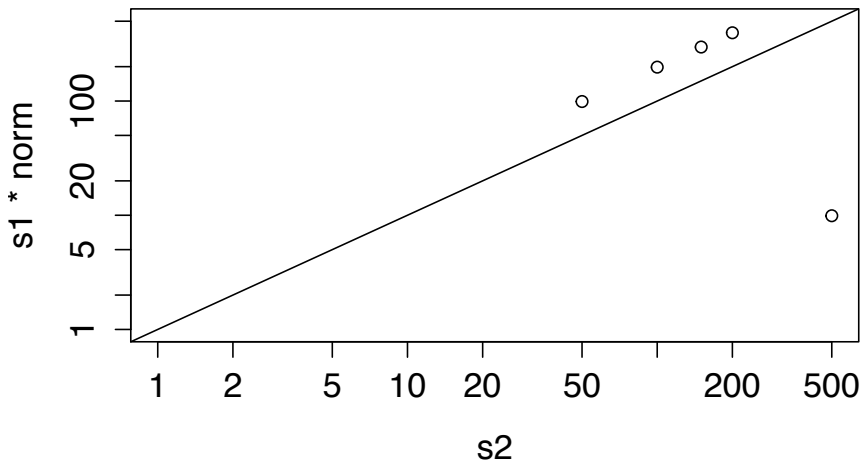
$> s1 = c(100, 200, 300, 400, 10)$
$> s2 = c(50, 100, 150, 200, 500)$

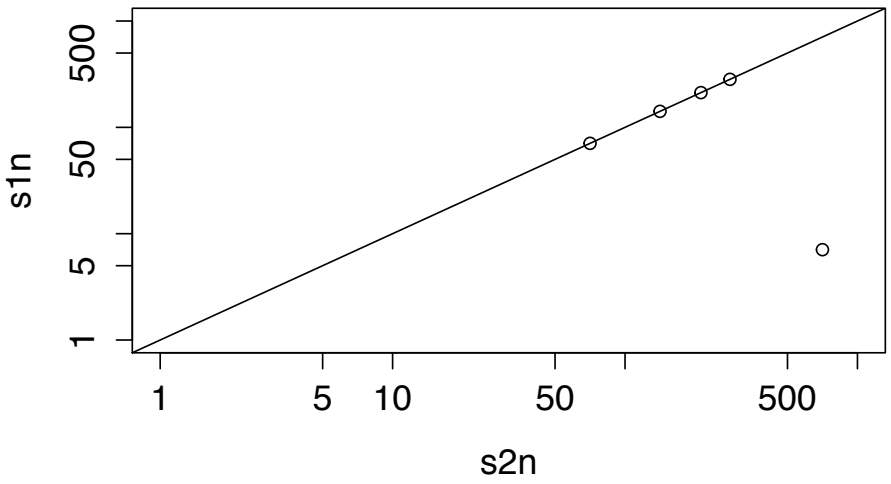
Similar read number,
one transcript many fold changed

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

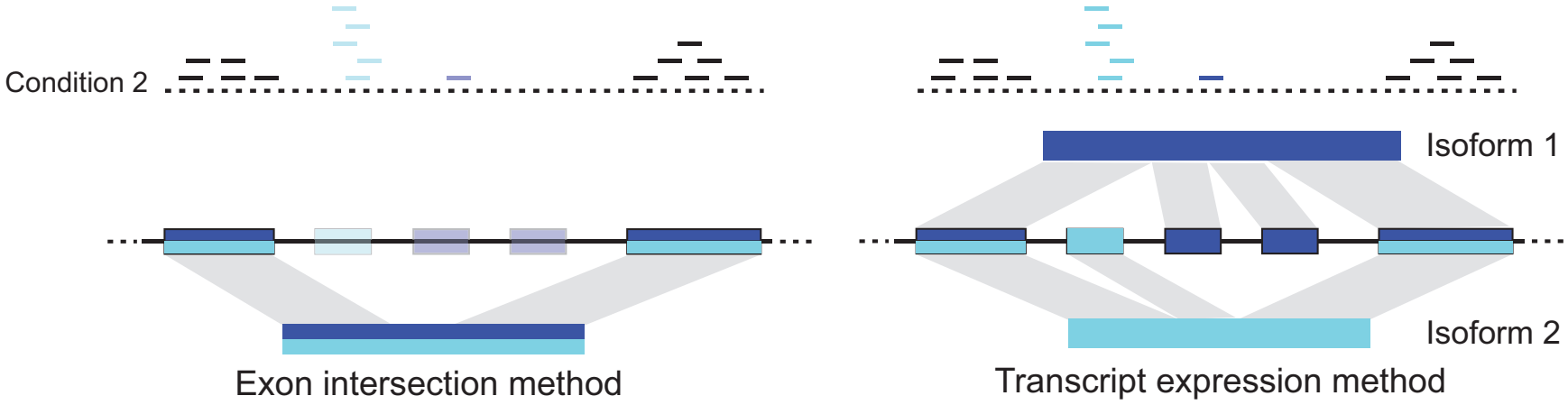
Size normalization results in 2-fold changes in *all* transcripts

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





But, how to compute counts for complex gene structures?



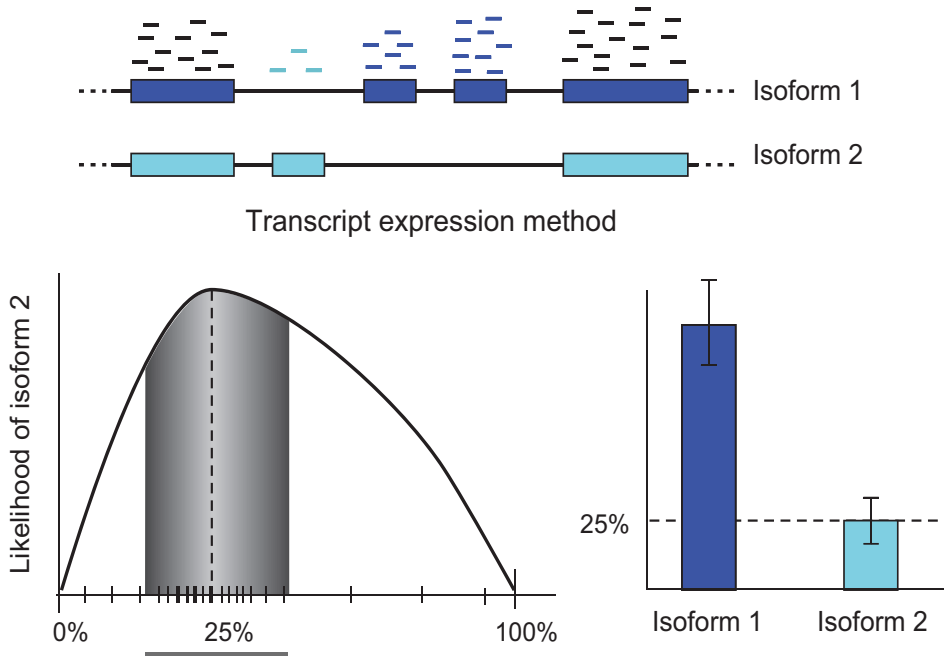
Three popular options:

Exon *intersection* model: Score constituent exons

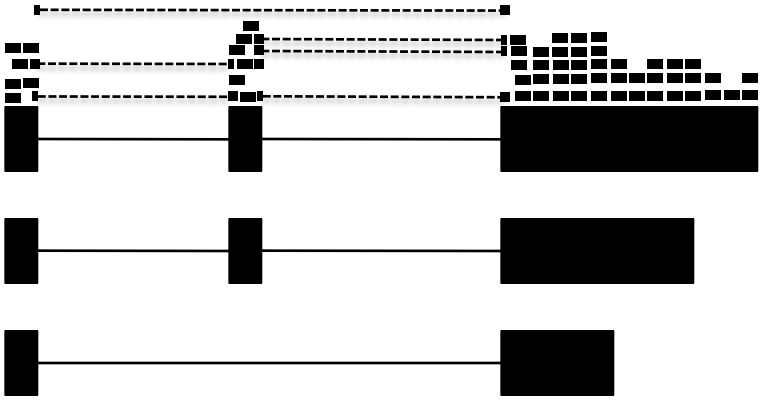
Exon *union* model: Score the the "merged" transcript

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

Quantification: read assignment method

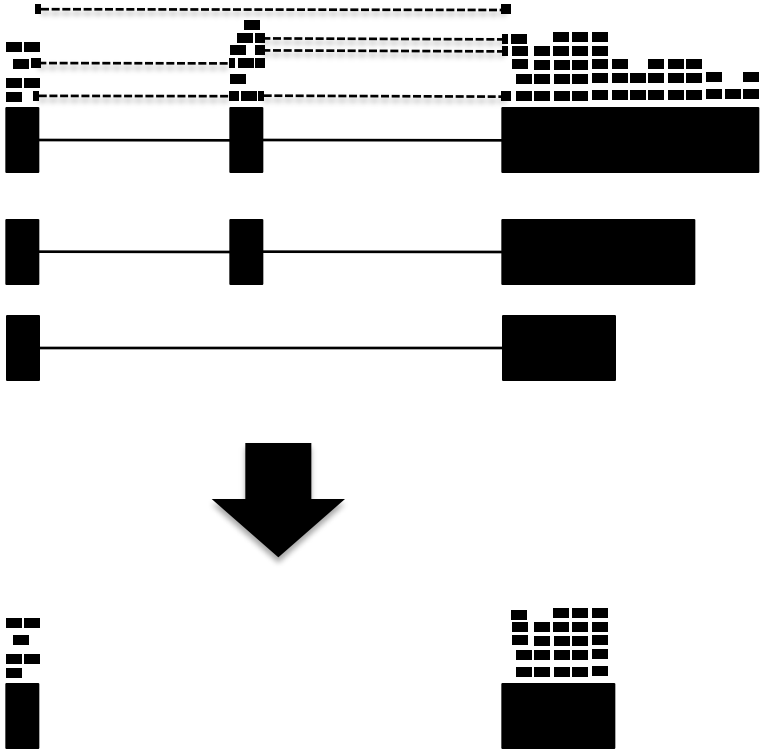


Quantification with multiple isoforms



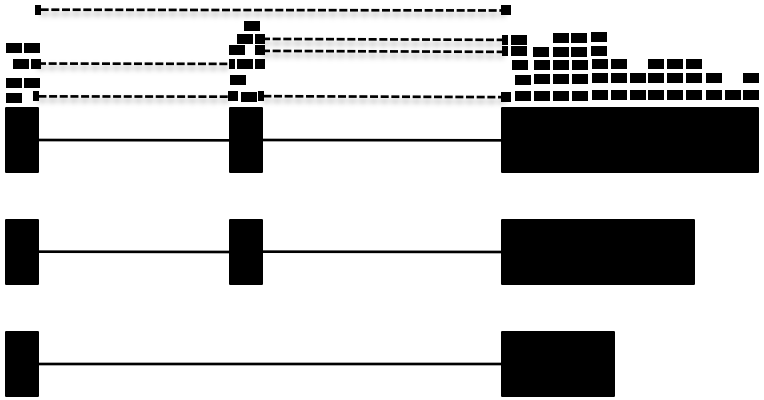
How do we define the gene expression?
How do we compute the expression of each isoform?

Computing gene expression

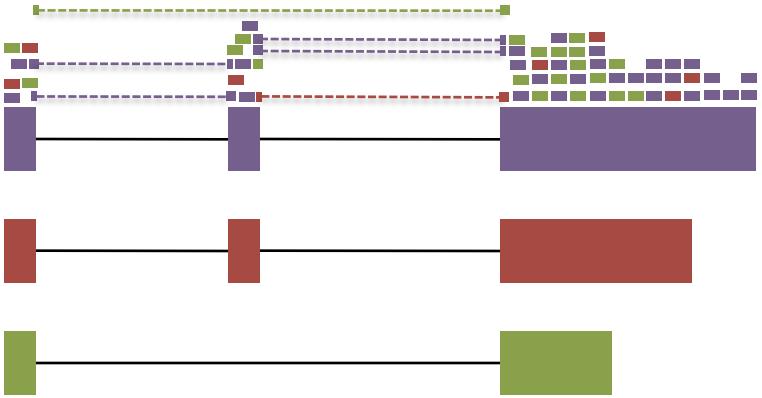


Idea1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)

Computing gene expression — isoform deconvolution



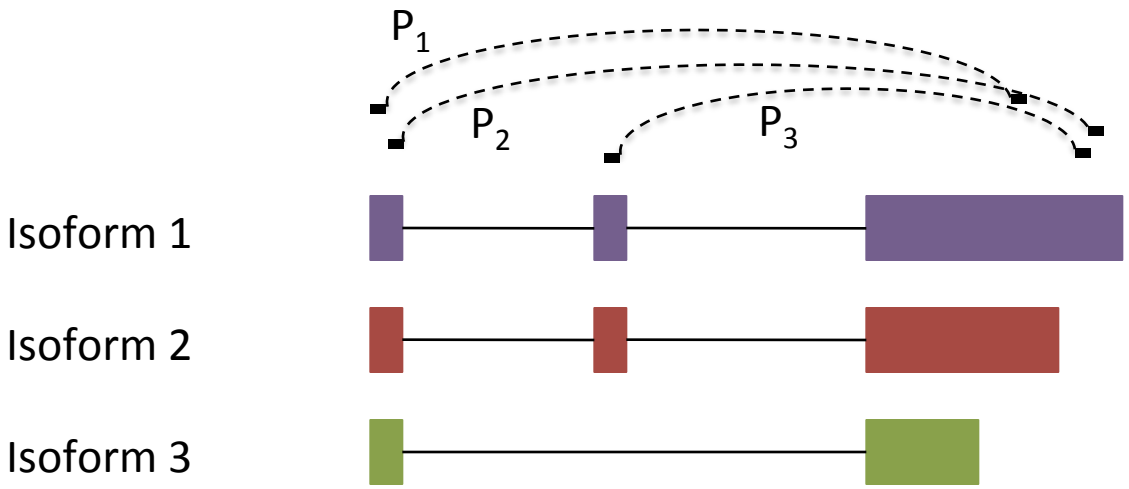
Computing gene expression — isoform deconvolution



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.

$$E = RPKM_1 + RPKM_2 + RPKM_3$$

Paired-end reads are easier to associate to isoforms

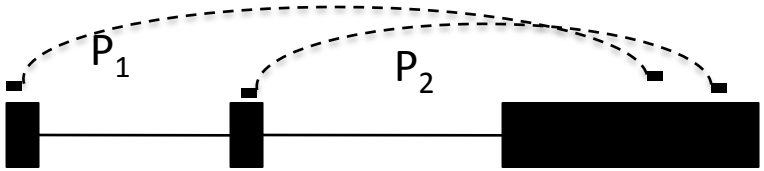


Paired ends increase isoform deconvolution confidence

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

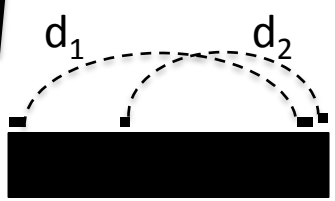
Do paired-end reads also help identifying reads originating in isoform 3?

We can estimate the insert size distribution

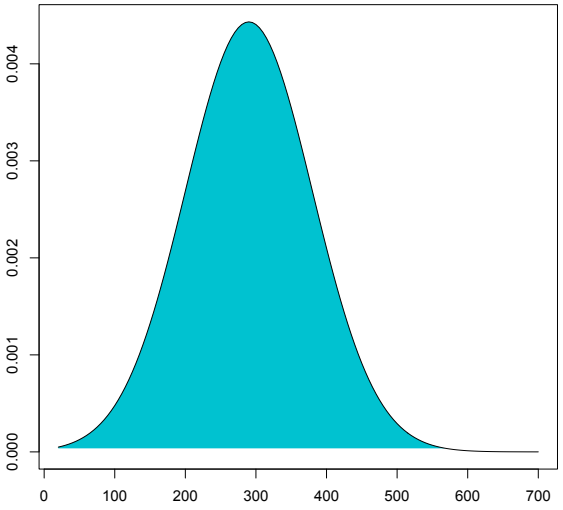


Get all single isoform reconstructions

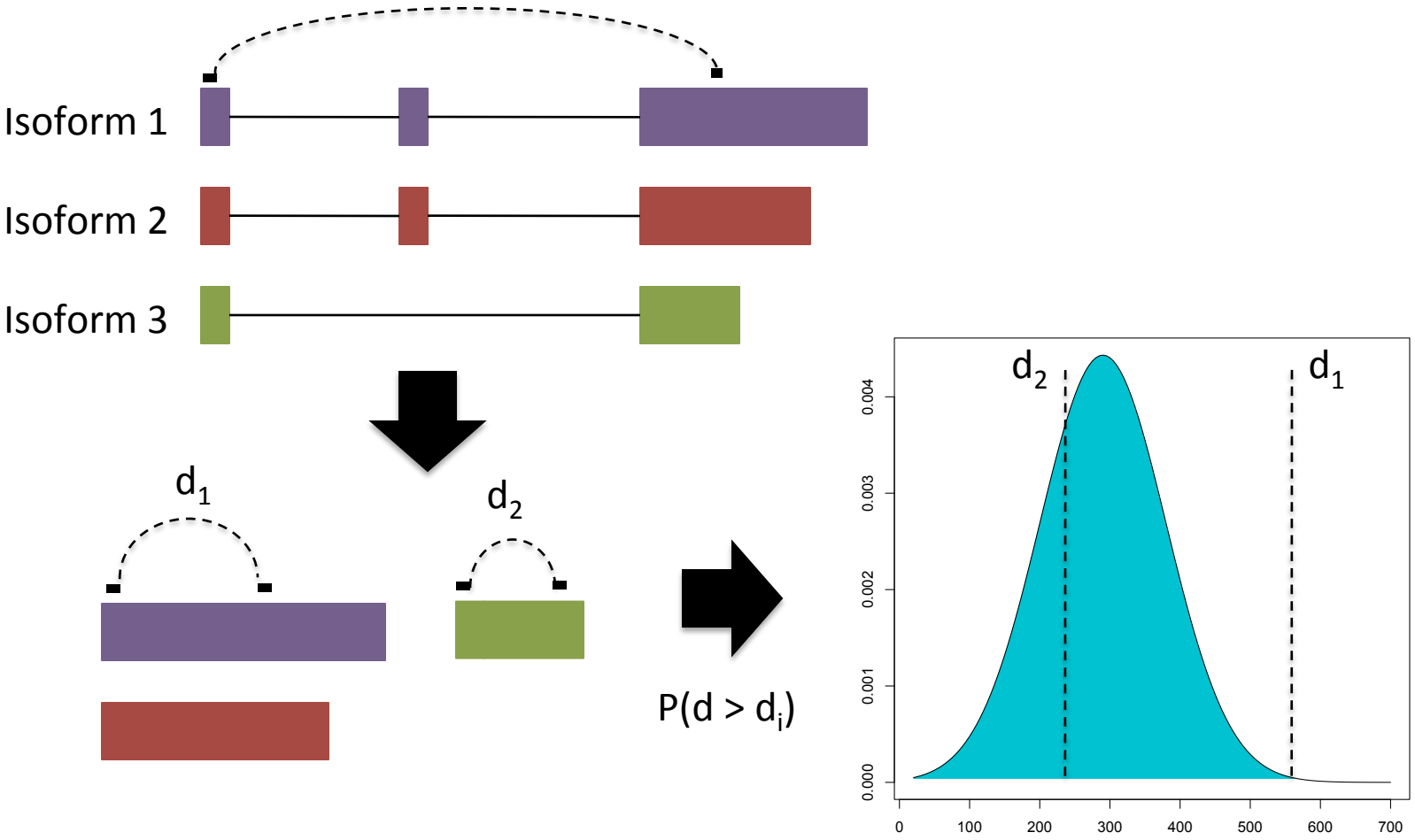
Splice and compute insert distance



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

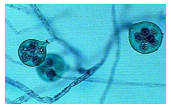
RNA-Seq quantification summary

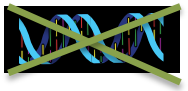
- Counts must be estimated from ambiguous read/ transcript assignment.
  - Using simplified gene models (intersection)
  - Probabilistic read assignment
- Counts must be normalized
  - RPKM is sufficient for intra-library comparisons
  - More sophisticated normalizations to account for differences in library composition for inter-library comparisons.

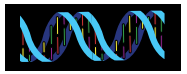
Programs to measure transcript expression

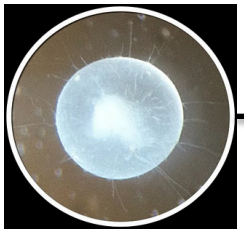
	Implemented method
Alexa-seq	Gene expression using intersection model
ERANGE	Gene expression using union model
Scripture	Gene expression using intersection model
Cufflinks	**Transcript deconvolution by solving the maximum likelihood problem**
MISO	**Transcript deconvolution by solving the maximum likelihood problem**
RSEM	**Transcript deconvolution by solving the maximum likelihood problem**

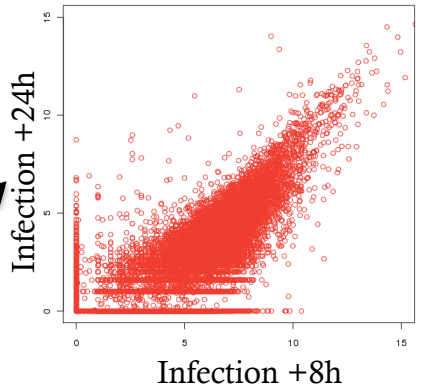
Advantages of RSEM, DESeq



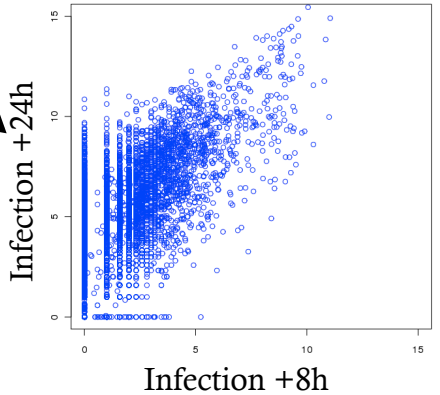








Trout



Saprolegnia

Infected fish cells

Brian Haas, Rays Jiang, Carsten Russ

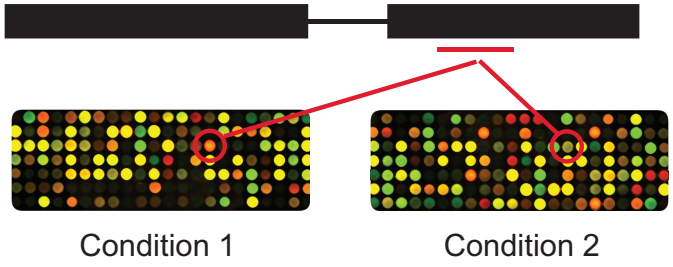
Analysis of counting data requires 3 broad tasks

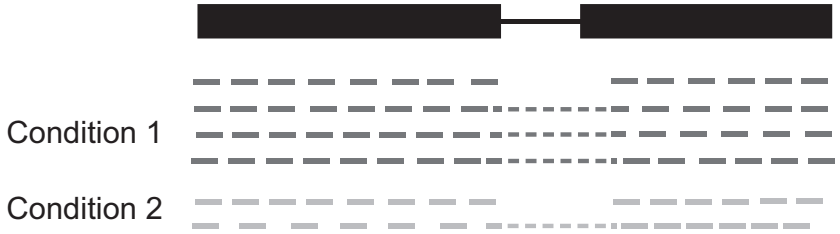
- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

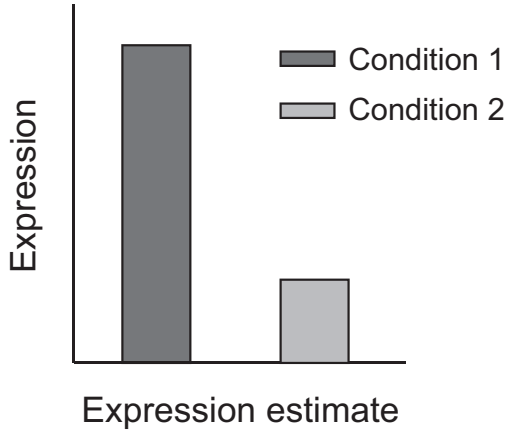
Differential Gene Expression Questions

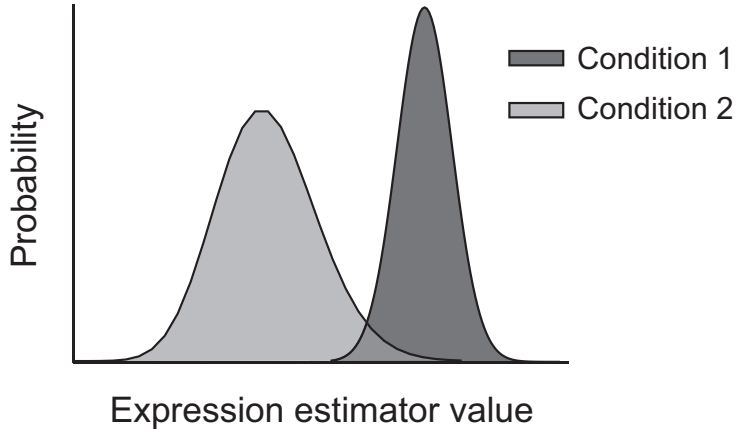
- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3’ UTRs

Differential gene expression using RNA-Seq









•(Normalized) read counts ←→ Hybridization intensity

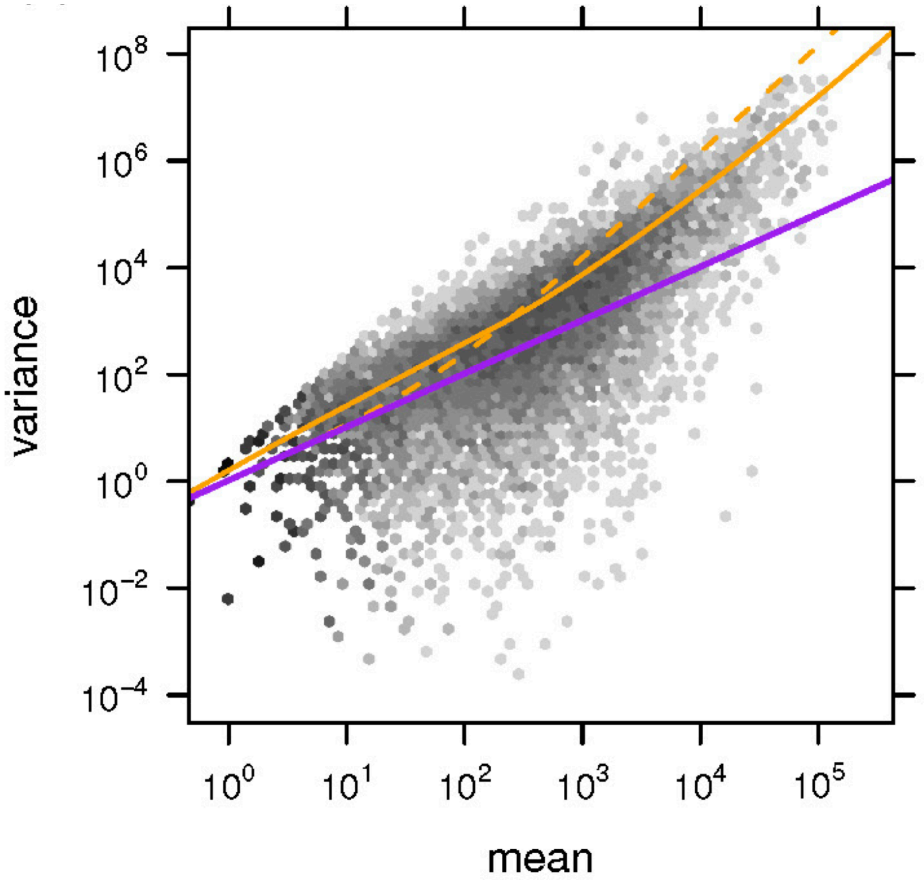
Differential analysis strategies

- Use read counts
  - Standard Fisher exact test

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

  - Model read counts (Poisson, negative binomial) and test whether models are distinct
  - Use empirical approaches that do not rely on parametric assumptions (more on this later)

Poisson model does not work



Adapted from Anders, 2010

Biological variance does not follow a Poisson model

Using a parametric model (DESeq, Cufflinks)

Because of overdisperssion DESeq and Cufflinks uses a Negative binomial to model read counts

$$K_{g,s} \sim \mathcal{N}(K_{g,s}, \sigma_{g,s});\ \sigma_{g,s} = K_{g,s} + \nu_{g,s}$$

Given observed counts for two samples in replicates

$$k_{g,s_1} \ldots k_{g,s_n};\ k_{g,t_1} \ldots k_{g,t_m}$$

DESeq tests the null hypothesis that all counts are sampled from the same distribution

$$P(\sum_i k_{g,s_i} + \sum_j k_{g,t_j} | \mu_s = \mu_t)$$

Cufflinks differential isoform usage

Let a gene G have *n* isoforms and let $p_1, \ldots, p_n$ the estimated fraction of expression of each isoform.

Call this a the isoform expression distribution *P* for G

Given two samples the differential isoform usage amounts to determine whether H_0: $P_1 = P_2$ or H_1: P_1 != P_2 are true.

To compare distributions Cufflinks utilizes an information content based metric of how different two distributions are called the Jensen-Shannon divergence:

$$JS(p^1, \ldots, p^m) = H\left(\frac{p^1 + \cdots + p^m}{m}\right) - \frac{\sum_{j=1}^{m} H(p^j)}{m}.$$

$$H(p) = -\sum_{i=1}^{n} p_i log p_i.$$

The square root of the JS distributes normal.

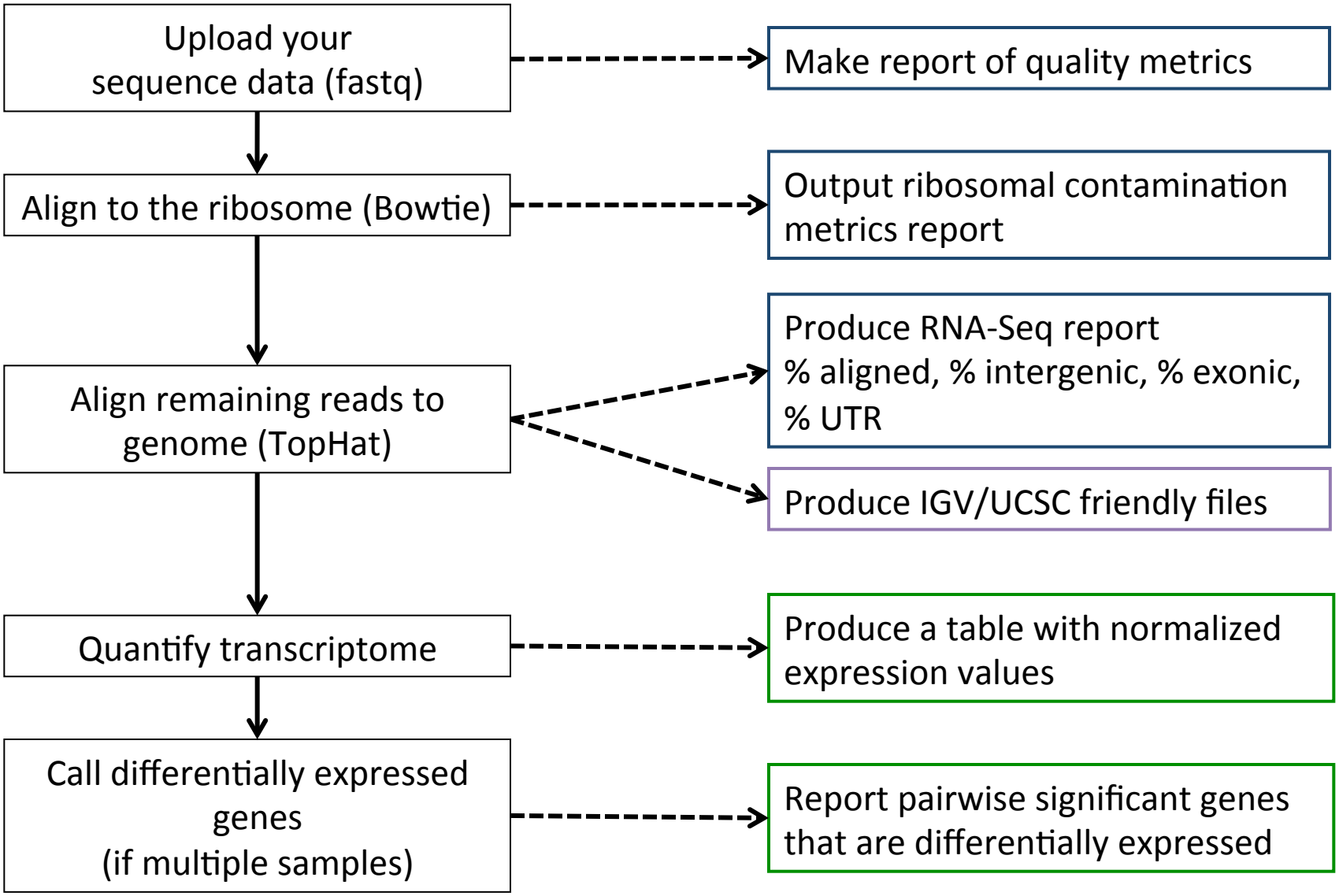
RNA-Seq differential expression software

	Underlying model	Notes
DegSeq	Normal. Mean and variance estimated from replicates	Works directly from reference transcriptome and read alignment
EdgeR	Negative Bionomial	Gene read counts table
DESeq	Negative Bionomial	Gene read counts table
Cufflinks	~~Poisson~~ Negative Bionomial	Works directly from the alignments
Myrna	Empirical	Sequence reads and reference transcriptome

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
- 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

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

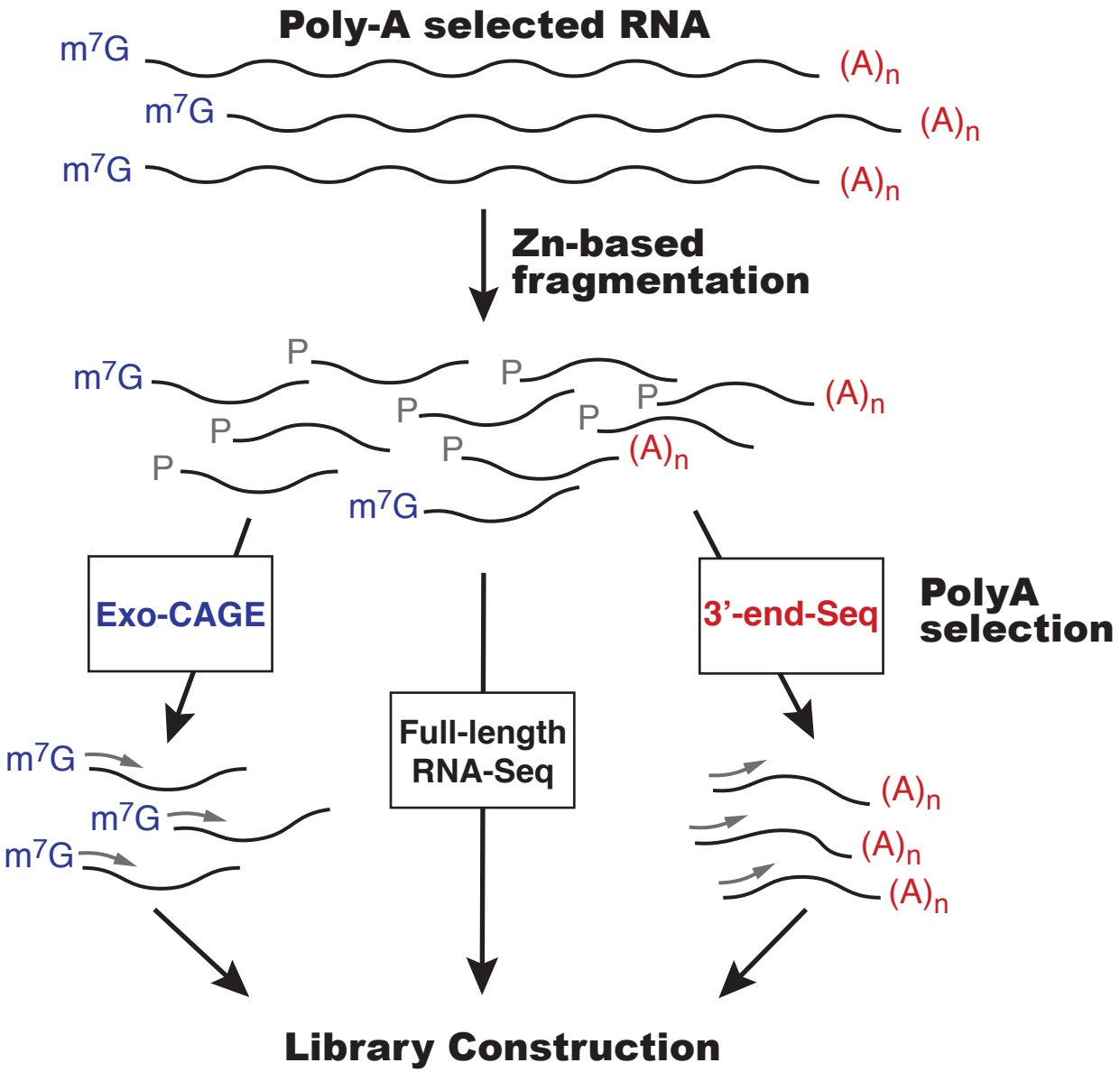


Final considerations on quantification

- Using different libraries:
  - Targeting the 3' end
  - Targeting 5' end
- What depth do we really need?

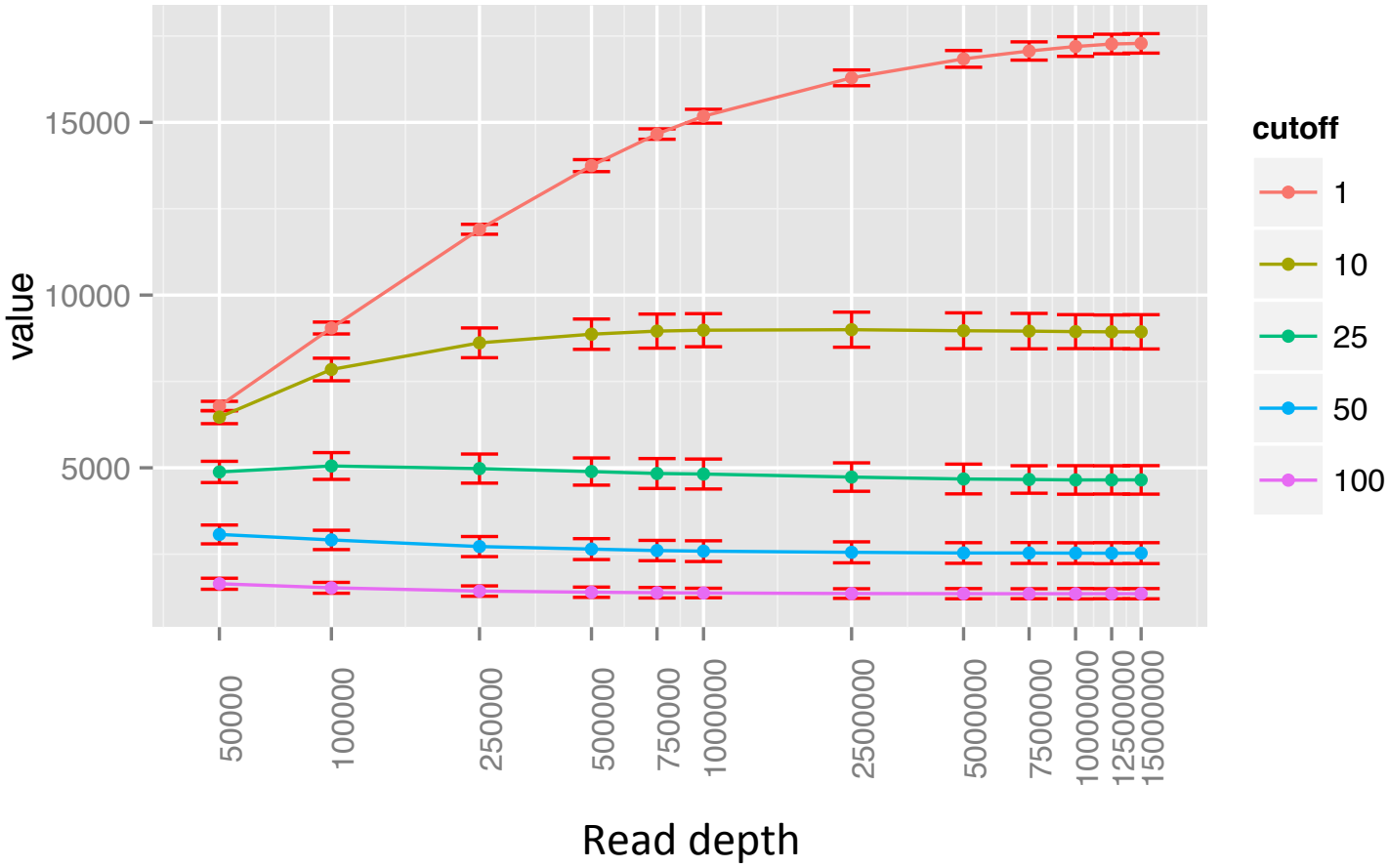
Alper Kucukural
Sabah Kadri
Maxim Artyomov

RNA-Seq libraries: Summary



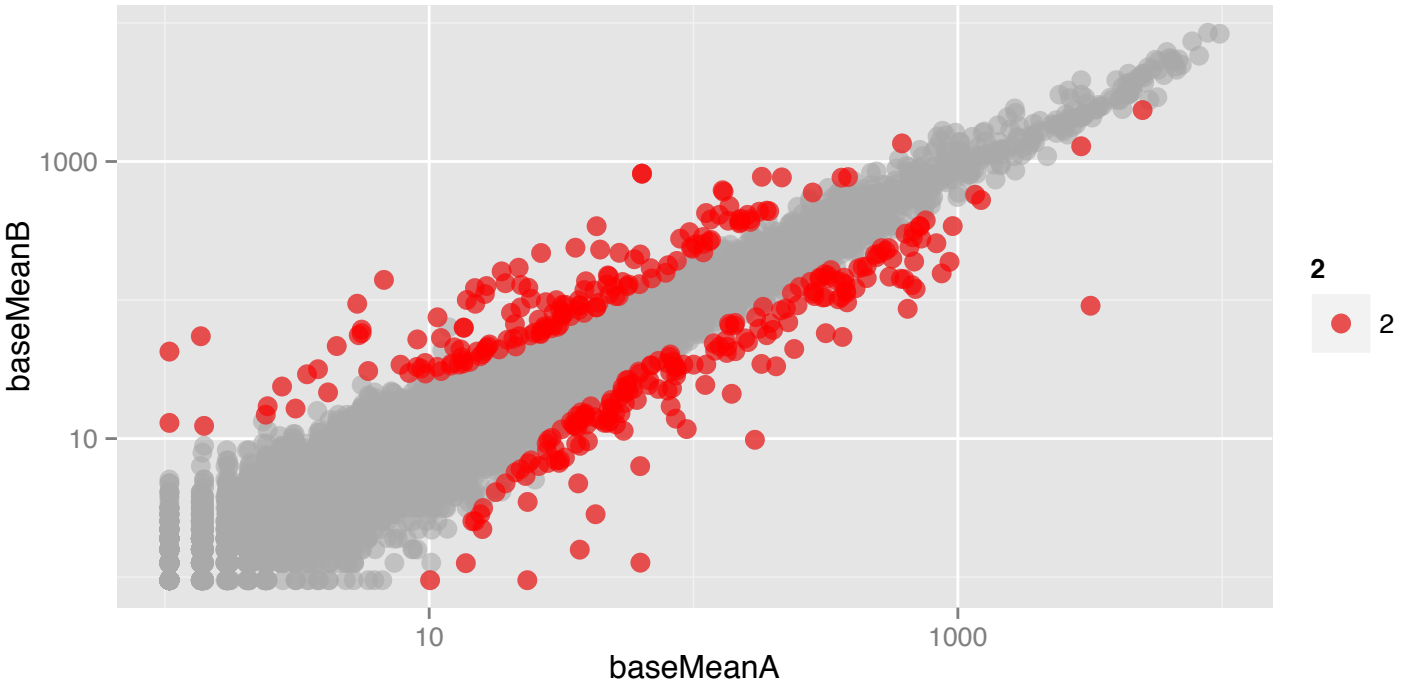
Sabah Kadri
Maxim Artyomov
Sebastian Kadener

Robustness to low depth: Transcripts detected

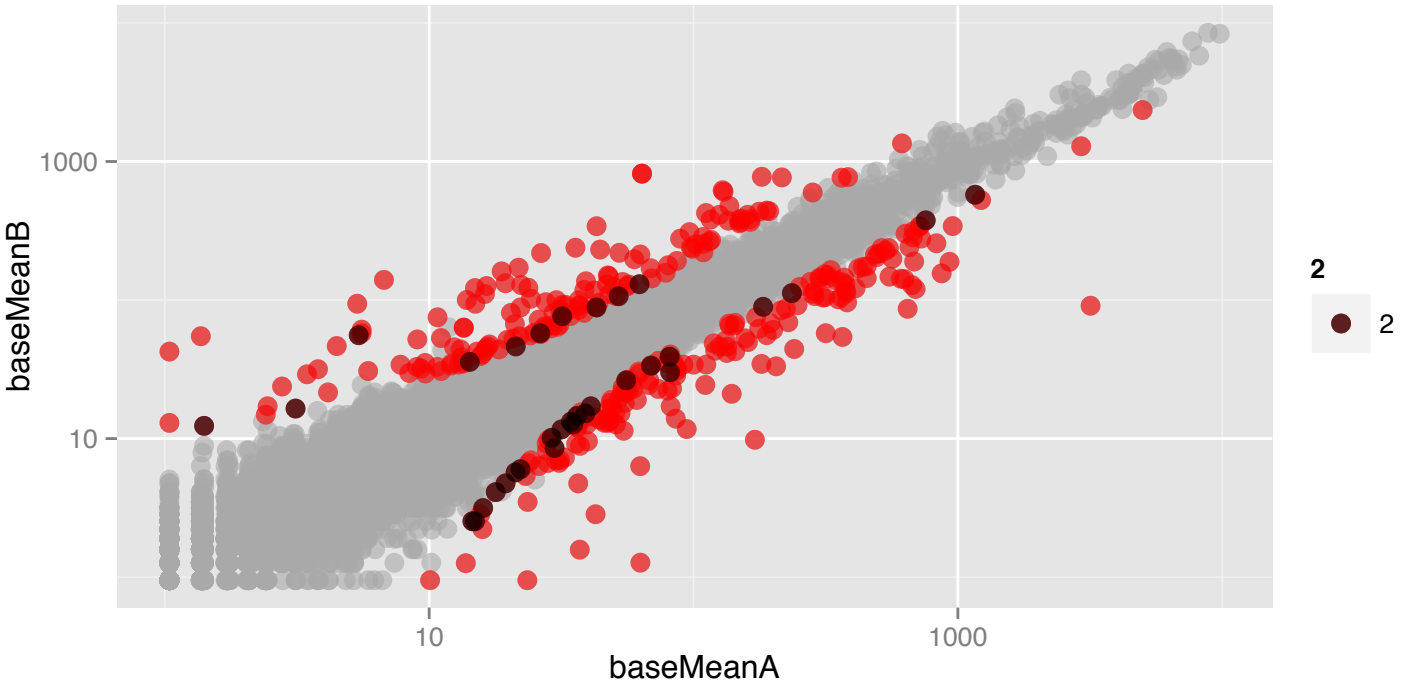


Alper Kucukural

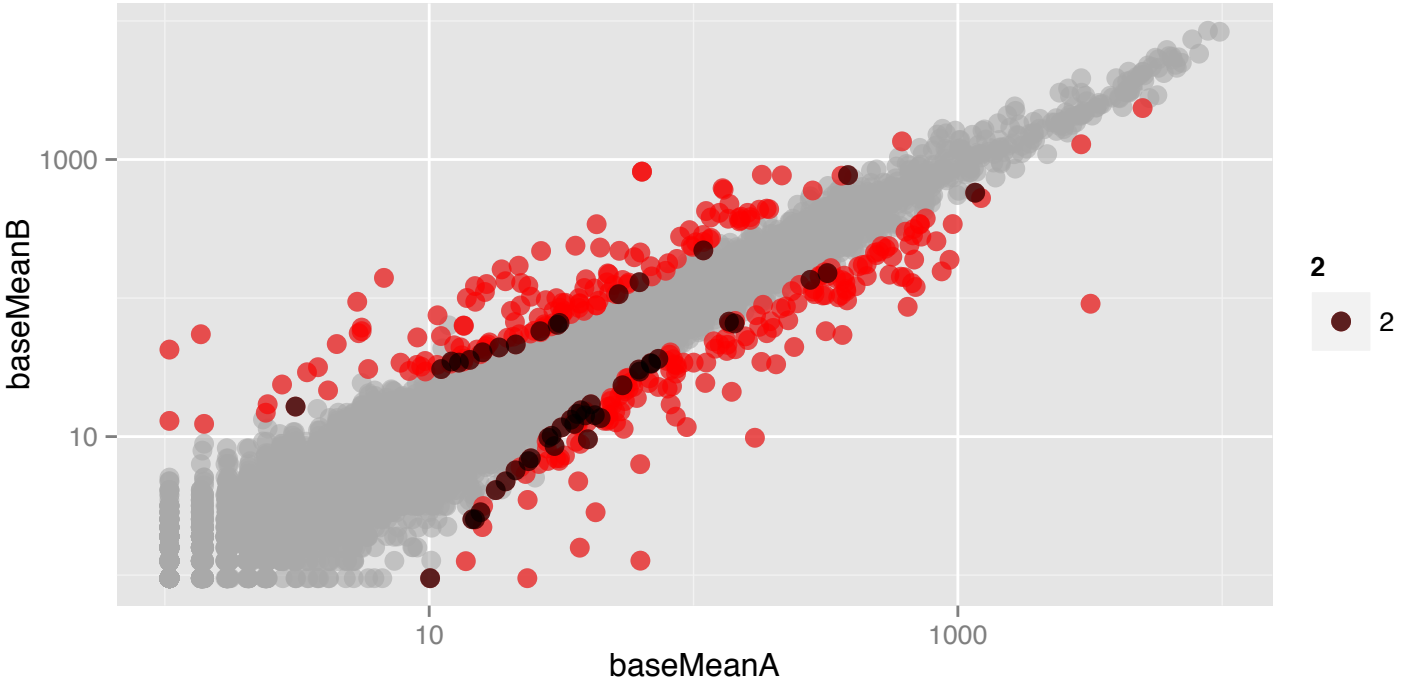
RSEM/DESeq: 15 mill reads in worm



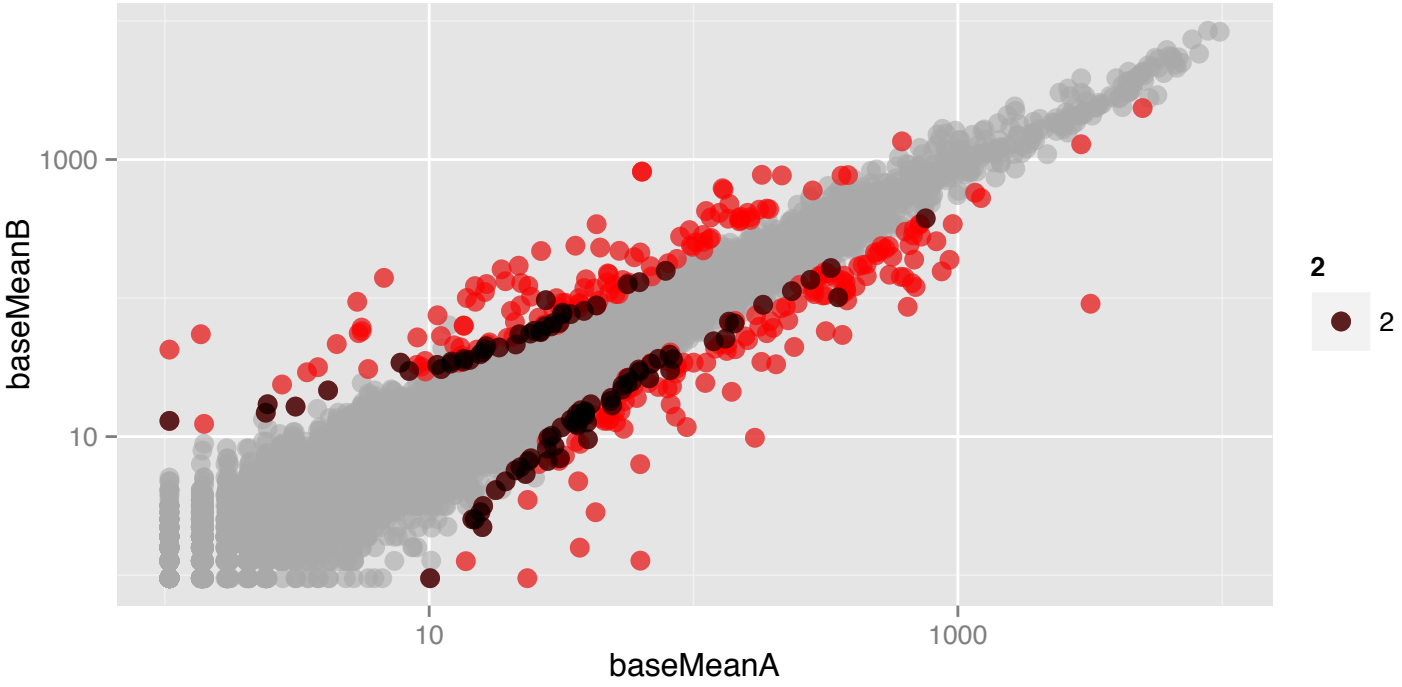
RSEM/DESeq: 10 mill reads in worm



RSEM/DESeq: 7.5 mill reads in worm

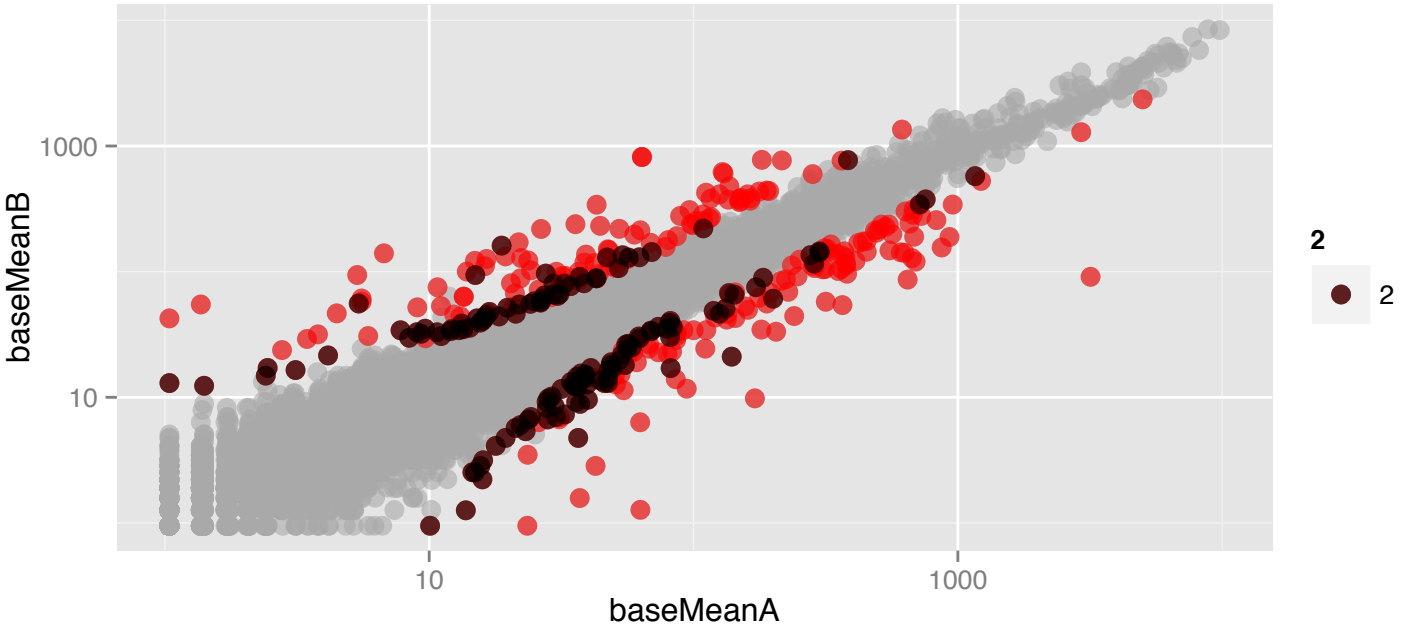


RSEM/DESeq: 5 mill reads in worm

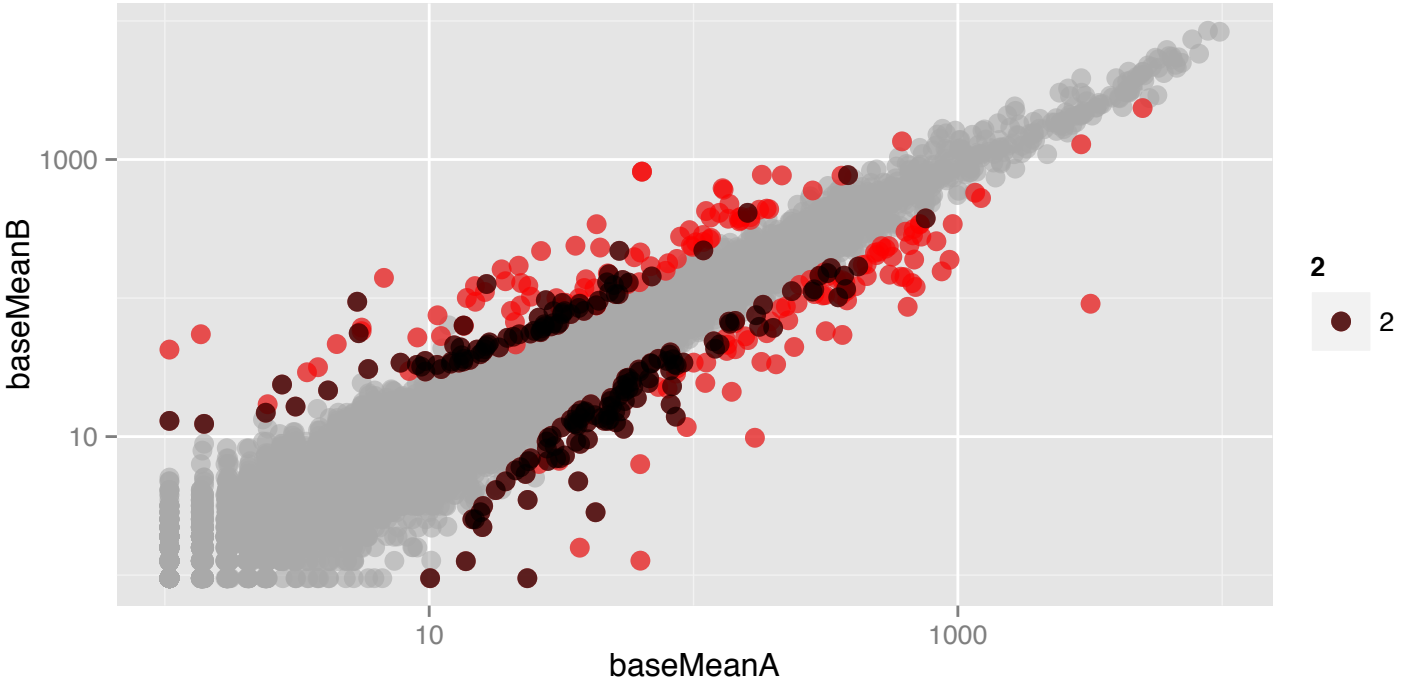


Alper Kucukural

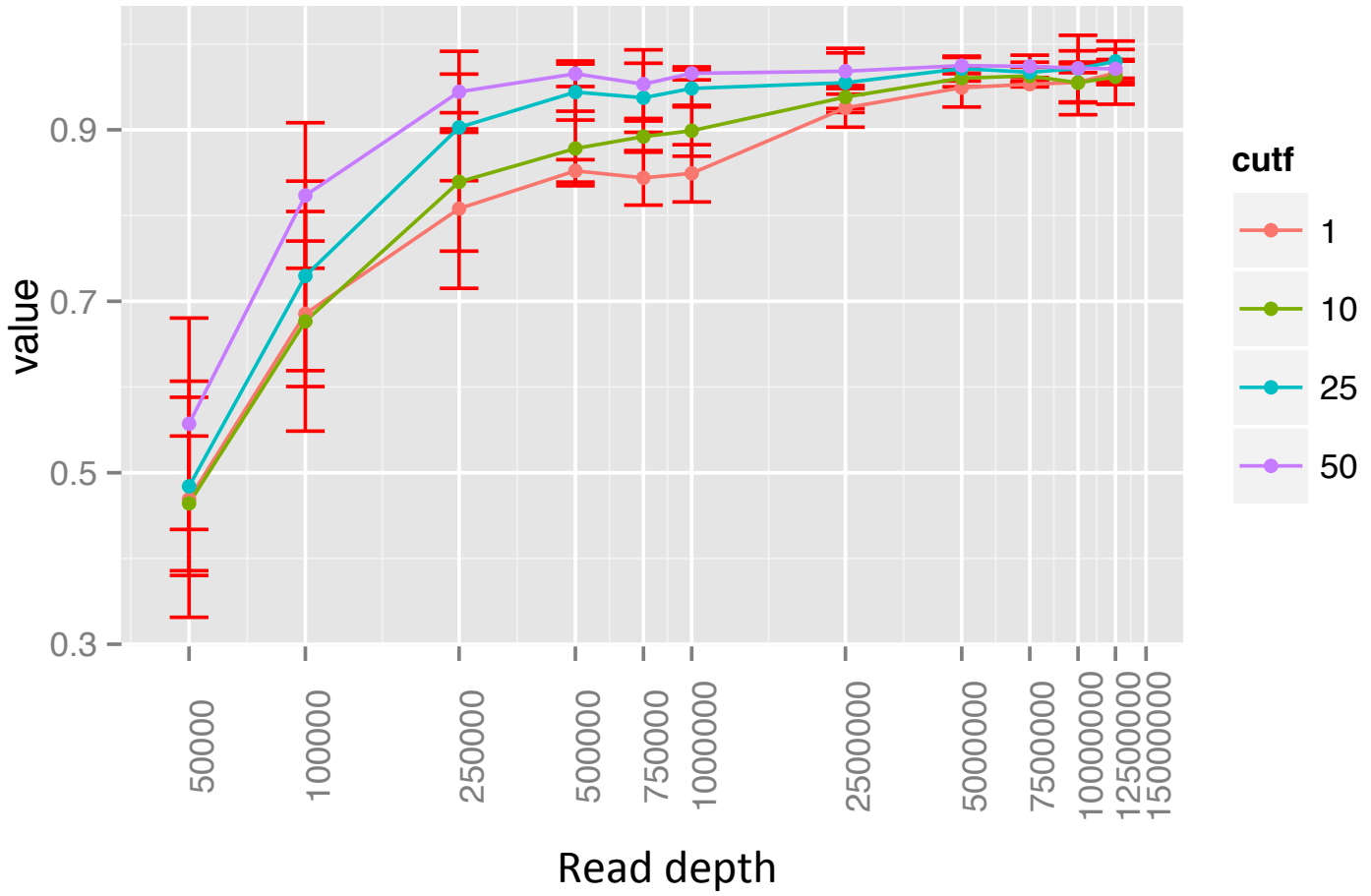
RSEM/DESeq: 2.5 mill reads in worm



RSEM/DESeq: 1 mill reads in worm



Robustness of DGE to low depth



Final considerations: The steps of Sequencing analysis

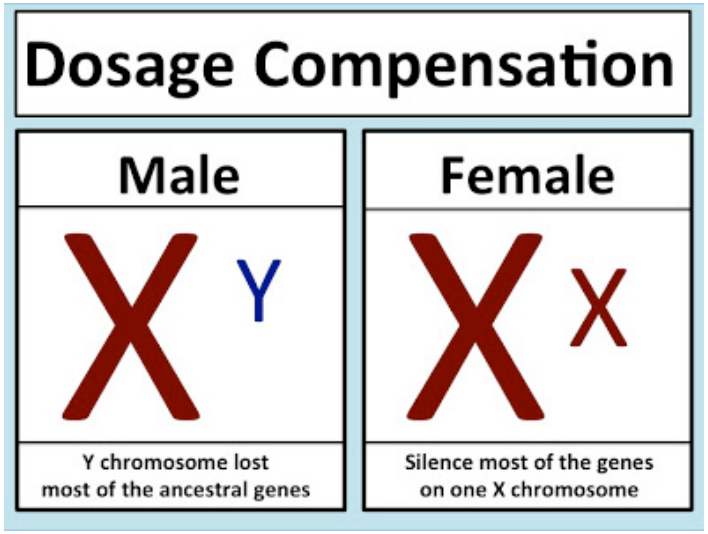
- Filter reads (fastq file) by removing adapter, splitting barcodes.
  - Evaluate overall quality, look for drop in quality at ends. Trim reads if ends are of low quality
- Alignment to the genome
  - Use transcriptome if available
  - Filter out likely PCR duplicates (reads that align to the same place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

A Vignette: Large non-coding RNA, are they an evolutionary playground?

Stefan Washietl
Manolis Kellis
http://genome.cshlp.org/content/early/2014/01/15/gr.165035.113?top=1

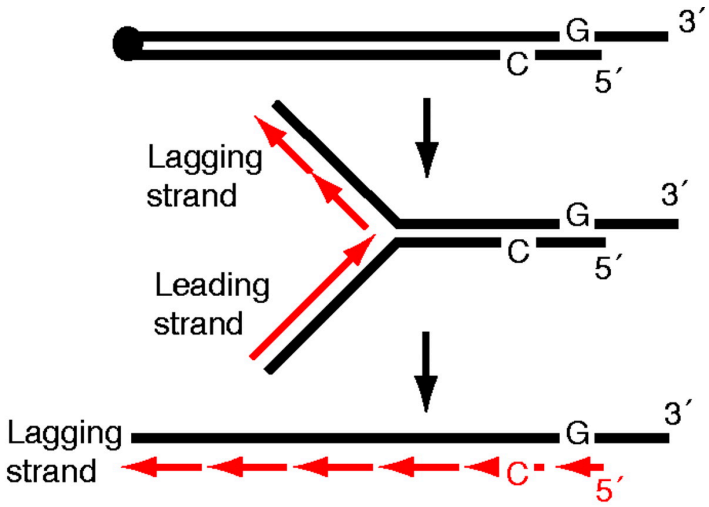
What do we know about lncRNA function?

- How to deal with XY vs XX?



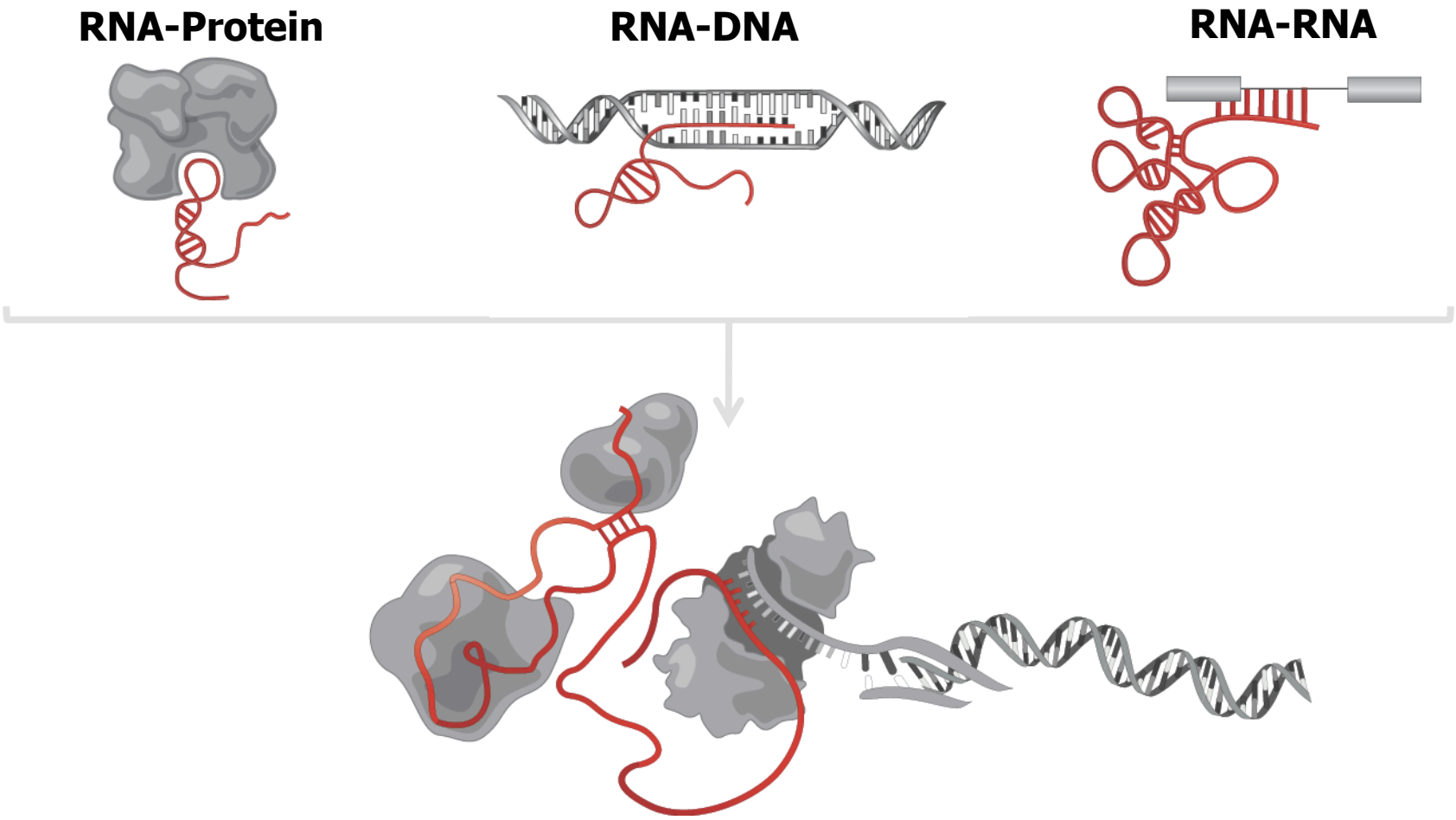
- Dosage compensation is regulated by XIST (Ballabio et al, 1987)
- XIST scaffolds large protein complexes
- XIST is a 17 Kb non-coding RNA

- How to keep telomeres?



- Telomerase (Greider & Blackburn 1985)
- Telomerase is a **Ribo**nucleoprotein (Greider & Blackburn 1989)
- TERC is 550 bases

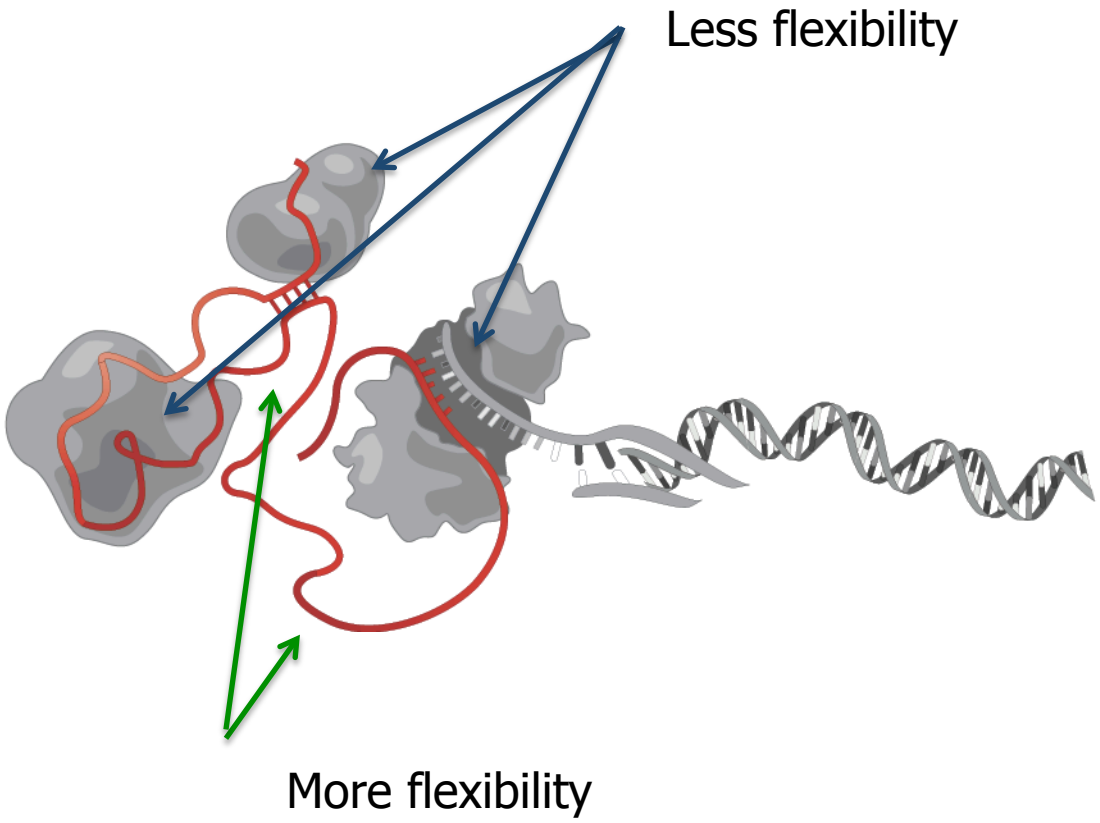
How to think about lincRNAs as functional units?



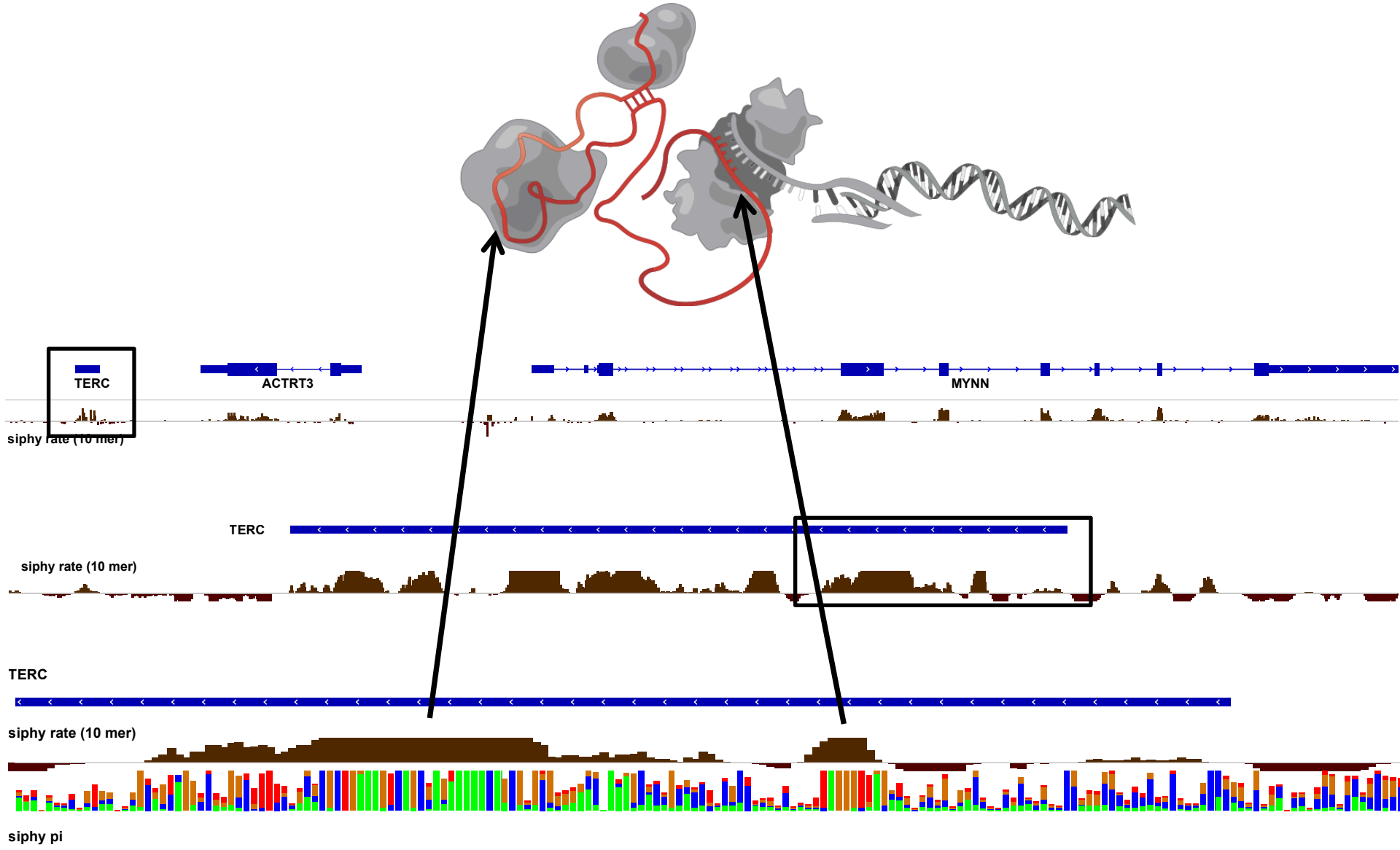
Example: Telomerase RNA

Not all sequences are functionally equivalent

RNA as a flexible malleable molecule



TERC has clear conserved patterns



lincRNAs play key roles in biological processes

The Noncoding RNA *Taurine Upregulated Gene 1* Is Required for Differentiation of the Murine Retina

Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes

lincRNAs act in the circuitry controlling pluripotency and differentiation

Rsx is a metatherian RNA with *Xist*-like properties in X-chromosome inactivation

lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs

Circadian changes in long noncoding RNAs in the pineal gland

A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes

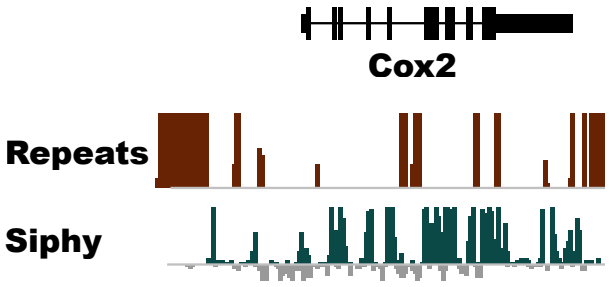
Morten Møller[f], mparative Sequencing

Susan Carpenter,[1,2] Maninjay Atianand,[1] Daniel Aiello,[1] Emiliano P. Ricci,[3] Pallavi Gandhi,[1] Lisa L. Hall,[4] Meg Byron,[4] Brian Monks,[1] Meabh Henry-Bezy,[1] Jeanne B. Lawrence,[4] Luke A. J. O'Neill,[2] Melissa J. Moore,[3] Daniel R. Caffrey,[1*†] Katherine A. Fitzgerald[1*†]

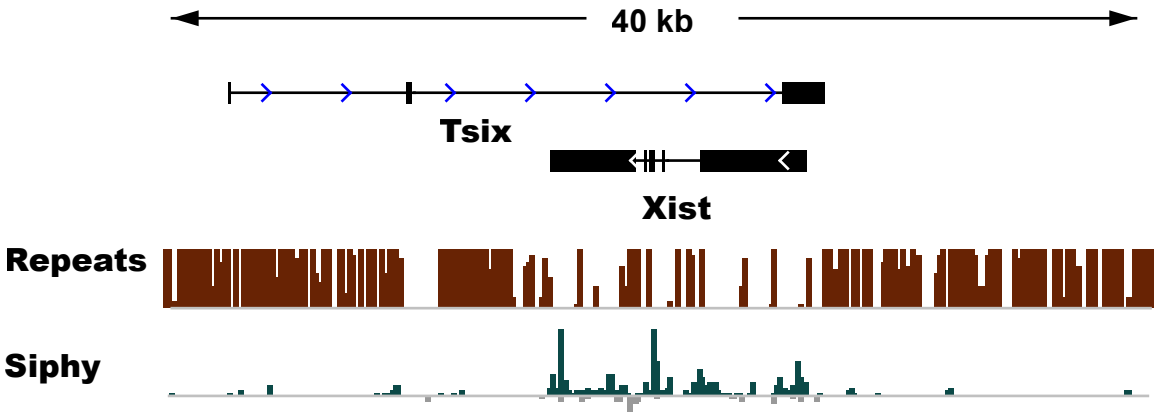
Play important roles in a variety of biological process

- Development
- Cancer
- Immunity
- Differentiation
- Circadian cycle

lincRNA are an evolutionary puzzle



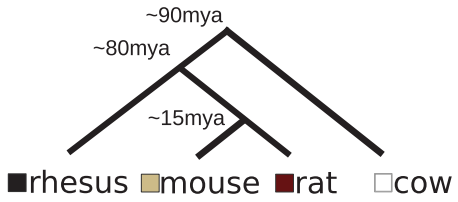
Intergenic **lincRNA**



Using expression to assess conservation

- Hypothesis 1: lincRNAs are under "patchy" constraint or a rapidly evolving (Xist?)
  - Sequence conservation would be underestimating lincRNA conservation
  - Evidence of syntenic **conserved expression**
- Hypothesis 2: lincRNAs are young and many are transcriptional noise
  - **Expression is species specific**
  - Sequence conservation not informative
- Hypothesis 3: lincRNAs are easily replaceable by functional orthologs (linc-cox2?)
  - Sequence conservation not informative
  - Syntenic conserved expression not informative

Evolutionary profiling RNA-Seq dataset



3 Individuals per species

Merkin el al. Science 2012

ARTICLE

doi:10.1038/nature10532

The evolution of gene expression levels in mammalian organs

David Brawand[1,2]*, Magali Soumillon[1,2]*, Anamaria Necsulea[1,2]*, Philippe Julien[1,2], Gábor Csárdi[2,3], Patrick Harrigan[4], Manuela Weier[1], Angélica Liechti[1], Ayinuer Aximu-Petri[5], Martin Kircher[5], Frank W. Albert[5]†, Ulrich Zeller[6], Philipp Khaitovich[7], Frank Grützner[8], Sven Bergmann[2,3], Rasmus Nielsen[4,9], Svante Pääbo[5] & Henrik Kaessmann[1,2]

Human
+
Chimp

A human centric approach

All GENCODE long noncoding transcripts

11790

No overlap with annotated coding genes or pseudogenes

5603

A human centric approach

All GENCODE long noncoding transcripts

11790

↓

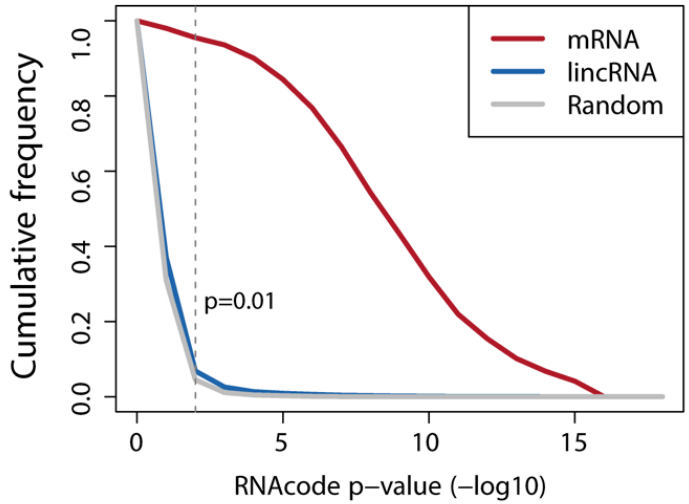
No overlap with annotated coding genes or pseudogenes

5603

↓

No detectable coding potential

5206



A human centric approach

All GENCODE long noncoding transcripts

11790

↓

No overlap with annotated coding genes or pseudogenes

5603

↓

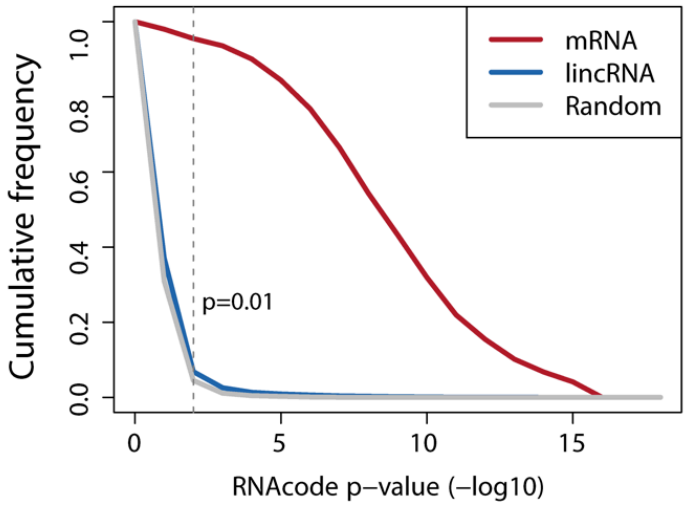
No detectable coding potential

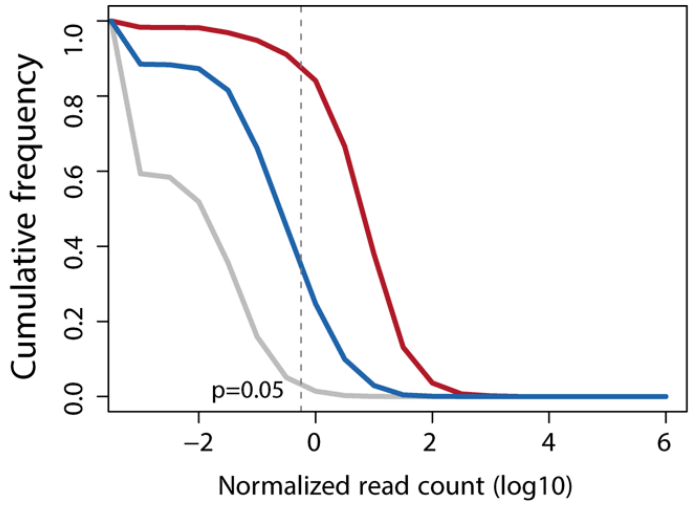
5206

↓

Significantly expressed in human in tissues used in this study

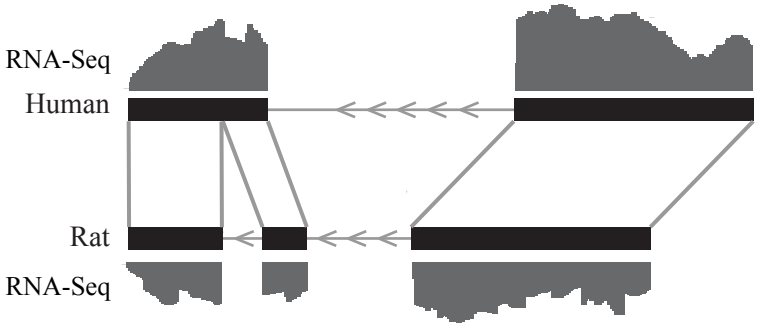
1898



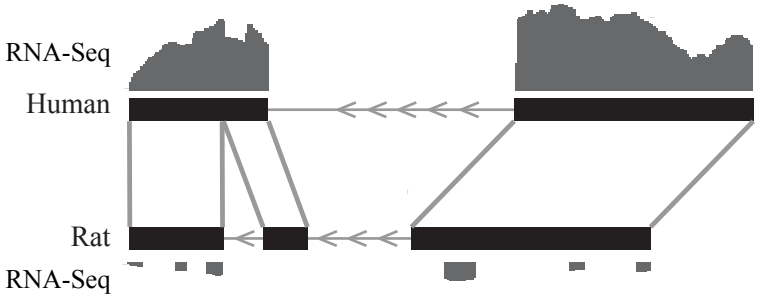


Stefan Washietl

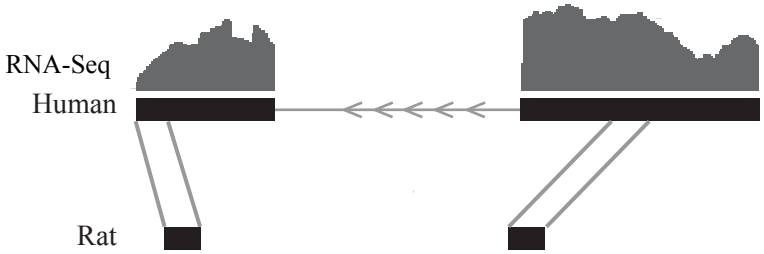
Assessing orthologous expression



Can find orthologous loci with significant expression

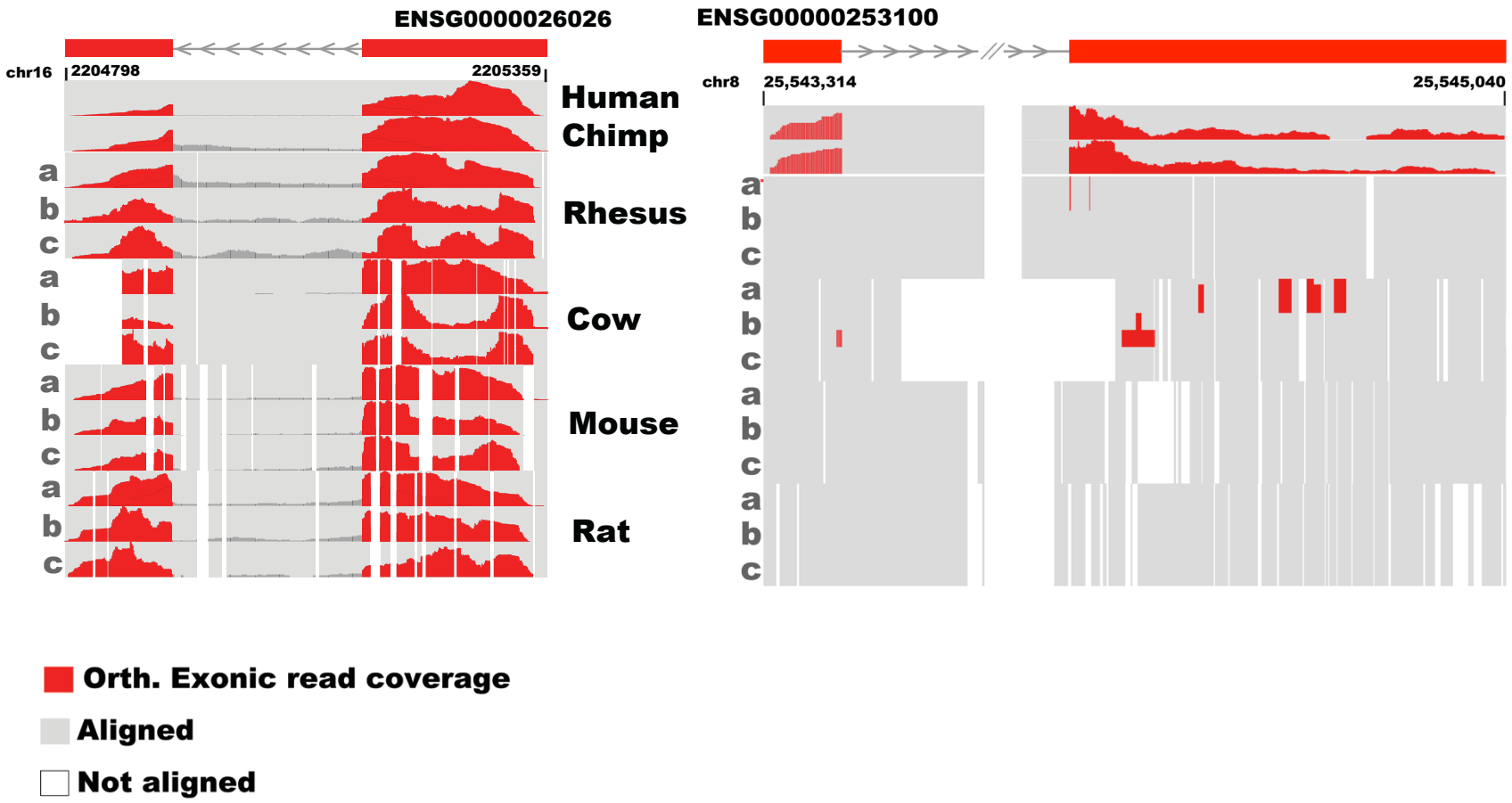


Can find orthologous loci but without significant expression

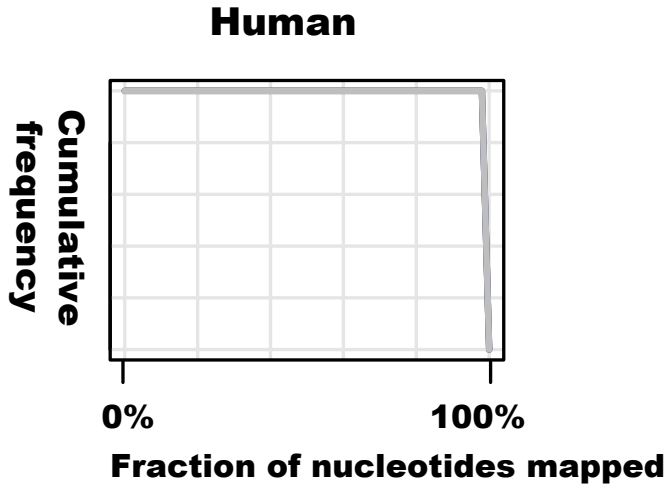


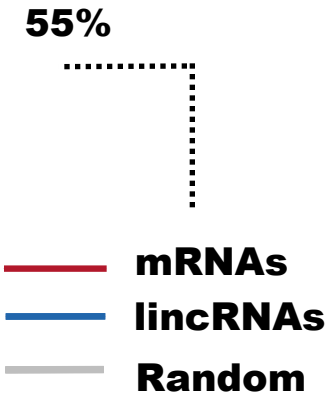
Cannot find orthologous loci

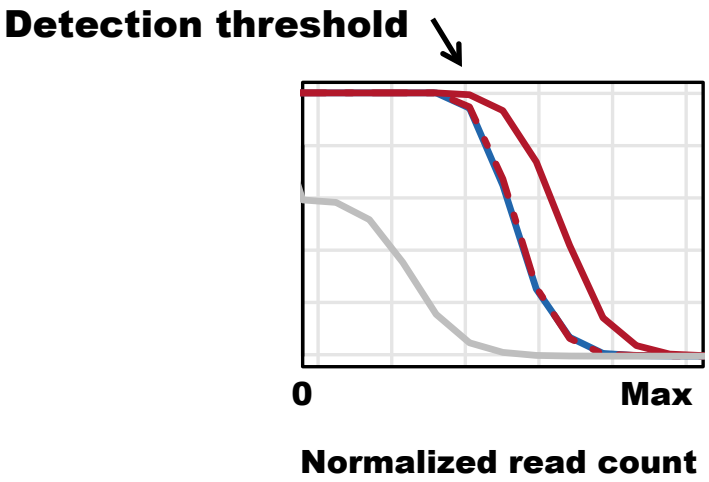
How many are lost or gained?

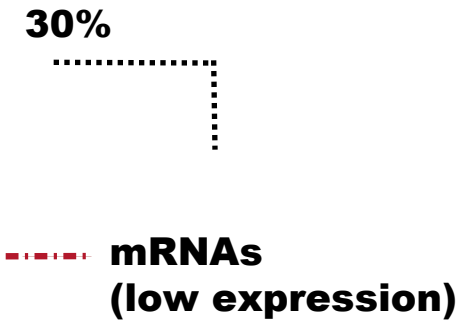


Conservation of Expression decays rapidly



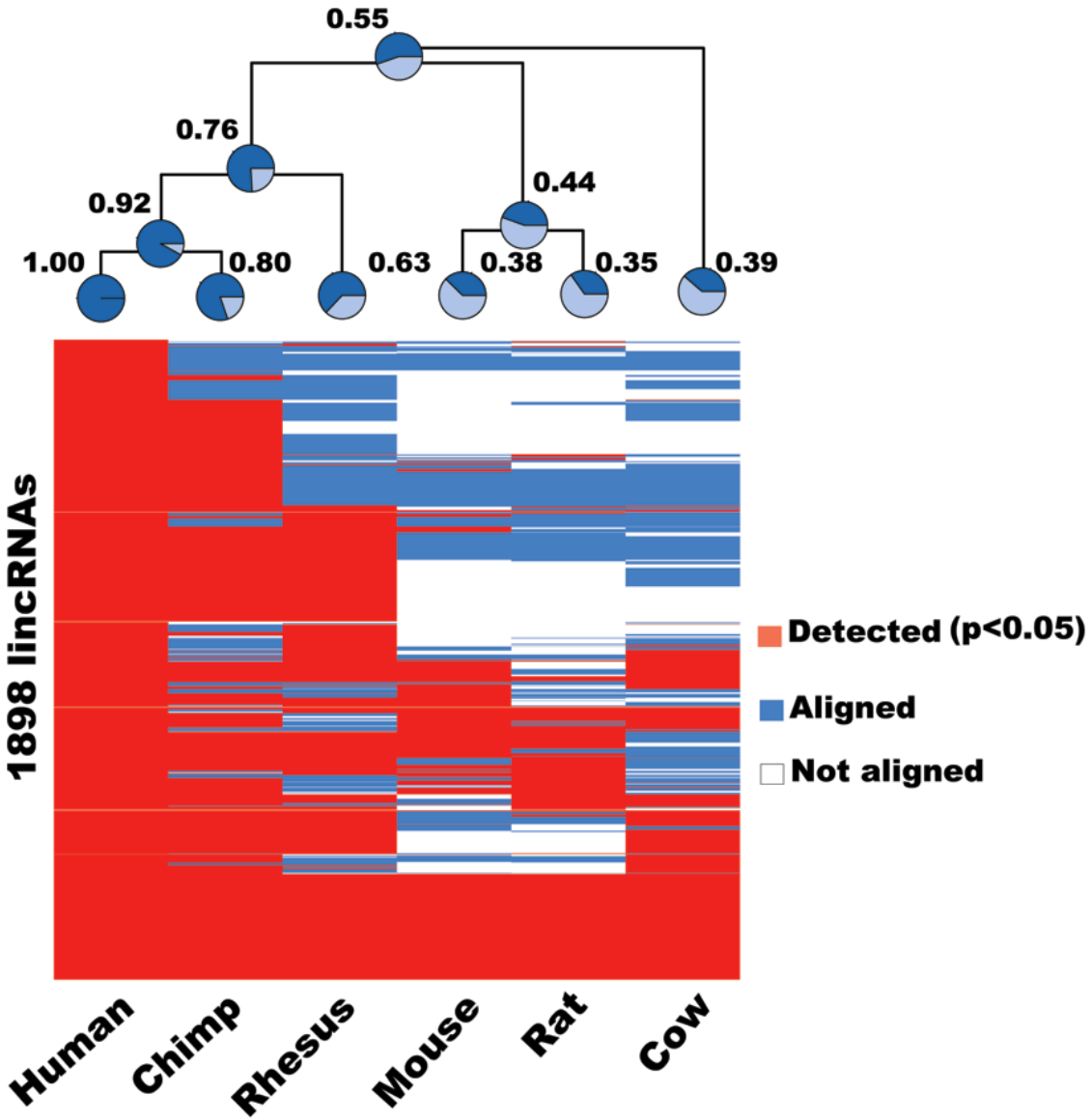




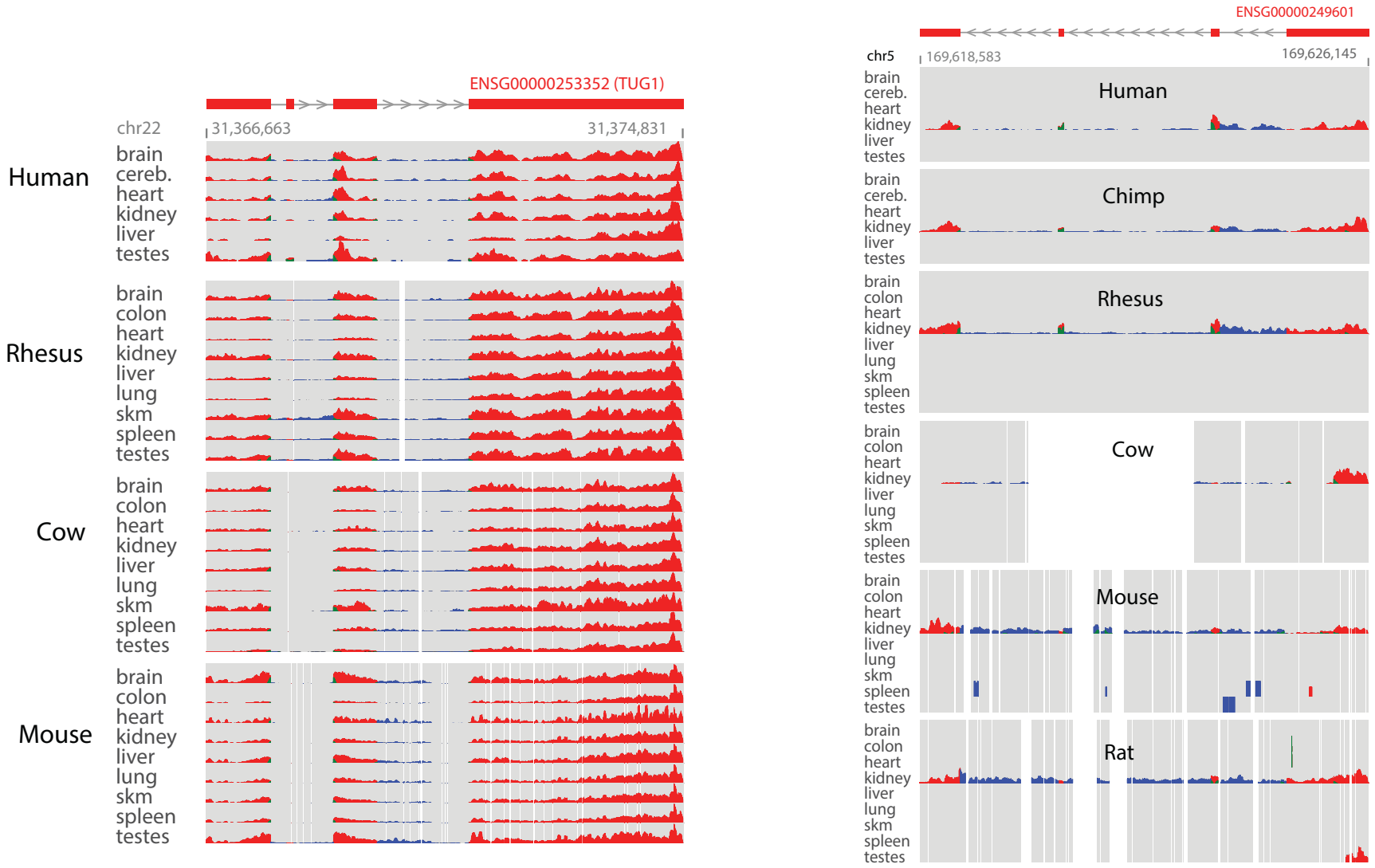


lincRNAs are lost much faster than predicted by their sequence conservation

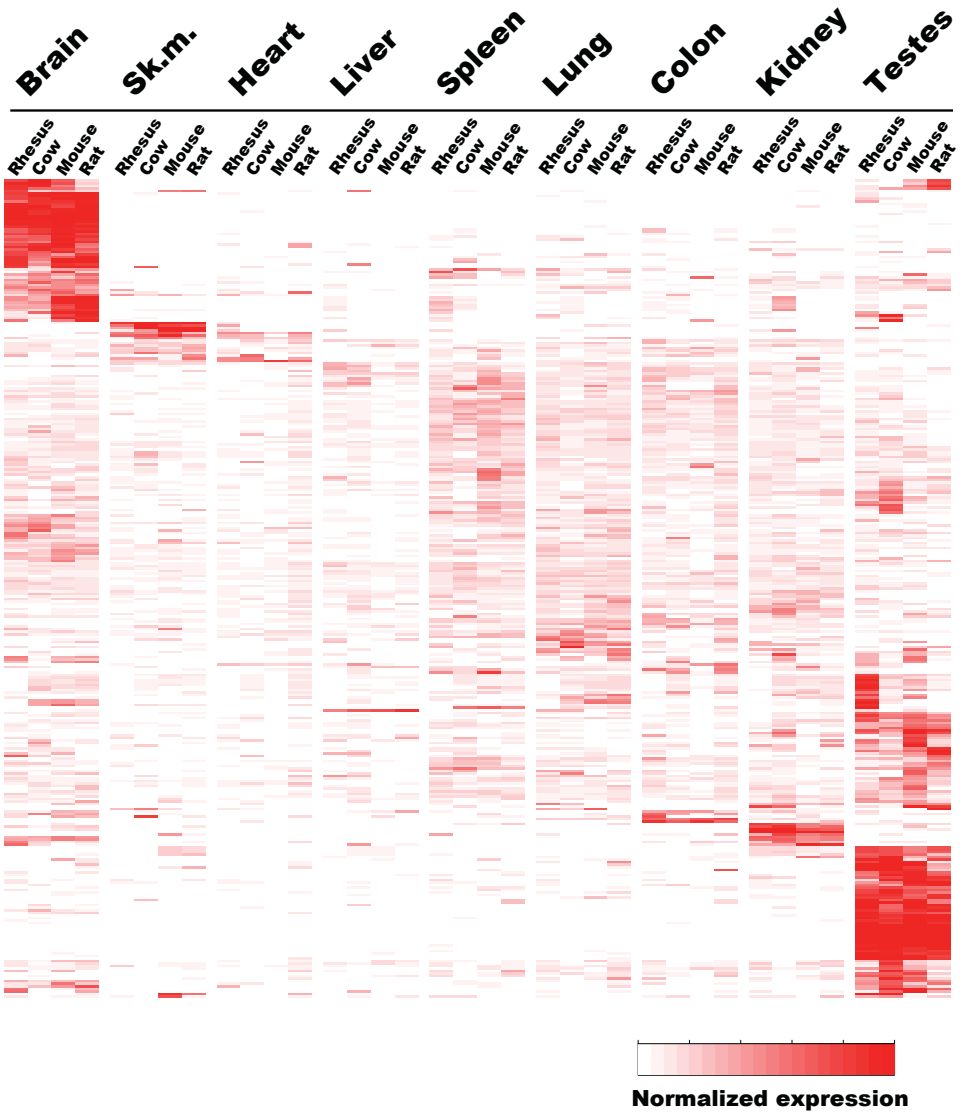
Rapid gain and loss of lincRNAs



Orothologous lincRNAs preserve their tissue specificity



Orthologous lincRNAs preserve their tissue specificity



lncRNAs

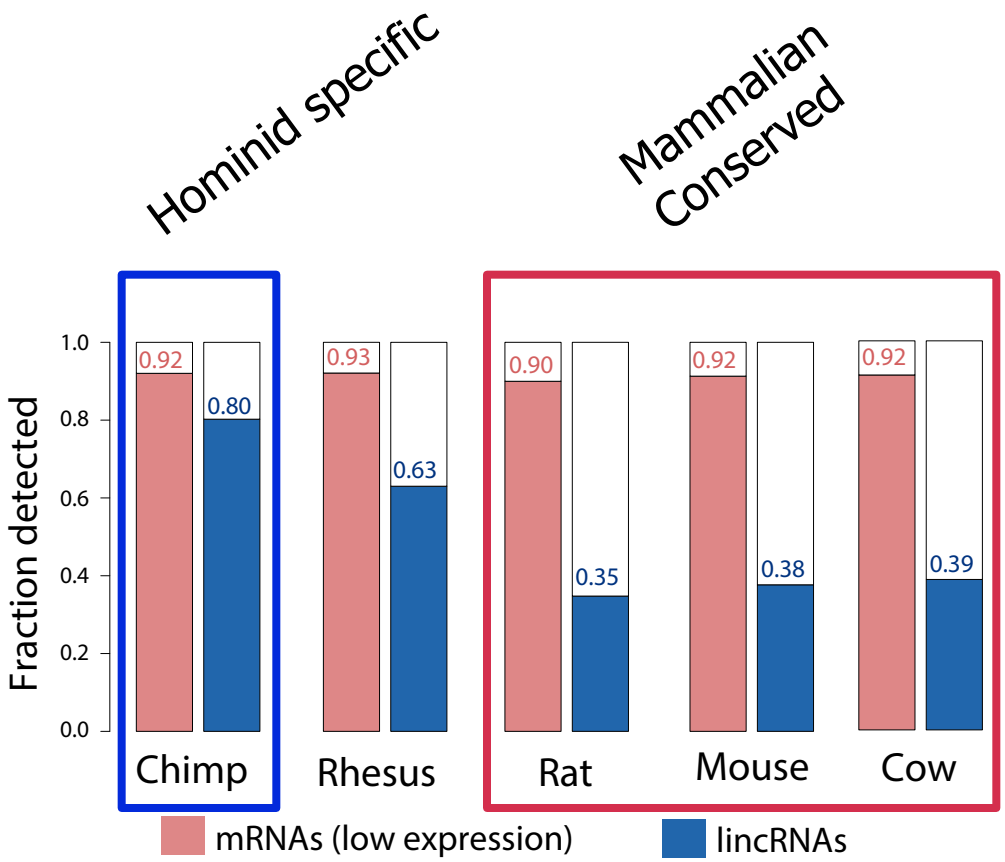
	Rhesus	Cow	Mouse	Rat
Rhesus	1.00	0.53	0.50	0.45
Cow		1.00	0.52	0.46
Mouse			1.00	0.66
Rat				1.00

mRNAs

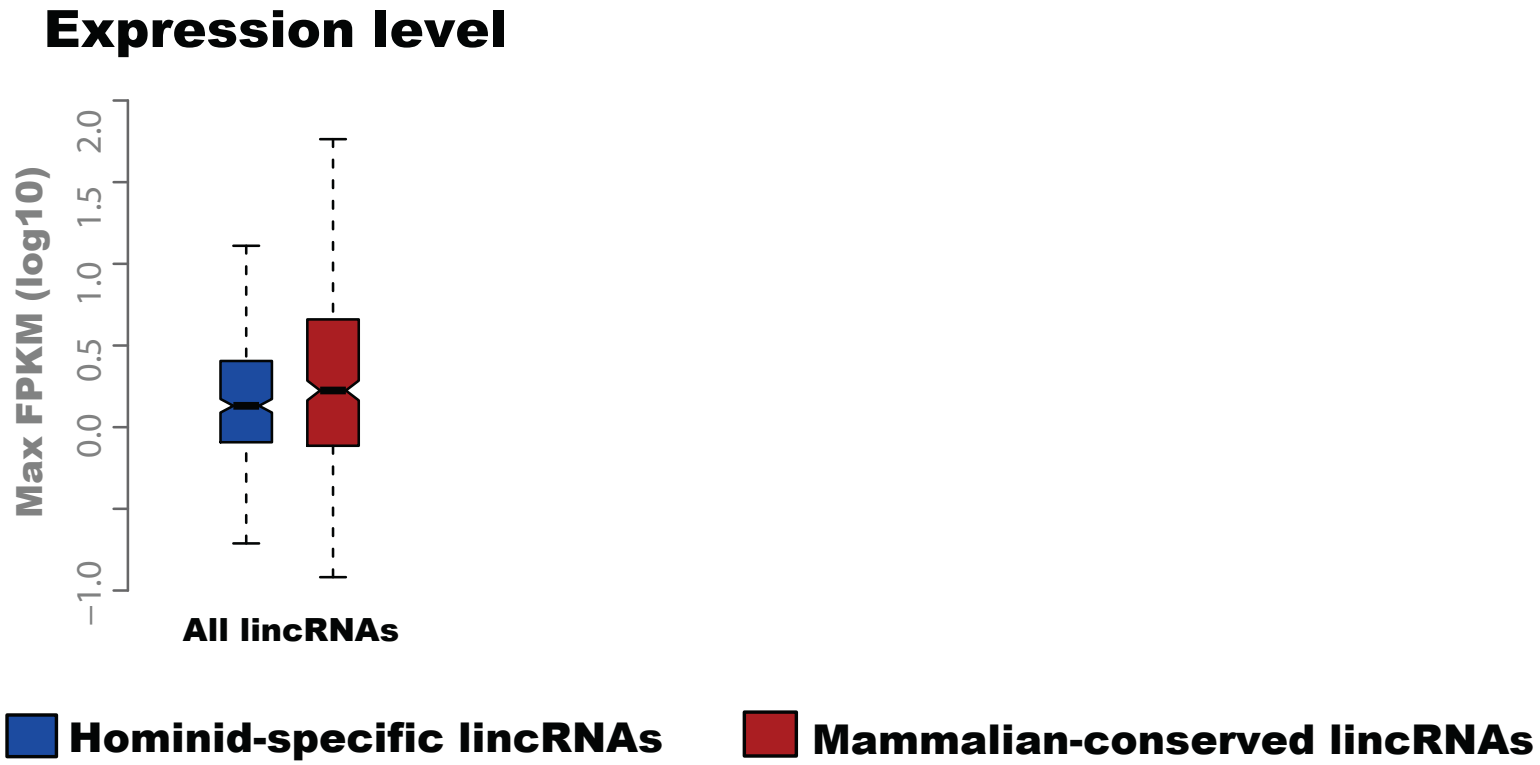
	Rhesus	Mouse	Rat	Cow
Rhesus	1.00	0.53	0.53	0.54
Mouse		1.00	0.54	0.54
Rat			1.00	0.72
Cow				1.00

Conserved lincRNAs have conserved regulation

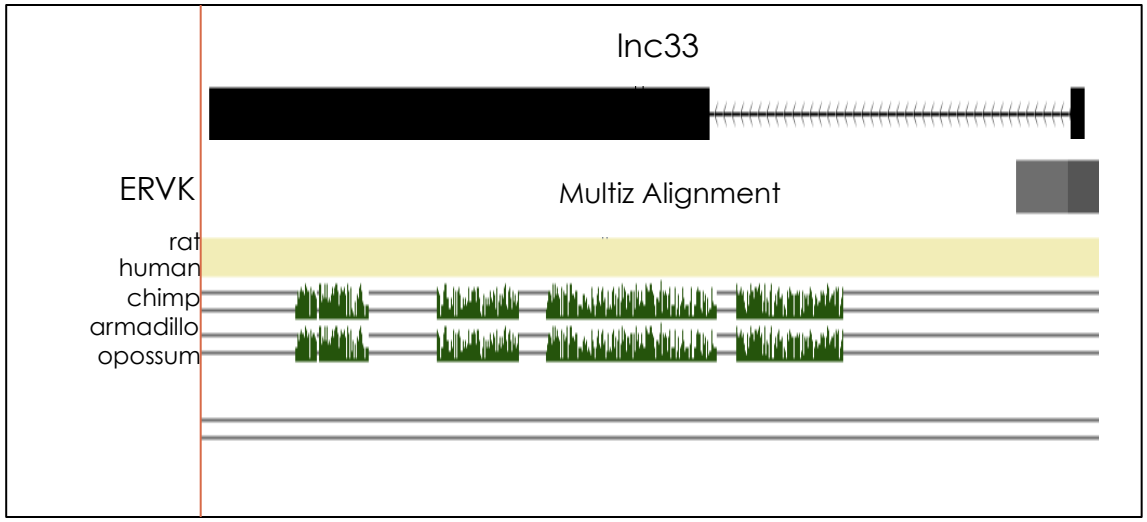
Young vs conserved lincRNAs



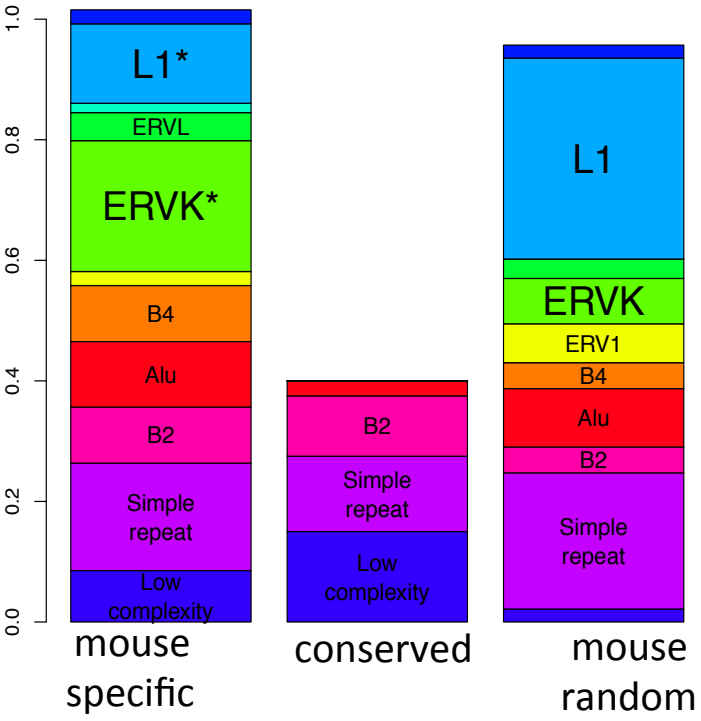
Similarly expressed yet more tissue specific



How are lincRNAs created?



How are lincRNAs created?



XIST like counterpart

LETTER

doi:10.1038/nature11171

Rsx is a metatherian RNA with *Xist*-like properties in X-chromosome inactivation

Jennifer Grant[1], Shantha K. Mahadevaiah[1], Pavel Khil[2], Mahesh N. Sangrithi[1], Hélène Royo[1], Janine Duckworth[3], John R. McCarrey[4], John L. VandeBerg[5], Marilyn B. Renfree[6], Willie Taylor[1], Greg Elgar[1], R. Daniel Camerini-Otero[2], Mike J. Gilchrist[1] & James M. A. Turner[1]

- Female Specific
- Large non-coding (> 20Kb)
- Coats the Xi
- Inactive in Germline cells
- Contains tandem repeats
- Capable of inactivating autosomes

Grant al. *Nature* 2012

Is Rsx a functional ortholog of XIST

Key observations

- lincRNAs have a very rapid rate of gain and loss
- Rapid gain/loss makes the XIST/RSX model were lincRNAs may be easily replaced appealing until … proven wrong or a more reasonable model arises
- Repetitive sequence could be a driving force in the genesis of lincRNAs
- Gene structure seems to be preserved only when junctions may play a functional role and turnover very rapidly when they not.
- Evolutionary signatures can distinguish lncRNA categories

Thanks

Stephan Washietl (MIT)

Jenny Chen (MIT/ Broad)

Manolis Kellis (MIT)

Garber lab

Xiaopeng Zhu

Barbara Tabak

Jenny Chen

Sabah Kadri

Bioinformatics core

A. Kucukural

Postdocs invited!

Garberlab.umassmed.edu