University of Massachusetts Medical School

UMASS

RNA-Seq primer (part II)

Sequencing: applications

Counting applications

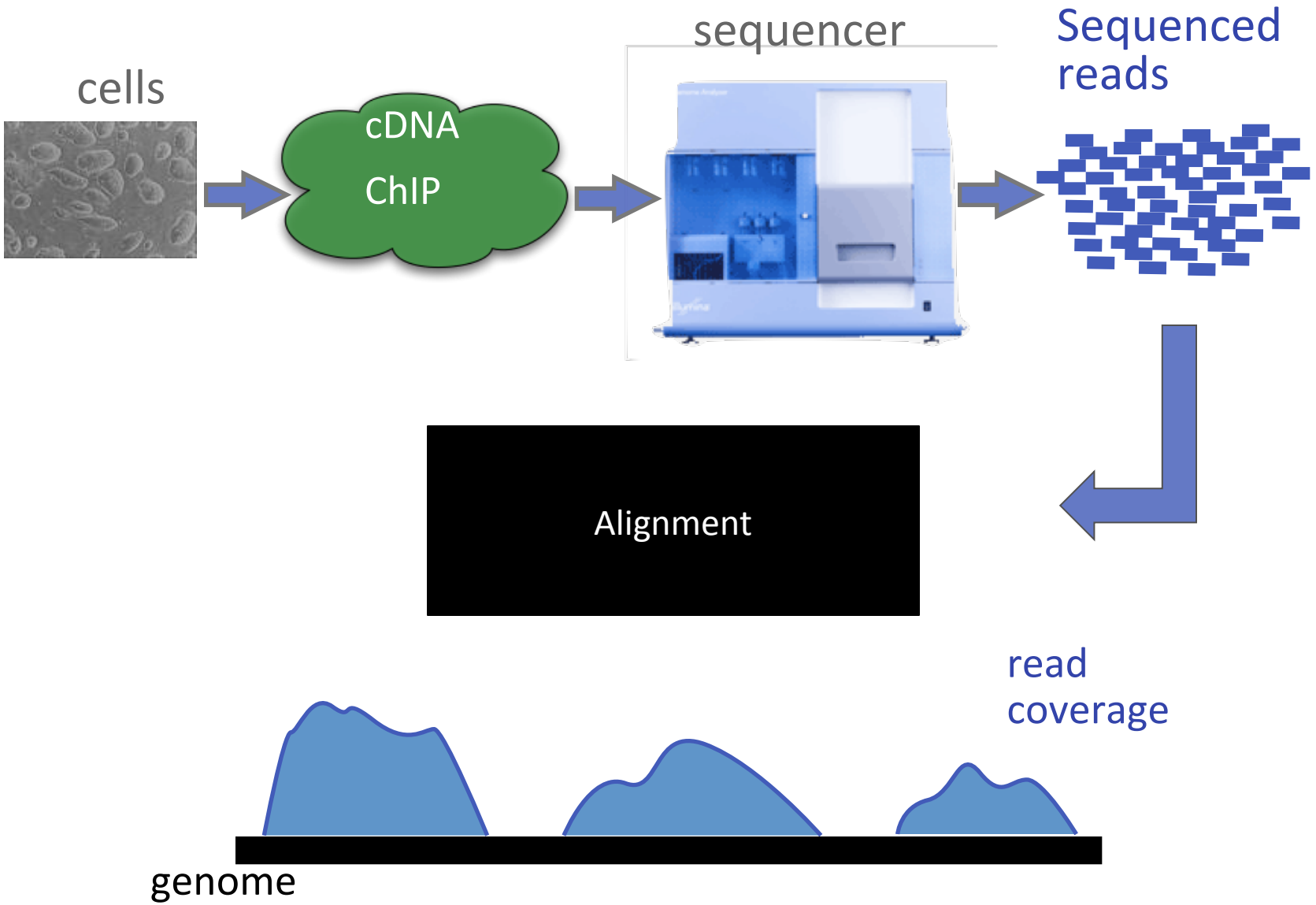
- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
  - 3D genome conformation
- Nucleic acid Interactions

Polymorphism/mutation discovery

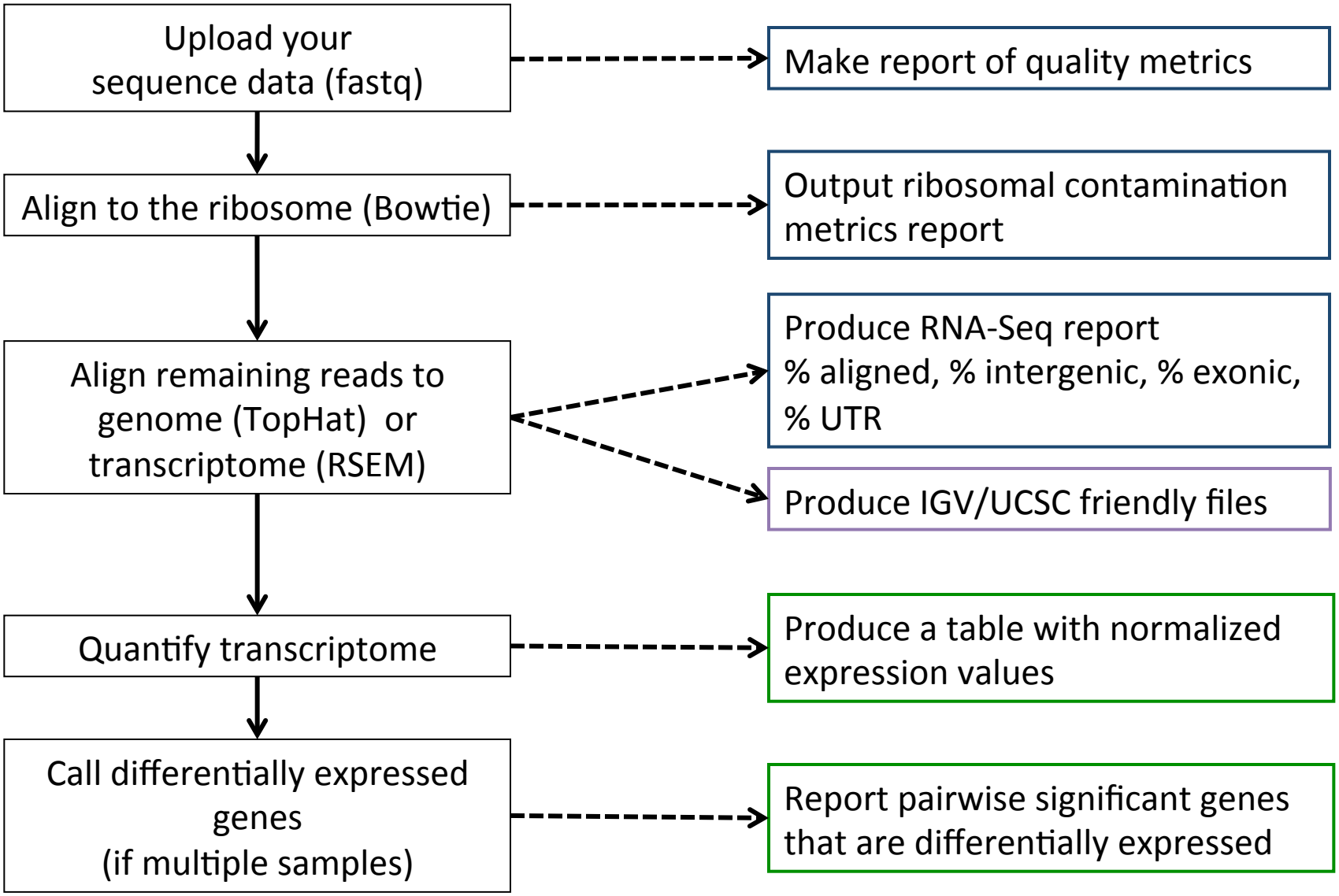
  - Bacteria
  - Genome dynamics
  - Exon (and other target) sequencing
  - Disease gene sequencing
- Variation and association studies
- Genetics and gene discovery

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

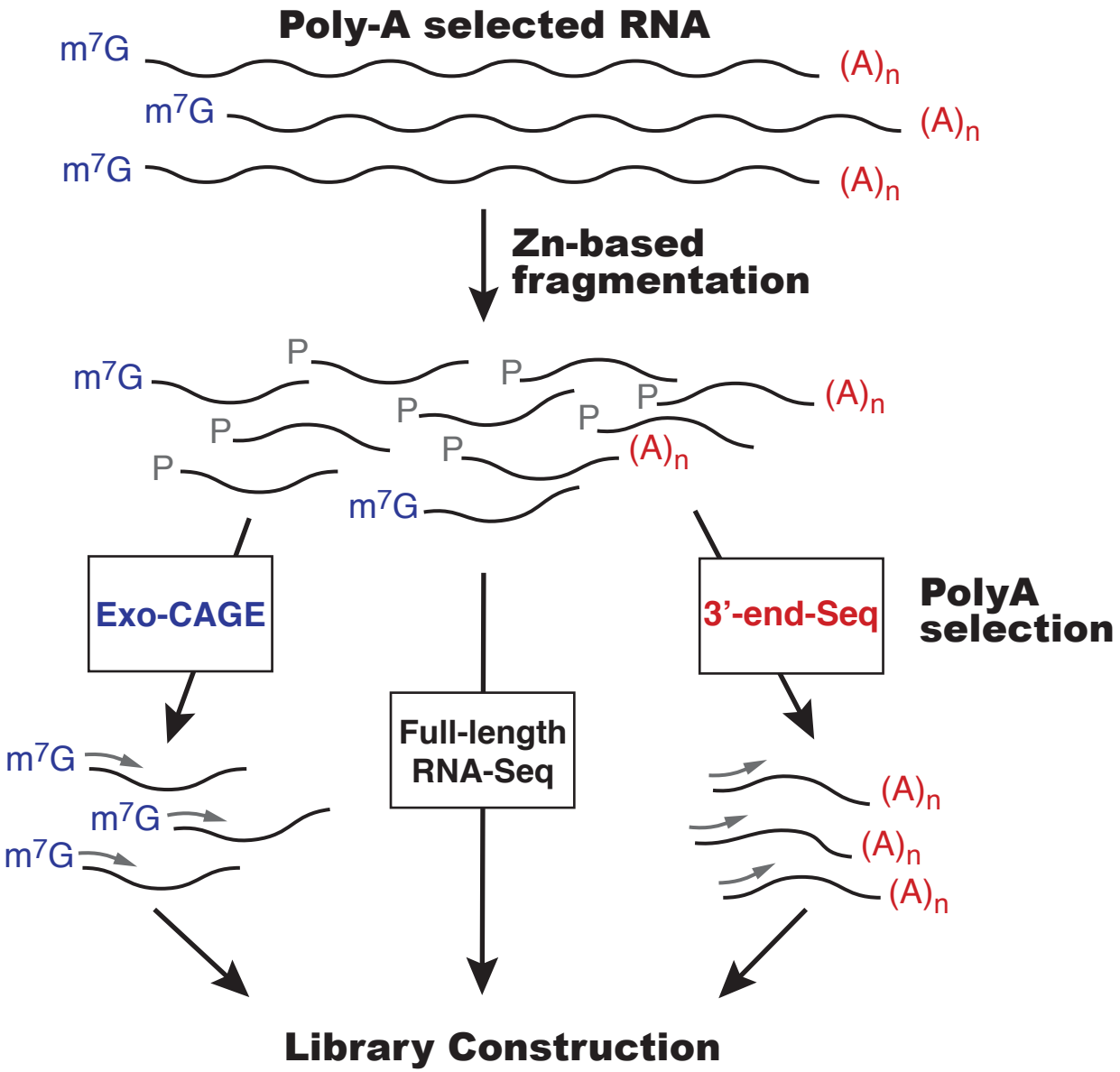
Counting applications



Our typical RNA quantification pipeline

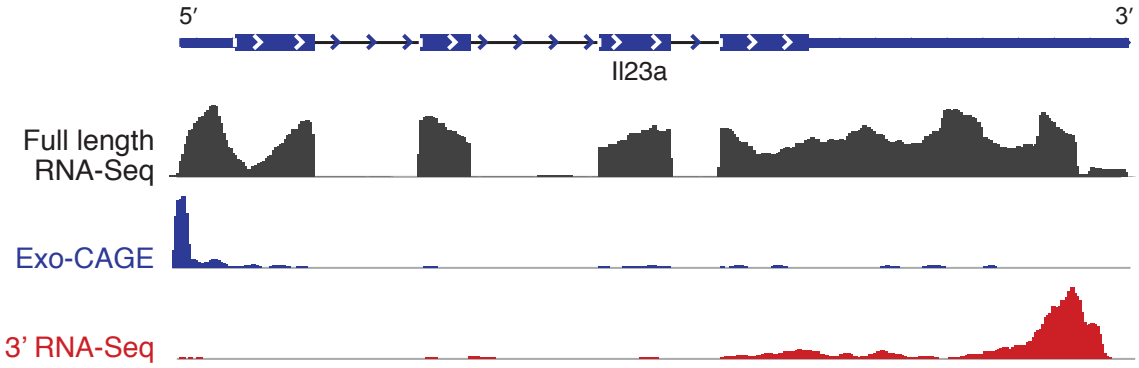


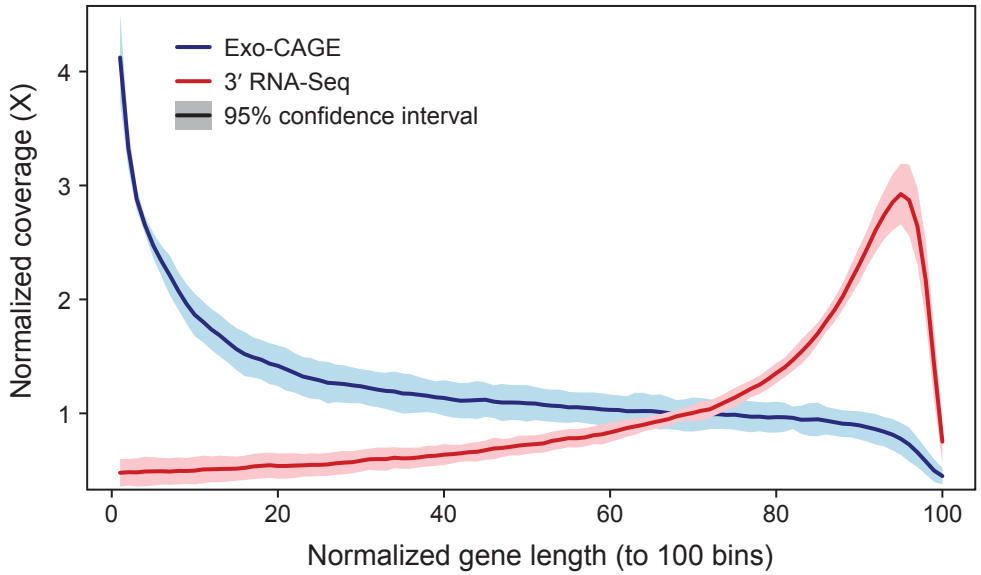
RNA-Seq libraries: Summary



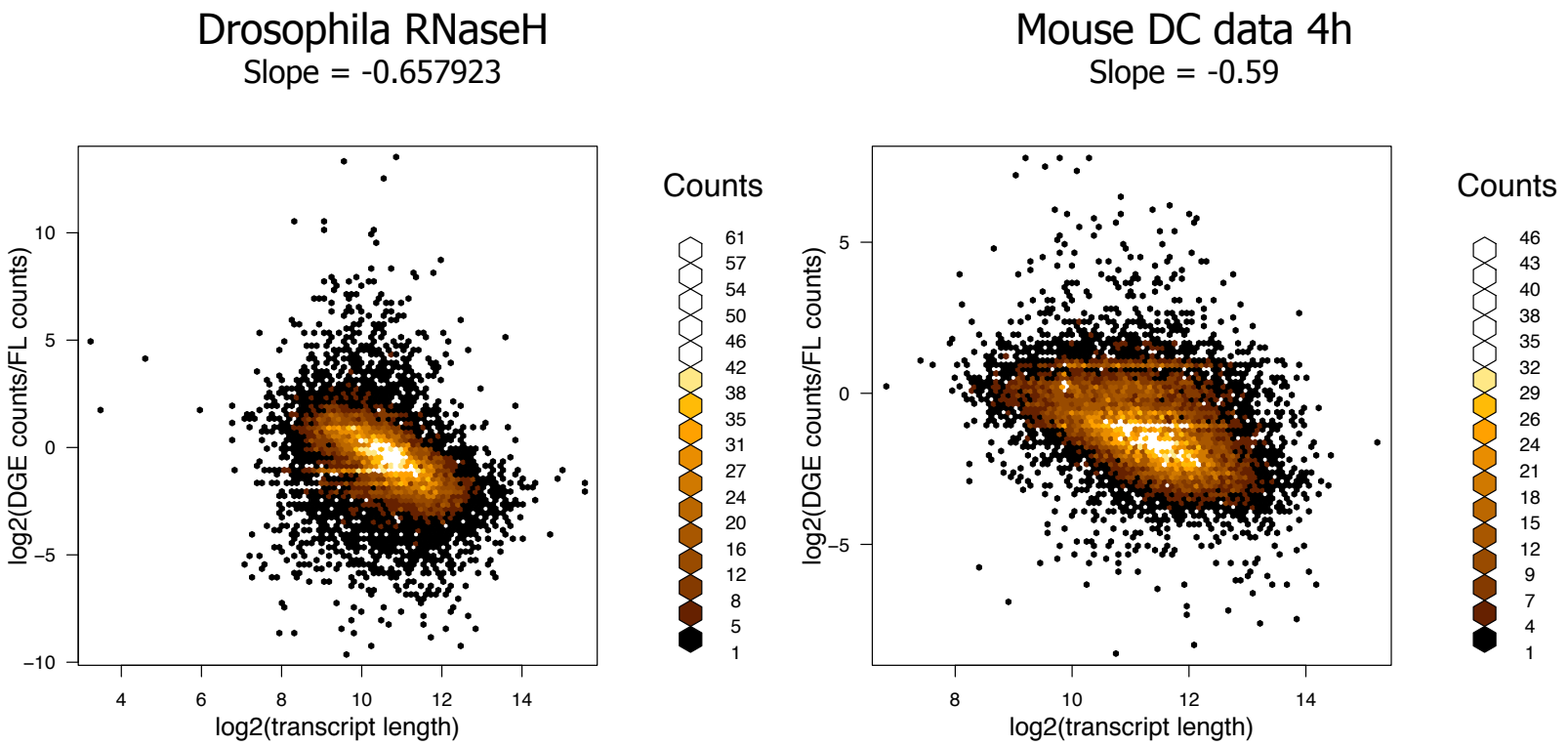
Sabah Kadri
Maxim Artyomov
Sebastian Kadener

RNA-Seq libraries for quantification

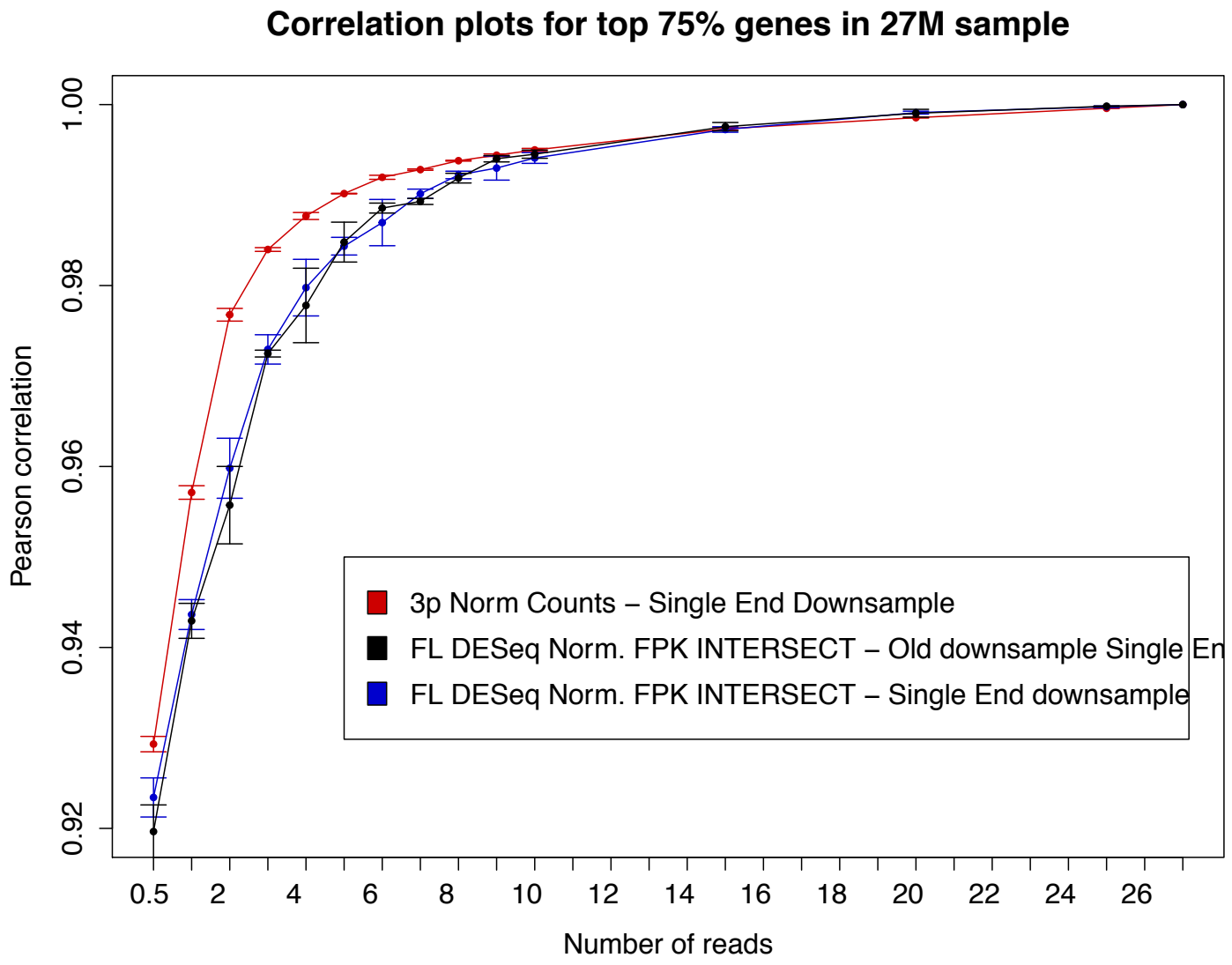




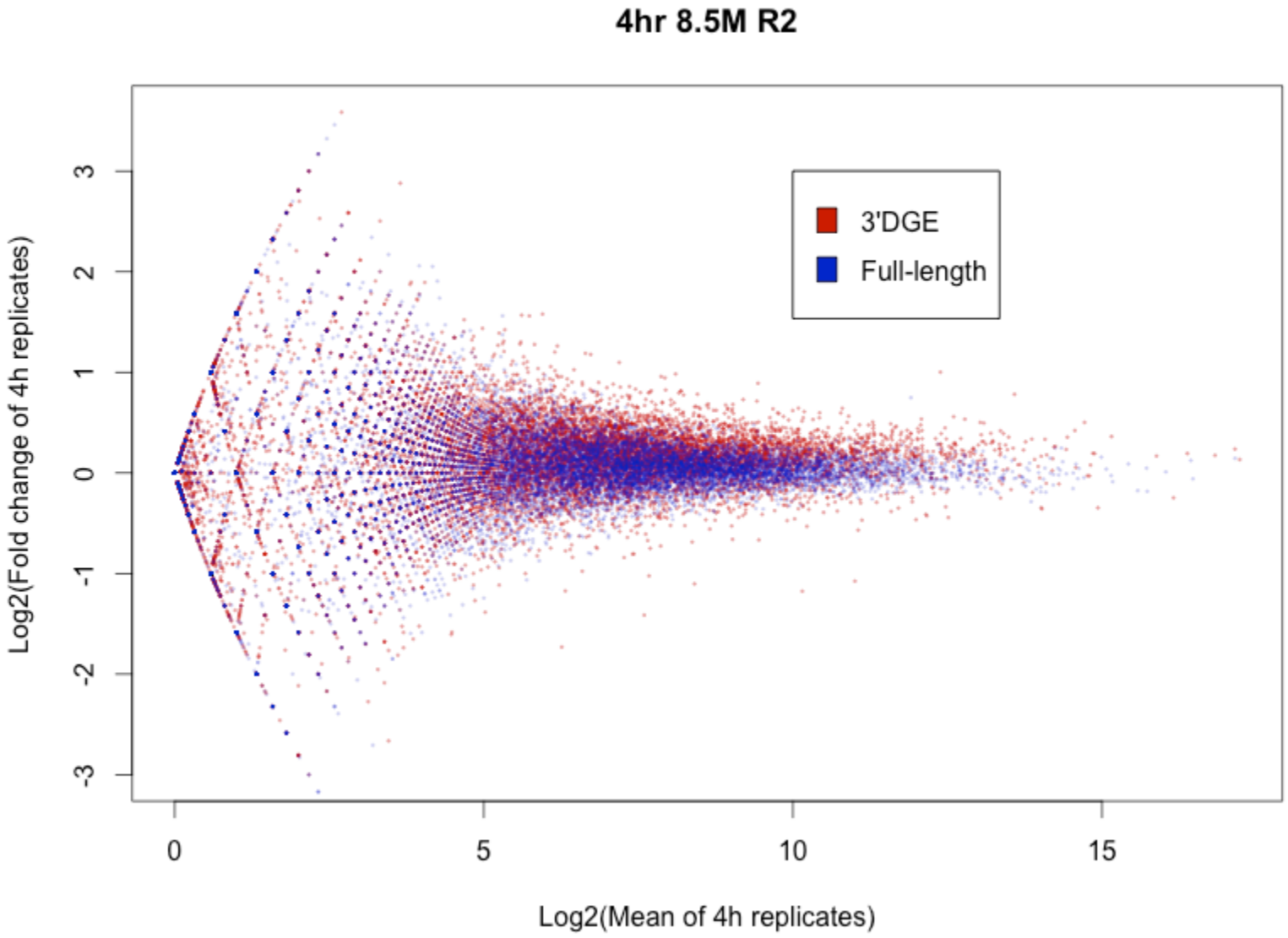
End sequence controls for length bias



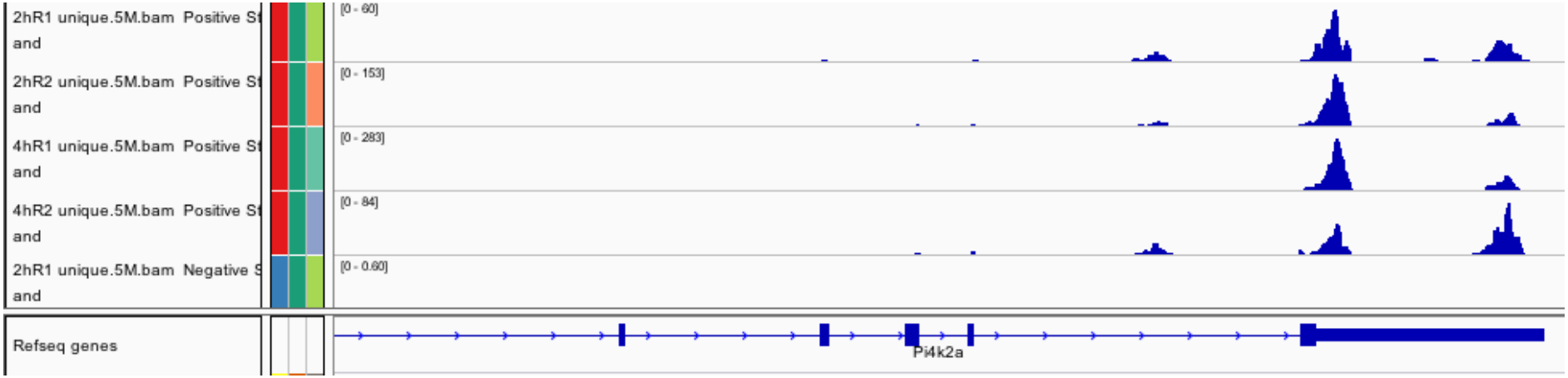
However no significant gain in power



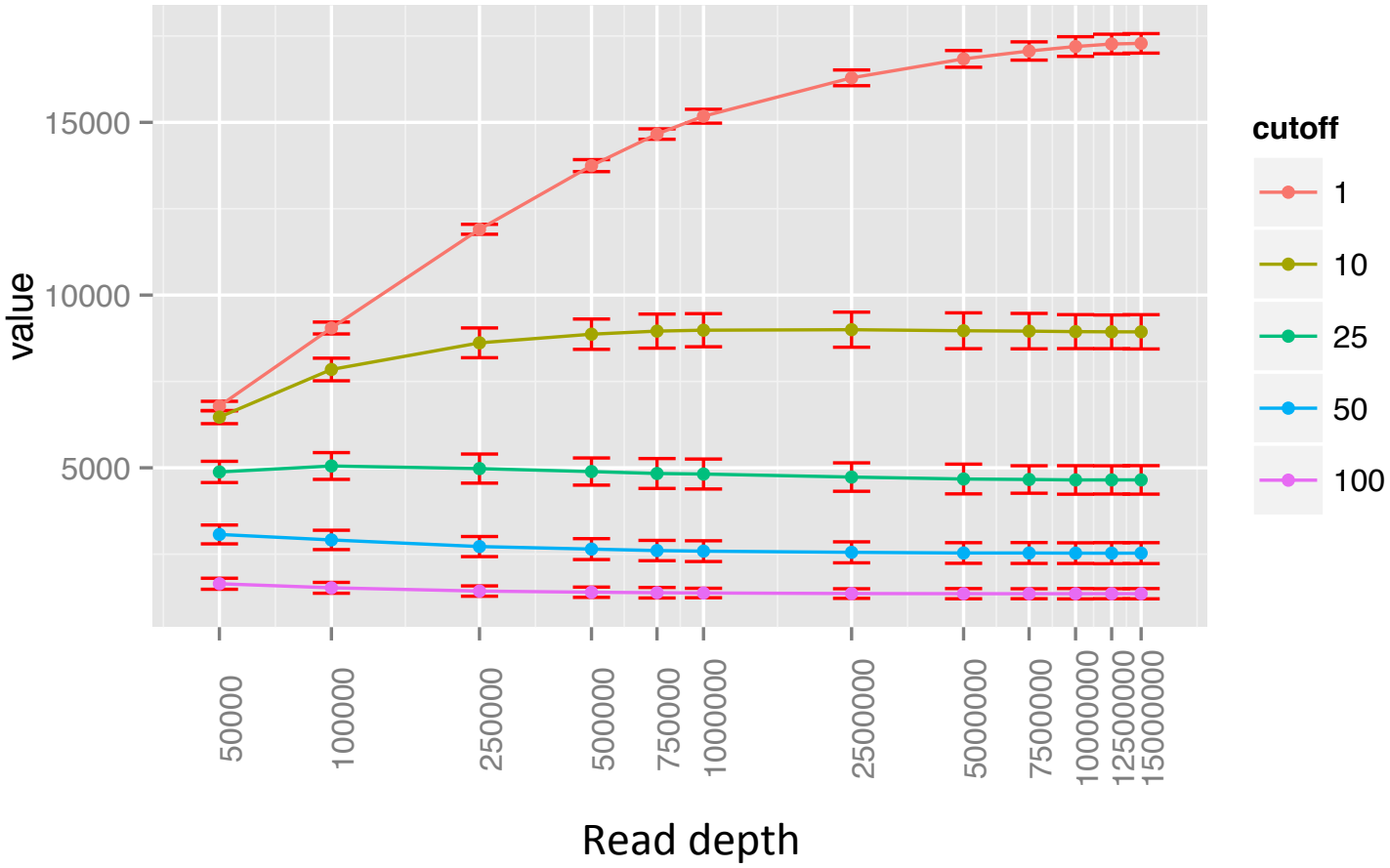
Surprisingly higher variance in DGE



3’ library are more variable

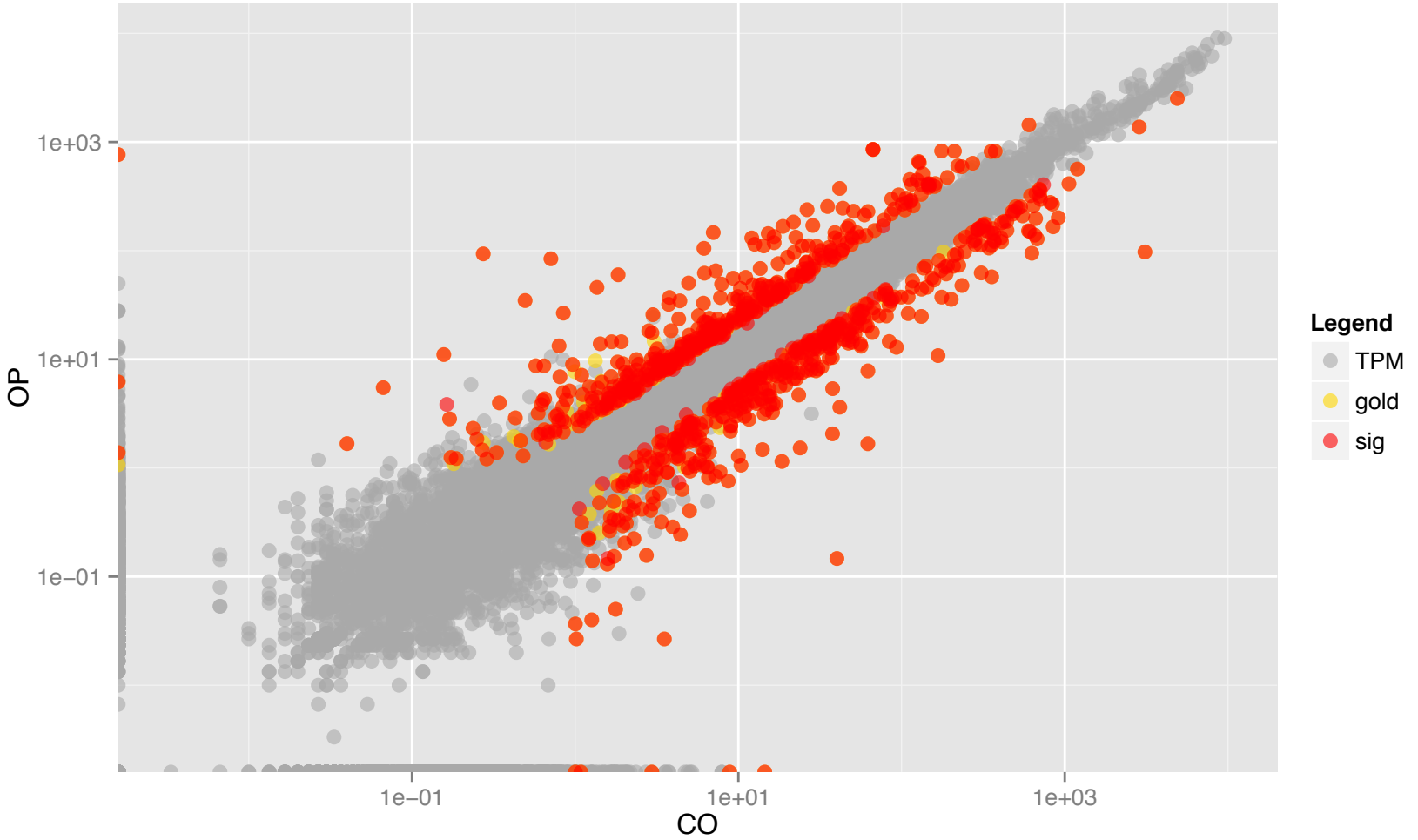


Robustness to low depth: Transcripts detected



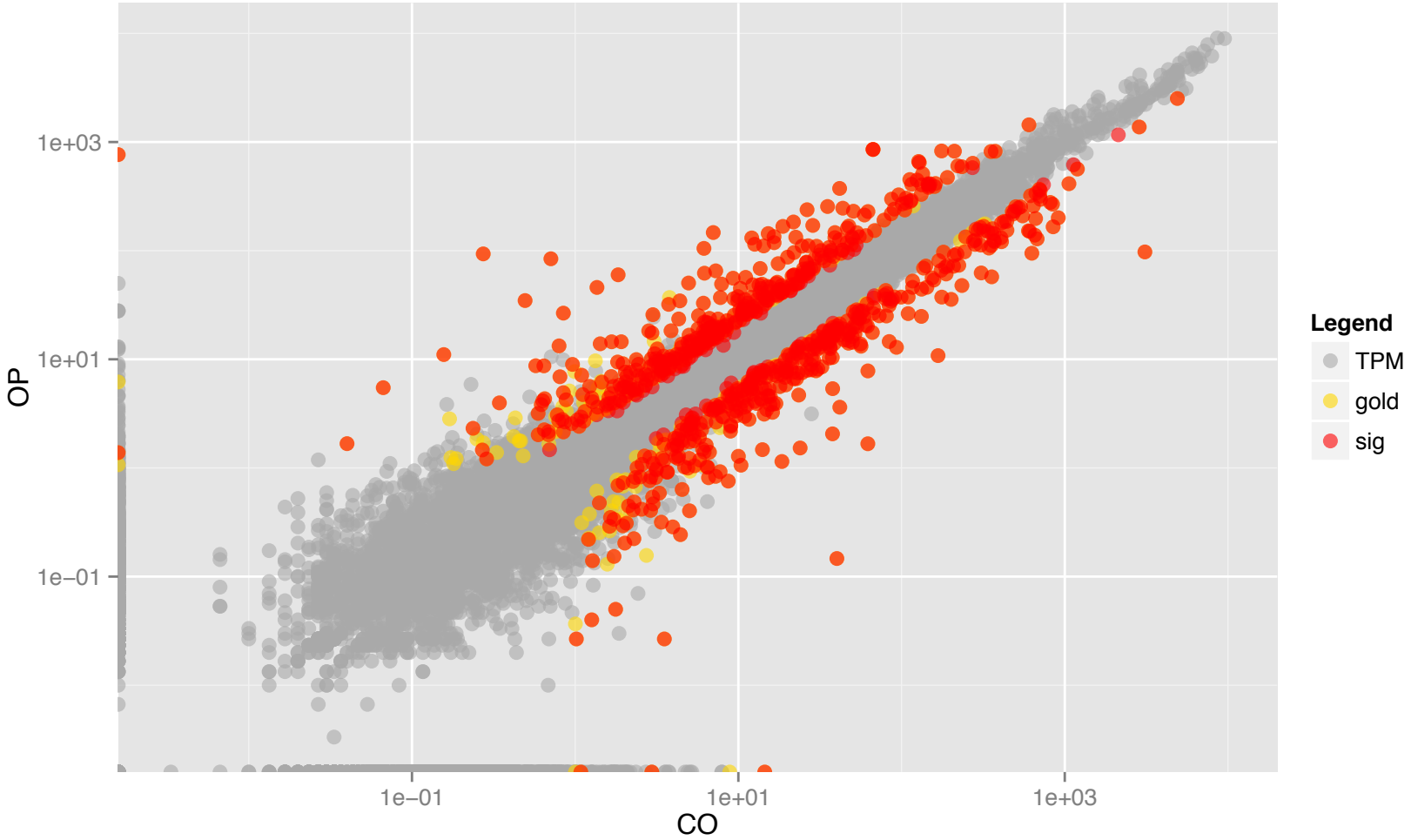
Alper Kucukural

RSEM/DESeq: 15 mill reads in worm



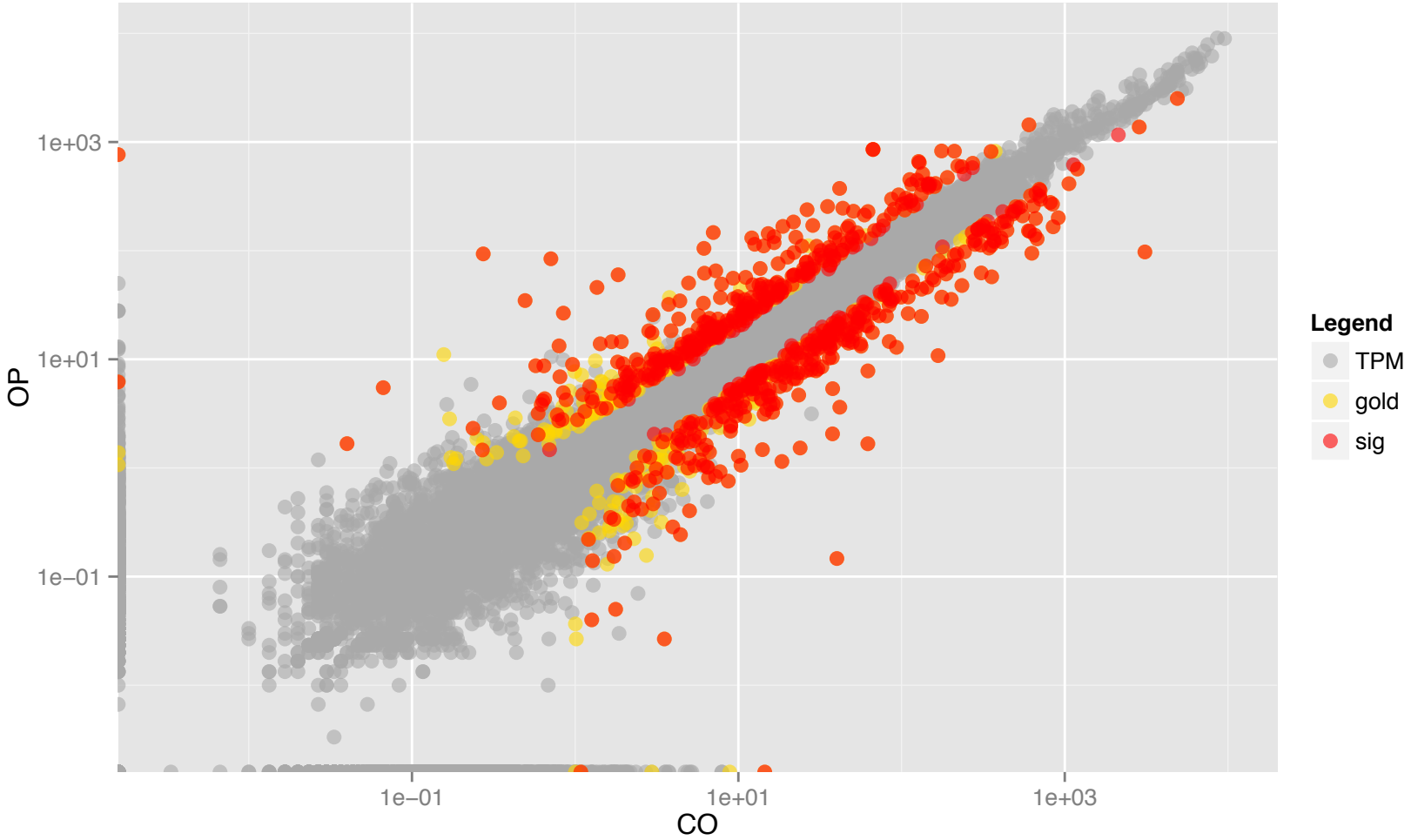
Alper Kucukural

RSEM/DESeq: 10 mill reads in worm



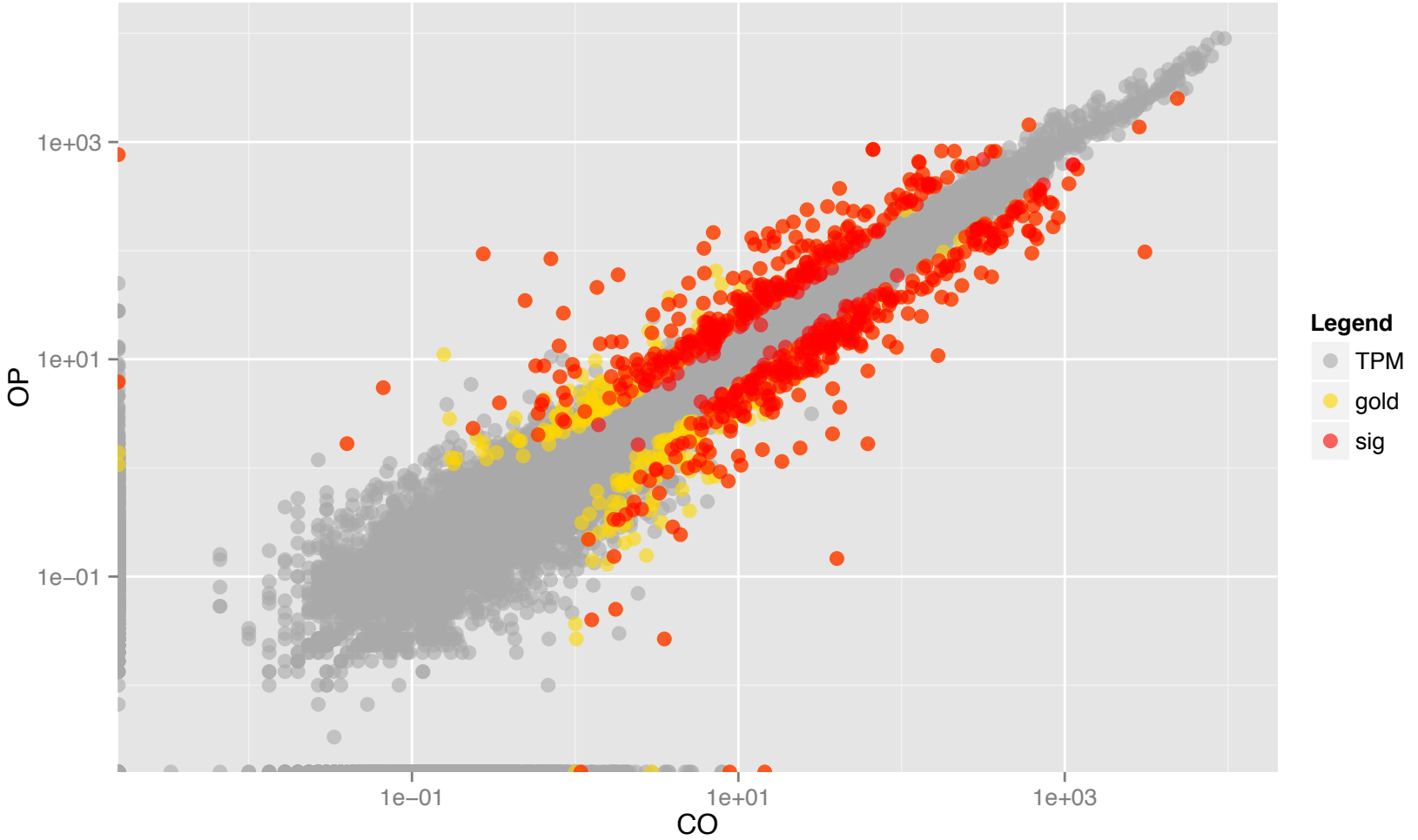
Alper Kucukural

RSEM/DESeq: 7.5 mill reads in worm



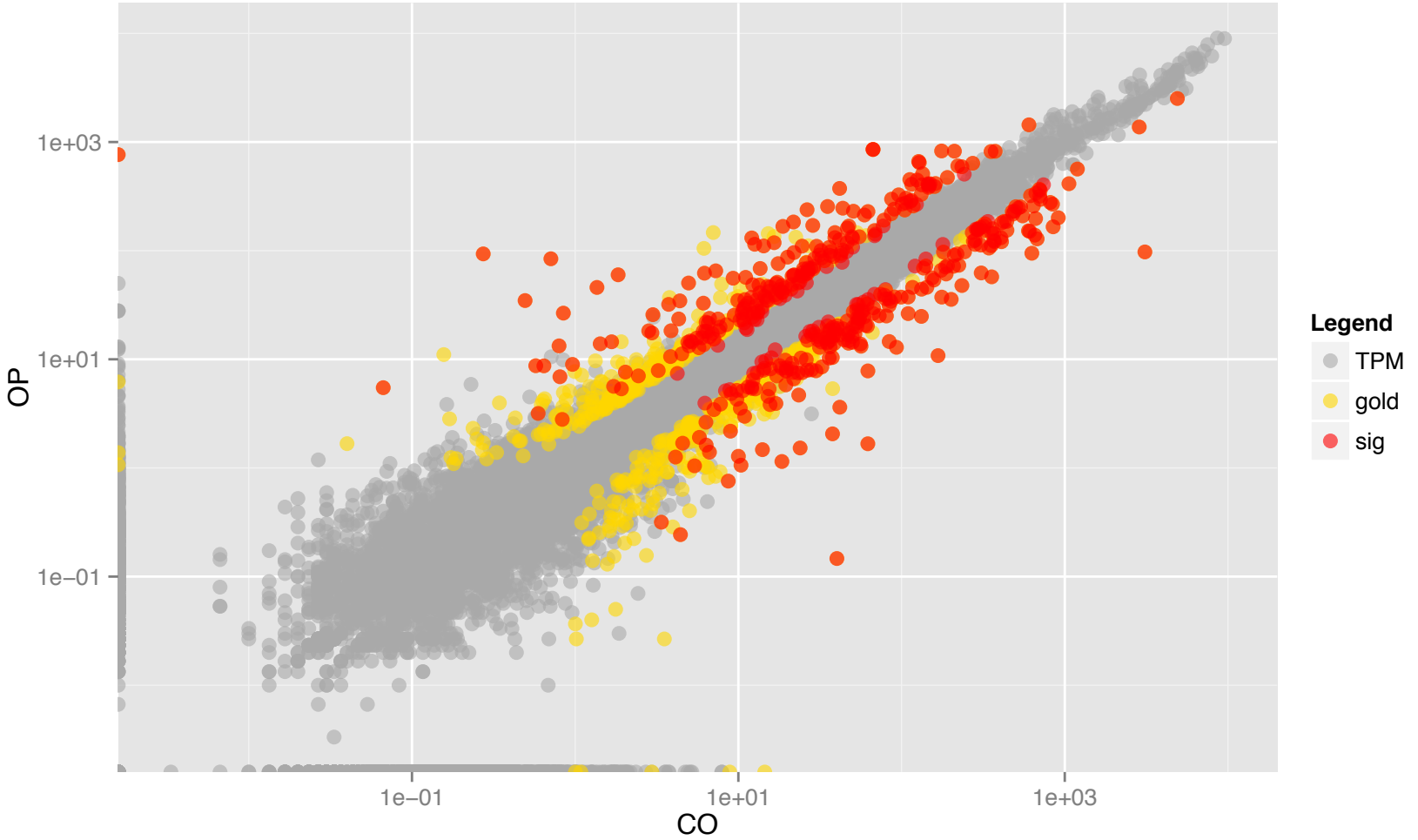
Alper Kucukural

RSEM/DESeq: 5 mill reads in worm

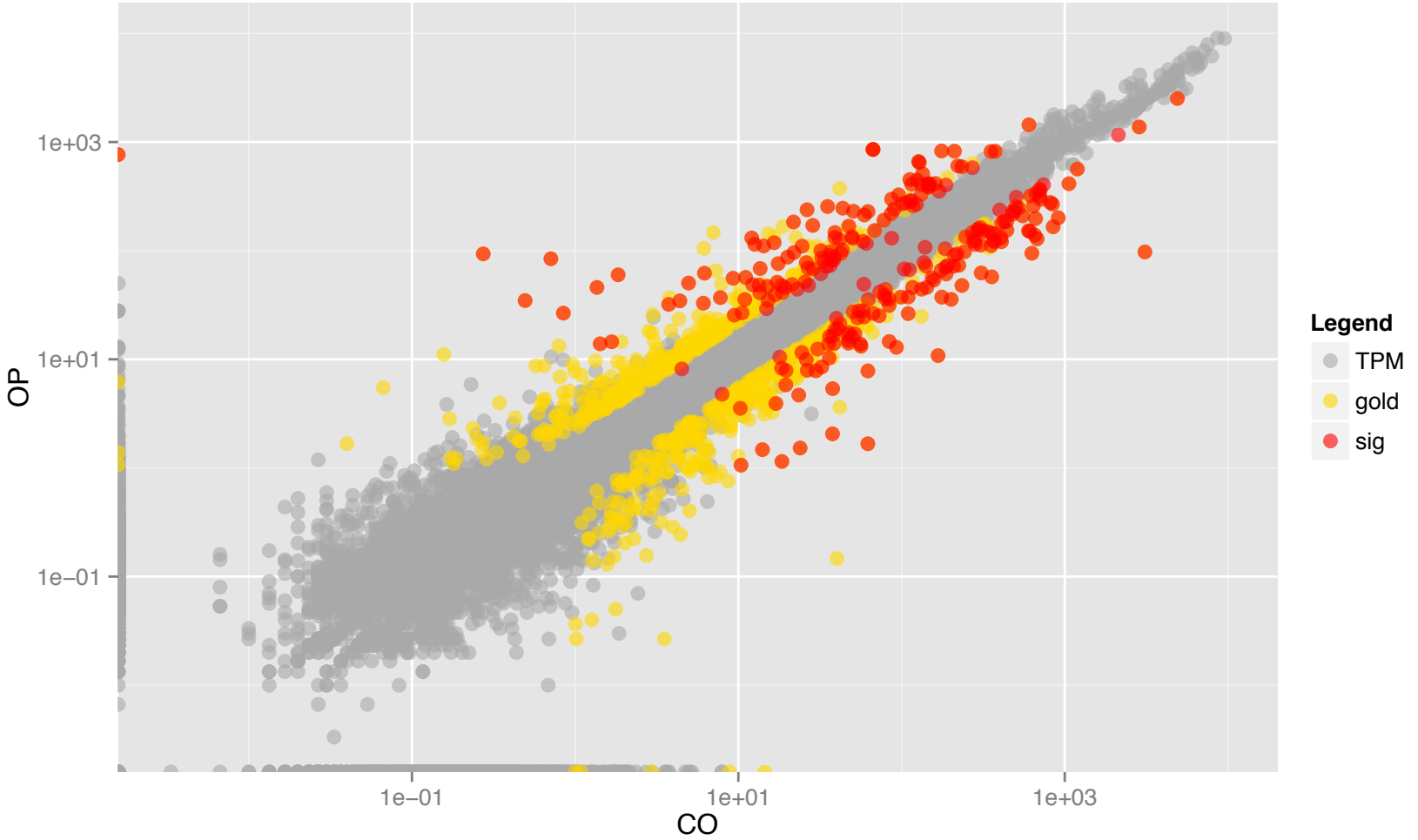


Alper Kucukural

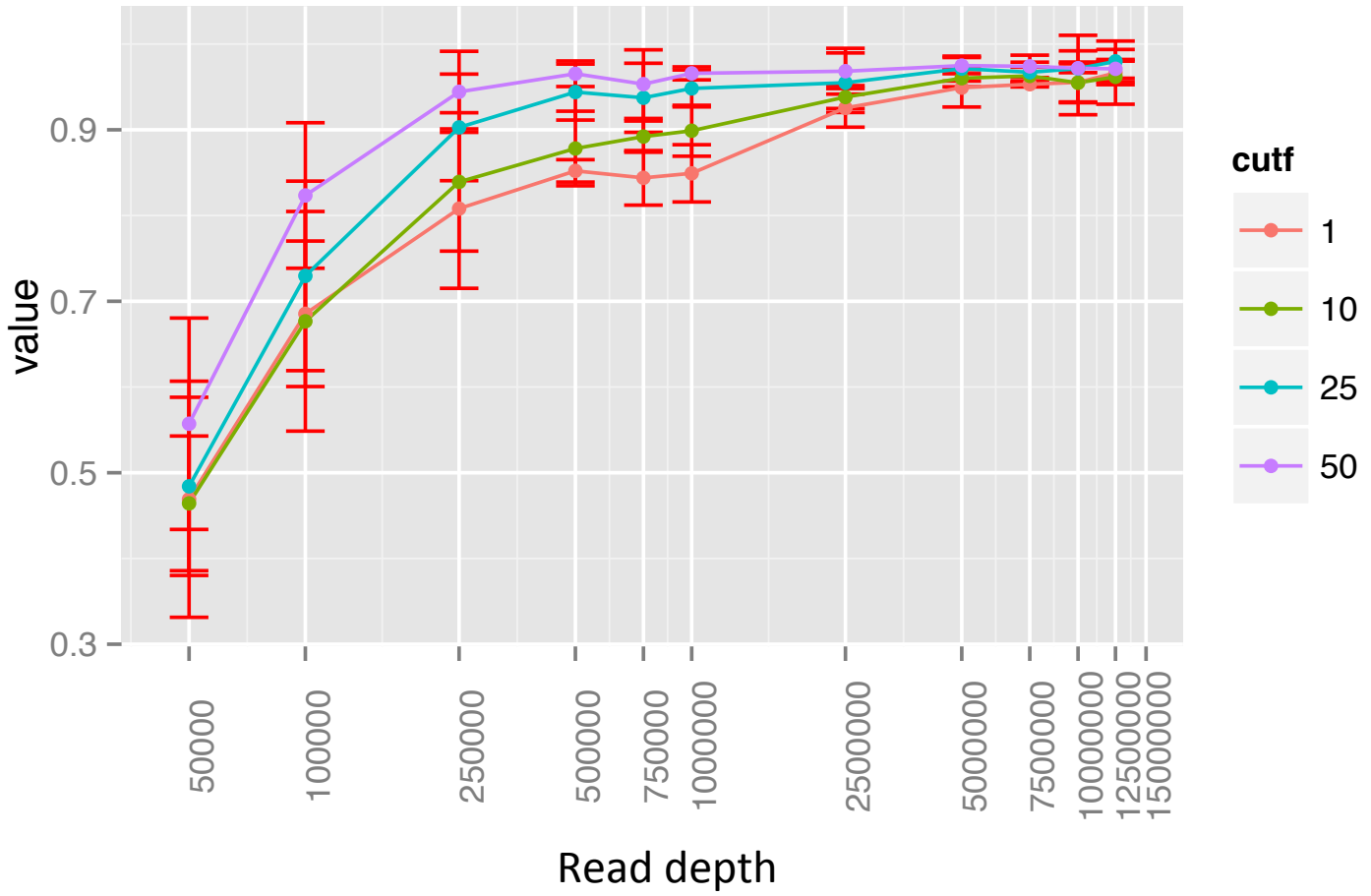
RSEM/DESeq: 2.5 mill reads in worm



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

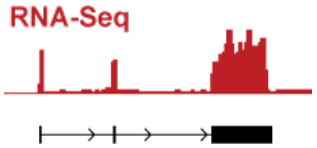
What does significance means?

- RNA-Seq: The gene is expressed
- ChIP-Seq: Factor binds the region
- CLIP-Seq: Protein binds RNA region

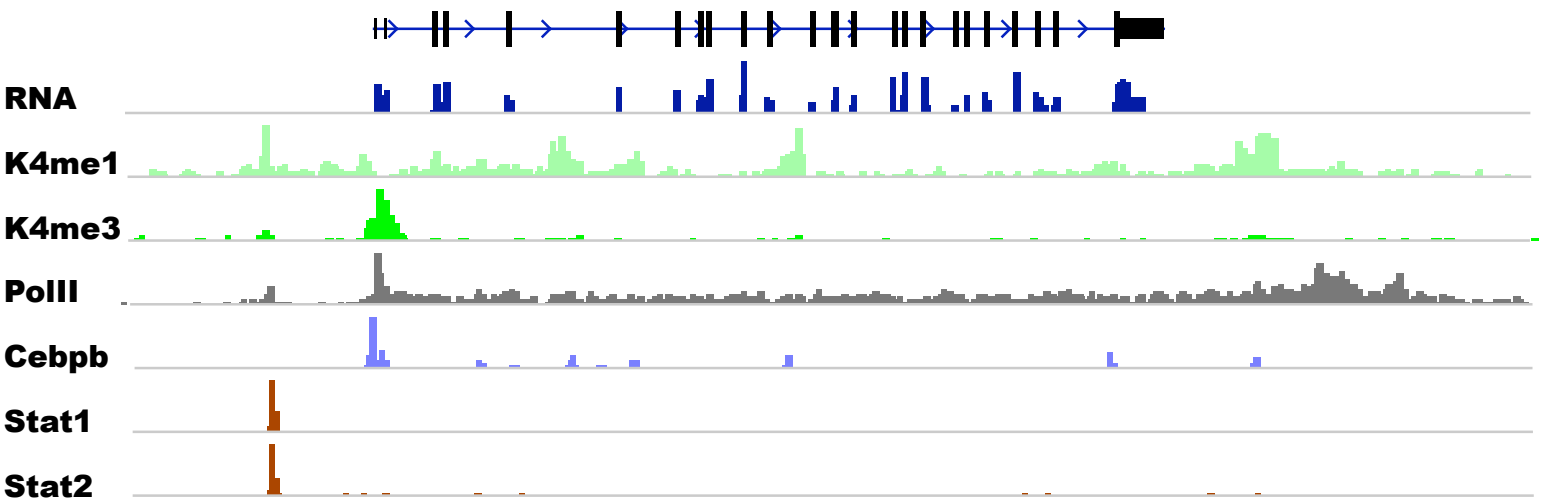
How do we find peaks?

Short modification

Long modification

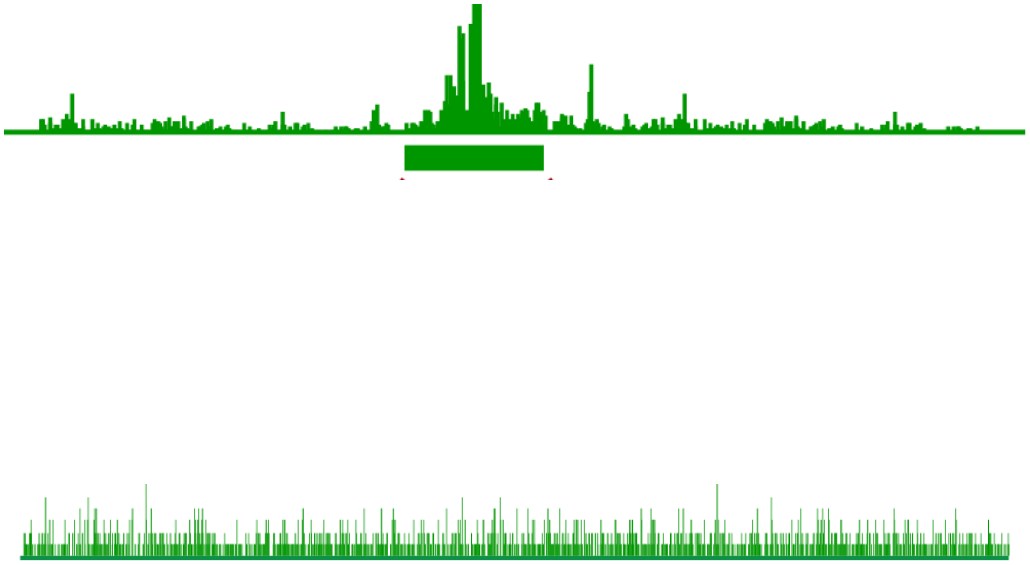


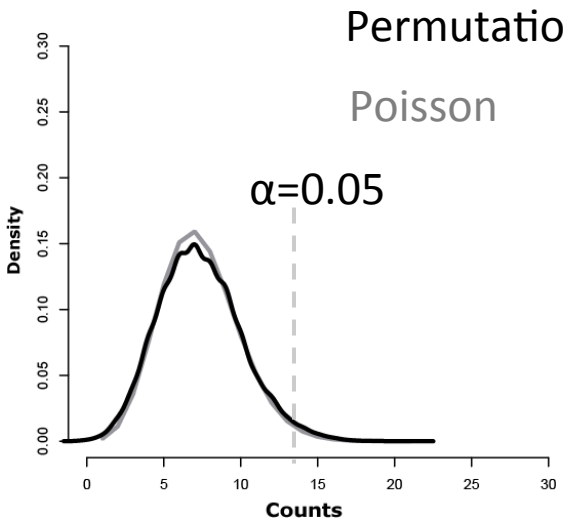
Discontinuous data



Scripture is a method to solve this general question

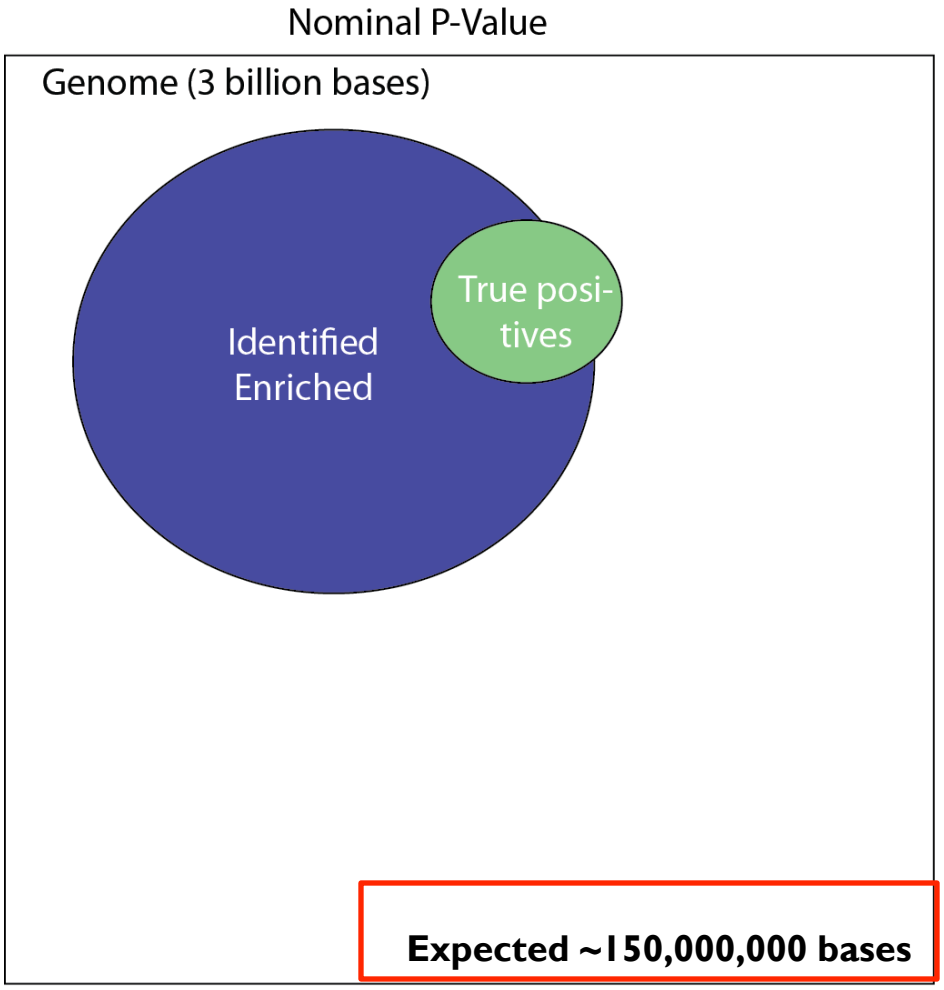
Our approach





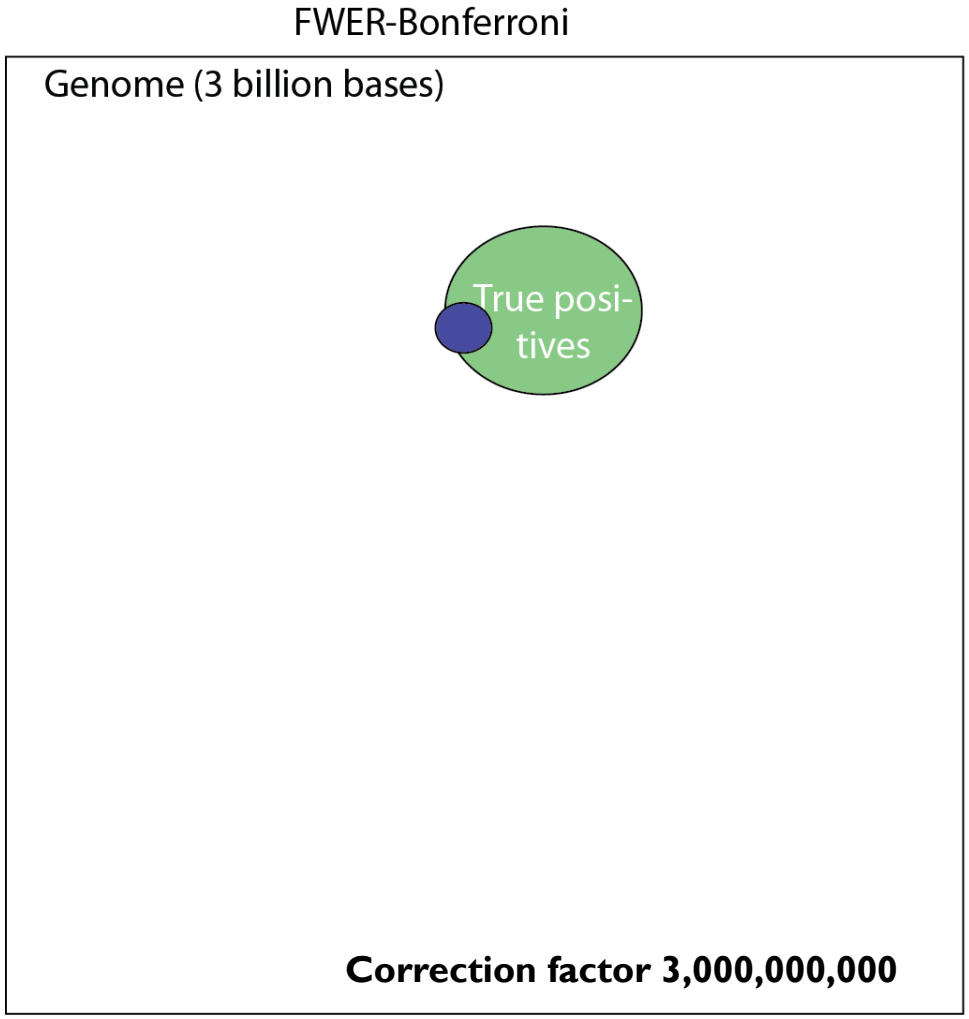
We have an efficient way to compute read count p-values …

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative

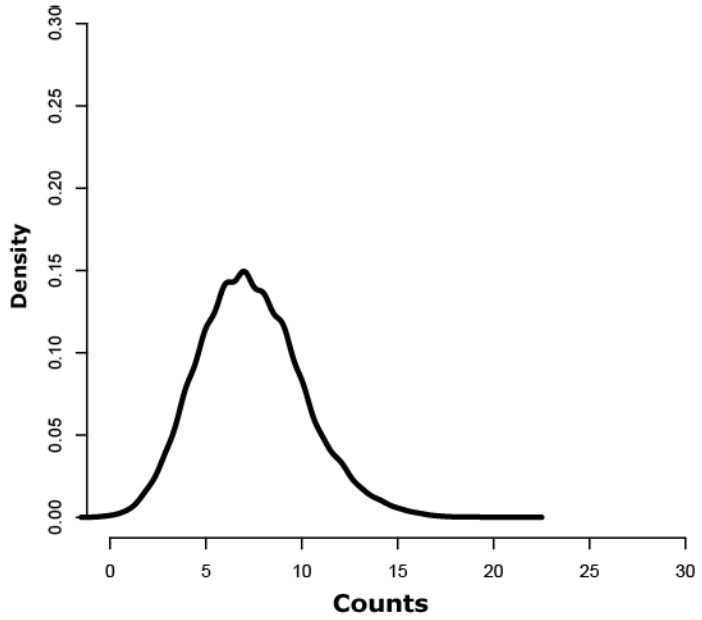


Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

Controlling FWER

Max Count distribution

α=0.05 α_{FWER}=0.05



Count distribution (Poisson)

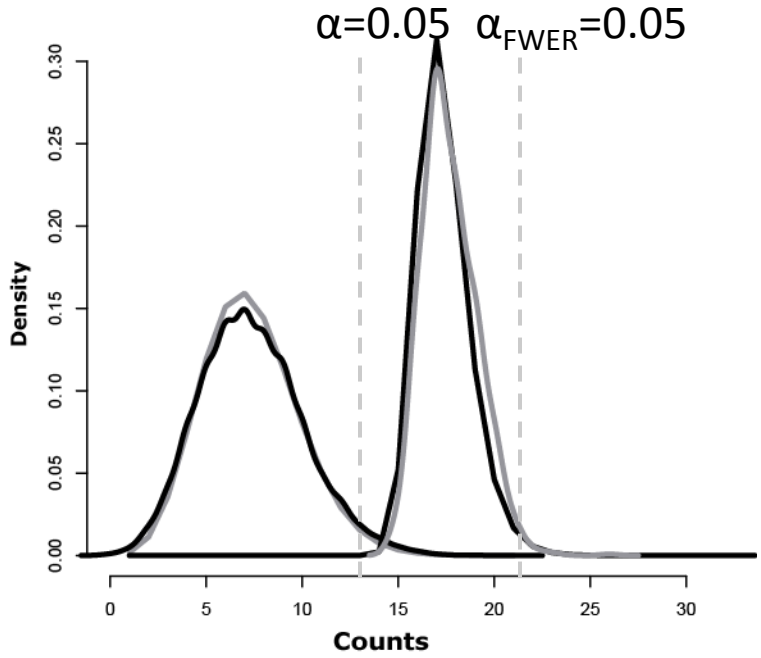
Given a region of size w and an observed read count n. What is the probability that one or more of the $3x10^9$ regions of size w has read count >= n under the null distribution?

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

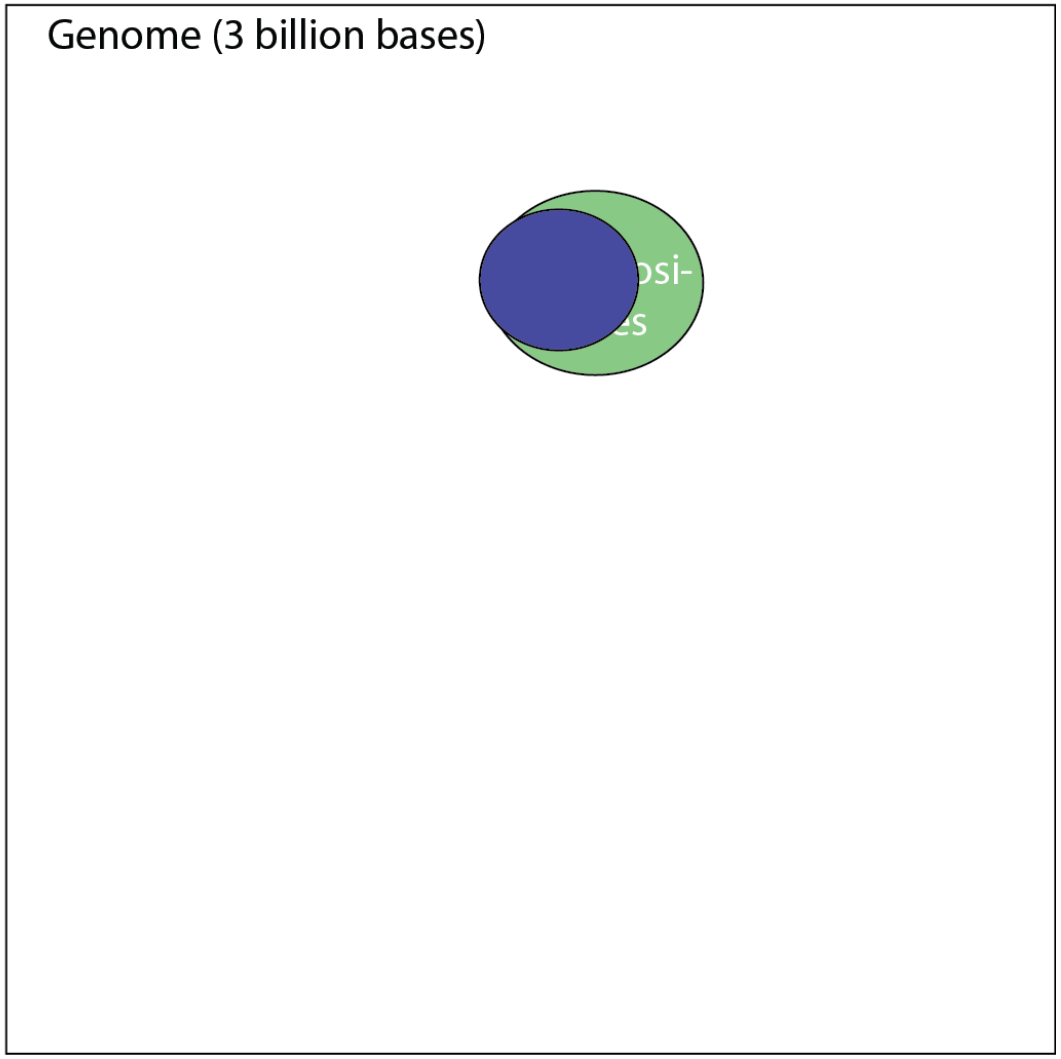
$P(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ counts given an expected count of λw

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

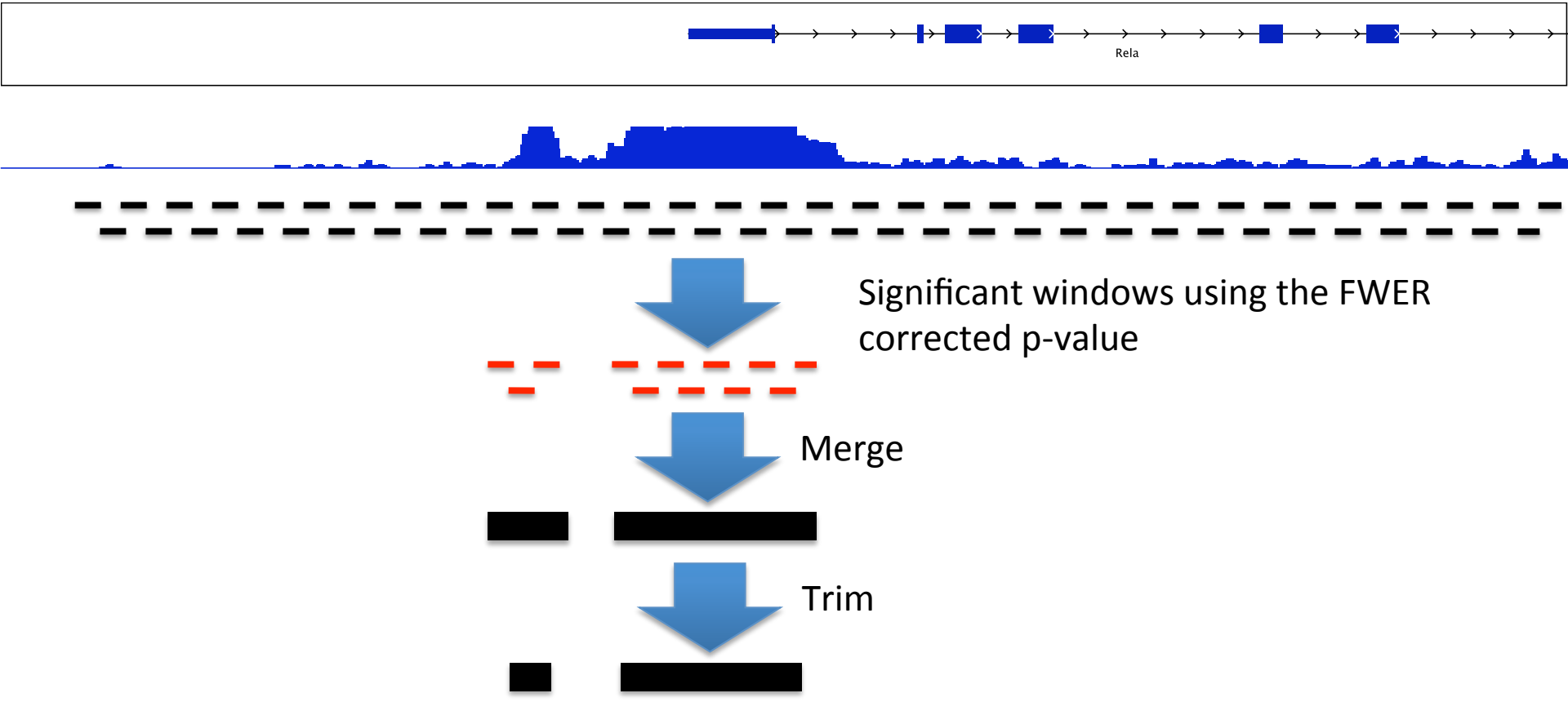
FWER-Scan Statistics



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

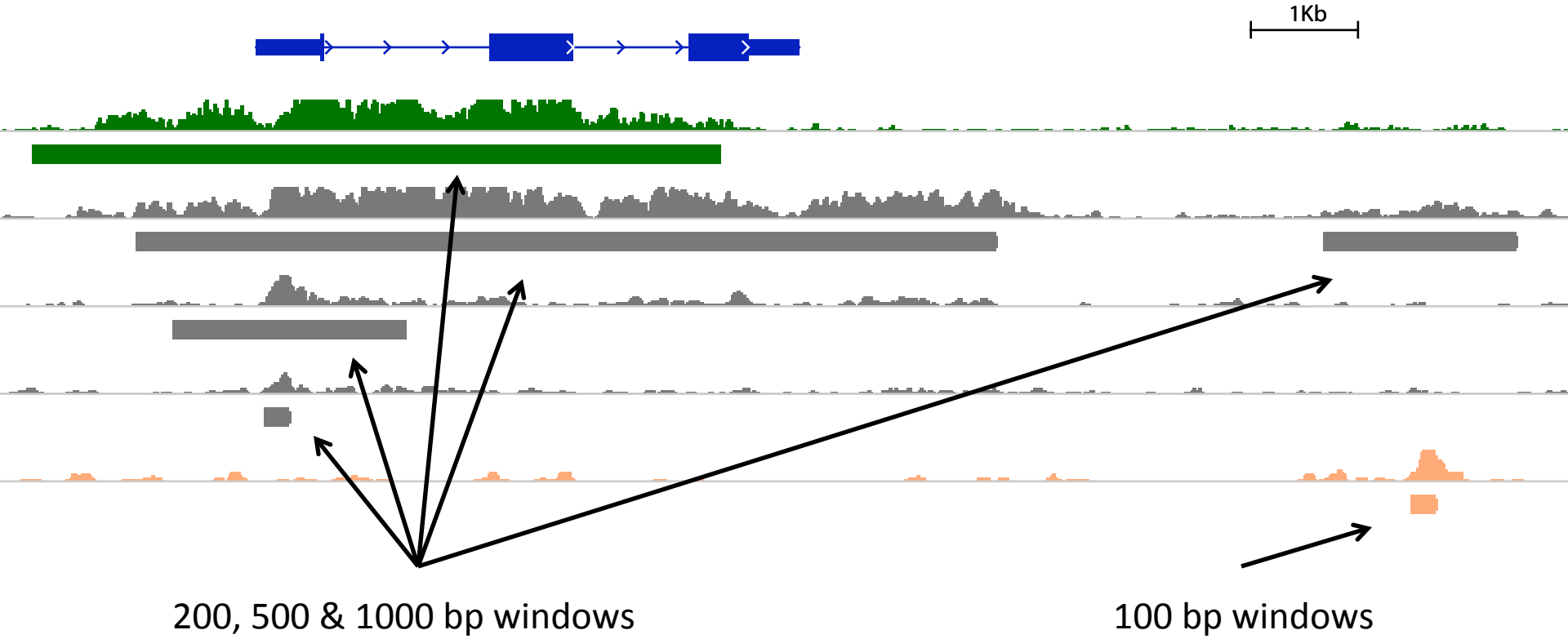


But, which window?

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

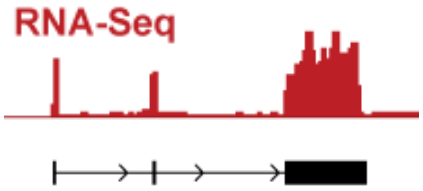


Can we identify enriched regions across different libraries?

Short modification ✓

Long modification ✓

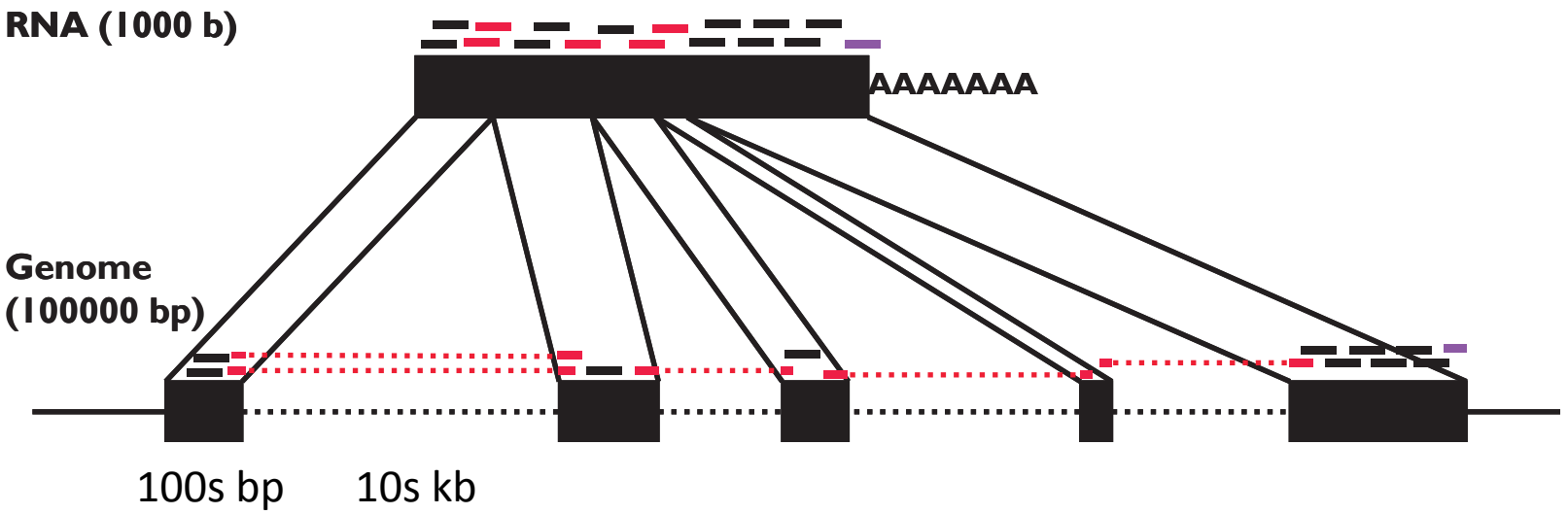
Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?



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

Scripture for RNA-Seq:
Extending segmentation to discontiguous regions

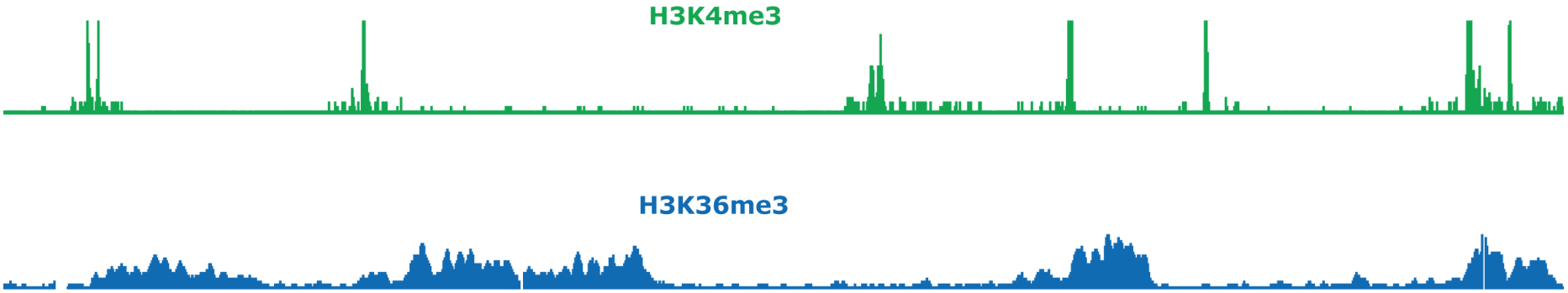
Transcript reconstruction problem as a segmentation problem



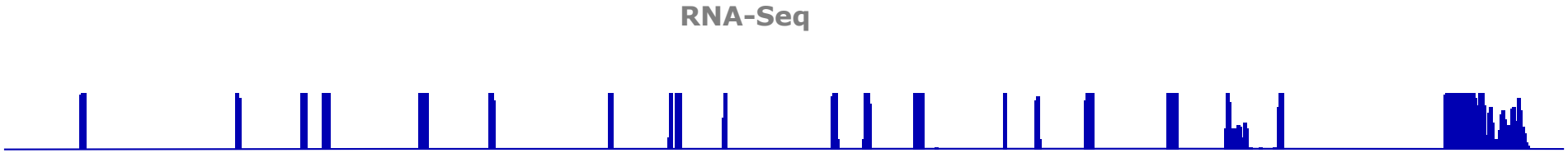
Challenges:

- 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

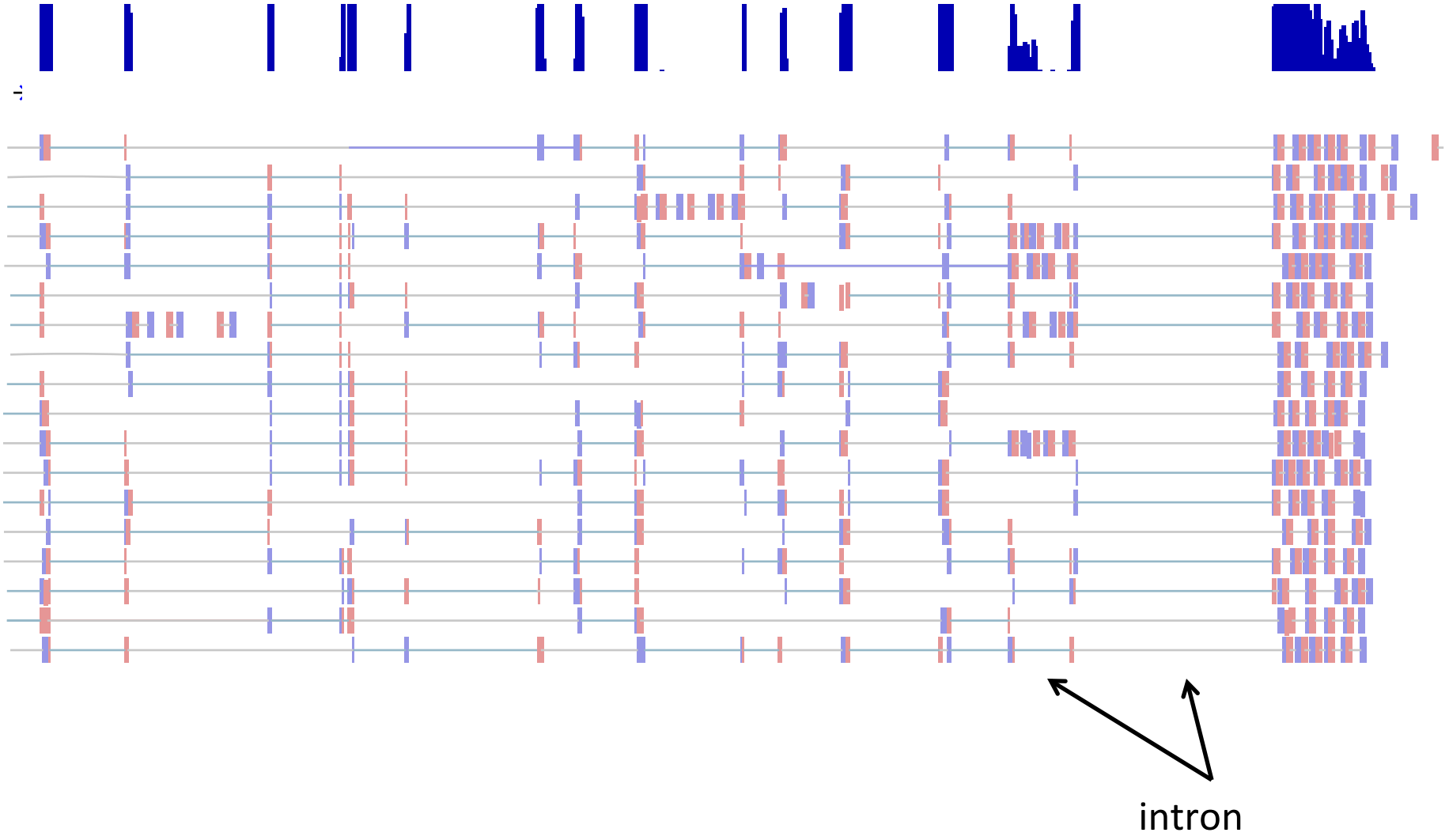


Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



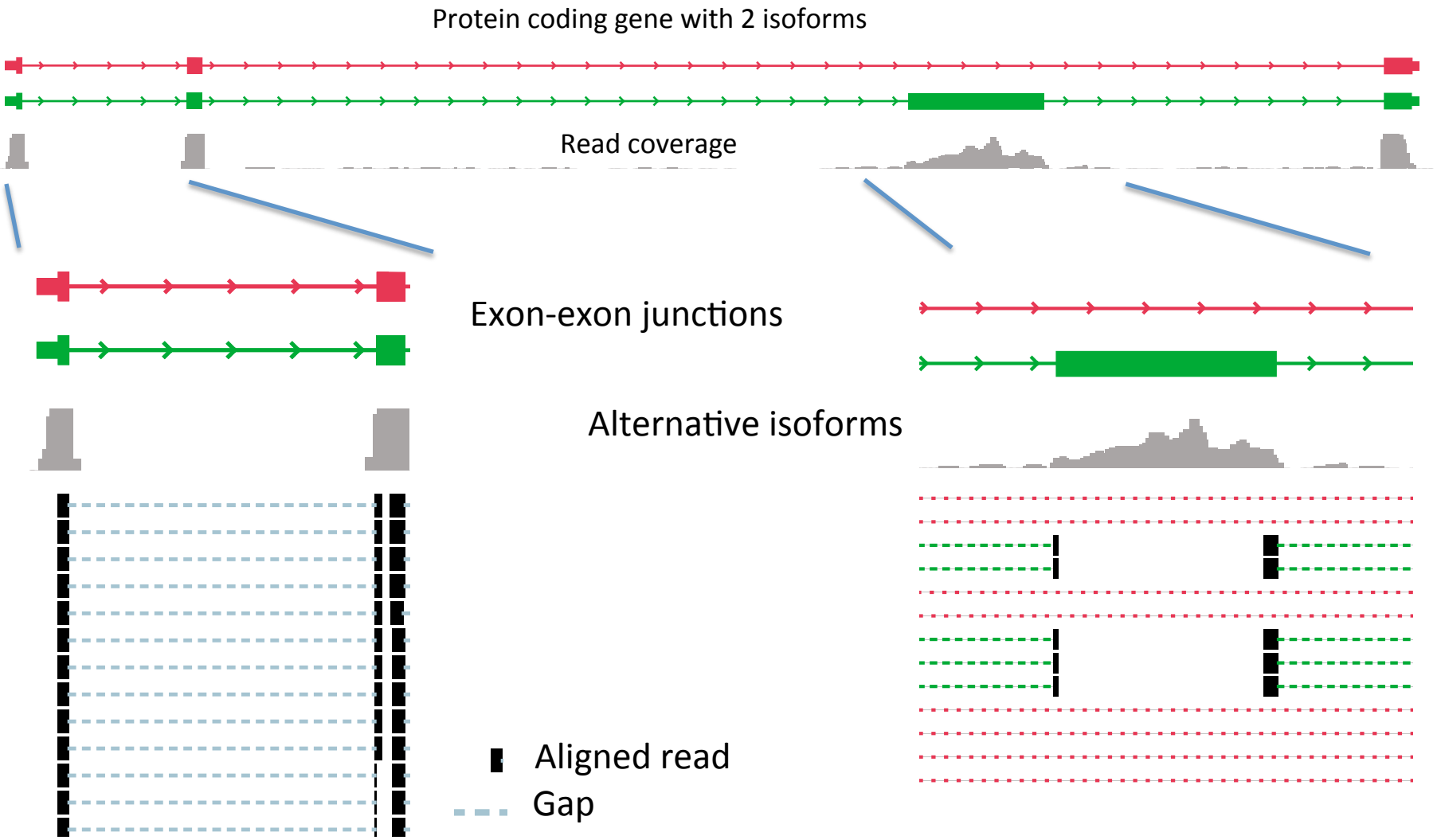
If we know the connectivity of fragments, we can increase our power to detect transcripts

Longer (76) reads increased number of junction reads



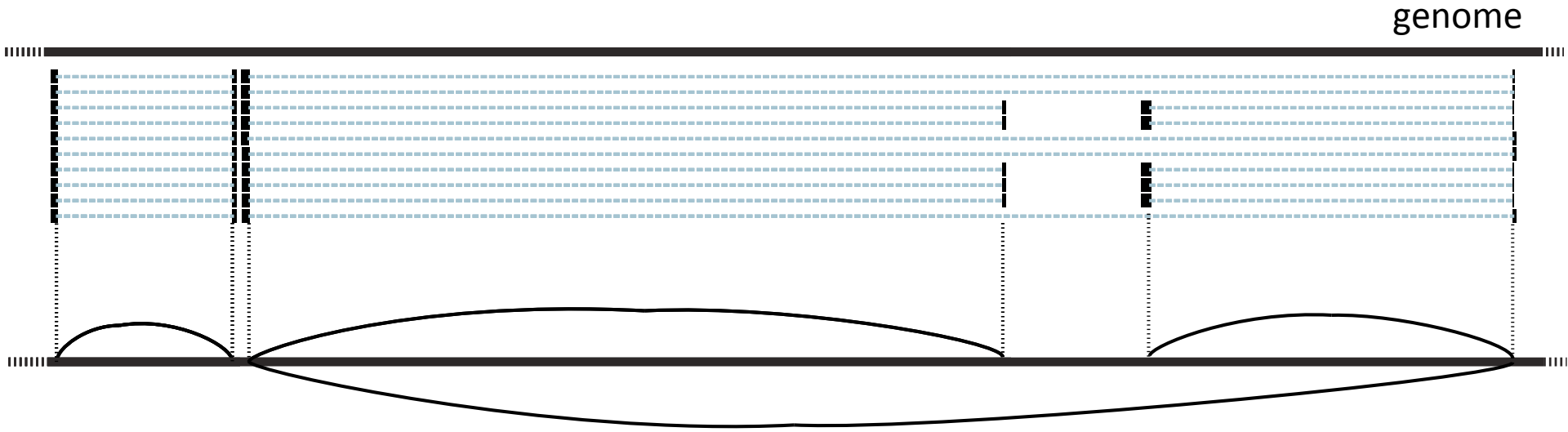
Exon junction spanning reads provide the connectivity information.

The power of spliced alignments



Statistical reconstruction of the transcriptome

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

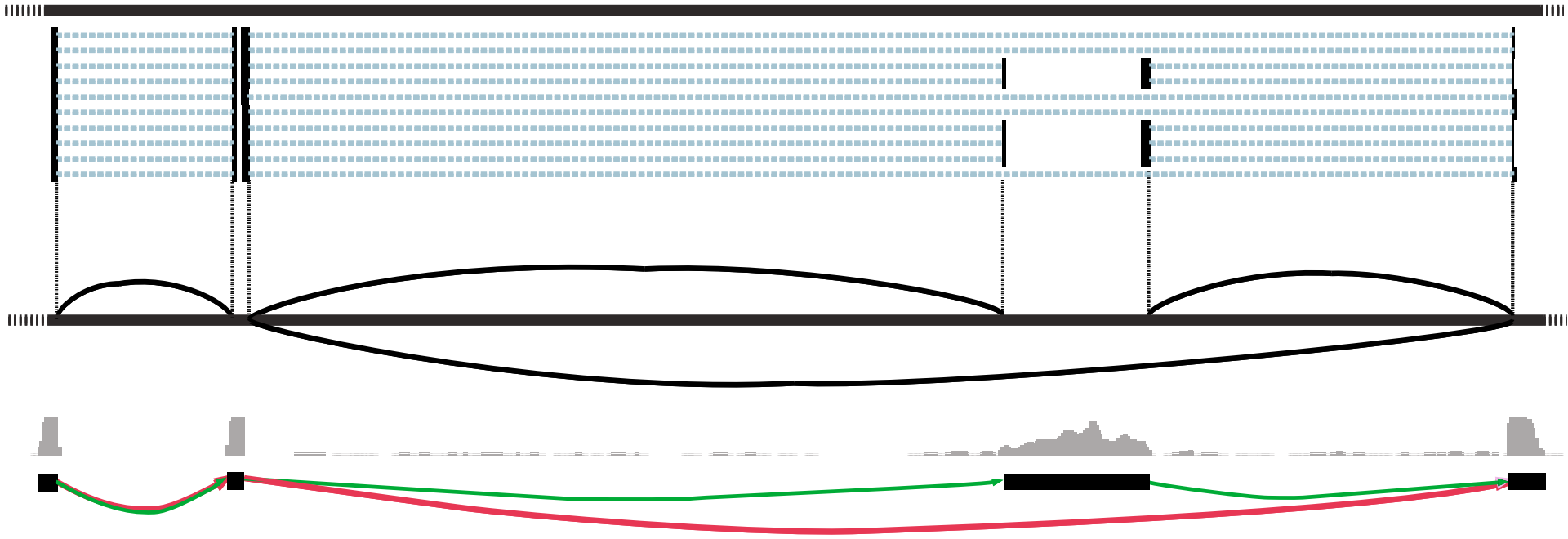


Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

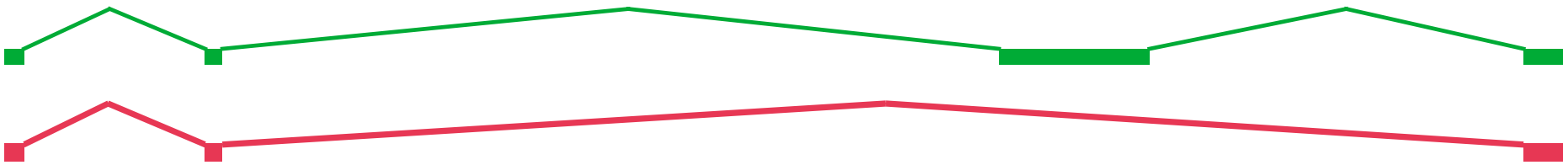
The "connectivity graph" connects all bases that are directly connected within the transcriptome

Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



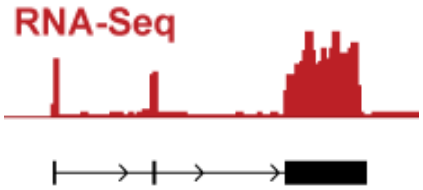
Step 4: Find significant segments



Can we identify enriched regions across different data types?

Short modification ✓

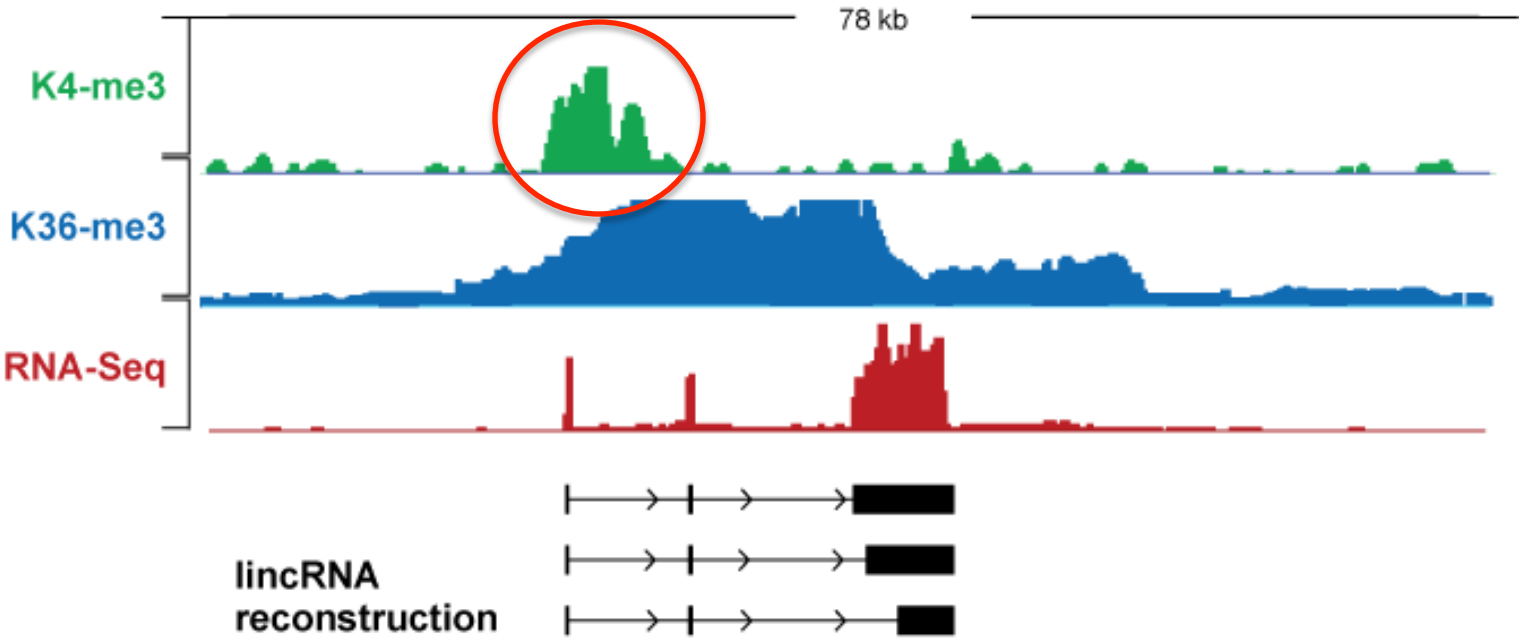
Long modification ✓



Discontinuous data ✓

Are we really sure reconstructions are complete?

RNA-Seq data is incomplete for comprehensive annotation

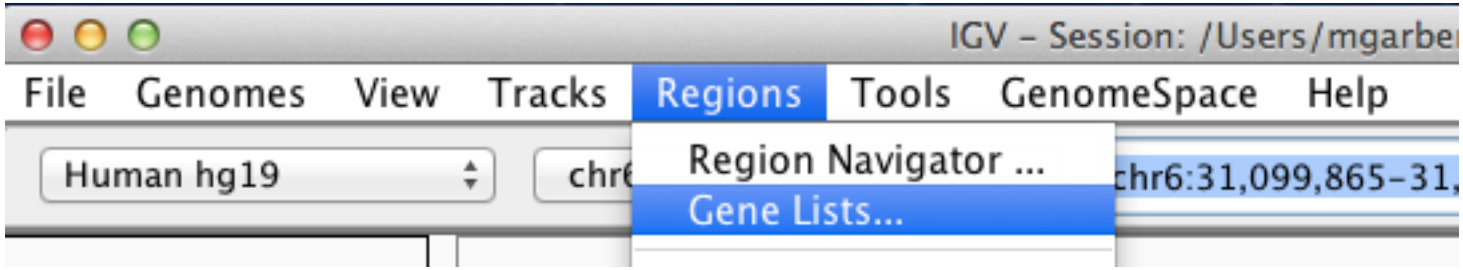


Library construction can help provide more information. More on this later

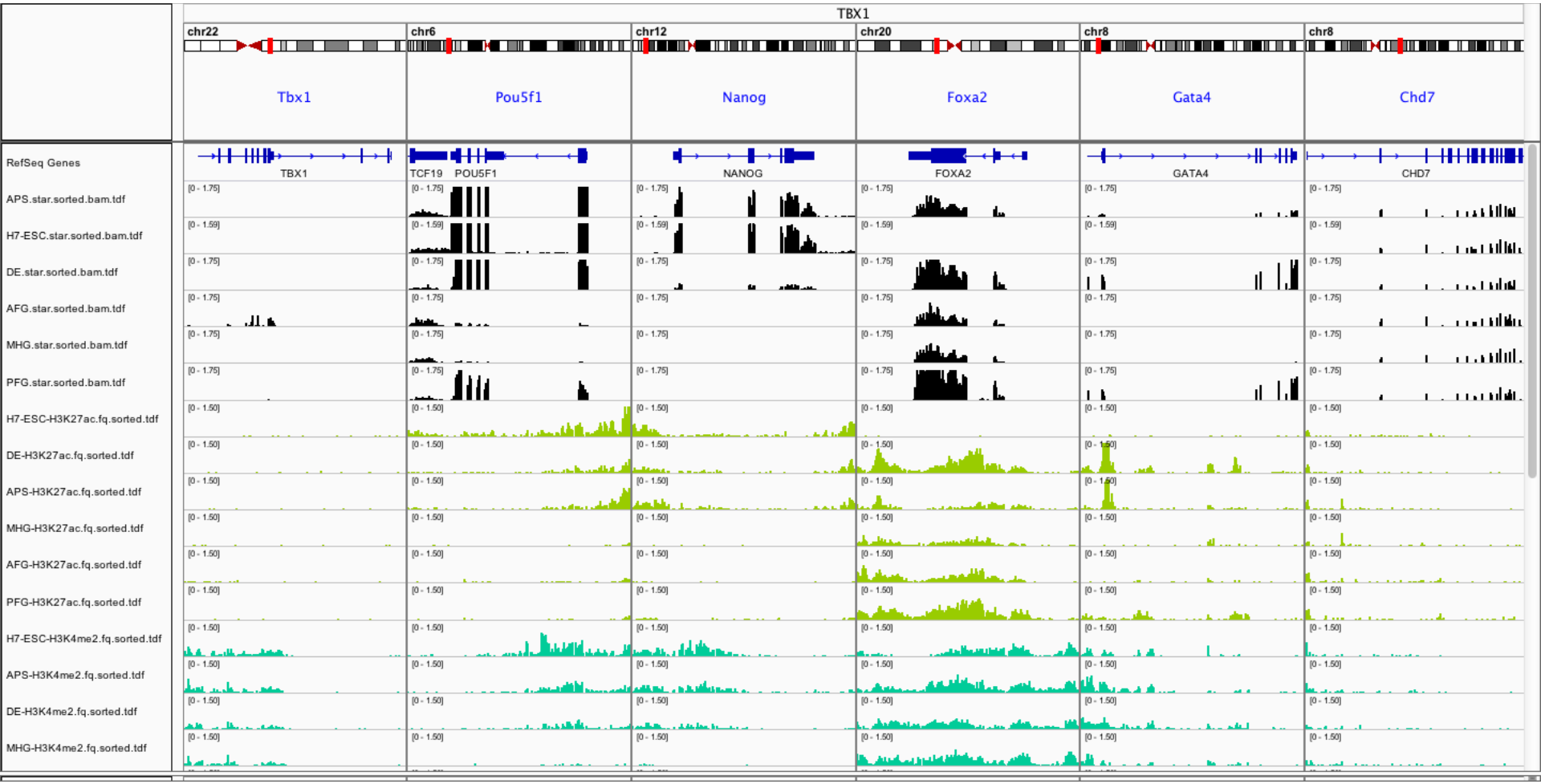
Visualization tricks & Tips

- Viewing normalized data
- Downsampling reads to avoid crashes
- Gene lists
- Sessions

Viewing several loci simulateneouly: Gene lists

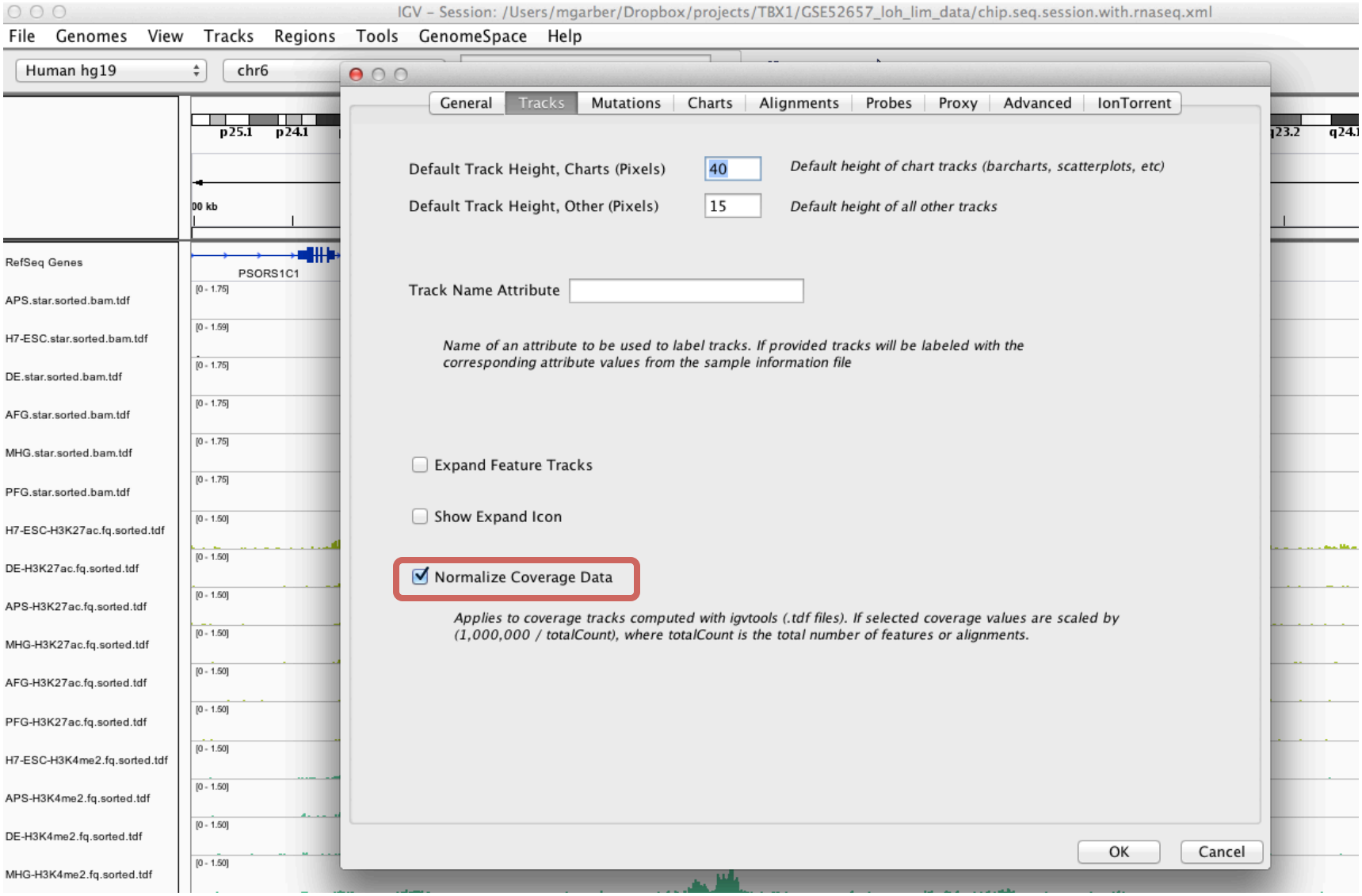


Viewing several loci simulateneouly: Gene lists



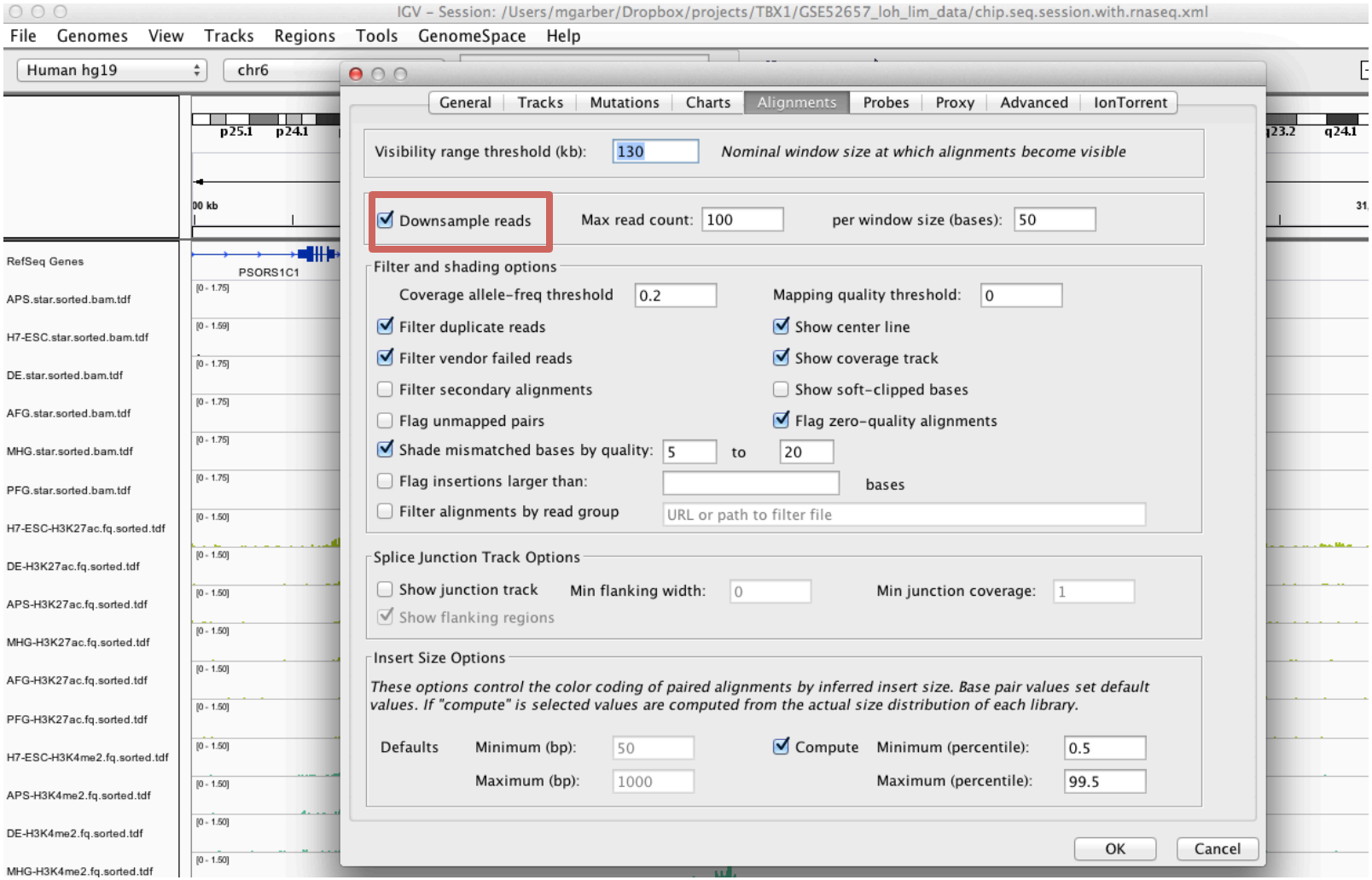
Normalizing tracks

You can use simple read depth normalization for comparison of different tracks



Configure your alignment display

Downsampling reads is critical when loading the full alignments.
When you are loading reads, downsampling ensures that regions with high coverage result in IGV running out of memory.



Saving sessions

Sessions allows you to store a set of desired tracks along with any setting you want