

# RNA Sequencing

---

Kasper D Hansen

[<khansen@jhsph.edu>](mailto:khansen@jhsph.edu)

Statistical Methods for Next Generation Sequencing

ENAR 2012

Slides contain material from  
Ben Langmead, Margaret Taub

# The Question(s) and the essay

# RNAs

---

poly-adenylated (coding) RNAs, "genes"

short non-coding RNAs (ncRNA), "microRNA"

long non-coding RNAs

ribosomal RNA



Total RNA

# RNAs

---

poly-adenylated (coding) RNAs, "genes"

short non-coding RNAs (ncRNA), "microRNA"

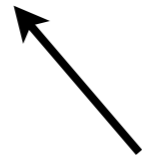
long non-coding RNAs

ribosomal RNA

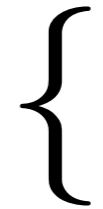


Total RNA

Enrichment



Most of the RNA in the cell



polyA capture  
ribominus

# RNAs

---

poly-adenylated (coding) RNAs, "genes"

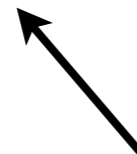
short non-coding RNAs (ncRNA), "microRNA"

long non-coding RNAs



Total RNA

ribosomal RNA



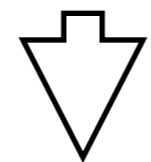
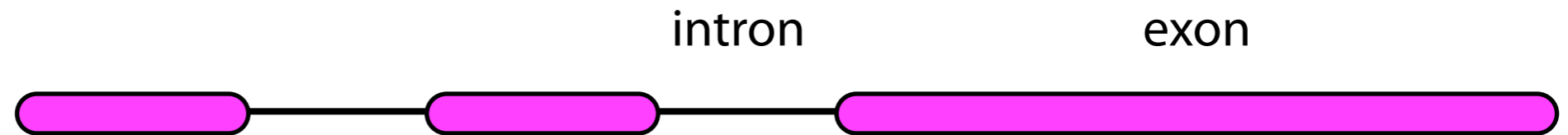
Most of the RNA in the cell



Enrichment

polyA capture  
ribominus

pre-mRNA



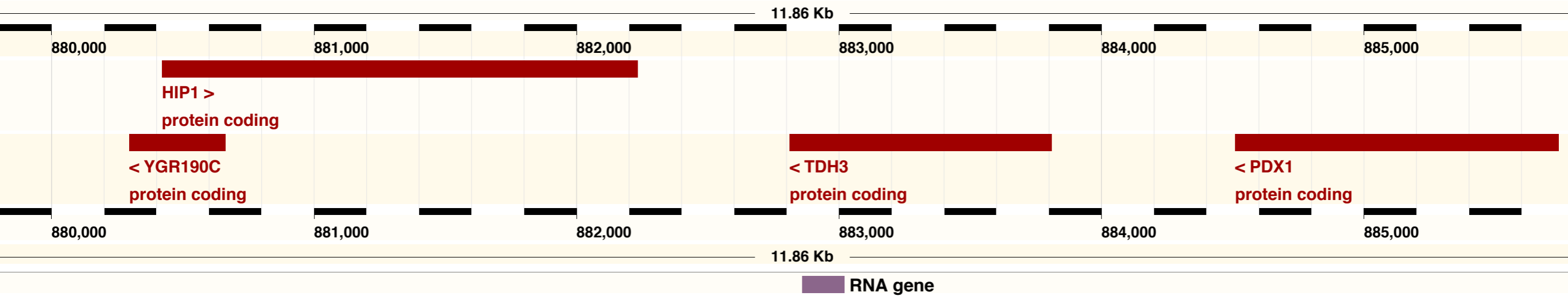
splicing

mature mRNA



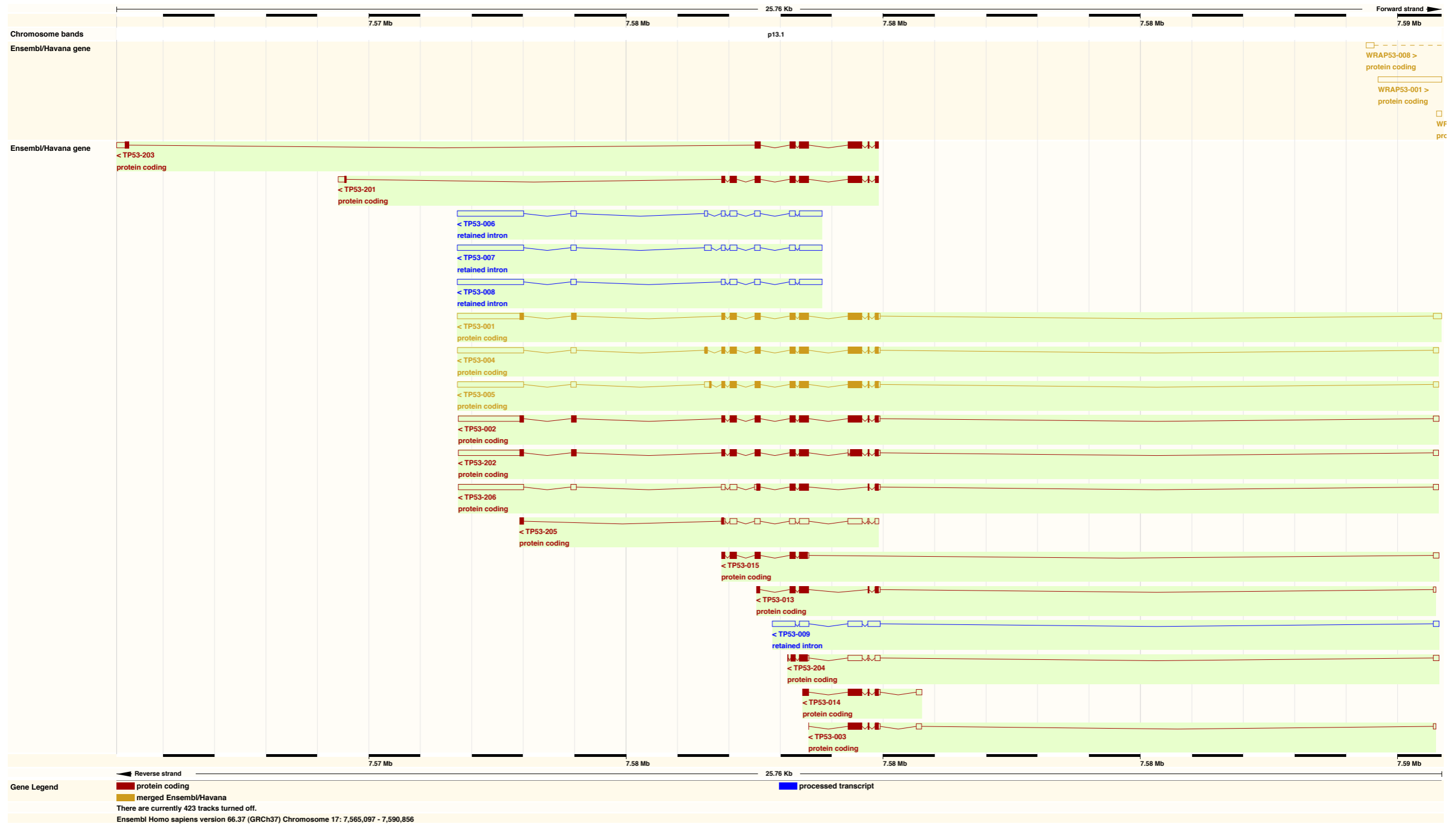
# Yeast

---



- Only one transcript per gene
- No (little) splicing
- Overlapping genes
- Little space between genes

# TP53 (human gene)



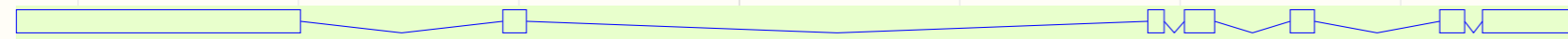
# TP53 (human gene)



< TP53-006  
retained intron



< TP53-007  
retained intron



< TP53-008  
retained intron



< TP53-001  
protein coding



< TP53-004  
protein coding



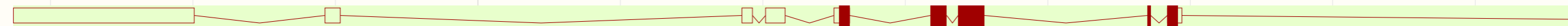
< TP53-005  
protein coding



< TP53-002  
protein coding



< TP53-202  
protein coding



< TP53-206  
protein coding



# Questions

---

What is the structure of known and unknown transcripts

Changes in splicing

Gene expression

Transcript expression

Allele specific expression

# Technical variation (broad overview)

---

Take tissue sample from individual

Extract RNA from tissue sample

Convert RNA to DNA

Sequence DNA

# Technical variation (broad overview)

---

Take tissue sample from individual

Extract RNA from tissue sample

Convert RNA to DNA

Sequence DNA

All of the above induces technical variation.

Several steps are independent of technology.

Also: day-to-day, laboratory, experimenter, machine

# Standard protocol

---

The current standard protocol for RNA-Seq is

Extraction of RNA, polyA purification

Fragmentation of RNA

Reverse transcription of RNA to cDNA (using random hex.)

Ligation of adapters

Size selection ~ 200bp (perhaps ~300bp)

PCR amplification (15 rounds or so)

Injection into flowcell

This produces reads from polyadenylated RNA without strand information.

# Standard protocol

---

The current standard protocol for RNA-Seq is

Extraction of RNA, polyA purification

Fragmentation of RNA

Reverse transcription of RNA to cDNA (using random hex.)

Ligation of adapters

Size selection ~ 200bp (perhaps ~300bp)

PCR amplification (15 rounds or so)

Injection into flowcell

This produces reads from polyadenylated RNA without strand information.

## Variants

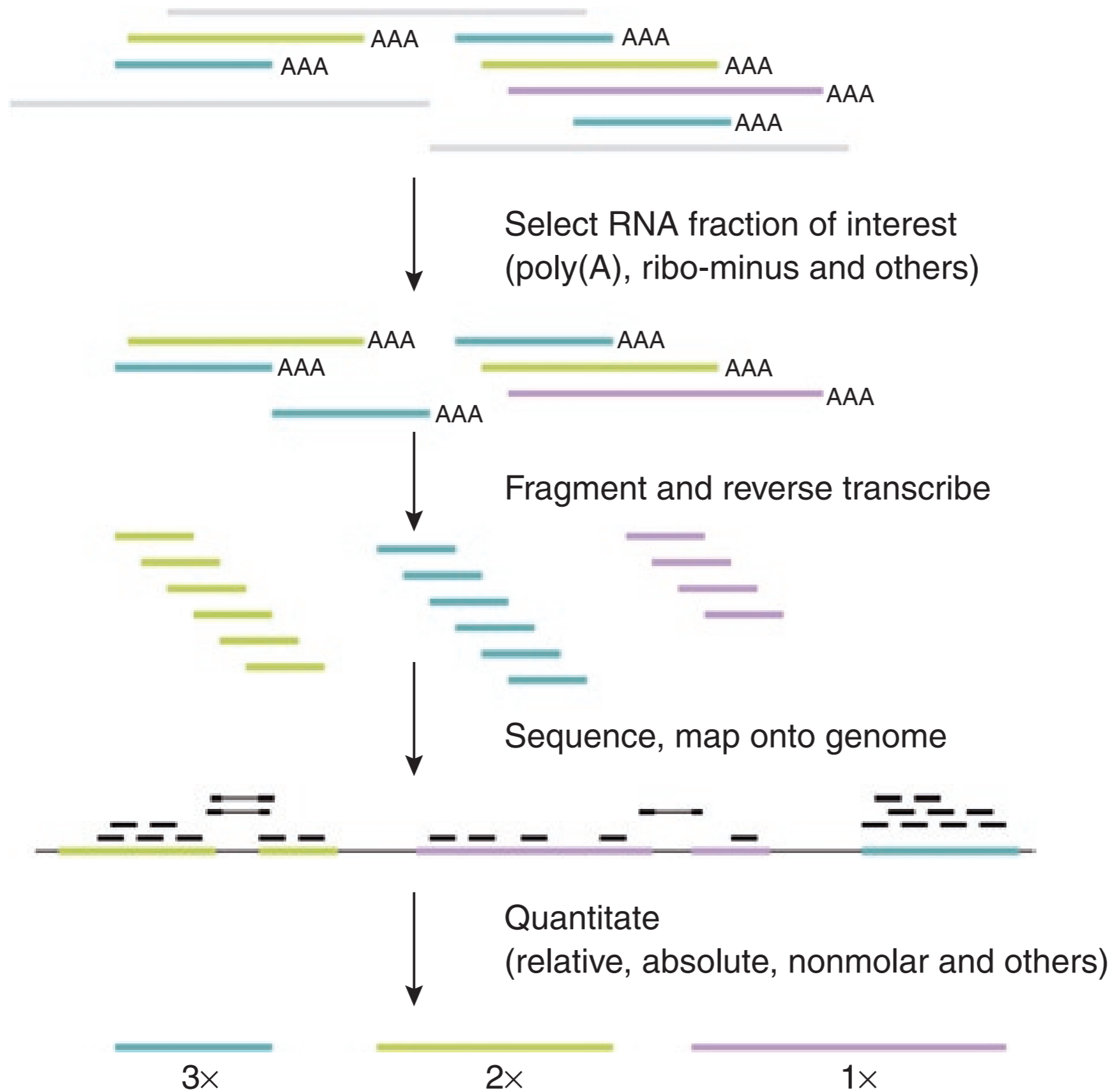
ribominus instead of polyA purification

strand specificity

small RNA sequencing (direct ligation of adaptors to RNA)

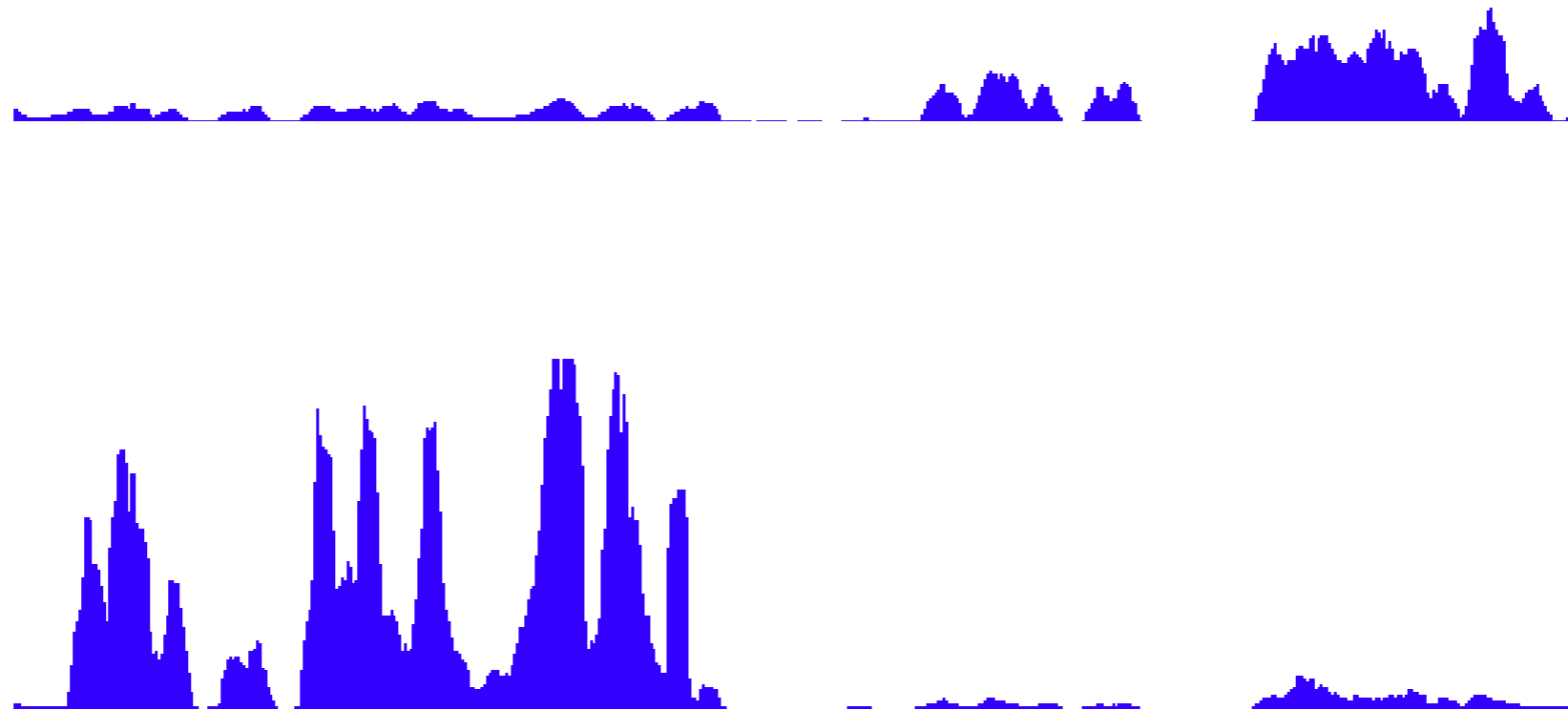
[ oligo(dT) priming instead of random hexamer priming ]

# Overview



# Data from *D. melanogaster*

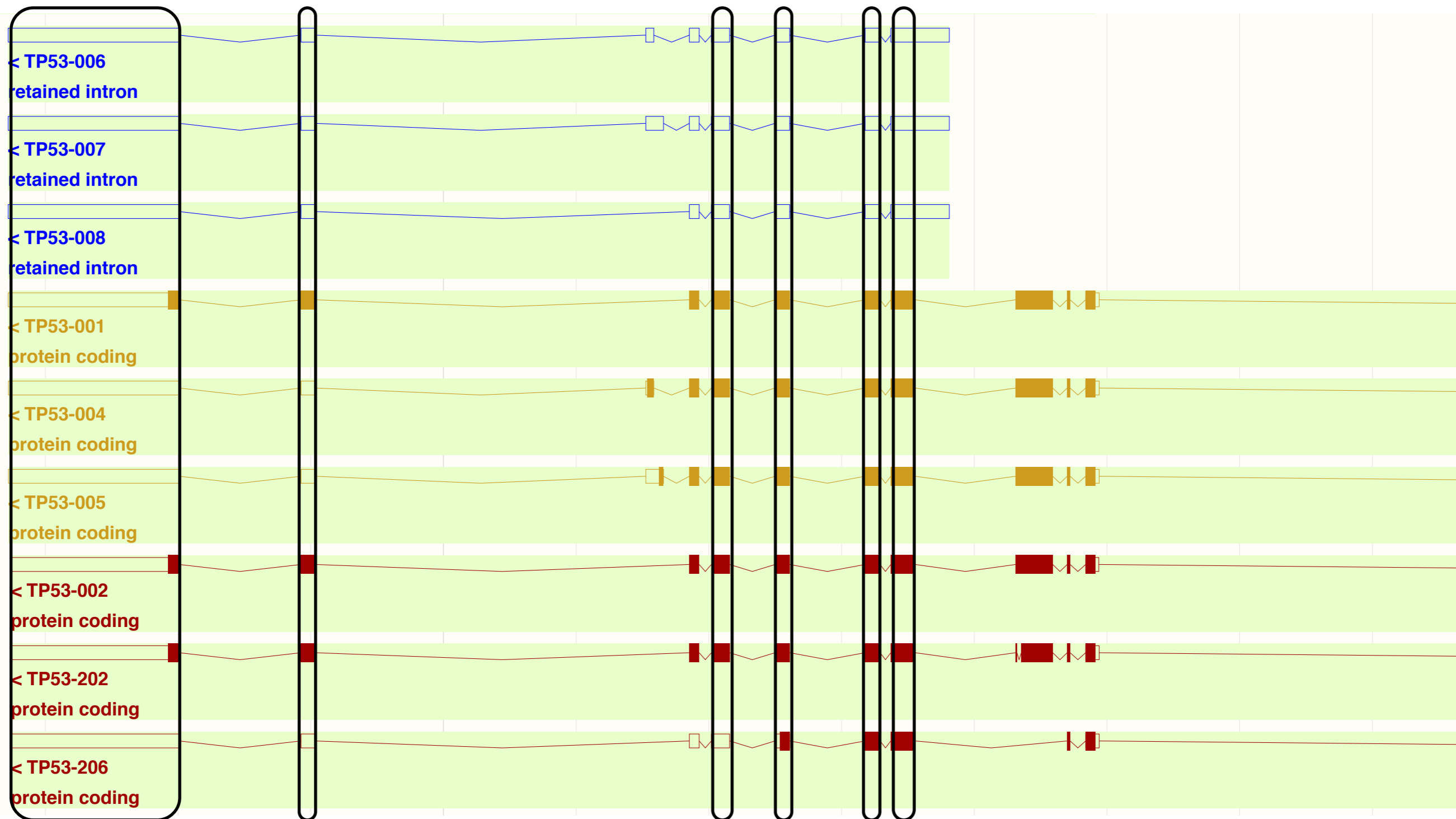
---



# Gene by Sample



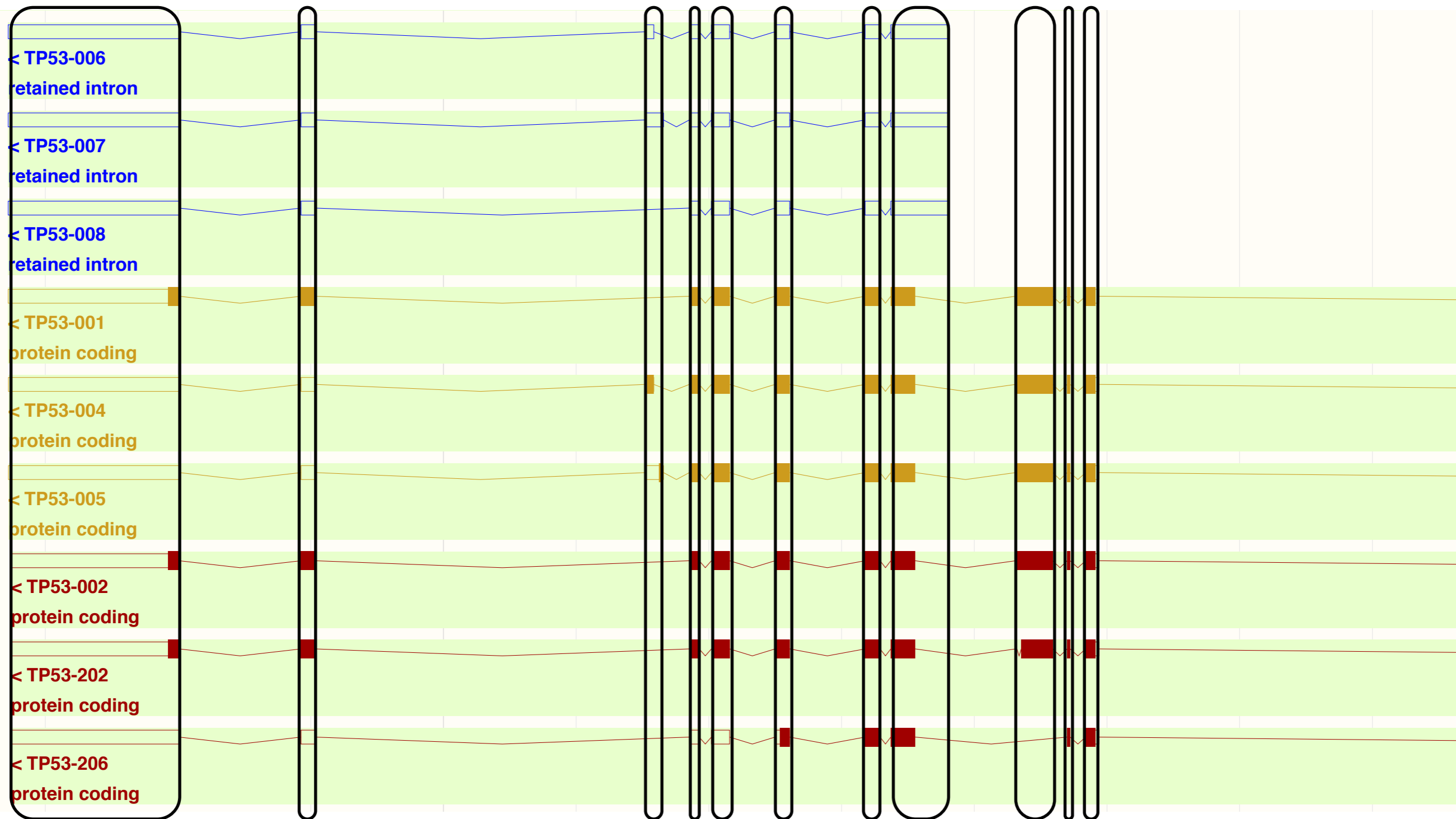
# Gene models



Union-intersection

("every base belonging to every transcript")

# Gene models



Union

("every base belonging to any transcript")

# Gene level data

---

Gene model + overlap rule = gene x sample matrix

(like microarrays)

Much work by statisticians re. inferring changes between conditions (differential expression).

Count data (many zeroes, very large range)

how do we model biological variability

# RPKM

---

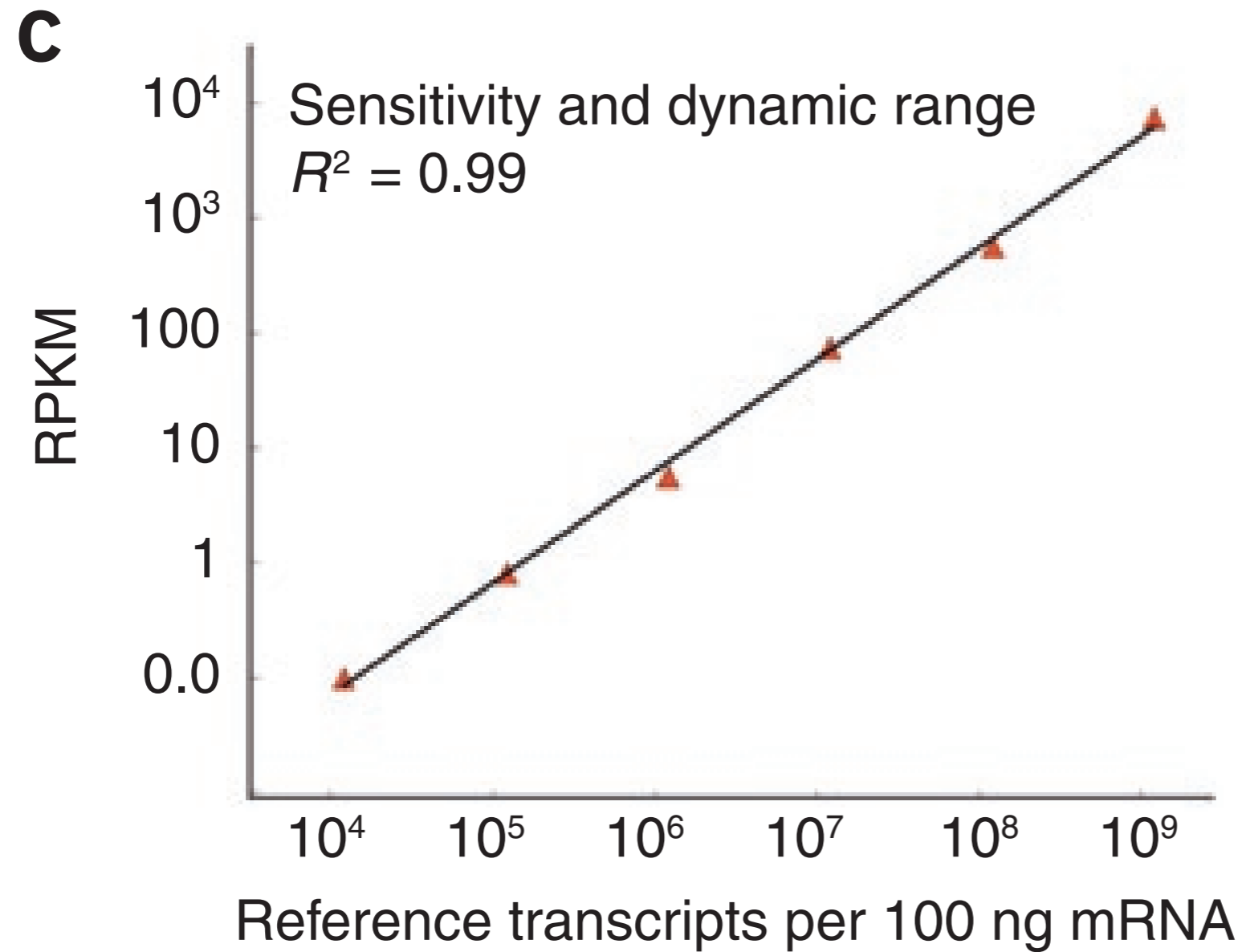
We need to control for  
sequencing depth  
gene length

Mortazavi (2008) Nat. Methods introduced “RPKM”.

$$\text{RPKM}(g, i) = \frac{X(g, i)}{L(g)N(i)}$$

# The big deal

---



# Poisson

---

Marioni (2008) Genome Res. showed that technical replicates are poisson.

$$X(g, i) \sim \text{Poisson}(\lambda_g N(i))$$

Bullard (2010) BMC Bioinformatics confirmed and extended to library preparation.

None of these papers looked at biological replicates or RNA extraction. Only the technical variation introduced by the sequencing machine.

# Negative Binomial

---

Several papers have considered more complicated count models, especially the negative binomial.

We have tricks for borrowing strength across genes.

$$X(g, i) \sim F\left(\theta(g), N(i)\right)$$

Key papers are

Anders (2010) Genome Biology ["DESeq"]

Hardcastle (2010) BMC Bioinformatics ["baySeq"]

McCarthy (2012) Nucleic Acids Res ["edgeR"]

Implementations in Bioconductor. Things change fast.

# The size factor

---

We need values of  $N(i)$  (“sequencing depth”) or (“size factor”)

Naive estimates:

- Number of reads

- Number of mapped reads

Several (scale) normalization methods exist.

- Bullard (2010) BMC Bioinformatics (“upper quartile”)

- Robinson (2010) Genome Biology (“TMM”)

- Anders (2010) Genome Biology

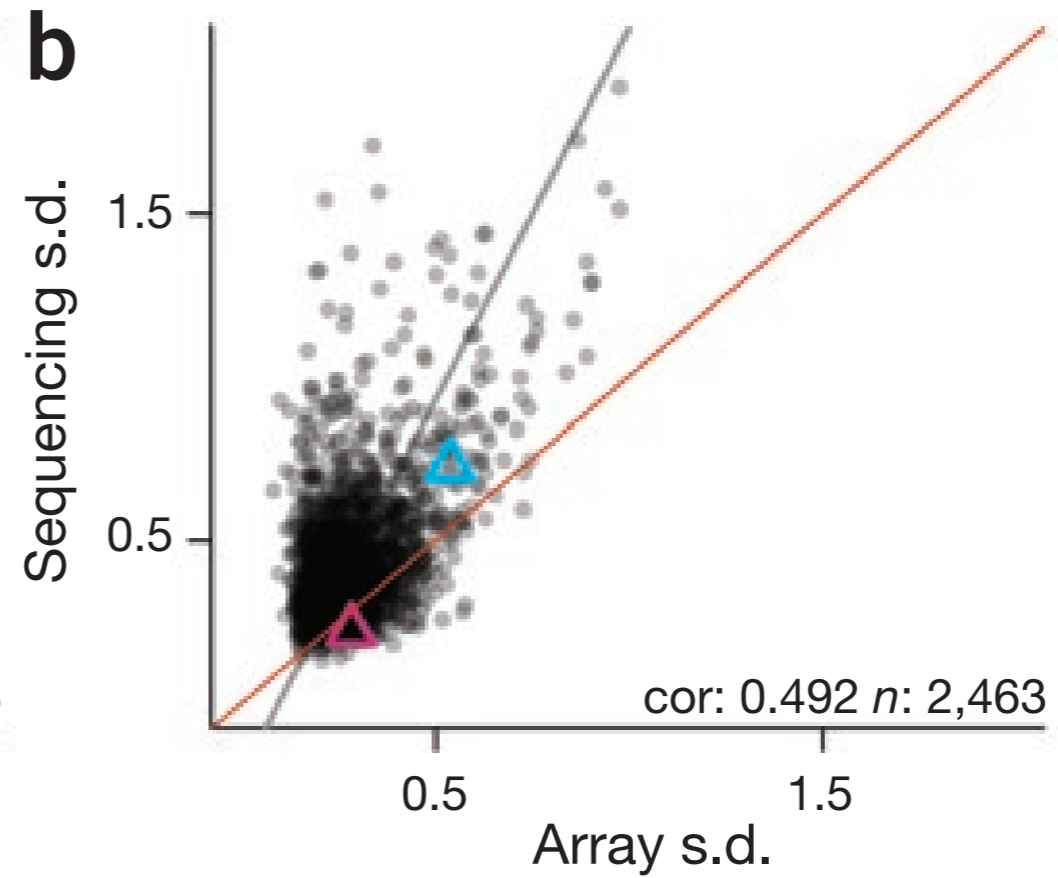
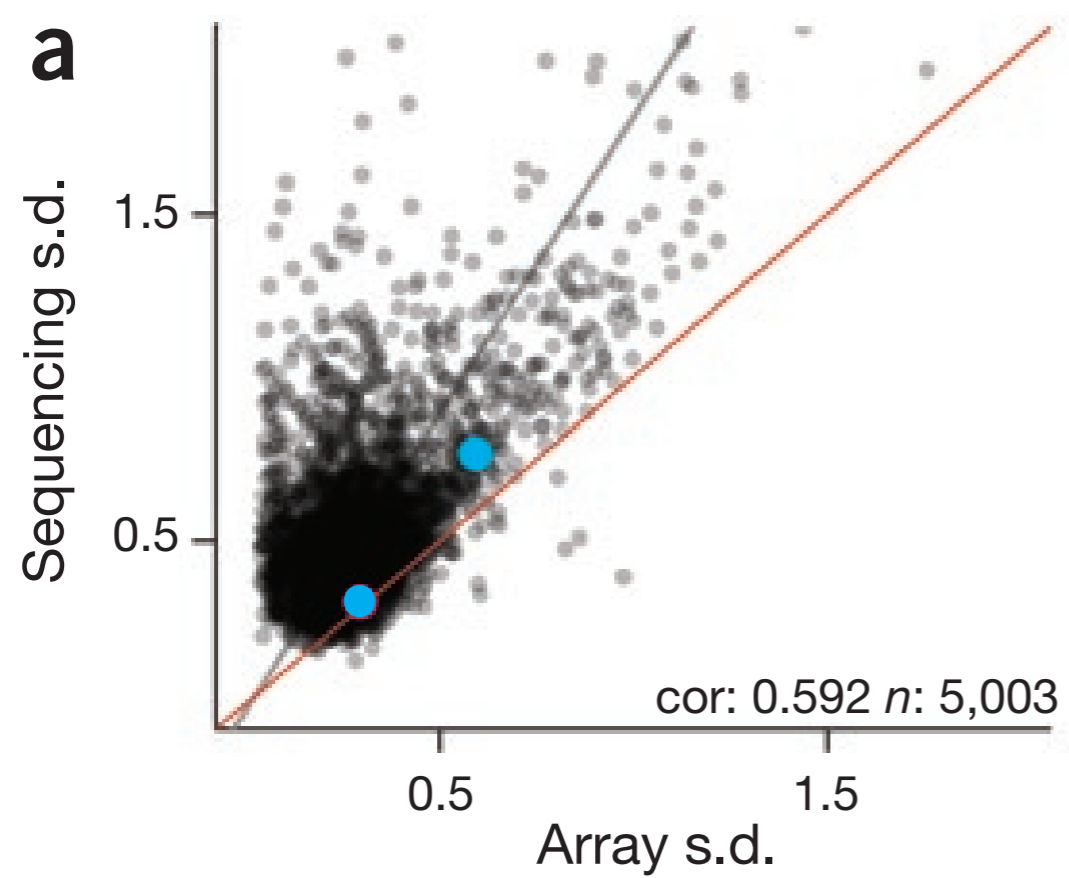
This is especially an issue when comparing very different samples. For example, between tissue types.

Langmead (2010) Genome Biology shows that it may be a good idea to use a gene-specific normalization factor.



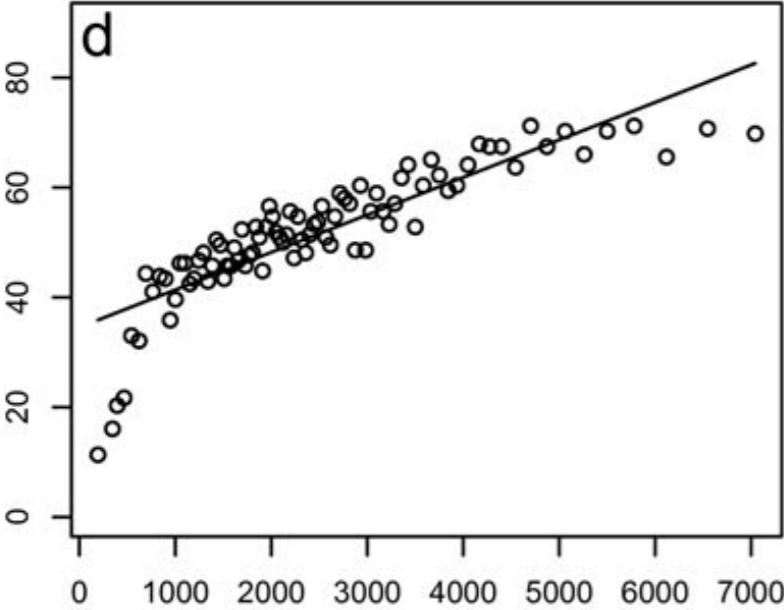
# Biological variability

---

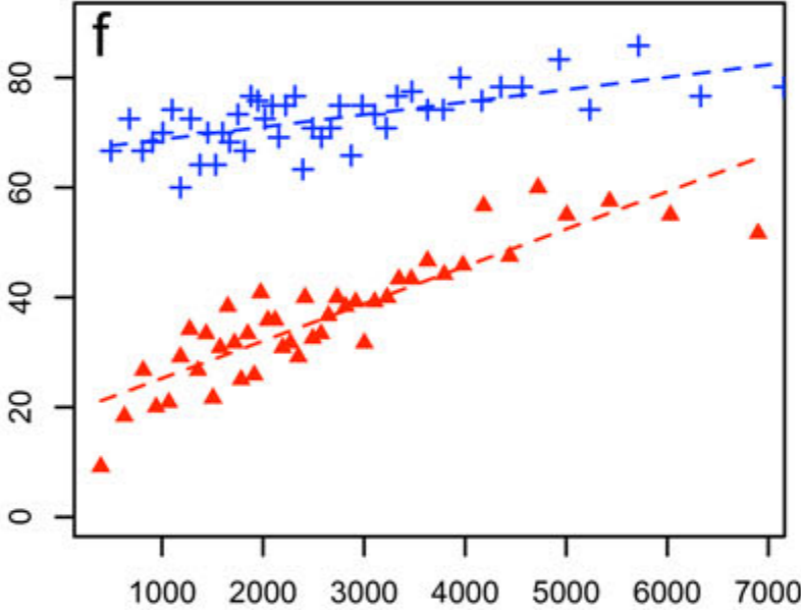


# Problems: length bias

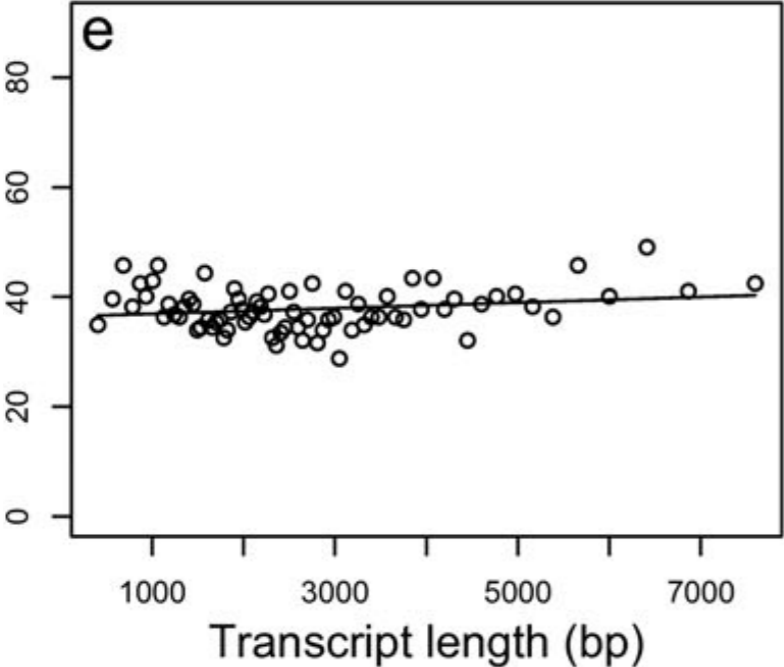
Sequencing Data (Marioni)



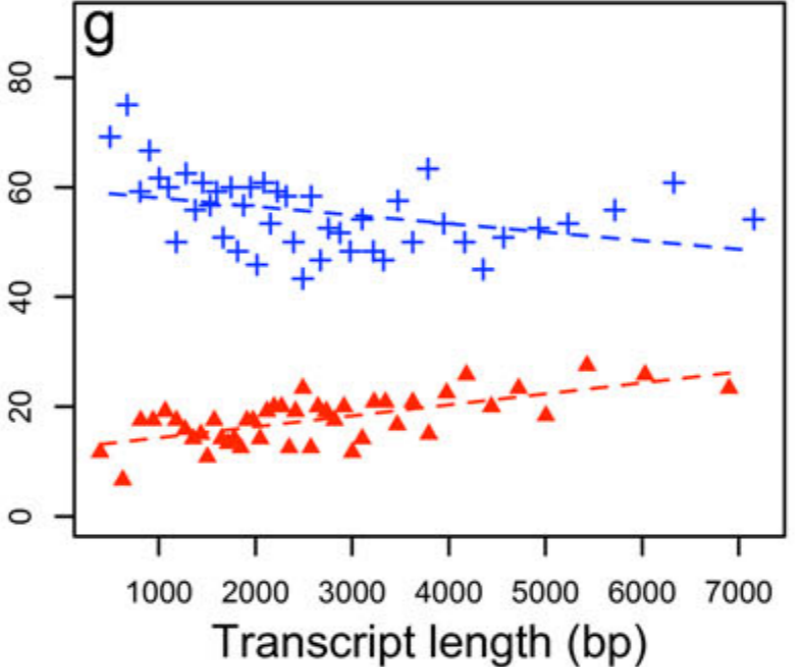
Sequencing Data (Marioni)



Array Data (Marioni)

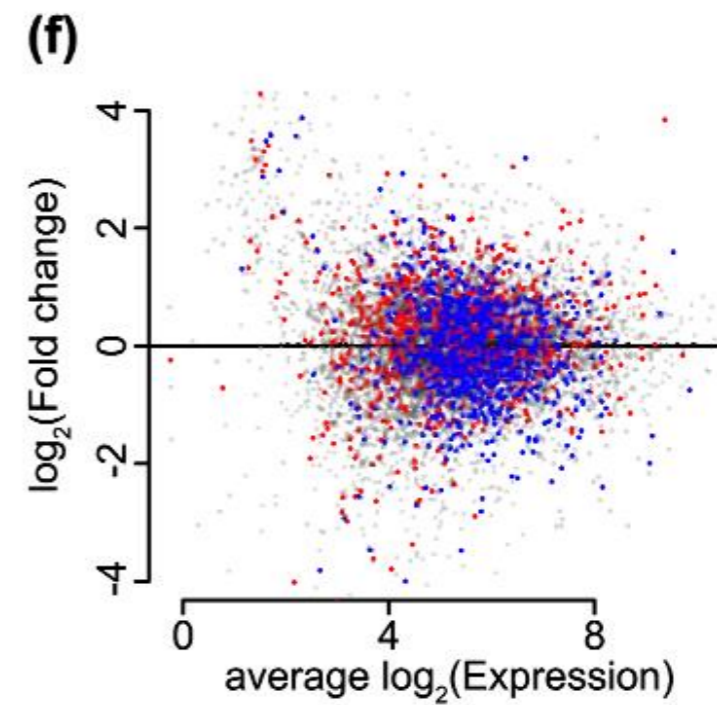
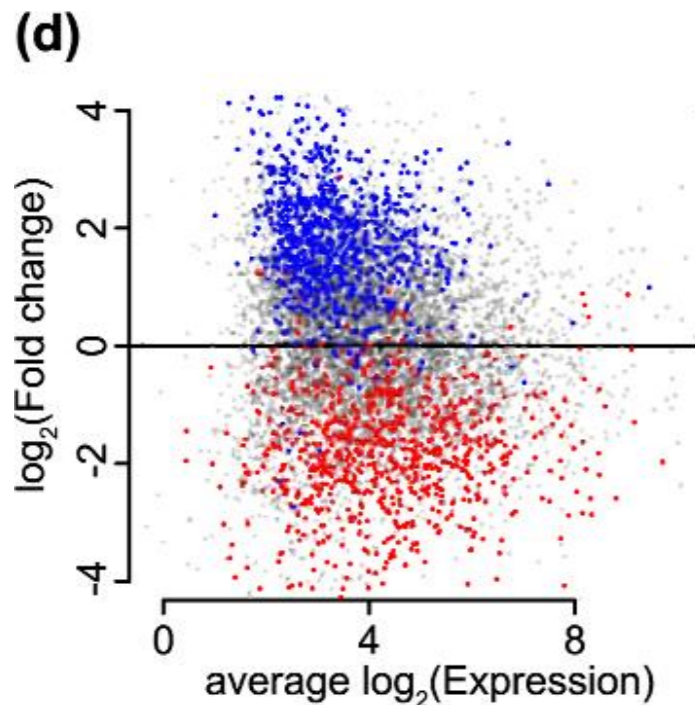
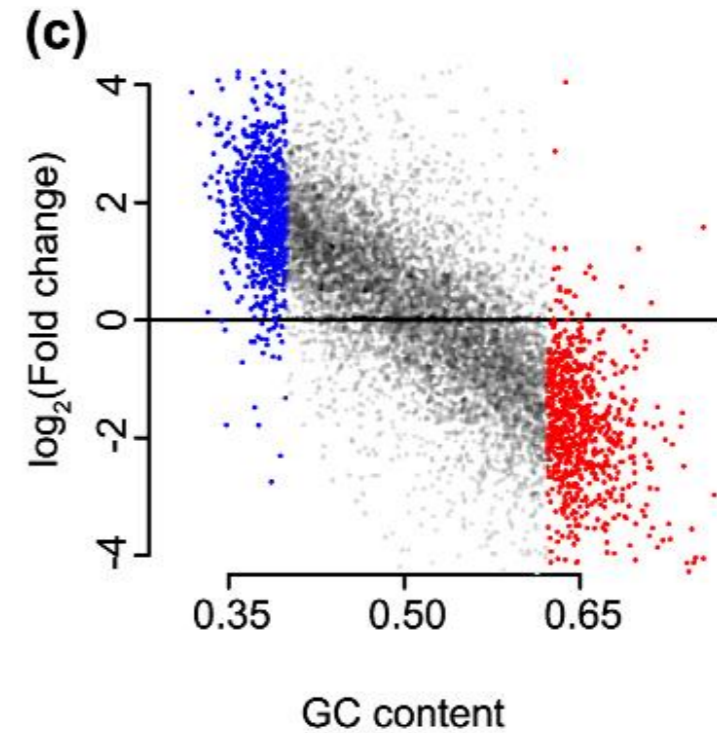
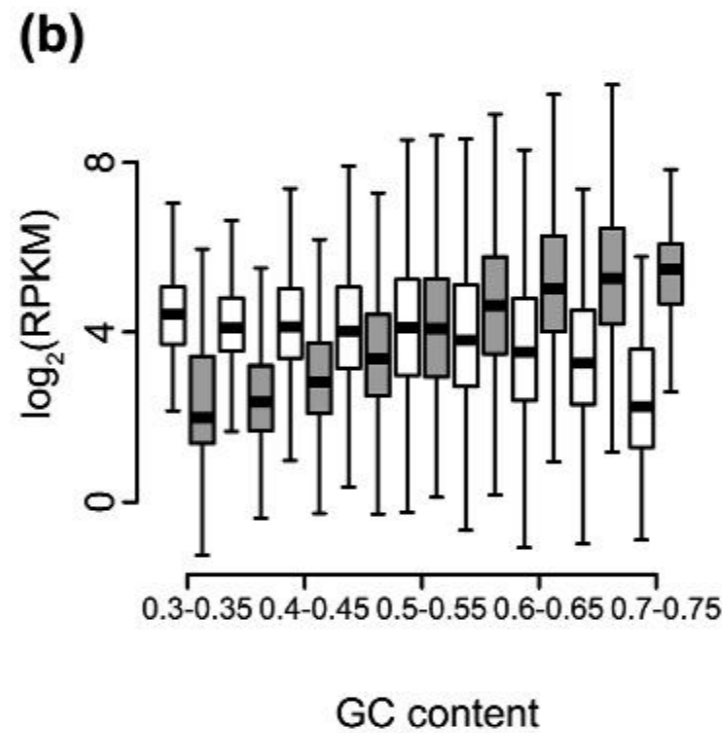


Array Data (Marioni)



# Problems: GC content bias

## Sample-specific GC content effect



Hansen (2012) Biostatistics ("CQN")

Also Risso (2011) BMC Bioinformatics ("EDAseq")

# Lessons

---

Biological variability

Need for normalization

Issues with length, GC content, ?

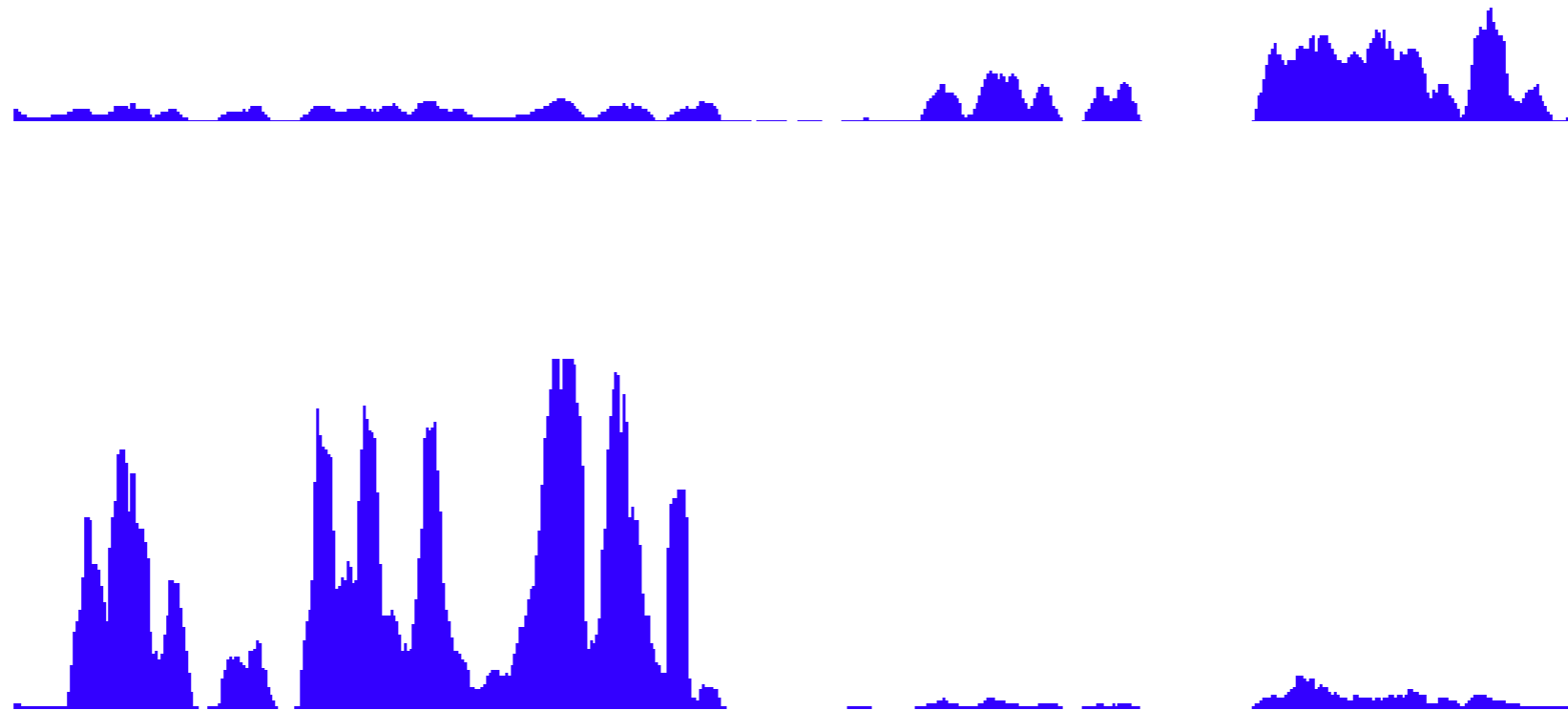
Models for count data, borrowing strength across genes

... but all of this addresses a question we could have answered using microarrays

**A look at the data**

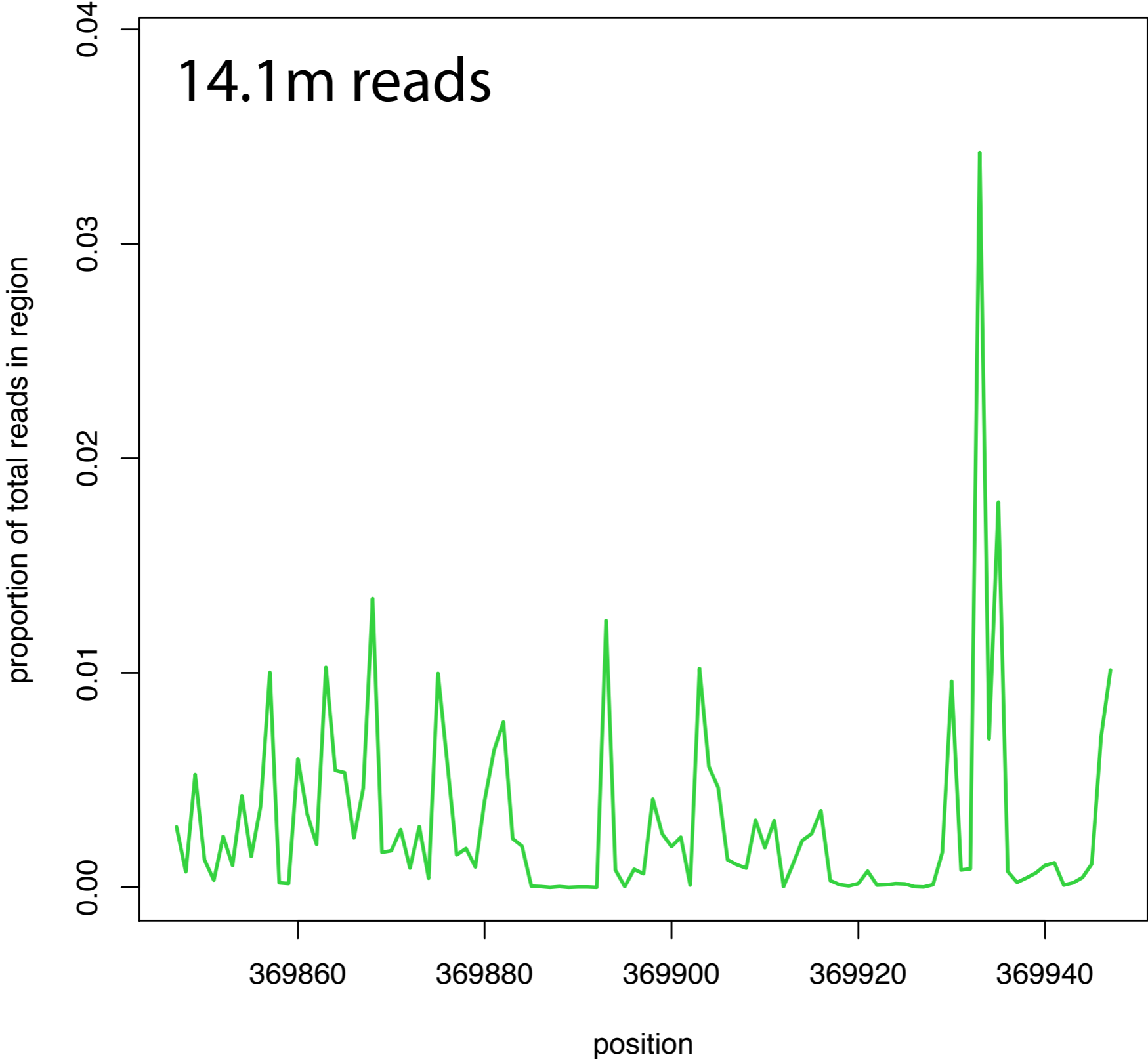
# Data from *D. melanogaster*

---

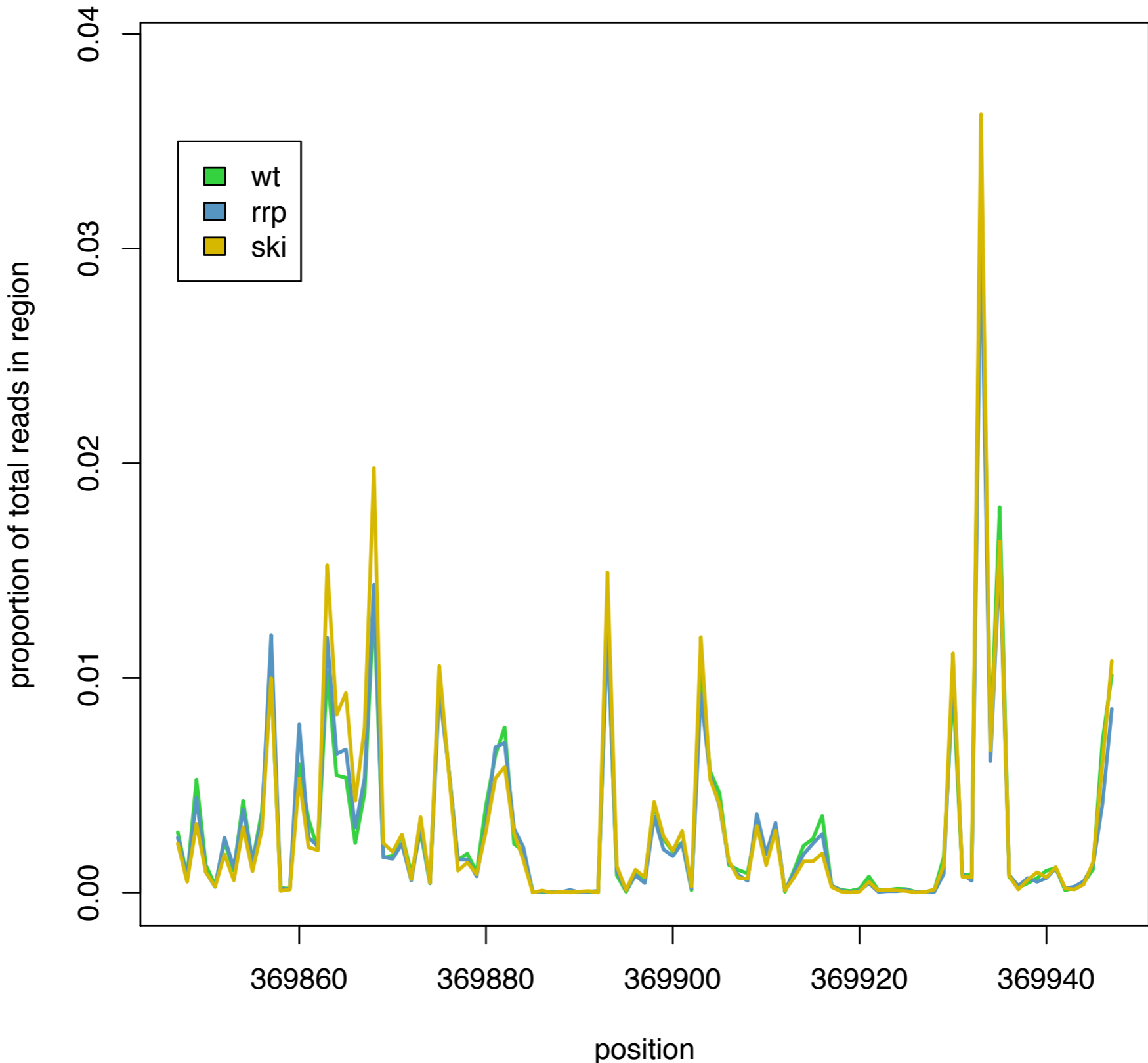


# Base effect - single sample

---

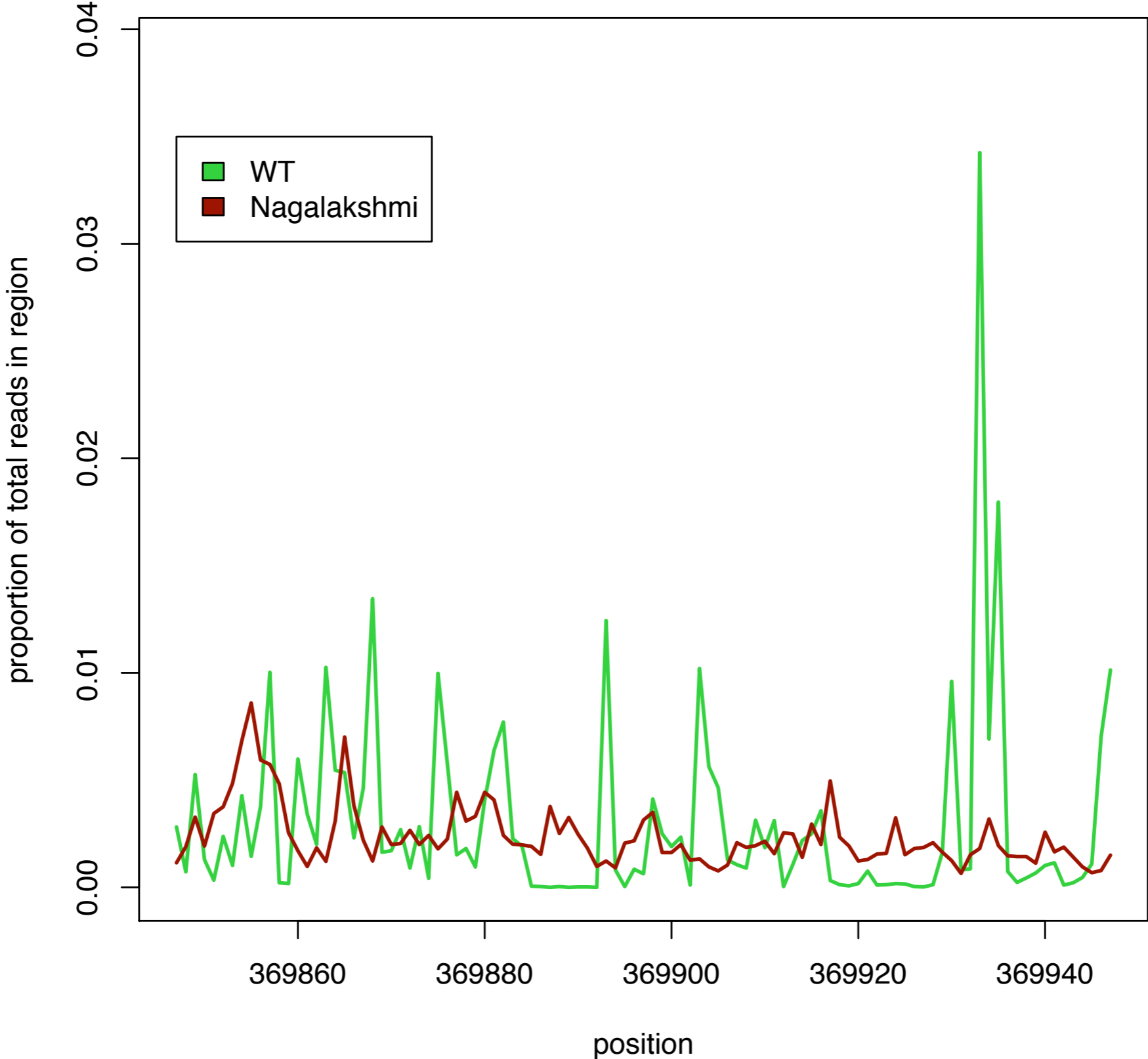


# Base effect - multiple samples





# Base effect - different study (and prep)



# Base effect - conclusions

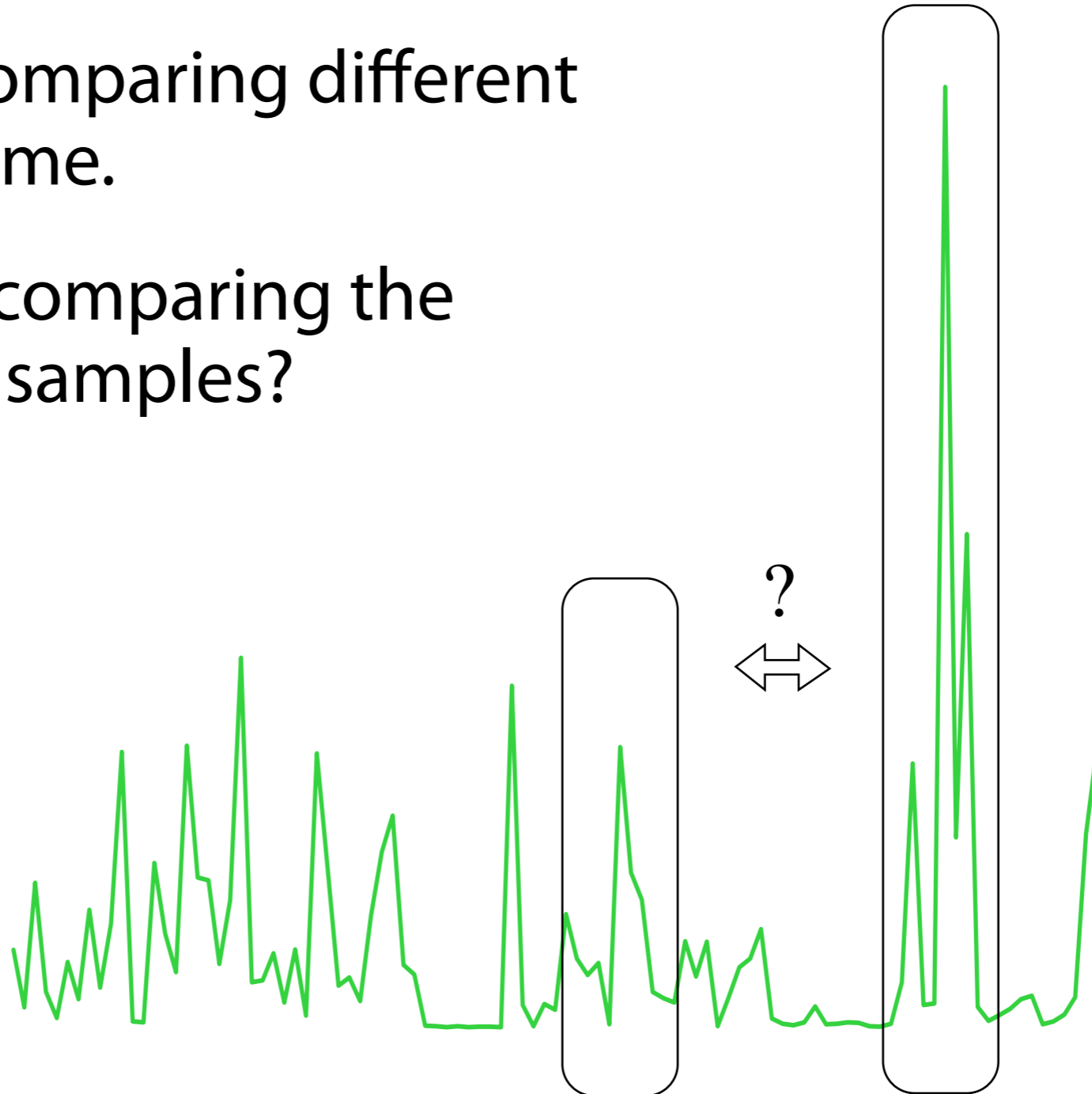
---

Reproducible base effect - like probe affinities in microarrays.

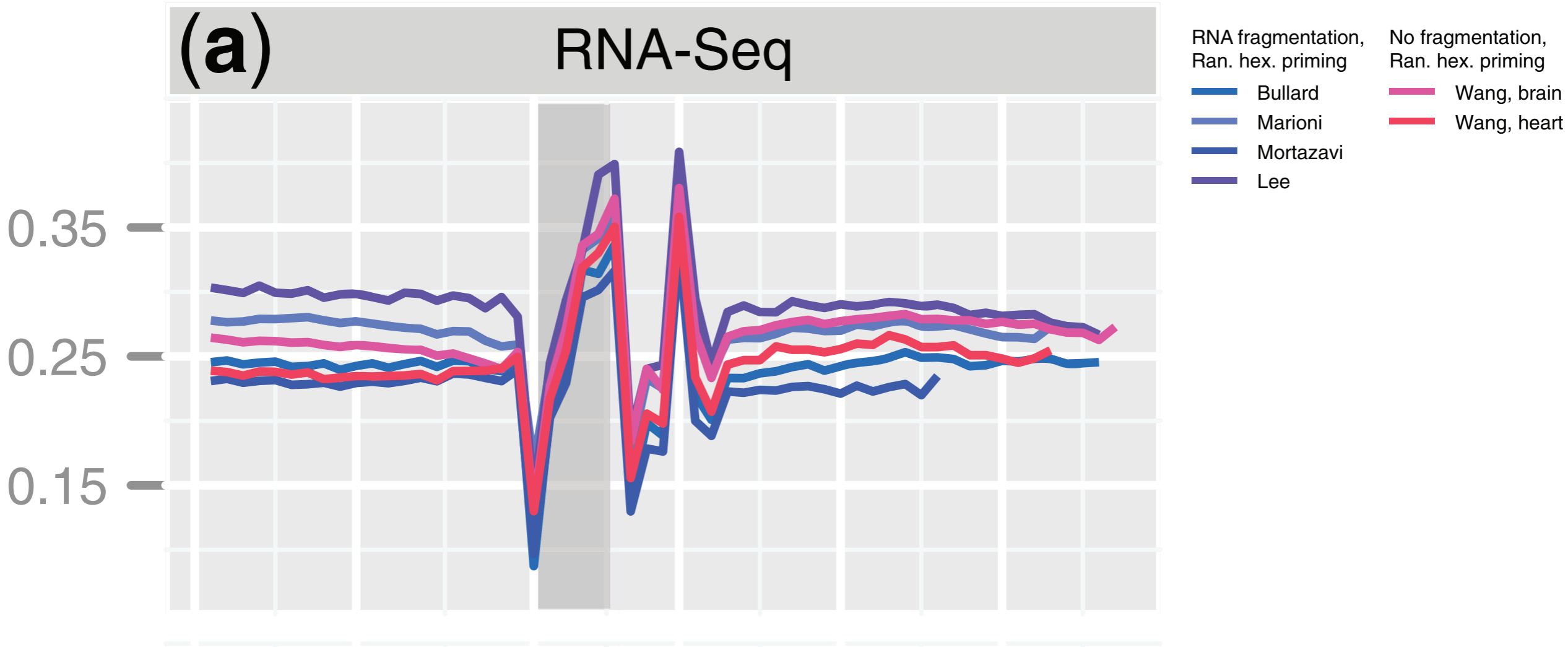
Seems to be prep dependent.

Creates issues for comparing different regions in the genome.

Less of an issue for comparing the same region across samples?



# Nucleotide content bias



# Correcting for spatial heterogeneity

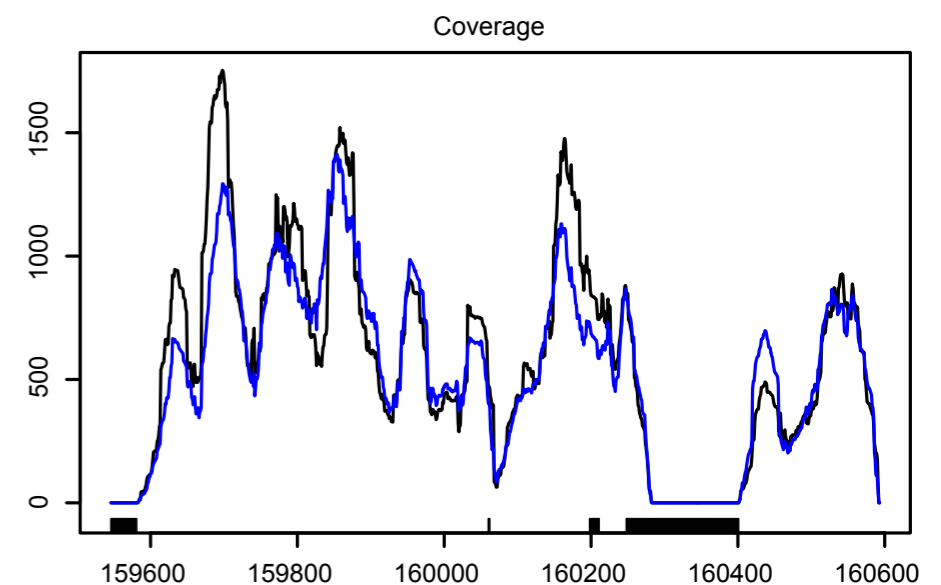
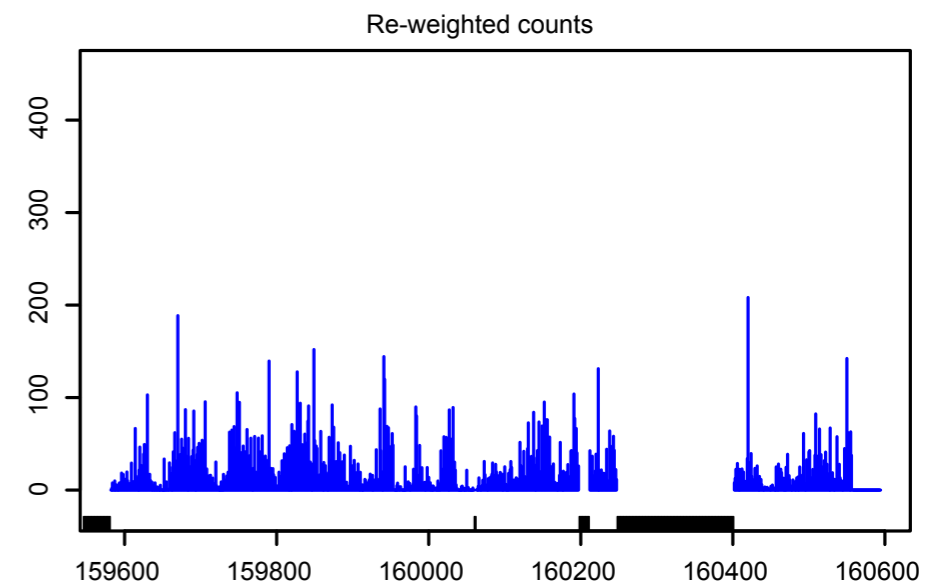
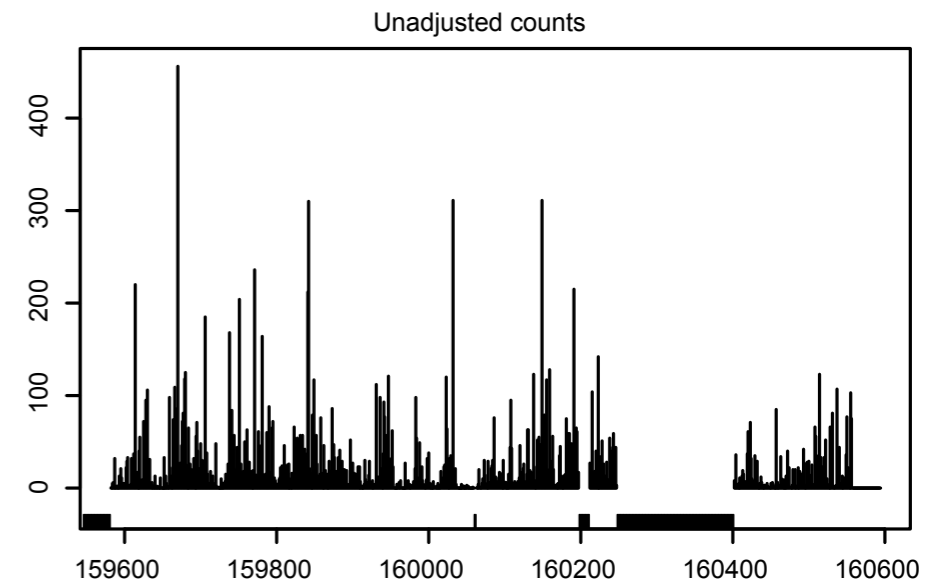
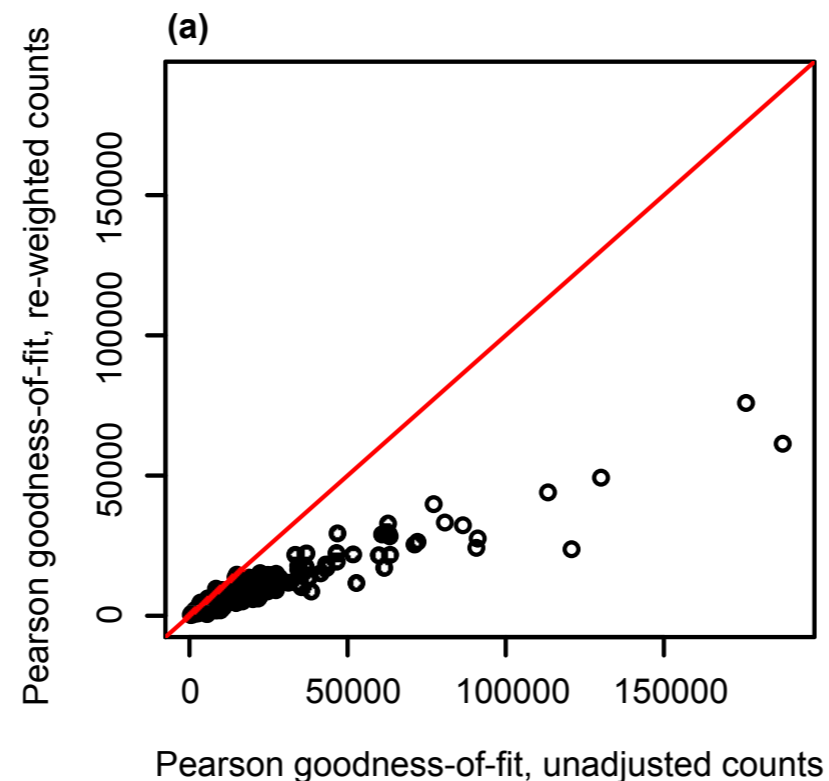
## A sample of papers

Hansen (2010) Nucleic Acids Res

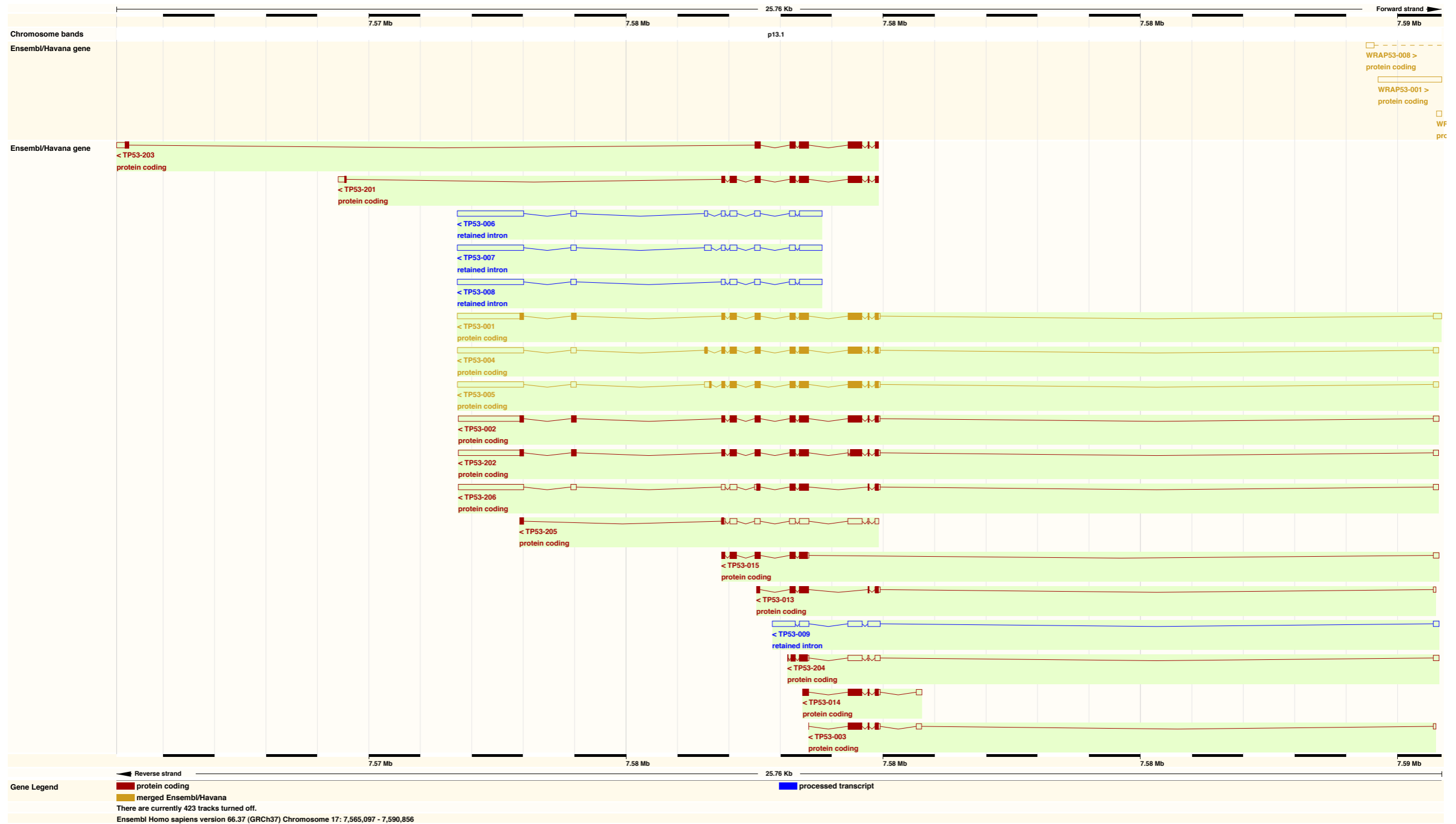
Li (2010) Genome Biology

Roberts (2011) Genome Biology

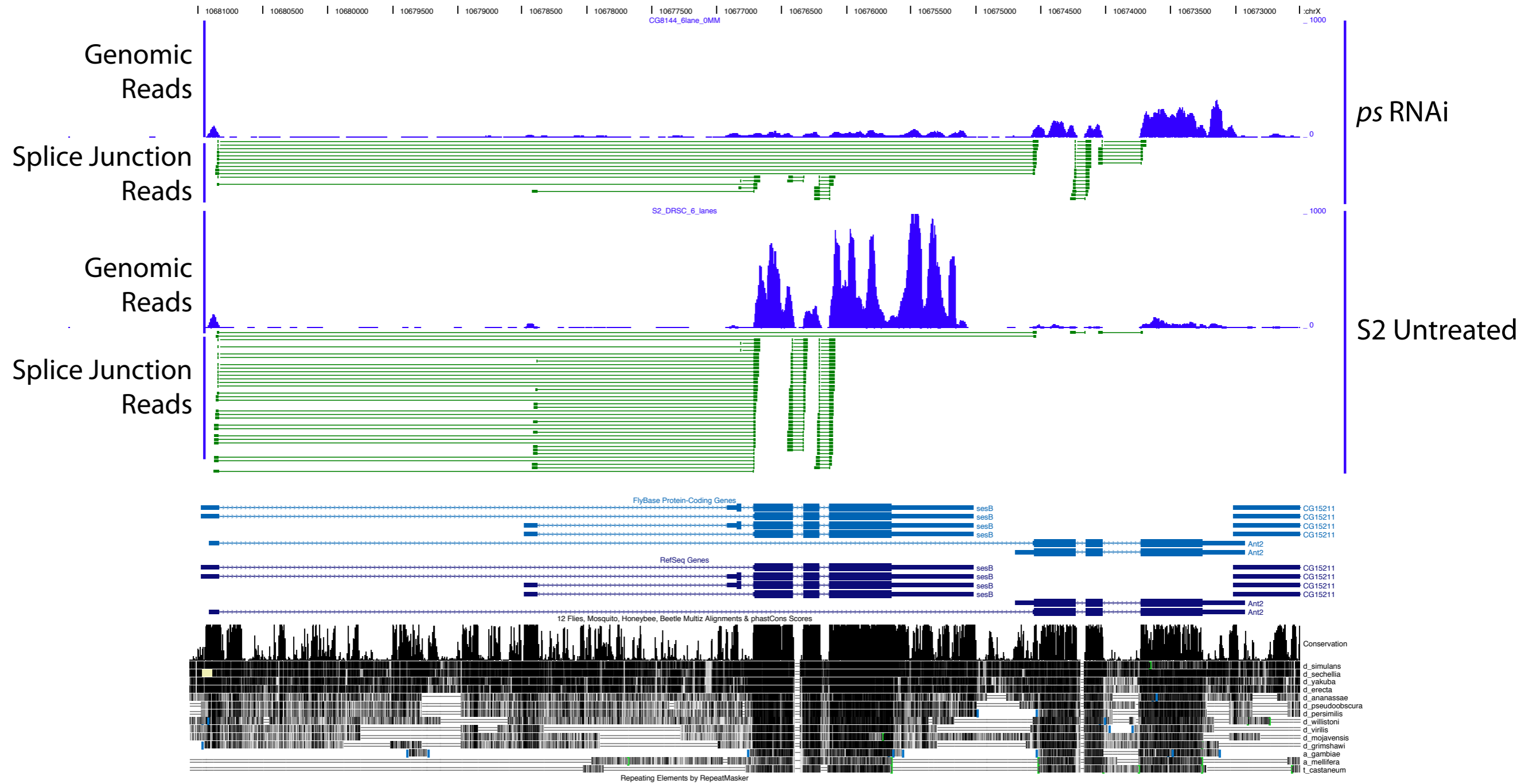
Jones (2012) Bioinformatics



# TP53 (human gene)

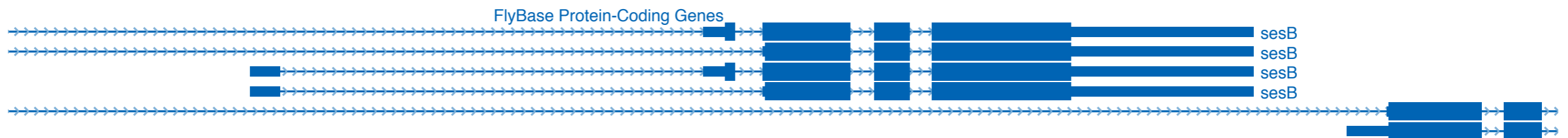
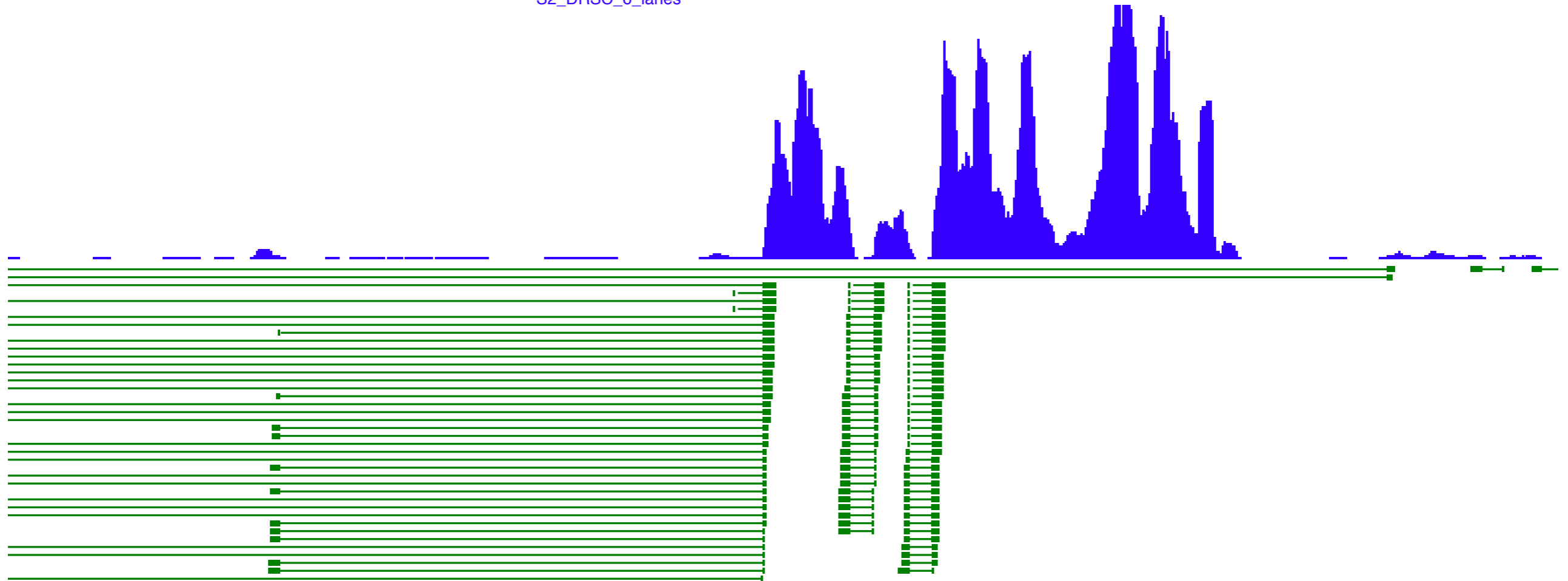


# Junction reads



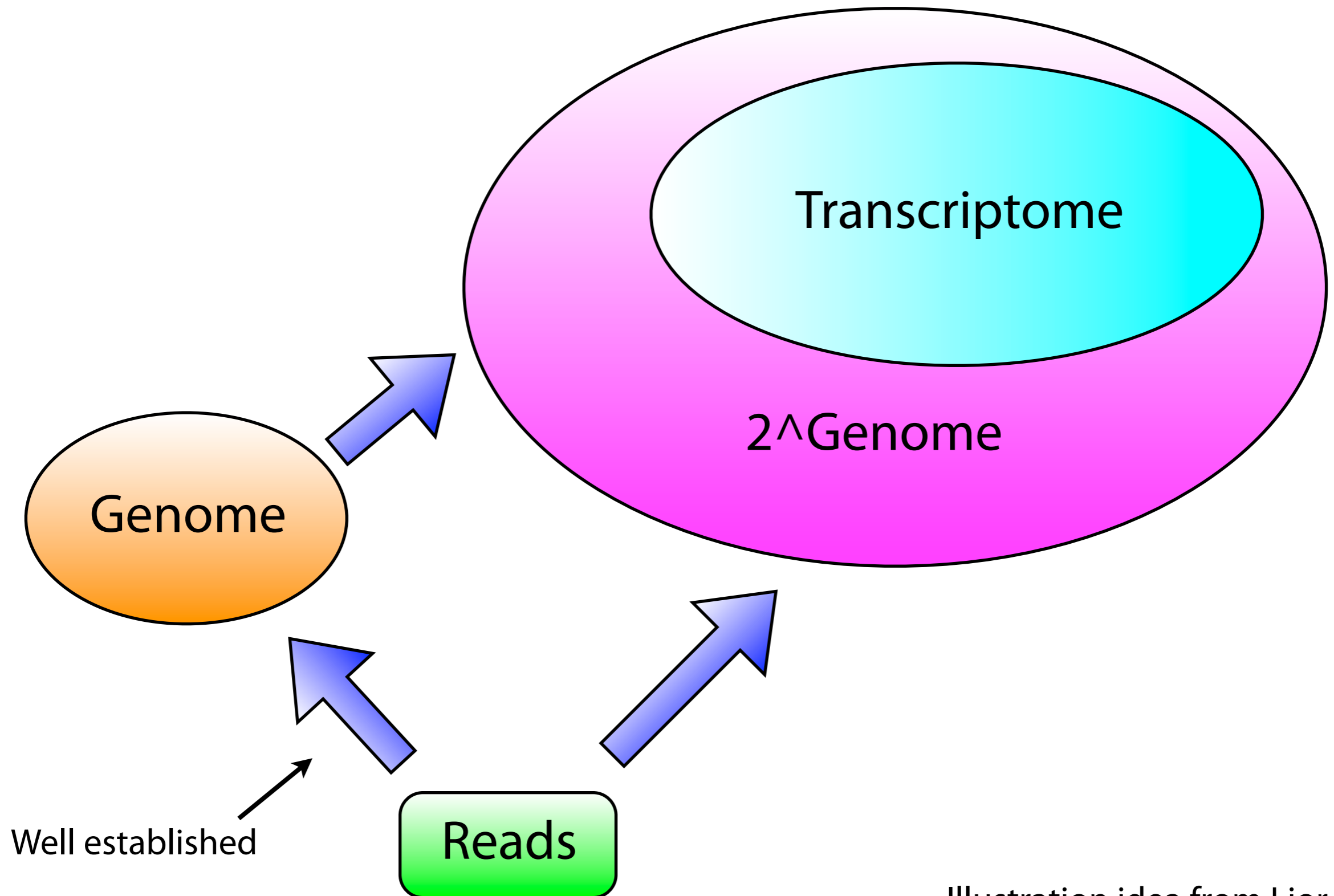
# Junction reads, zoom

S2\_DRSC\_6\_lanes



# Mapping reads to the transcriptome

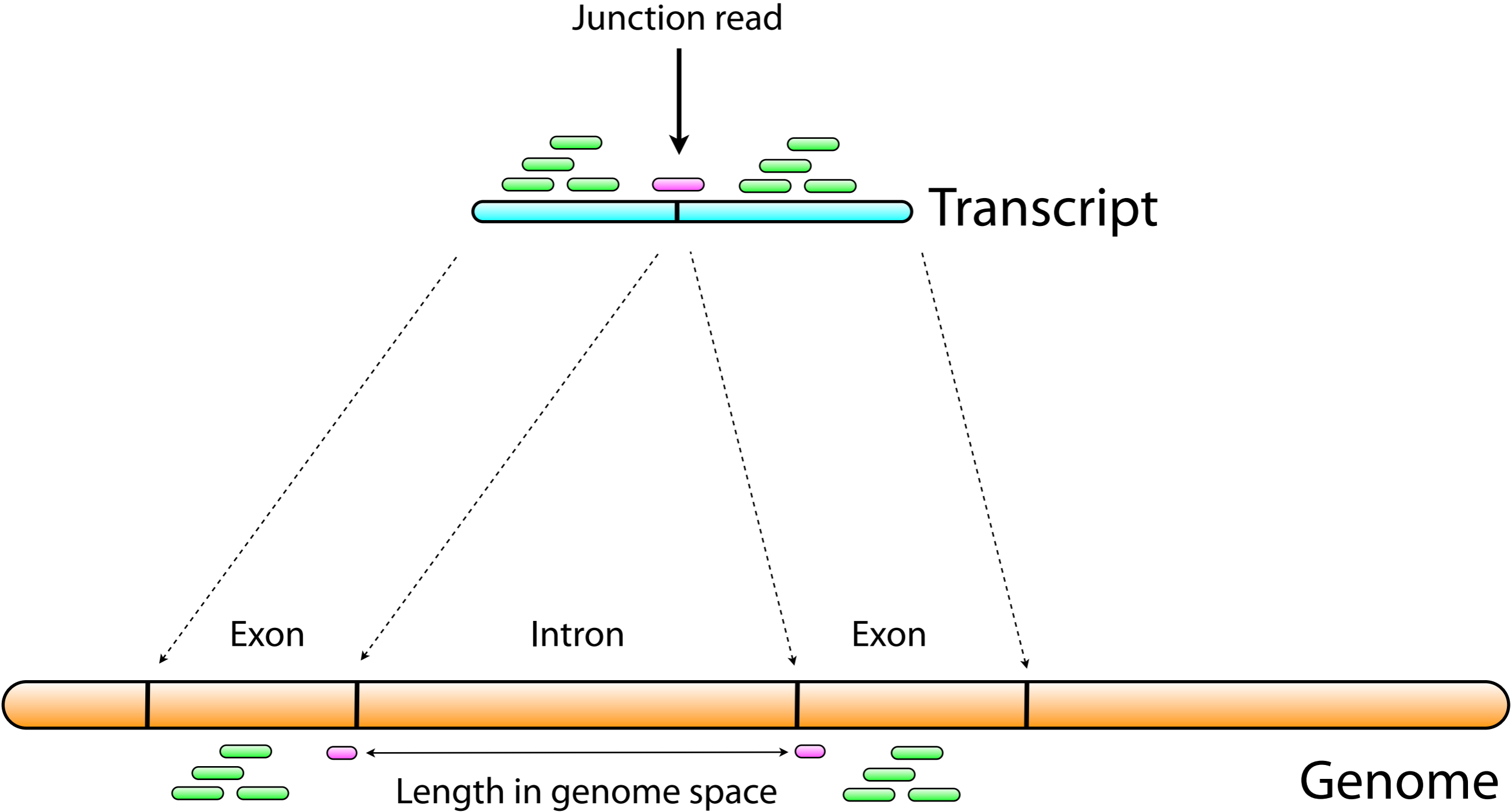
---





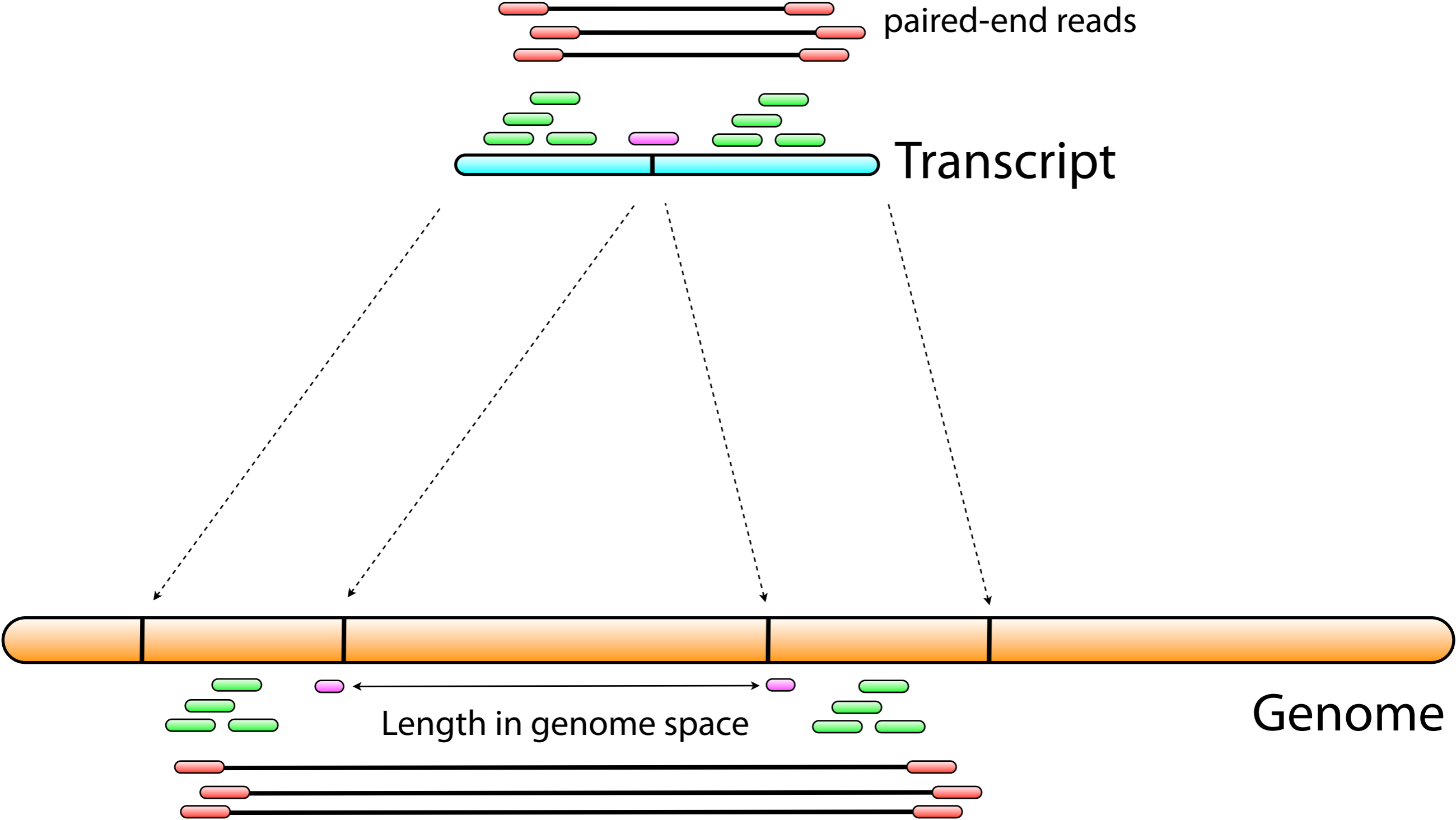
# Mapping transcripts

---



# Mapping transcripts

---



# The basic approaches

---

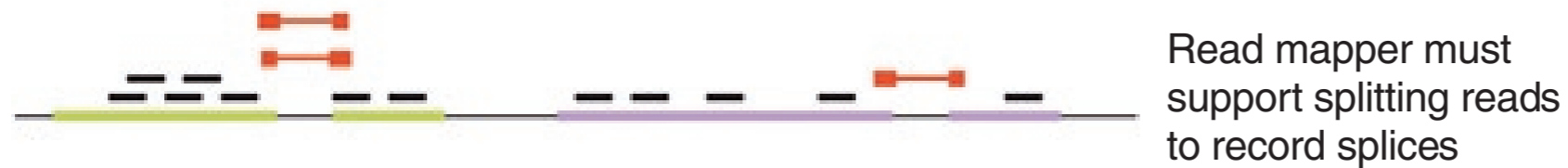
**a**

*De novo* assembly of the transcriptome



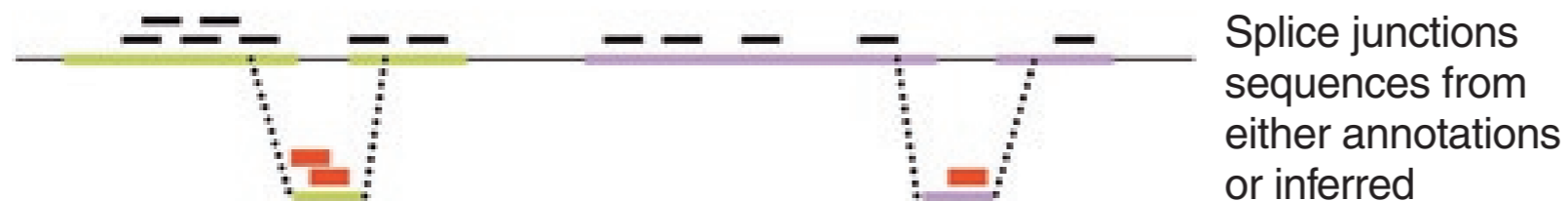
**b**

Map onto the genome



**c**

Map onto the genome and splice junctions



# Strategies for mapping to junctions

---

Map to known junctions (or to known transcripts, but that involves a lot of bookkeeping).

Map to combination of known exons.

Map completely de-novo using canonical acceptor and donor sites. (huge!)

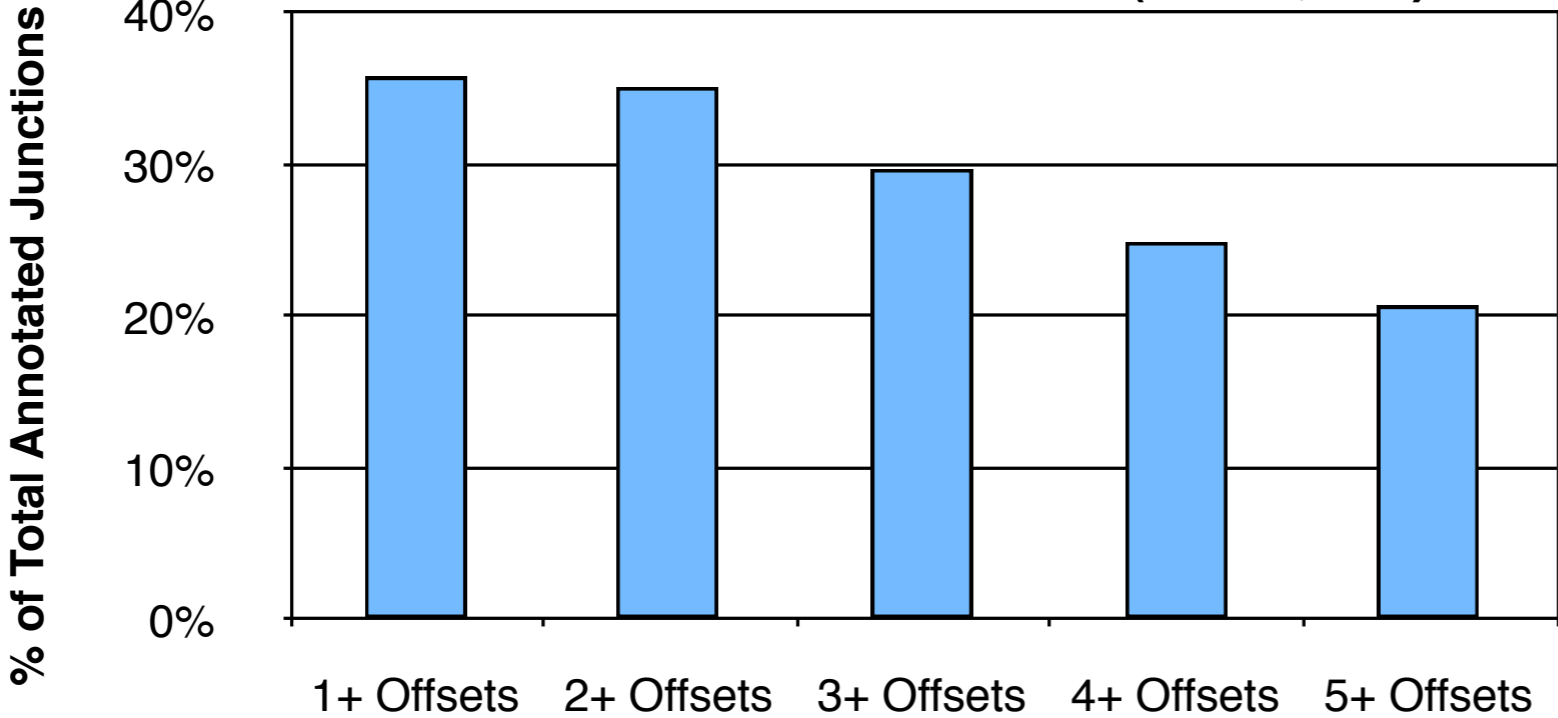
Map de-novo, but constrain the search to canonical acceptor and donor sites between and in transcribed region: transcript assembly. (TopHat).

Paired-end data will help with this.

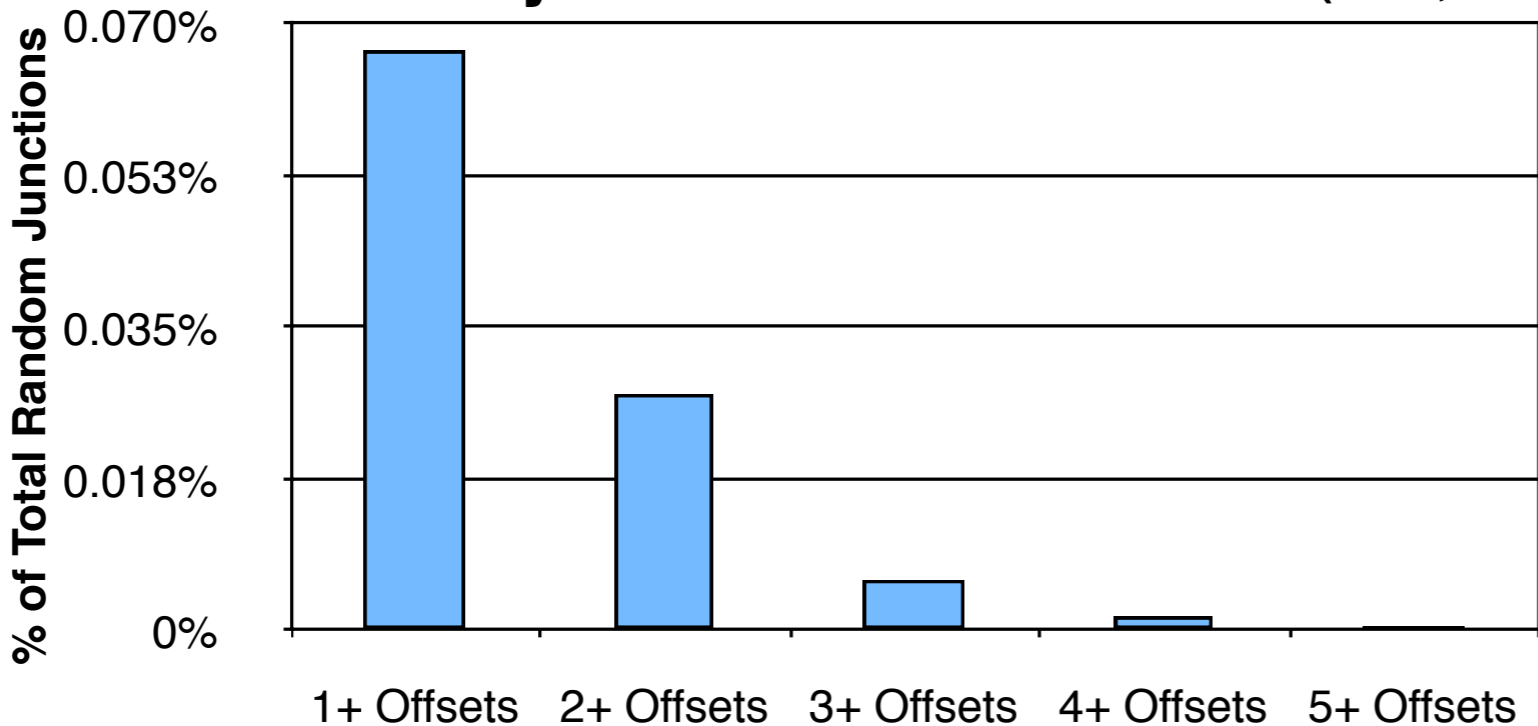
# FP rates for junctions

---

### Annotated Junctions (n= 58,212)



### Randomly Generated Junctions (n=5,409,600)



# Issues

---

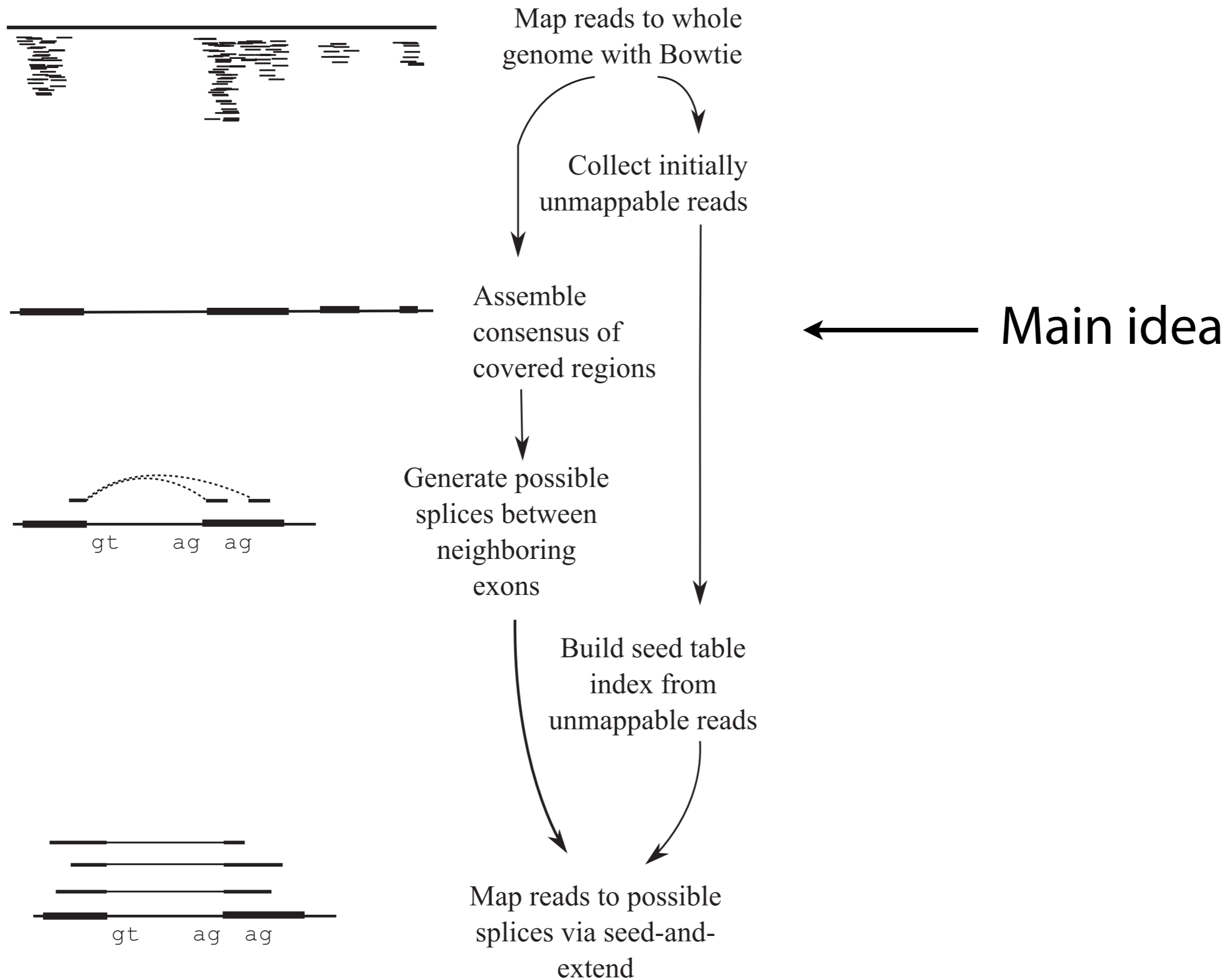
Hard to map near splice sites (both de-novo and known)

Similar regions of the genome +  
error in reads +  
differences between sample and reference  
= possibility of mapping errors. Still no real understanding.

Do not underestimate this aspect of the data.

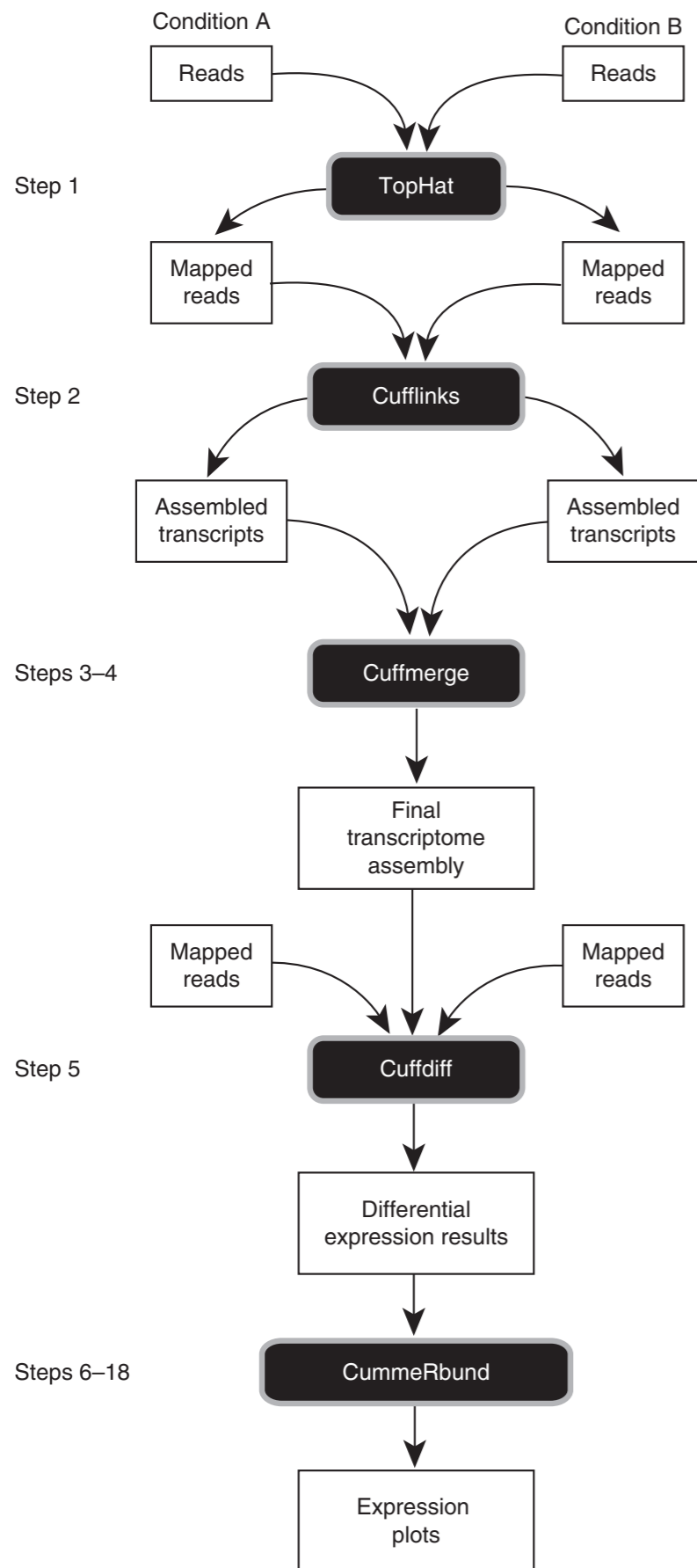
**Assembly**


# TopHat







# Tuxedo tools




  
**Bowtie**  
Extremely fast, general purpose short read aligner

  
**TopHat**  
Aligns RNA-Seq reads to the genome using Bowtie  
Discovers splice sites

  
**Cufflinks package**

- Cufflinks  
Assembles transcripts
- Cuffcompare  
Compares transcript assemblies to annotation
- Cuffmerge  
Merges two or more transcript assemblies
- Cuffdiff  
Finds differentially expressed genes and transcripts  
Detects differential splicing and promoter use

  
**CummeRbund**  
Plots abundance and differential expression results from Cuffdiff