RNA-Seq: Sequencing the Transcriptome

Kasper Daniel Hansen
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
FHCRC, Seattle 12th-14th of November 2008

RNA-Seq: Comparison with Microarrays

Potential for surveying the entire transcriptome, including novel, un-annotated regions.

Potential for determining gene structure and isoform level expression using reads mapping to splice junctions.

Potential for making better presence/absence calls on regions.

Potential for allelle specific expression combined with SNP calling.

Con: the assay is dependent on sequencing effort, low expressed regions will be missed.

Protocol

The current standard protocol for RNA-Seq is

Extraction of RNA, polyA purification
Fragmentation of RNA
RT of RNA to cDNA
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.

Attempts are being made to make the assay strand specific and to assay total RNA as well.

Data from *D. melanogaster*

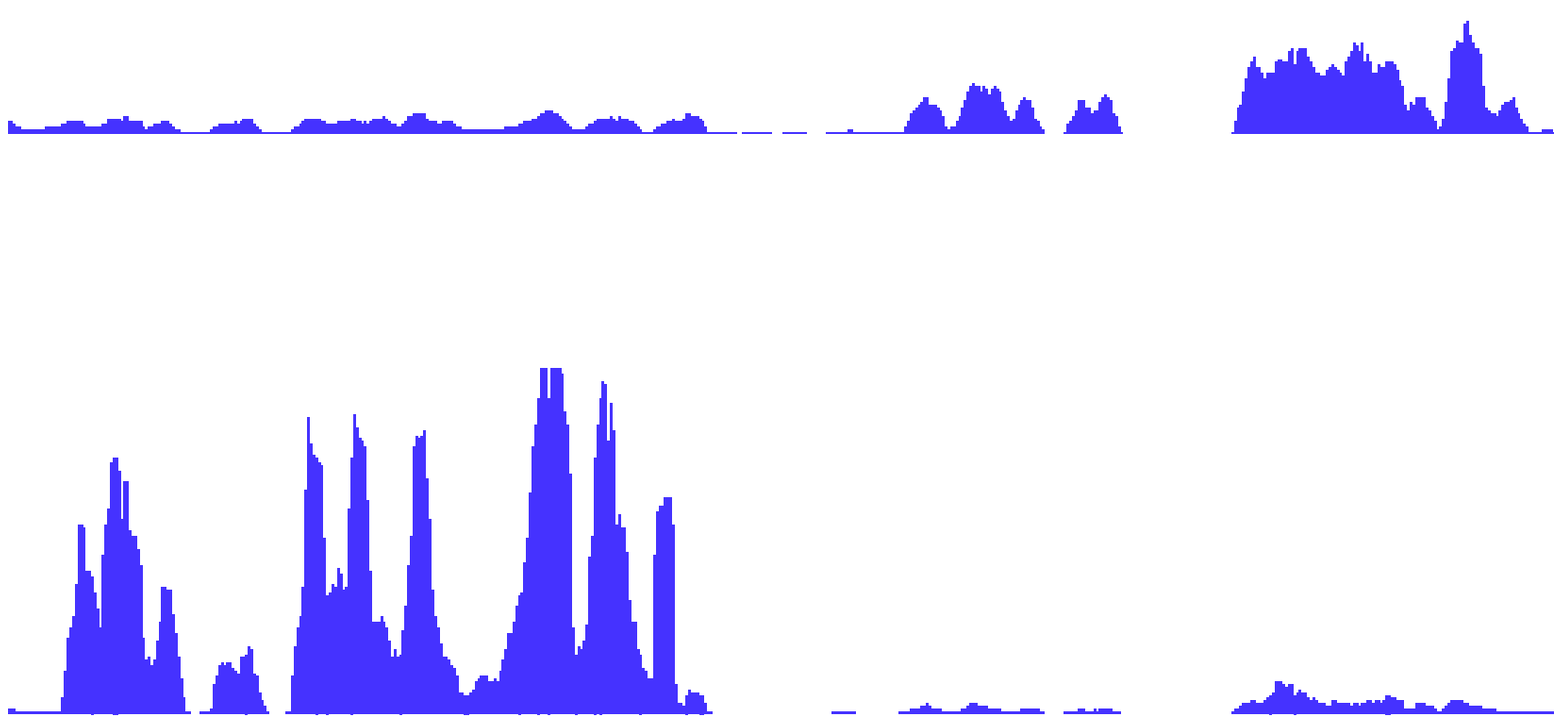
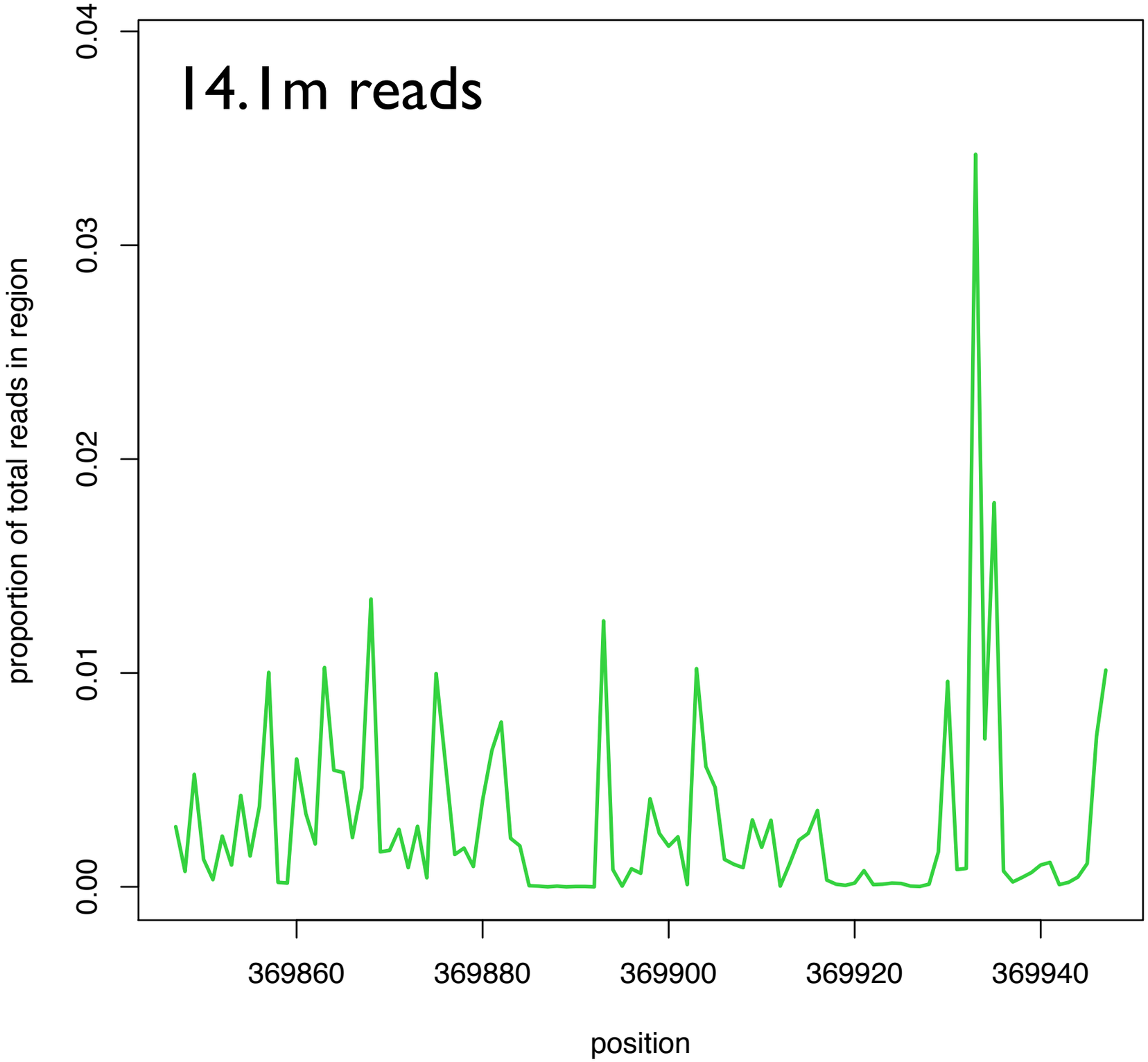
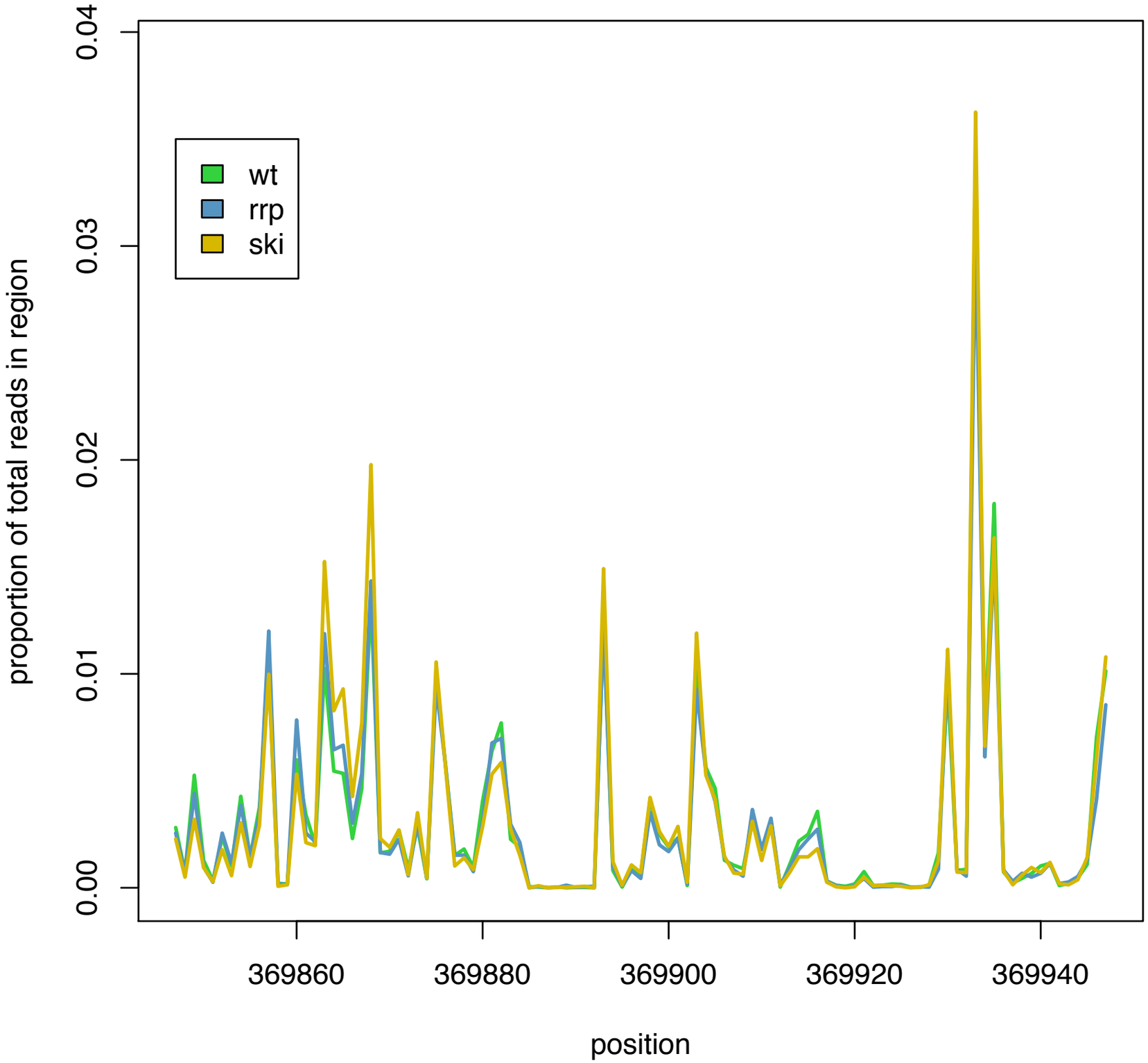


Image from Brenton Gravely

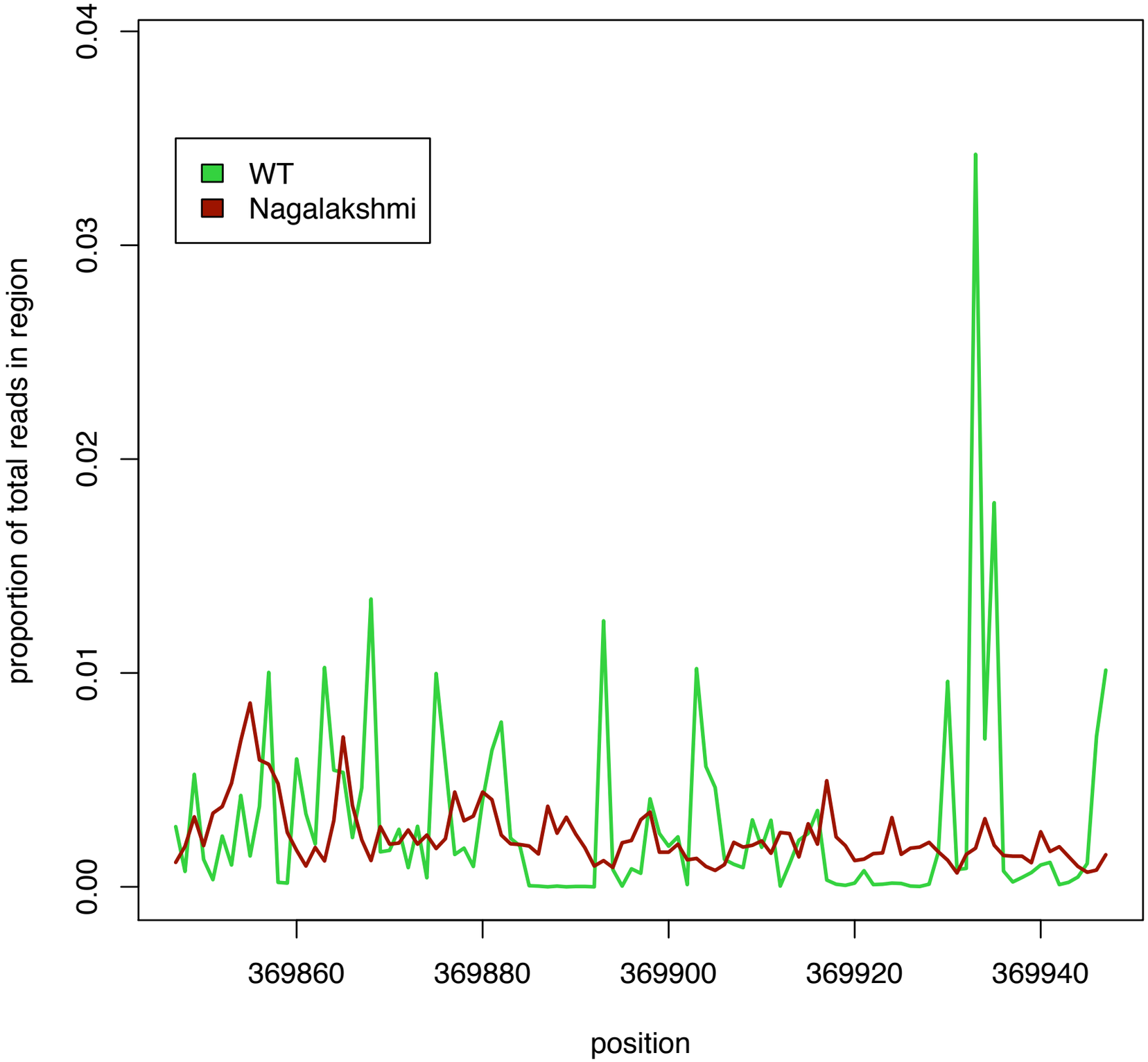
Base effect - single sample



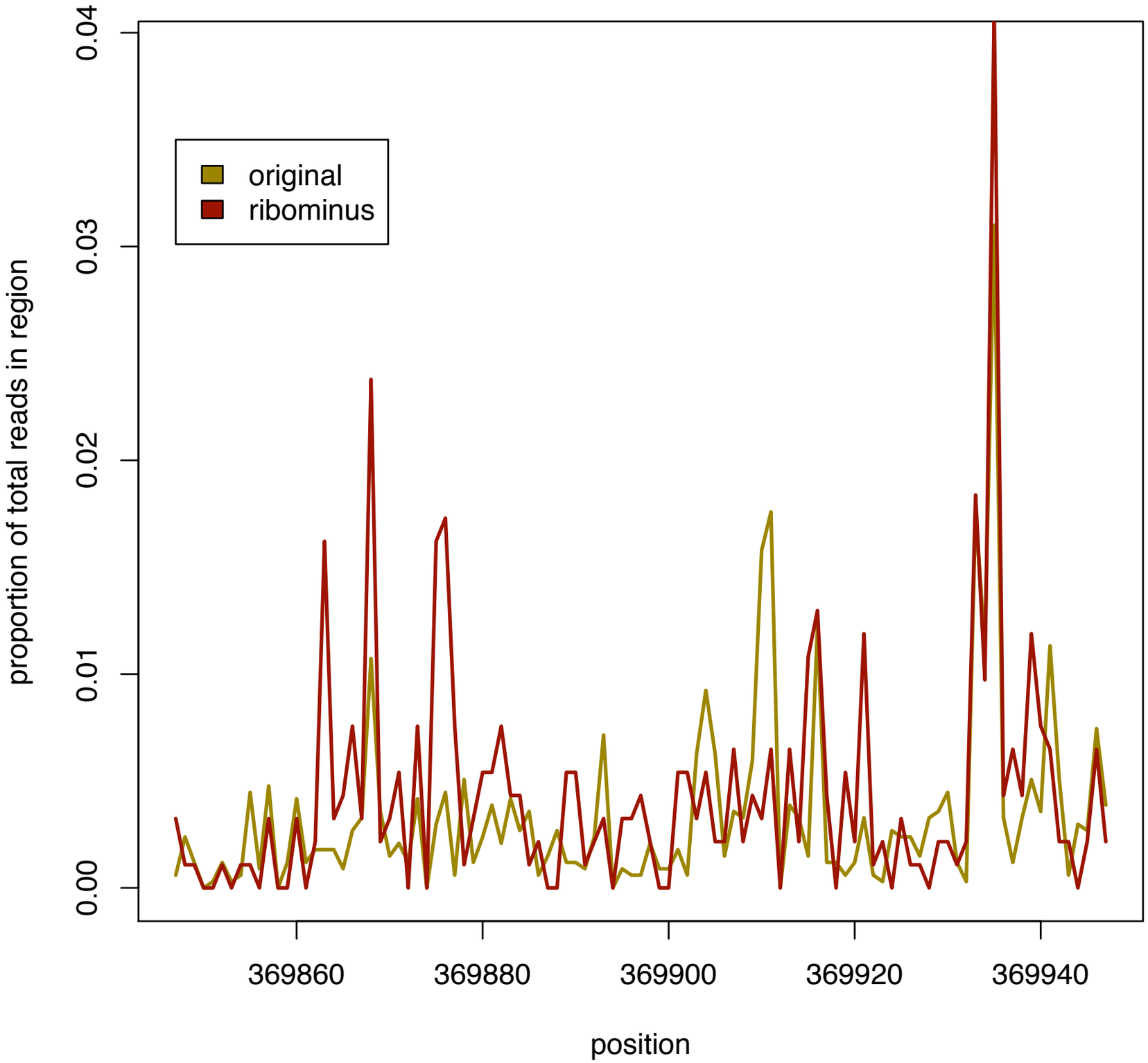
Base effect - multiple samples



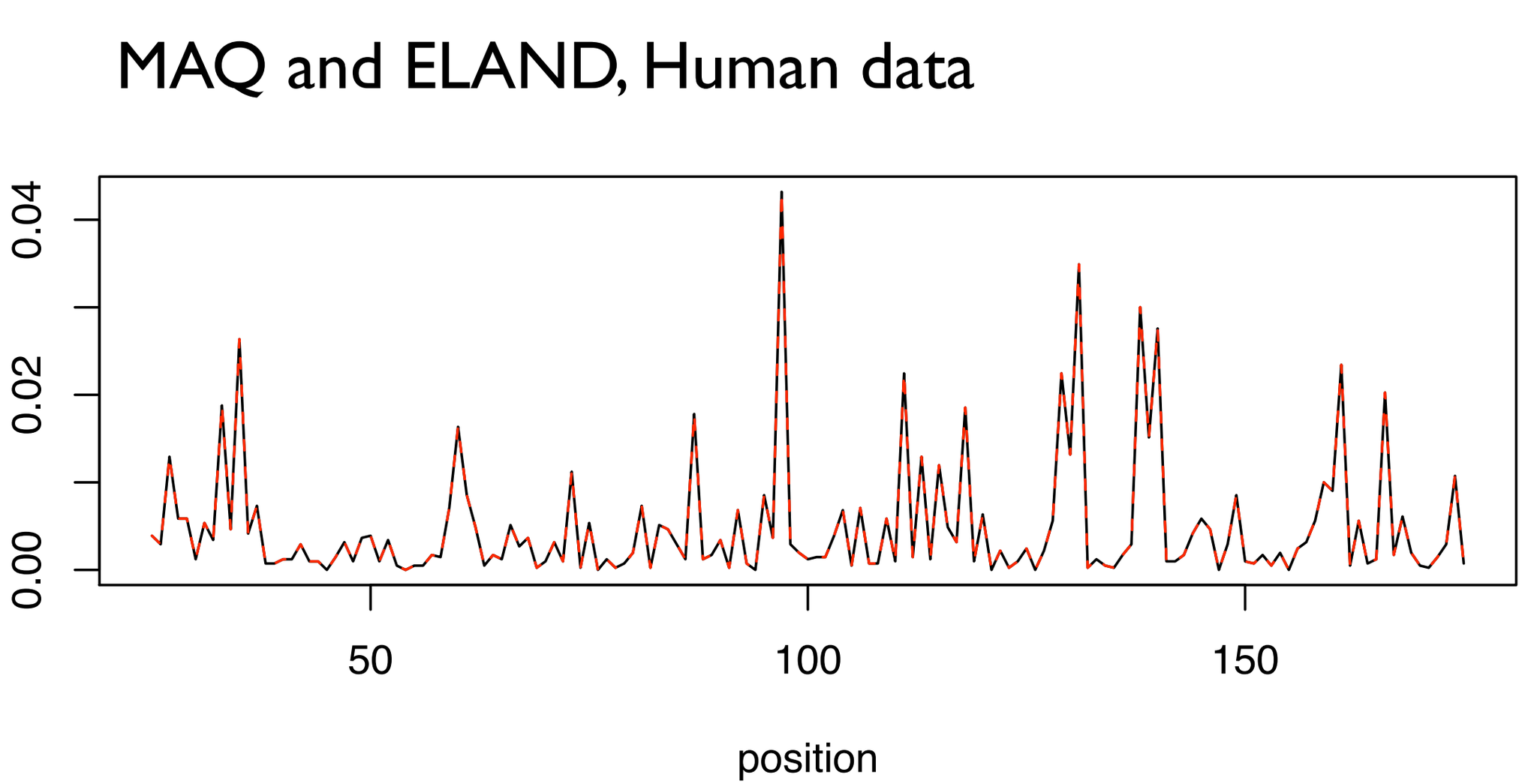
Base effect - different study (and prep)



Base effect - different prep



Base effect - different aligners



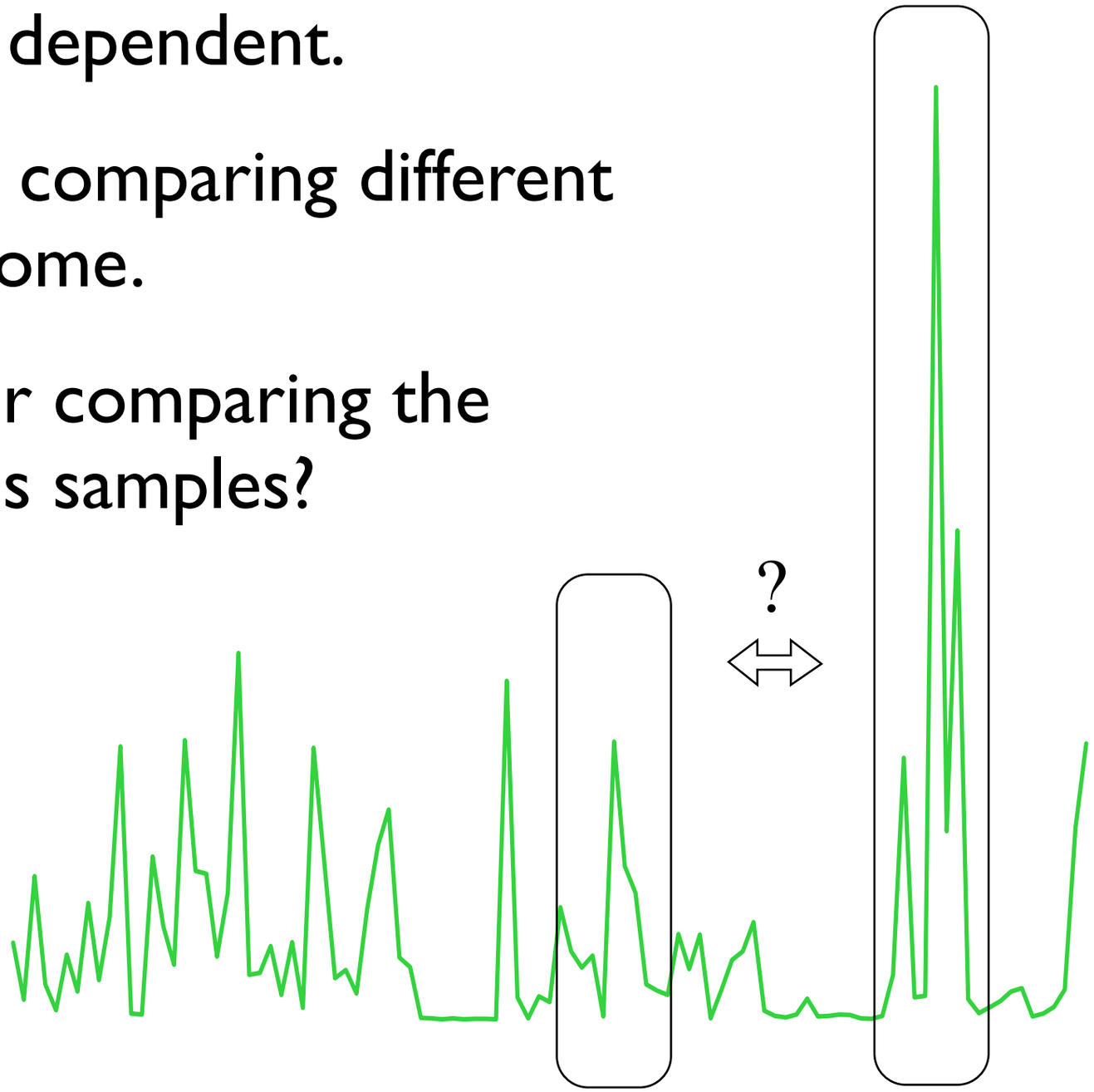
Base effect - conclusions

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?



Mapping reads to the transcriptome

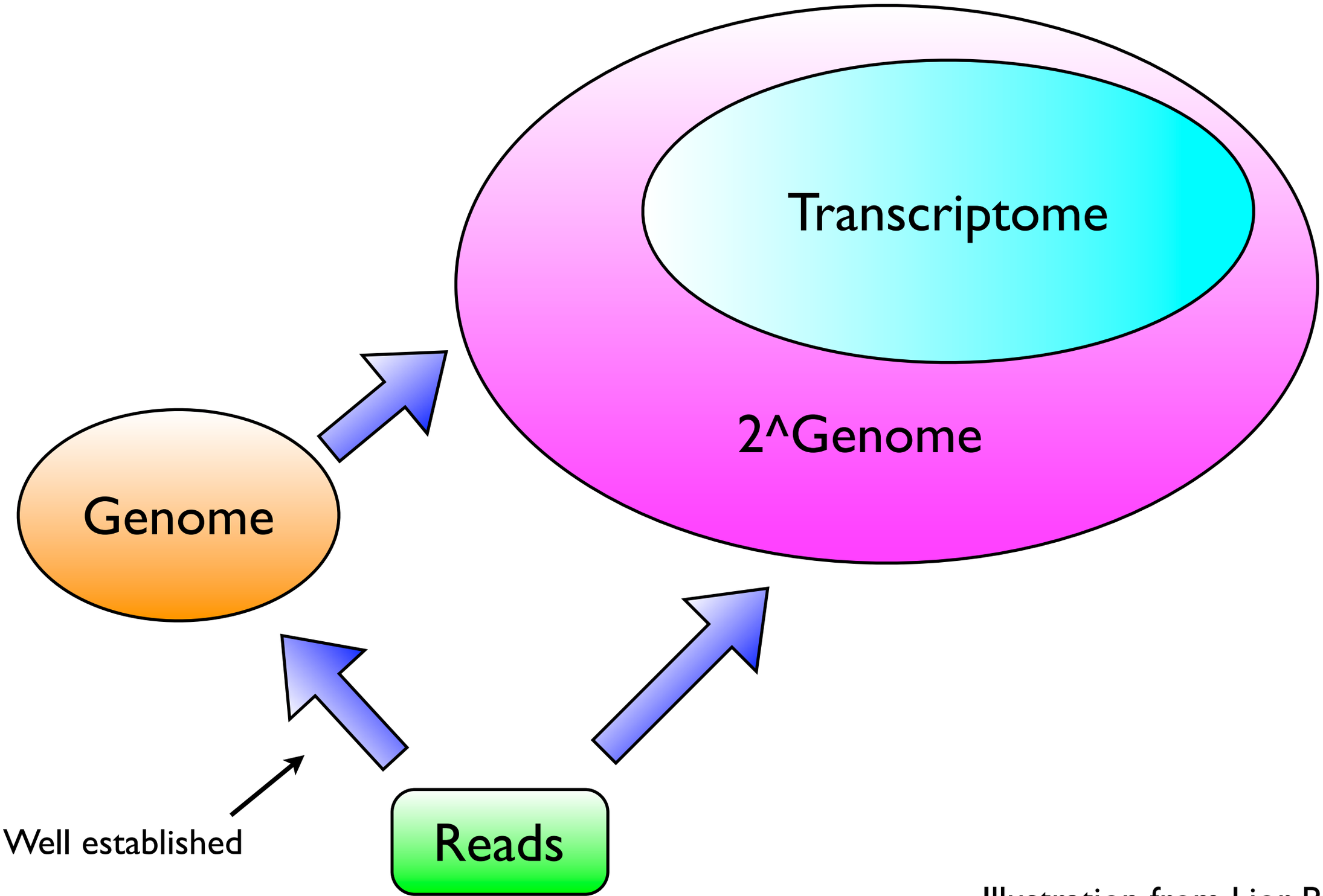
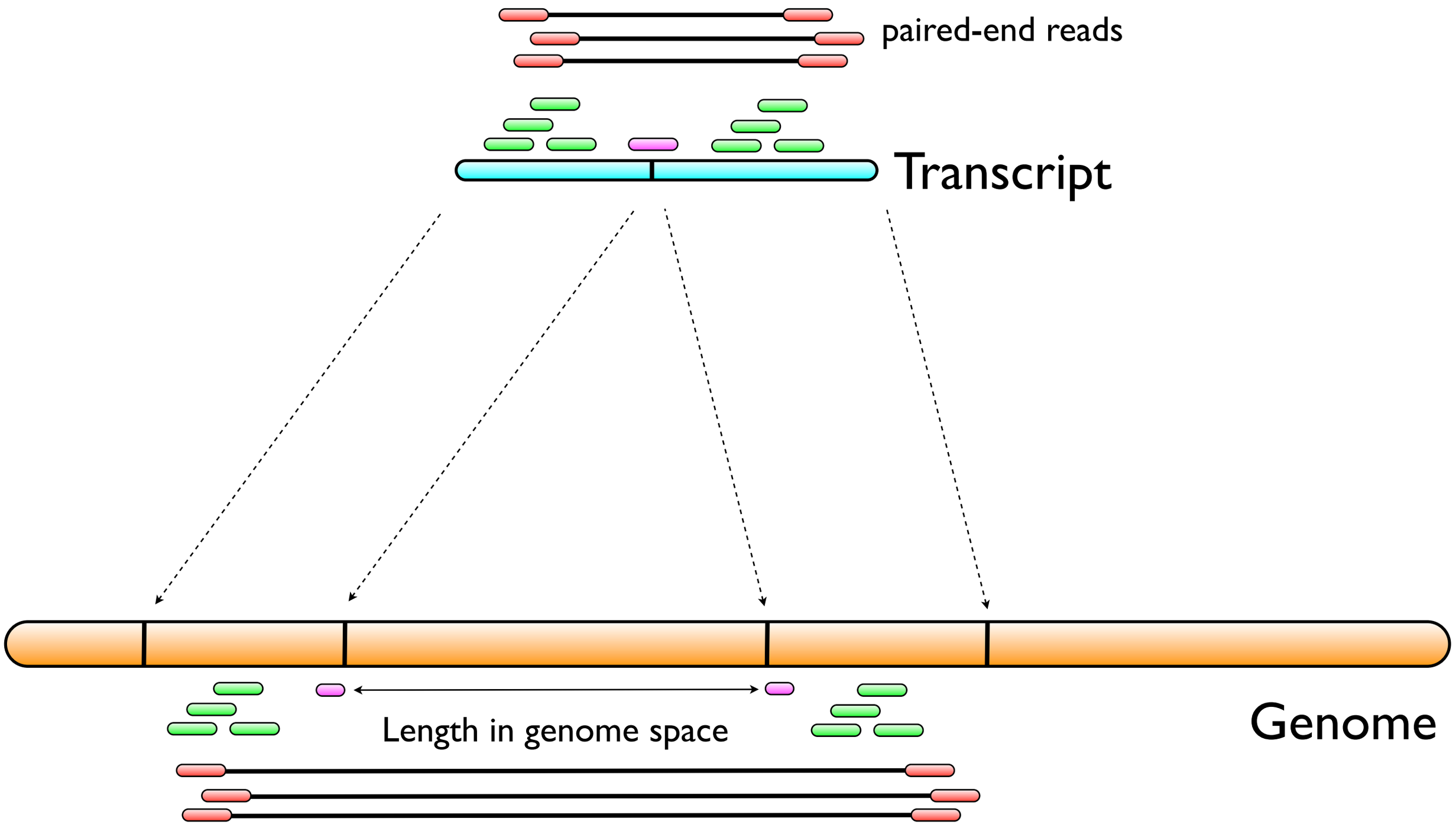


Illustration from Lior Patcher

Mapping transcripts



Junction reads

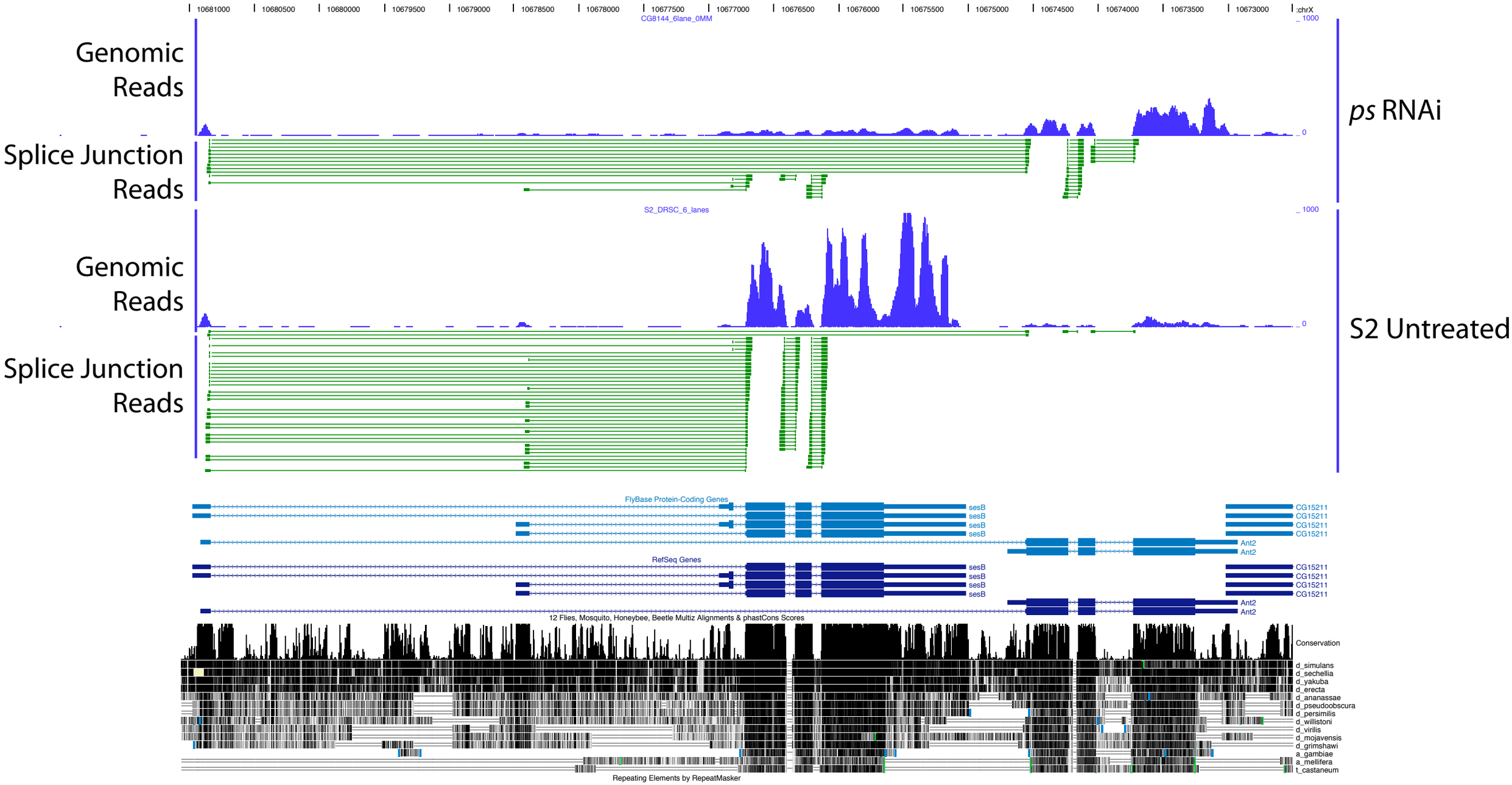


Image from Brenton Gravely

Junction reads, zoom

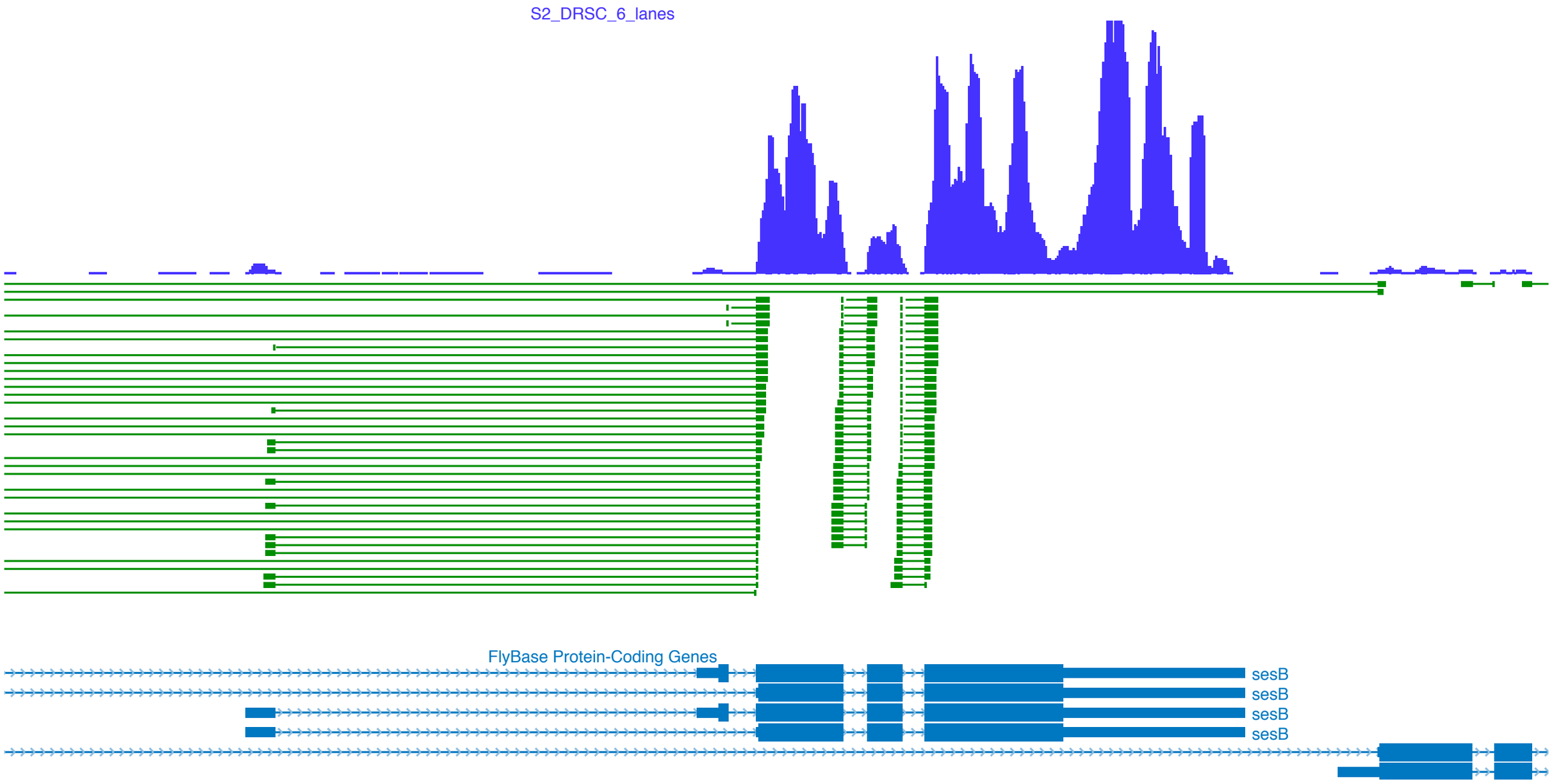
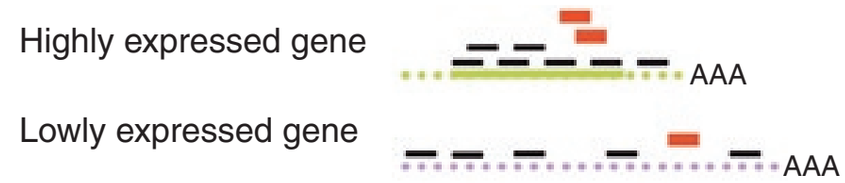


Image from Brenton Gravely

The basic approaches

a

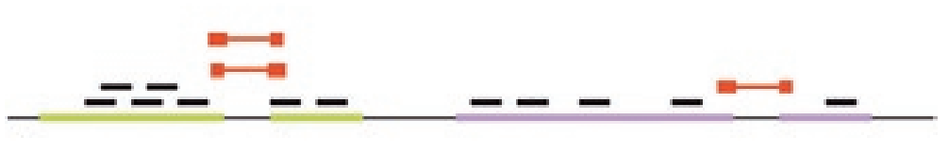
De novo assembly of the transcriptome



Read coverage must be high enough to build EST contigs (solid bar)

b

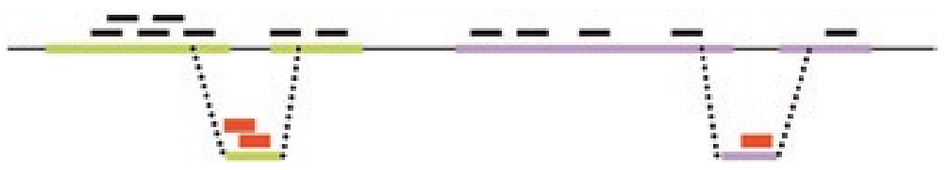
Map onto the genome



Read mapper must support splitting reads to record splices

c

Map onto the genome and splice junctions



Splice junctions sequences from either annotations or inferred

From Pepke (2009 Nat Methods)

Strategies for mapping to junctions

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

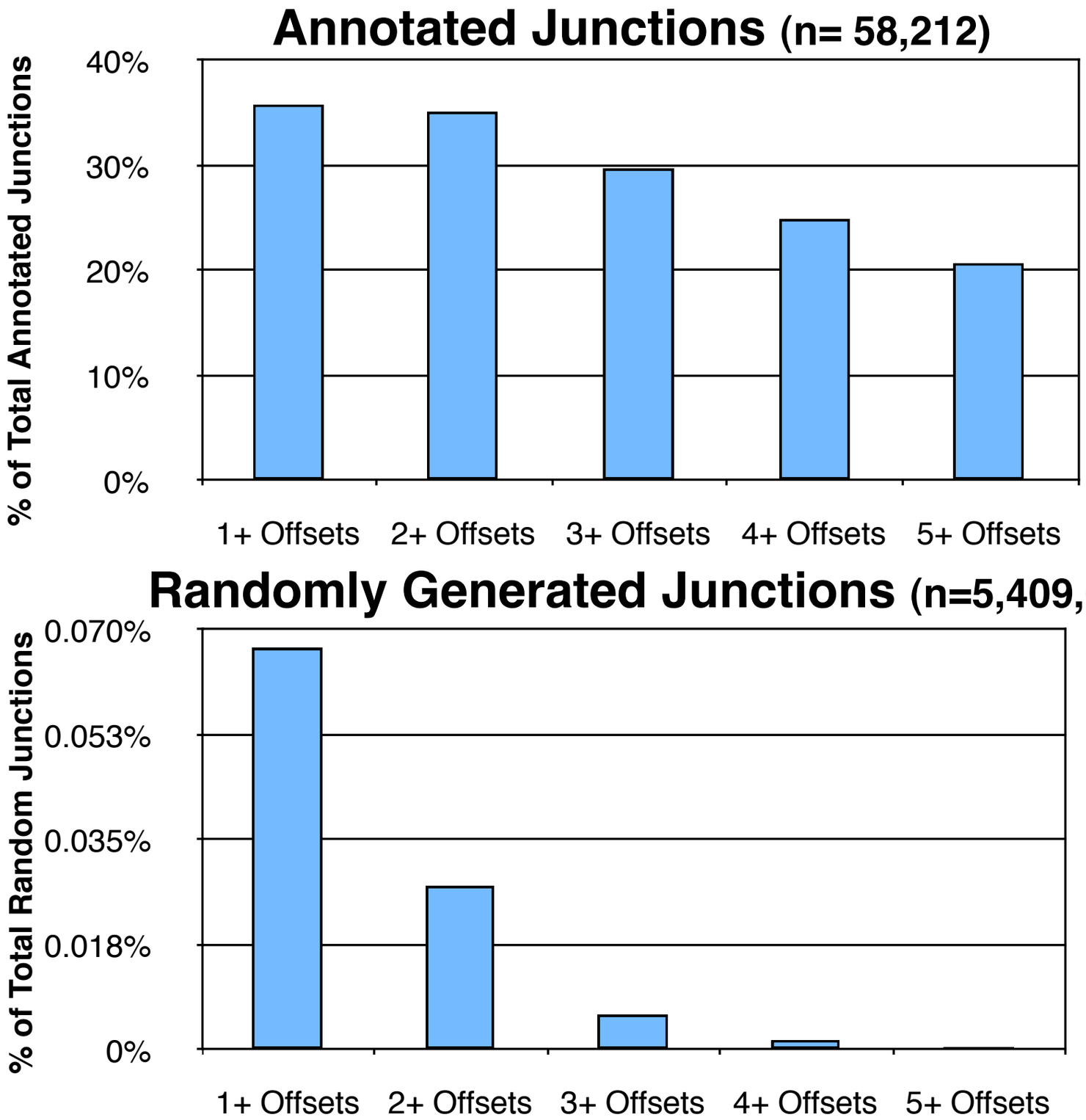
Map to combination of known exons.

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 does this).

Paired-end data will help with this.

FP rates for junctions

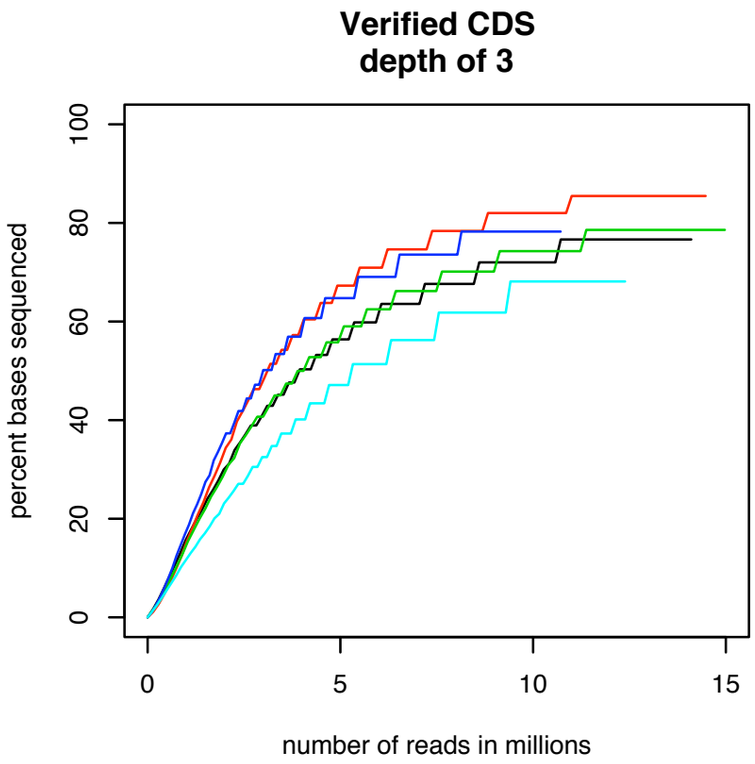


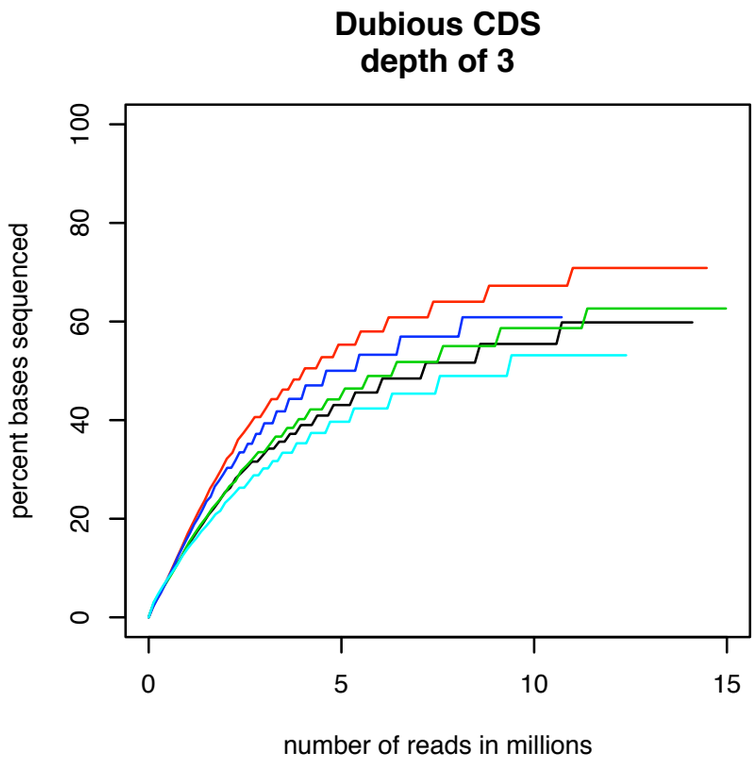
Mapping - conclusions

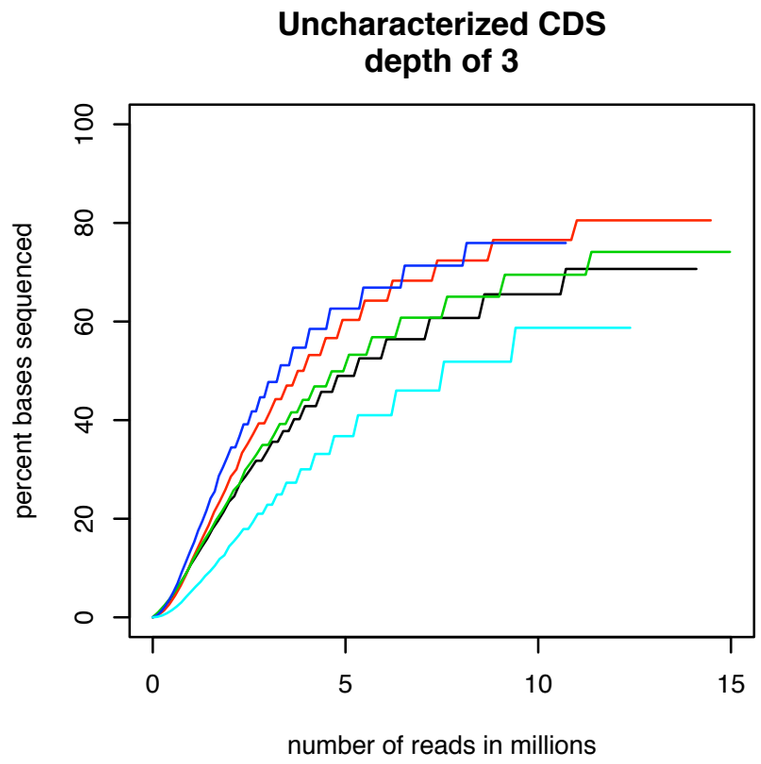
Mapping to transcript space is not easy.

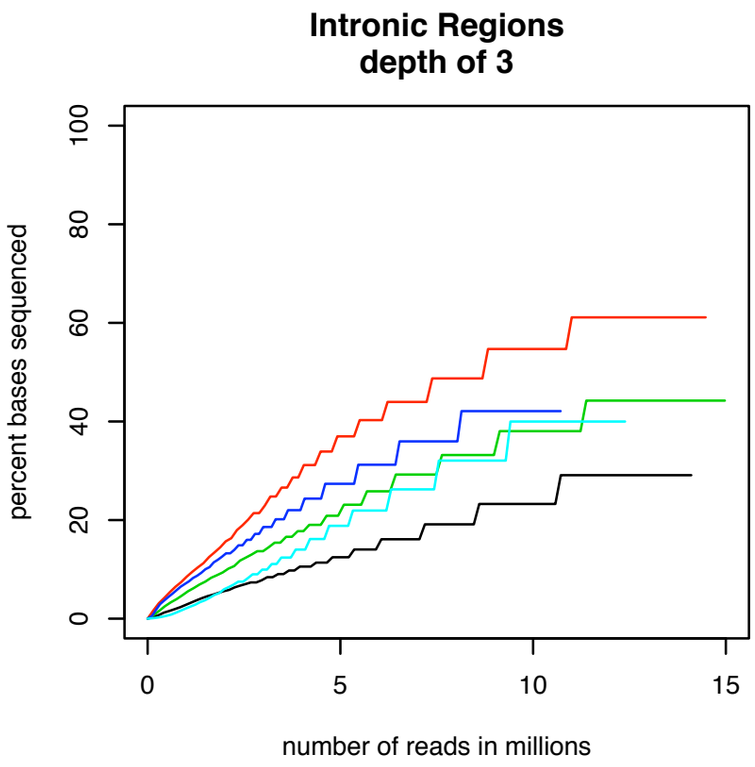
But essential for really understanding alternative splicing.

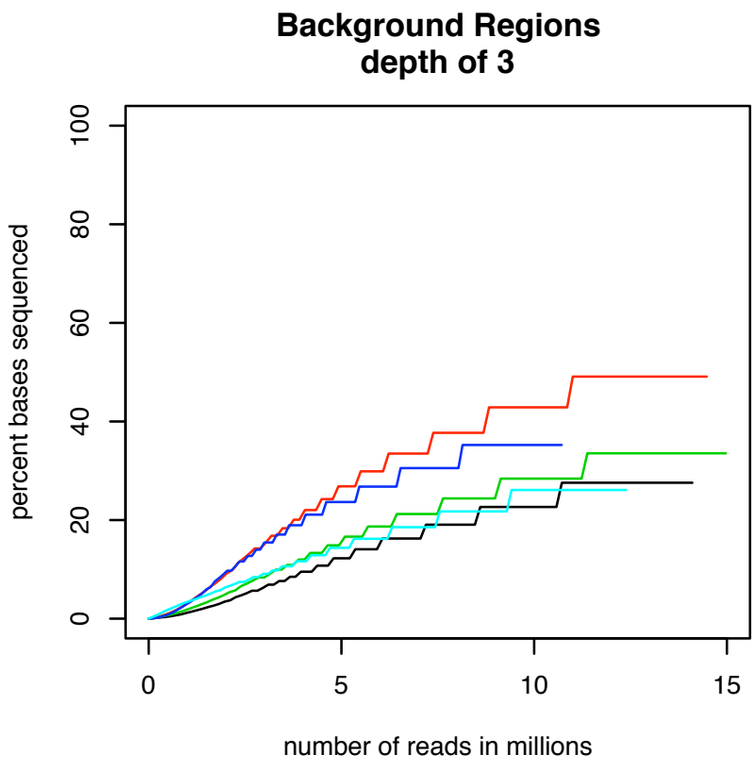
Coverage

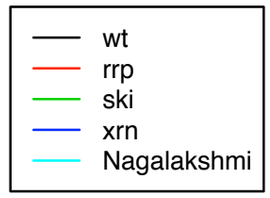

Verified CDS
depth of 3
percent bases sequenced
0
20
40
60
80
100
0
5
10
15
number of reads in millions


Dubious CDS
depth of 3
percent bases sequenced
0
20
40
60
80
100
0
5
10
15
number of reads in millions

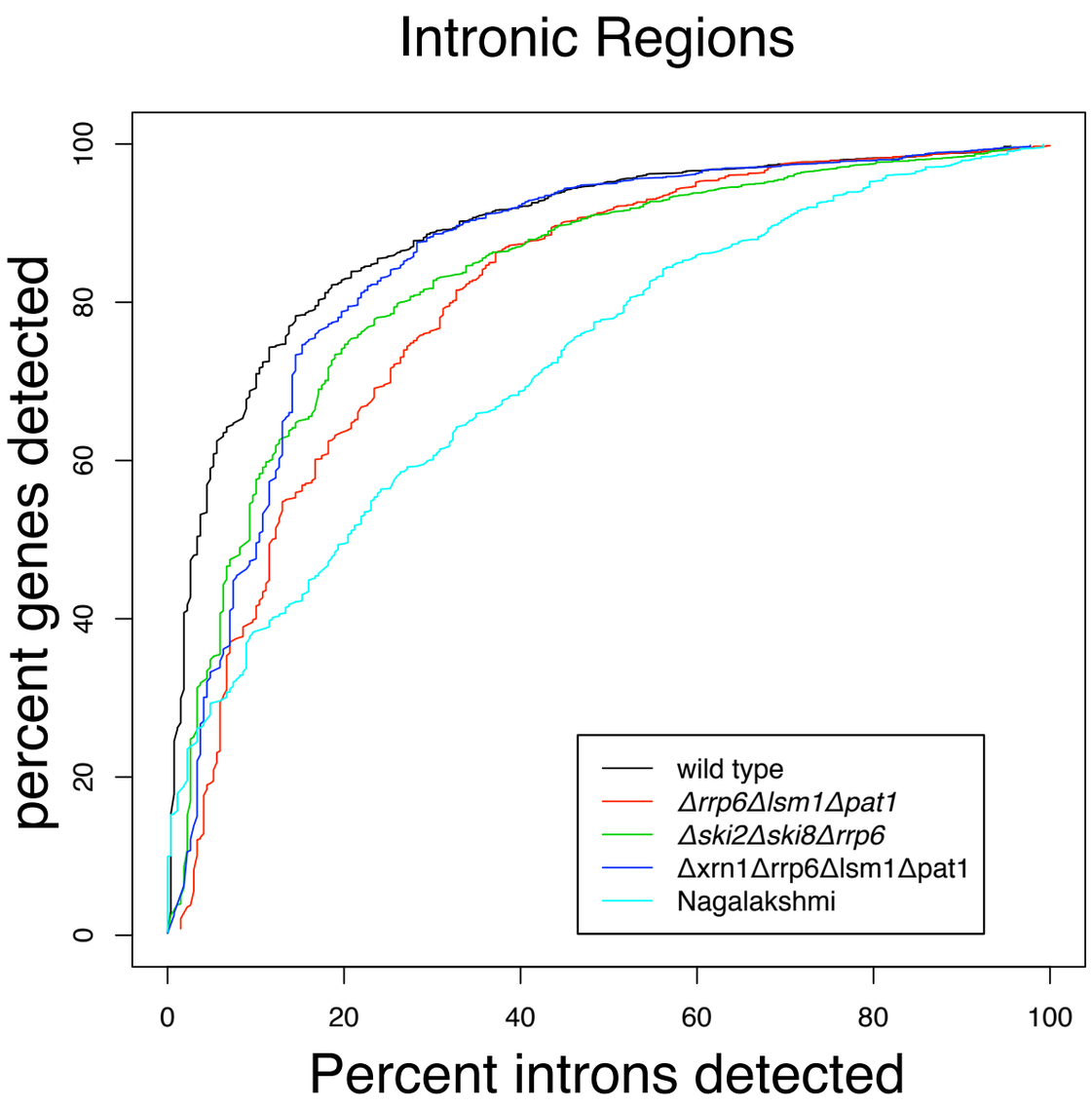

Uncharacterized CDS
depth of 3
percent bases sequenced
0
20
40
60
80
100
0
5
10
15
number of reads in millions

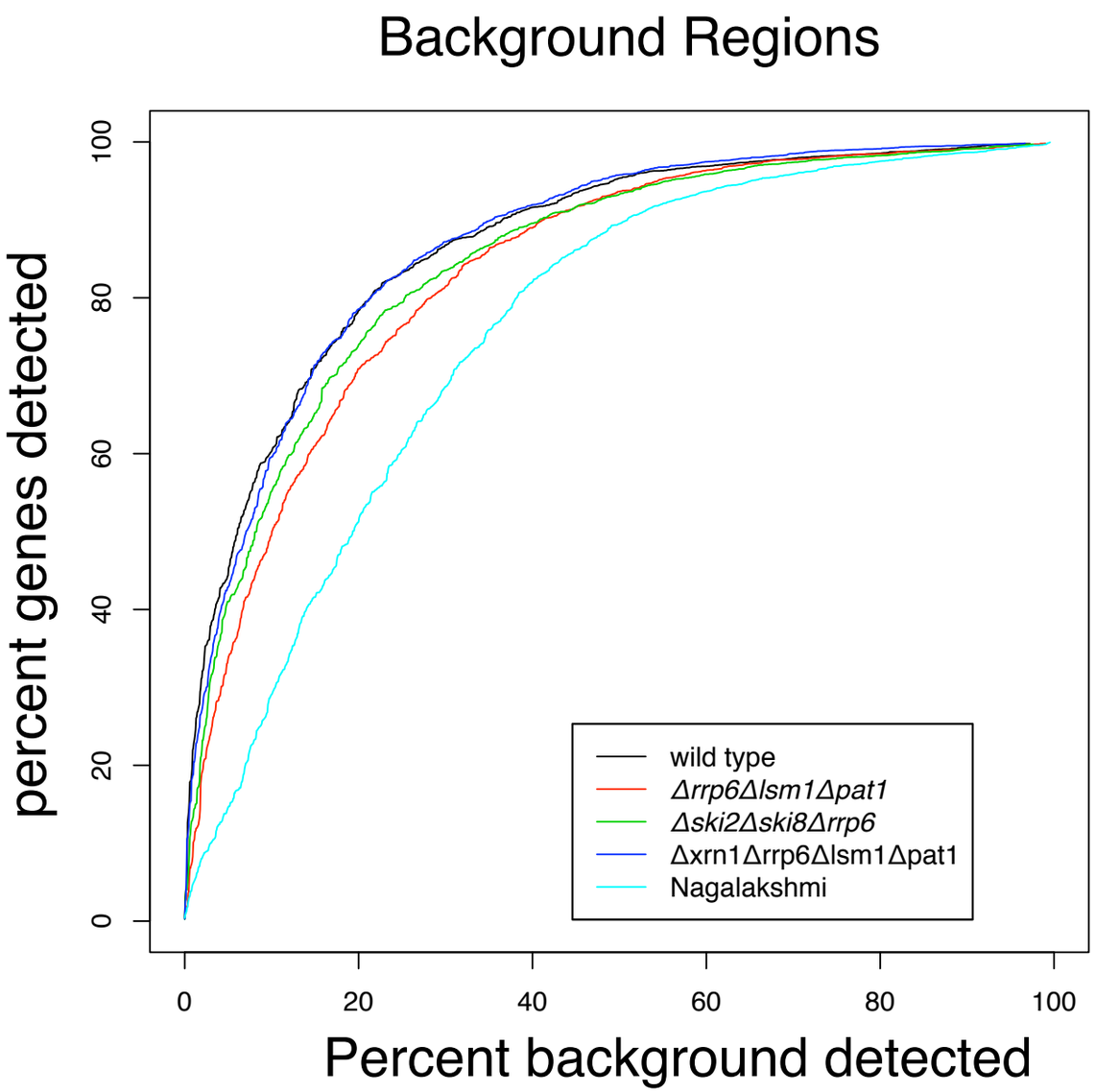

Intronic Regions
depth of 3
percent bases sequenced
0
20
40
60
80
100
0
5
10
15
number of reads in millions


Background Regions
depth of 3
percent bases sequenced
0
20
40
60
80
100
0
5
10
15
number of reads in millions


wt
rrp
ski
xrn
Nagalakshmi

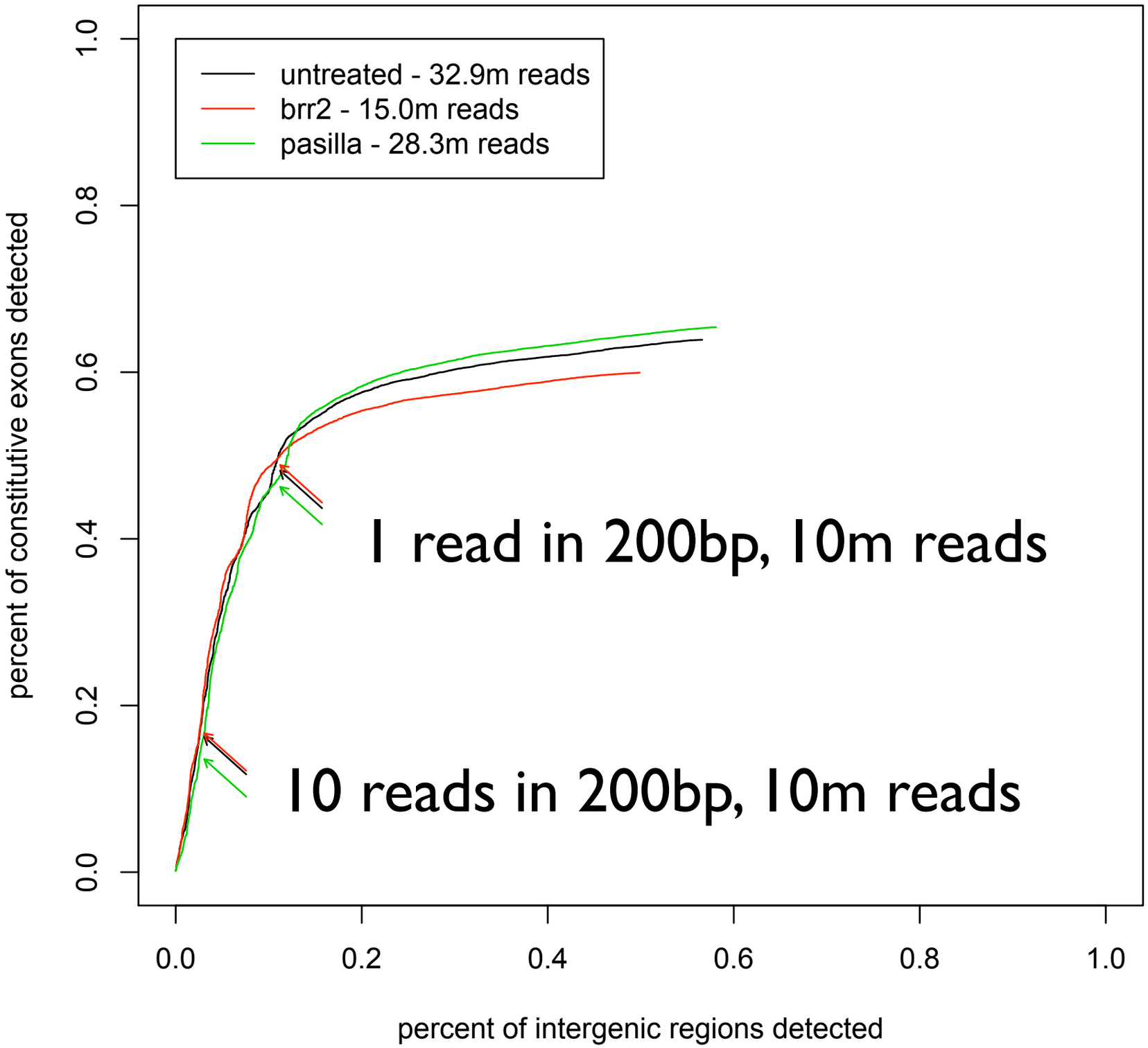
Detection in *Cerevisiae*





Background: outside any transcribed feature, subtracted a boundary, subtracted any region detected as transcribed in recent studies

Detection in *Drosophila*



Replication

Sources of variation

Lane variation

Flowcell variation

Library prep variation

Biological variation

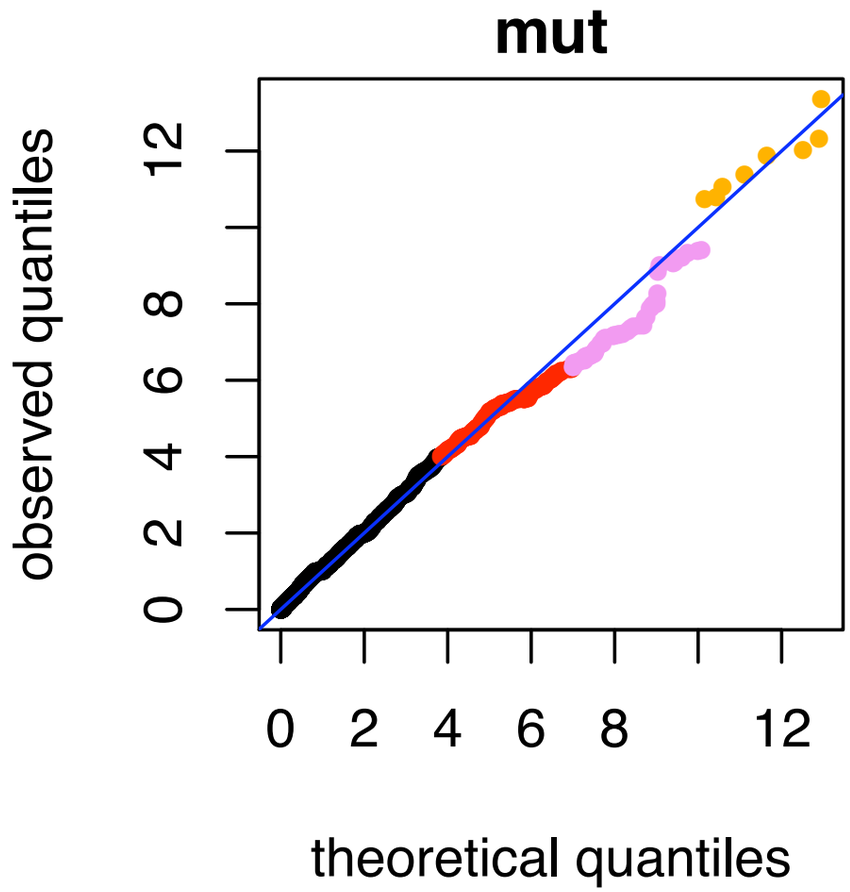
good fit

less good fit

Systematic differences

?: Is absolute quantification possible

Poisson model



Differential expression (DE)

Various methods have been proposed, all variants on a Poisson model.

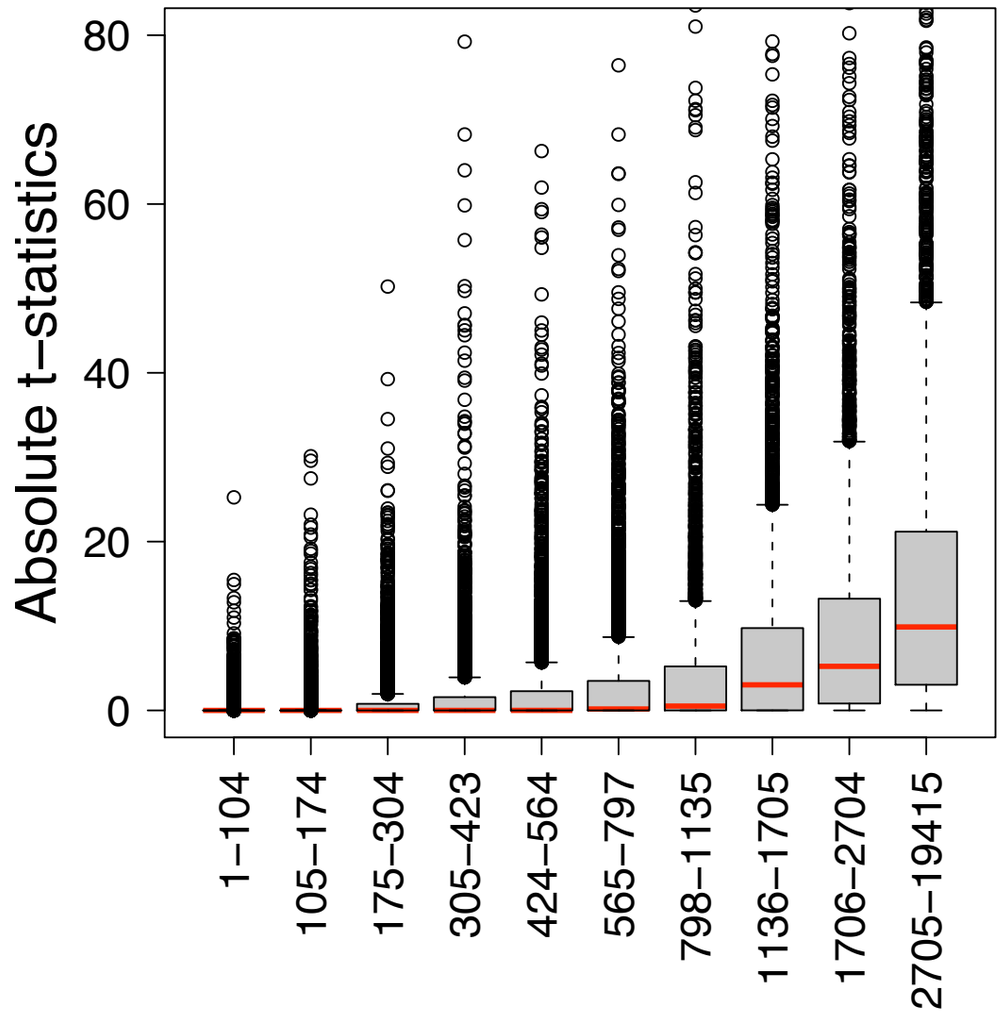
We find that Fisher's test or a GLM based LR test performs well. Of these two, the GLM based model is more flexible.

Normalization matters a lot (later). We suggests a simple upper-quartile global normalization; quantile normalization might be necessary for more noisy datasets.

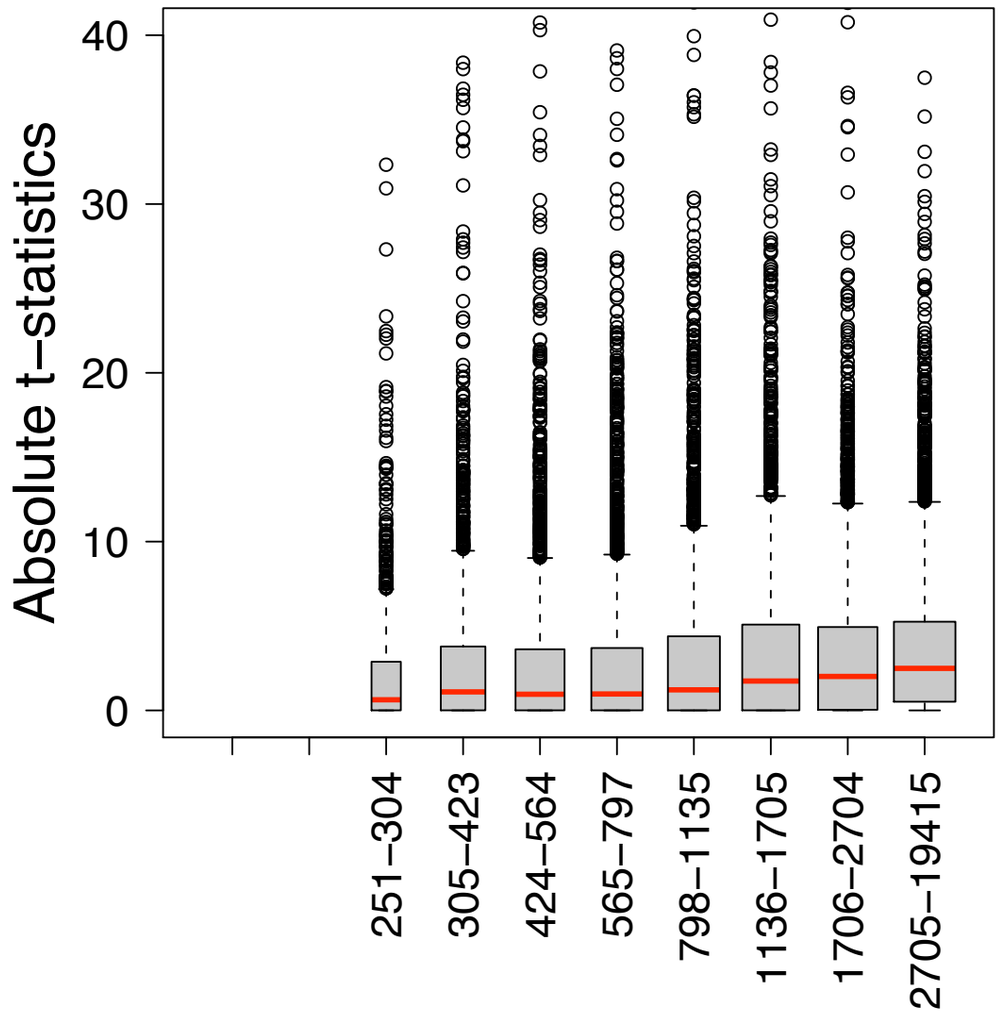
Most datasets only makes it possible to estimate the technical variance; the biological is ignored. This underestimates the variance.

In general, there is a significant flow cell effect, but the effect is small.

Bias based on gene length



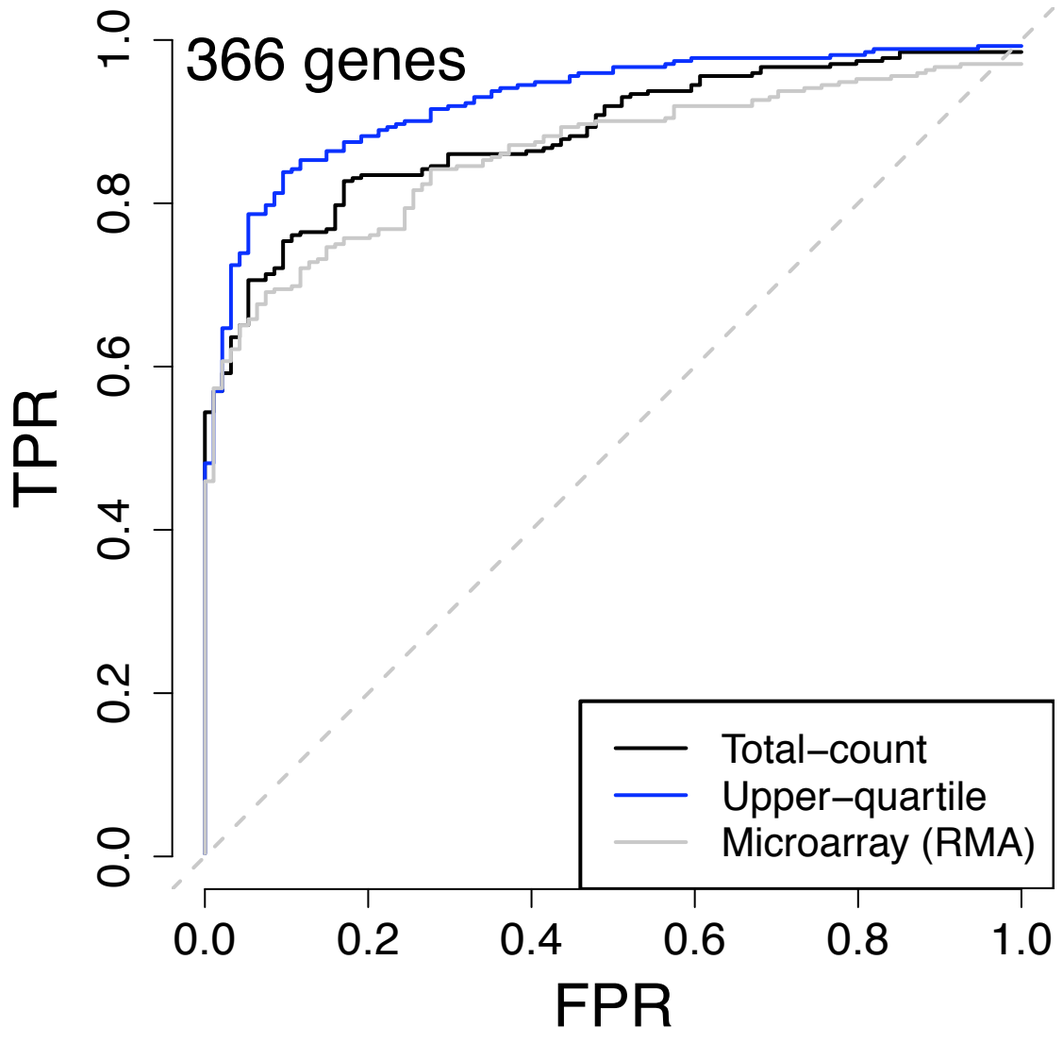
(a) Full-length UI genes



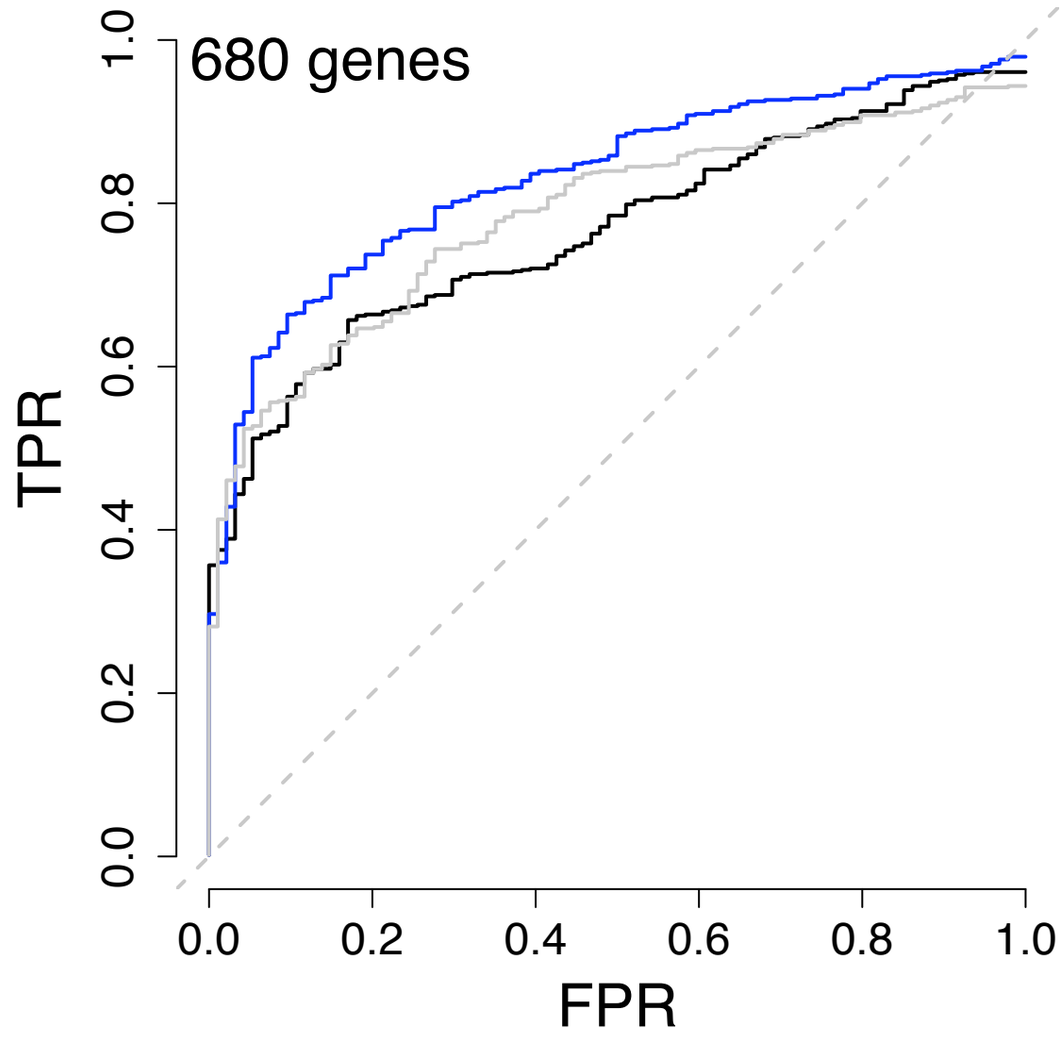
(b) 250-bp UI gene regions

Bad for interpretation

DE, the effect of normalization

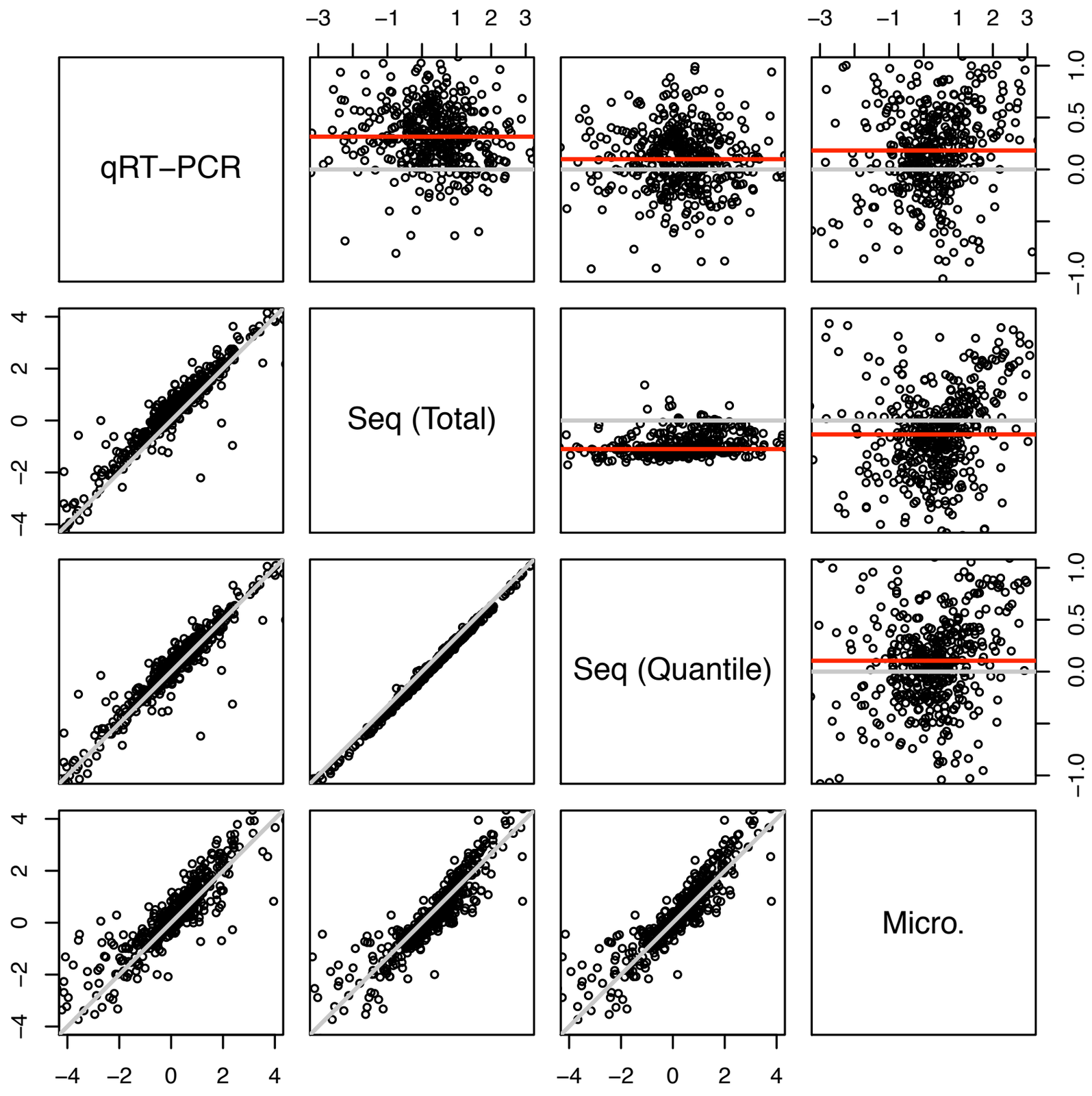


(a) qRT-PCR positives: LR$>$ 2



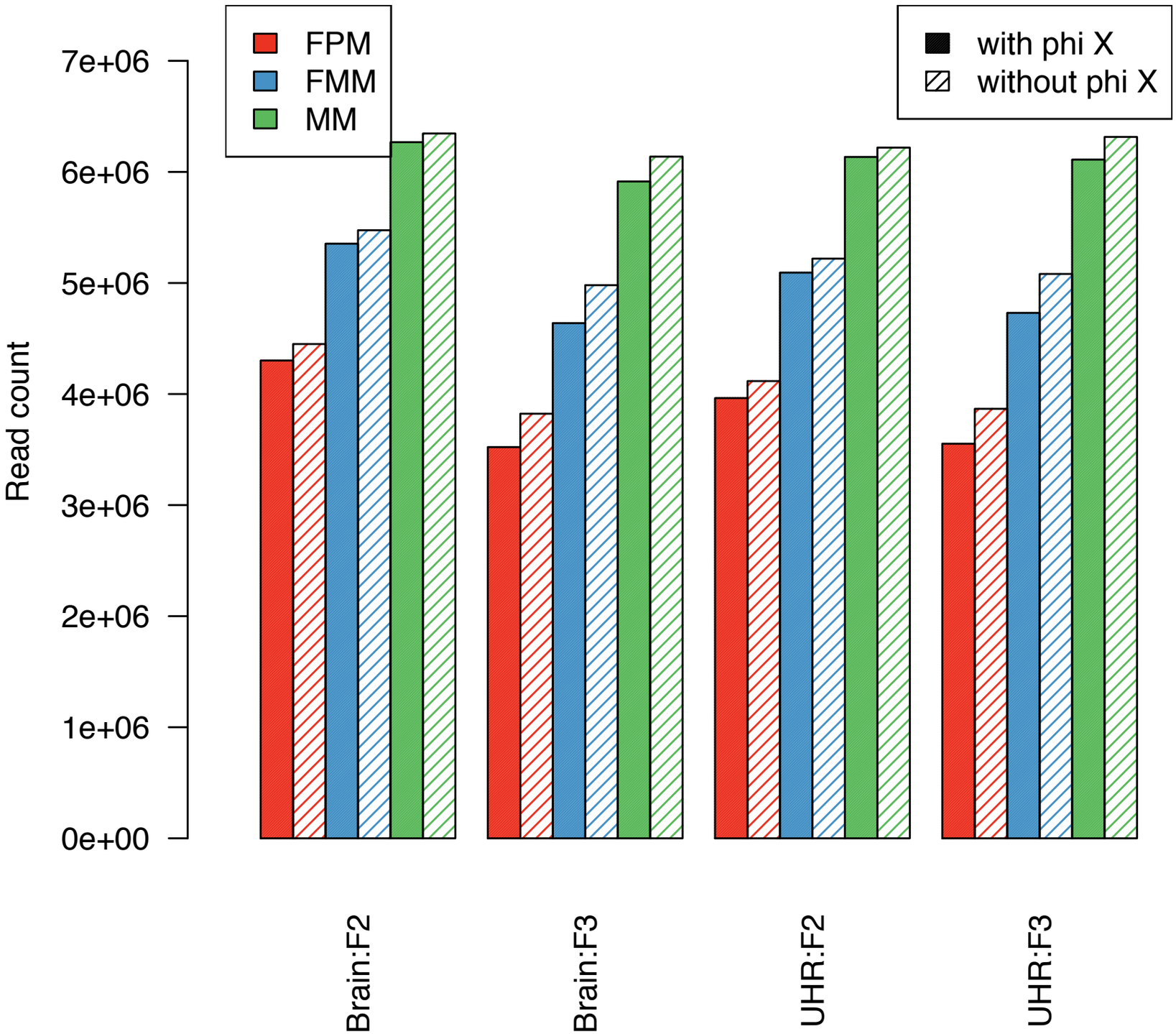
(b) qRT-PCR positives: LR$>$ 0.5

Normalization

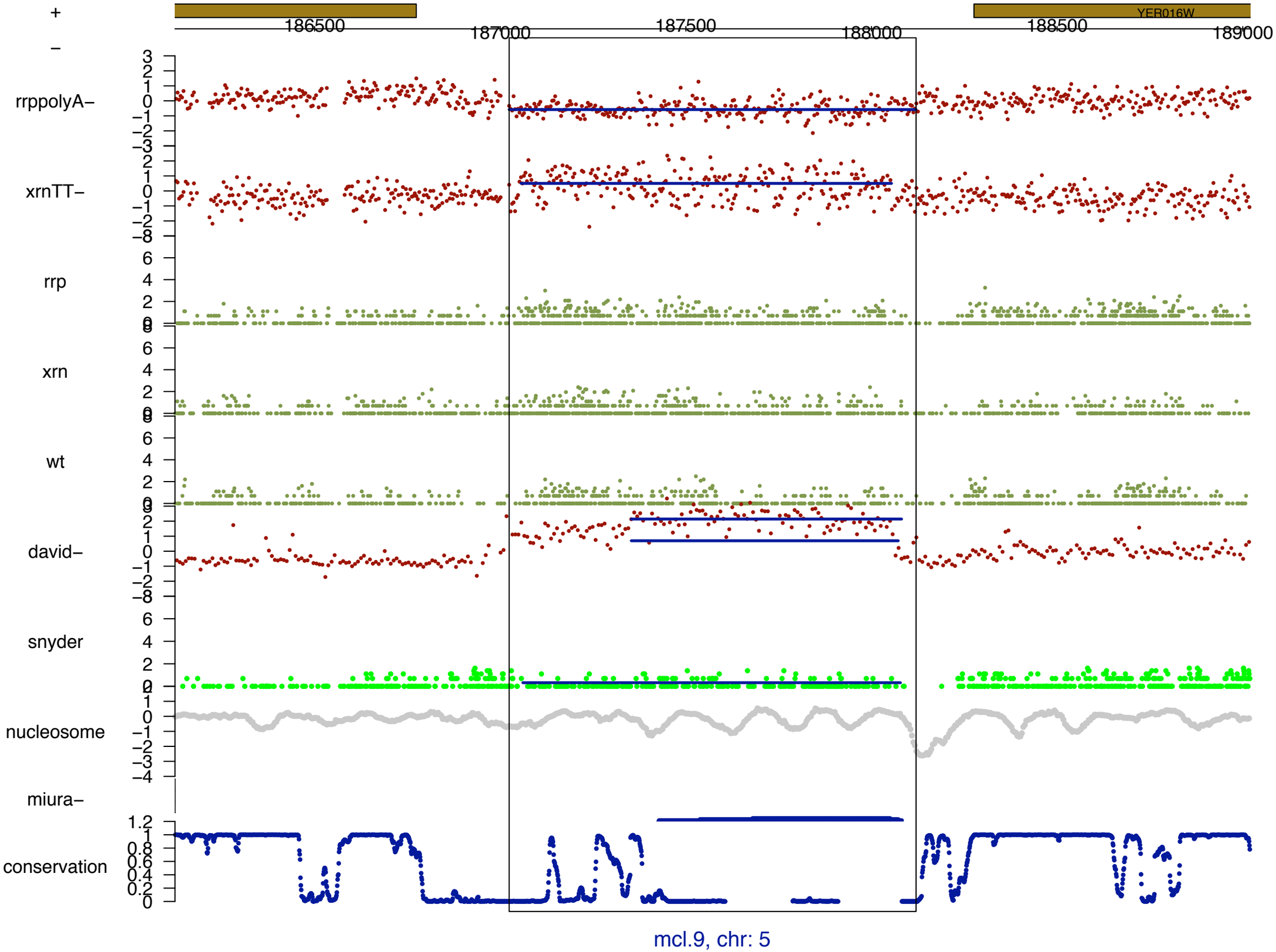


Seq (total) is essentially RPKMs

Running phi X does not seem necessary



Genome Graphs, example



References

Normalization, PhiX, DE comparison
Bullard, Purdom, Hansen, Dudoit (2009, tech report),
www.bepress.com/ucbbiostat/paper247/

Gene length bias
Oshlack, Wakefield (2009, Biology Direct)

Yeast data, coverage
Lee, Hansen, Bullard, Dudoit, Sherlock (2009, PLoS Genetics)

Current review
Pepke, Wold, Mortazavi (2009, Nat Methods)

A classic
Mortazavi, Williams, McCue, Schaffer, Wold (2008, Nature)

Acknowledgements

Statistics

Jim Bullard
Sandrine Dudoit
Elizabeth Purdom
Margaret Taub
Steffen Durinck
Terry Speed

RNA assembly

Cole Trapnell

S. Cerevisiae

Gavin Sherlock
Albert Lee

D. Melanogaster

Brenton Gravely
Mike Duff
Li Yang
Steven Brenner
Angela Brooks