RNA-Seq: Sequencing the Transcriptome

Kasper Daniel Hansen
Division of Biostatistics, UC Berkeley
Using Bioconductor for ChIP-Seq experiments
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.

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)
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*

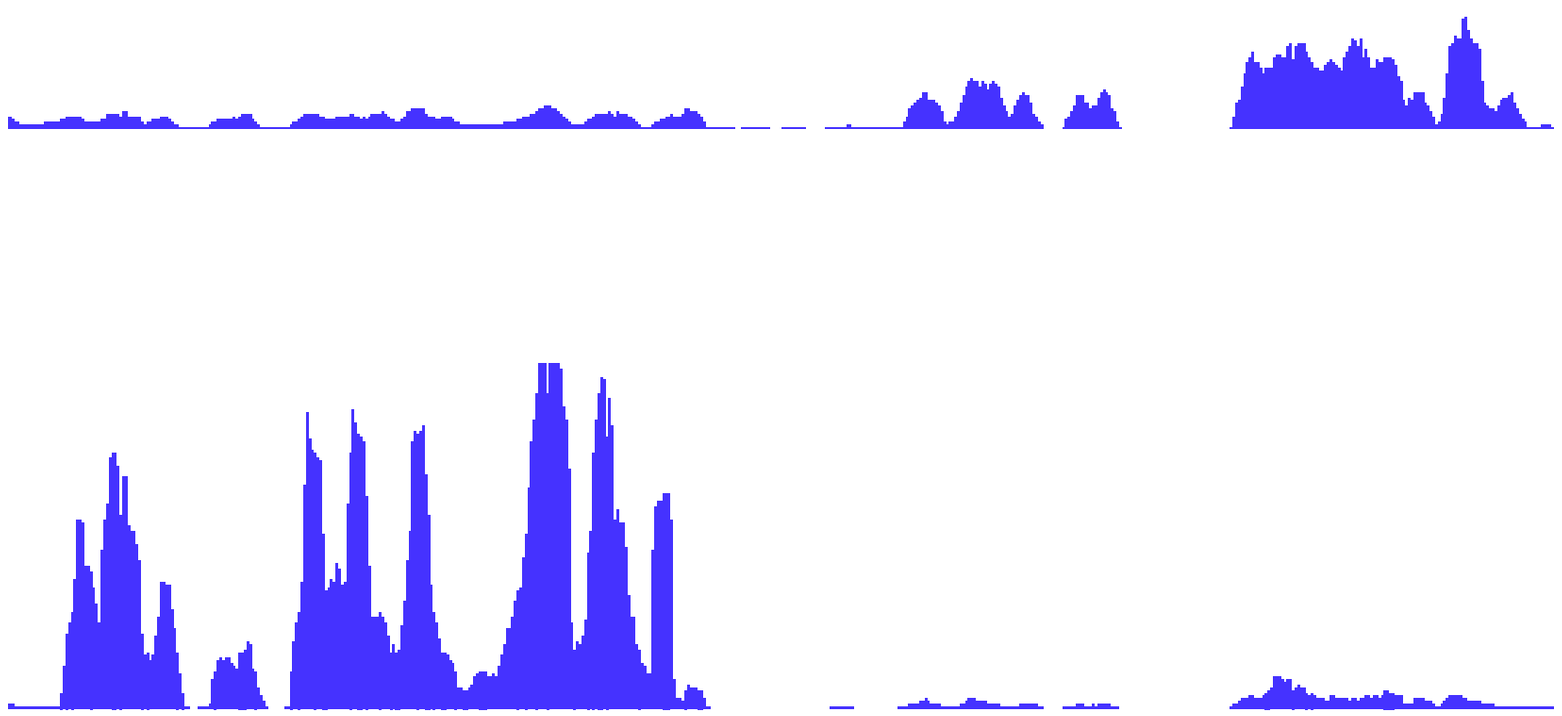
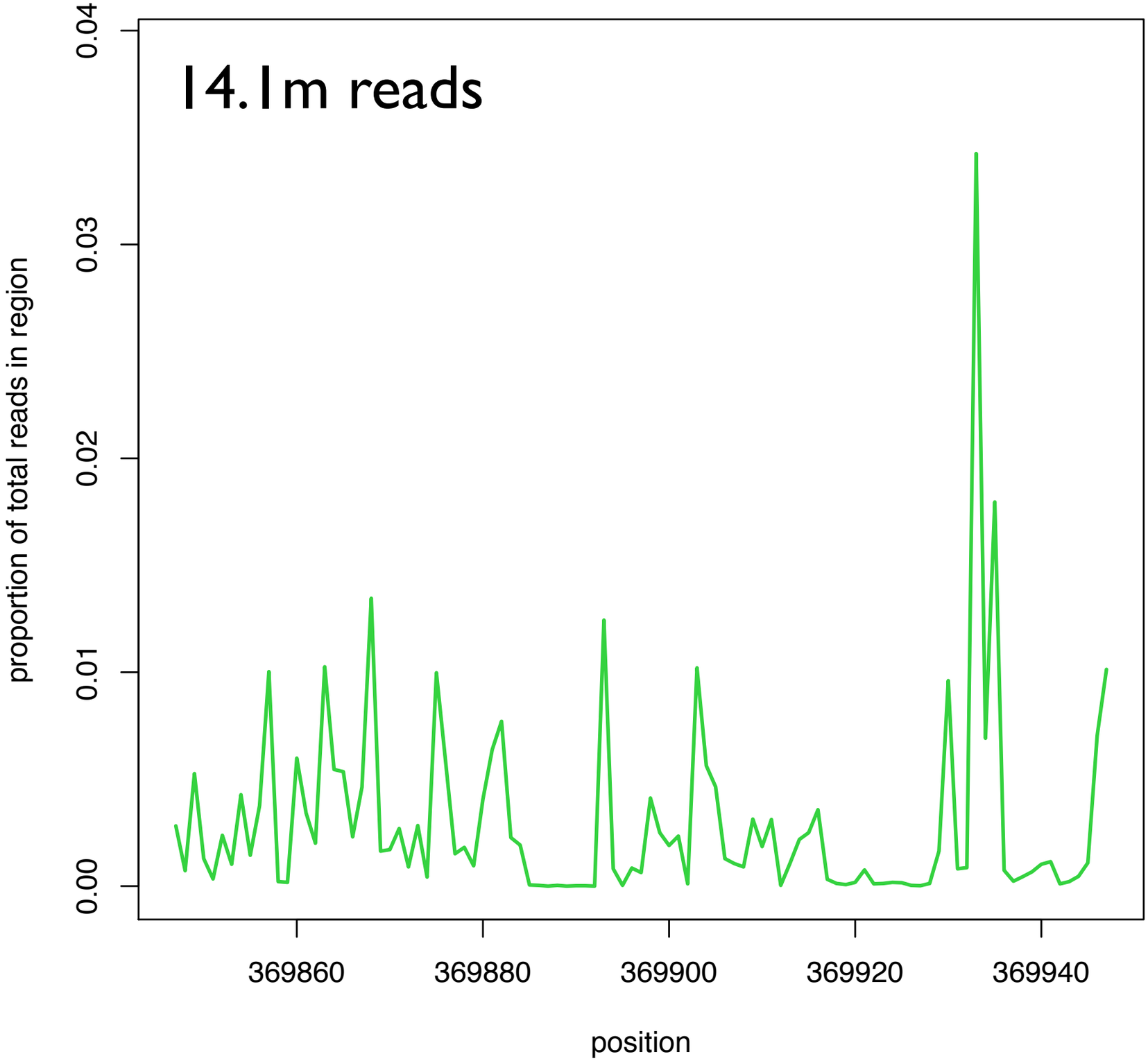
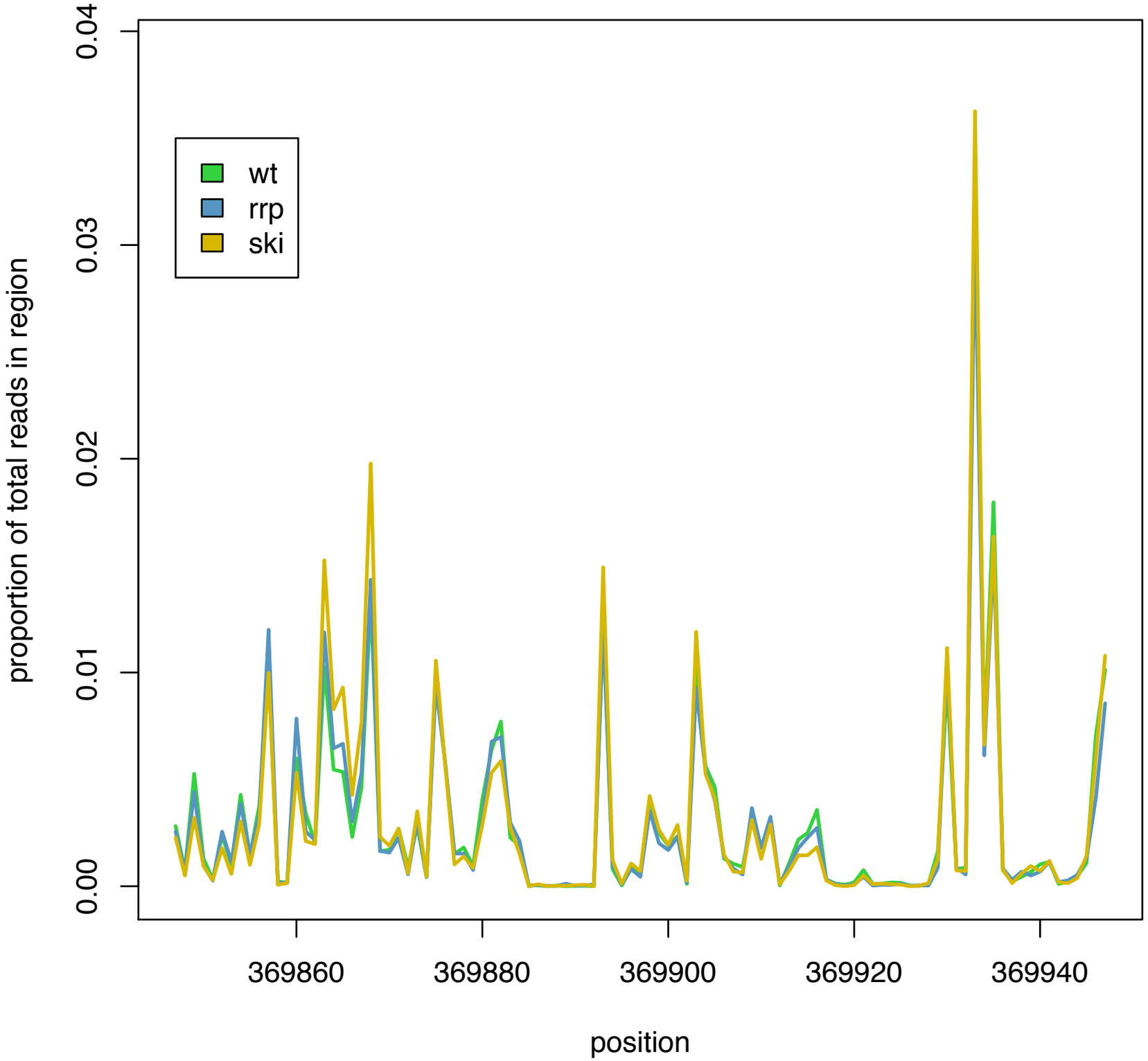


Image from Brenton Gravely

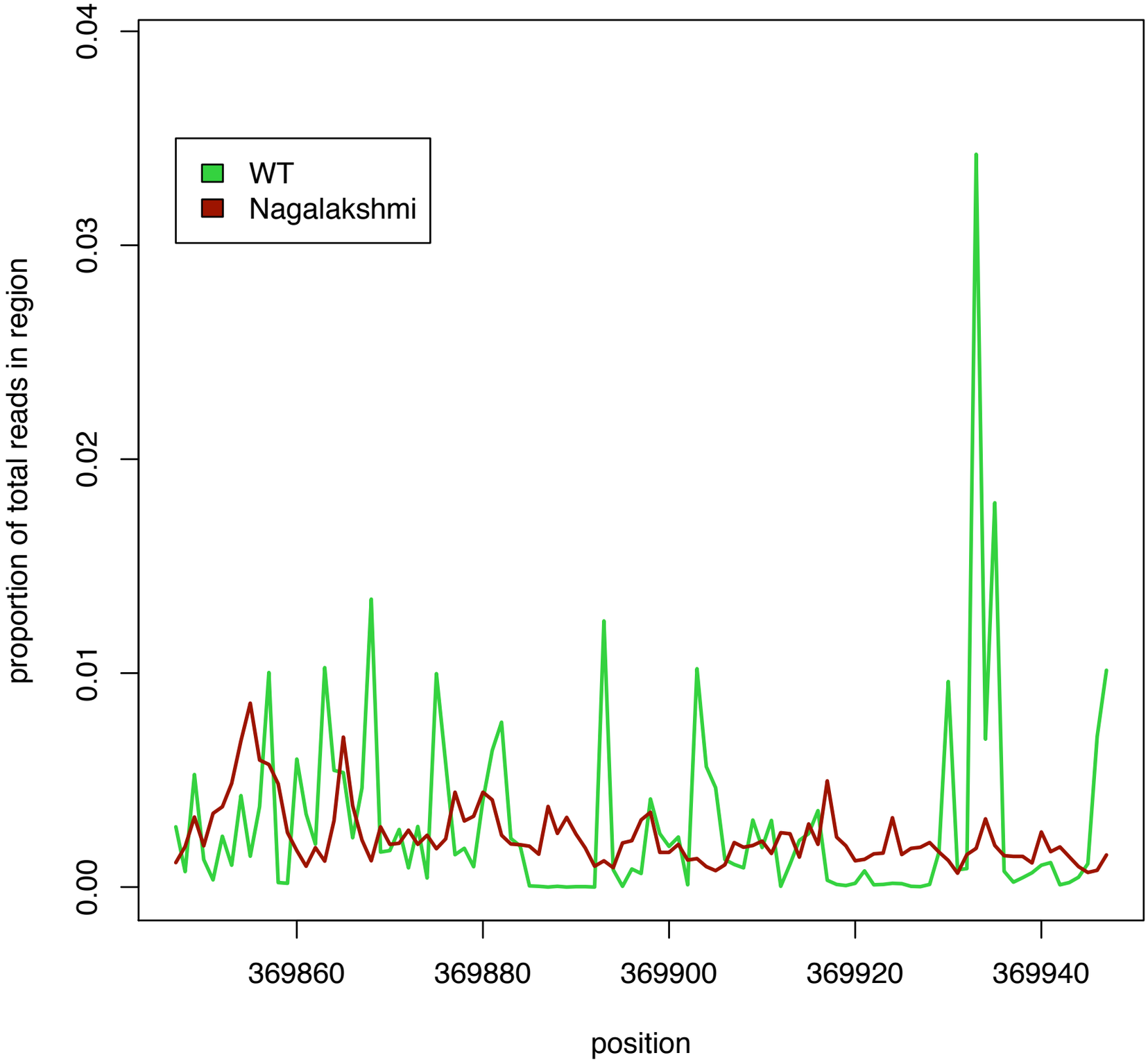
Base effect - single sample



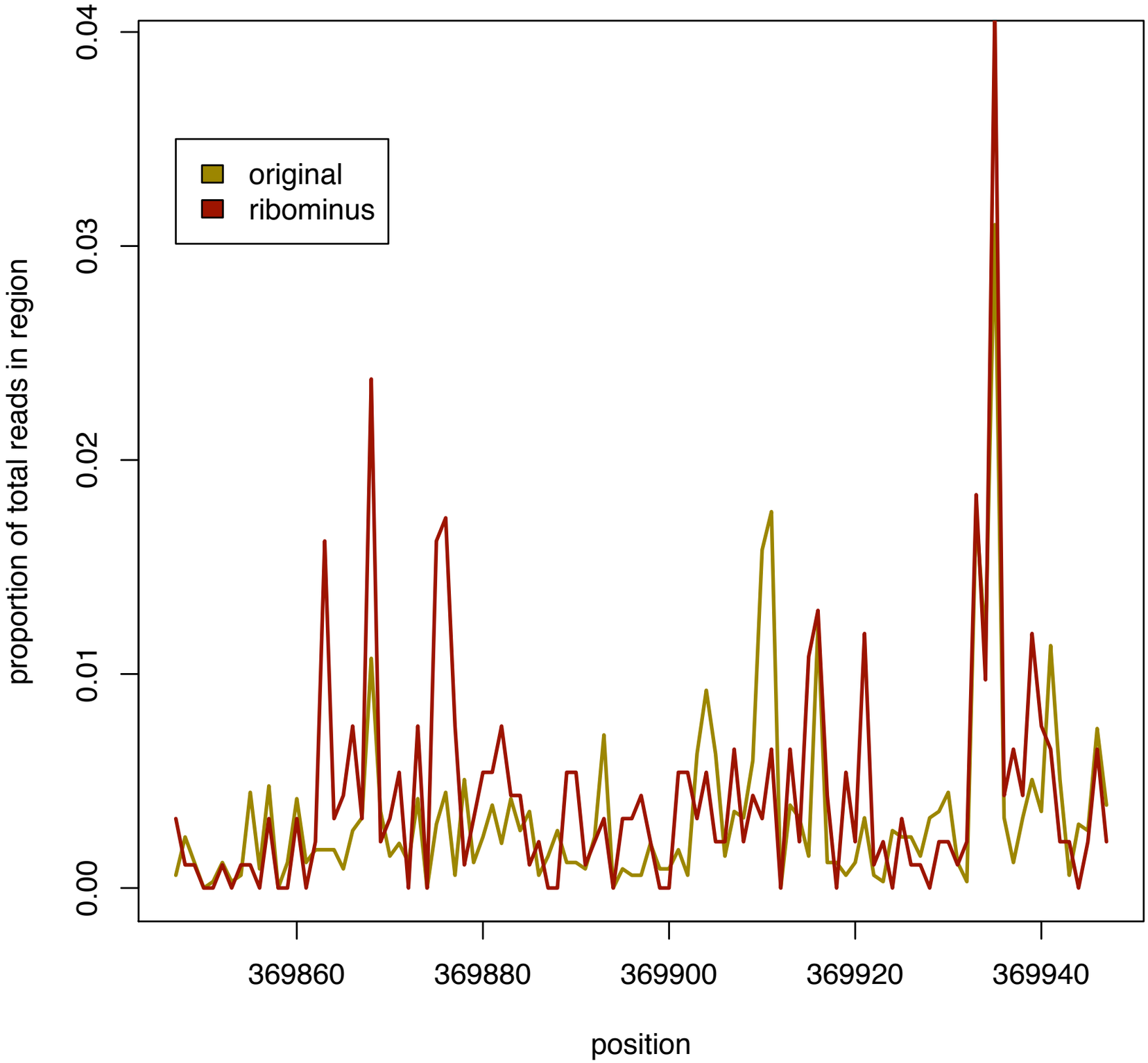
Base effect - multiple samples



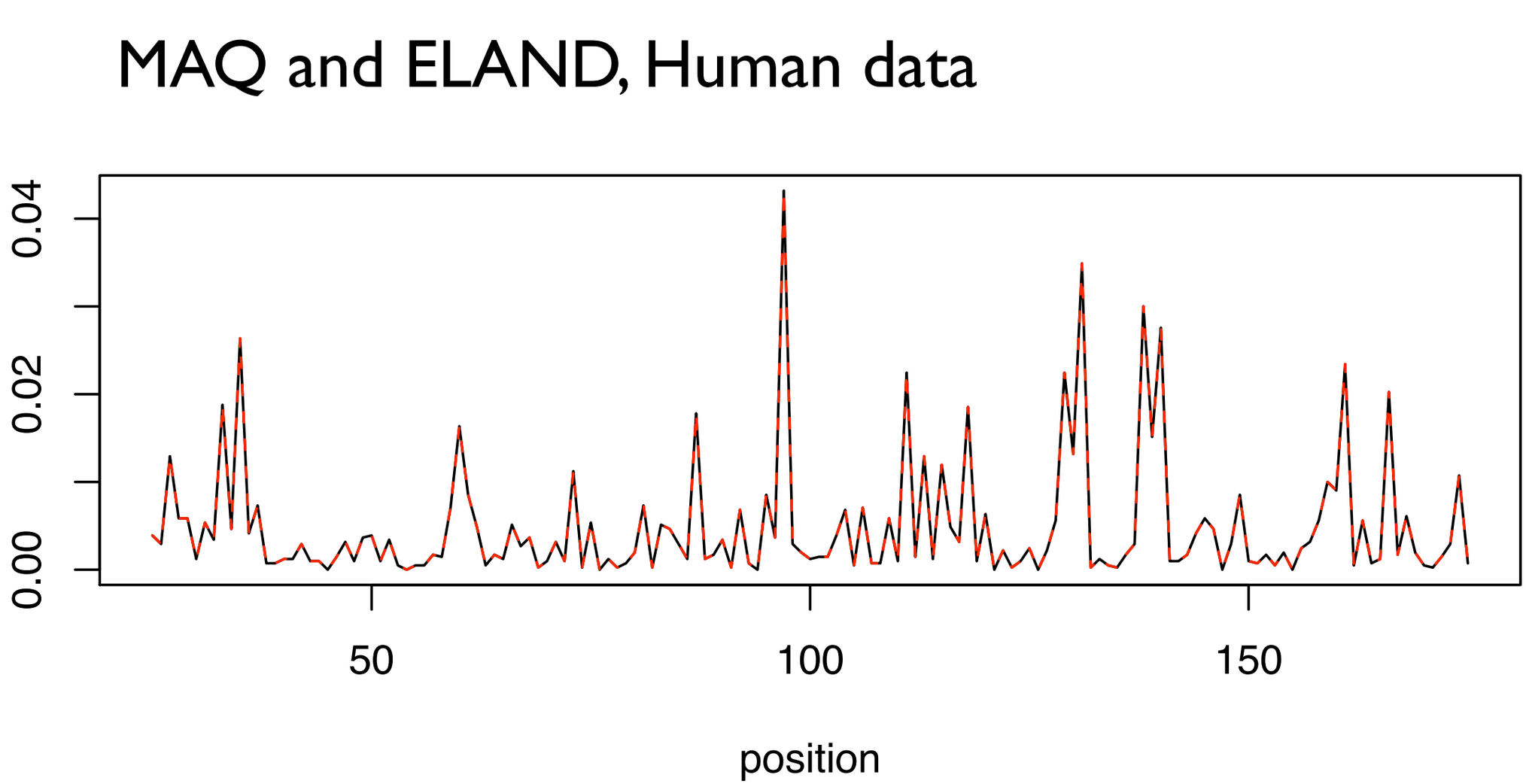
Base effect - different study (and prep)



Base effect - different prep



Base effect - different aligners



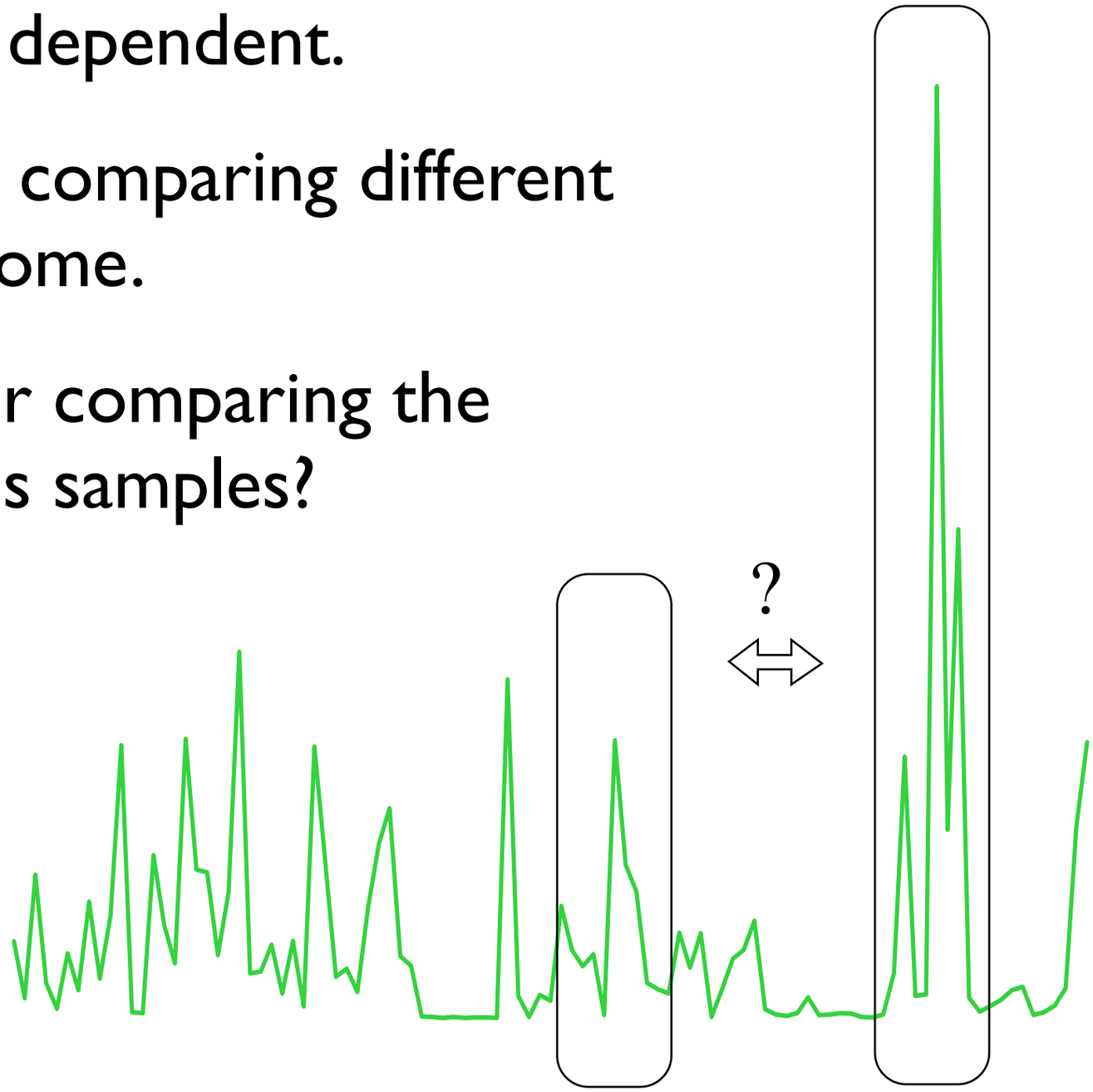
Base effect - conclusions

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

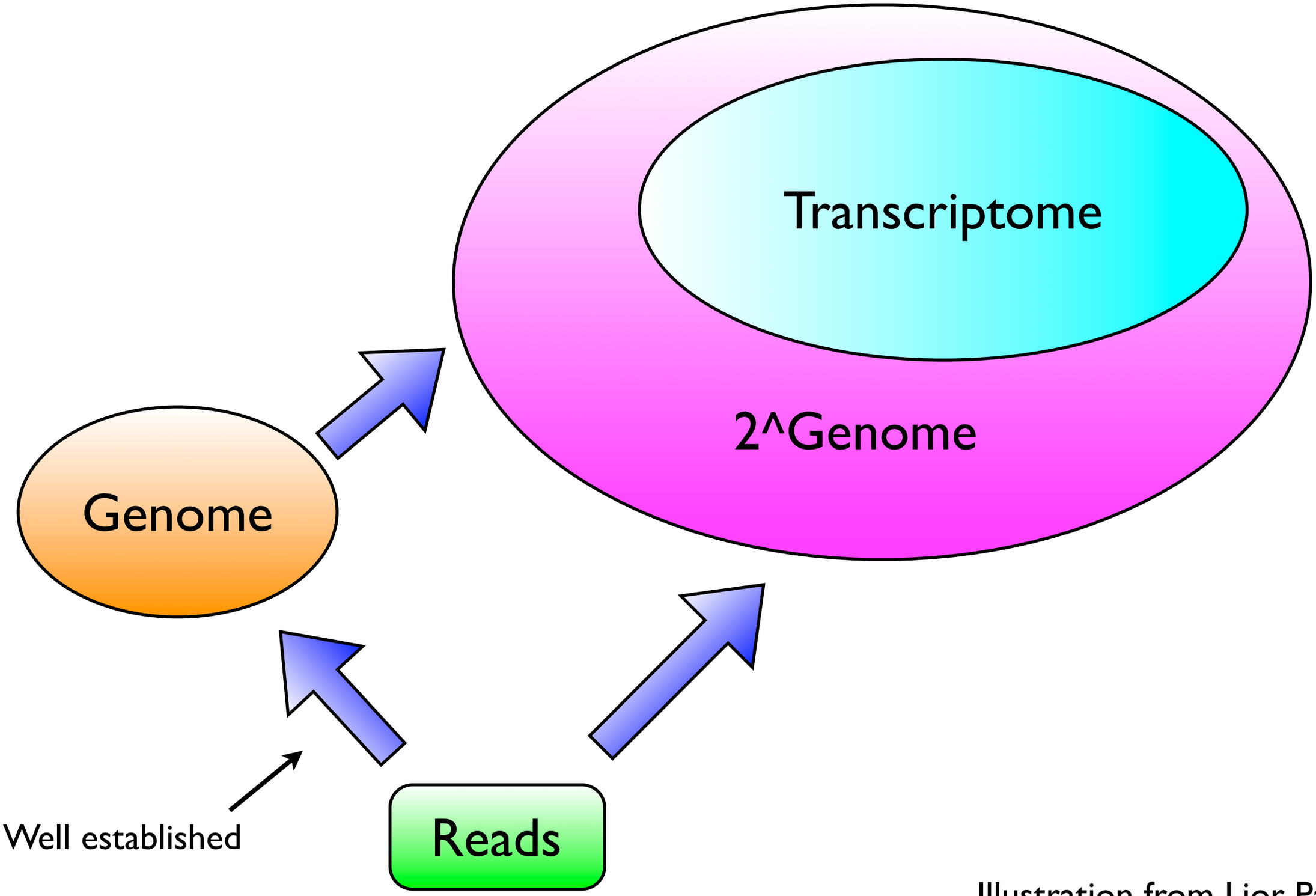
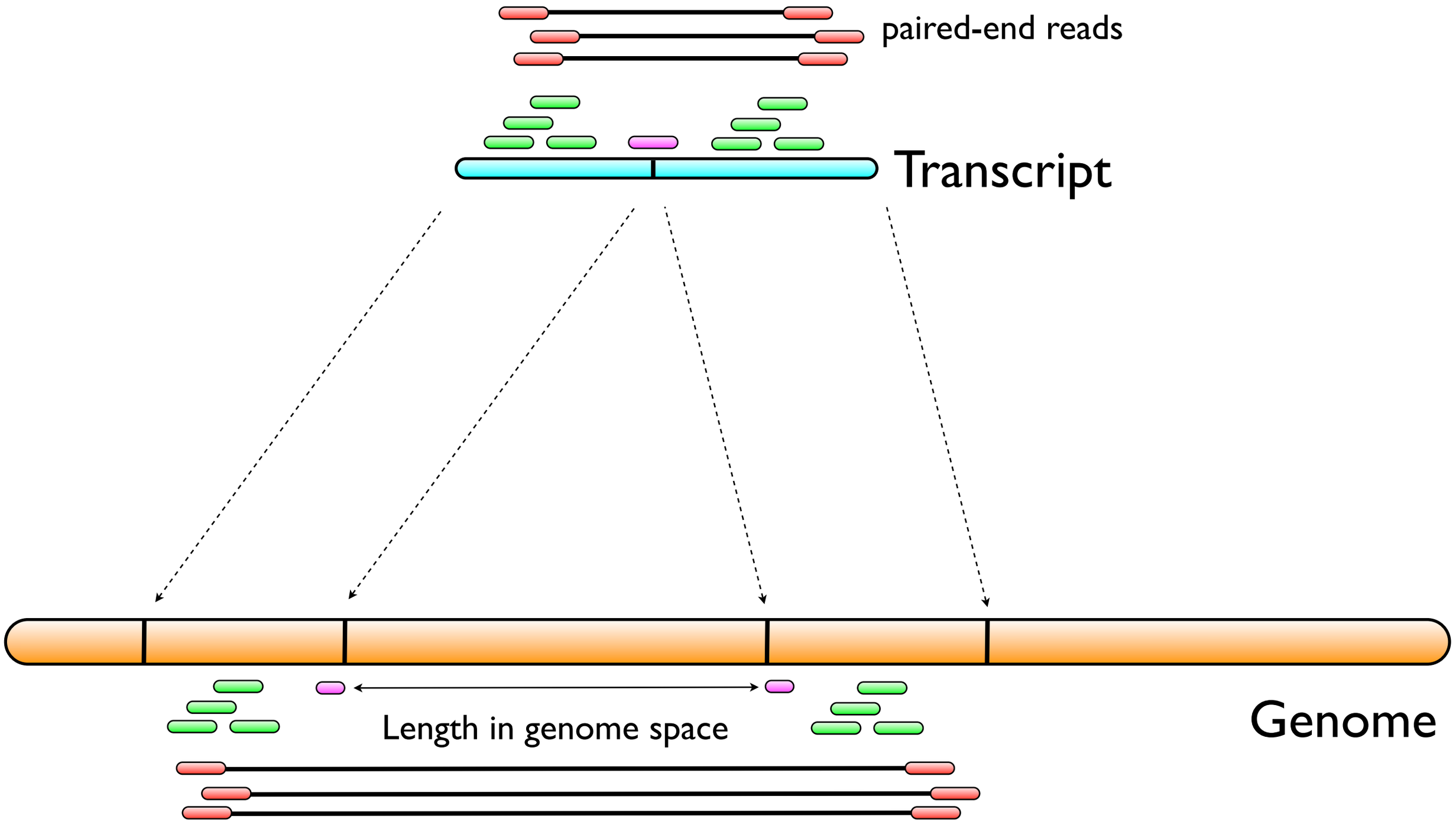


Illustration from Lior Patcher

Mapping transcripts



Junction reads

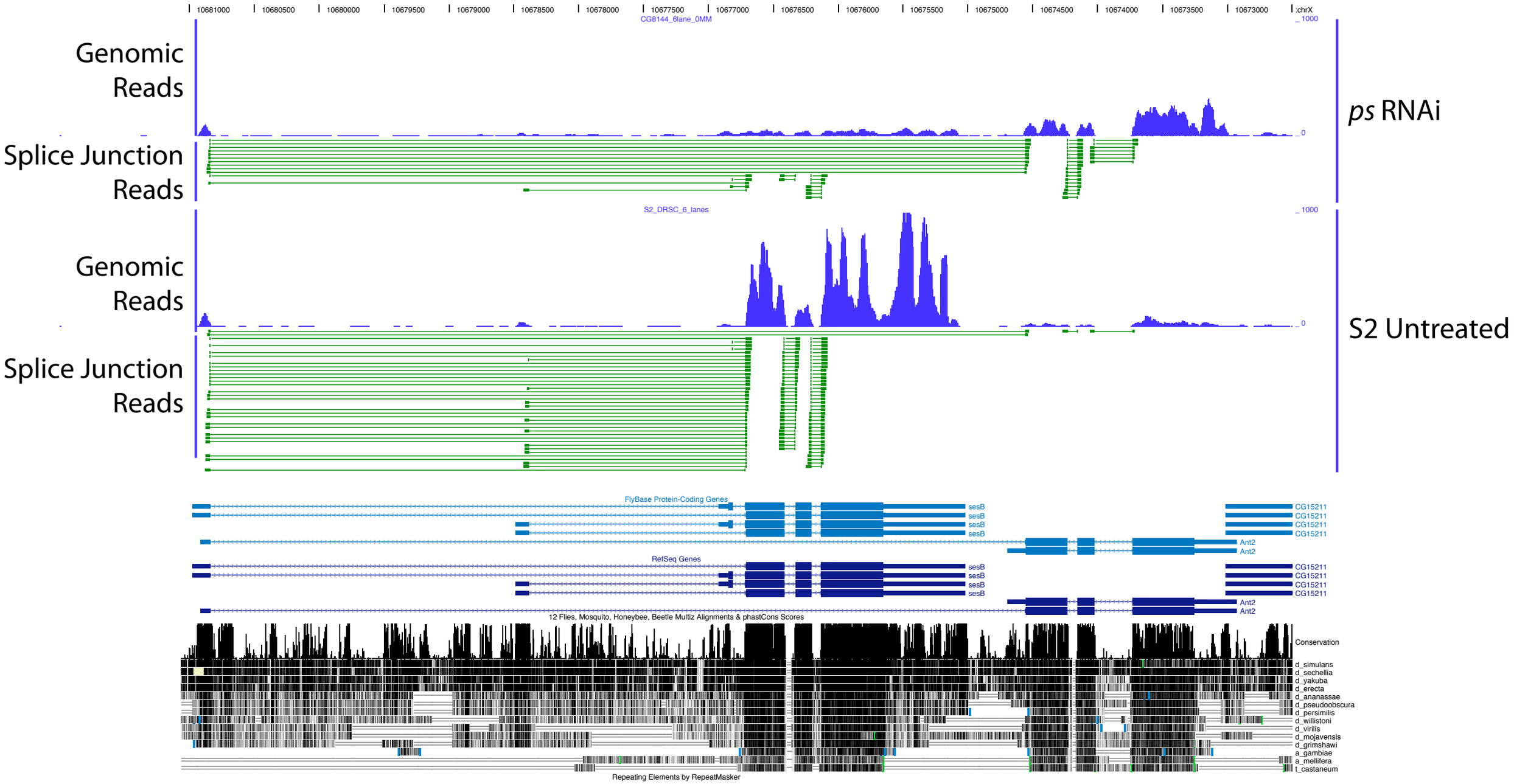


Image from Brenton Gravely

Junction reads, zoom

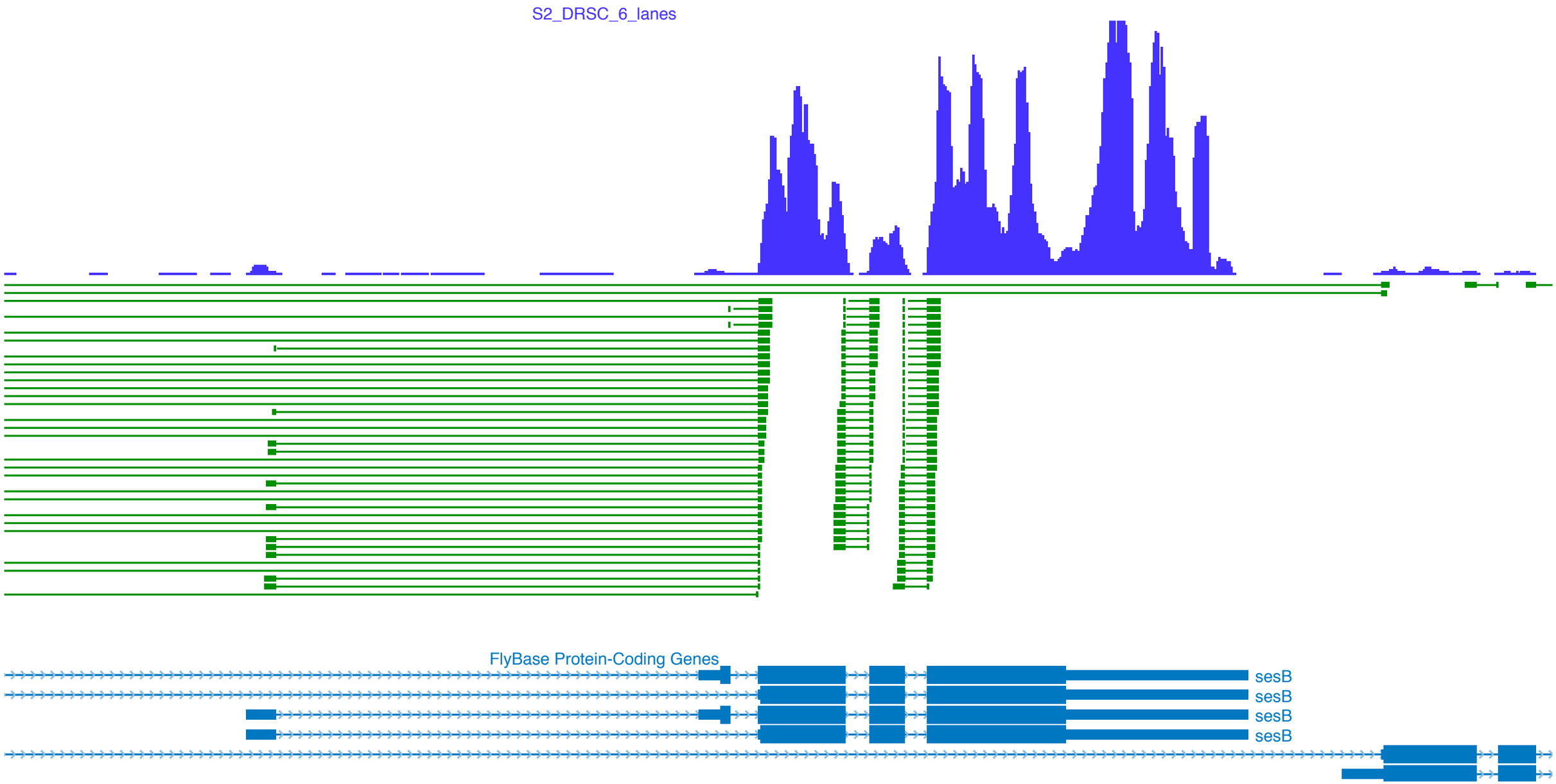


Image from Brenton Gravely

Strategies for mapping to junctions

Map to known junctions.

Map to combination of known exons.

Map completely de-novo using canonical acceptor and donor sites. The combinatorics makes this an intimidating approach.

Map de-novo, but constrain the search to canonical acceptor and donor sites between and in transcribed region: transcript assembly. This is the approach taken by TopHat.

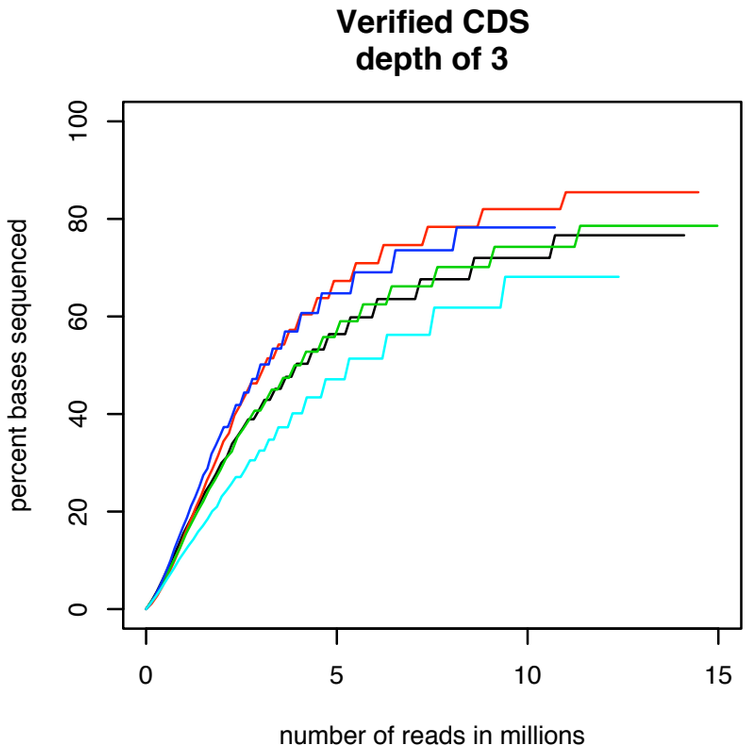
Paired-end data will make de-novo mapping a real possibility.

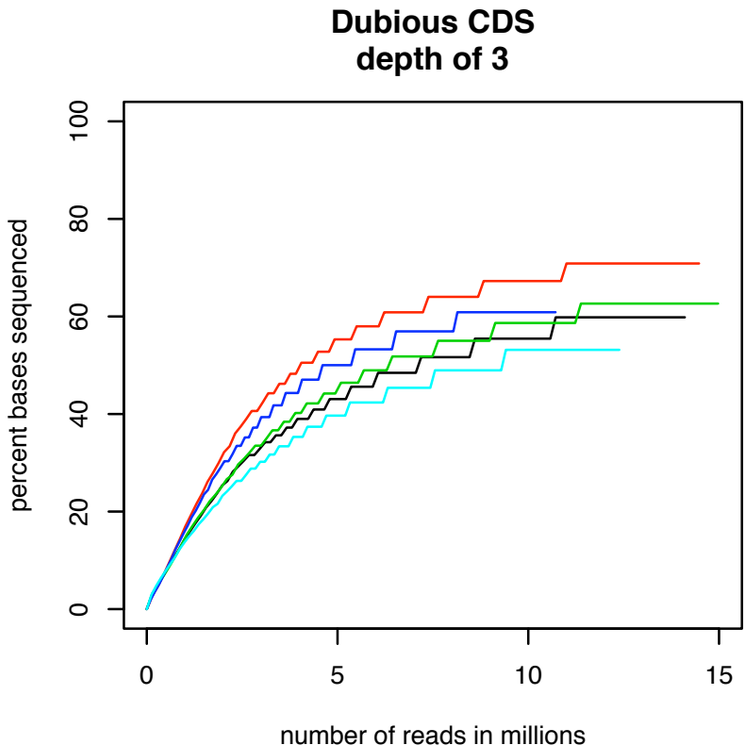
Mapping - conclusions

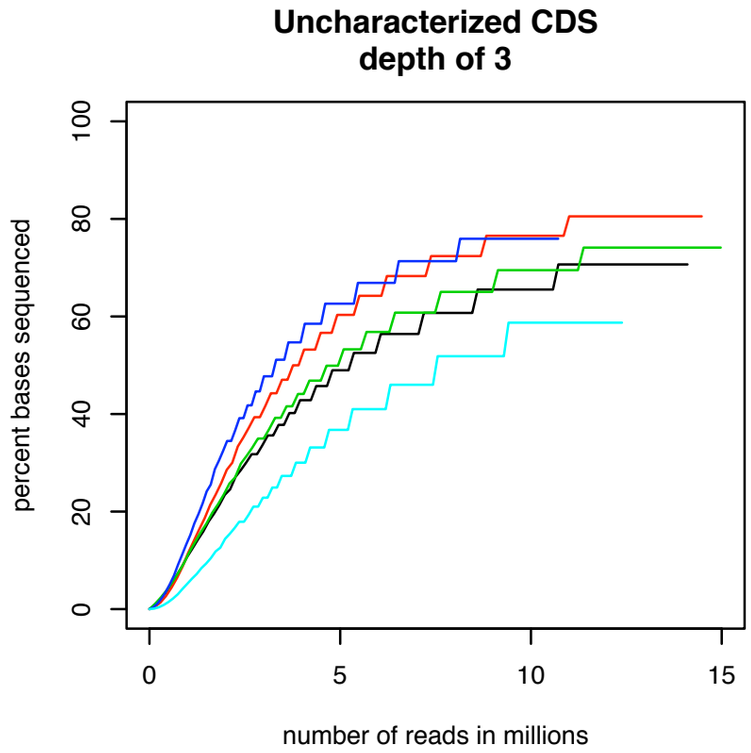
Mapping to transcript space is not easy.

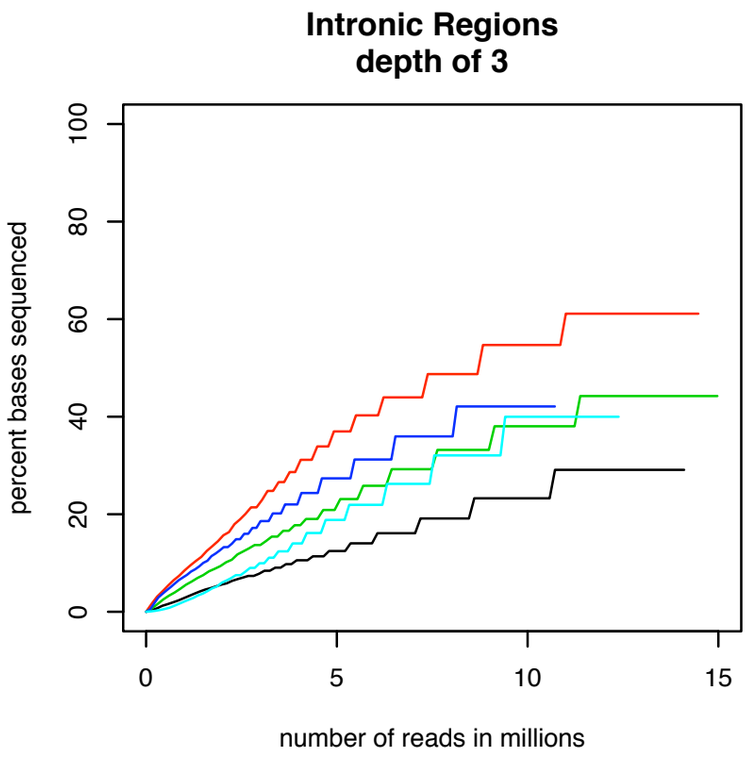
But essential for really understanding alternative splicing.

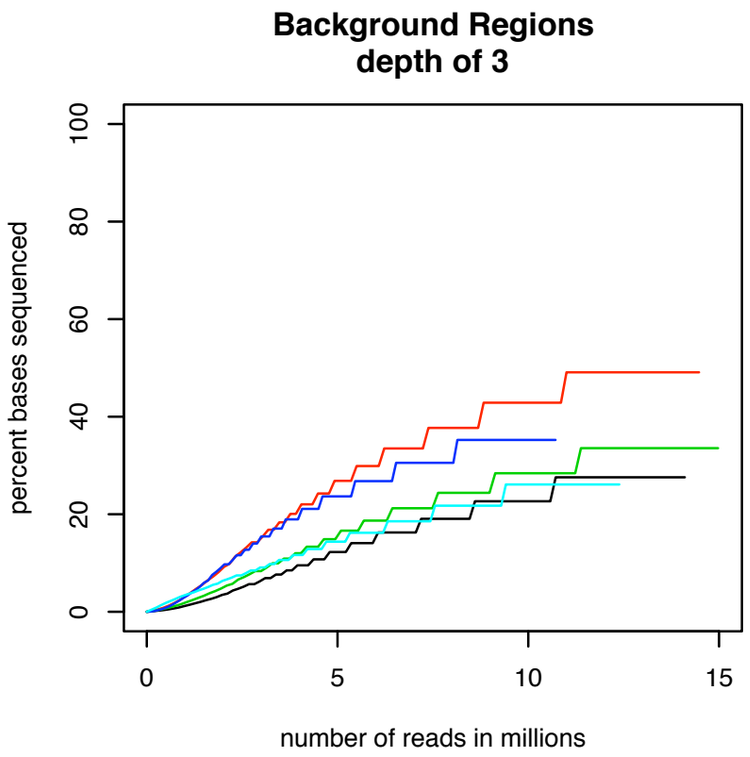
Constructing all novel splice junctions based on canonical splice sites but only accepting splicing within genes (and small regions upstream/downstream of the gene) in *D. Melanogaster* yields 605,000,000 splice junctions.

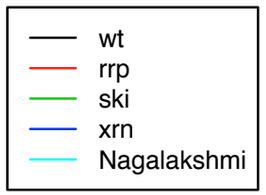

Verified CDS
depth of 3
percent bases sequenced
number of reads in millions
0
20
40
60
80
100
0
5
10
15


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

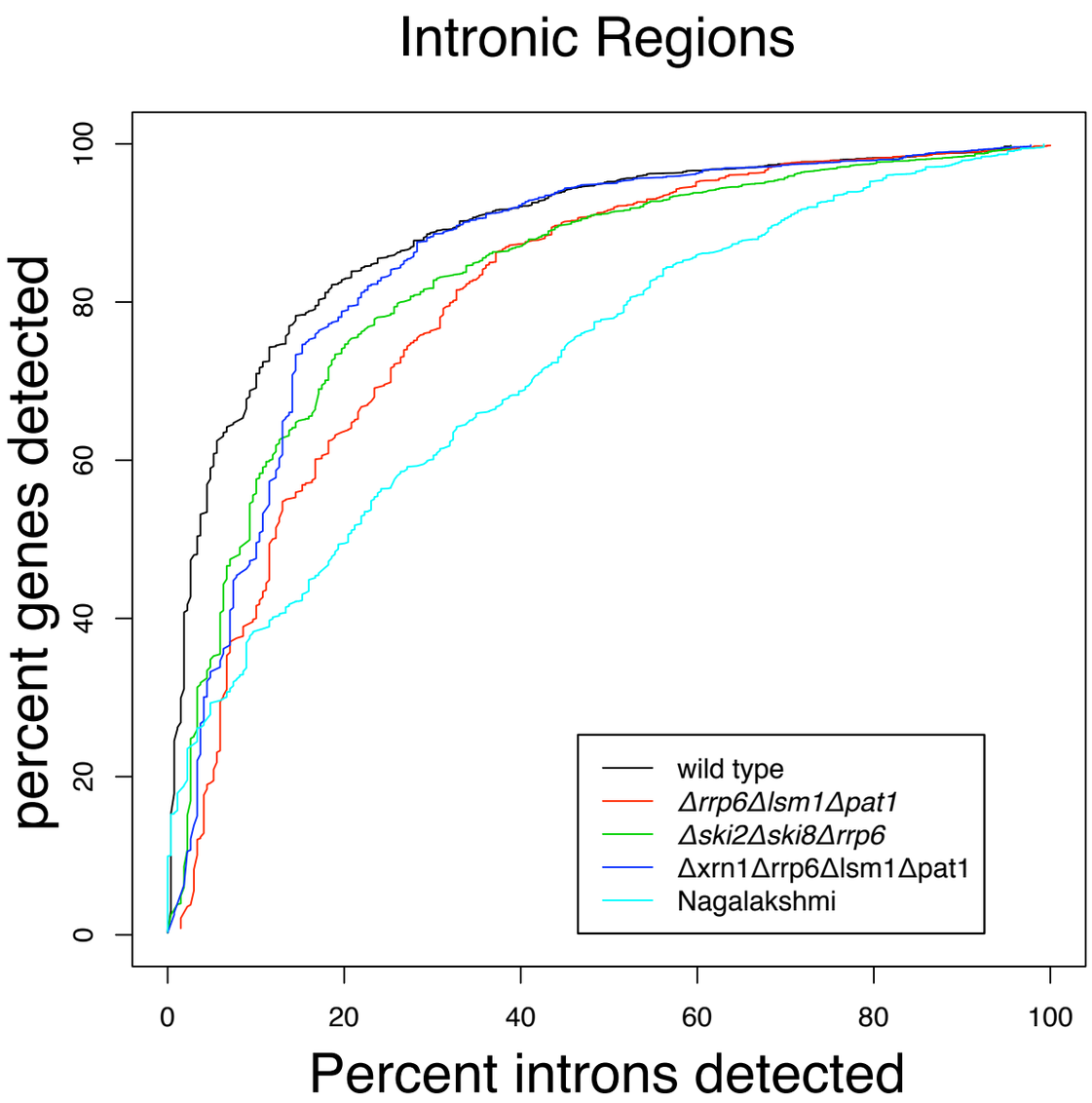

Uncharacterized CDS
depth of 3
percent bases sequenced
number of reads in millions
0
20
40
60
80
100
0
5
10
15

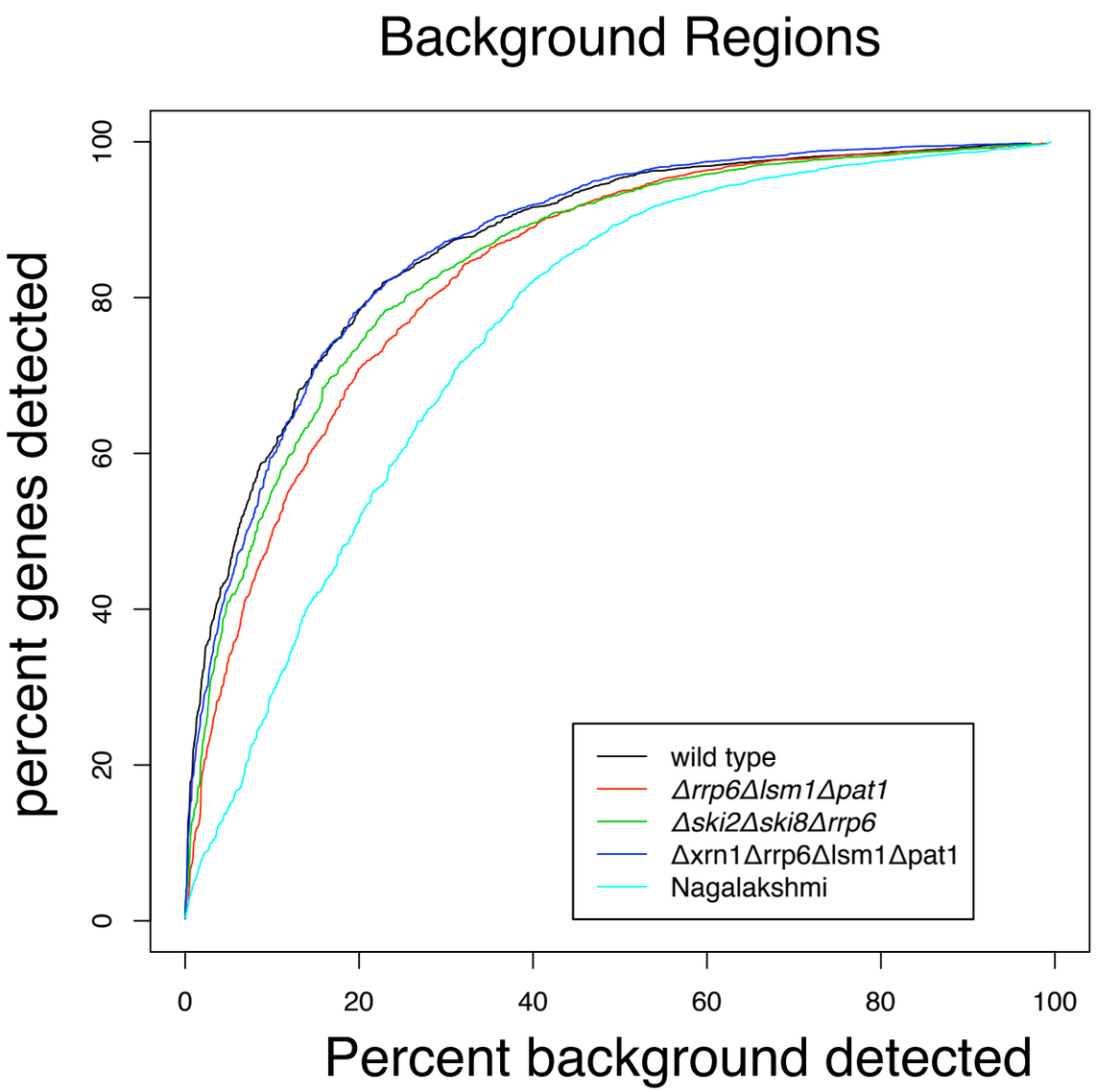

Intronic Regions
depth of 3
percent bases sequenced
number of reads in millions
0
20
40
60
80
100
0
5
10
15


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


wt
rrp
ski
xrn
Nagalakshmi

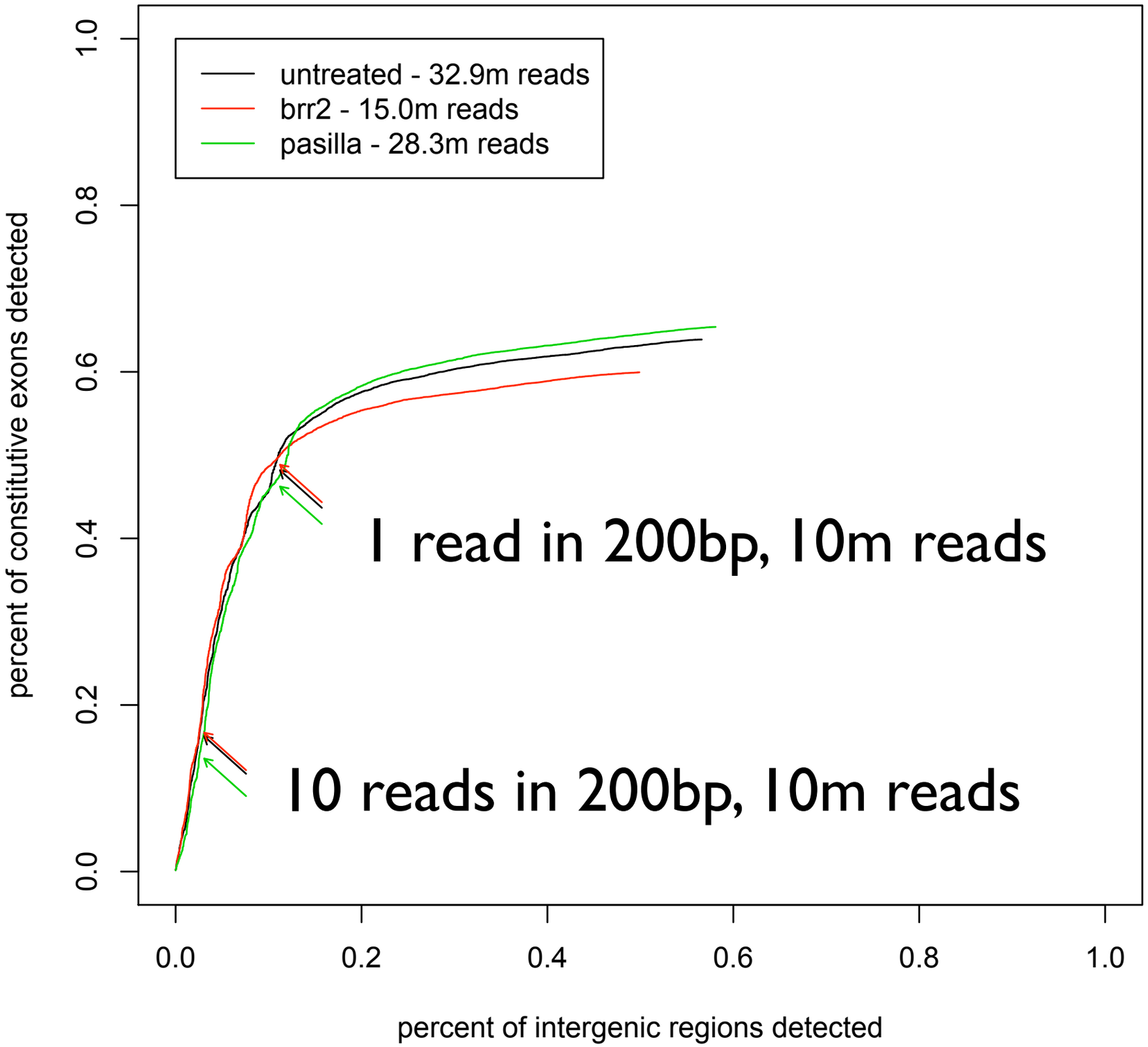
Detection in *Cerevisiae*





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

Detection in *Drosophila*



Replication

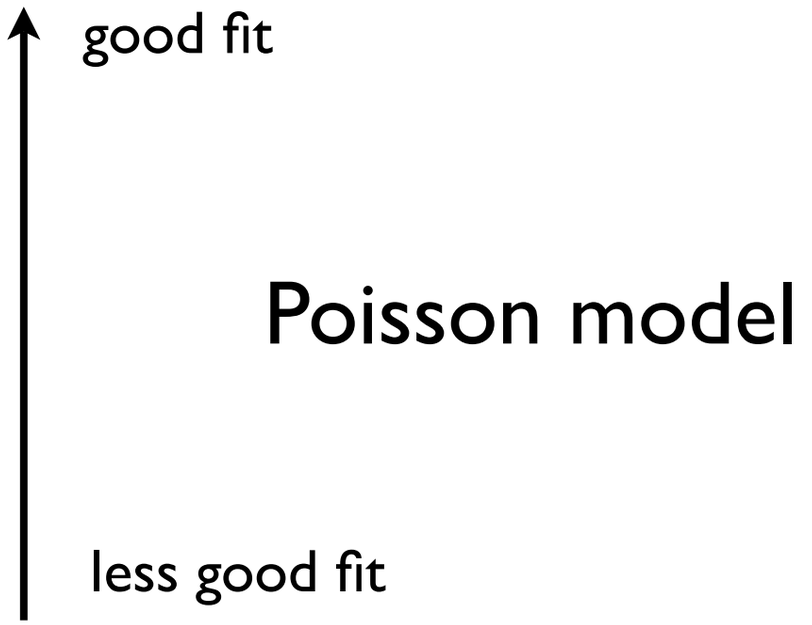
<u>Sources of variation</u>

Lane variation

Flowcell variation

Library prep variation

Biological variation



Systematic differences

?: Is absolute quantification possible

Software

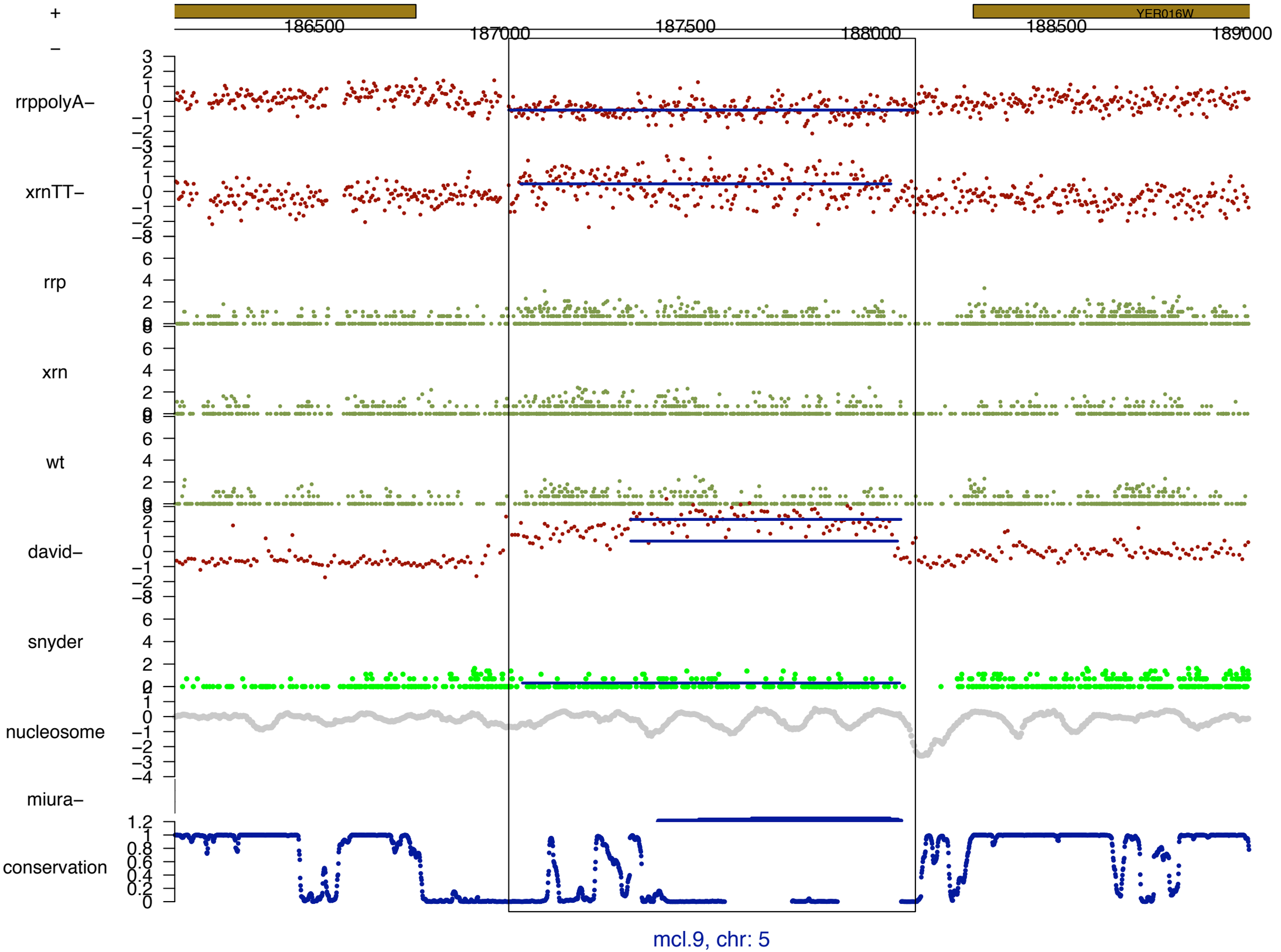
We have developed two **R** packages to help us

GenomeGraphs (Durinck, Bullard) : plots annotation and experimental data along a genome. Makes it easy to construct high quality images as well as to do data exploration. Available from Bioconductor.

Genominator (Bullard, Hansen) : provides support for managing, accessing and analyzing data oriented along a chromosome, together with annotation. Uses a SQLite backend. Works very well for unpaired reads mapped to the genome. Available from our home page.

R> summarizeByAnnotation(expData, annoData, fx)

Genome Graphs, example



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