

# Using Next Generation Sequence Data to Analyze Cancer Genomes

İnanç Birol

NRC-PBI Workshop – 21 January 2011

# BC Genome Sciences Centre

- Part of BC Cancer Agency

	Lifetime probability of			
	Developing cancer		Dying of cancer	
	%	One in	%	One in
Females	39	2.5	24	4.2
Males	45	2.2	29	3.5

*Canadian Cancer Society Statistics, 2008*



# High Throughput Sequencing

## Sequencing human genome

- Human genome project:  
1990 – 2003
- 18 Gb data:  
~ 6-fold coverage
- Cost:  
~ \$3G
- 2<sup>nd</sup> Generation Sequencers:  
every week
- 100 Gb data:  
~ 33-fold coverage
- Cost:  
~ \$10k



# Why sequence cancer genomes and transcriptomes?

- To identify cancer driver mutations and pathways
- To identify targets for development of new therapies
- To improve diagnostic precision and prognostic accuracy
  - e.g. “breast cancer” describes several diseases
- To match patients to treatments
  - Optimize treatment modalities based on individual genes and genomes
- To understand differences in treatment response
  - Outright failure
  - Remission / relapse (treatment resistance)
- To manage treatment failure
  - Alternative existing therapies



# Current NGS Platforms at GSC



- 8 GA IIx



- 9 HiSeq 2000

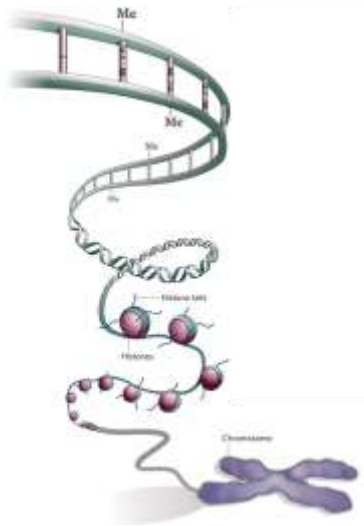


- 11 SOLiD 4.0



# Sequencing Library Types

Highly versatile platforms



Whole Genome Shotgun

Exome

Amplicon

Targeted capture

ChIP

MeDIP

MRE

Bisulfite

FAIRE

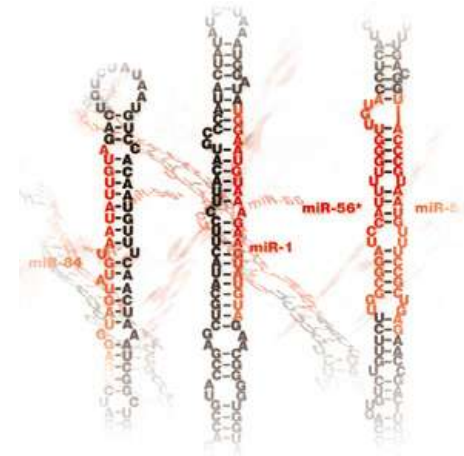
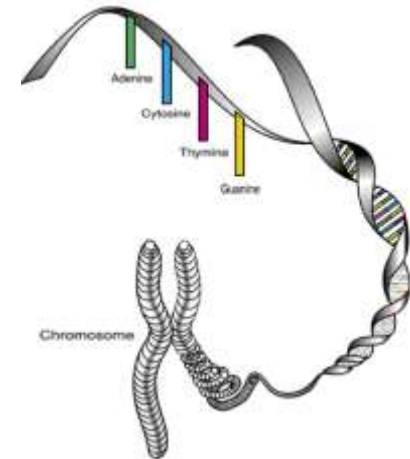
Whole transcriptome

PolyA RNA

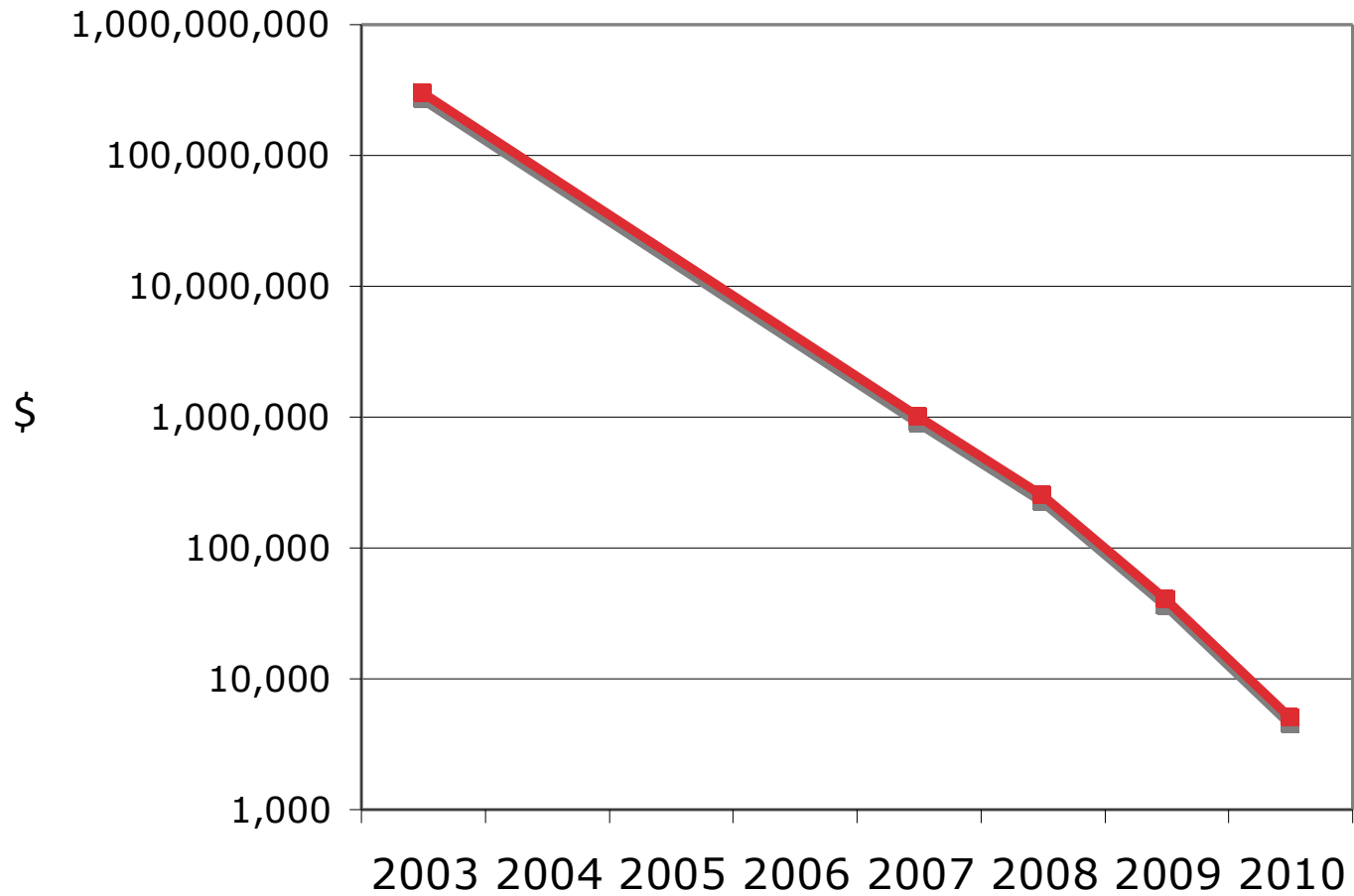
miRNA

Small RNA

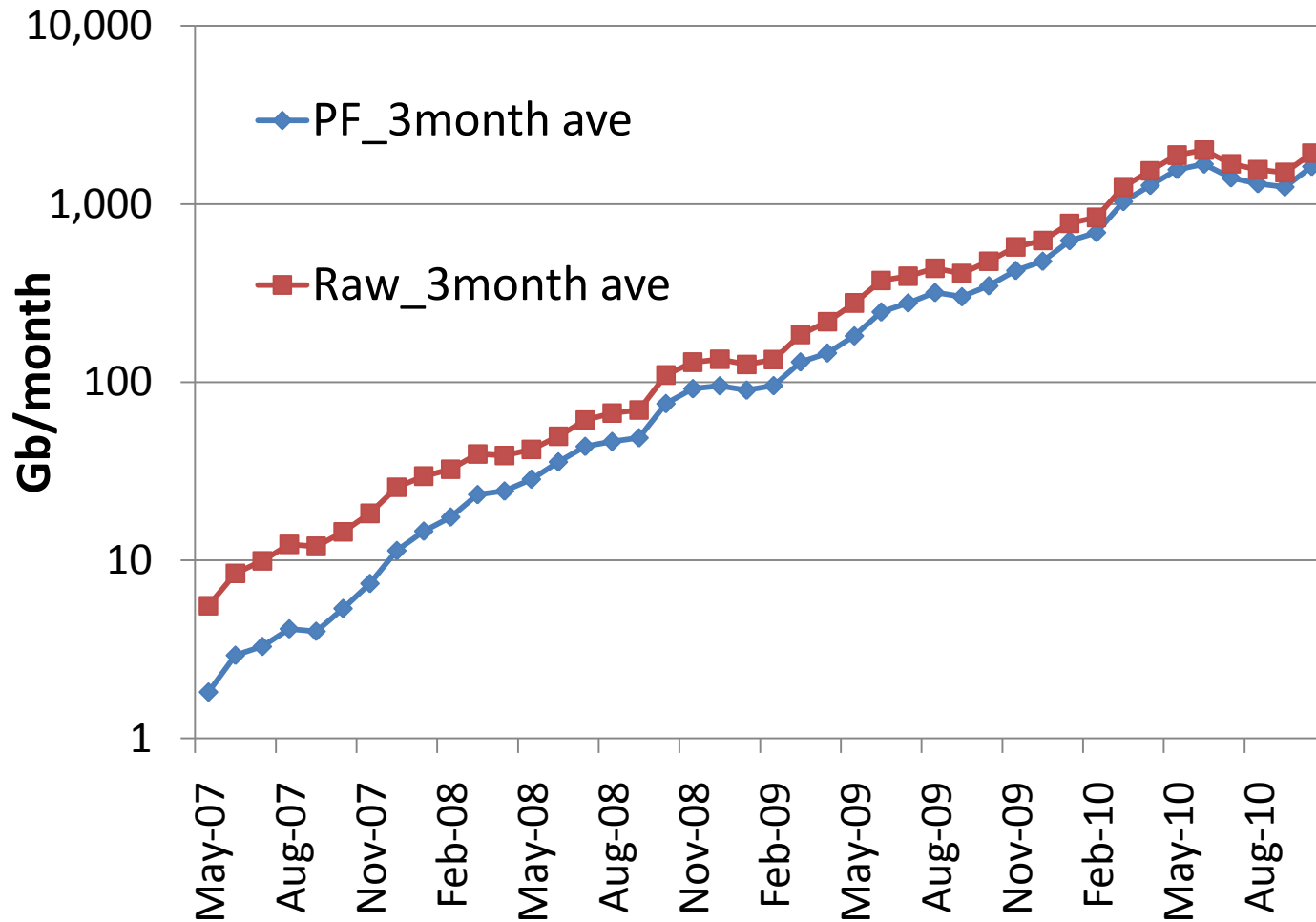
SAGE...



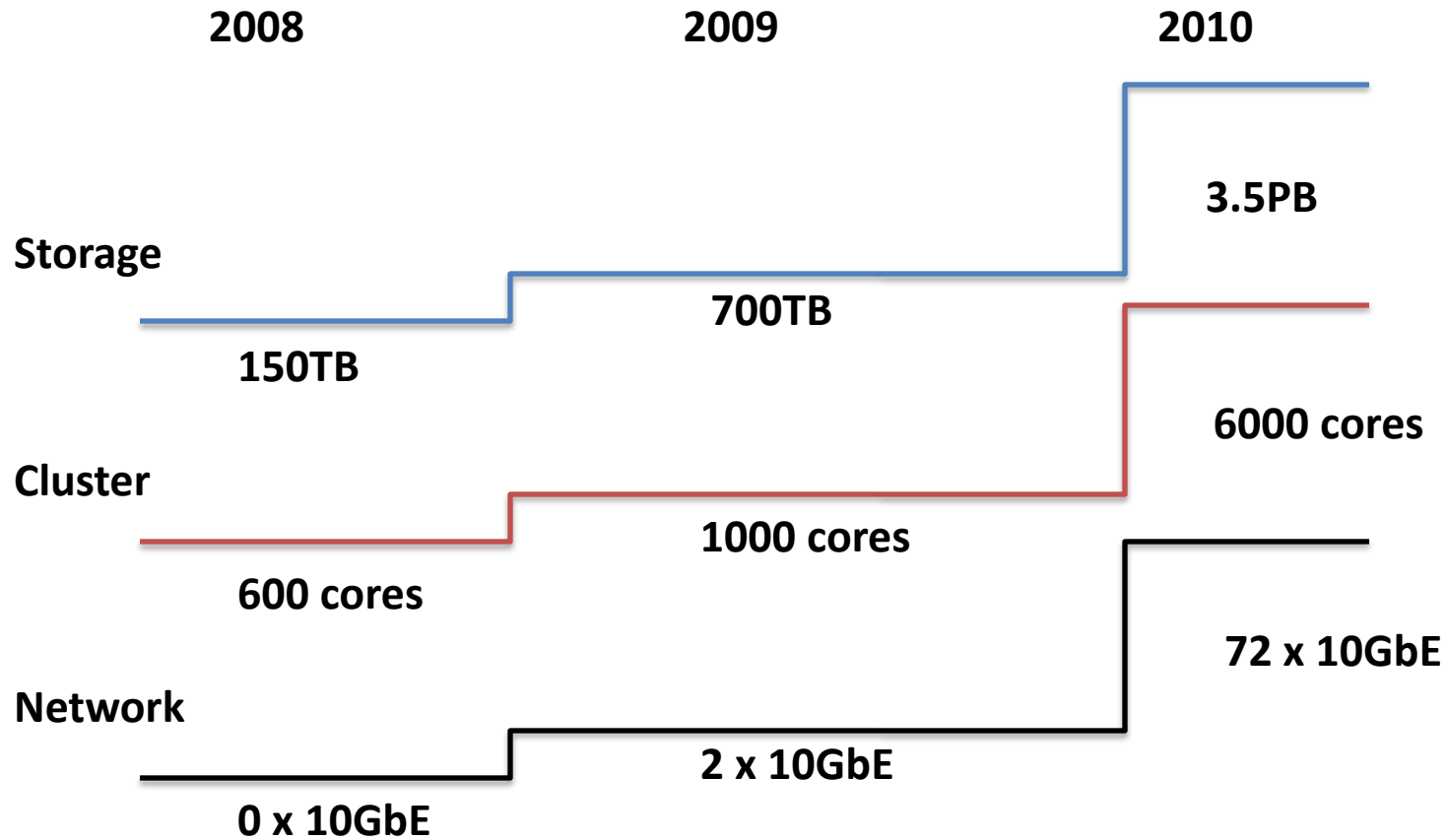
# Cost of Sequencing Human Genome



# Illumina Throughput



# Compute Resources



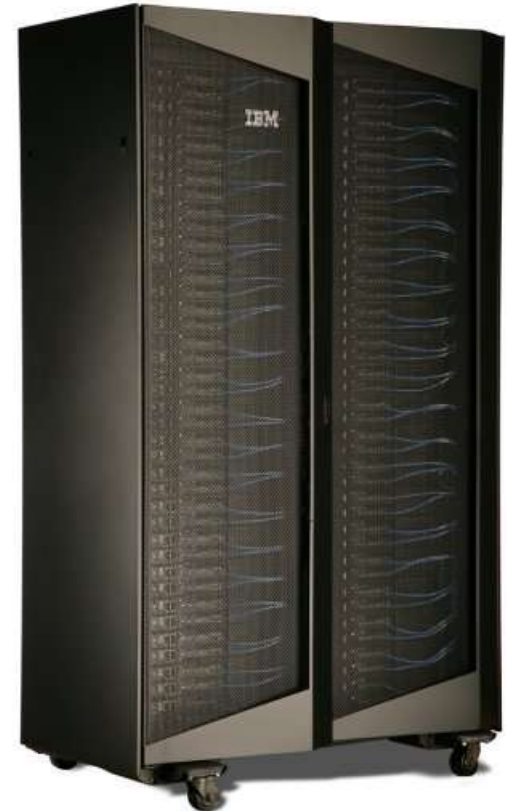
# Archive Storage

- OnTap 8.0C Cluster
  - 12 NetApp 6080 storage heads
- 1.2PB of usable storage
- 12GB/s read and 6GB/s write throughput

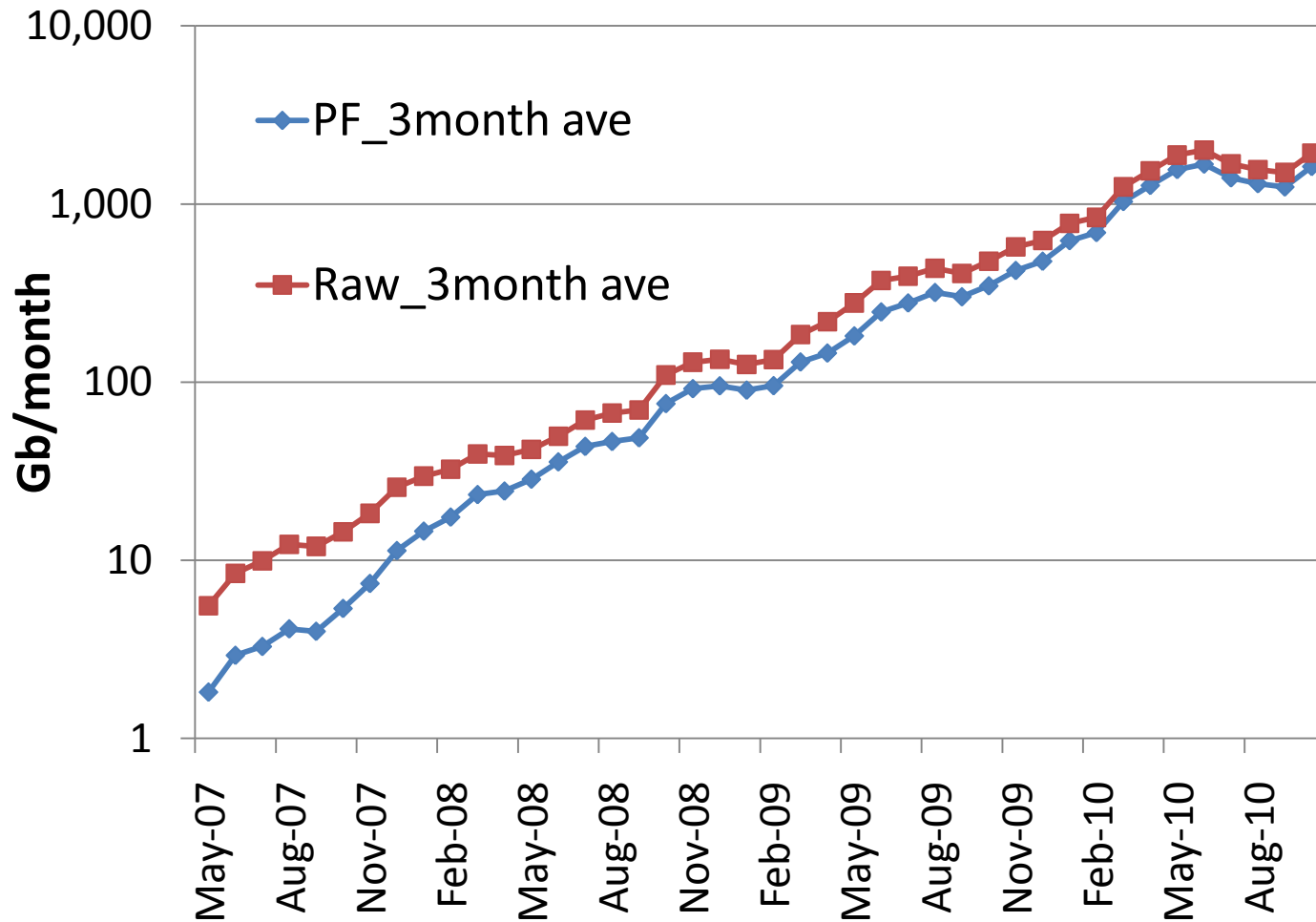


# Cluster Upgrade

- 420 compute nodes
- Intel Xeon X5650 2.66GHz (hyperthreaded) Processors
- 2 sockets/node
- 6 cores/socket
- 24 threads / node
- 48GB / node
- #178 in Top 500 Cluster list  
[www.top500.org](http://www.top500.org)
- # 19 in Top 500 Green Cluster list  
[www.green500.org](http://www.green500.org)



# Illumina Throughput



# Implications

$$\frac{d \left( \begin{array}{c} \text{data} \\ \text{generation} \end{array} - \begin{array}{c} \text{data} \\ \text{processing} \end{array} \right)}{dt} > 0^*$$

Operation Models

FIFO  
FIFO  
FISH

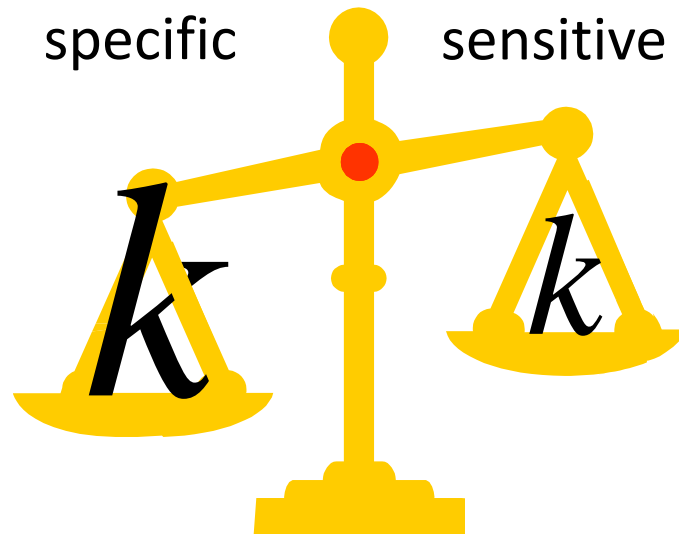


\* Unless bioinformaticians do something about it!



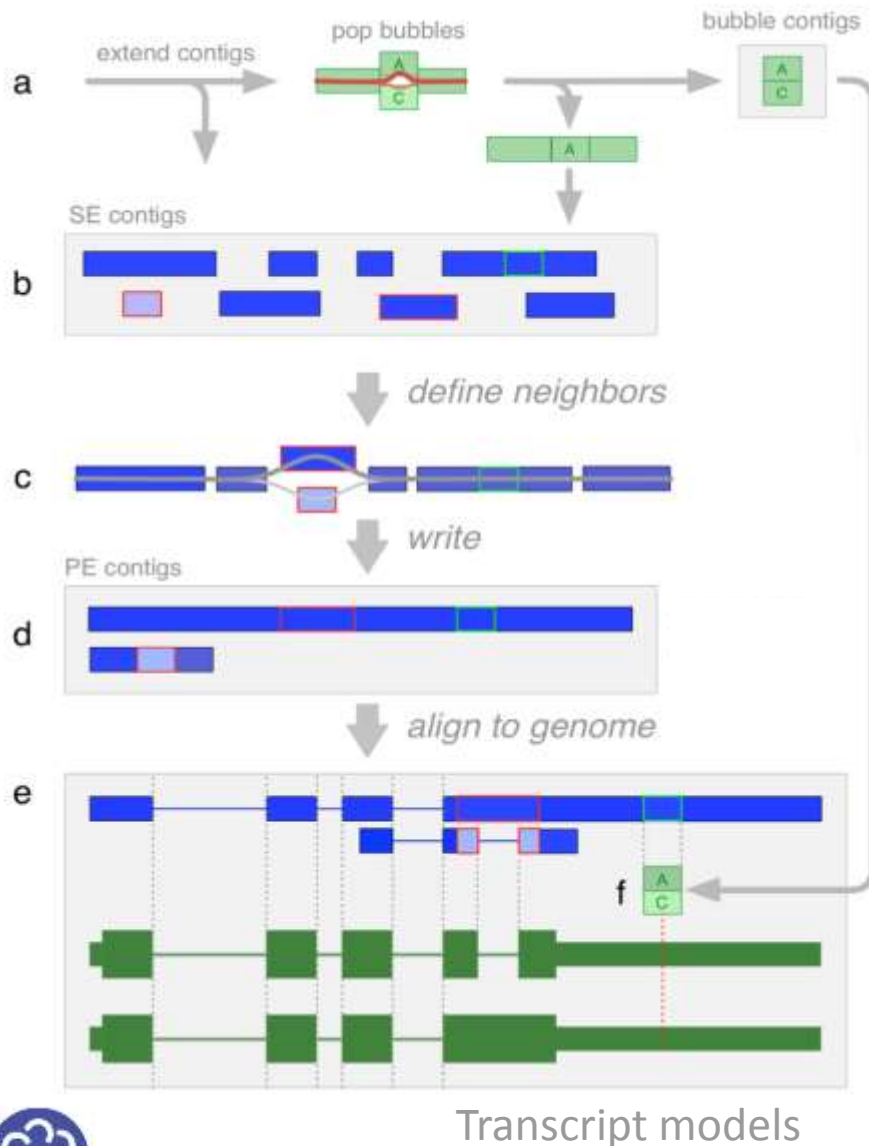
# ABySS

- Assembly by Short Sequences
  - de Bruijn graph SE assembly (a.k.a.  $k$ -mer extension)
  - followed by PE contig merging and scaffolding





# Transcriptome Assembly

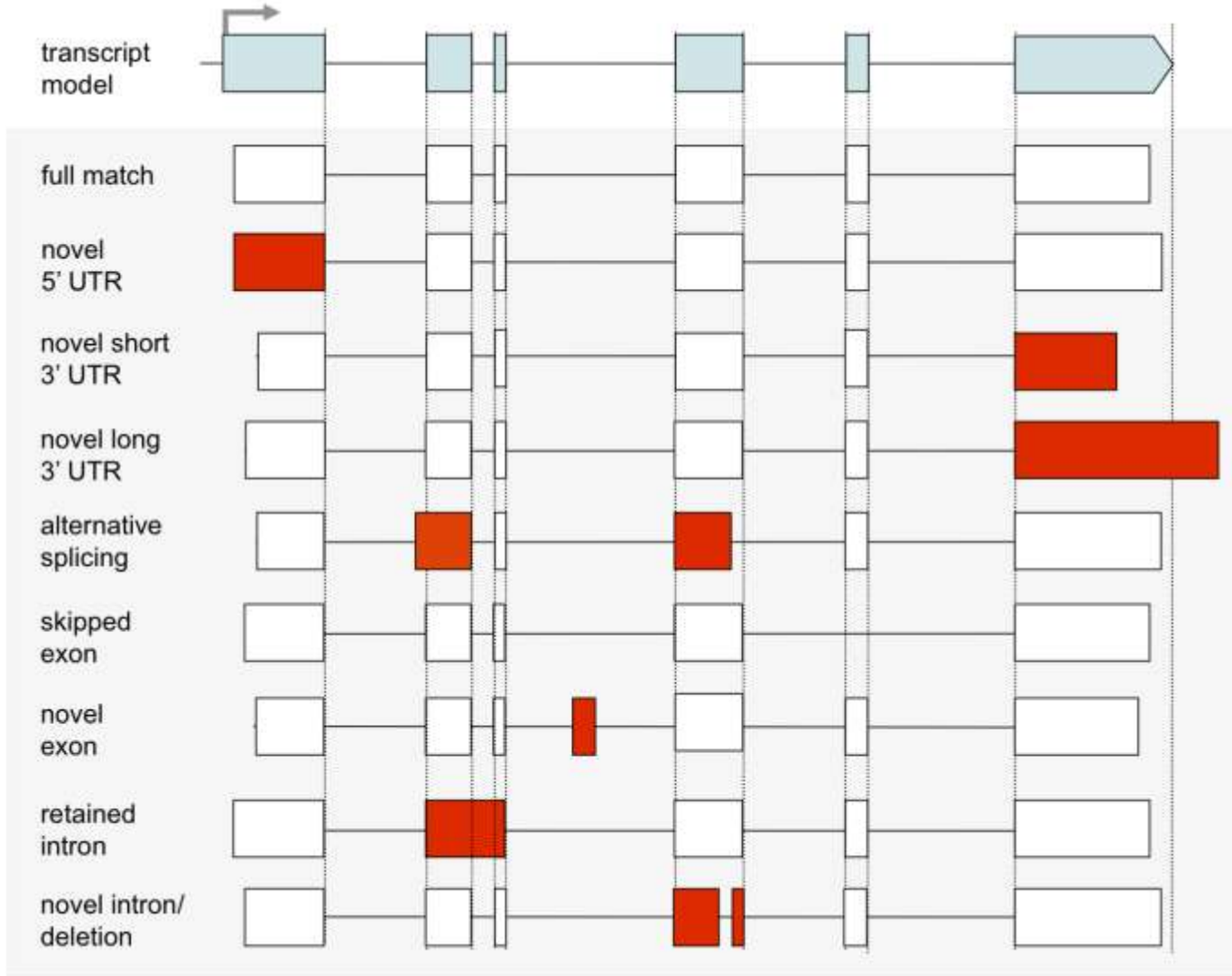


Transcriptome assembly is different from genome assembly

- varying coverage levels  
⇒ varying expression levels
- split assembly paths  
⇒ isoforms/splice variants
- small contig sizes  
⇒ small product sizes



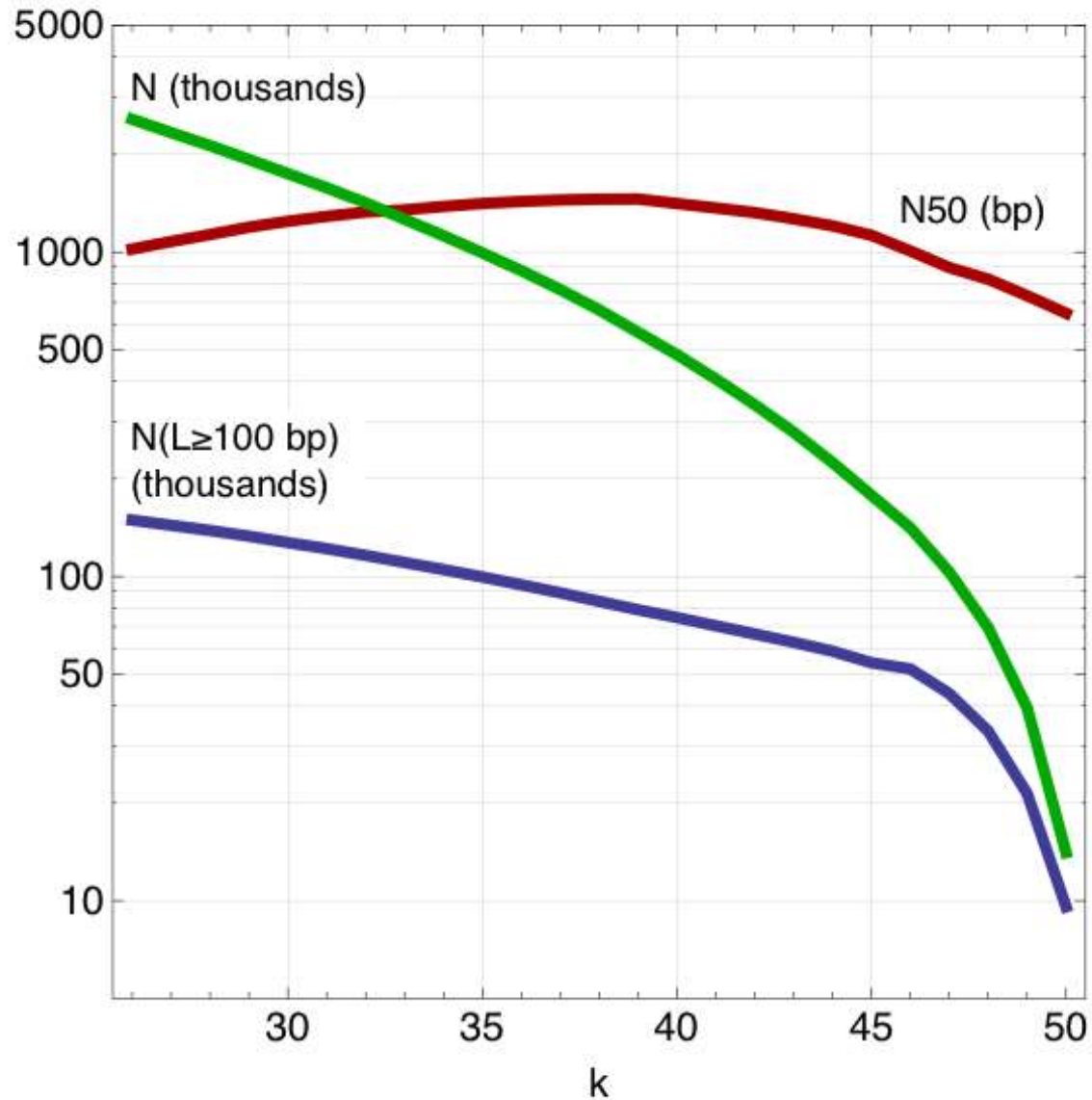
# Events



+ chimeric transcripts

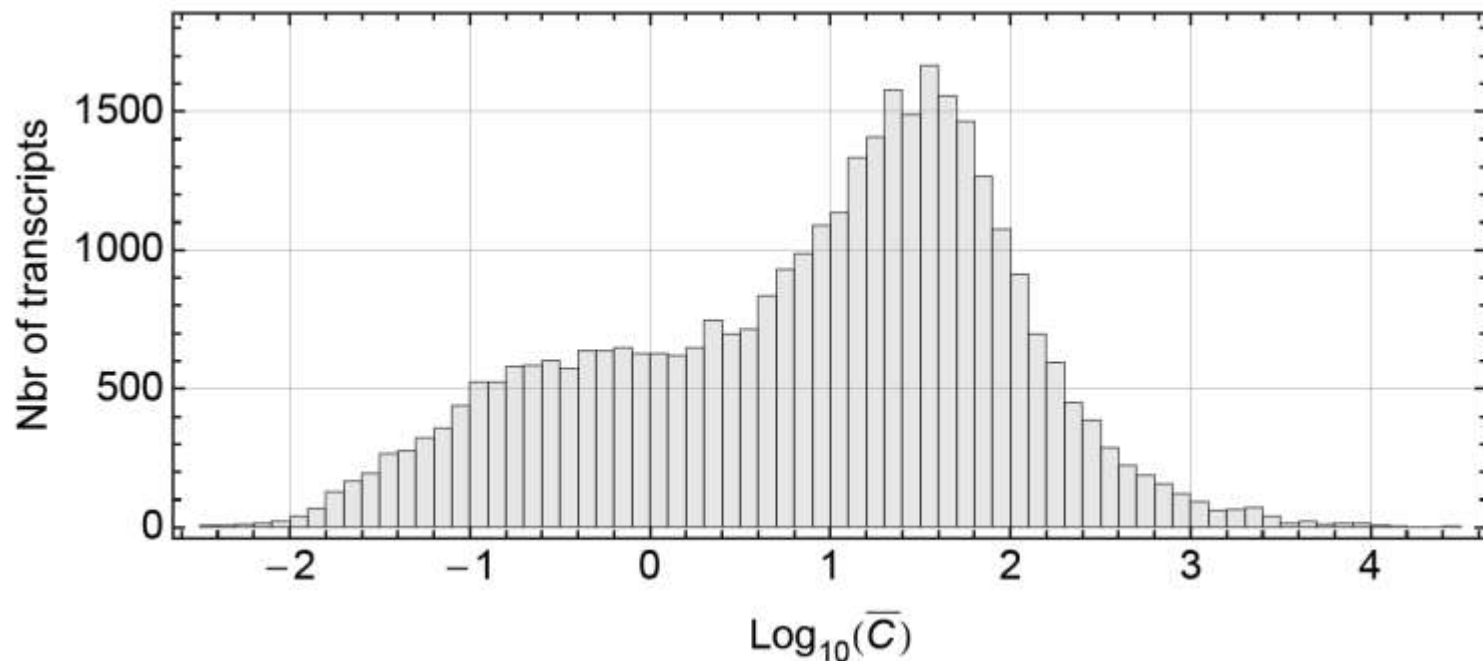


# What Overlap to Choose?

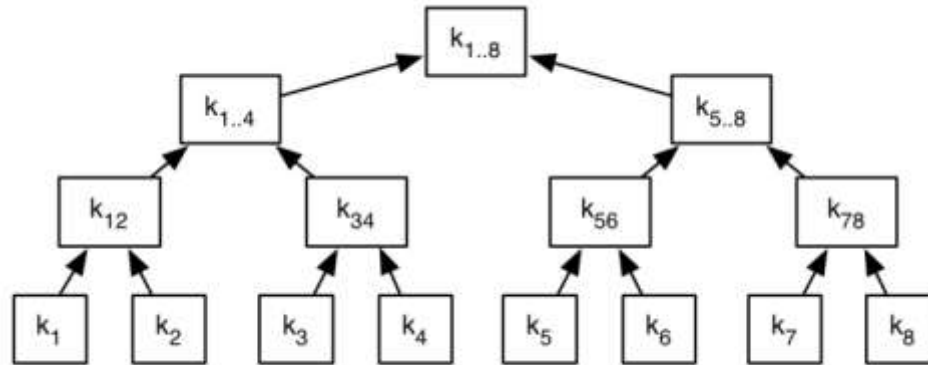


# Multi- $k$ for Varying Expression Levels

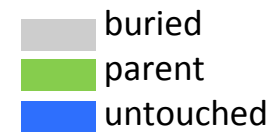
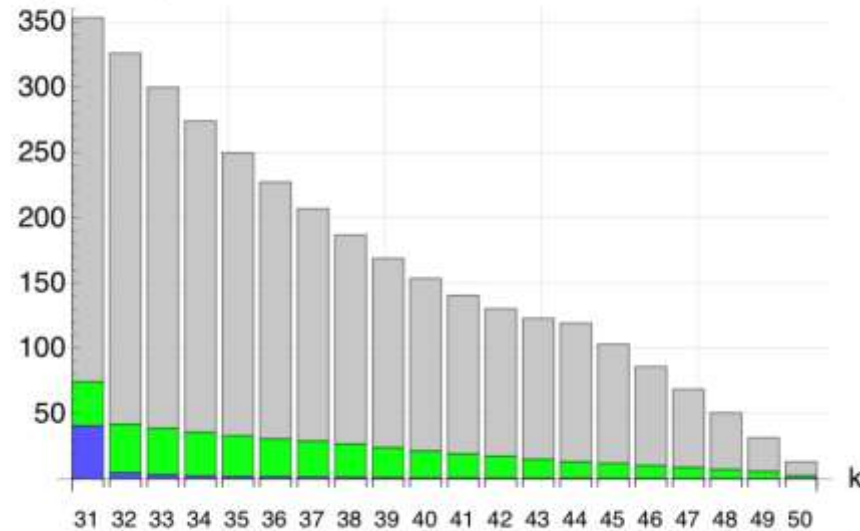
- Selection of parameter  $k$  depends on read coverage depth
- Expression levels vary over 5 orders of magnitude



# Assembly Merging



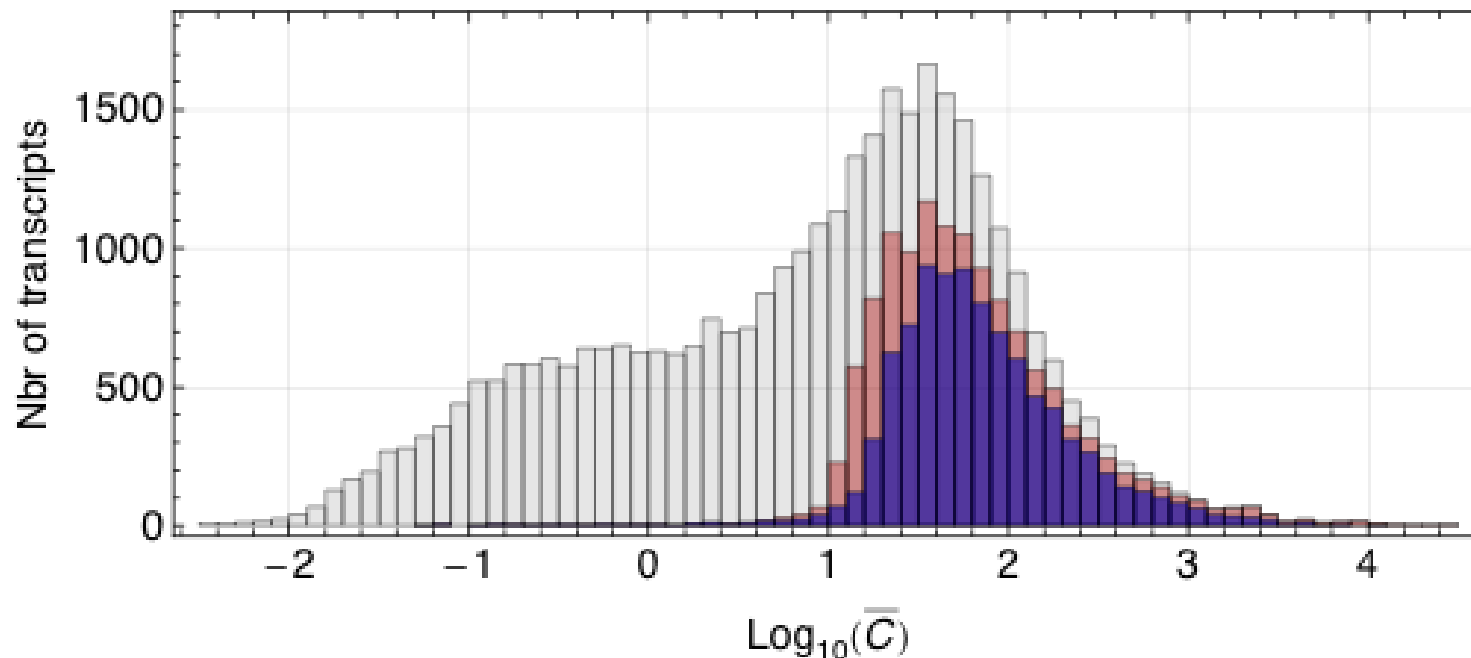
Contigs (1000s)



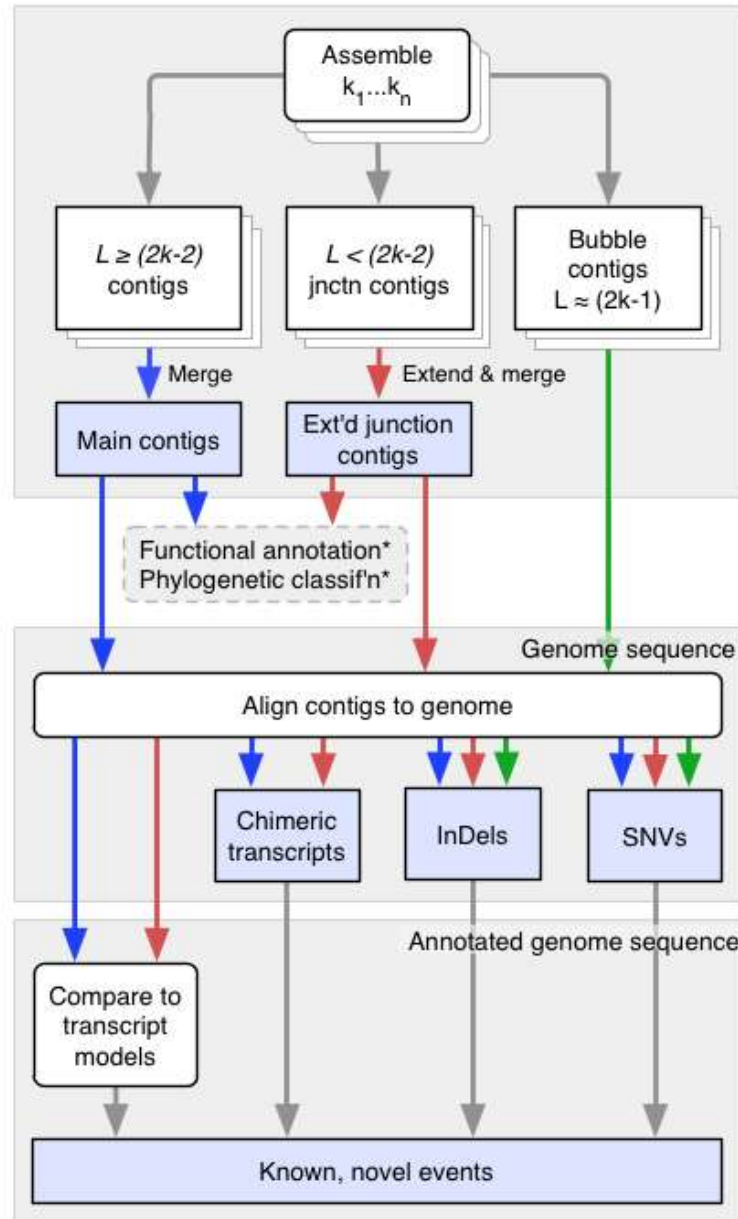
# Multi- $k$ Assembly

We capture a wide range of expression levels

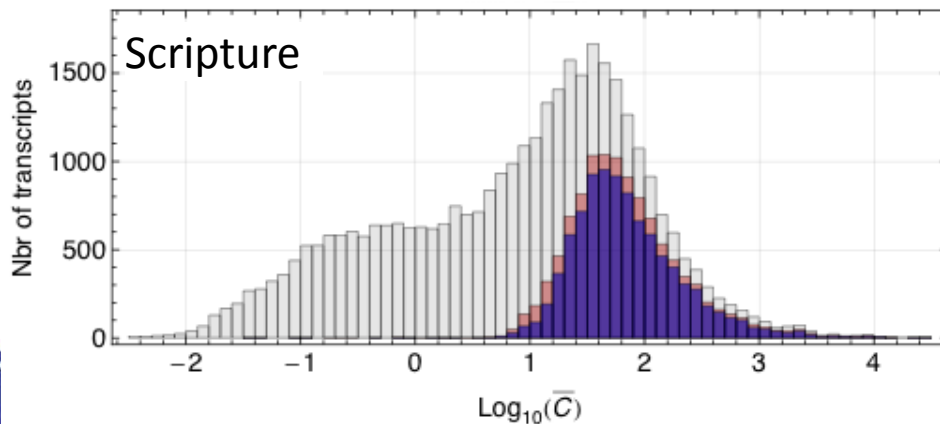
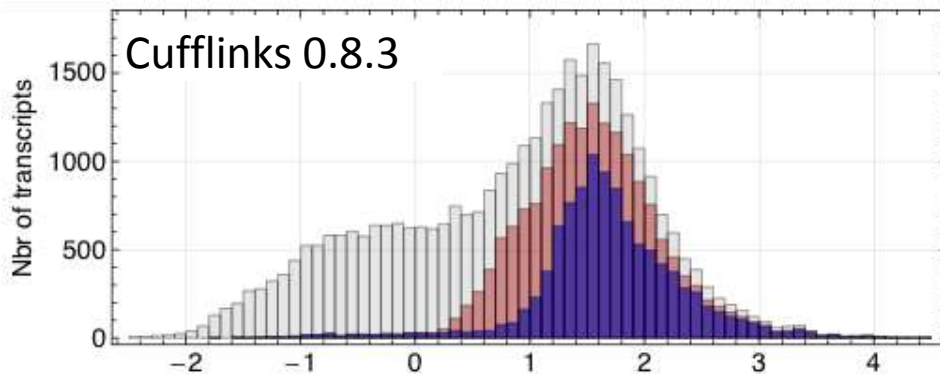
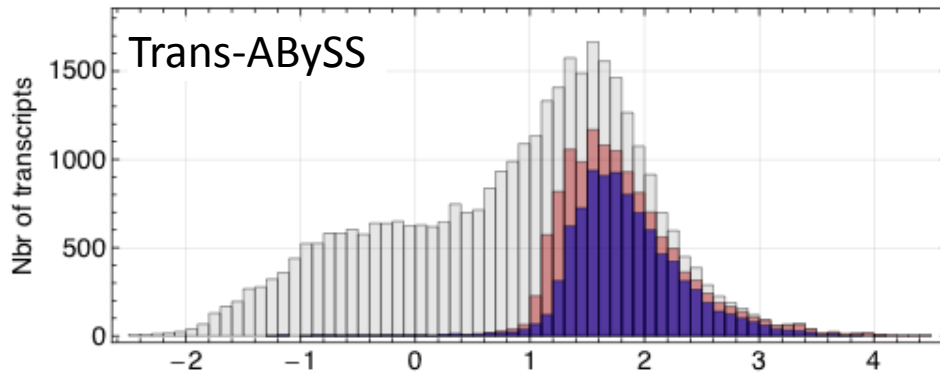
- Gray: all transcripts with a read alignment
- Blue: at least 80% of a transcript in a single contig
- Red: at least 80% of a transcript is reconstructed



# Trans-ABYSS Pipeline



# Transcriptome Assembly



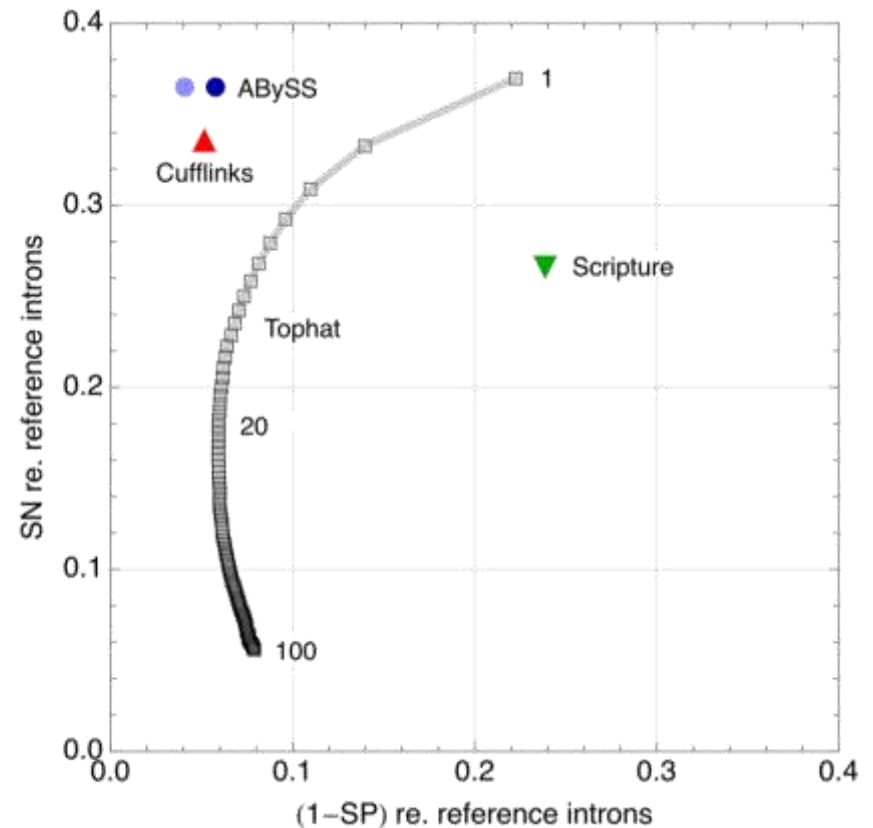
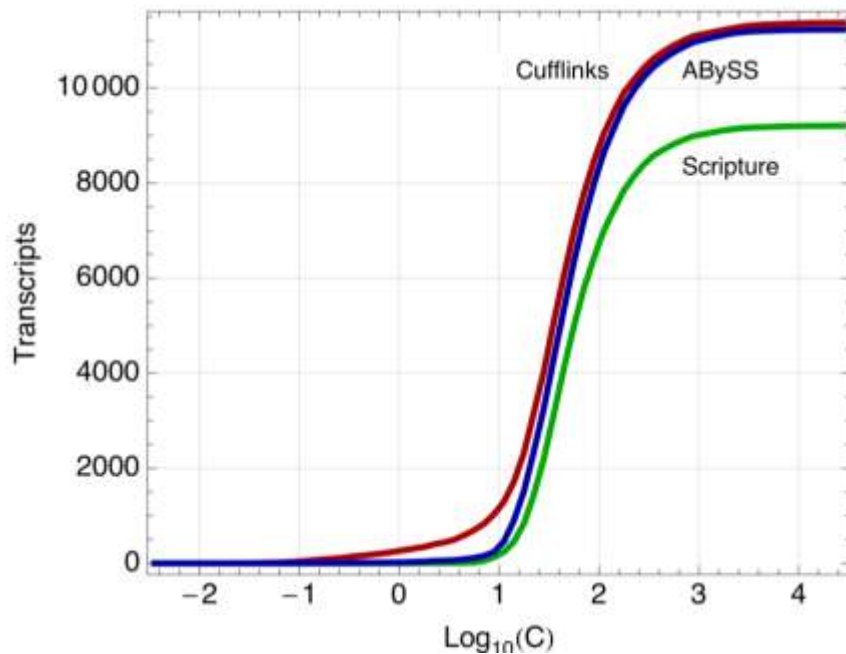
*De novo* assembly based on ABySS

Reference-based assembly based on TopHat alignments

[Trapnell et al., 2010; Guttman et al., 2010; Trapnell et al., 2009]

# Performance

- Trans-ABYSS constructs
  - as many transcripts
  - with better sensitivity and specificity



# Case Study – NBL

- 650 new cases per year in the US
- Most common extracranial solid cancer in childhood
- Most common cancer in infancy
- A neuroendocrine tumour  
(between the hormonal and nervous systems)
- One of the few human malignancies known to demonstrate spontaneous regression
- Stratified into three risk categories
  - low, intermediate, high



# NBL – Cause

- Certain cases run in families
  - very rare germline mutation in ALK (Mossé et al. 2008)
- Risk factors proposed
  - parental factors
    - exposure to chemicals, smoking, alcohol, medicinal drugs during pregnancy and birth factors (Olshan and Bunin, 2000)
    - maternal use of hair dye (McCall et al. 2005; Heck et al. 2009)
    - hormones and fertility drugs (Olshan et al. 1999)
  - atopy and exposure to infection early in life (Menegaux et al. 2004)
- Results are inconclusive



# NBL TARGET Project

- GSC:
  - 10 WGSS Tumour 30x
  - 10 WGSS Matched Constitutional (blood) 30x
  - 10 RNA-seq Tumour 12 Gb raw
- Broad Inst.:
  - 100 tumour exome capture sequencing
- NCI:
  - ~100 tumour RNA-seq



# Analysis

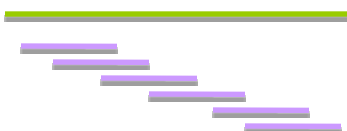
millions of short paired-end reads



*de novo* assembly  
ABySS

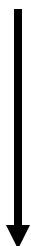


blat conigs to genome

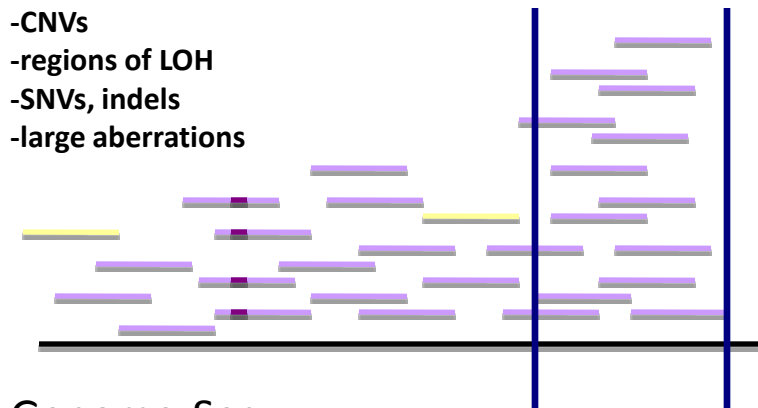


- translocations
- gene fusions
- novel splice variants

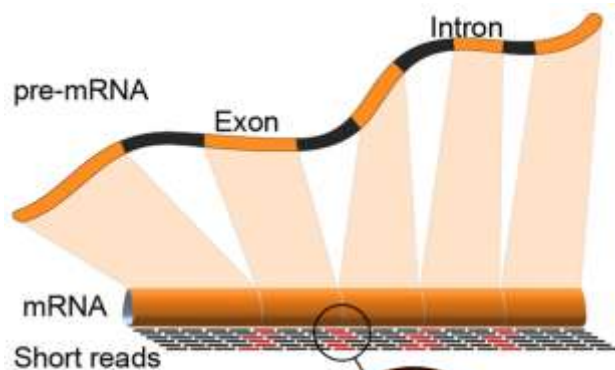
alignment  
(BWA)



- CNVs
- regions of LOH
- SNVs, indels
- large aberrations



Genome-Seq



- SNVs, indels
- gene fusions
- gene-level expression
- exon-level expression
- alternative splicing

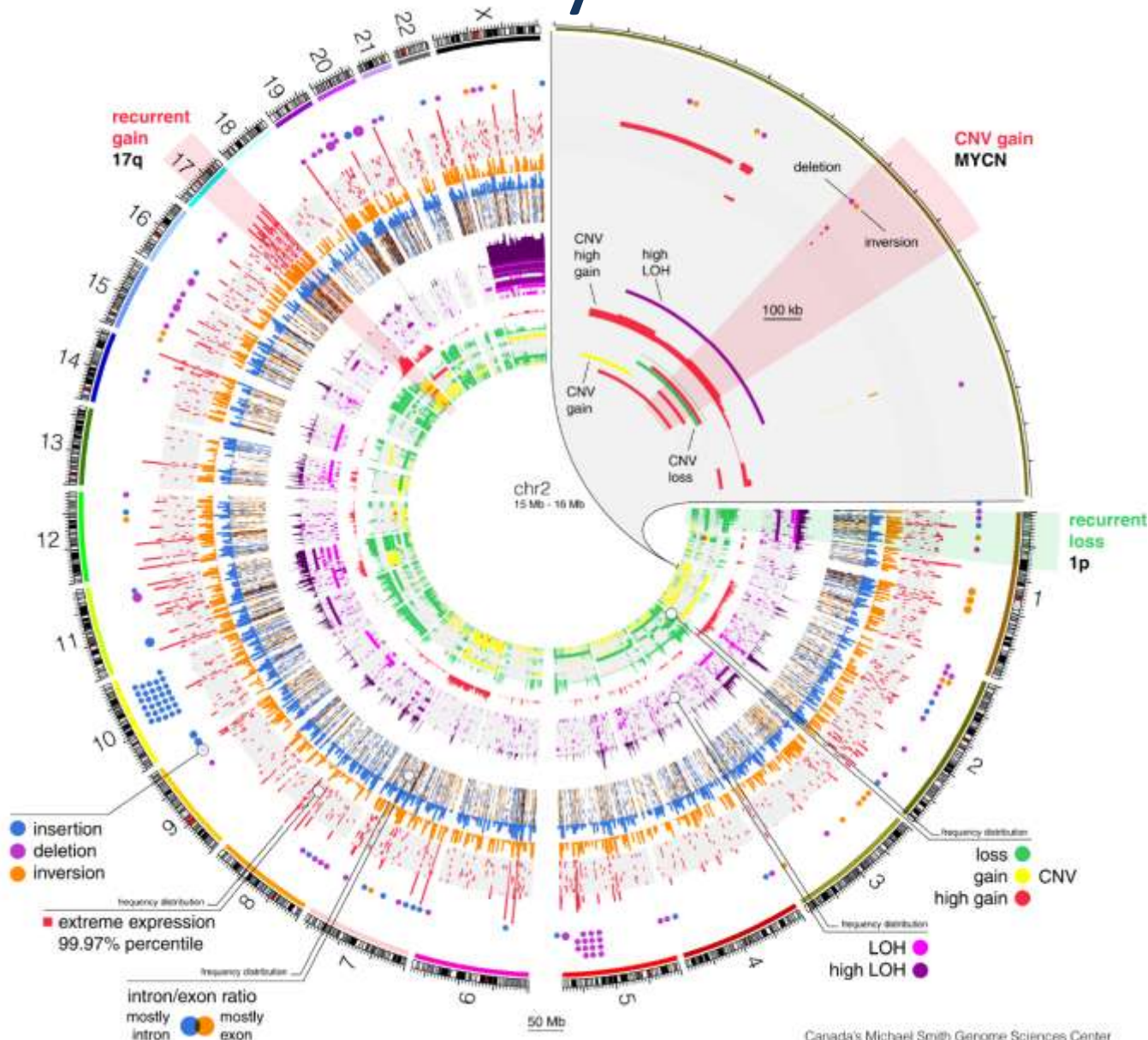
Short read is split by intron when aligning to reference Genome

Transcriptome (RNA-Seq)

Olena Morozova



# Analysis Results

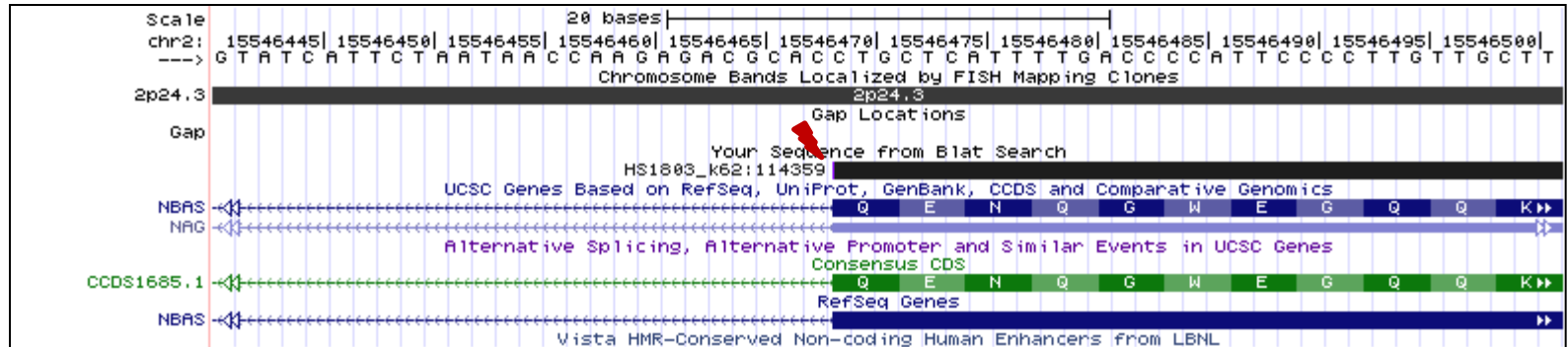
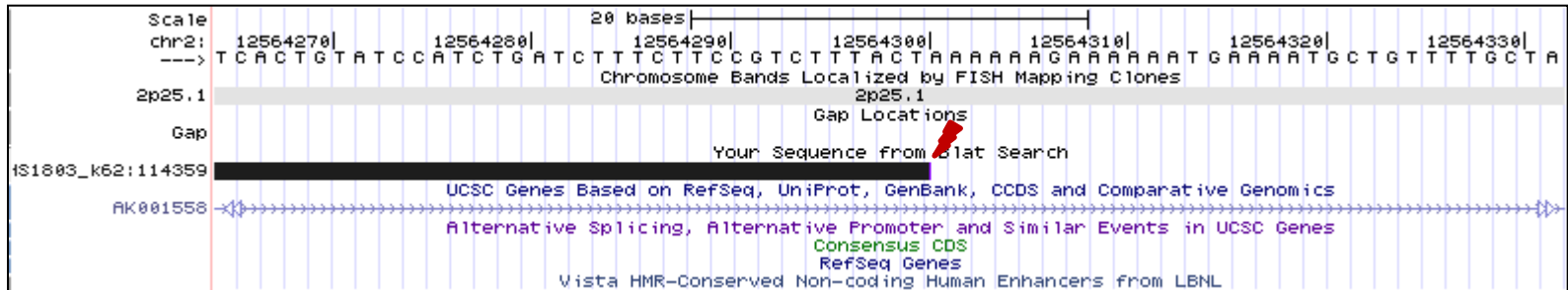


Richard Corbett  
Sa Li  
Olena Morozova  
Martin Krzywinski



# 3Mb deletion on chr2 disrupts NBAS

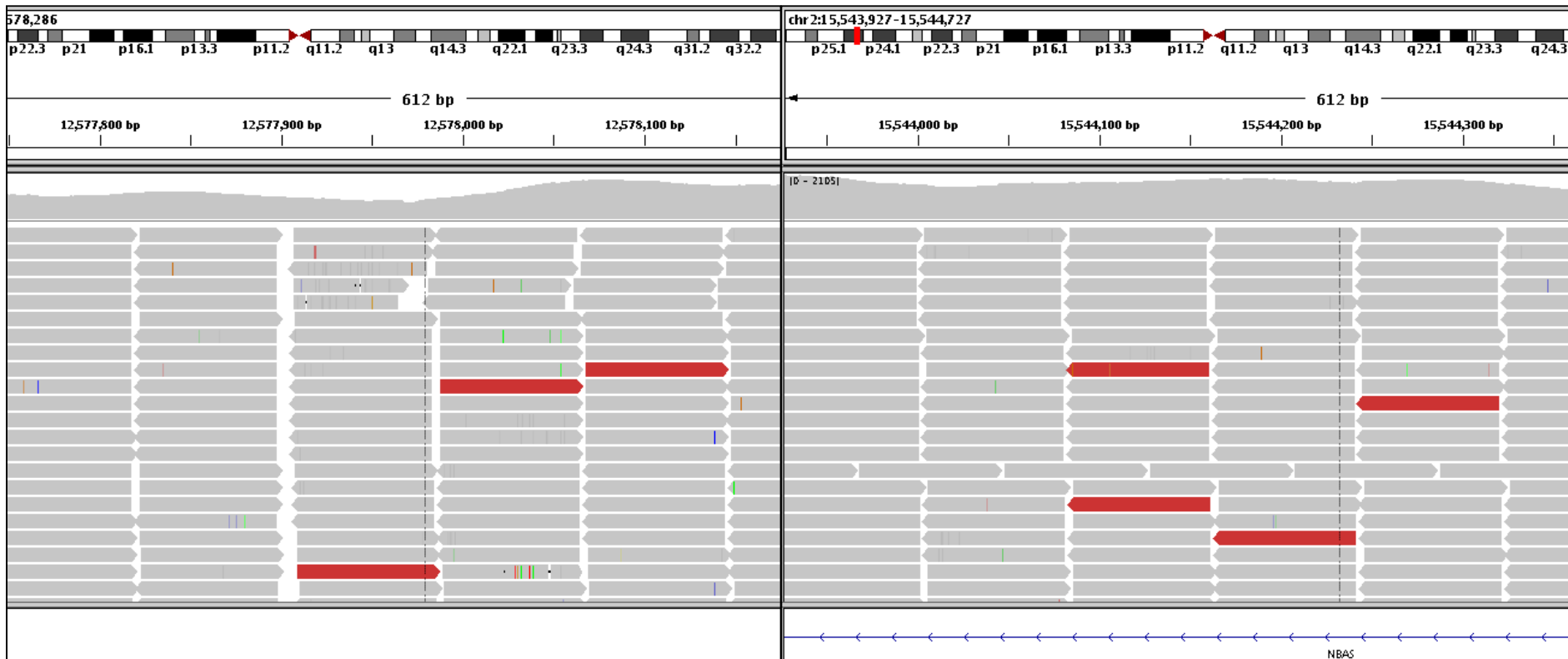
k62:114359(127bp) TARGET:chr2:12564233-12564300,chr2:15546469-15546529 CONTIG:1-68,67-127 +, READPAIRS:10 SPAN\_READS:12



# Genomic data supports 3Mb deletion in NBAS

chr2:12,577,750-12,578,200

chr2:15,543,950-15,544,350



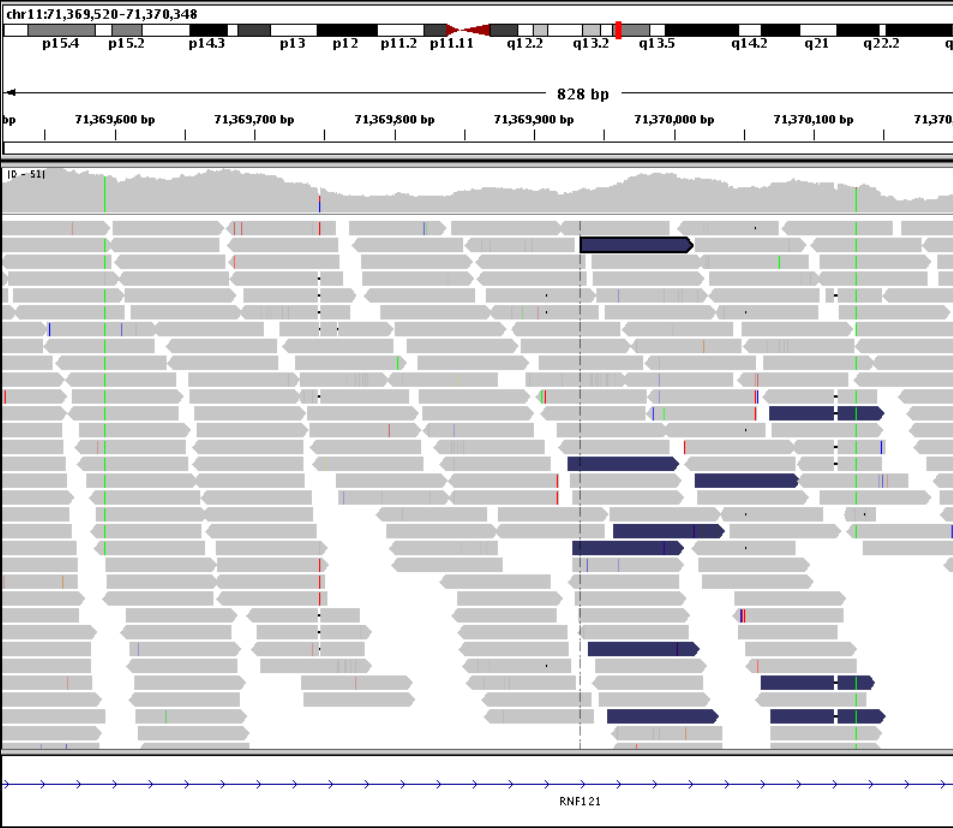
Coloured reads bridge the deletion on chr2



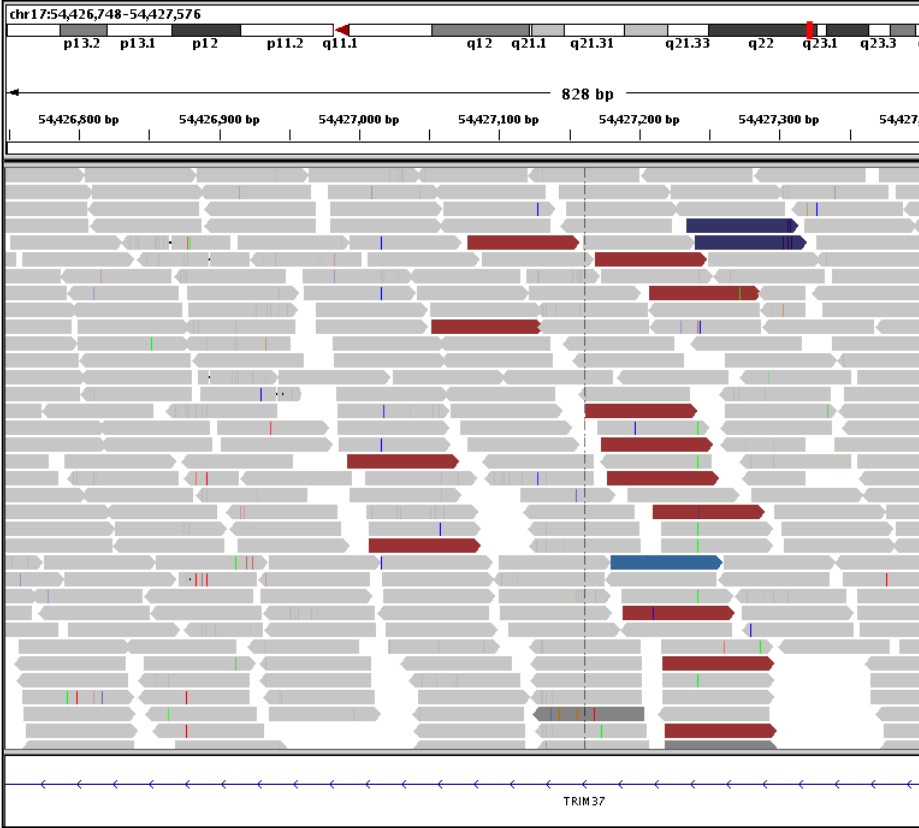


# Genomic data indicates translocation

chr11



chr17



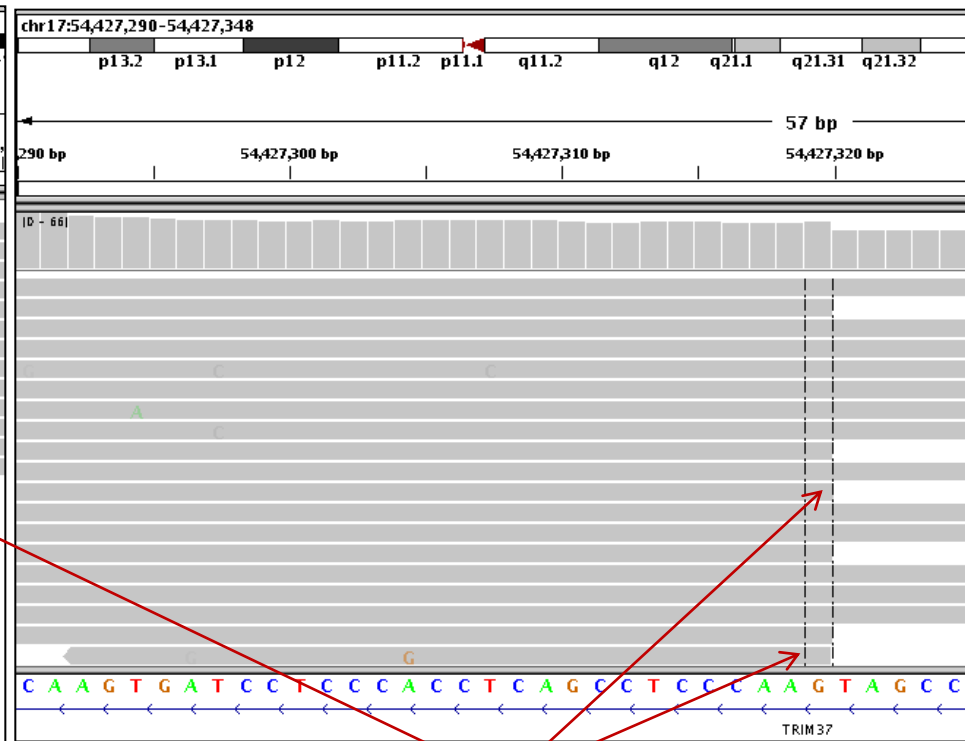
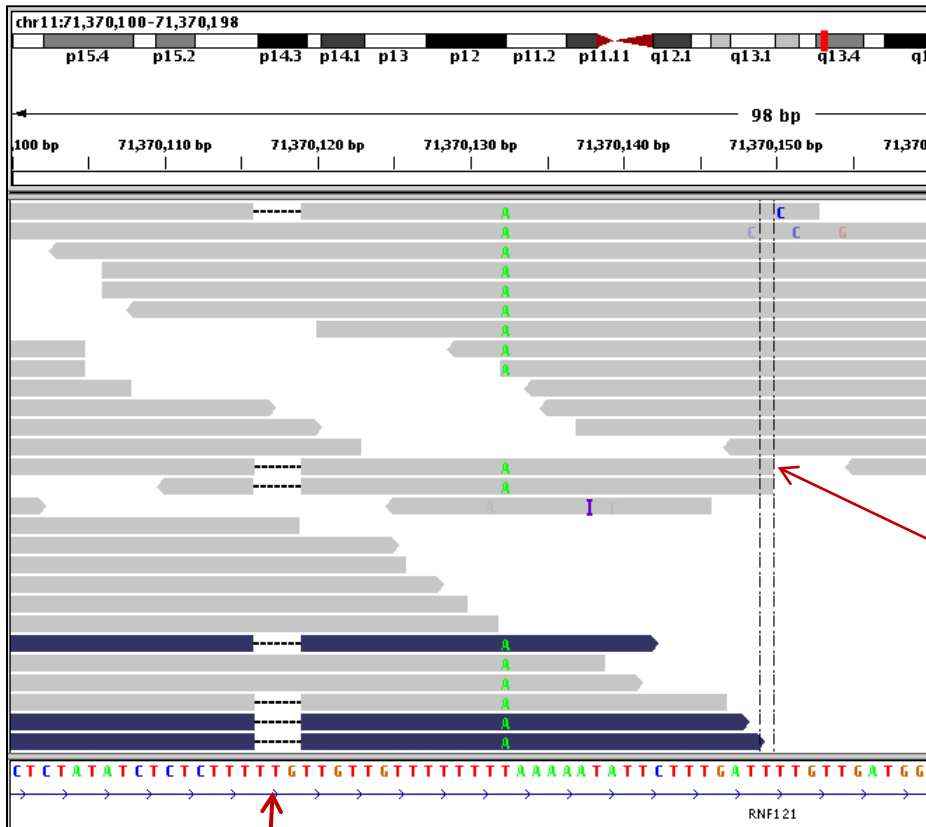
Coloured reads are mate pairs that span the translocation breakpoint



# Genomic breakpoint identified with IGV

Genomic breakpoint chr11:71,370,150

Genomic breakpoint chr17:54,427,320



Reads bridging the translocation contain a novel 3bp deletion TTG

Truncated reads found in genomic BAM file cross the translocation breakpoint

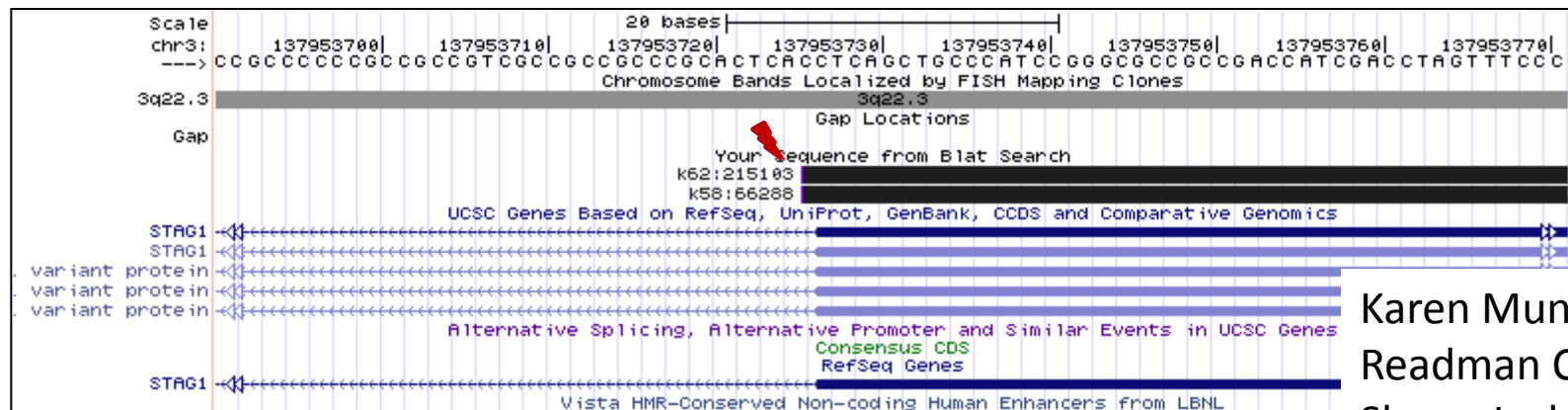
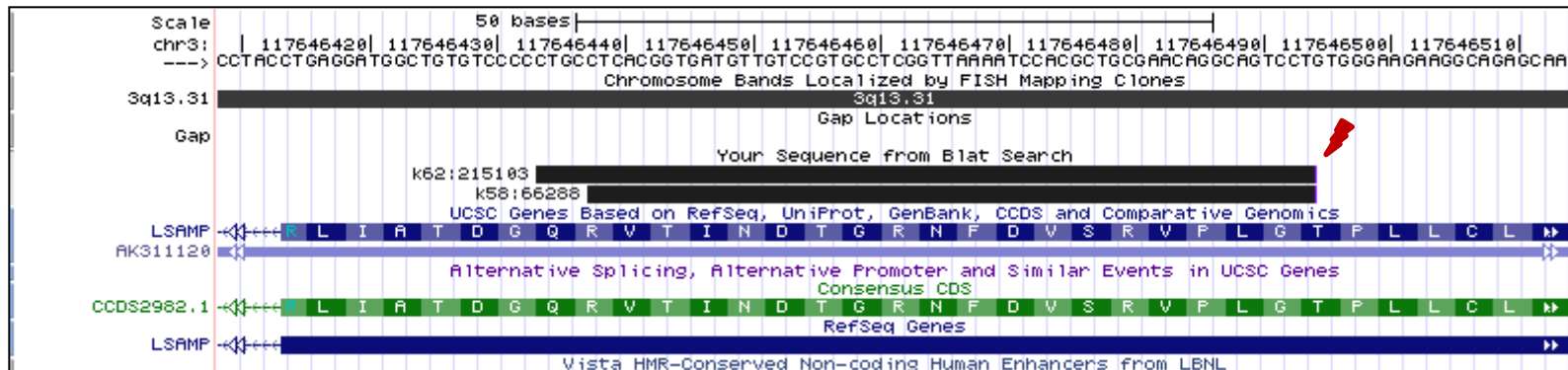
Karen Mungall





# 20Mb deletion creates a LSAMP/STAG1 fusion

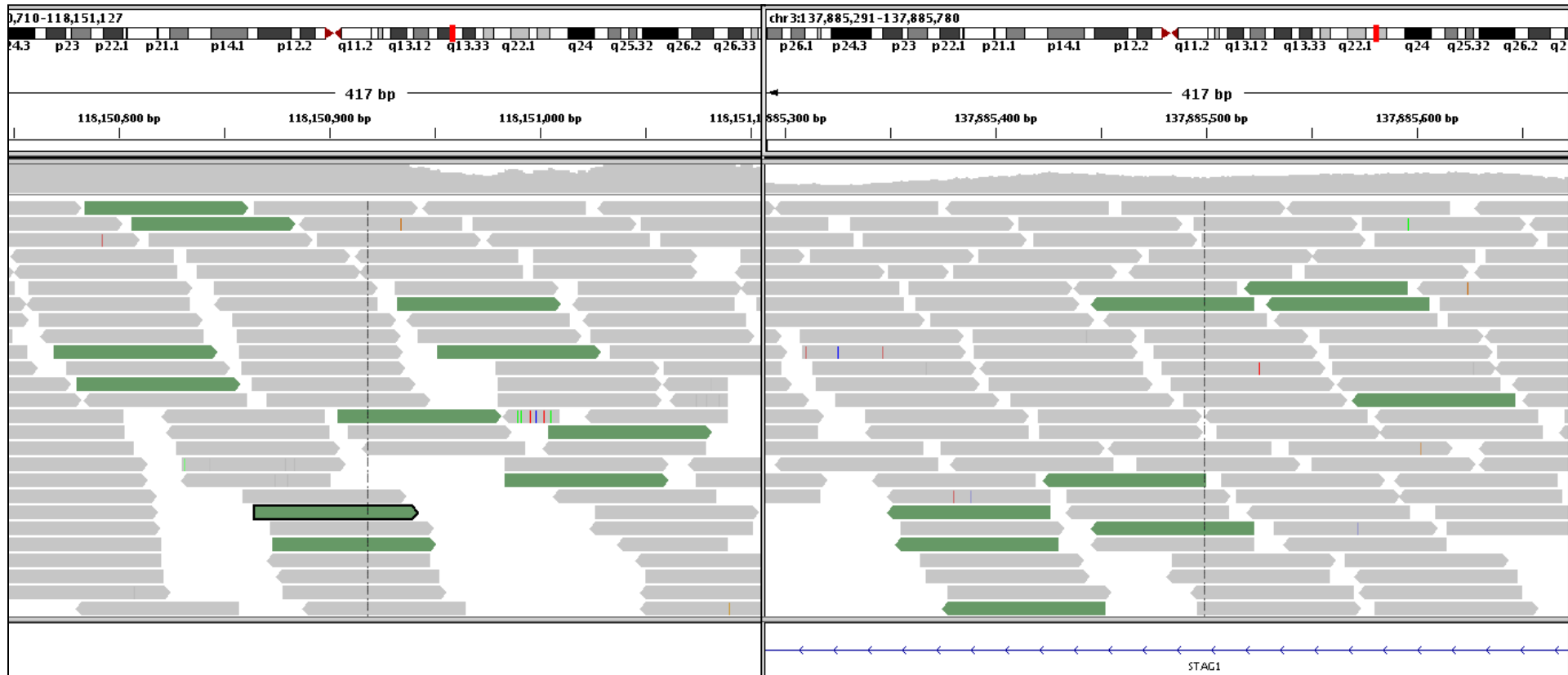
k62:215103(135bp) TARGET:chr3:137953726-137953802,chr3:117646434-117646494 CONTIG:1-77,75-135 -,- READPAIRS:9 SPAN\_READS:3  
 k58:66288(133bp) TARGET:chr3:137953726-137953804,chr3:117646438-117646494 CONTIG:1-79,77-133 -,- READPAIRS:9 SPAN\_READS:6



Karen Mungall  
 Readman Chiu  
 Shaun Jackman  
 Jenny Qian



# Genomic data confirms the deletion affecting STAG1 and places the breakpoint upstream of LSAMP



chr3:118,150,800-118,151,100

chr3:137,885,300-137,885,650

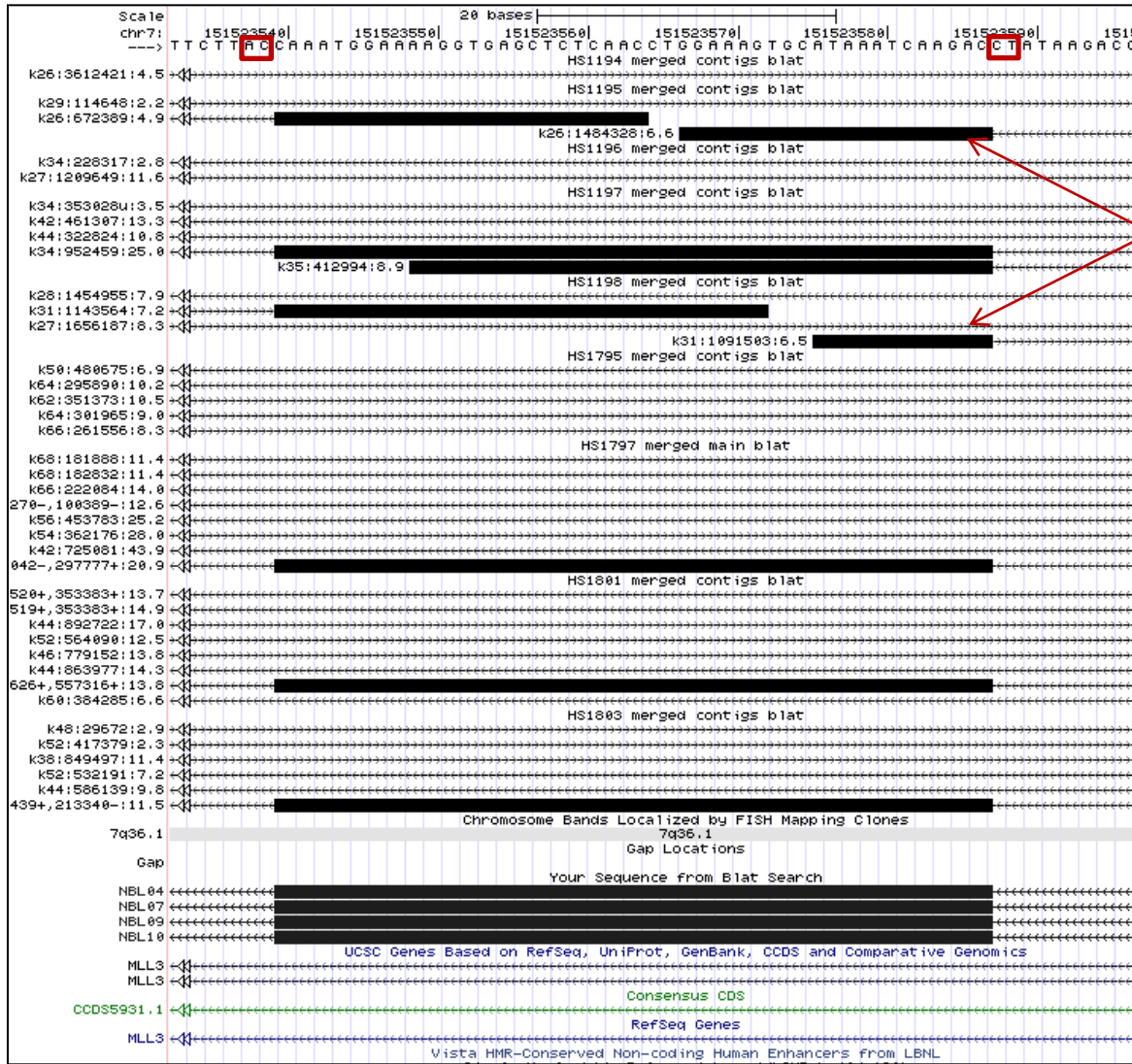
Karen Mungall



# Novel exon in MLL3 in 4 libraries

NBL04 novel\_exon k34:952459 uc003wkz.1 MLL3 32,33 10 chr7:151523540-151523587 48 orf:good,3064-3111;DDL...PHG,1046aa,0-3140,3140nt,1.00,1  
 NBL07 novel\_exon k54:13434+,133042-,297777+ uc003wkz.1 MLL3 32,33 7 chr7:151523540-151523587 48 orf:good,682-729;GWS...RKA,381aa,0-1145,1145nt,1.00,1  
 NBL09 novel\_exon k42:503365-,722626+,557316+ uc003wkz.1 MLL3 32,33 7 chr7:151523540-151523587 48 orf:good,670-717;EGE...AKL,460aa,1-1383,1382nt,1.00,1  
 NBL10 novel\_exon k38:95065+,154439+,213340- uc003wkz.1 MLL3 32,33 5 chr7:151523540-151523587 48 orf:good,318-365;KRS...MNG,282aa,1-849,848nt,1.00,1

MLL3



Evidence for novel exon in NBL02 and NBL05 in addition to the 4 other libraries

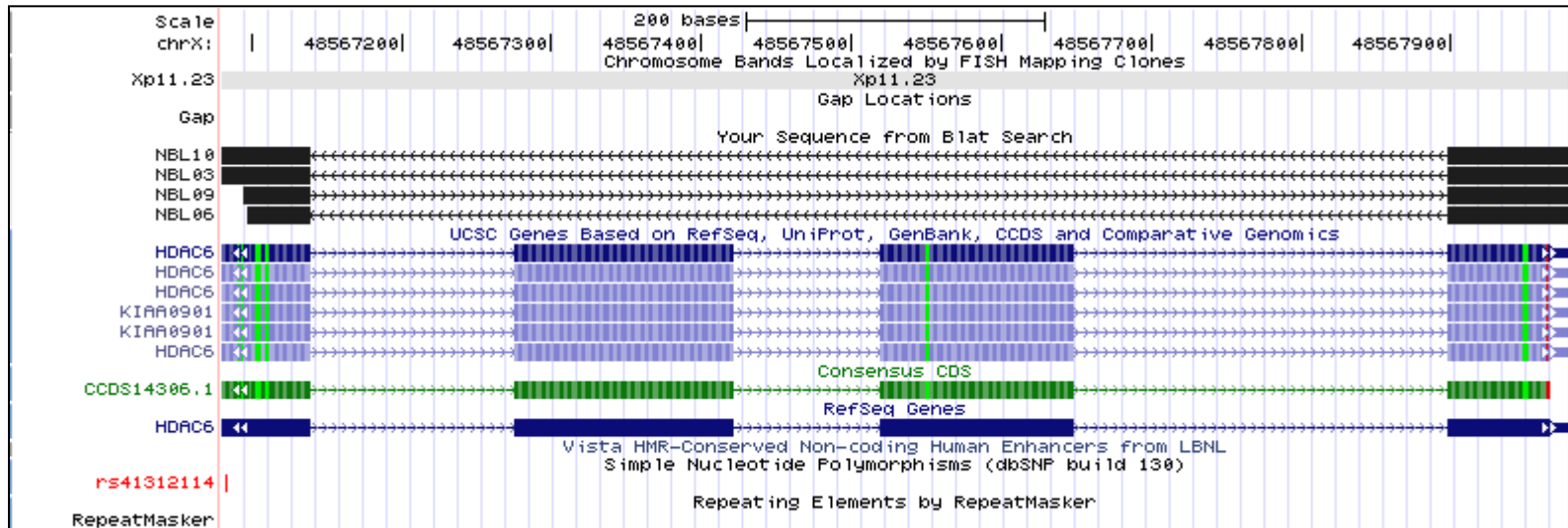
Some splicing support seen in R2G file HS1197



# 2 skipped exons in HDAC6 found in 4 libraries

NBL03 skipped\_exon k26:163806-,843903+,1264643+ uc004dks.1 HDAC6 27,28 4,5 chrX:48567276-48567648 orf:1004-1565;FNP...HPH,394aa,379-1563,1184nt,0.76,-1  
 NBL06 skipped\_exon k44:143578+,708557+,296001+ uc004dks.1 HDAC6 27,28 1,2 chrX:48567276-48567648 orf:0-957;GVG...LSS,114aa,223-567,344nt,0.36,1  
 NBL09 skipped\_exon k46:145635+,145689-,499421- uc004dks.1 HDAC6 27,28 1,2 chrX:48567276-48567648 orf:0-667;YHE...YKM,91aa,390-665,275nt,0.41,1  
 NBL10 skipped\_exon k40:89073-,724463+,717355+ uc004dks.1 HDAC6 27,28 4,5 chrX:48567276-48567648 orf:1033-1579;FNP...HPH,394aa,393-1577,1184nt,0.75,-1

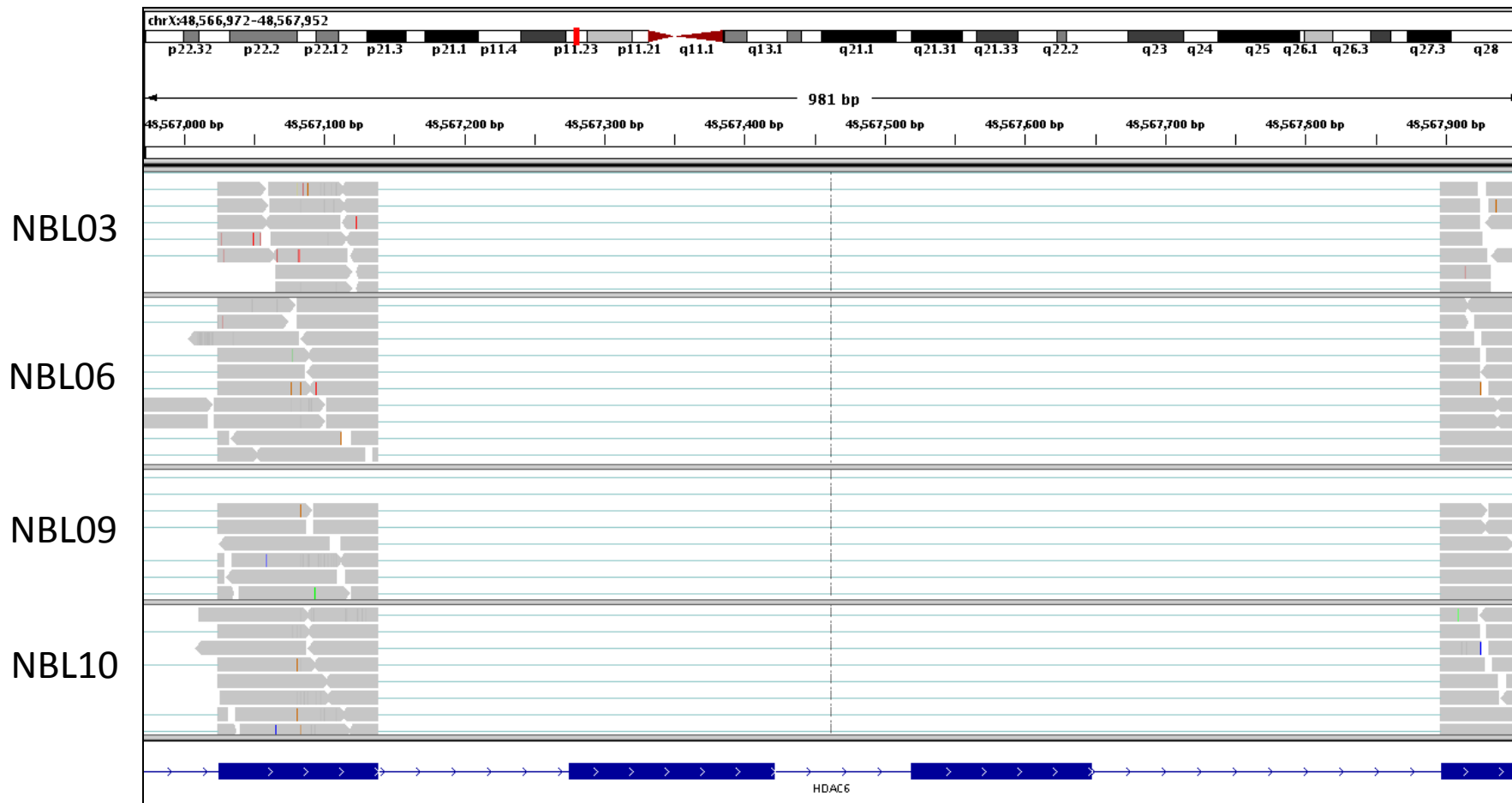
HDAC6



Evidence seen in R2G BAM file

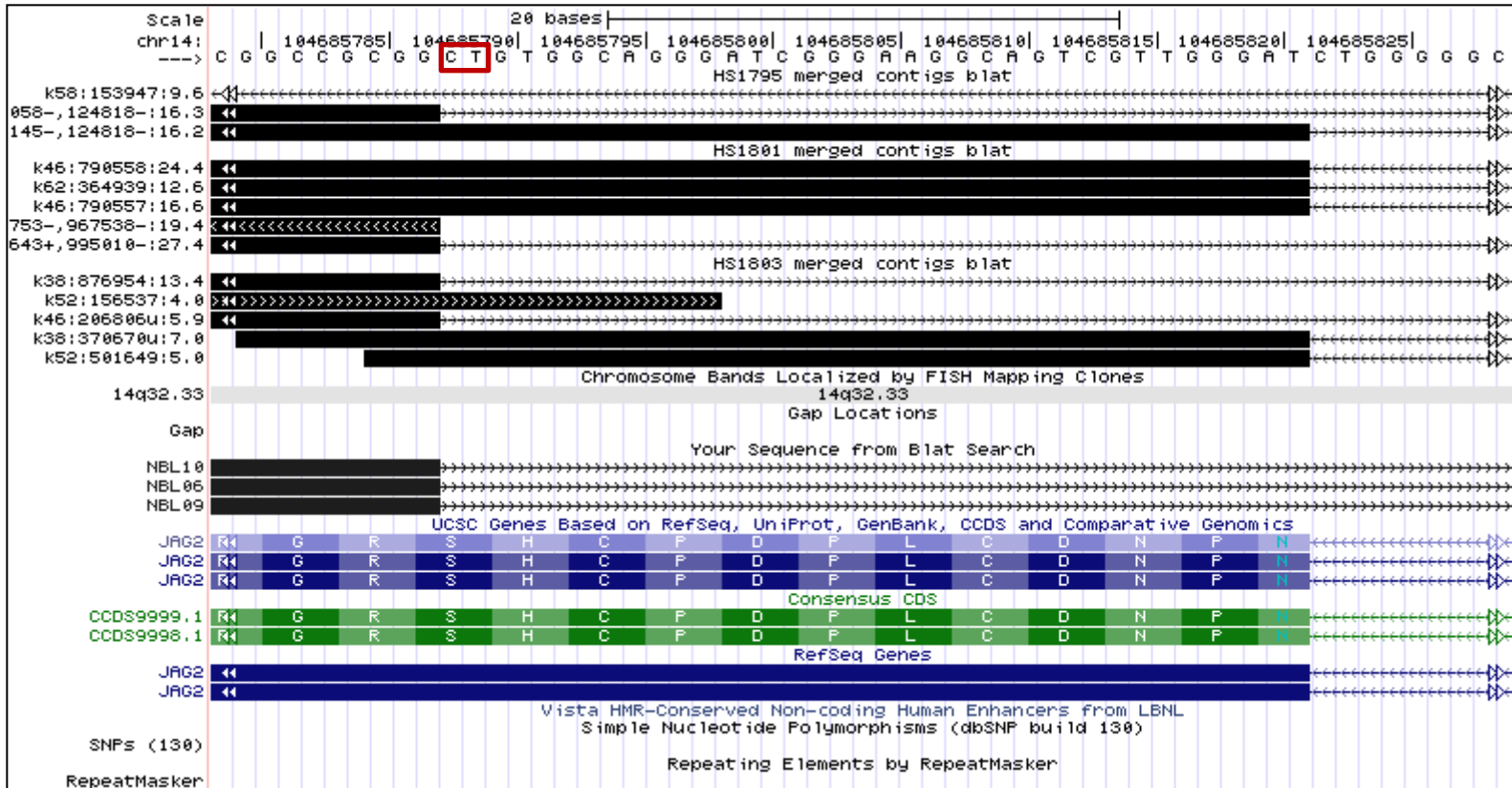


# Split read alignments confirm 2 exon skip in HDAC6



# Alt 3' splice in JAG2 found in 3 libraries

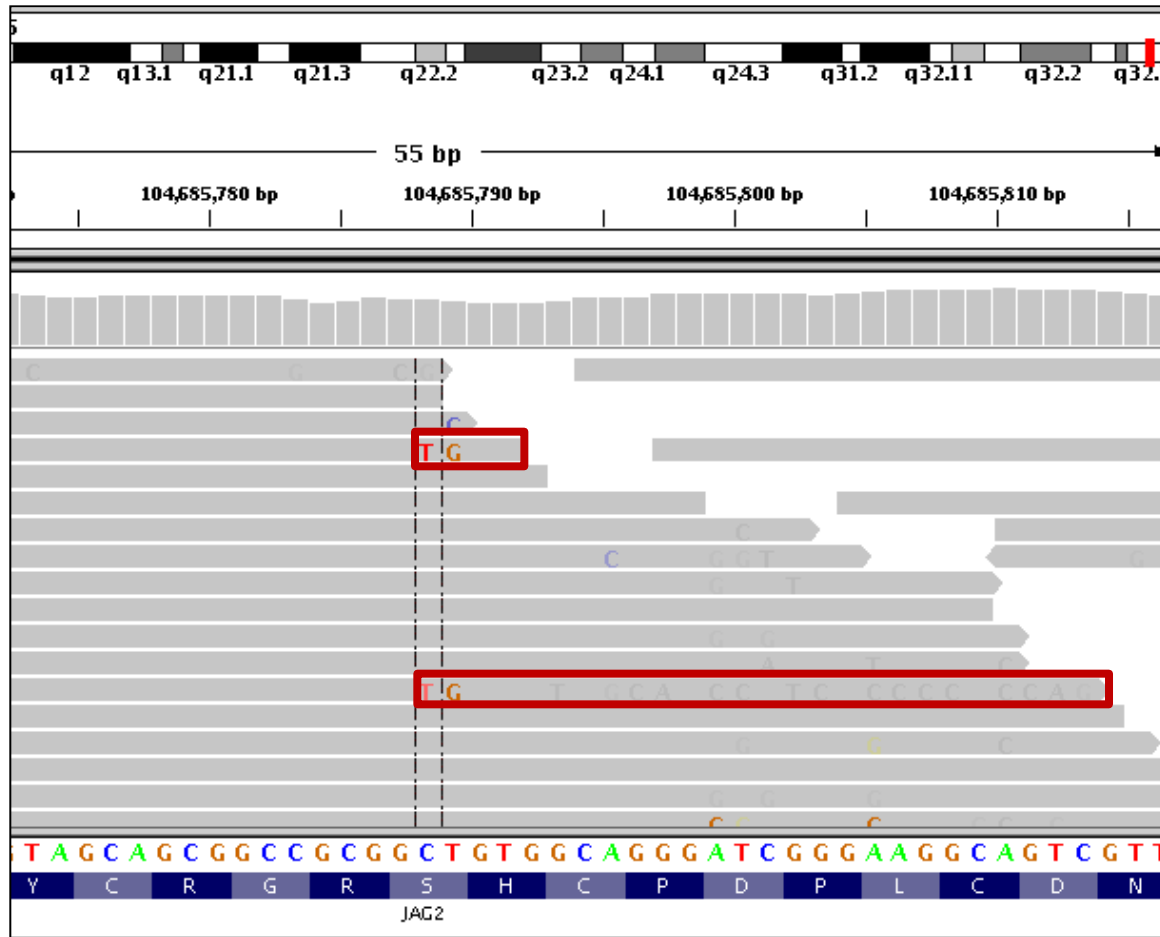
NBL06AS3 k38:79305-,499058-,124818- uc001yqg.1 JAG2 11 11 chr14:104685708-104685787 orf:good,2537-2616;ARA...TPV,635aa,2387-4294,1907nt,0.44  
 NBL09AS3 k38:134124+,557643+,995010- uc001yqf.1 JAG2 11 5 chr14:104685708-104685787 orf:good,345-424;VND...TPV,288aa,195-1061,866nt,0.81,-1  
 NBL10 AS3 k38:876954 uc001yqg.1 JAG2 11 11 chr14:104685708-104685787 orf:good,2538-2617;SFT...TPV,609aa,2388-4217,1829nt,0.43,-1  
 NBL10 AS3 k46:206806u uc001yqf.1 JAG2 11 2 chr14:104685708-104685787 orf:good,9-88;PCH...HAS,197aa,0-593,593nt,0.99,-1



Evidence seen in R2G BAM



# Read support for alternate splice in JAG2

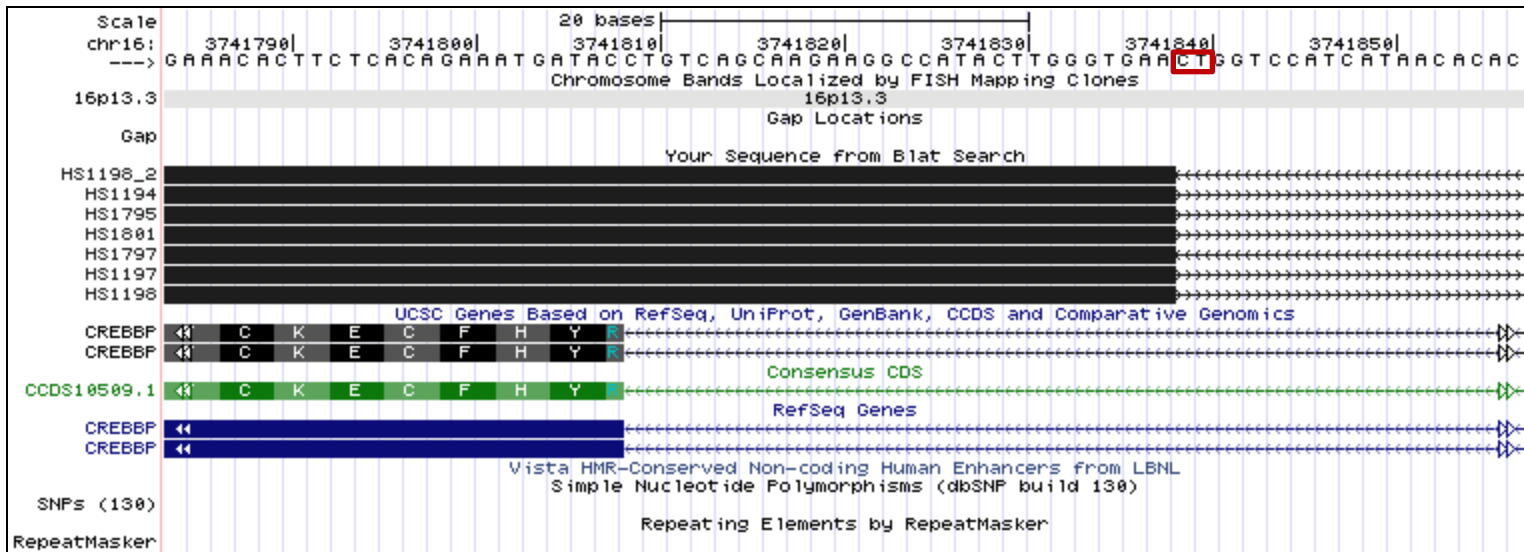


Sequence in red aligns at next exon



# Alt 3' splice in CREBBP found in 6 libraries

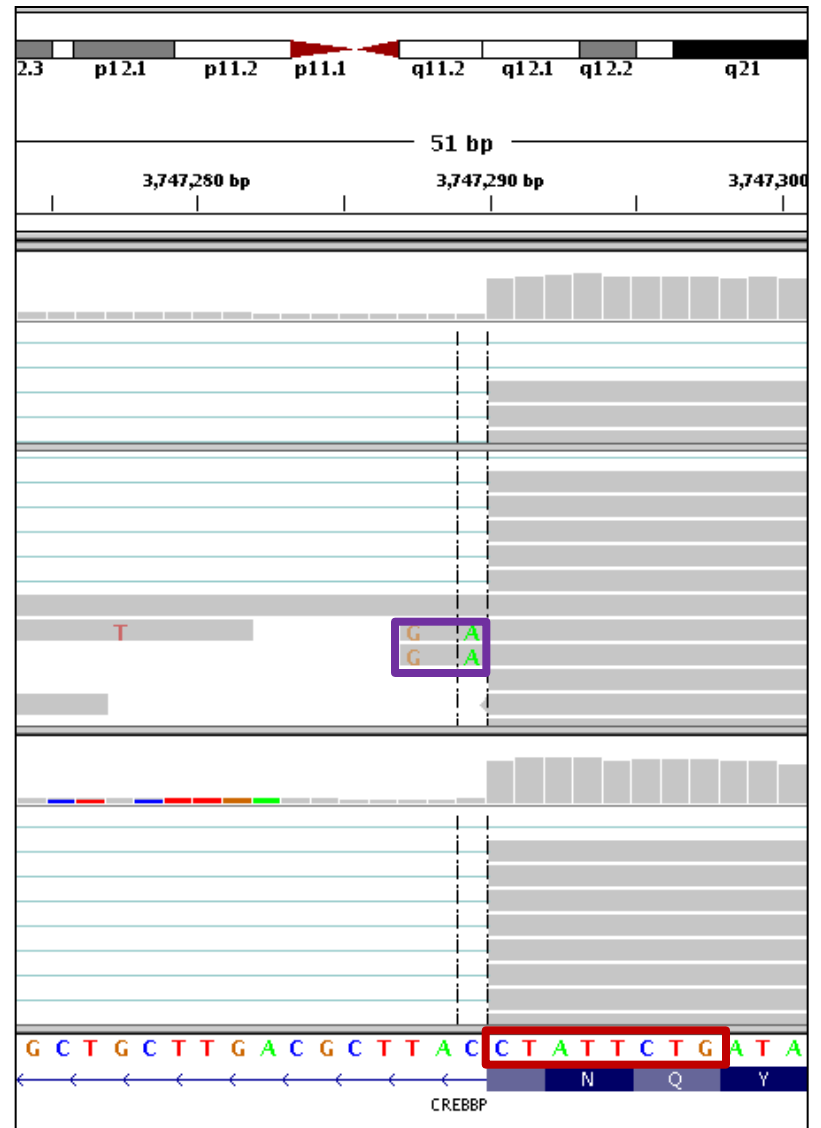
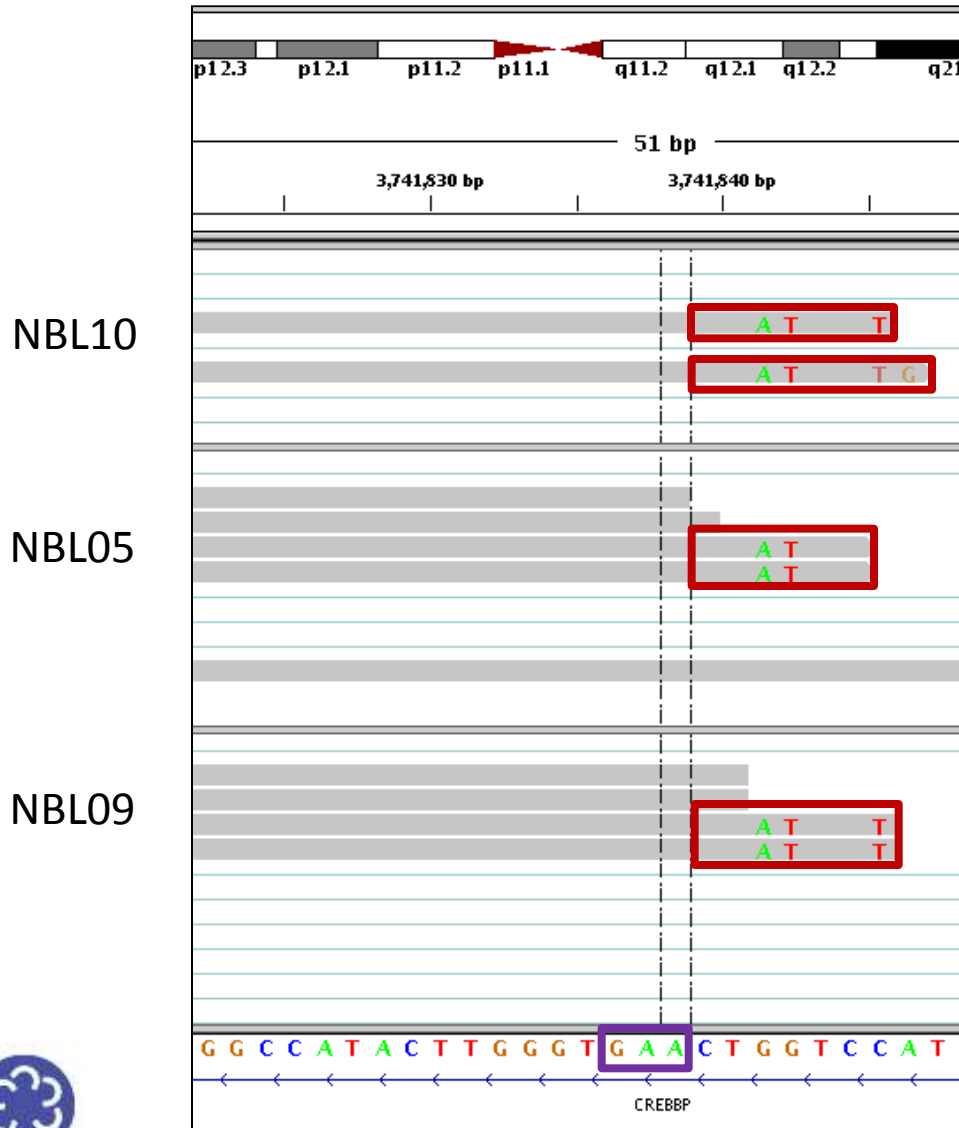
NBL01 AS3 k34:671268+,264091+,871201- uc002cvv.1 CREBBP 12 7 chr16:3741728-3741838 orf:good,532-642;STS...LDS,373aa,0-1121,1121nt,1.00,-1  
NBL04AS3 k33:530544+,414942-,999817- uc002cvv.1 CREBBP 12 3 chr16:3741728-3741838 orf:good,89-199;YEF...CGR,95aa,1-288,287nt,0.99,-1  
NBL05 AS3 k34:256068-,165611-,387773- uc002cvv.1 CREBBP 12 2 chr16:3741728-3741838 orf:good,33-143;PTV...FEK,364aa,2-1096,1094nt,1.00,-1  
NBL05 AS3 k42:396876 uc002cvv.1 CREBBP 12 11 chr16:3741728-3741838 orf:good,1212-1322;QAE...LMD,671aa,1-2016,2015nt,1.00,1  
NBL06 AS3 k62:99512+,235840+,185139+ uc002cvv.1 CREBBP 12 5 chr16:3741728-3741838 orf:good,264-374;ESS...KFL,293aa,2-883,881nt,1.00,-1  
NBL07 AS3 k40:757985-,722866-,790202+ uc002cvv.1 CREBBP 12 5 chr16:3741728-3741838 orf:good,249-359;KKI...HLE,268aa,0-806,806nt,1.00,1  
NBL09 AS3 k52:241464+,104576+,295260- uc002cvv.1 CREBBP 12 5 chr16:3741728-3741838 orf:good,254-364;NGT...RVN,286aa,1-861,860nt,1.00,-1

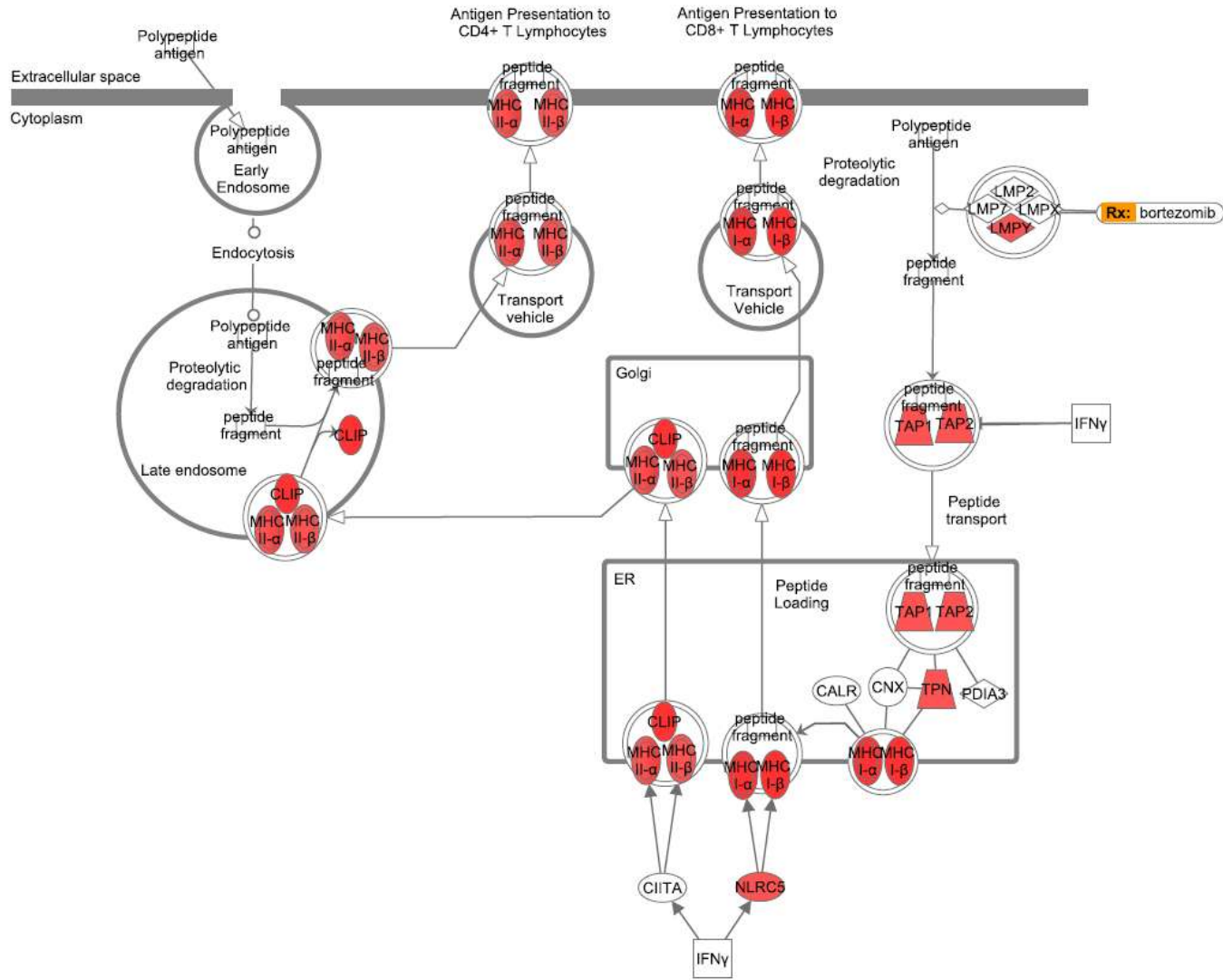


Evidence seen in R2G BAM for alternate splice for NBL10 as well



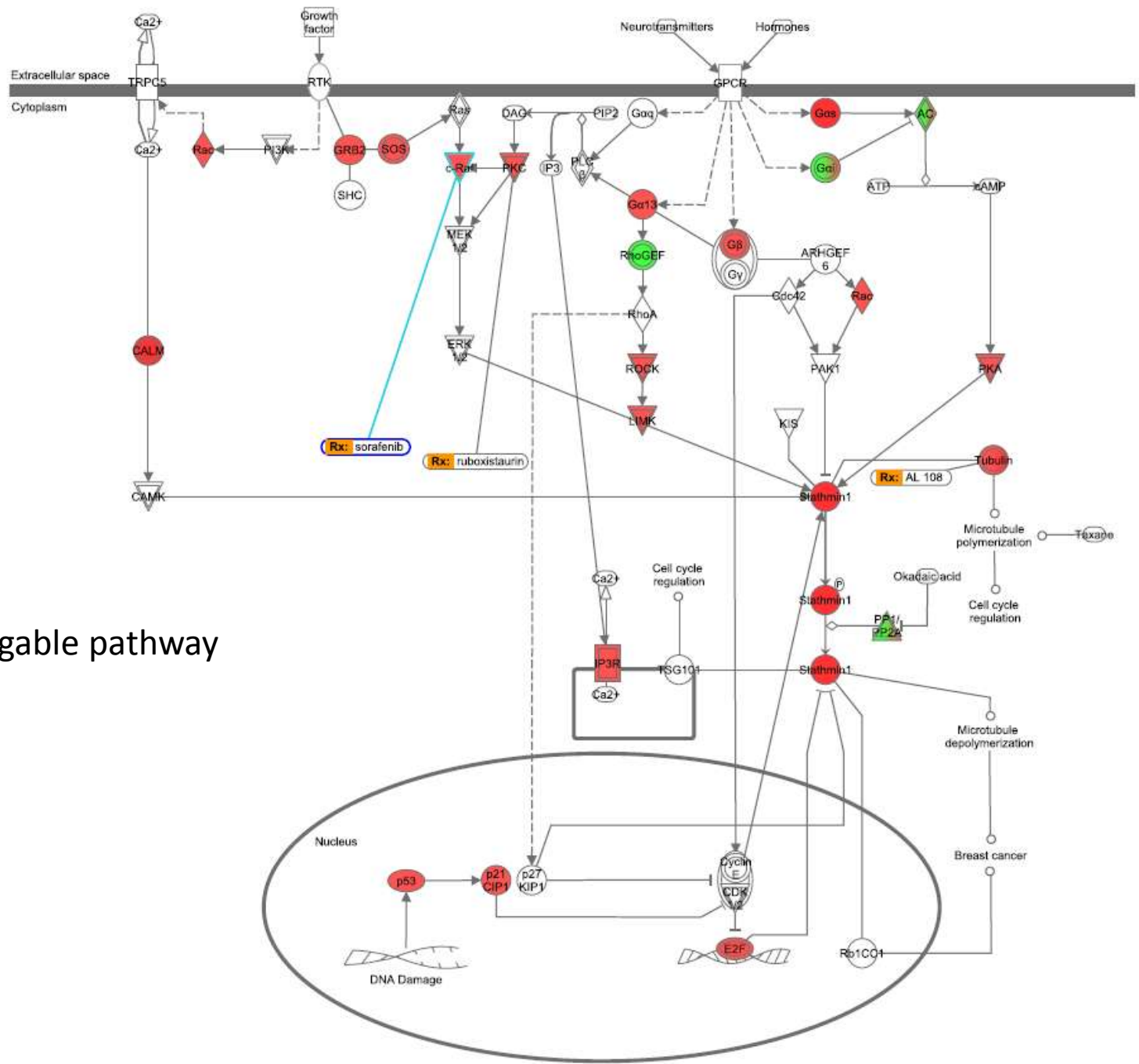
# Reads support the novel 3' splice site in CREBBBP





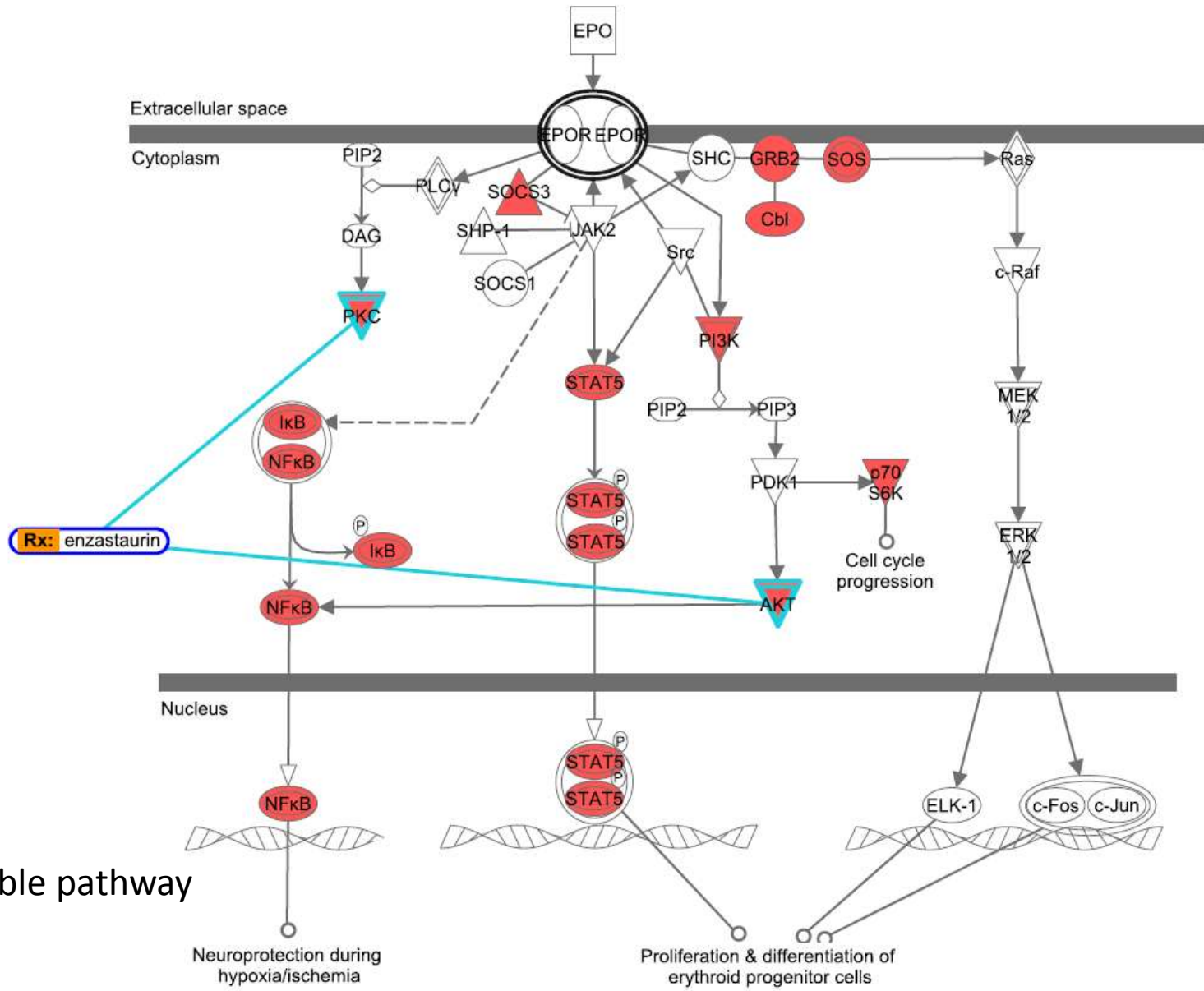
Top druggable pathway in NBL1





Top druggable pathway in NBL2





Top druggable pathway  
in NBL7



# Summary

- Sequencing technology is changing rapidly
- Enabling new and exciting applications
- To reap the benefits, bioinformatics has to catch up
  
- Assembly of Illumina data yields significant insights in studying cancer genomes and transcriptomes
- High throughput assembly and analysis is feasible



# Acknowledgements

Marco Marra  
Steve Jones  
Rob Holt

Richard Moore &  
Sequencing team

Yongjun Zhao &  
Library Core team

Greg Stazyk &  
Systems team

Shaun Jackman  
Jenny Qian  
Rong She  
Readman Chiu  
Karen Mungall  
Gordon Robertson

Olena Morozova  
Richard Corbett  
Sa Li



Canada's Michael Smith  
**GENOME  
SCIENCES**  
CENTRE



BC Cancer Agency  
CARE & RESEARCH