Aligners

J Fass | 21 June 2017

Definitions

Assembly:

I’ve found the shredded remains of an important document; put it back together!

Definitions

Alignment:

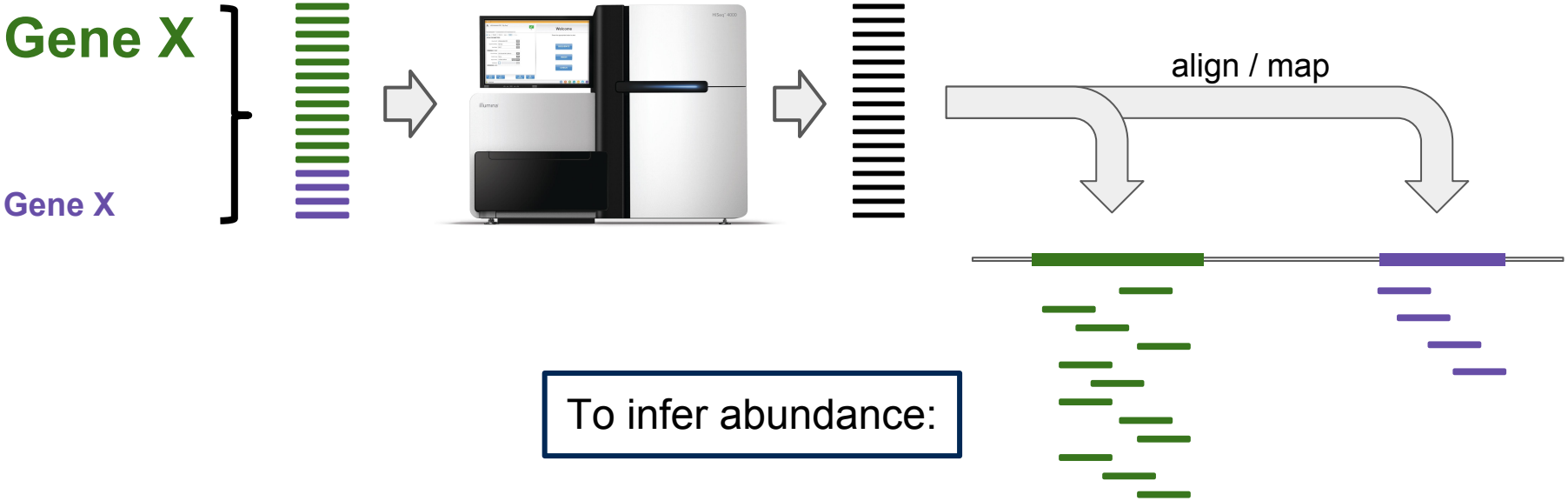
Somebody plagiarized parts of my document; where did they copy paragraphs from and where did each of the words come from?

Definitions

Mapping:

Somebody plagiarized parts of my document; where did they copy paragraphs from ~~and where did each of the words come from~~?

Why align (or map)?



To infer abundance:

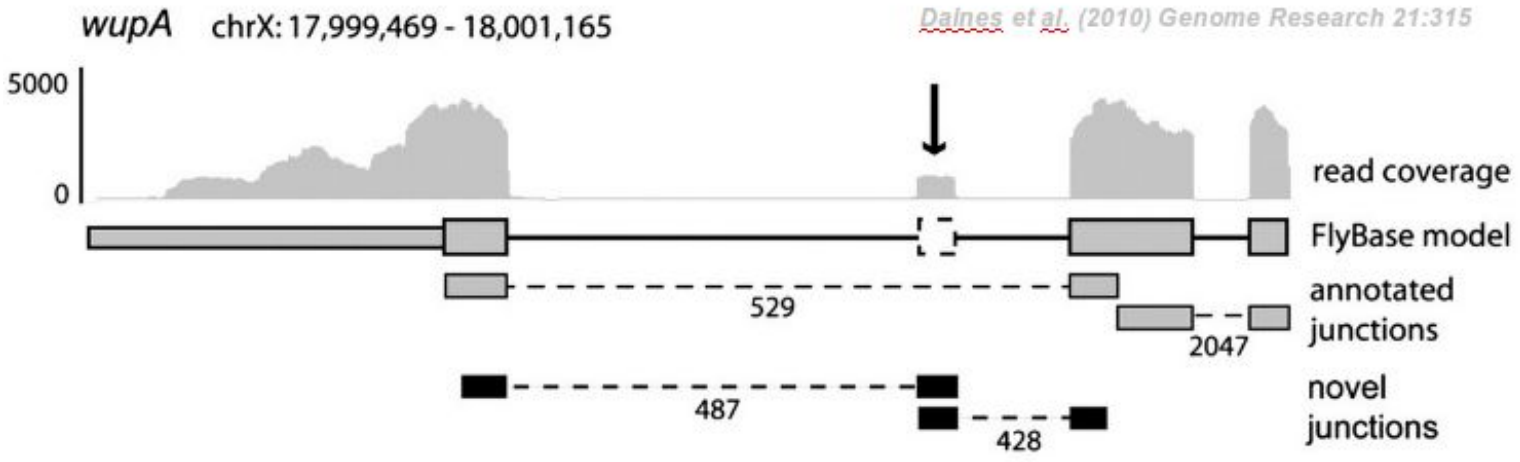
Why align (~~or map~~)?

Individual A →  ← **Individual B**

```
ATGATAGCATCGTCGGGTGTCTGCTCAATAATAGTGCCGTATCATGCTGGTGTTATAATCGCCGCATGACATGATCAATGG
                                CAATAAAAGTGCCGTATCATGCTGGTGTTACAATCGCCGCA
                                          CGTATCATGCTGGTGTTACAATCGCCGCATGACATGATCAATGG
                            TGTCTGCTCAATAAAAGTGCCGTATCATGCTGGTGTTACAATC
                    ATCGTCGGGTGTCTGCTCAATAAAAGTGCCGTATCATG--GGTGTTATAA
                                CTCAATAAGAGTGCCGTATCATG--GGTGTTATAATCGCCGCA
                                                    GTTATAATCGCCGCATGACATGATCAATGG
```

To measure variation

Why align (~~or map~~)?



To discover transcribed sequence.

Timeline

“Global” and “local”

Global aligners try to align all provided sequence, end to end, both “query” and “subject / target” …

E.g.

- Aligning two *Salmonella* genomes
- Aligning human and gorilla orthologous coding regions

“Global” and “local”

Local aligners try to find “hits” or chains of hits *within* each provided sequence …

E.g.

- Finding mitochondrial “splinters” in nuclear chromosomes
- Finding genes that share a domain with a gene of interest

“Glocal … ?”

Short read aligners generally assume that the *whole read* came from somewhere within the target (reference) sequence …

… so, ***global*** with respect to the read, and ***local*** with respect to the reference.

Short Read (Non-splicing) Aligners

Li, H and Homer, N (2010) *Briefings in Bioinformatics* 11:473
"A survey of sequence alignment algorithms for next-generation sequencing"

Table 1:

Popular short-read alignment software

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

These two were fastest, at ~7 Gbp (vs human) per CPU day … HiSeq 2500 generated 50-100 Gbp per day (at the time)

(*Fall '12-'13*) … 150-180 Gbp per day

(*Summer '16*) … 600 Gbp per day

(*Summer '17*) … 1-3 Tbp per day

https://www.illumina.com/systems/sequencing-platforms.html

Burrows-**W**heeler Aligners

Burrows-Wheeler Transform used in bzip2 file compression tool; FM-index (Ferragina & Manzini) allow efficient finding of substring matches within compressed text – algorithm is *sub-linear* with respect to time and storage space required for a certain set of input data (reference 'ome, essentially).

Reduced memory footprint, faster execution.

BWA

BWA is a fast gapped aligner. Long read aligners (bwasw and mem) also fast, and can perform well for 454, Ion Torrent, Sanger, and PacBio reads. BWA is actively maintained and has a strong user community.

bio-bwa.sourceforge.net

‘bwa aln’ (BWA “backtrack”) for reads < 70 bp

‘bwa bwasw’

‘bwa mem’ (seeds with *maximal exact matches*, extends via *Smith-Waterman*)

Bowtie

(now Bowtie 2) … comparable to BWA.
Bowtie is part of a suite of tools (Bowtie, Tophat, Cufflinks, CummeRbund) that address RNAseq experiments.

http://bowtie-bio.sourceforge.net

Written by same folks as Tophat … so, full compatibility.

Tophat(2)

Aligns full reads to genome, to determine “coverage islands.” Creates simulated spliced exons based on these islands, then aligns remaining short reads to the simulated cDNA (to find reads crossing splice junctions).

HISAT2

Improved replacement for Tophat2, using ...

STAR

Spliced Transcripts Aligned to a Reference

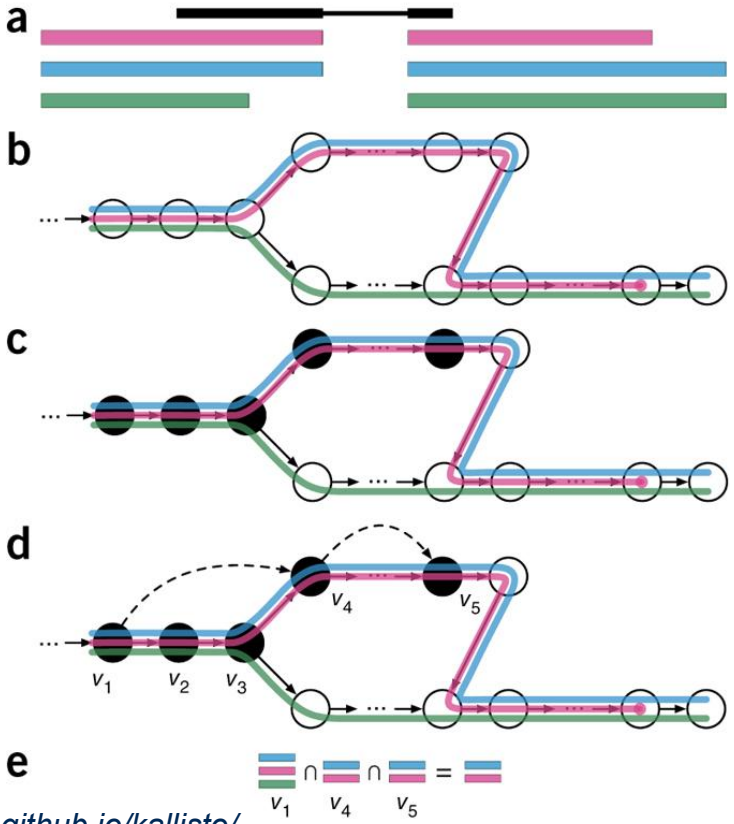
Aligns short reads and full length cDNA

Similar to BWA MEM algorithm, but searches uncompressed version (suffix array) of genome (faster, but more RAM required!). Claims better sensitivity and specificity than previous short read aligners, and 50x speed vs TopHat2. Requires ~27 GB RAM for human genome!

Kallisto

A pseudoaligner (mapper? see Pachter *blog post*) that compares read k-mers (all overlapping subsequences) to a *transcriptome de Bruijn graph (T-DBG)* to find transcripts compatible with the read. Also uses expectation maximization (EM) and bootrapping to determine most likely transcript and abundance uncertainty.

See *sleuth* for DGE analysis.

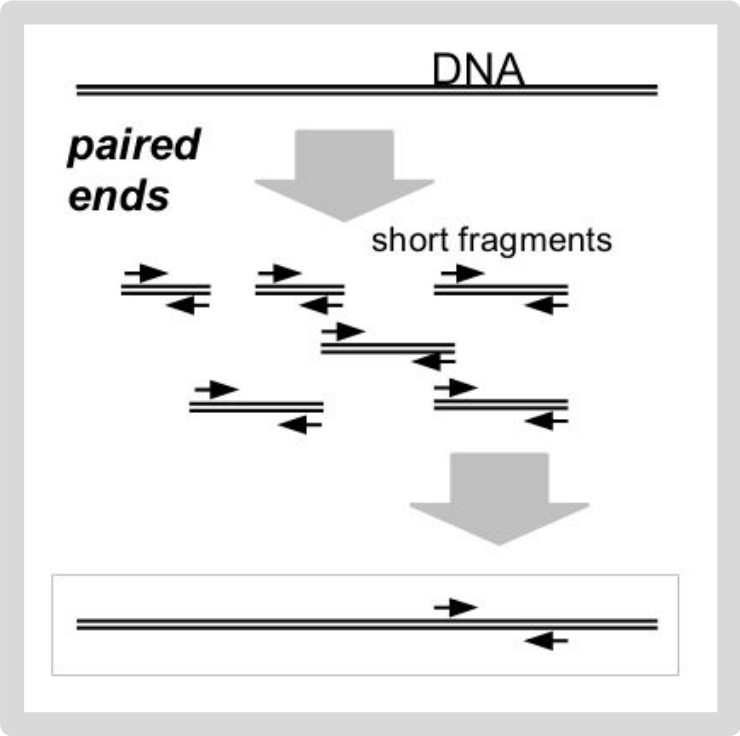


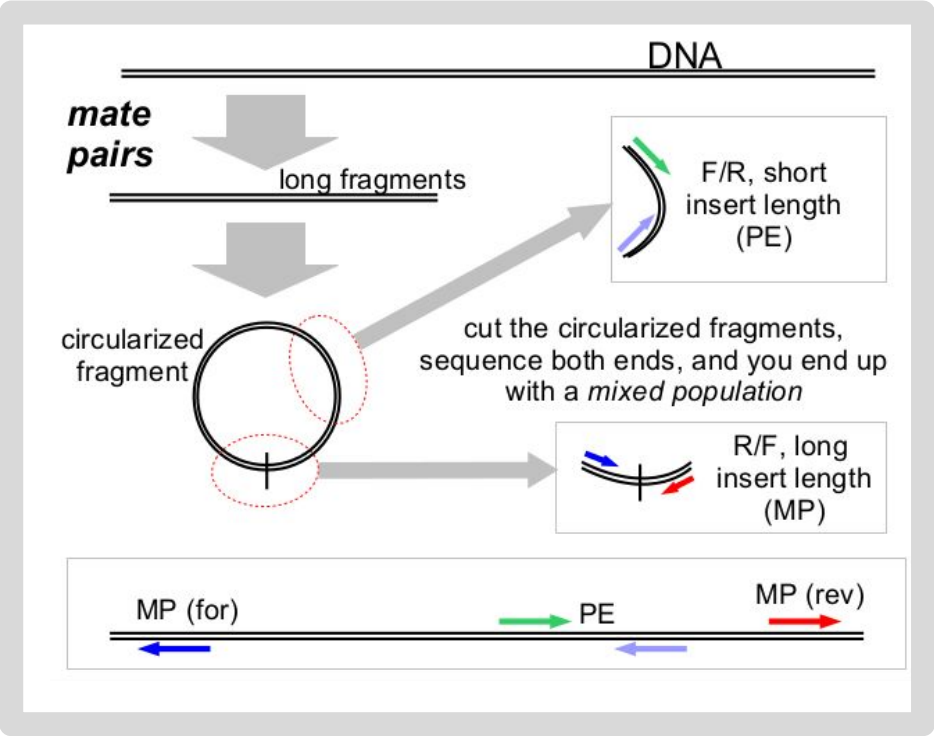
http://pachterlab.github.io/kallisto/
http://pachterlab.github.io/sleuth/

Salmon

Performs *quasi-mapping* (yeeesh, another term) to a set of transcripts (not genome), similar in methods to kallisto.

General Alignment Parameters / Concepts





General Alignment Parameters / Concepts

Edit Distance:

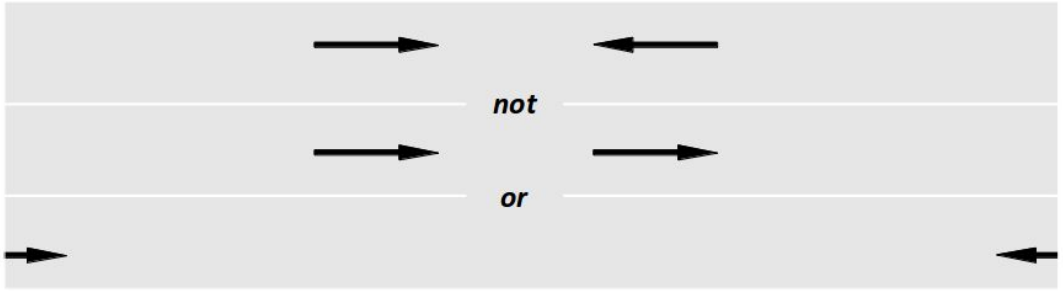
```
ATCGACCGCGCTAA-TATTAGTC . . .
  CGACGGCGCTAACTATTA
```

edit distance = 2

Mapping Quality:

prob. of incorrect position = $10^{-MQ/10}$ ... *(BWA)*

Proper Pairs:



General Alignment Parameters / Concepts

Insert Size:

Insert Size (ISIZE)

Think: "inserted" between adapter sequences.

Inner Distance:

Inner Distance

Think: what's the length of the "inner" part of my DNA fragment, not accessed directly by reads?

General Alignment Parameters / Concepts

Multimappers:

Reads that align *equally well* to more than one reference location.

Generally, multimappers are discounted in variant detection, and often discounted (ignored) in counting applications (RNA-Seq).

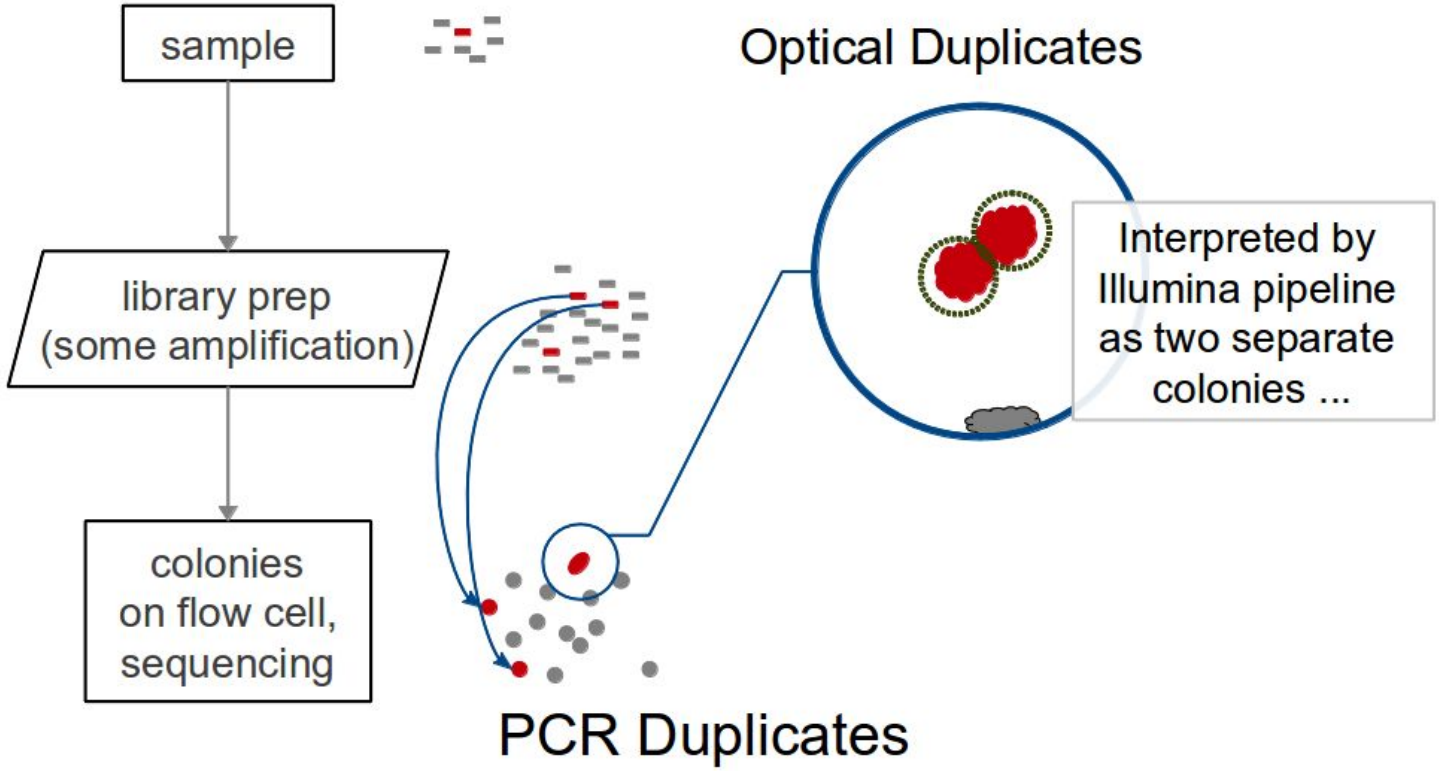
Note: *multimapper "rescue"* in some algorithms.

Duplicates:

Reads or read pairs arising from the same library fragment, either during library preparation (PCR duplicates) or colony formation (optical duplicates).

Generally, duplicates are discounted in variant detection , and ignored in counting applications (RNA-Seq).

General Alignment Parameters / Concepts



General Alignment Parameters / Concepts

Edit distance

Mapping Quality

Proper pairs

Insert size (template length) versus “inner distance”

Multimappers

Duplicates (don’t try de-duping without PE … applies in RNA-Seq as well … e.g. with phiX, 5k positions, but adding second read results in millions of possible fragments) … PCR versus optical (defunct)

General Alignment Parameters / Concepts

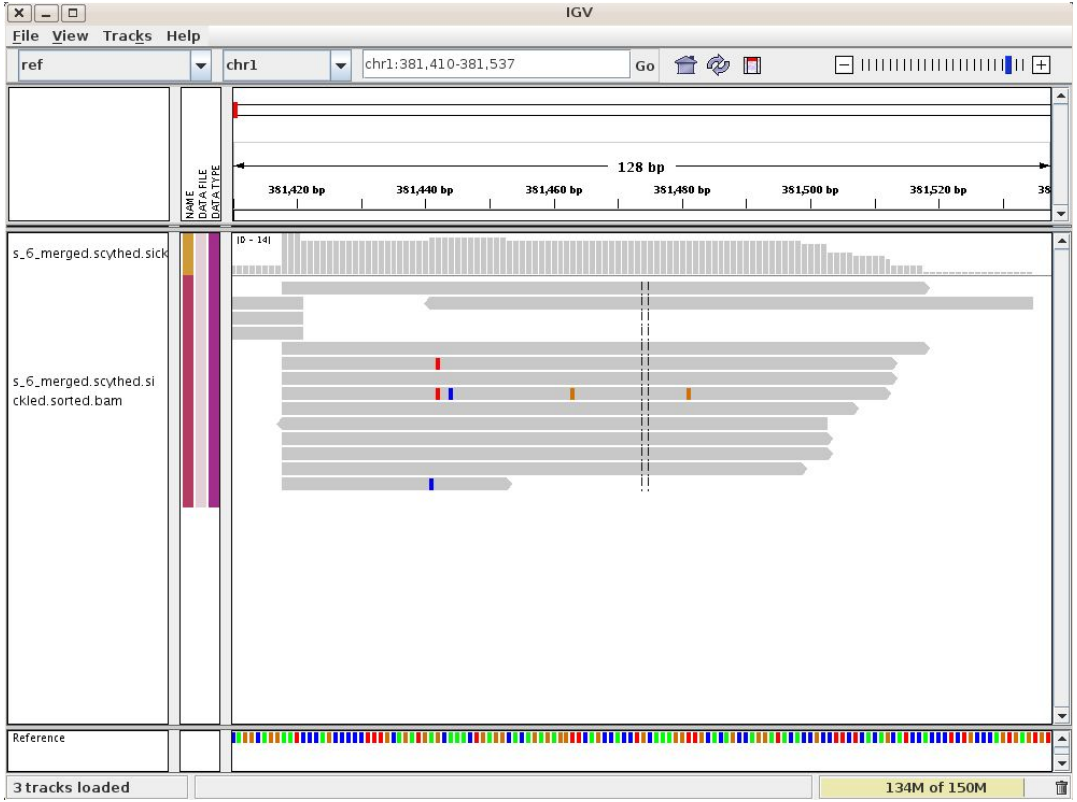
Splice junction inference?

Splice junction track?

Clipping. H / S. … yes, need for SAM slides

Alignment Viewers

- IGV (Integrated Genomics Viewer)
  - www.broadinstitute.org/igv/
- BAMview, tview (in SAMtools), IGB, GenomeView, SAMscope ...
- UCSC Genome Browser, GBrowse



IGV

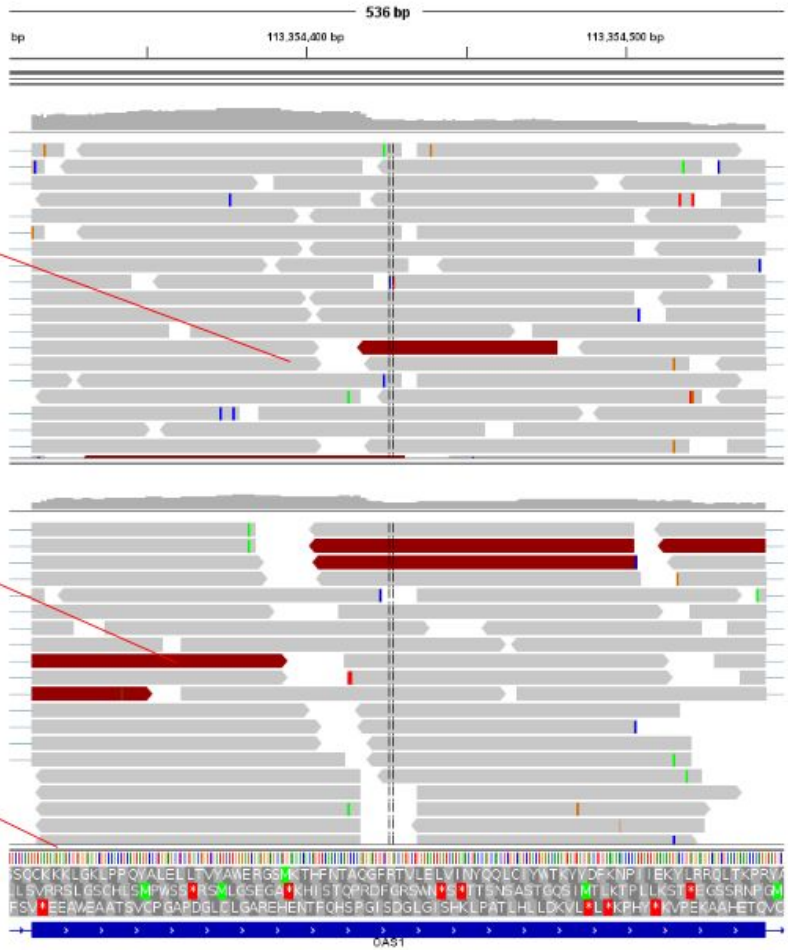

red box indicates region of reference in view below
16 kb
coverage track: read coverage depth plot
read alignments: (various view styles - "squished" shown here) … read positions, orientations, pairing, sequence that disagrees with reference highlighted, improper pairs highlighted, etc.
annotation tracks (GTF, BED, etc.)

IGV

colored bases where they disagree with reference (substitution, indel, etc.)

improper pairs (mate aligns far away, in wrong orientation, or on another chromosome)

reference sequence, reading frames, etc.



IGV

More on IGV's interface, file formats, and display can be found here:

http://www.broadinstitute.org/igv/AlignmentData

More on interpreting and customizing IGV's display can be found here:

http://www.broadinstitute.org/software/igv/interpreting_insert_size