Aligners

J Fass | 20 June 2018

Definitions

Assembly:

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

Definitions

Alignment:

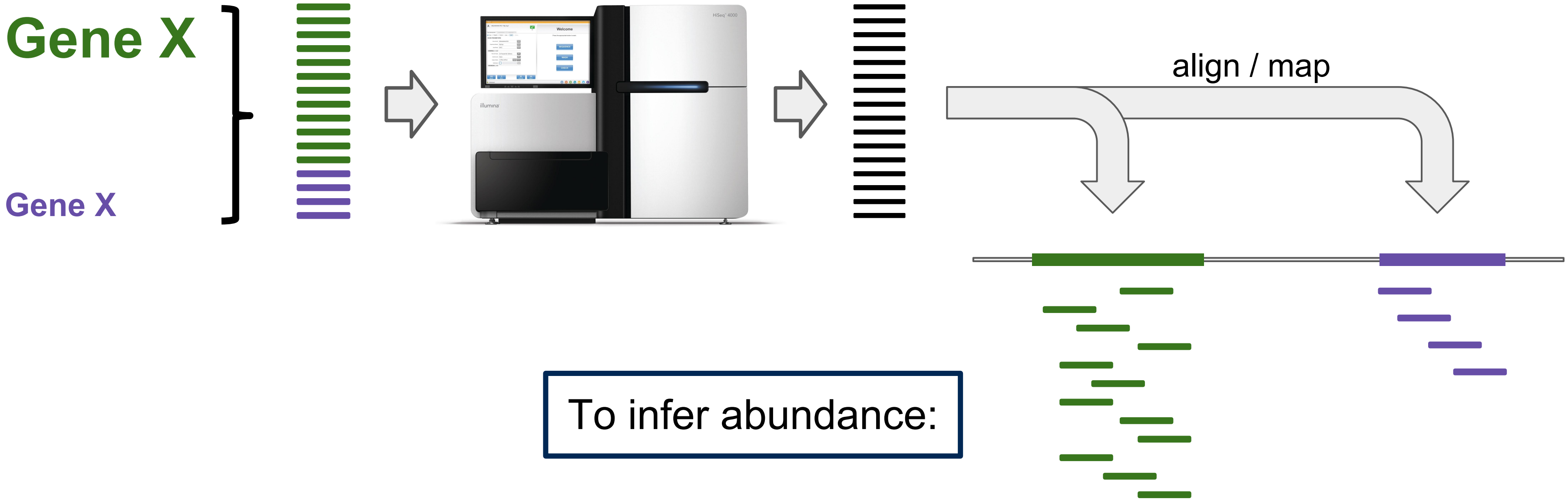
Somebody plagiarized parts of my document; where did they copy paragraphs from and where were each of the words and letters copied (perhaps with mistakes or changes) from?

Definitions

***Mapping*:**

Somebody plagiarized parts of my document; where did they copy paragraphs from ~~and where were each of the words and letters copied (perhaps with mistakes or changes) from~~?

Why align (or map)?



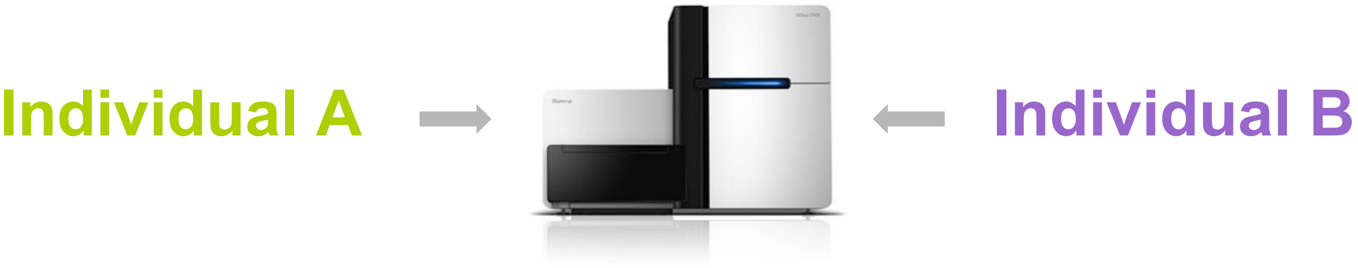
Gene X

Gene X

align / map

To infer abundance:

Why align (~~or map~~)?



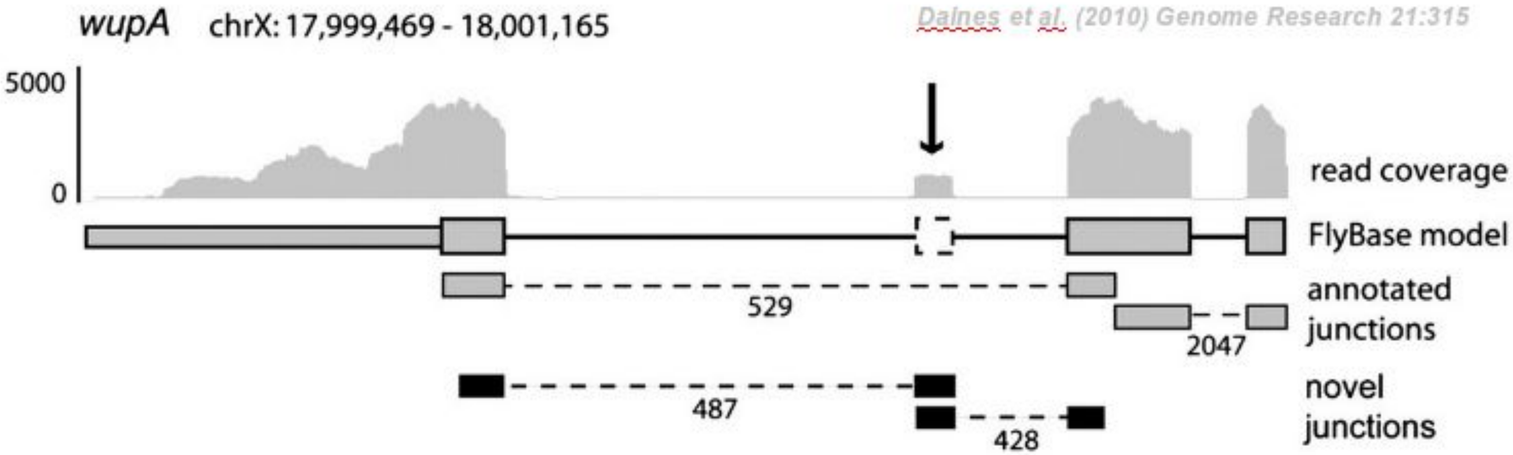
Individual A

Individual B

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

To measure variation

Why align (~~or map~~)?



Daines et al. (2010) Genome Research 21:315

To discover transcribed sequence.

More Definitions: “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 … ?”

Early short read aligners generally assumed 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 = “***glocal***”.

Modern aligners are generally ***local***, with respect to both read and reference, which generally allows them to ignore poor alignment in low quality 3’ tails and/or adapter and linker sequences. Do we need to trim anymore?

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”

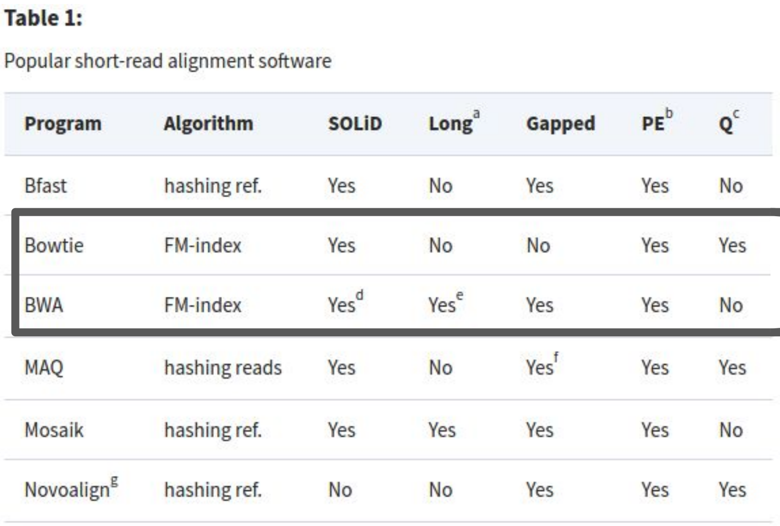


Table 1:

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-Wheeler 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.

Tophat2

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).

“Please stop using Tophat” -- *its author (PI)*

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!

HISAT2

Improved replacement for Tophat2, multiple high level and low level indexes of compressed sequence (similar to graph-based assemblers using De Bruijn graphs to represent overlapping k-mers). Graph structure can represent multiple sequence paths, i.e. a population of references, allowing rapid genotyping of samples.

HISAT2 is currently probably the best combination, *in a full aligner*, of speed (faster than STAR) and memory (~4 GB for human genome).

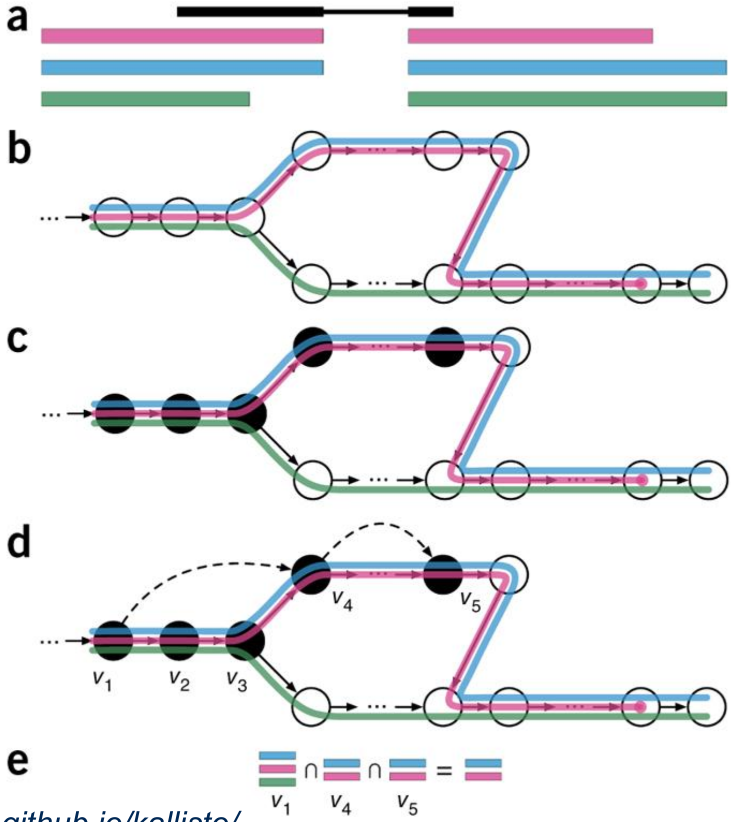
Kim, Langmead, Salzberg (2015) *Nature Methods* 12:357

Kallisto

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

Bray … Pachter (2016) *Nat Biotech* 34:525

See [sleuth] for DGE analysis.



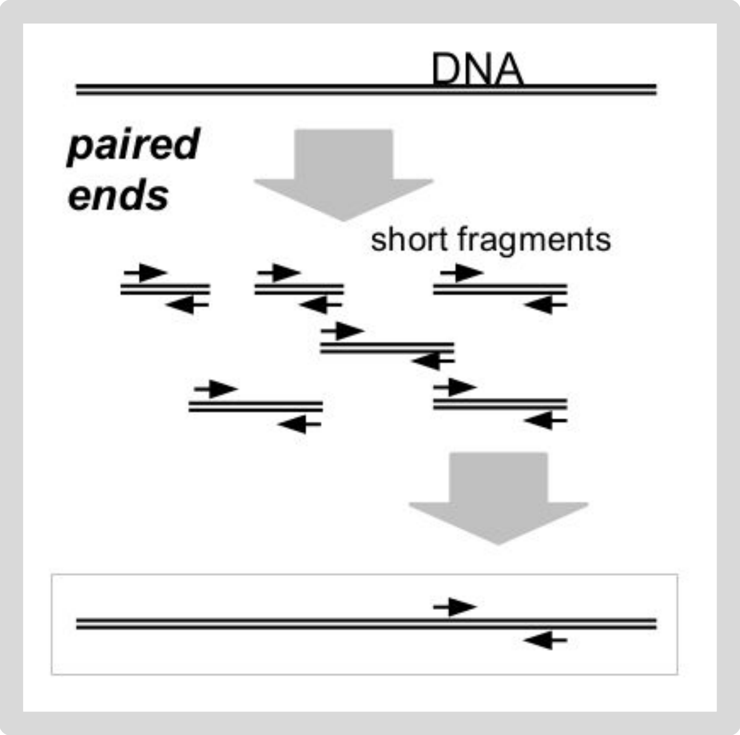
http://pachterlab.github.io/kallisto/
http://pachterlab.github.io/sleuth/

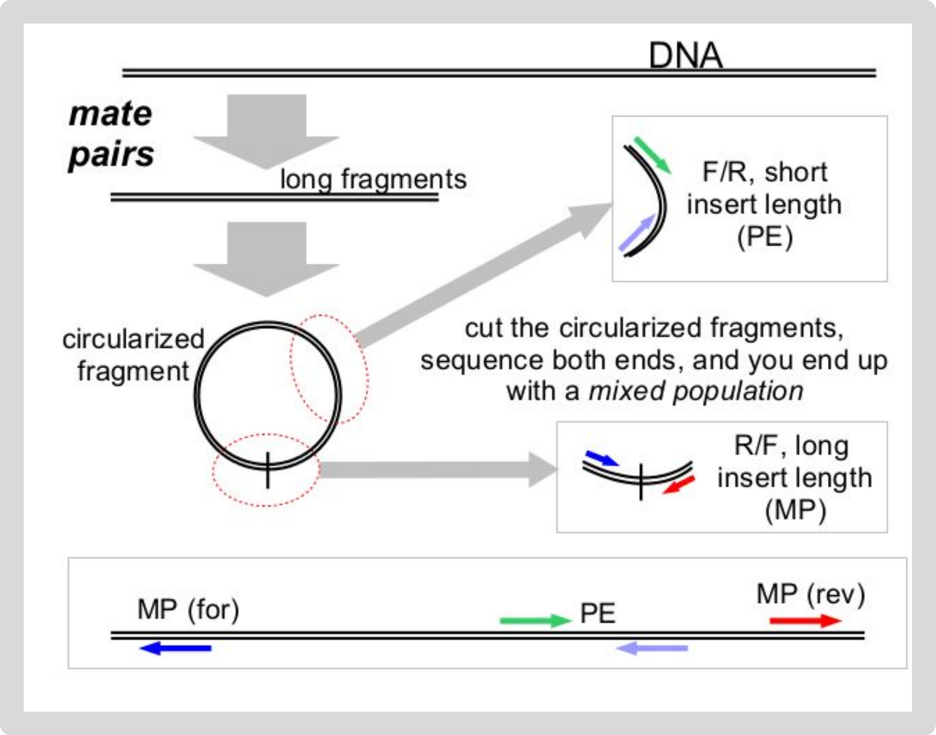
Salmon

Supersedes Sailfish, an older k-mer based transcript analysis tool. Performs *quasi-mapping* to a set of transcripts (not genome), similar in methods to kallisto. Also performs bias correction for multiple modes of bias (sequence, position) to more accurately determine abundance.

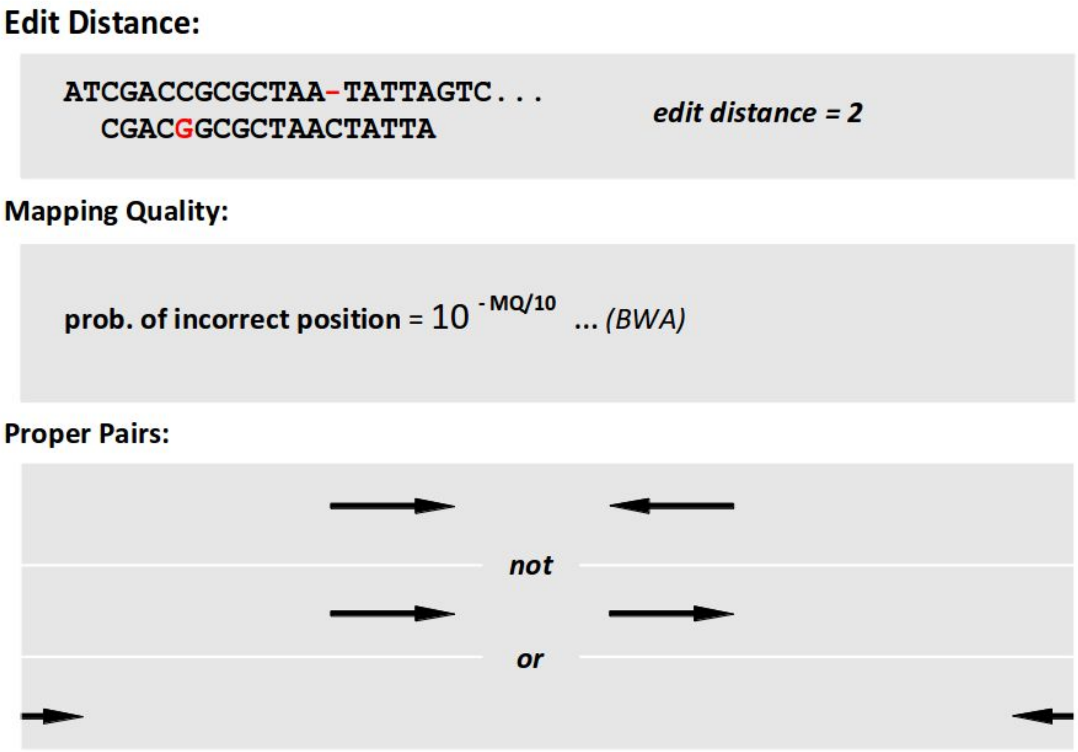
Patro, Duggard, Kingsford (2015) *bioRxiv* [*http://dx.doi.org/10.1101/021592*](http://dx.doi.org/10.1101/021592)

General Alignment Parameters / Concepts




DNA
mate pairs
long fragments
circularized fragment
F/R, short insert length (PE)
cut the circularized fragments, sequence both ends, and you end up with a *mixed population*
R/F, long insert length (MP)
MP (for)
PE
MP (rev)

General Alignment Parameters / Concepts



Edit Distance:

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

edit distance = 2

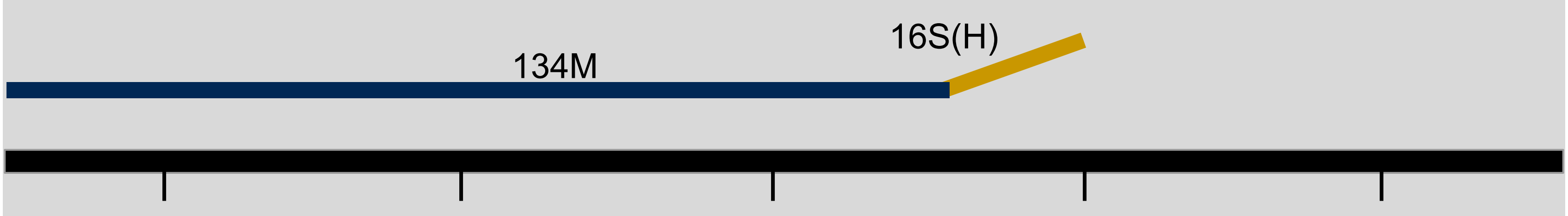
Mapping Quality:

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

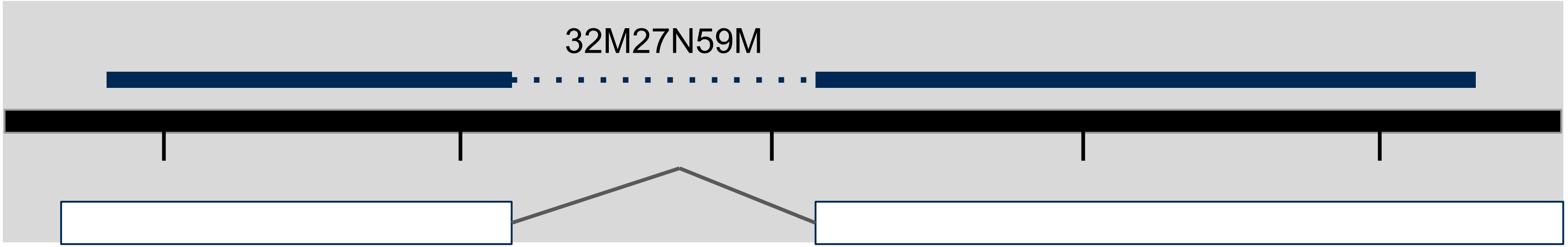
Proper Pairs:

General Alignment Parameters / Concepts

Clipping:

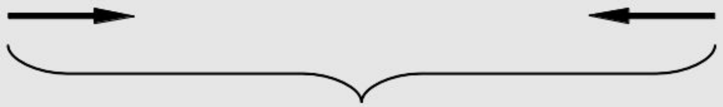


Splicing:



General Alignment Parameters / Concepts

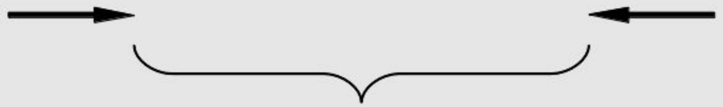
Insert Size:



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 are often discounted in counting applications (like RNA-Seq … would "cancel" out anyway).

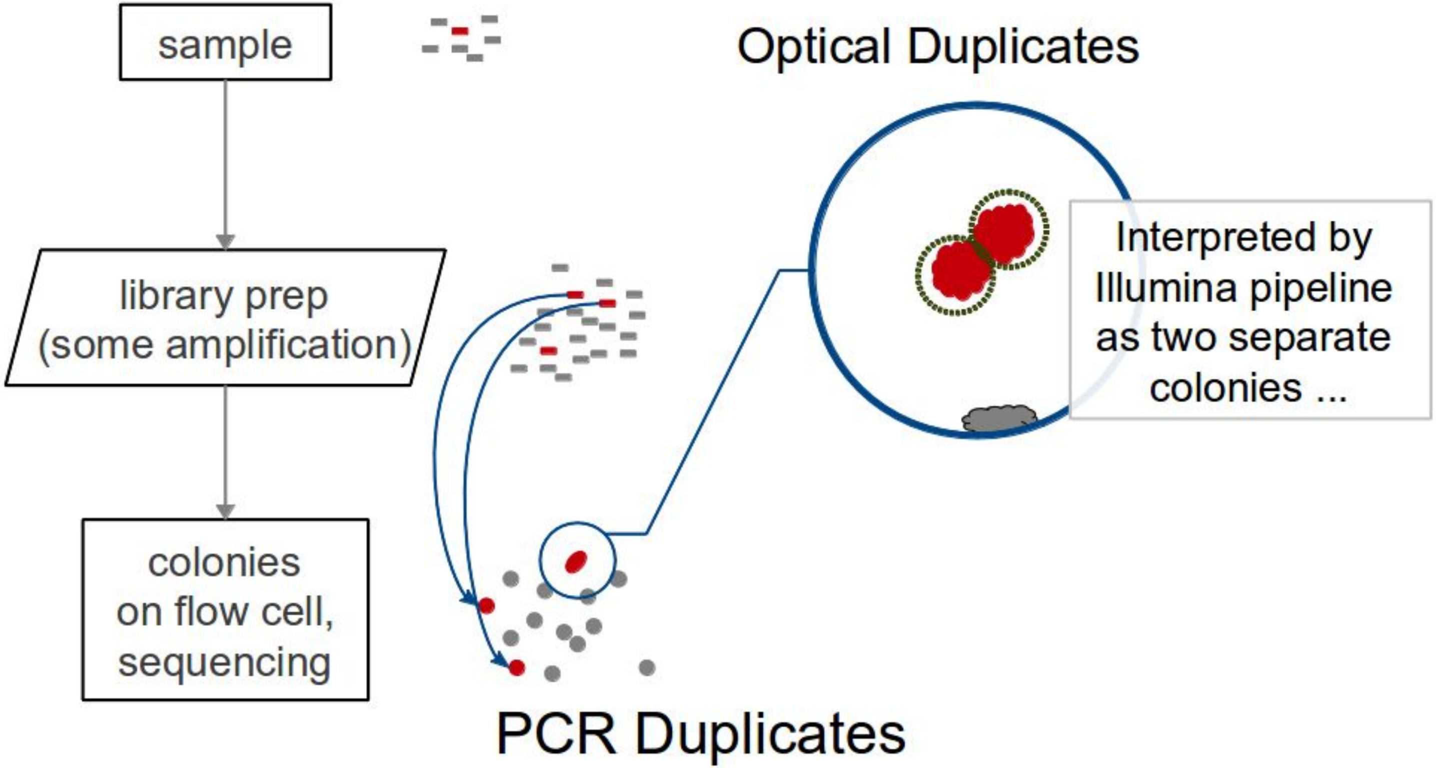
Note: *multimapper "rescue"* in some algorithms (RSEM, Express?).

Duplicates:

Reads or read pairs arising from the same original library fragment, either during library preparation (PCR duplicates) or colony formation (optical duplicates; not an issue anymore).

Generally, duplicates can only be detected reliably with paired-end sequencing. If PE, they're discounted in variant detection, and discounted in counting applications (like RNA-Seq).

General Alignment Parameters / Concepts



sample

library prep (some amplification)

colonies on flow cell, sequencing

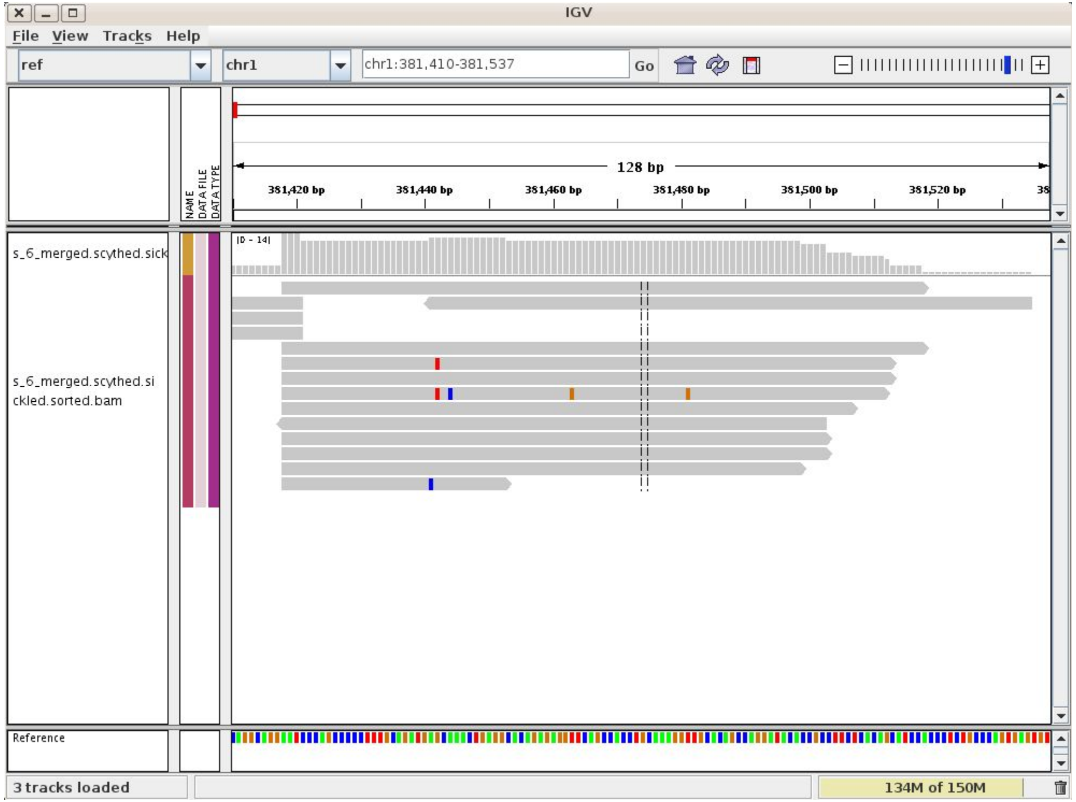
Optical Duplicates

Interpreted by Illumina pipeline as two separate colonies ...

PCR Duplicates

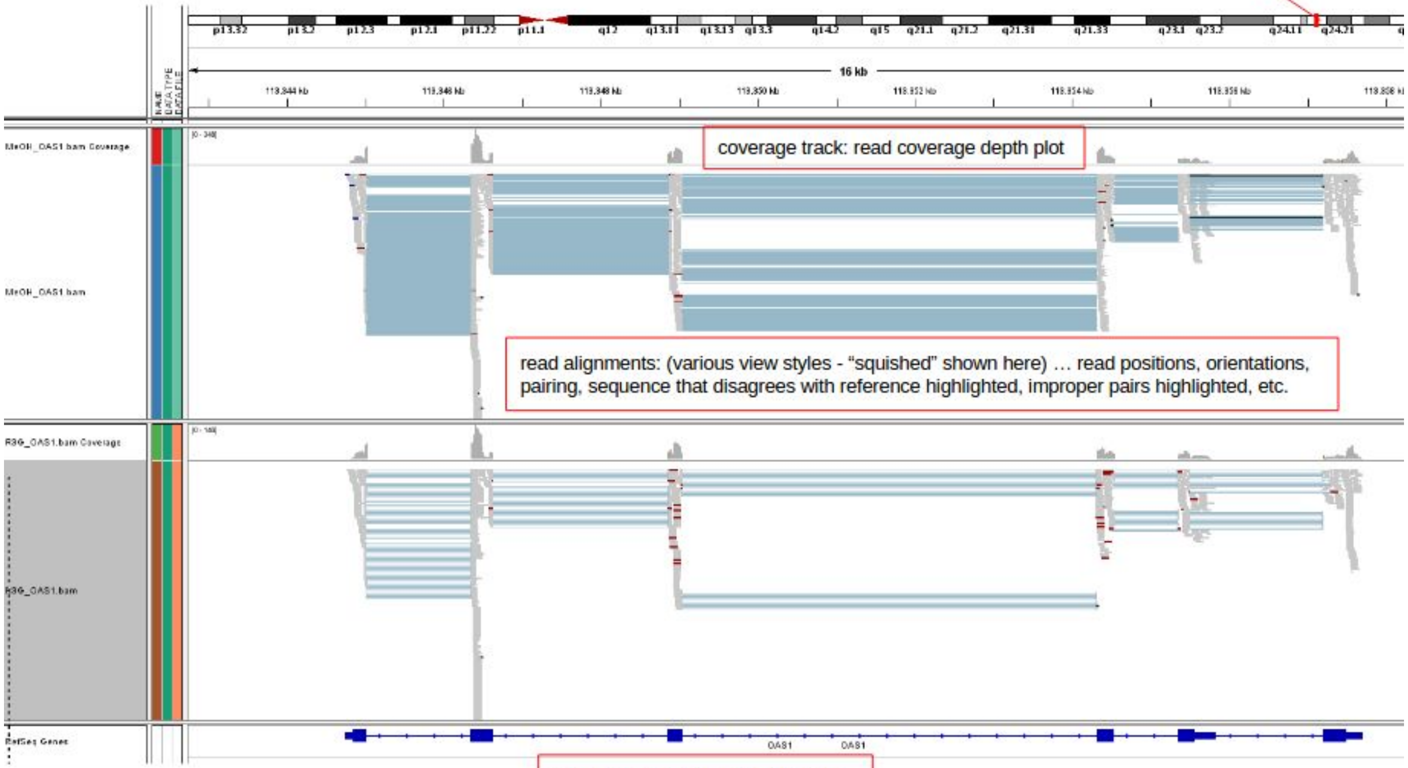
Alignment Viewers

- 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

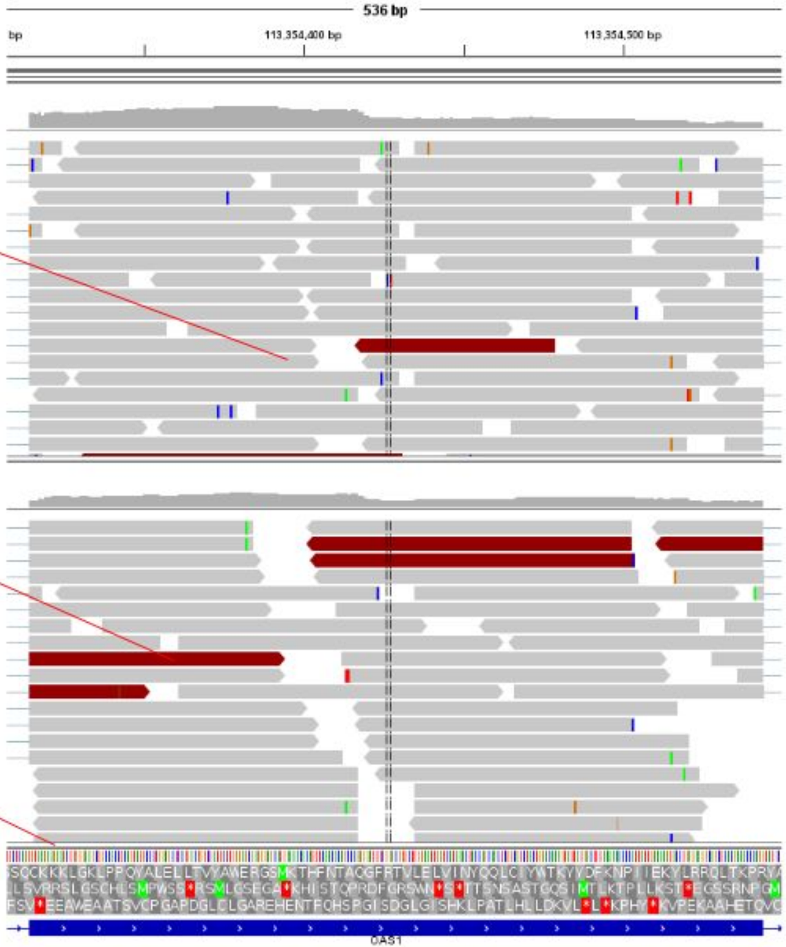


coverage track: read coverage depth plot

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