Aligners

J Fass | 23 August 2017

Definitions

Assembly:

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

Definitions

Alignment:

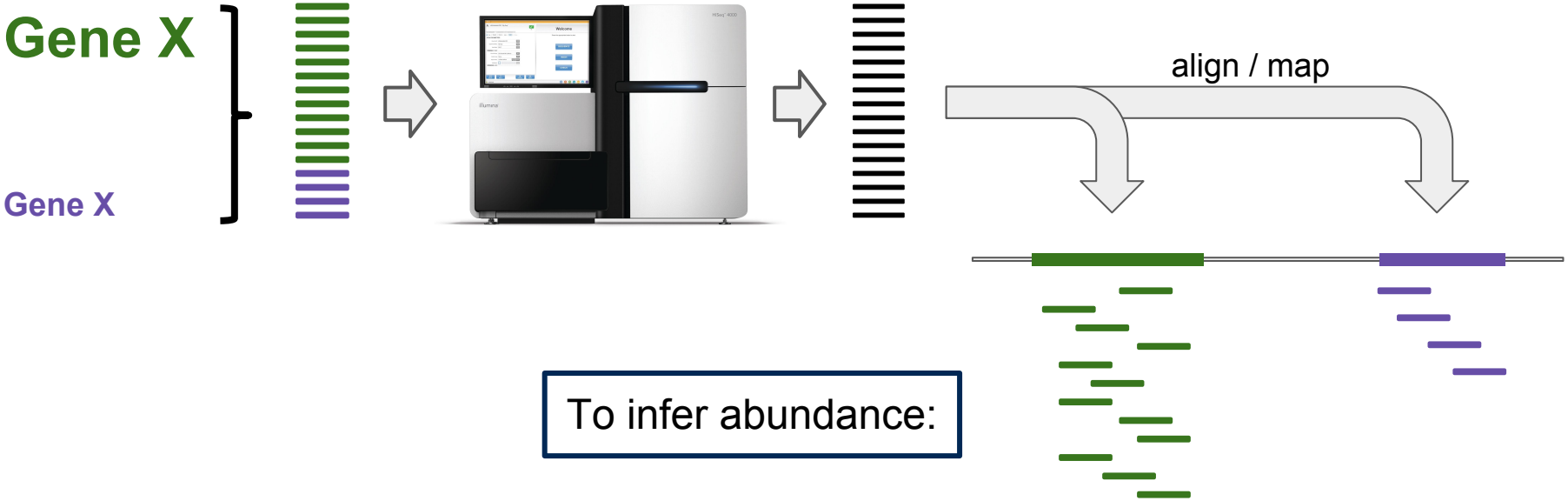
Somebody plagiarized parts of my document; where did they copy paragraphs from and where did each of the words come from?

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:

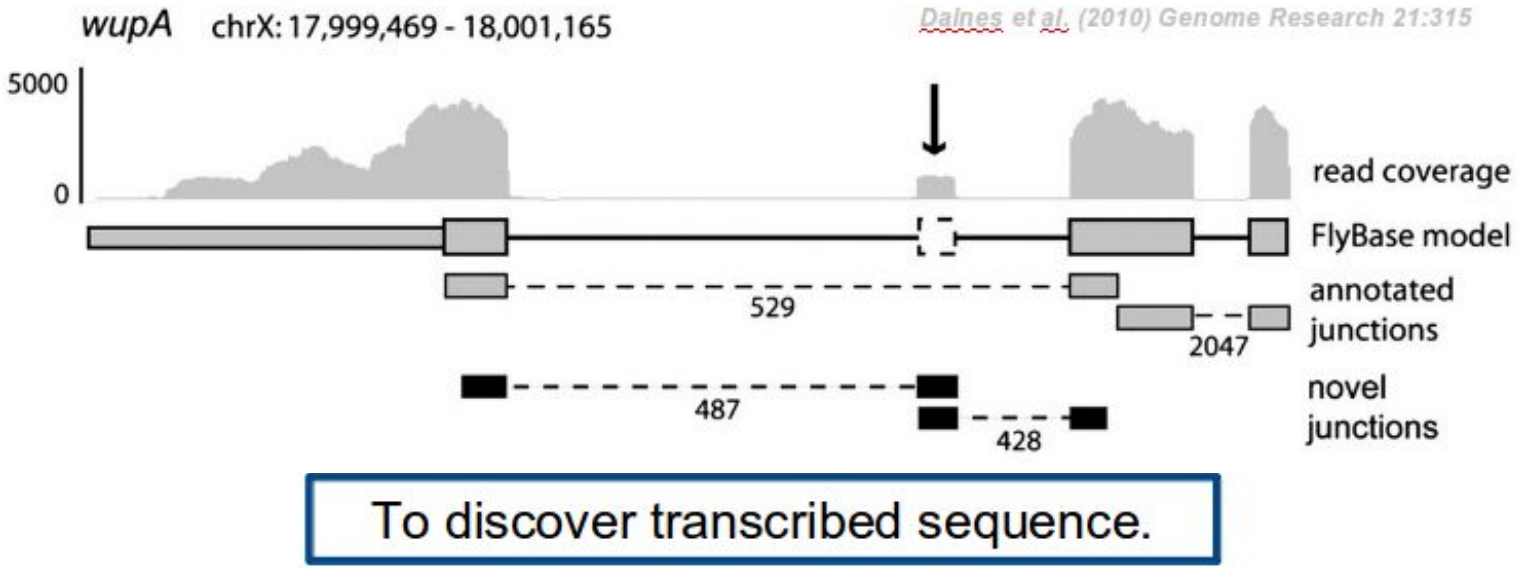
Why align (~~or map~~)?

Individual A →  ← **Individual B**

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

To measure variation

Why align (~~or map~~)?



To discover transcribed sequence.

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 … ?”

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-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.

Long Read Aligners

BLASR (**B**asic **L**ocal **A**lignment with **S**uccessive **R**efinement)

DALIGNER (of DAZZLER assembler; by Gene Myers, author of BLAST)

Minimap2 (by Heng Li, author of BWA)

GraphMap (nanopore read aligner; even tougher than PacBio)

NGM-LR (co**N**vex **G**ap-cost align**M**ents for **L**ong **R**eads)

Long Read Aligners

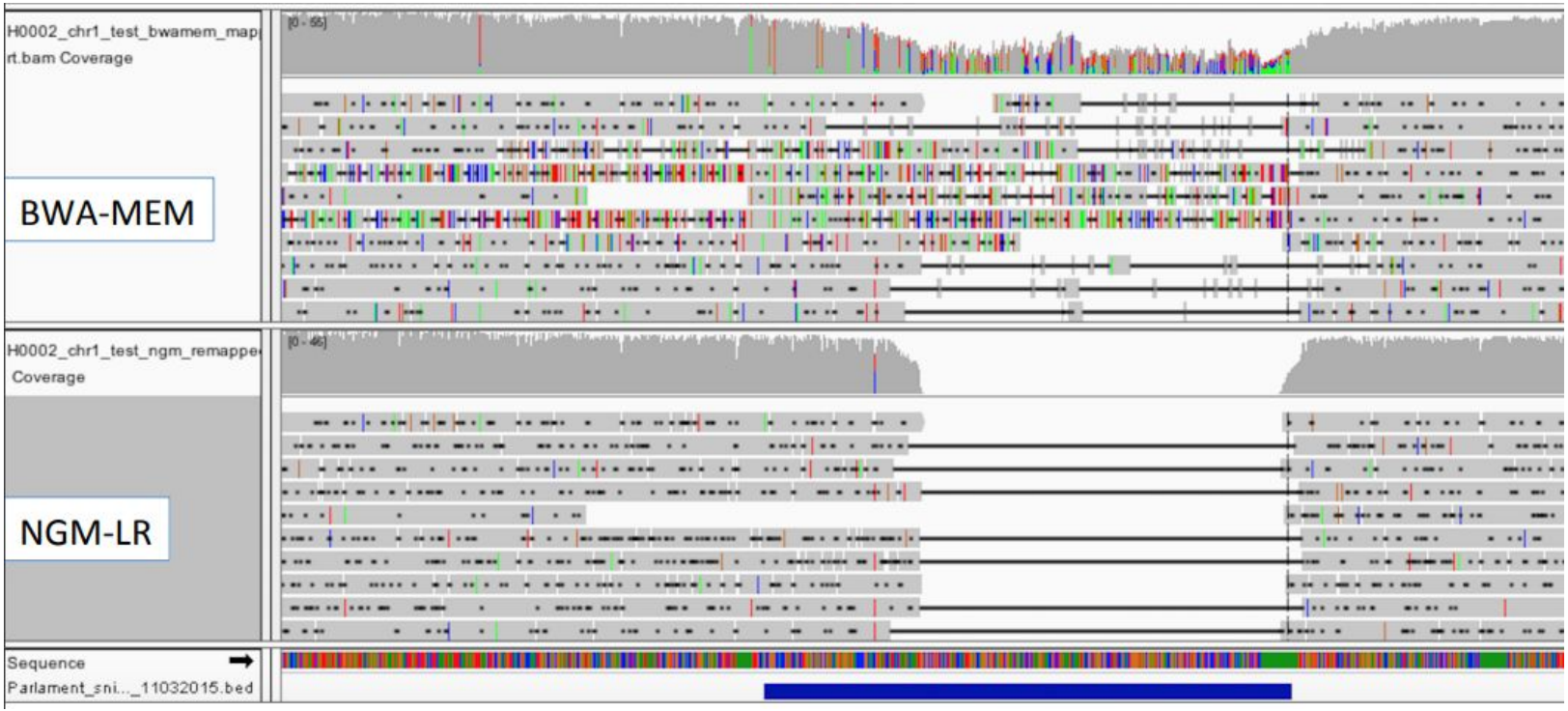
NGM-LR (co**N**vex **G**ap-cost align**M**ents for **L**ong **R**eads)

“linear” … gap open / extension cost is equal and constant

“affine” … gap open cost ≠ gap extension cost

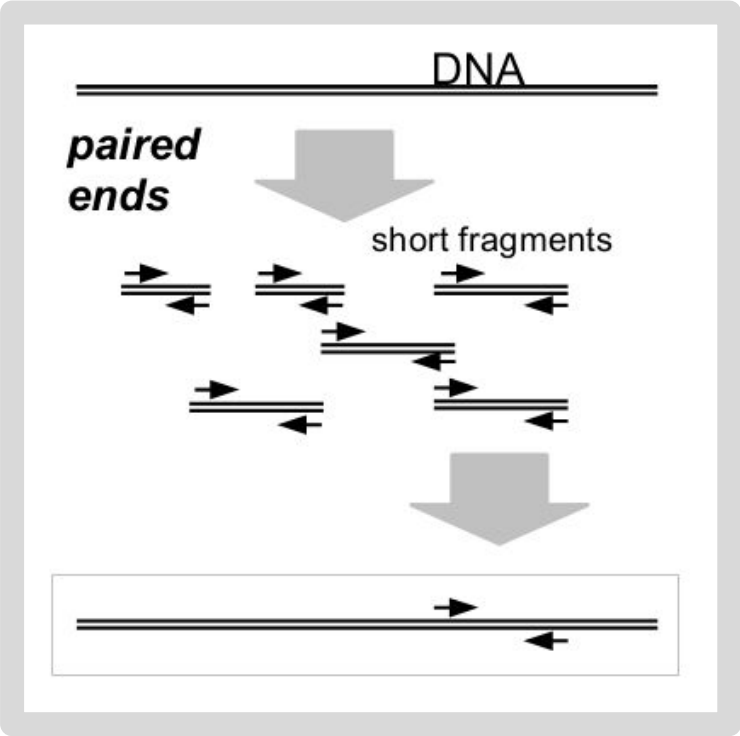
“convex” … similar to affine, but extension cost decreases with gap size

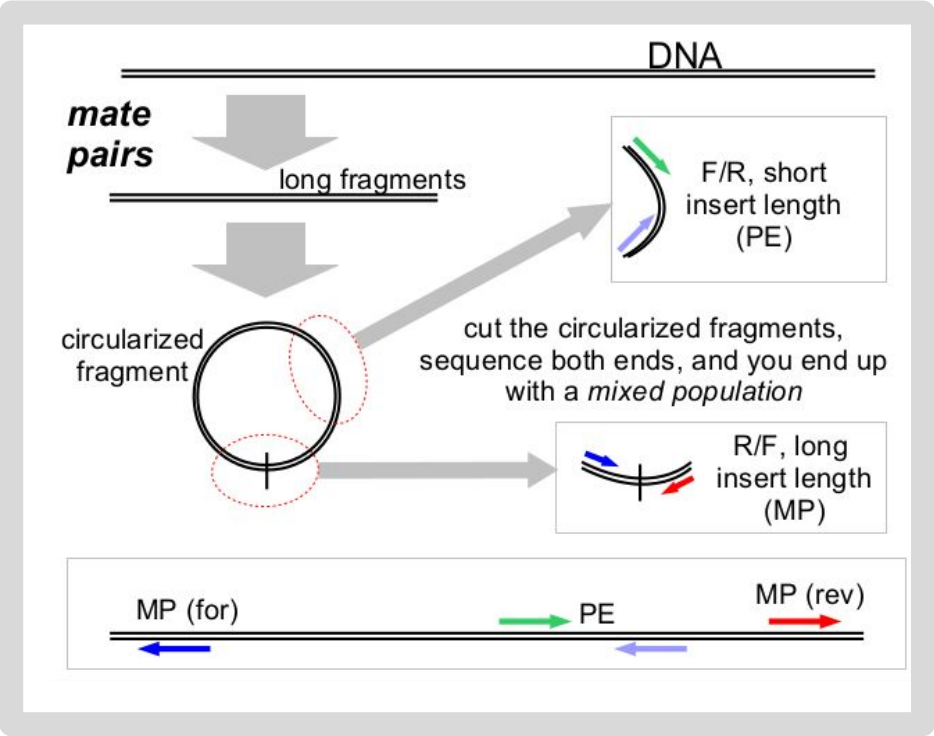
Long Read Aligner - NGM-LR



talk by Fritz Sedlazeck at Biological Data Science conf., CSHL, 2016

General Alignment Parameters / Concepts





General Alignment Parameters / Concepts

Edit Distance:

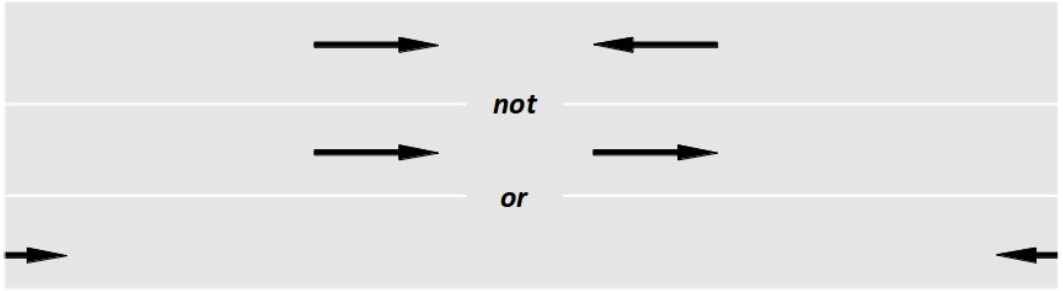
```
ATCGACCGCGCTAA-TATTAGTC . . .
  CGACGGCGCTAACTATTA
```

edit distance = 2

Mapping Quality:

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

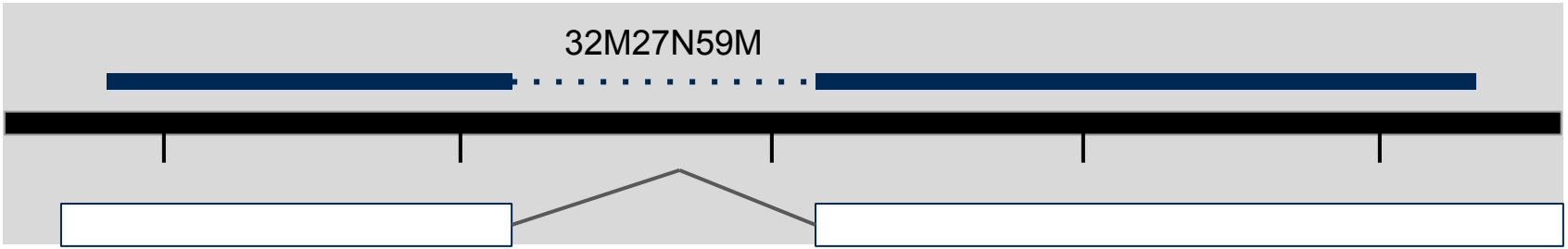
Proper Pairs:



General Alignment Parameters / Concepts

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).

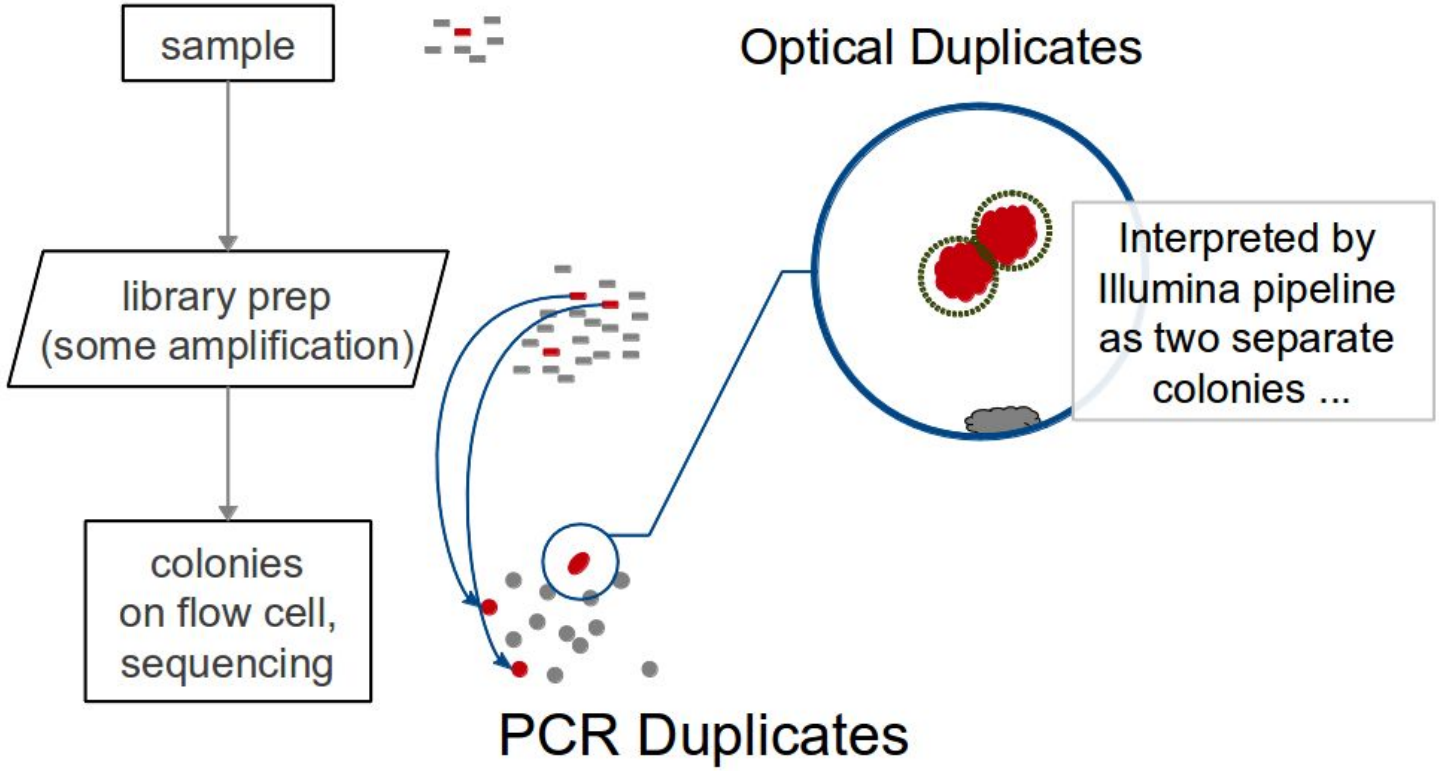
Note: *multimapper "rescue"* in some algorithms (RSEM, Express?).

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).

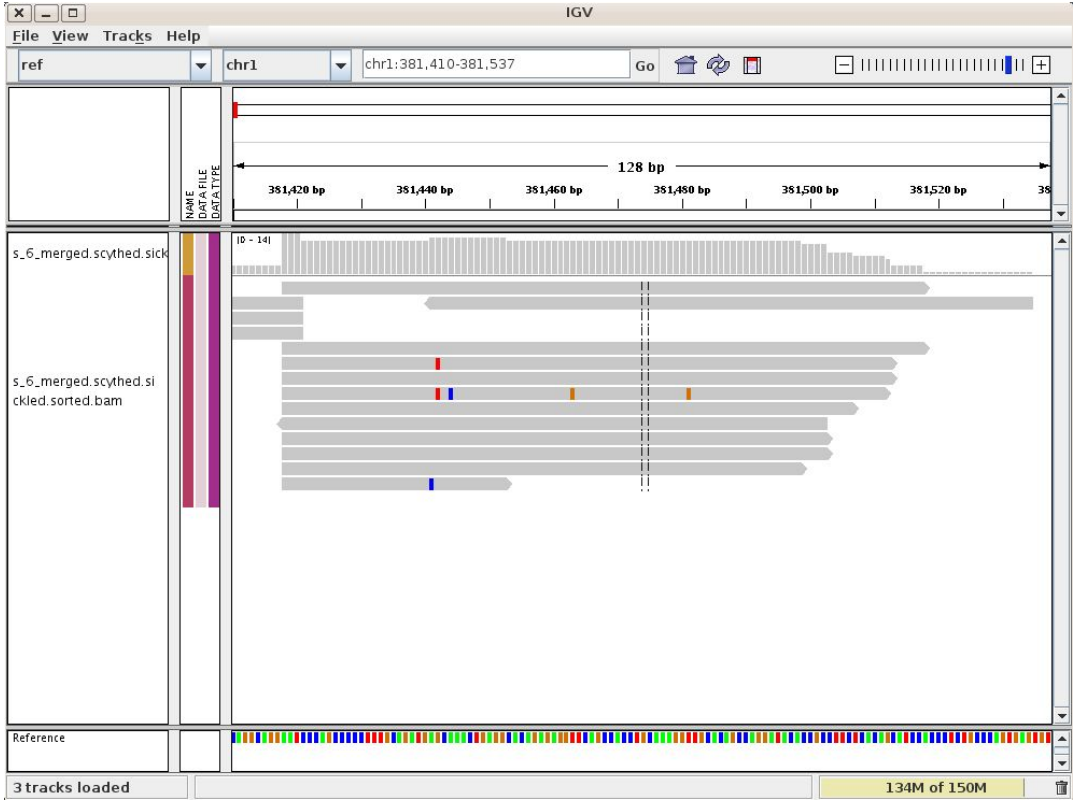
Generally, duplicates can only be detected reliably with paired-end sequencing. If PE, they're discounted in variant detection, and discounted in counting applications (like RNA-Seq).

General Alignment Parameters / Concepts



Alignment Viewers

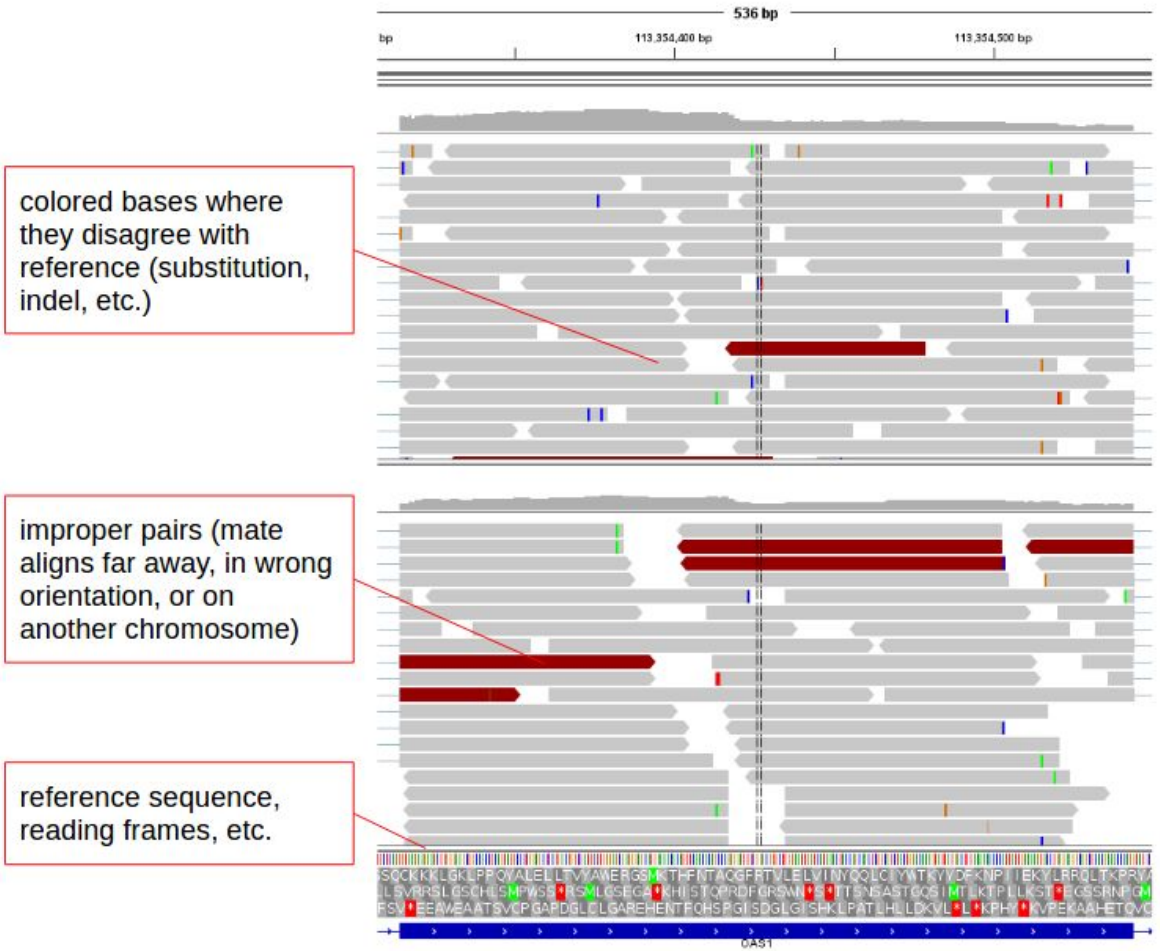
- IGV (Integrated Genomics Viewer)
  - www.broadinstitute.org/igv/
- BAMview, tview (in SAMtools), IGB, GenomeView, SAMscope ...
- UCSC Genome Browser, GBrowse



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.
536 bp
bp
113,354,400 bp
113,354,500 bp
CAS1

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