Medizinische Genomanalysen

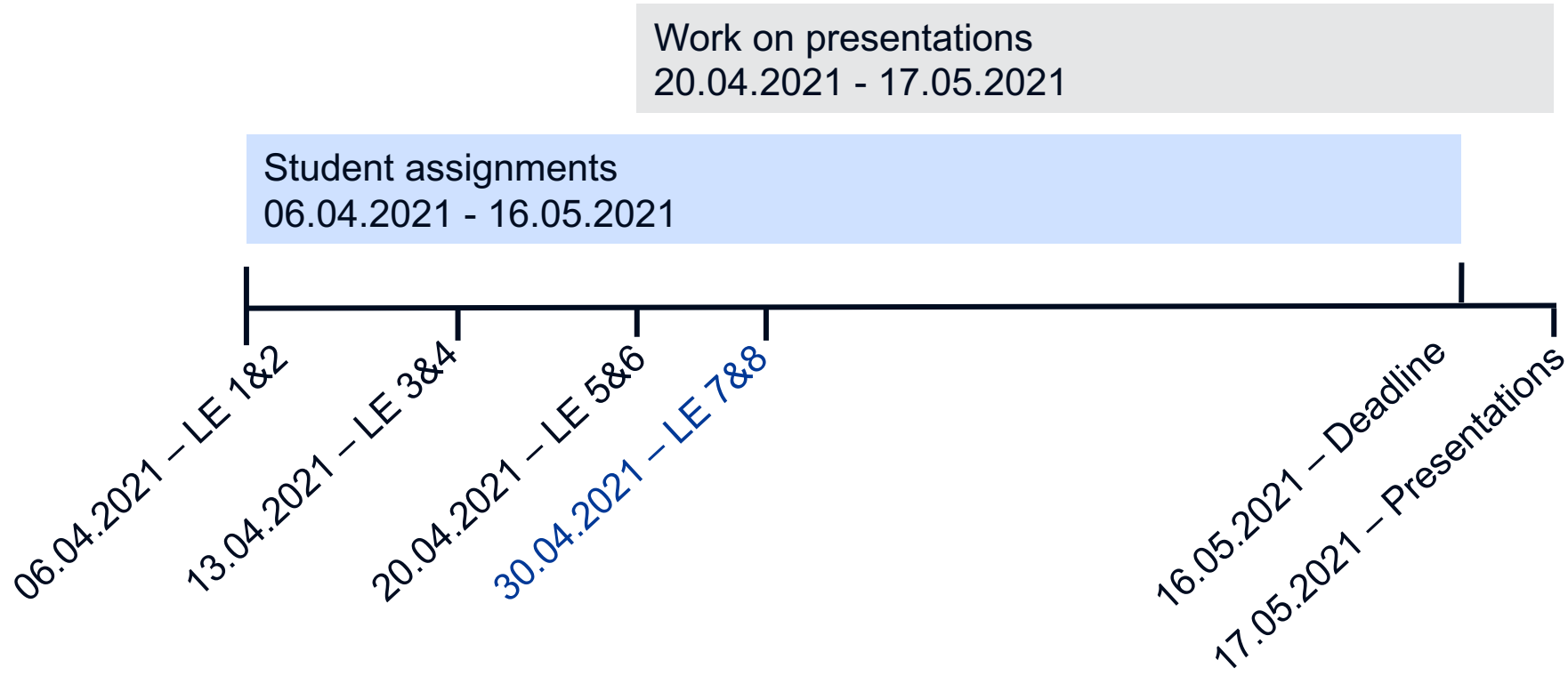
LE 3 (13.04.2021 – 18:00)

Stephan Pabinger
stephan.pabinger@gmail.com

Assignment 1

Status update

Timeline


Work on presentations
20.04.2021 - 17.05.2021
Student assignments
06.04.2021 - 16.05.2021
06.04.2021 – LE 1&2
13.04.2021 – LE 3&4
20.04.2021 – LE 5&6
30.04.2021 – LE 7&8
16.05.2021 – Deadline
17.05.2021 – Presentations

Recap

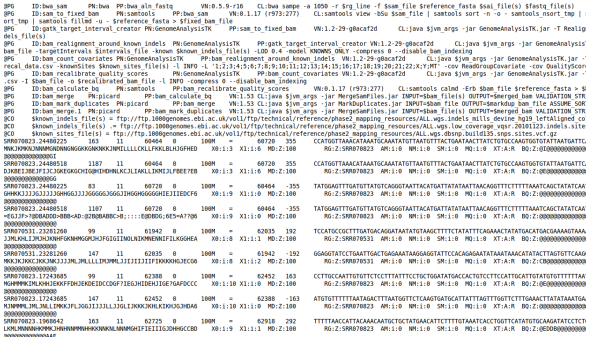
Recap

History

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology
1977	DNA sequencing
1985	*In Vitro* Amplification of DNA (PCR)
2001	The Human Genome Project

Different sequencing technologies

SAM/BAM files



Structure

LE 1

- History, terms, sequencing, SAM/BAM

LE 2

- Reference genome, FASTQ, cleaning

LE 3

- SAMtools, genetic variation, variant calling

LE 4

- Variant callers, structural variations & callers

LE 5

- CNVs, somatic mutations, filtering, annotation

LE 6

- Visualization, pipelines

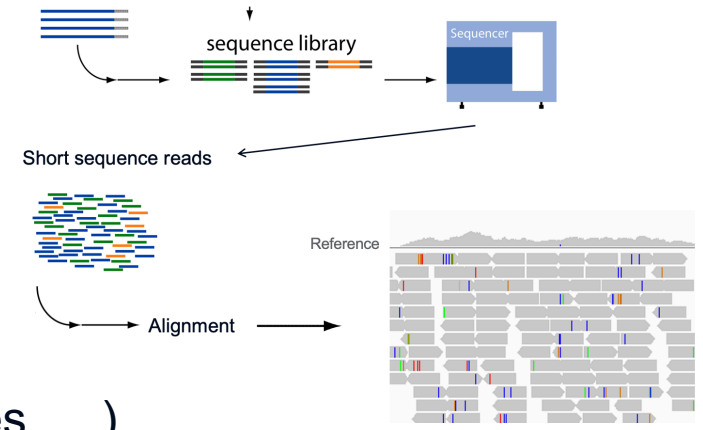
LE 7 & 8

- Albert Kriegner

LE 9 & 10

- Presentations

Cleaning up FASTQ files



Before mapping make sure

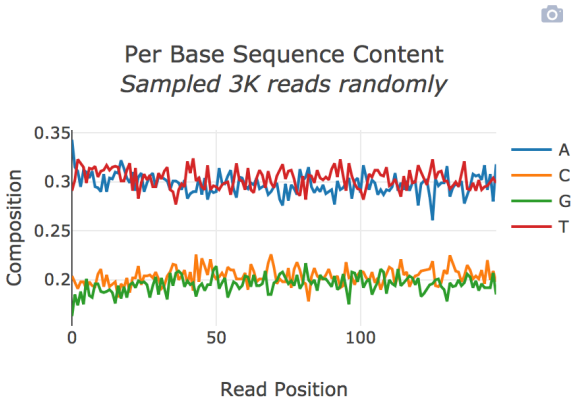
- Adapter sequences are removed
- Non genomic sequences are removed (barcodes, …)
- Clean contaminations (PRINSEQ, DeconSeq)
- Trim Ns
- Trim bad quality reads

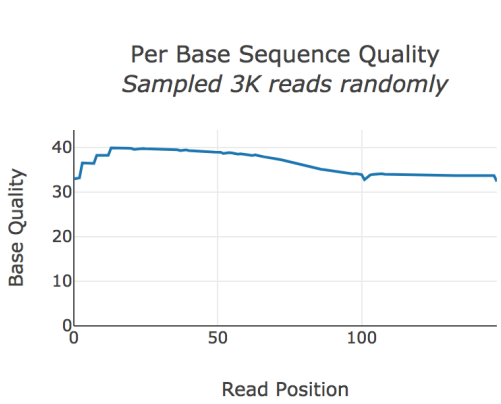
Skip cleaning → less read mapping; if not randomly distributed some areas won't get enough coverage

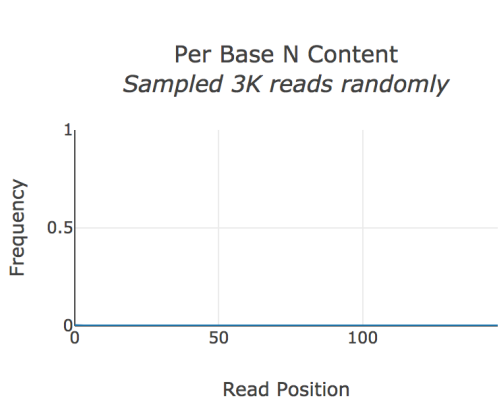
Tools for FASTQ manipulation

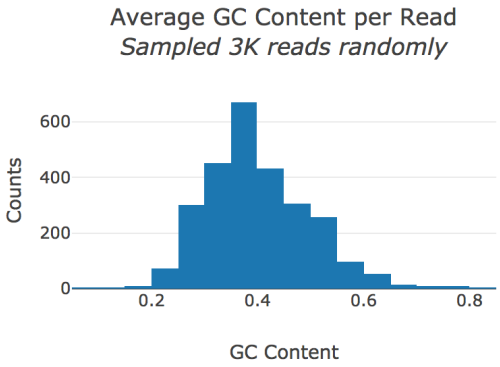
- FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  HTML output
- Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit/
  Lots of tools, charts, trimming, clipping, filtering
- Cutadapt https://code.google.com/p/cutadapt/
  Remove adapter sequences
- DeconSeq http://deconseq.sourceforge.net/
  User friendly interface, coverage plots, metagenomics datasets
- PRINSEQ: http://prinseq.sourceforge.net/
  HTML output, trimming, filtering, contaminations
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic
  Paird-end trimming
- Atropos https://github.com/jdidion/atropos
  Multi-threading, paired-end, bisulfite-seq, …
- ...

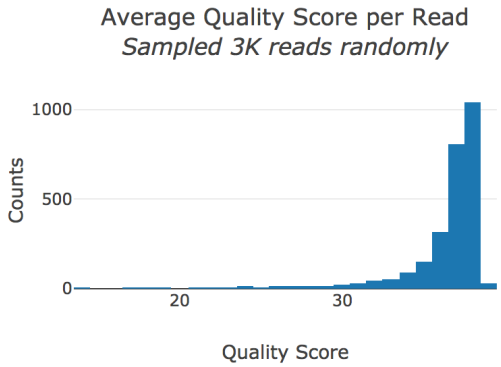
www.fastq.bio

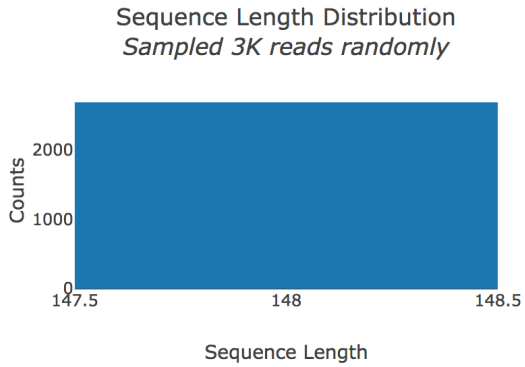








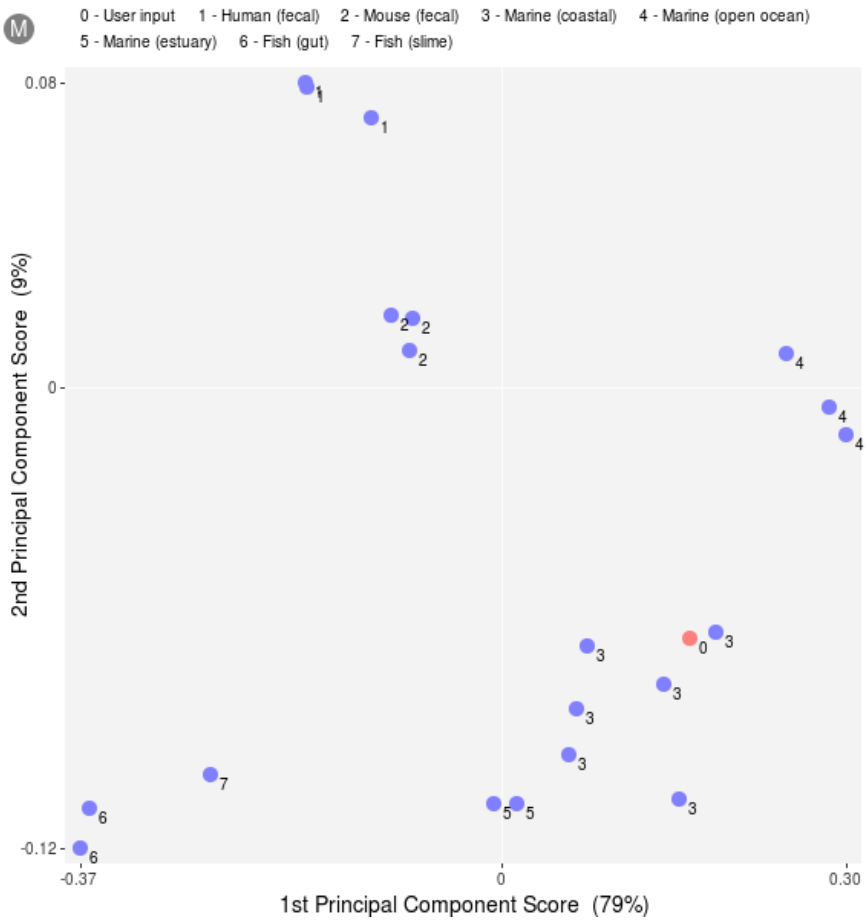




Check for contamination

PRINSEQ

- Dinucleotide (e.g., TA, GC, …) odds ratios
- Principal component analysis (PCA) to group metagenomes

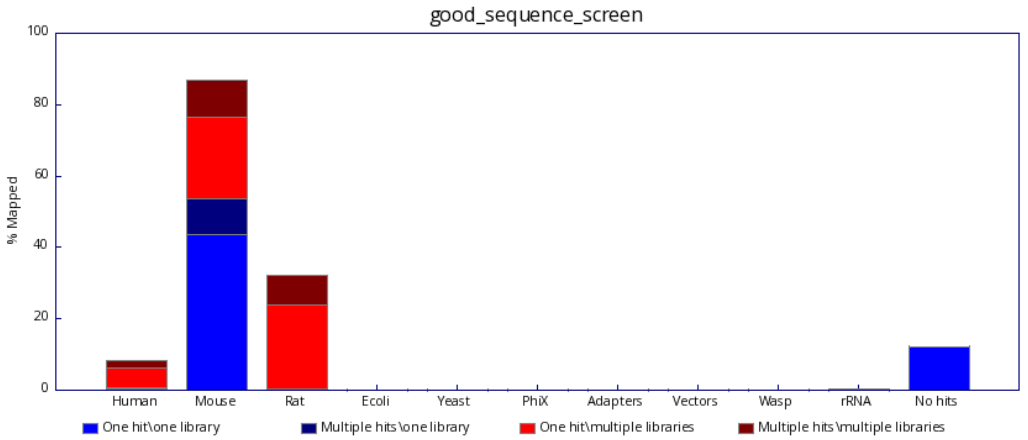


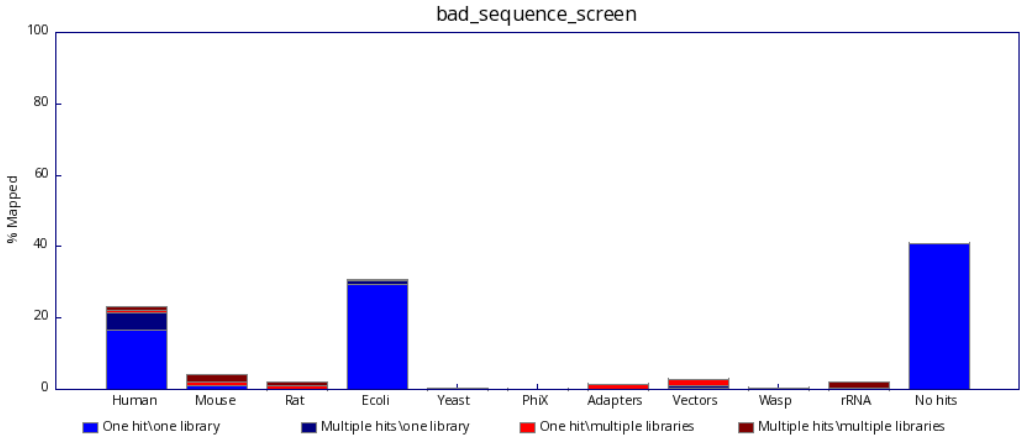
Check for contamination

FastQ Screen

Screen against a set of sequence databases:

- genomes of all of the organisms you work on
- PhiX
- Vectors
- …





Graph aligner

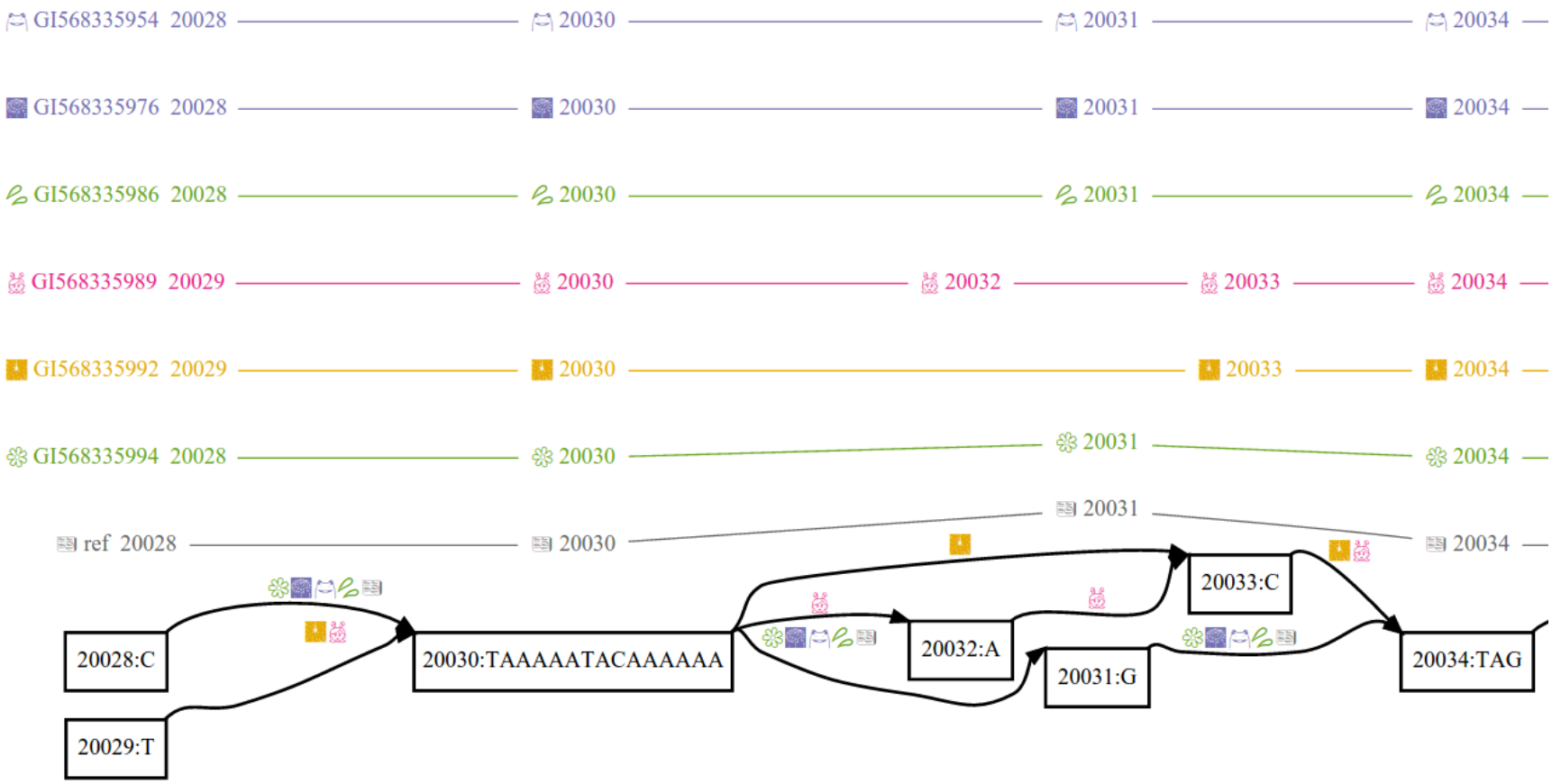
Linear references might lead to reference bias: a tendency to miss alignments or report incorrect alignments for reads containing non-reference alleles

vg: a toolkit of computational methods for creating, manipulating, and utilizing graph structures

- Full vg graph uses 3.92 GB when serialized to disk, and contains 3.181 Gbp of sequence
- Complete file sizes including indices vary from 25GB to 63GB
- https://github.com/vgteam/vg

Variation graph toolkit improves read mapping by representing genetic variation in the reference
Nat Biotechnol. 2018 Oct; 36(9): 875–879.

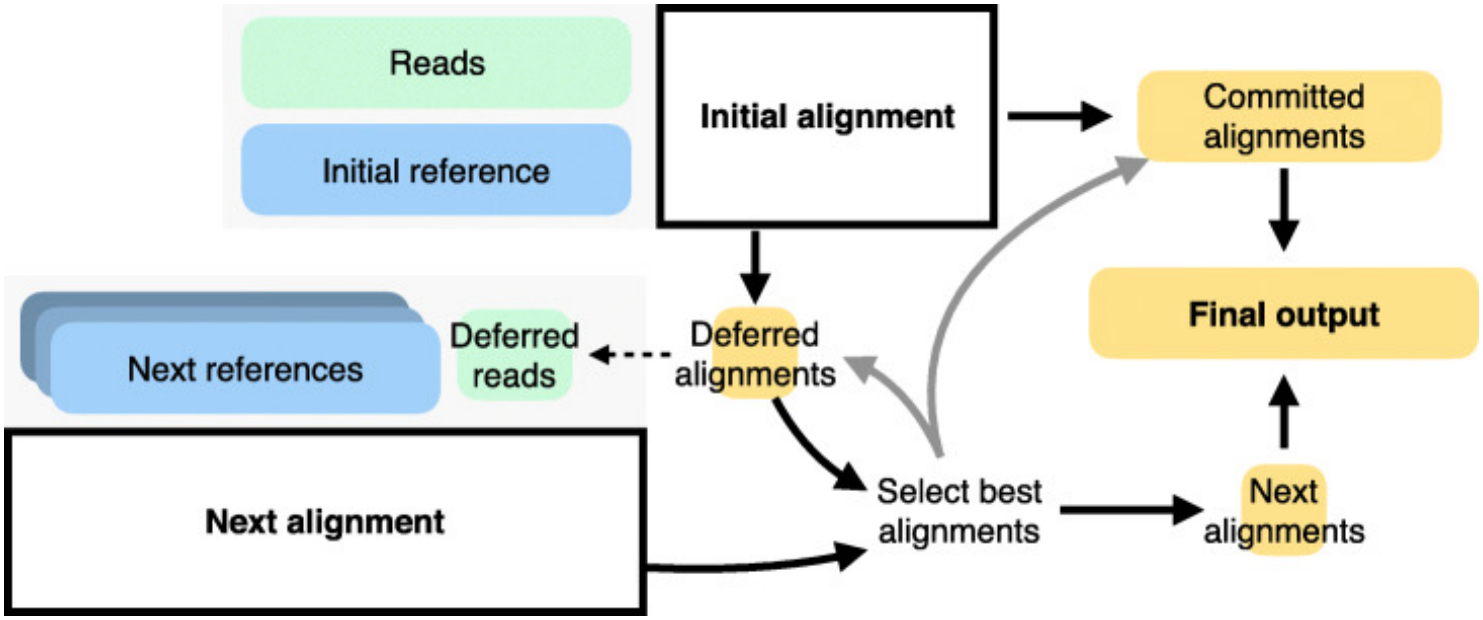
Graph aligner



Variation graph toolkit improves read mapping by representing genetic variation in the reference
Nat Biotechnol. 2018 Oct; 36(9): 875–879.

Reference flow alignment - multiple population ref-genomes

- Uses a collection of references chosen to cover known genetic variants



Reference flow: reducing reference bias using multiple population genomes.
Genome Biology, 2021
PMID: 33397514

SAM Properties

SAM - Recap

1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

2 FLAG

- Bitwise
- Picard (online) – explain flags

Examples

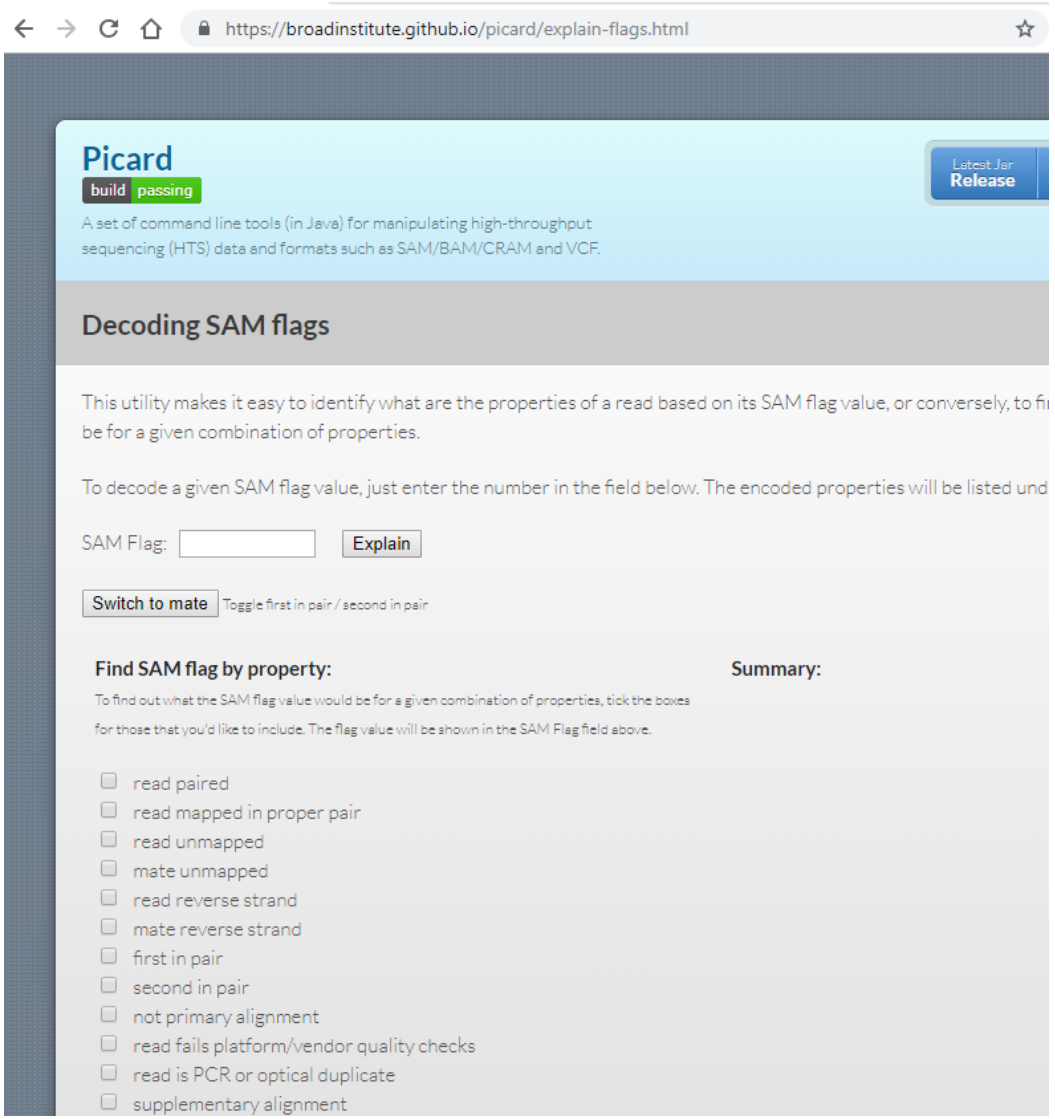
1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE

2 FLAG

- Bitwise
- Picard (online) – explain flags

Examples

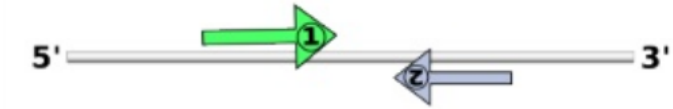
1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE



https://broadinstitute.github.io/picard/explain-flags.html

2 FLAG

Reads mapped in proper pair

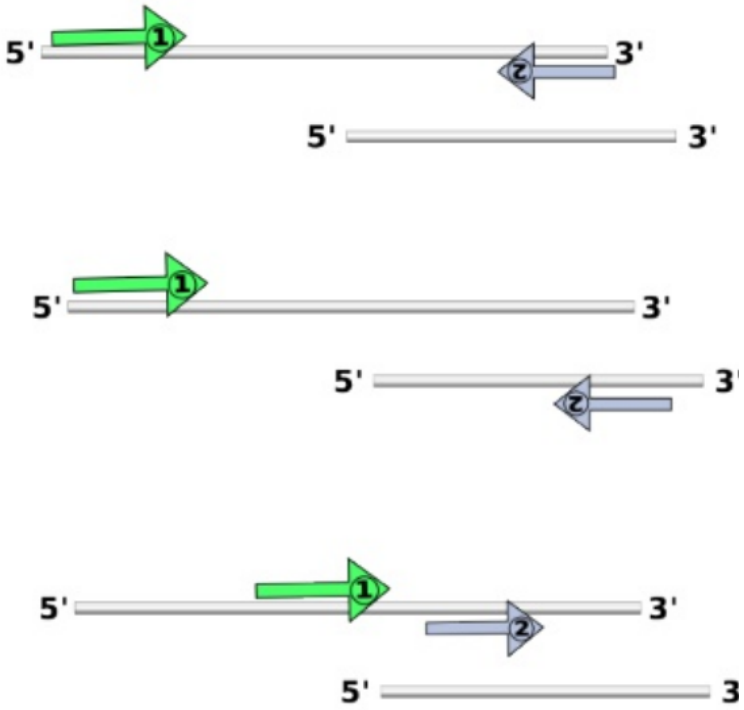


Flag: 67 Explain

Explanation:

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [ ] mate reverse strand
- [x] first in pair

Reads not mapped in proper pair



3 RNAME & 4 POS & 5 MAPQ

RNAME

- Reference name
  FASTA sequence name (e.g.: chr4)

POS

- Mapping position
  leftmost position in reference (e.g.: 142345)
  ! Reverse strand

MAPQ

- Mapping quality
- Phred score
- Depends on mapping program

Multiple mapping - MAPQ

Bowtie 1

- 255 for uniquely mapped reads (default)
- 0 for multiply mapped reads

Bowtie 2

- MAPQ filter of >=40 to get reads which had only 1 convincing alignment
- Lower filter to allow multi-mapped reads; secondary alignment with varying degrees of difference to the primary

BWA

- Phred scores as MAPQ values
- 0 if read maps to multiple positions

6 CIGAR

Used as a compact way to represent sequence alignment

Ref `ACTCAGTG--GT`

Read `ACGC-TGCAGTTATATAGG`

Cigar 4M1D3M2I2M7S

?

Example

Ref `CCTTAG`

Read `CCT-AA`

Cigar 3M1D2M

Op	BAM	Description
`M`	0	alignment match (can be a sequence match or mismatch)
`I`	1	insertion to the reference
`D`	2	deletion from the reference
`N`	3	skipped region from the reference
`S`	4	soft clipping (clipped sequences present in SEQ)
`H`	5	hard clipping (clipped sequences NOT present in SEQ)
`P`	6	padding (silent deletion from padded reference)
`=`	7	sequence match
`X`	8	sequence mismatch

Coordinate systems

Coordinate system

- 0 based → 0, 1, 2, … 9 | 1 based → 1, 2, 3, … 10

- BED – 0 based
- GFF – 1 based
- Ensembl uses a one-based coordinate system - UCSC use a zero-based coordinate system

	1 based	0 based
Third element	3	2
First ten	1, 10	0, 10
Second ten	11, 20	10, 20
One base long at 10	10,10	9,10
Interval	end – start + 1	end – start
Five elements at 100	100, 104	99, 104

Conversion tool

convert_zero_one_based

- Python CLI
- convert between zero and one based coordinate systems
- Use: `convert_zero_one_based --help`

https://github.com/griffithlab/convert_zero_one_based

Alignment

Quality check

Quality check of alignment

Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

Metrics

- Coverage / nucleotide distribution
- Reads mapped outside of a target (e.g., Exome sequencing)
- Number of mapped reads (wrong reference genome?)
- Insert size statistics
- Mapping quality - rule of thumb: Anything less than Q20 is not useful data

Tools

- Qualimap 2
  http://qualimap.bioinfo.cipf.es/
- bamstats
  http://bamstats.sourceforge.net/

Read Group Tag (RG)

RG - Meta information

- ID: unique e.g., SRA number (Sequence read archive)
- PL: Sequencing platform
- PU: Platform unit (run name / flowcell-barcode-lane)
- LB: Library name
- PI: Insert size (Predicted mean insert size)
- SM: Sample (Individual)
- CN: Sequencing center

PICARD

Use Picard to add RG tags to your SAM files

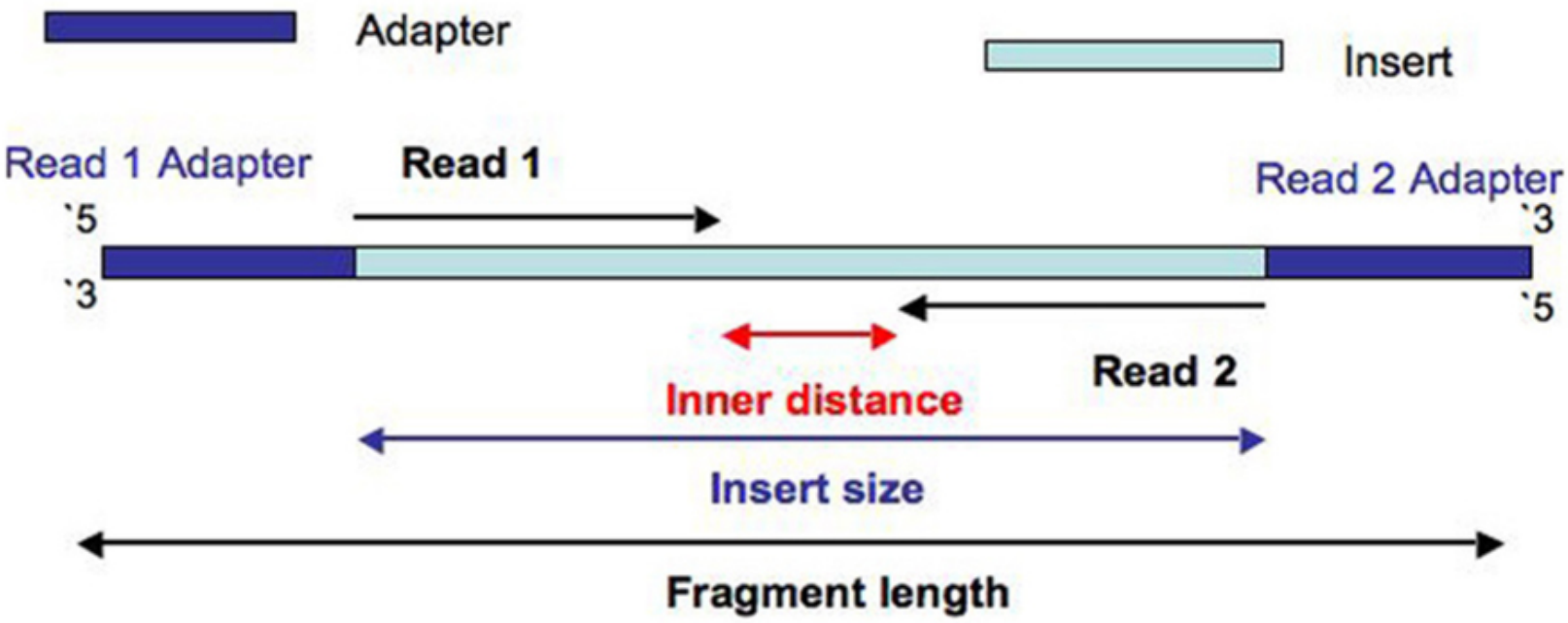
Read Group Tag – why important?

- It refers to a set of reads that were generated from a single run of a sequencing instrument.
- When multiplexing is involved, then each subset of reads originating from a separate library run on that lane will constitute a separate read group.
- To see the read group information for a BAM file, use the following command:

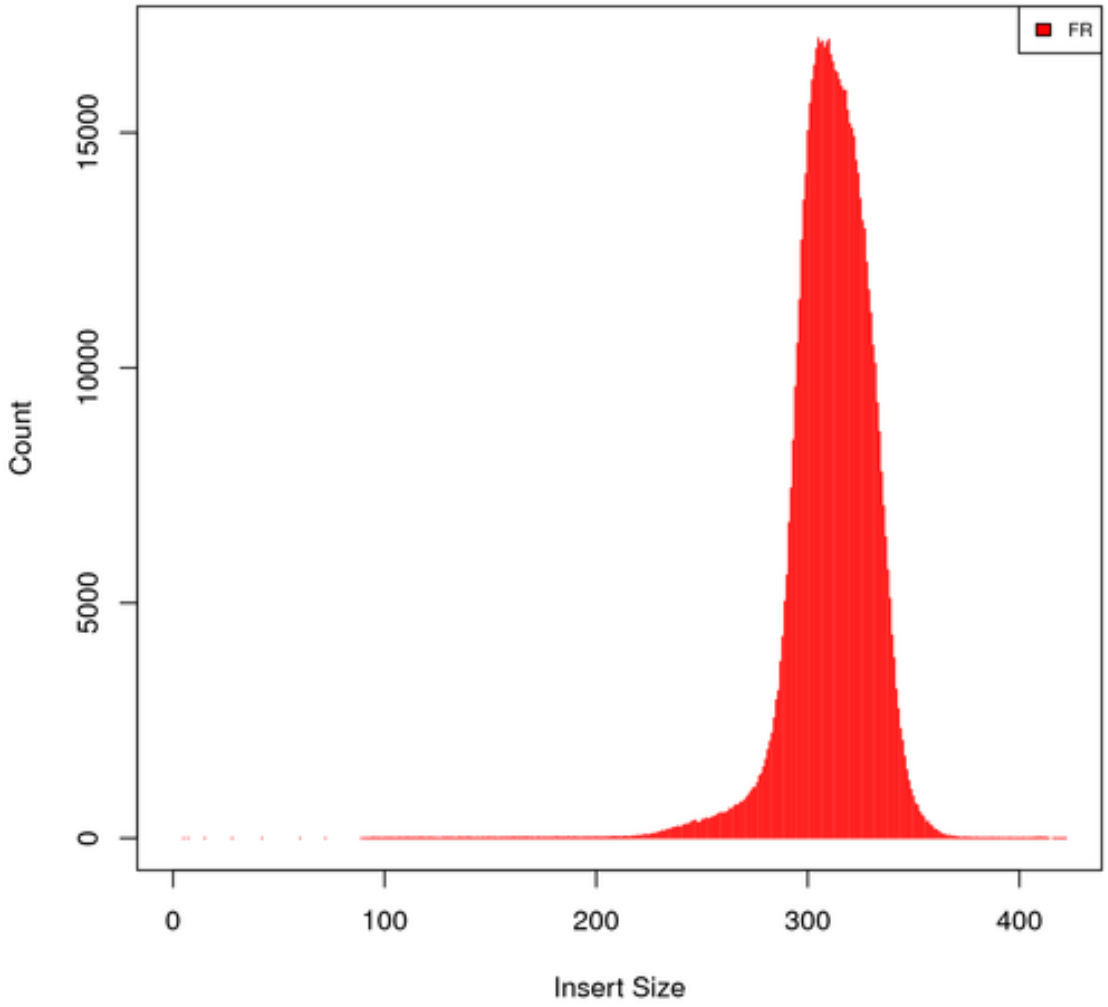
```
samtools view -H sample.bam | grep '@RG’
```

- Check that your FASTQ files have appropriate RQ when performing demultiplexing
- Important to have a correct RG tag → required by bioinformatics analysis tools (GATK, …)

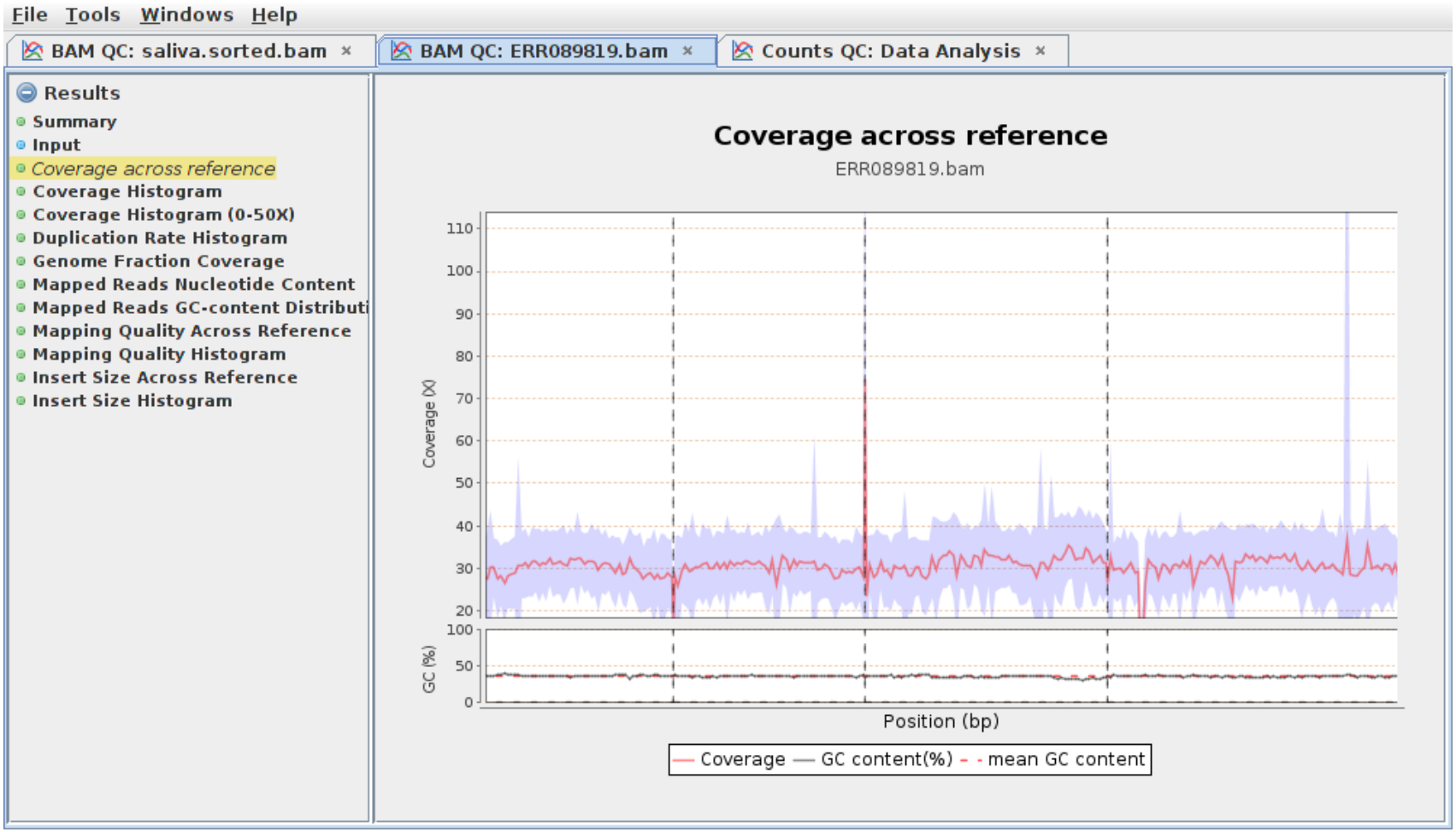
Insert size


Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length

Insert size histogram



Qualimap 2


File Tools Windows Help
BAM QC: saliva.sorted.bam
BAM QC: ERR089819.bam
Counts QC: Data Analysis
Results
Summary
Input
Coverage across reference
Coverage Histogram
Coverage Histogram (0-50X)
Duplication Rate Histogram
Genome Fraction Coverage
Mapped Reads Nucleotide Content
Mapped Reads GC-content Distributi
Mapping Quality Across Reference
Mapping Quality Histogram
Insert Size Across Reference
Insert Size Histogram
Coverage across reference
ERR089819.bam
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

Read duplicates

Origin

- PCR amplification step in library preparation
  1. Get DNA pieces (shatter / enrich DNA)
  2. Ligate adapters to both ends of the fragments
  3. PCR amplify the fragments with adapters
  4. Put fragments on beads or across flowcells
  5. Amplify fragments
  6. Sequence

Identification

- Have the same starting position

http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965708/

Read duplicates

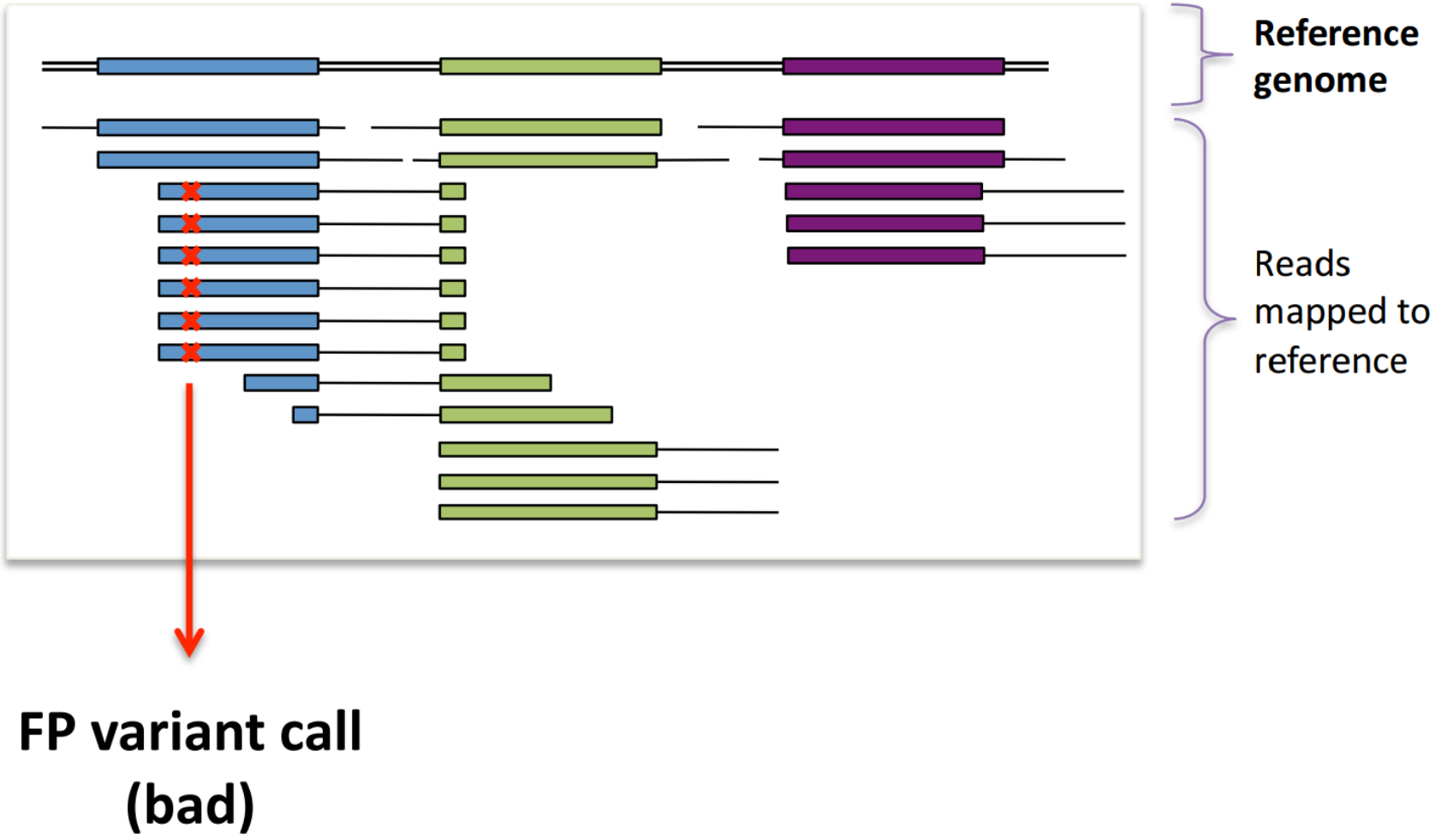
Problems/causes

- More steps during PCR amplification with little input material → more duplicates
  - This may result in duplicate DNA fragments in the final library

- Higher rates (~30%) arise when too little starting material is used → more amplification of the library is needed

- May result in false SNP calls (statistical model gets mixed up)

Read duplicates



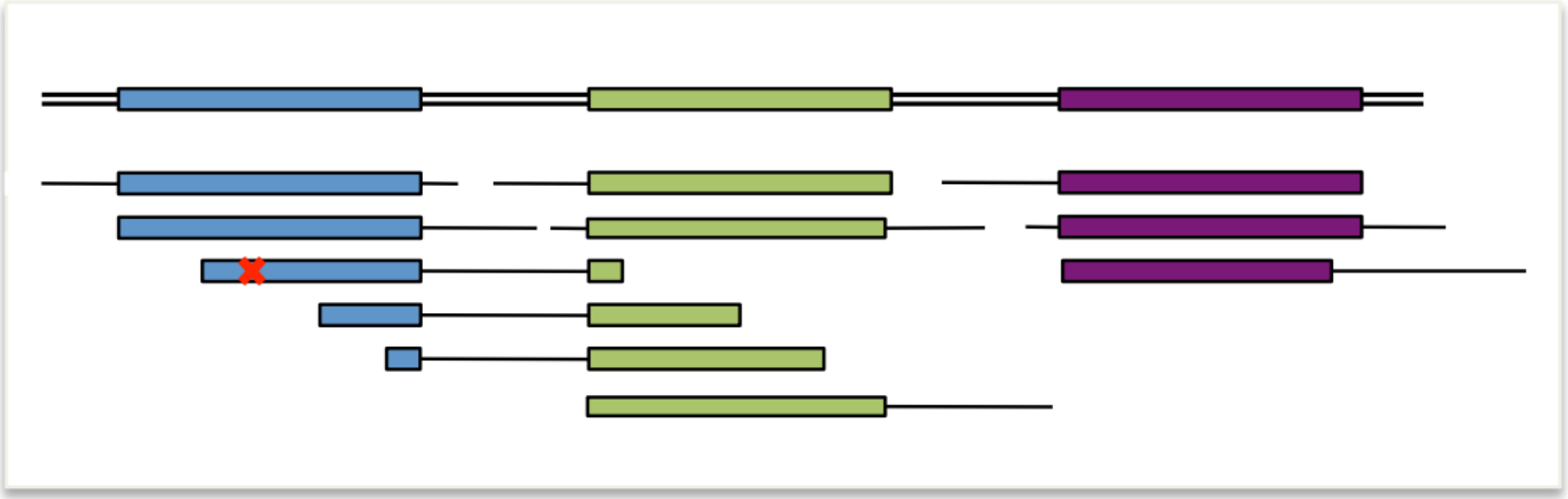
http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

Read duplicates - removal

→ Identify reads that map to the same location

→ Remove all but one

After marking duplicates



Read duplicates - removal

Attention!

Do not remove for

- Haloplex enrichment (nonrandom fragmentation method)
- PCR based enrıchment

! You would remove results

Base Quality Score Recalibration

- Various sources of systematic error
  → over / under-estimated base quality scores

- Quality score assigned to single base in isolation
  (assigned by the sequencing machines)

- Variant calling algorithms rely heavily on base quality scores

Solution → Correct base quality scores

Base Quality Score Recalibration

Apply machine learning to model these errors empirically and adjust the quality scores accordingly

- First the program builds a model of covariation based on the data
  - Reported quality score
  - Position in the read (cycle)
  - Preceding and current base – sequence context (homopolymer, ...)
- and a set of known variants (1000g, dbSNP, large private cohort)
  - discount most of the real genetic variation

- First pass: calculate new QS based on the model
  Second pass: adjust the base quality scores

→ Visual inspection with before/after plots

Good explanation: http://zenfractal.com/2014/01/25/bqsr/

Base Quality Score Recalibration

WES

- For WES restrict to capture targets
  off-target sites are likely to have higher error rates

Organism with no known variants?

- Call variants -> apply stringent filter -> use these for recalibration
- Repeat previous steps

https://www.broadinstitute.org/gatk/events/slides/1506/GATKwr8-B-3-Base_recalibration.pdf

Informative NGS QC page

Data storage

Guidelines for diagnostic next-generation sequencing

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795226/

Eur J Hum Genet. 2016

- Should stick to the standard open file formats FASTQ, BAM, and VCF
- Full-log files have to be stored in addition to the analysis results
  - complete as possible
  - making the whole analysis from FASTQ data to the diagnostic report reproducible.
- No (international) consensus yet on what should be stored

ACMG clinical laboratory standards for next-generation sequencing

https://www.acmg.net/docs/acmg_lab_standards_next_generation_sequencing_sept2013.pdf

- Recommend that the laboratory consider a minimum of 2-year storage
- File type that would allow regeneration of the primary results as well as reanalysis
- Retention of the VCF, along with the final clinical test report interpreting the subset of clinically relevant variants

Working with SAM/BAM files

SAMTOOLS

http://samtools.sourceforge.net

View

```
samtools view -h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view -f 0x02 <file.bam> > <only_proper_paired.sam>
```

Sort

```
samtools sort -o <sorted.bam> <aln.bam>
```

Index

```
samtools index <sorted.bam>
(only for BAM)
```

SAMTOOLS

Simple stats (reads mapped, reads paired, ...)

```
samtools flagstat <file.bam>
```

Stats for each chr/contig - reads mapped and unmapped

```
samtools idxstats <sorted.bam> (and indexed)
```

Convert SAM to BAM

```
samtools view -S -h -b <aln.sam> > <aln.bam>
```

Converting BAM to a sorted BAM file (without intermediate file)

```
samtools view -bSh <file.sam> | samtools sort -o
file_sorted.bam -
```

http://davetang.org/wiki/tiki-index.php?page=SAMTools

Other tools

SAMBAMBA - „Multithreaded" SAMtools - https://github.com/lomereiter/sambamba

- view
- sort
- index
- merge
- flagstat
- markdup

elprep - https://github.com/exascience/elprep

- High-performance tool for preparing .sam / .bam / .cram files
- In-memory and multi-threaded application
- Requires lots of memory (WGS ~256GB)
- Replacement for SAMTOOLS & Picard

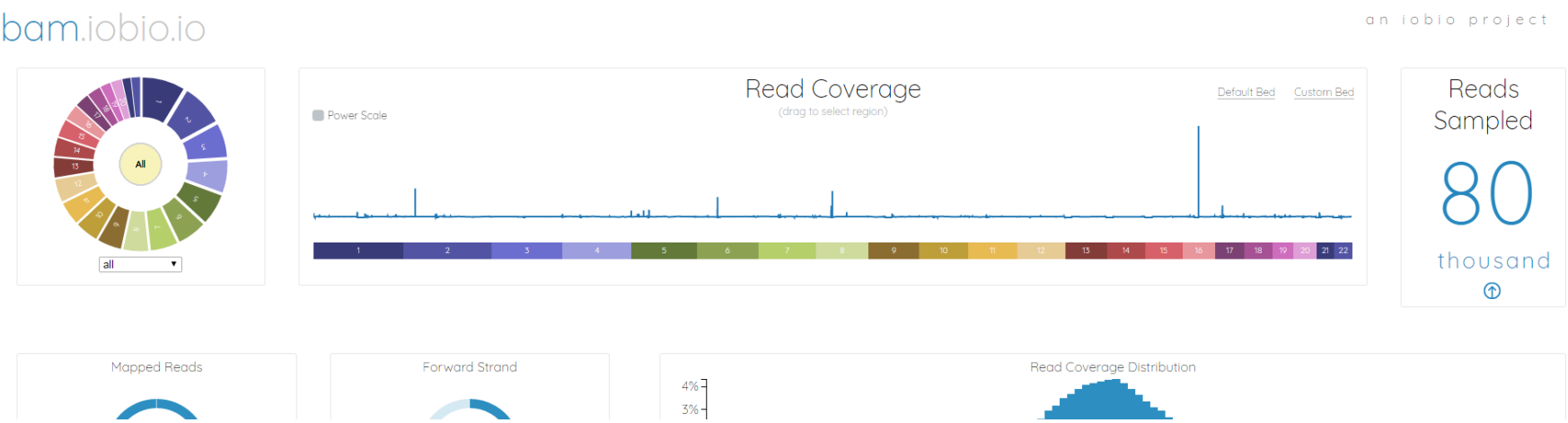
PICARD & iobio

PICARD

- JAVA based tool (http://picard.sourceforge.net/)
- BuildBamIndex, FastqToSam, MergeSamFiles, …

bam.iobio.io

- Web-based (http://bam.iobio.io)
- Coverage overview
- Mapping overview



Genetic variations

Genetic variation

A regular diploid human cell contains 46 chromosomes (23 pairs)

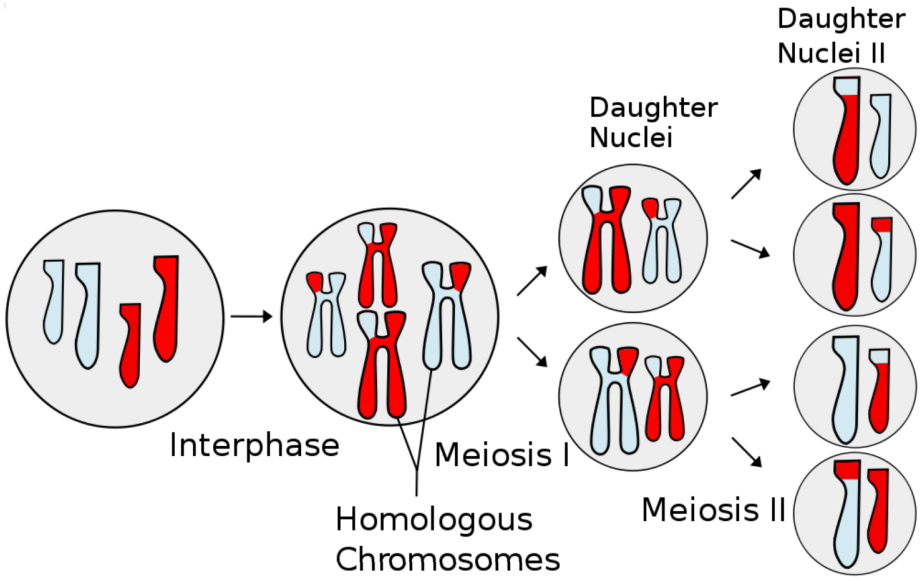
- 22 pairs + sex chromosomes XX(female) XY(male)
- One set of chromosomes inherited from each parent

Germline

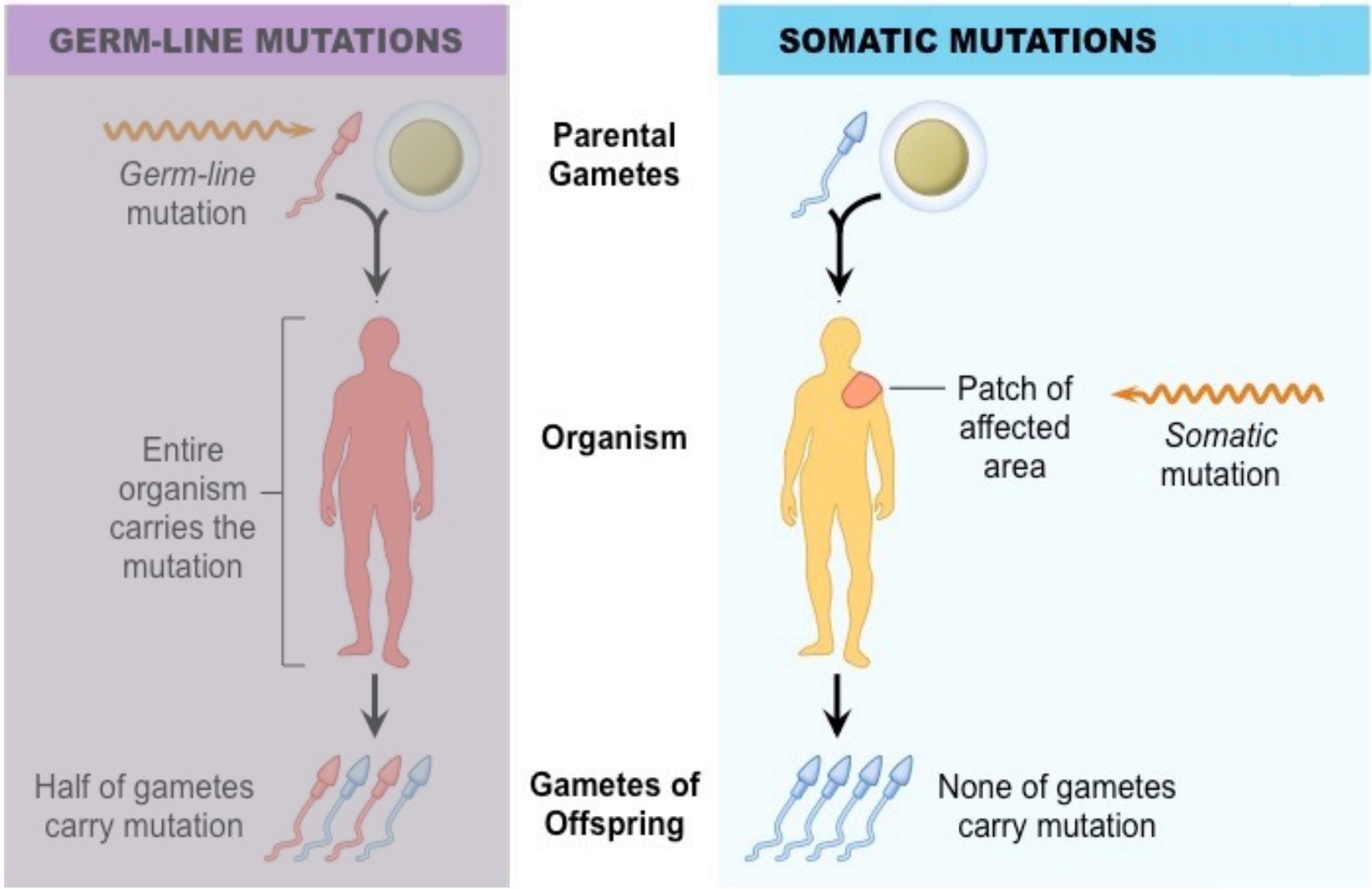
- Meiosis → four genetically unique **haploid** gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell

Somatic mutation

- Not inherited from parent
- Acquired from spontaneous mutations during DNA replication



Somatic mutations



Types of genetic variations

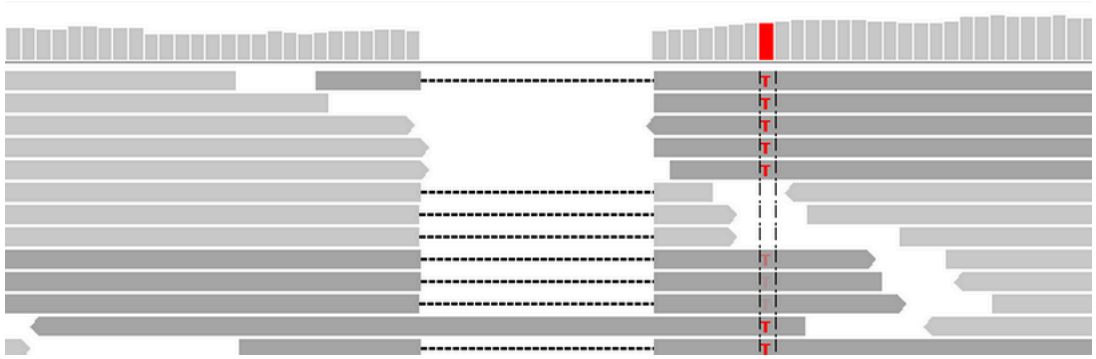
SNV / SNP

- A single nucleotide — A, T, C or G — in the genome differs between members of a population or chromosome pairs
- Originally defined as occurring at least in one individual of the population (these definitions may shift in time)
- SNV (single nucleotide variant) if observed very rarely
- SNP, SNV → may fall within
  - coding sequences of genes
  - non-coding regions of genes
  - intergenic regions

Types of genetic variations

INDEL

- Insertion / deletion of bases
- Coding regions of the genome - produce a frameshift mutation (unless multiple of 3)
- There are approximately 190-280 frameshifting INDELs in each person.
  A map of human genome variation from population-scale sequencing". Nature 467 (7319)



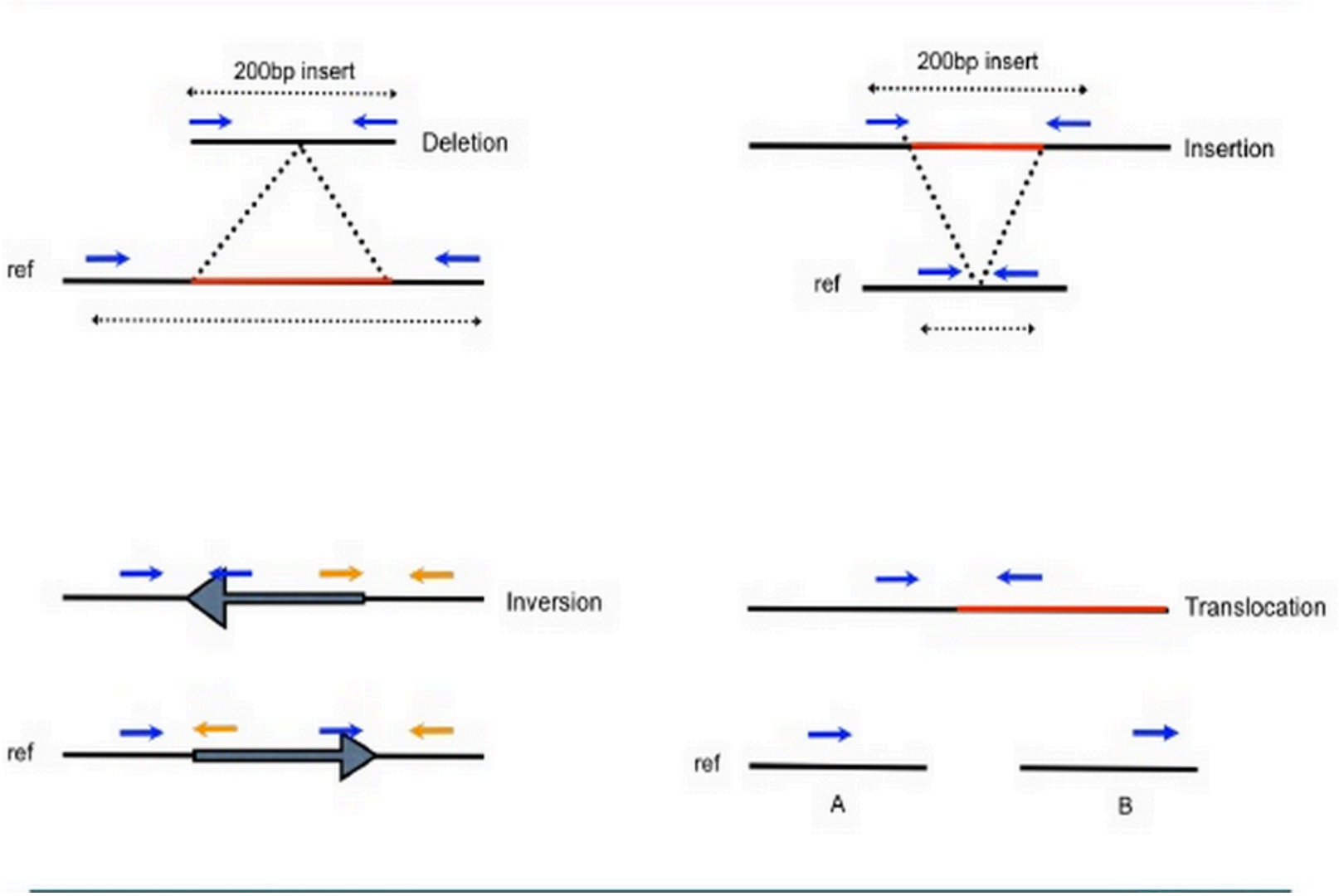
Types of genetic variations

Structural variations (SV)

- Variation in structure of an organism's chromosome

- Insertions
  Deletions
  CNV
  Inversions
  Translocations

Types of structural genetic variations



Human genome

- Typical genome differs from the reference human genome at **4.1 million to 5.0 million sites**.

- **~99.9%** of variants consist of **SNVs and INDELs**

- Structural variants affect more bases
  - 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~1100 insertions)
  - Affecting ~20 million bases of sequence

- African ancestry populations harbor the greatest numbers of variant sites (as predicted by the out-of-Africa model of human origins)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/

The 1000 genome project; A global reference for human genetic variation; Nature 2015

What kind of data do you have?

Genotyping

- Single samples
- Family → finding causing mutation of rare disease
- Time series

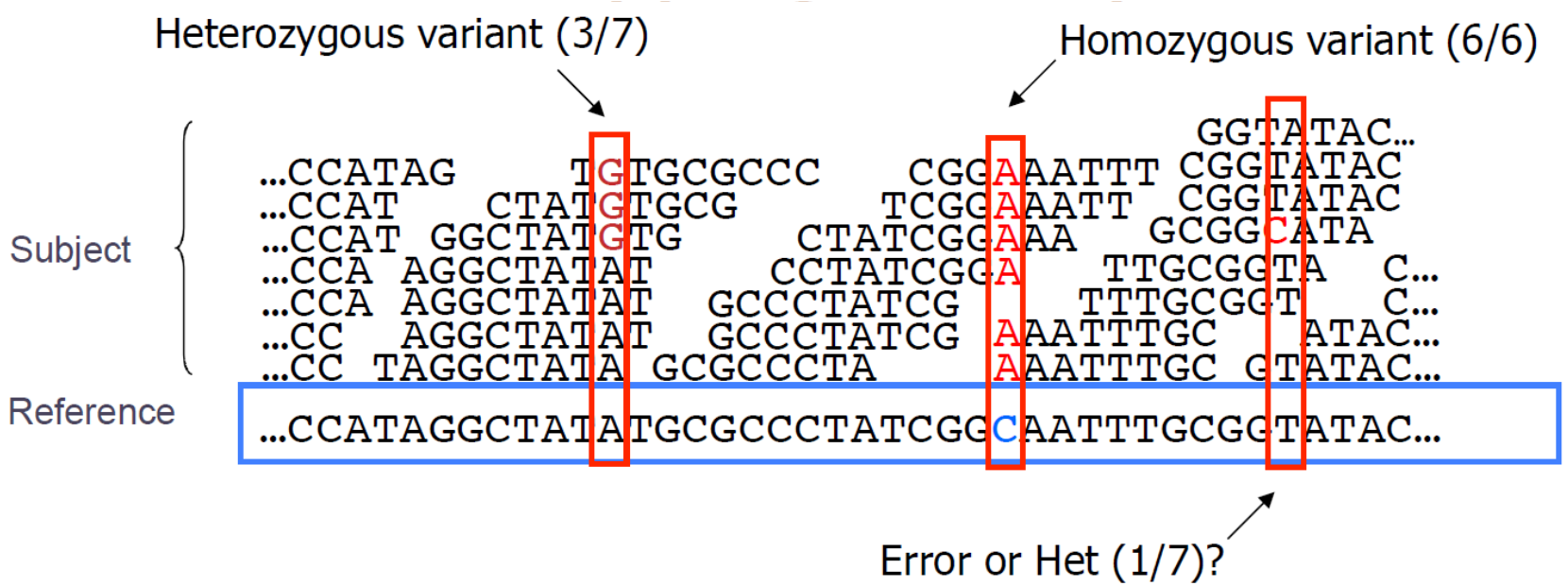
Somatic mutations

- Primary tumor tissue
- Blood samples

→ Different callers / strategies for respective samples

Variant Calling

Genotyping theory



- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!
- Sequencing instruments make mistakes
  – Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

What are variant calls?

Find differences to a reference (hg19, GRCh38)

Naive variant calling

- Check all the reads that cover base chr11:1234567
- Add up the bases at chr11:1234567
- e.g. 15 A's, 4 G's
- Is this an A/G heterozygous site or four sequencing errors?

Actual variant callers

- Estimate likelihood of a variant site vs a sequencing error
- Sequencing error rate
- Quality scores

(Other) reasons for a mismatch

- True SNP

OR

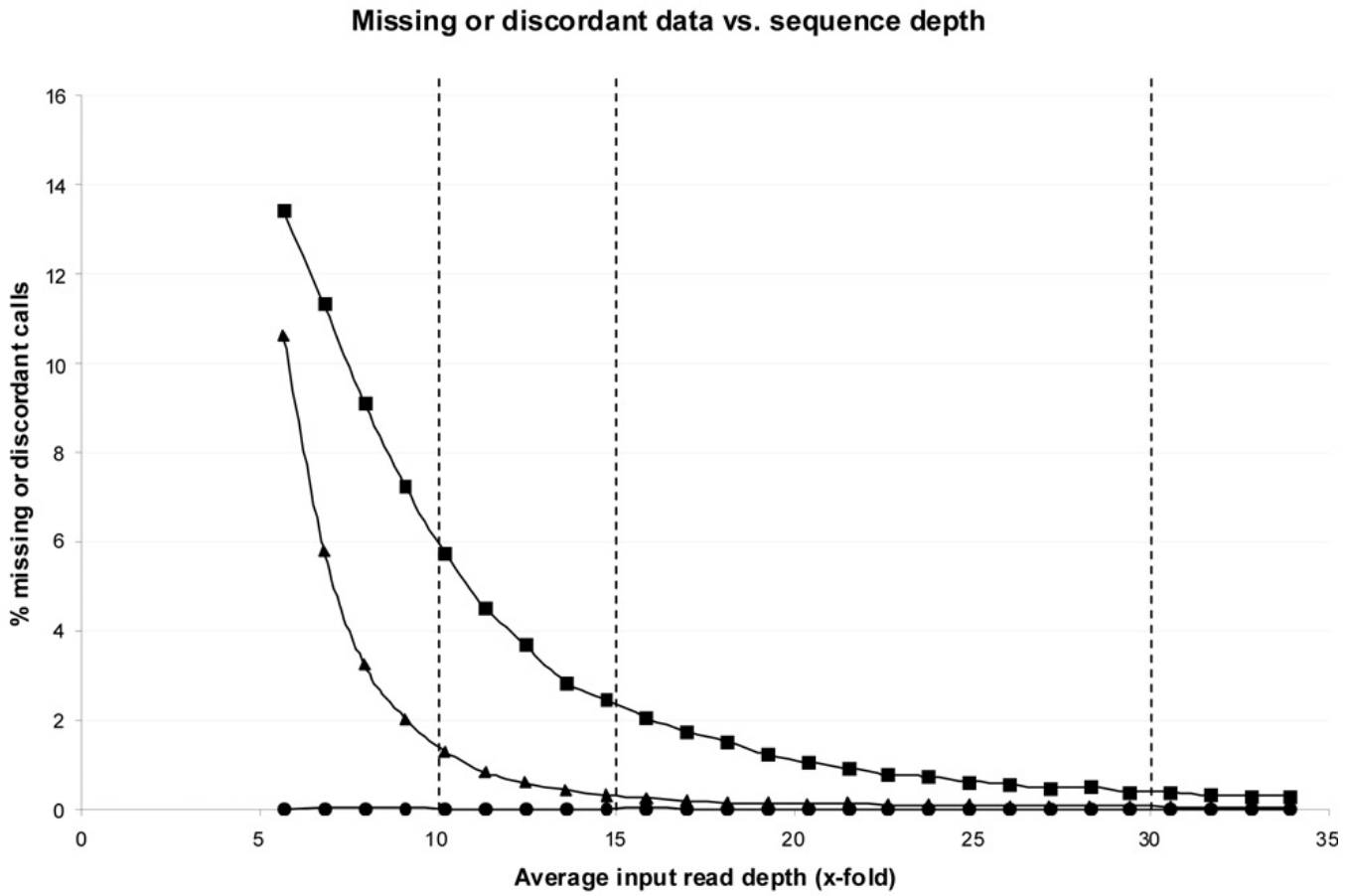
- Error generated in library preparation
- Base calling error
  - May be reduced by improved base calling methods, but cannot be eliminated
- Misalignment (mapping error)
  - Local realignment to improve mapping
- Error in reference genome sequence

Difficulties

- High depth - low quality regions → likely due to copy number or other larger structural events
- Repetitive regions → artefactual variants
- Regions of low complexity (~ 2% of genome)
  polypurine (AG), AT-rich regions, simple tandem repeats

Linderman et al. **Analytical validation of whole exome and whole genome sequencing for clinical applications.** *BMC Medical Genomics* 2014, **7**:20

Heng Li. **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** http://arxiv.org/pdf/1404.0929v1.pdf



Squares Not covered by sequence

Triangles Heterozygote undercalls in sequence data relative to genotype data

Circles Discordant SNV calls compared to genotypes

Undercall - (false negatives; true genotype is polymorphic yet the call is homozygous reference or 'missing')
Bentley, et al. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry, Nature

Genotype variant calling

Variant callers

Bayesian genotype model - evaluates probability of genotype given read data

Basic model - Bayes Theorem

$P(genotype|data) \propto P(data|genotype)\ P(genotype)$

P(genotype): prior probability for variant (Genome wide SNP rate)

P(data|genotype): likelihood for observed (called) allele type

Likelihood P(data|genotype) - what s' known to affect base calling

- Error rate increases as cycle numbers increase
- Error rate depends on substitution type (Ti/Tv)
- Error rate depends on local sequence environment
- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Proximity to INDEL

INDEL calling

In general – call INDELs based on the I and D events in BAM file

Consider

- Misalignment of the read
- Homopolymer runs
- Length of reads
- Zygosity

Approach to remove FP

- Create new haplotype (new reference) and realign the reads to this ref
- Count number of reads supporting this new haplotype
- → computationally extensive

Refined INDEL analysis

Examine sources of INDEL errors

- Experimental validation of INDELs called from 30x whole genome vs. 110x whole exome of the same sample
- Most of the errors due to short microsatellite errors introduced during exome capture, also misses most long INDELs
- Recommend WGS for INDEL analysis instead

	All INDELs	Valid	PPV	INDELs >5bp	Valid (>5bp)	PPV (>5bp)
Intersection	160	152	95.0%	18	18	100%
WGS	145	122	84.1%	33	25	75.8%
WES	161	91	56.5%	1	1	100%

Reducing INDEL calling errors in whole-genome and exome sequencing data
Fang, H, Wu, Y, Narzisi, G, O'Rawe, JA, Jimenez Barrón LT, Rosenbaum, J, Ronemus, M, Iossifov I, Schatz, MC § , Lyon, GL §
Genome Medicine (2014) 6:89. doi:10.1186/s13073-014-0089-z

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, INDELs) within exome-capture data.

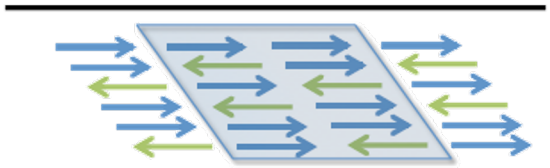
Features

- Combine mapping and assembly
- Exhaustive search of haplotypes
- De novo mutations

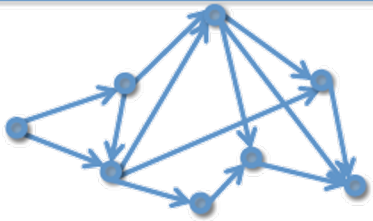
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods.* doi:10.1038/nmeth.3069

Scalpel Algorithm

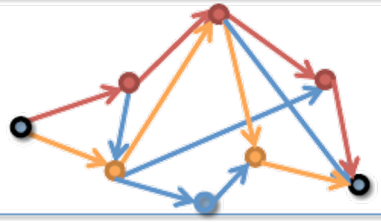
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations

Variant Call Format VCF

File format to store variant information

Specification

https://github.com/samtools/hts-specs

Variant calling data files

VCFv4.3.tex is the canonical specification for the Variant Call Format and its textual (VCF) and binary (BCF) encodings, while **VCFv4.1.tex** and **VCFv4.2.tex** describe their predecessors. **VCFv4.4.draft.tex** is a working draft of the upcoming version of VCF format and is under active revision. These formats are discussed on the vcftools-spec mailing list.

More information

- http://vcftools.sourceforge.net/VCF-poster.pdf
- https://www.biostars.org/p/12964/

VCF file format

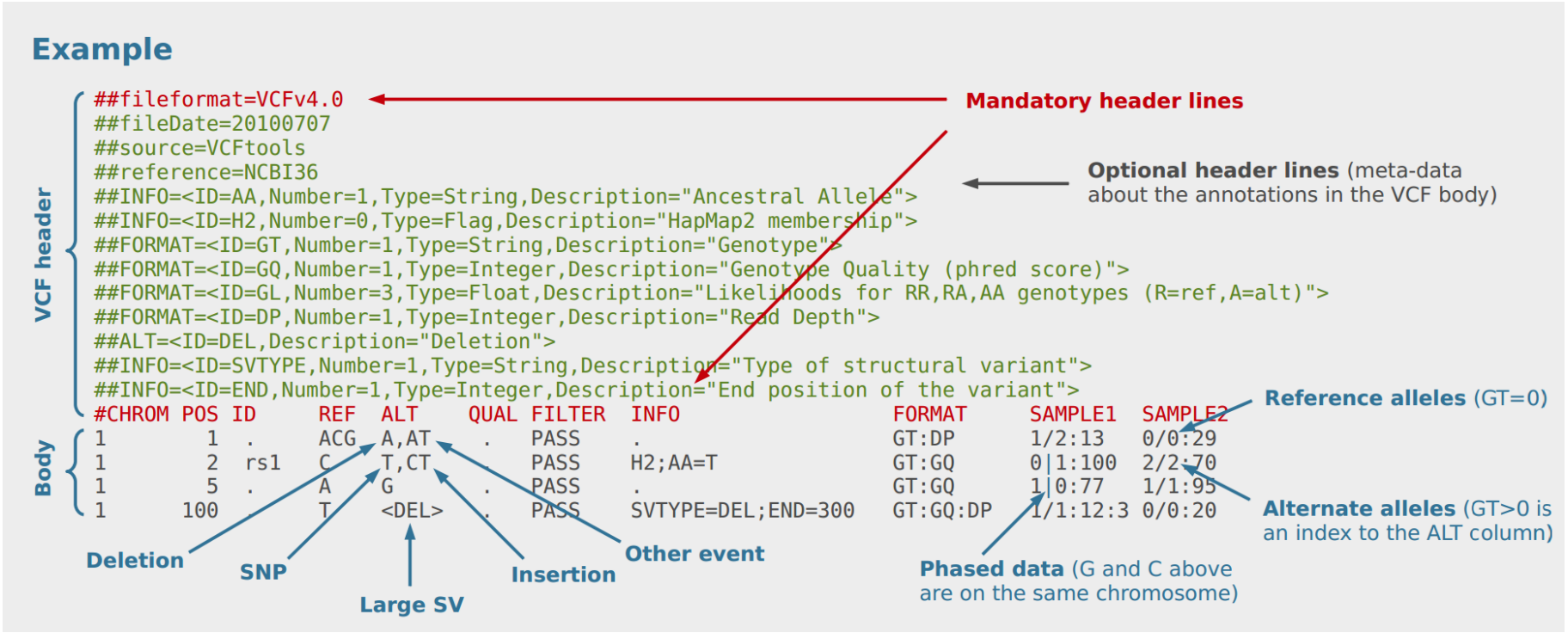
CHROM	chromosome / contig
POS	the reference position with the 1st base having pos 1 for INDELs this is actually the base preceding the event
ID	id, if dbSNP variant - rs number
REF	reference base for INDELs, the reference string must include the base before the event
ALT	comma separated list of alternate non-reference alleles called on at least one of the samples
QUAL	phred-scaled quality score of the assertion
FILTER	PASS if the position has passed all filter criteria, otherwise list why filter was not passed
INFO	additional information

QUAL - Variant Calling Quality

Factors

- Coverage at position (DP)
- Number of reads supporting the call
- Strand bias
- Base qualities at variant position

VCF - Example



http://vcftools.sourceforge.net/VCF-poster.pdf

Format fields

Specifies type of data present for each genotype

- e.g.: GT:DP:GQ:MQ
- fields defined in metadata header

GT	Genotype
DP	Read depth at position for sample
DS	Downsampled because of too much coverage
GQ	Genotype quality encoded as a phred quality
MQ	Mapping quality
QD	Variant quality score over depth

...

Genotype field

- GT: genotype, encoded as alleles separated by either | or /
  - 0 for the ref, 1 for the 1st allele listed in ALT, 2 for the second, etc
  - REF=A and ALT=T
- genotype 0/0 means homozygous reference A/A
- genotype 0/1 means heterozygous A/T
- genotype 1/1 means homozygous alternate T/T
  - /: genotype unphased and | genotype phased
    (Phased data are ordered along one chromosome https://www.biostars.org/p/7846/)
- ...

```
chr1    873762  .       T   G   [CLIPPED] GT:AD:DP:GQ:PL    0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   [CLIPPED] GT:AD:DP:GQ:PL    1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   [CLIPPED] GT:AD:DP:GQ:PL    0/1:1,3:4:25.92:103,0,26
```

http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk

VCF – Example

(taken from Thomas Keane)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- What version of the human reference genome was used?
- What does the DB INFO tag stand for?
- What does the ALT column contain?
- At position 17330, what is the total depth? What is the depth for sample NA00002?
- At position 17330, what is the genotype of NA00002?
- Which position is a tri-allelic SNP site?
- What sort of variant is at position 1234567?

Assignment 2

Types of variants

SNPs

Alignment
```
ACGT
ATGT
```

VCF representation

POS	REF	ALT
2	C	T

Insertions

Alignment
```
AC-GT
ACTGT
```

VCF representation

POS	REF	ALT
2	C	CT

Deletions

Alignment
```
ACGT
A--T
```

VCF representation

POS	REF	ALT
1	ACG	A

Complex events

Alignment
```
ACGT
A-TT
```

VCF representation

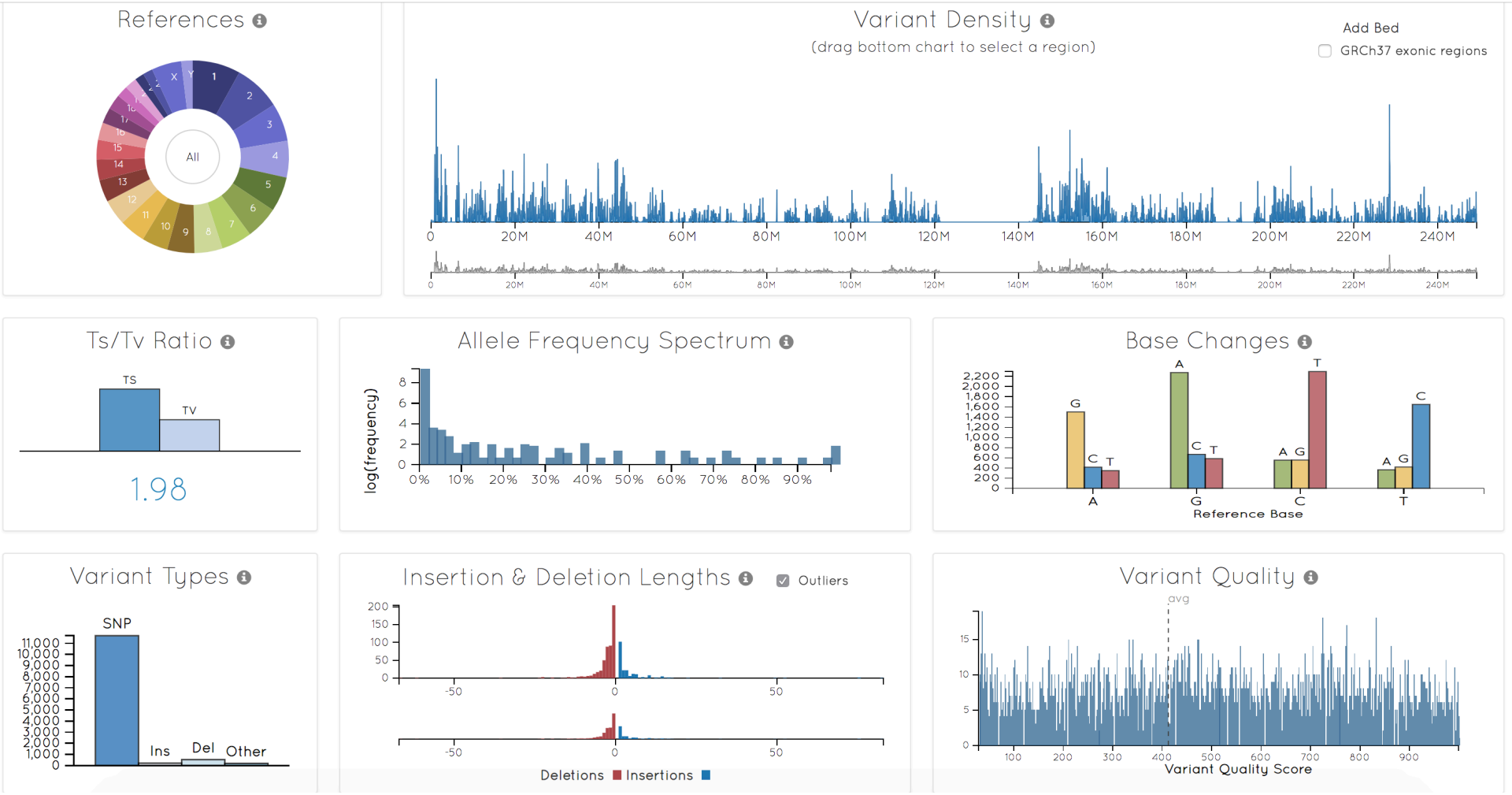
POS	REF	ALT
1	ACG	AT

Large structural variants

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

vcf.iobio.io


References
All
Variant Density
(drag bottom chart to select a region)
Add Bed
GRCh37 exonic regions
Ts/Tv Ratio
TS
TV
1.98
Allele Frequency Spectrum
log(frequency)
Base Changes
Reference Base
Variant Types
SNP
Ins
Del
Other
Insertion & Deletion Lengths
Outliers
Deletions
Insertions
Variant Quality
avg
Variant Quality Score

Realignment around INDELs

Why?

- Alignments tend to accumulate FP SNPs near true INDELs
- SNPs are often less penalized compared to INDELs

Realignment principles

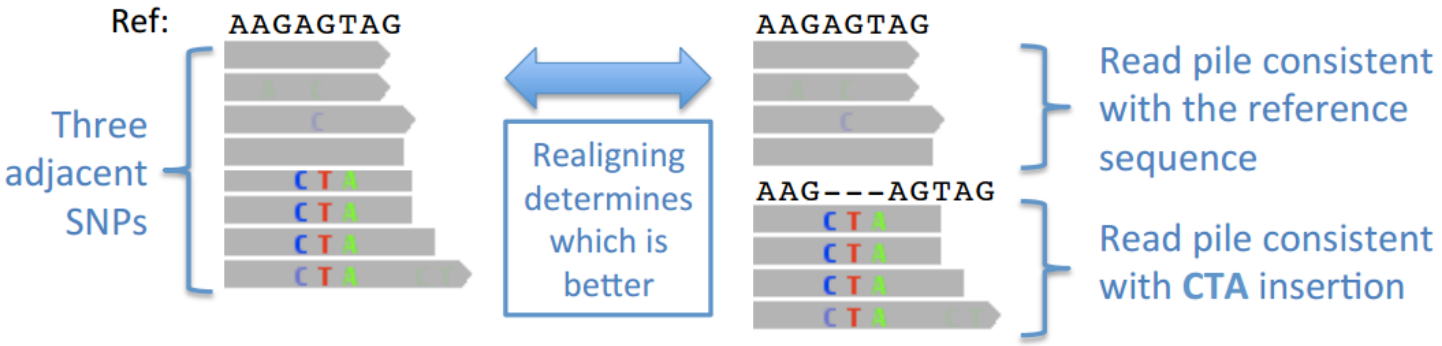
- Realign locally around INDELs→ GATK

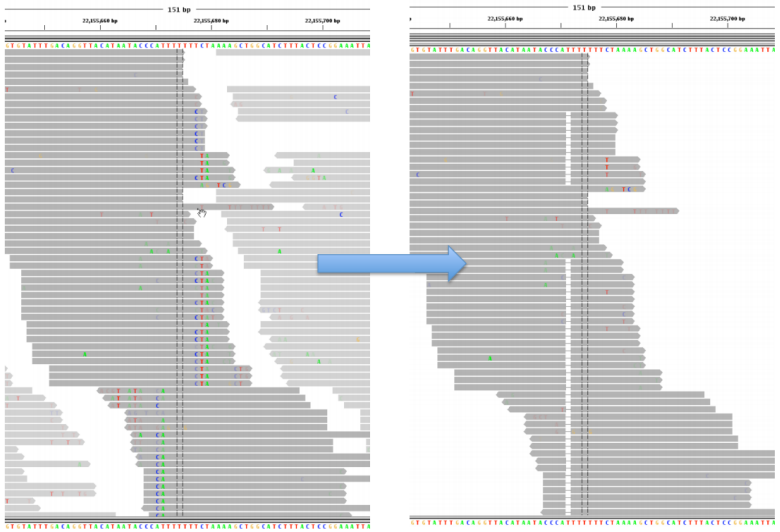
- Input set of known INDEL sites (dbSNP, 1000genomes)
- INDELs from alignments
- Presence of mismatches & softclips (BAM file)

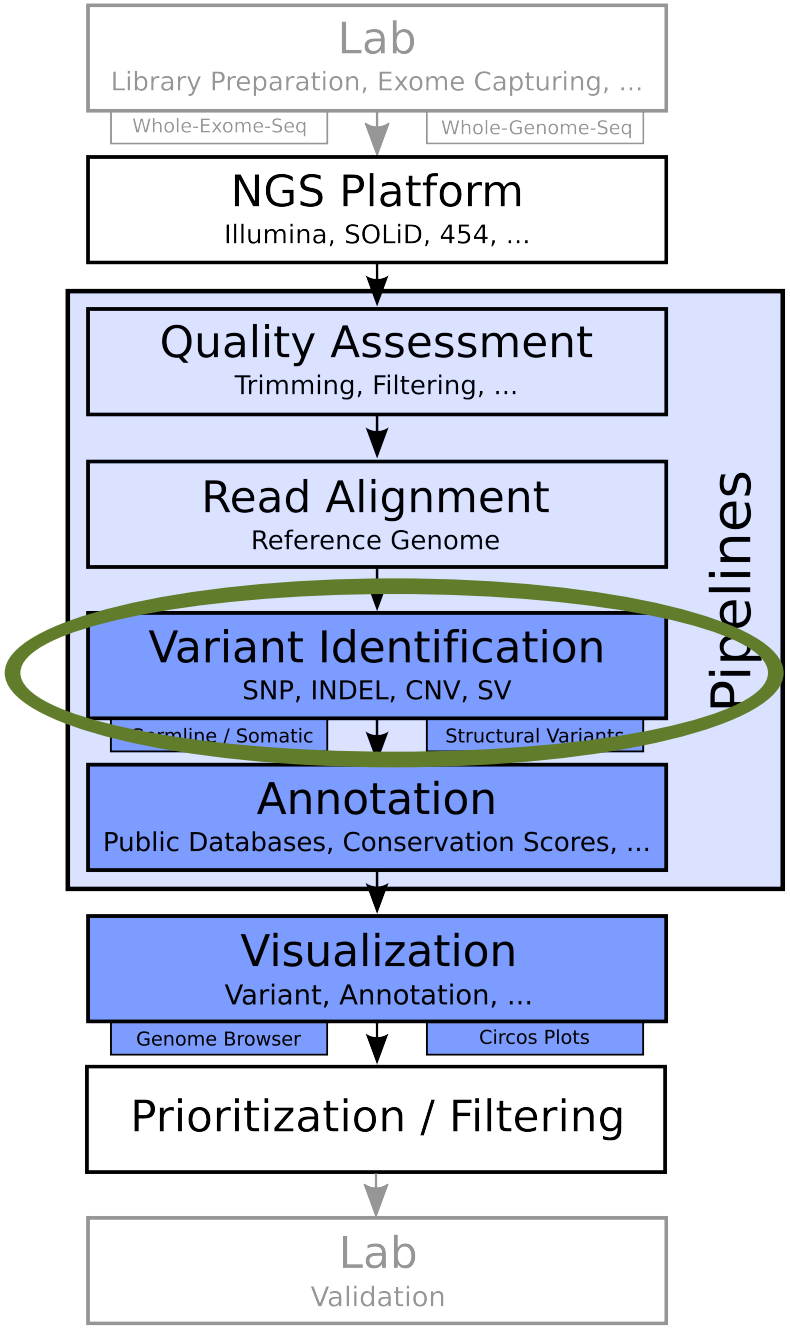
Realignment around INDELs

Realignment

- Model the INDEL haplotype
- → if score of alternative consensus is better than original use realigned







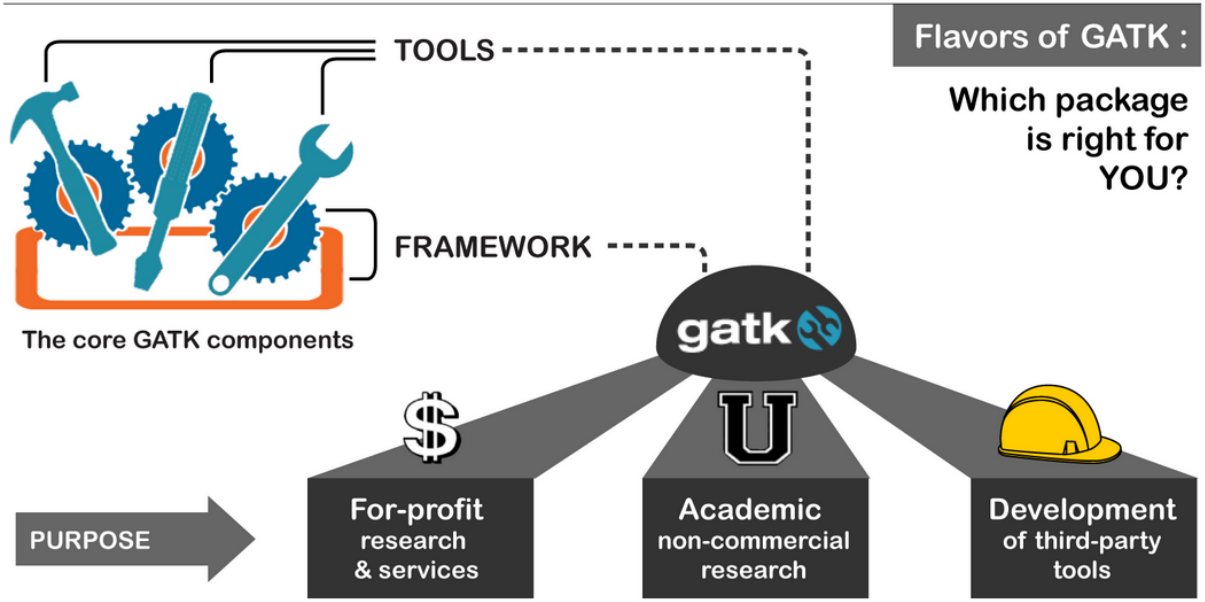
Variant callers

GATK - Genome Analysis Toolkit

Variant calling pipeline

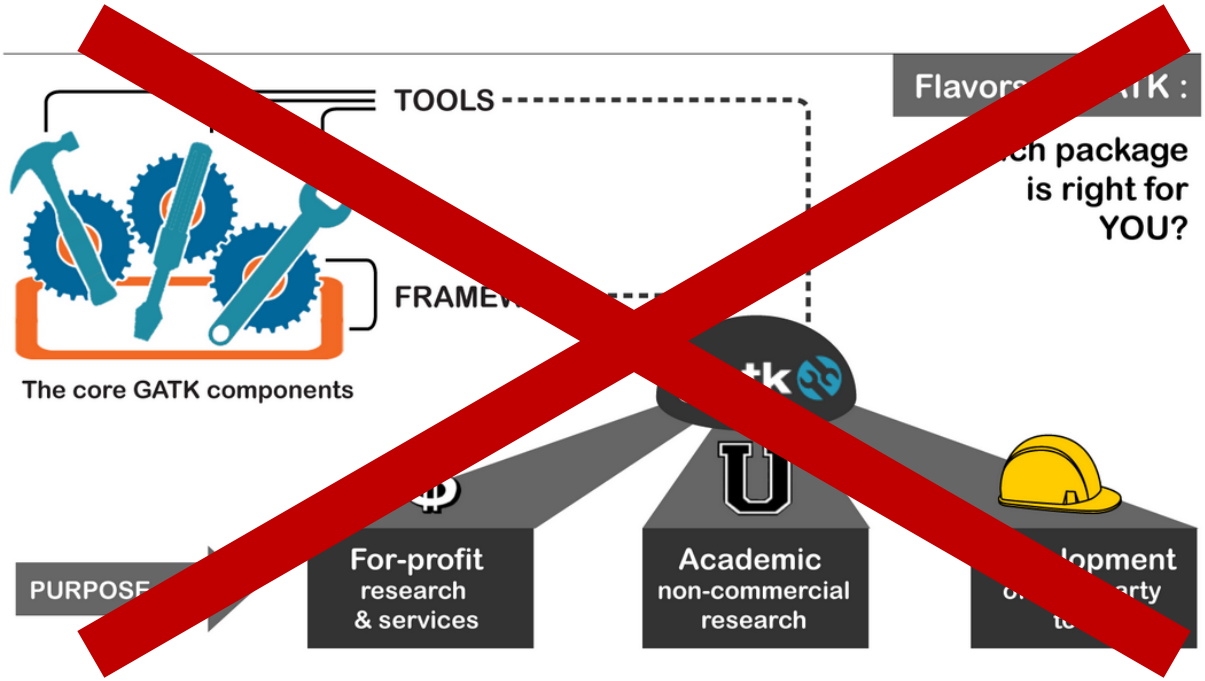
GATK

- JAVA, command line software
- Linux (Mac)
- Mixed closed/open-source model



GATK

- JAVA, command line software
- Linux (Mac)
- Mixed closed/open-source model

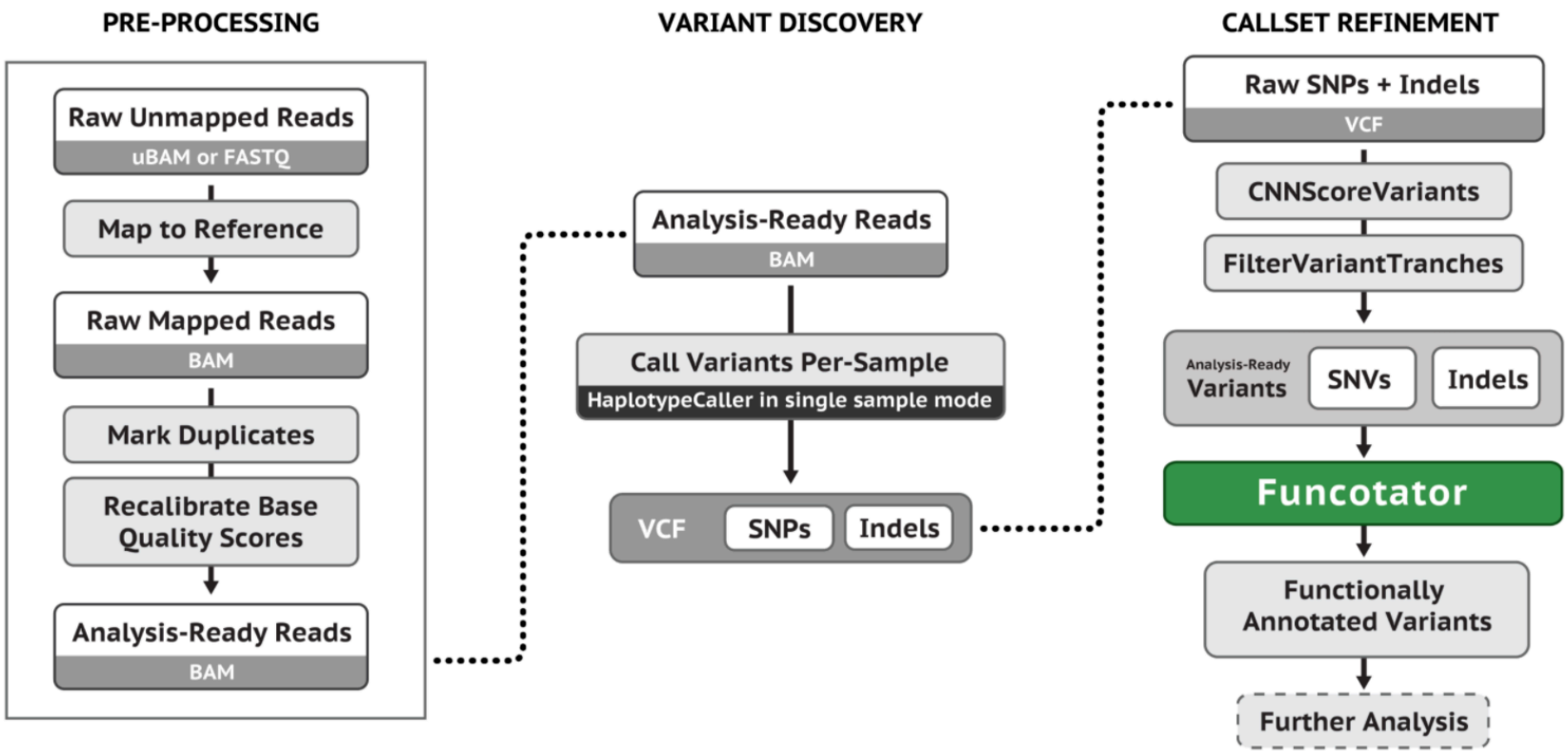


GATK

- JAVA, command line software
- Linux (Mac)
- GATK4 is open-source under a BSD 3-clause

Permissions	Limitations	Conditions
✔ Commercial use	✖ Liability	ℹ License and copyright notice
✔ Modification	✖ Warranty	
✔ Distribution		
✔ Private use		

GATK



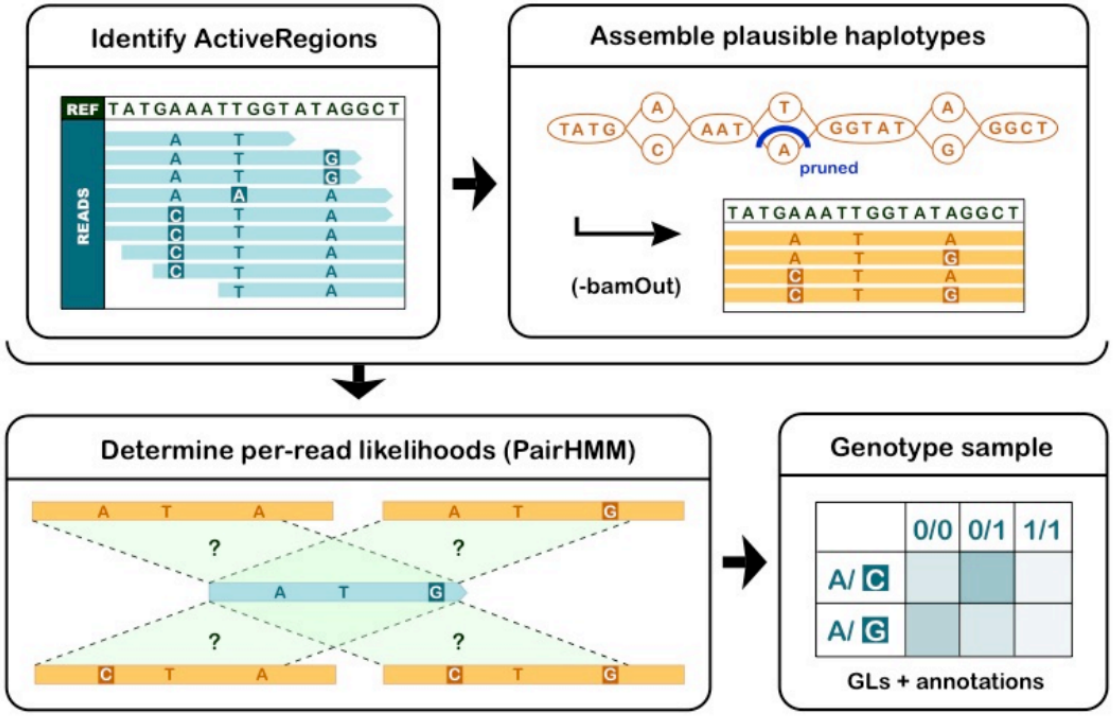
https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Funcotator (FUNCtional annOTATOR) analyzes given variants for their function (as retrieved from a set of data sources) and produces the analysis in a specified output file.

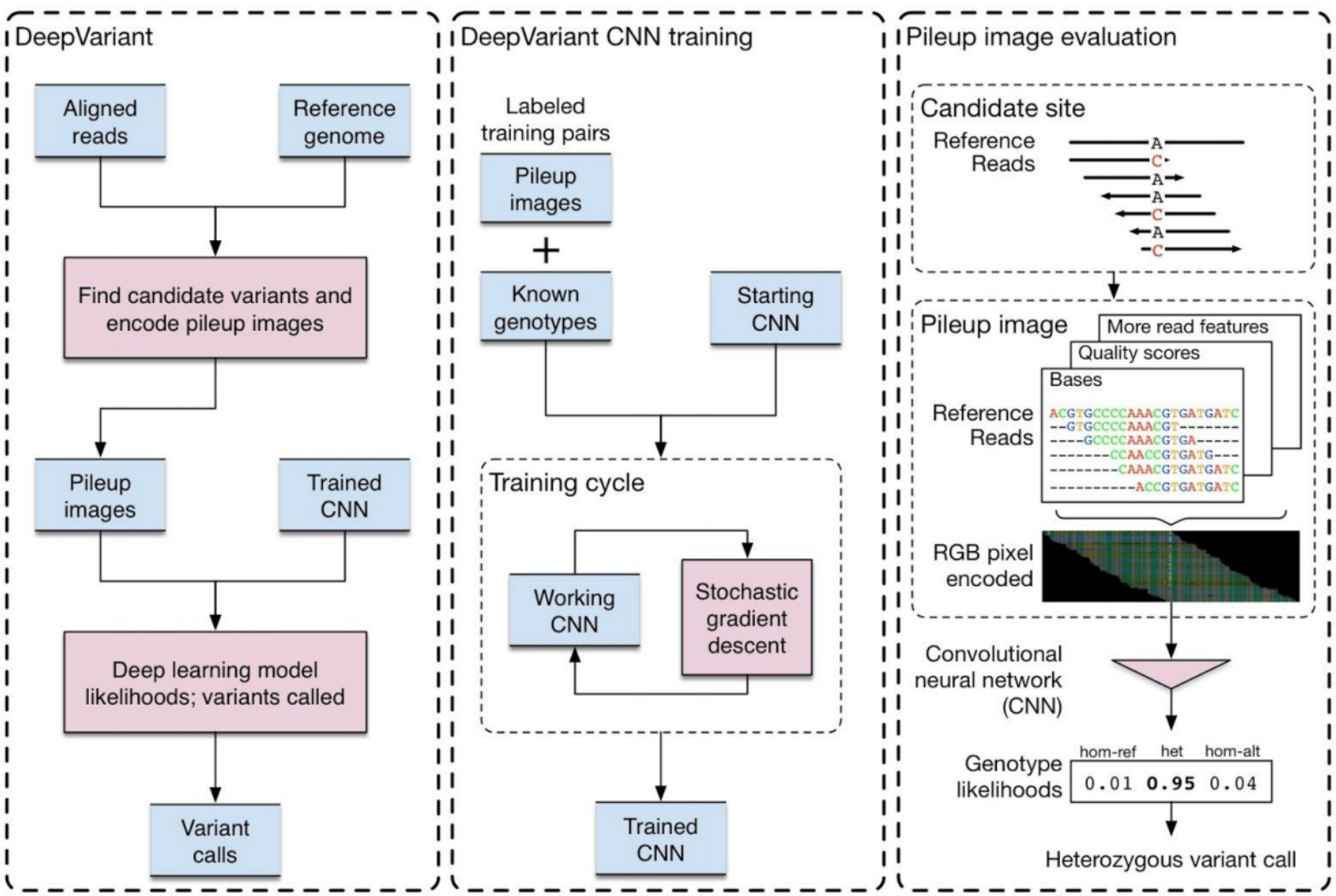
Haplotype Caller

Calls SNVs and INDELs

- **Identify**: sliding Window, count mismatches, INDELs
- **Assemble**: local re-assembly; collect most likely haplotypes; align with SW
- **Score**: use HMM model to score haplotypes
- **Genotype**: use Bayesian model to determine most likely haplotypes



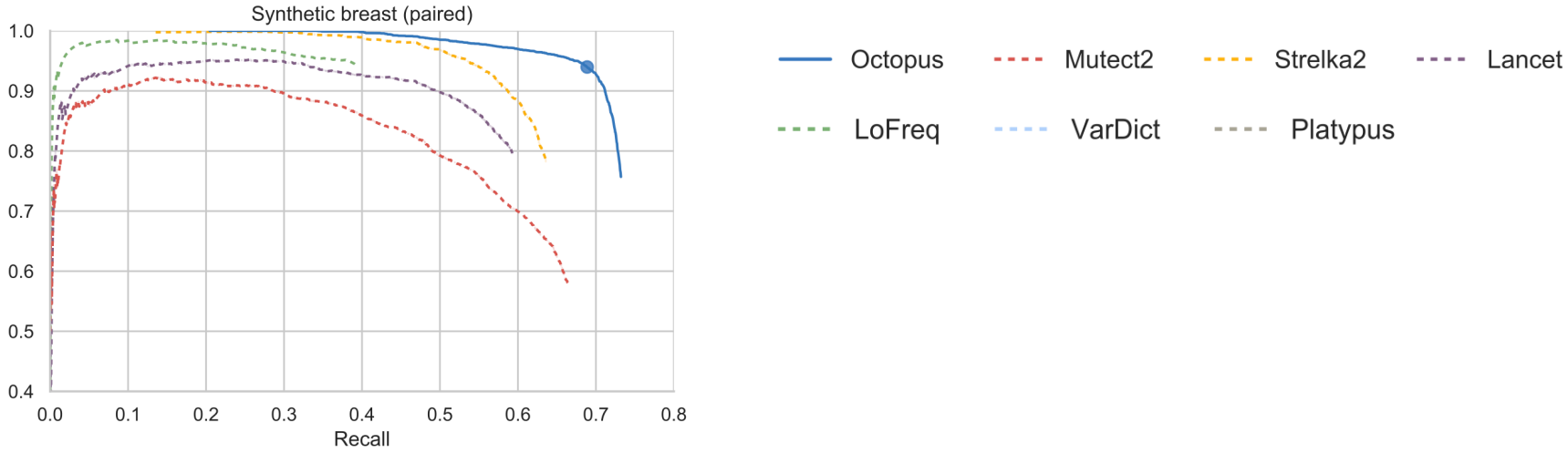
Deep Variant



A universal SNP and small-indel variant caller using deep neural networks
Poplin et al. (2018) https://www.nature.com/articles/nbt.4235

Octopus - A unified haplotype-based method for accurate and comprehensive variant calling

- Other haplotype variant callers for diploid genomes
- Octopus
  - Combines sequencing reads and prior information to phase-called genotypes of **arbitrary** ploidy
  - Accurately calls germline variants in individuals
  - More sensitive to low-frequency somatic variation, yet calls considerably fewer false positives than other methods



Cooke, D.P., Wedge, D.C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol* (2021).

Variant callers

https://github.com/deaconjs/ThousandVariantCallersRepo

README.md

ThousandVariantCallersRepo

The Thousand Variant Callers Project is a comprehensive survey of software available (~250 pack
mutations in human DNA. Included are links to published studies and software repositories, as w
each on benchmarking, algorithmic details, application notes, and so on.

See the wiki page or browse the markdown files for the surveys.

Tictac parallelizes variant calling, and contains scripts for a variety of callers from this survey.

This is a free, community-driven resource so if you see something incorrect or incomplete or hav
a hand! To add content please clone the repo, edit the wiki markdown files, and make a pull requ

SNP Variant Callers

caller	pubyear	from	study	source	algorithm
graphtyper	2017	deCODE genetics	study	source	Population-scale genotyping using pangenome graphs
muse	2016	MD Anderson Cancer Center	study	source	F81 Markov Substitution Model
sinvict	2016	Simon Frasier University, Canada	study	source	
multigems	2016	University of California, Riverside	study	source	Multinomial Bayesian, base and alignment quality priors
somaticseq	2015	Roche Bina	study	source	meta-caller, decision tree
discosnp	2015	Genscale France	study	source	reference-free, de bruijn graph
2kplus2	2015	Norwich Research Park, UK, Sainsbury lab	study	source	reference-free, de bruijn graph
exscalibur	2015	University of Chicago	study	source	
multisnv	2015	Cambridge Tavare	study	source	joint paired, timepoint pooling
rarevator	2015	University of Florence	study	source	Fisher's exact test, conserved loci only
snv-ppilp	2015	University of Helsinki, Finland	study	source	perfect phylogeny/integer linear programming
platypus	2014	U Oxford	study	source	Haplotype, bayesian, multi-sample, local realignment
baysic	2014	Baylor/Genformatic LLC	study	source	Meta-caller, Bayesian, unsupervised
haprnuc	2014	Kyoto University, Japan	study	source	Haplotype, Bayesian HMM
snpest	2014	U Copenhagen	study	source	reference-free, generative probabilistic
variantmaster	2014	Geneva Medical School, Switzerland	study	source	reference-free, pedigree inference
mutect	2013	Broad Getz	study	source	Beta-binomial, Variable Allele Fraction, filter population SNPs
niks	2013	Max Planck Institute for Plant Breeding Research, Germany	study	source	
ebcall	2013	Vanderbilt Zhao	study	source	Heuristic, multiple feature
shearwater	2013	U Cambridge/Welcome Trust	study	source	Beta-binomial, DeepSNV with aggregate control counts
shimmer	2013	NHGRI Larsen	study	source	Fisher's exact test, variant read count > N
bubbleparse	2013	Norwich Research Park Sainsbury Lab, UK	study	source	Reference-free, de Bruijn graph
cake	2013	Welcome Trust Adams	study	source	Meta-caller, simple 2x consensus, post-filter
denovogear	2013	WashU St Louis Conrad	study	source	Beta-binomial, pedigree
qsnp	2013	U Queensland	study	source	Heuristic, min 3 reads, post-filter
rvd	2013	Stanford University School of Medicine	study	source	Beta-binomial
seurat	2013	Translational Genomics Research Institute	study	source	Joint-paired, beta-binomial
snptools	2013	Baylor College of Medicine	study	source	Haplotype, Bayesian HMM
vcmm	2013	RIKEN Japan	study	source	Multinomial Bayesian, priors corrected Illumina q-score
vip	2013	Case Western, Li lab	study	source	Overlapping Pools
virmid	2013	UCSD Bafna	study	source	Joint-paired, Beta-binomial, purity estimation
varscan2	2012	WashU St Louis Wilson	study	source	Heuristic, min 3 reads, filter
jointsnvmix	2012	U British Columbia Vancouver	study	source	Joint-paired, Beta-binomial

Variant quality score recalibration

Purpose

- Assign a well-calibrated probability to each variant call
- Uses a list of **true variant sites** as input (HapMap, 1000Genomes, own set)

1 - Create recalibration file

- Takes the overlap of the training/truth resource sets and of your callset
- Models the distribution relative to specified annotations (depth, quality, read position, …) → group them into clusters
- Variants closer to cluster center → higher score than outliers

2 - Apply recalibration

- Use recalibration file to assign score
- Output field: VQSLOD

Detailed tutorials

(howto) Recalibrate base quality scores = run BQSR

Comments (27)

Objective

Recalibrate base quality scores in order to correct sequencing errors and other experimental artifacts.

Prerequisites

- TBD

Steps

1. Analyze patterns of covariation in the sequence dataset
2. Do a second pass to analyze covariation remaining after recalibration
3. Generate before/after plots
4. Apply the recalibration to your sequence data

1. Analyze patterns of covariation in the sequence dataset

Action

Run the following GATK command:

```
java -jar GenomeAnalysisTK.jar \
    -T BaseRecalibrator \
    -R reference.fa \
    -I realigned_reads.bam \
    -L 20 \
    -knownSites dbsnp.vcf \
    -knownSites gold_indels.vcf \
    -o recal_data.table
```

Expected Result

This creates a GATKReport file called `recal_data.grp` containing several tables. These tables contain the covariation data that will be used in a later step to recalibrate the base qualities of your sequence data.

It is imperative that you provide the program with a set of known sites, otherwise it will refuse to run. The known sites are used to build the covariation model and estimate empirical base qualities. For details on what to do if there are no known sites available for your organism of study, please see the online GATK documentation.

To consider ...

- Correctly formatted reference genome

Important note about human genome reference versions

If you are using human data, your reads must be aligned to one of the official b3x (e.g. b36, b37) or hg1x (e.g. hg18, hg19) references. The contig ordering in the reference you used must exactly match that of one of the official references canonical orderings. These are defined by historical karotyping of largest to smallest chromosomes, followed by the X, Y, and MT for the b3x references; the order is thus 1, 2, 3, ..., 10, 11, 12, ... 20, 21, 22, X, Y, MT. The hg1x references differ in that the chromosome names are prefixed with "chr" and chrM appears first instead of last. The GATK will detect misordered contigs (for example, lexicographically sorted) and throw an error. This draconian approach, though unnecessary technically, ensures that all supplementary data provided with the GATK works correctly. You can use ReorderSam to fix a BAM file aligned to a missorted reference sequence.

http://www.broadinstitute.org/gatk/guide/article?id=1213

- BAM file
  - sorted
  - indexed
  - with RG

Ensemble variant calling

Method

- Combine multiple VCF caller outputs into one callset
- Specify how many callers need to identify a variant (heuristic step)
- Use included and excluded variants to train a support vector machine
  → use this classifier to identify trusted variants

Validation

- Used a pair of replicates
- Compared to variants from a single calling method, the ensemble method produced **more concordant variants** when comparing the replicates, with **fewer discordants**

https://github.com/chapmanb/bcbio.variation.recall

Benchmarking Variant Callers

- Genome in a Bottle (GIAB) benchmarks

Reference Build	Benchmark Set	Reference Coverage	SNVs	Indels	Base pairs in Seg Dups and low mappability
GRCh37	v3.3.2	87.8	3,048,869	464,463	57,277,670
GRCh37	v4.2.1	94.1	3,353,881	522,388	133,848,288
GRCh38	v3.3.2	85.4	3,030,495	475,332	65,714,199
GRCh38	v4.2.1	92.2	3,367,208	525,545	145,585,710

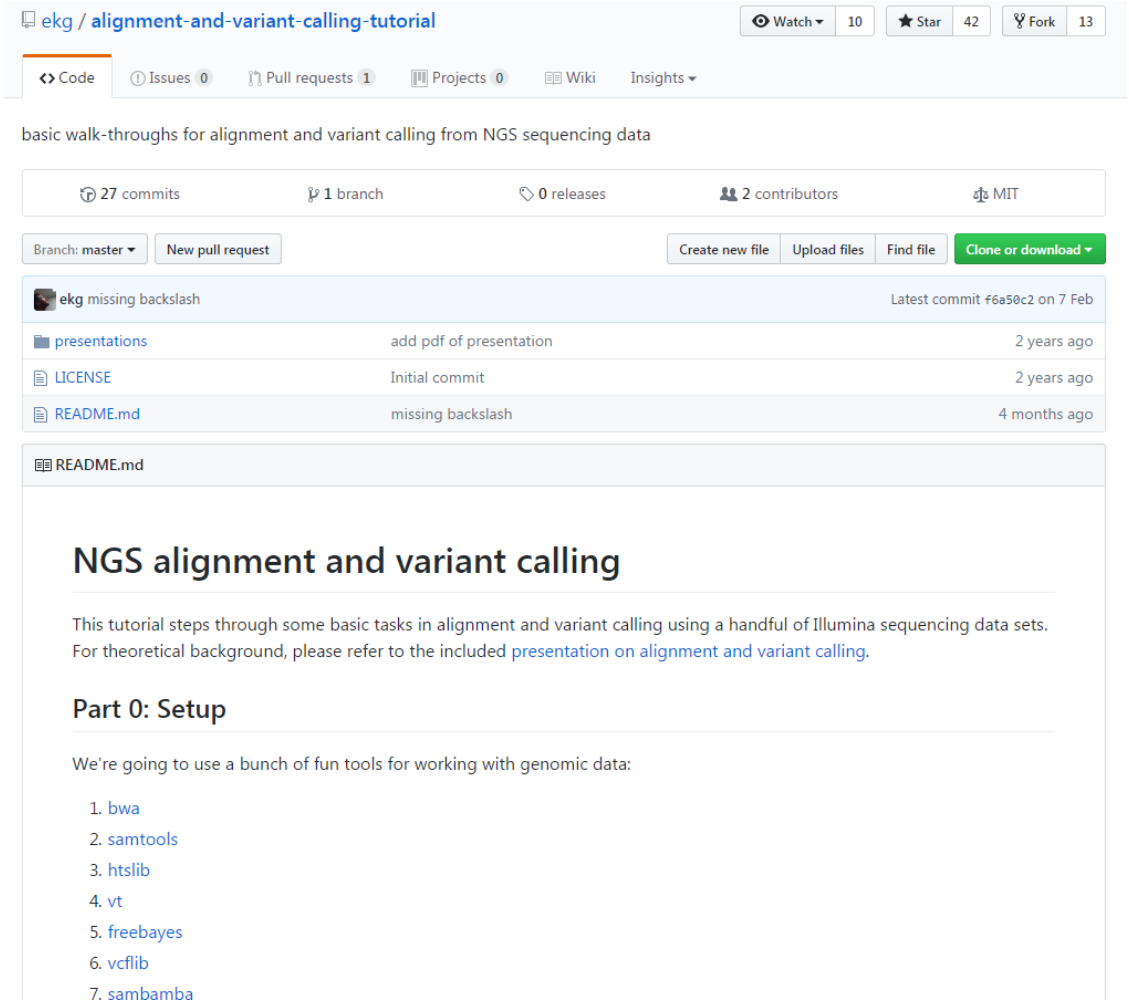
Benchmarking challenging small variants with linked and long reads

Justin Wagner, Nathan D Olson, Lindsay Harris, Ziad Khan, Jesse Farek, Medhat Mahmoud, Ana Stankovic, Vladimir Kovacevic, Aaron M Wenger, William J Rowell, Chunlin Xiao, Byunggil Yoo, Neil Miller, Jeffrey A Rosenfeld, Bohan Ni, Samantha Zarate, Melanie Kirsche, Sergey Aganezov, Michael Schatz, Giuseppe Narzisi, Marta Byrska-Bishop, Wayne Clarke, Uday S Evani, Charles Markello, Kishwar Shafin, Xin Zhou, Arend Sidow, Vikas Bansal, Peter Ebert, Tobias Marschall, Peter Lansdorp, Vincent Hanlon, Carl-Adam Mattsson, Alvaro Martinez Barrio, Ian T Fiddes, Arkarachai Fungtammasan, Chen-Shan Chin, Fritz J Sedlazeck, Andrew Carroll, Marc Salit, Justin M Zook, Genome in a Bottle Consortium

doi: https://doi.org/10.1101/2020.07.24.212712

Useful information on how-to perform variant calling

https://github.com/ekg/alignment-and-variant-calling-tutorial



ekg / alignment-and-variant-calling-tutorial

Watch 10 | Star 42 | Fork 13

Code | Issues 0 | Pull requests 1 | Projects 0 | Wiki | Insights

basic walk-throughs for alignment and variant calling from NGS sequencing data

27 commits | 1 branch | 0 releases | 2 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

ekg missing backslash — Latest commit f6a50c2 on 7 Feb

presentations	add pdf of presentation	2 years ago
LICENSE	Initial commit	2 years ago
README.md	missing backslash	4 months ago

README.md

NGS alignment and variant calling

This tutorial steps through some basic tasks in alignment and variant calling using a handful of Illumina sequencing data sets. For theoretical background, please refer to the included presentation on alignment and variant calling.

Part 0: Setup

We're going to use a bunch of fun tools for working with genomic data:

1. bwa
2. samtools
3. htslib
4. vt
5. freebayes
6. vcflib
7. sambamba

Other types of variant calling

Structural variant calling

- Identify large deletions, insertions, translocations, inversions

Copy Number Variation (CNV) calling

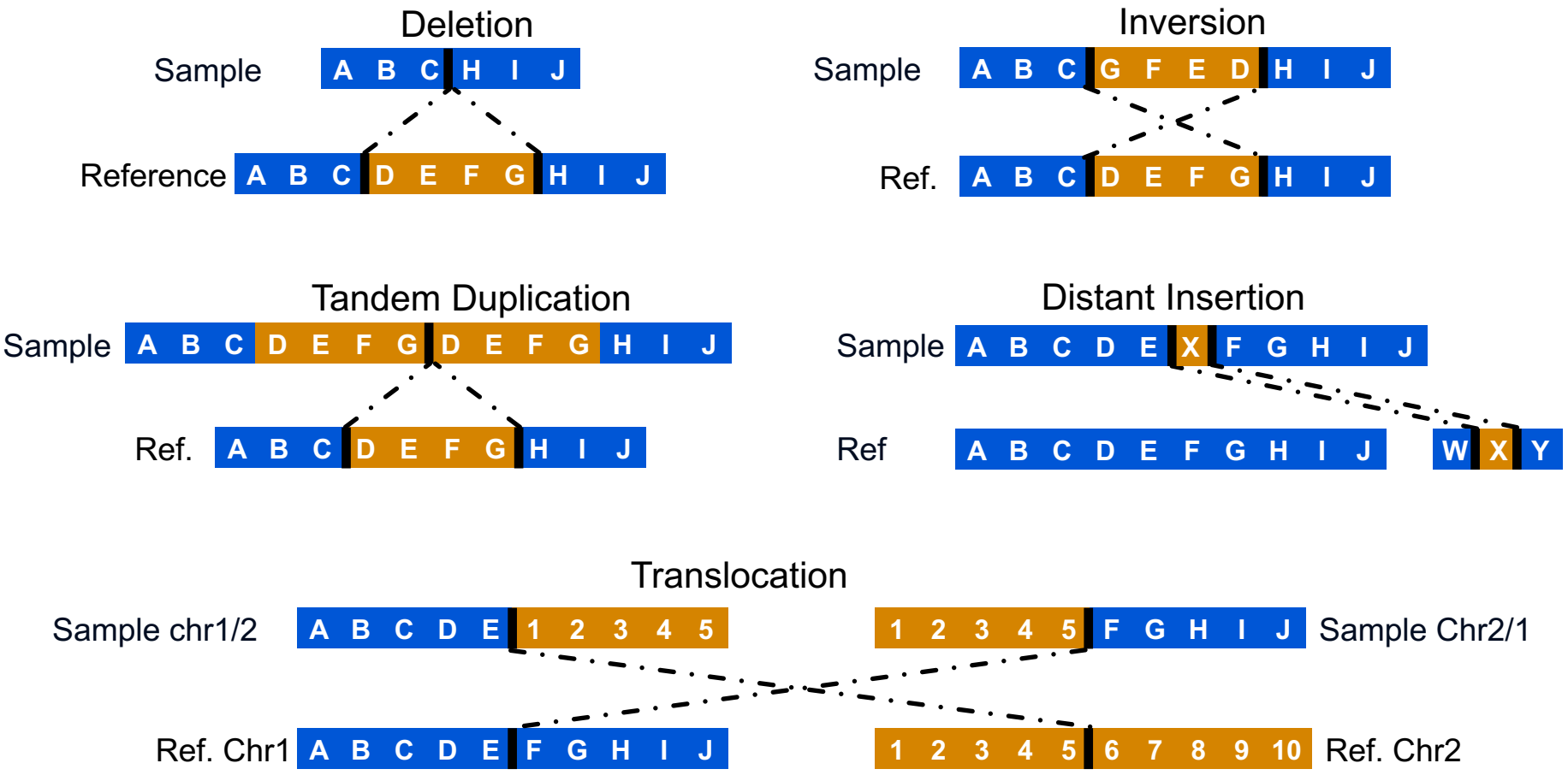
- Which parts of the genome are amplified or deleted?

Somatic variant calling

- Find acquired mutations

Structural variation calling

Structural variations



Why is structural variation relevant / important?

They are common and **affect a large fraction of the genome**

- In total, SVs impact more base pairs than all single nucleotide differences

They are a major **driver of genome evolution**

- Speciation can be driven by rapid changes in genome architecture
- Genome instability and aneuploidy: hallmarks of solid tumor genomes

SV and human disease phenotypes

Table 2 Examples of copy number variations (CNVs) and conveyed genomic disorders[a]

Phenotype	OMIM	Locus	CNV
Mendelian (autosomal dominant)[b]			
Williams-Beuren syndrome	194050	7q11.23	del
7q11.23 duplication syndrome	609757	7q11.23	dup
Spinocerebellar ataxia type 20	608687	11q12	dup
Smith-Magenis syndrome	182290	17p11.2/*RAI1*	del
Potocki-Lupski syndrome	610883	17p11.2	dup
HNPP	162500	17p12/*PMP22*	del
CMT1A	118220	17p12/*PMP22*	dup
Miller-Dieker lissencephaly syndrome	247200	17p13.3/*LIS1*	del
Mental retardation	601545	17p13.3/*LIS1*	dup
DGS/VCFS	188400/192430	22q11.2/*TBX1*	del
Microduplication 22q11.2	608363	22q11.2	dup
Adult-onset leukodystrophy	169500	*LMNB1*	dup
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	256100	2q13/*NPHP1*	del
Gaucher disease	230800	1q21/*GBA*	del
Pituitary dwarfism	262400	17q24/*GH1*	del
Spinal muscular atrophy	253300	5q13/*SMN1*	del
beta-thalassemia	141900	11p15/*beta-globin*	del
alpha-thalassemia	141750	16p13.3/*HBA*	del

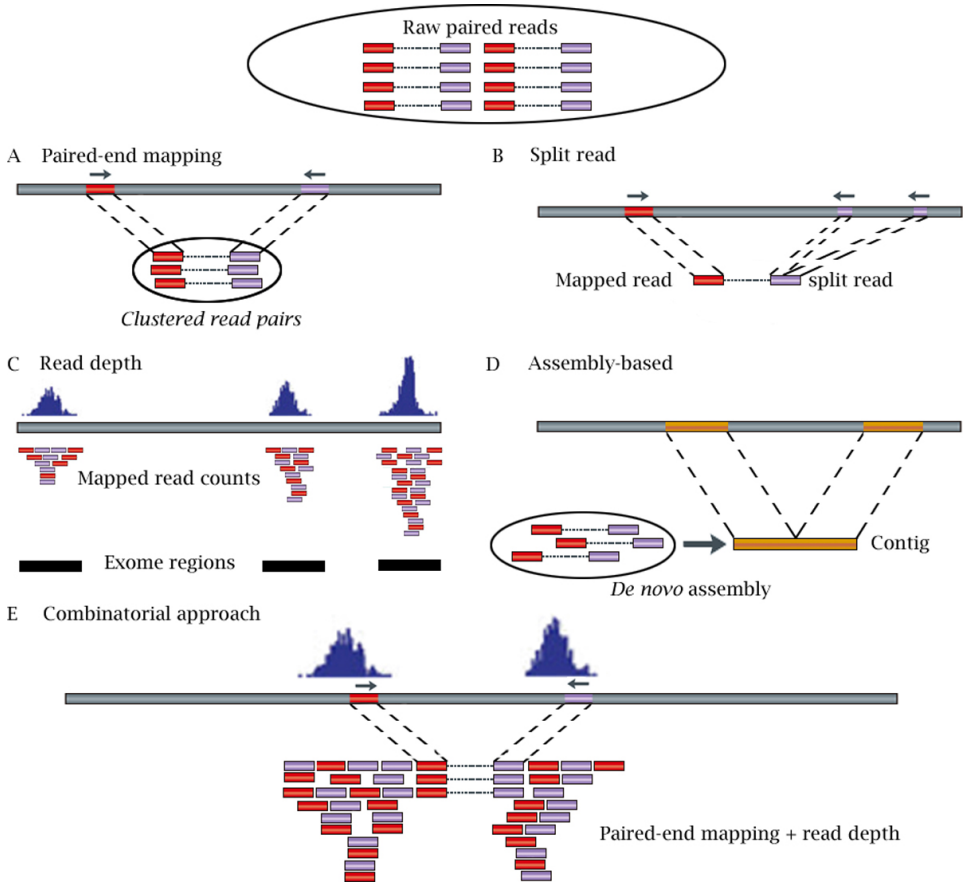
Zhang et al, 2009

SV calling

Use paired end information to detect these events

- Deviations of the expected insert size
- Presence/absence of mate pairs
- Read depth for CNVs

SV/CNV detection



A. Paired-end mapping (PEM) strategy detects SVs/CNVs through **discordantly mapped reads**. A discordant mapping is produced if the distance between two ends of a read pair is **significantly different from the average insert size**.

B. Split read (SR)-based methods use **incompletely mapped read** from each read pair to identify small SVs/CNVs.

C. Read depth (RD) approach detects by **counting the number of reads mapped** to each genomic region. In the figure, reads are mapped to three exome regions.

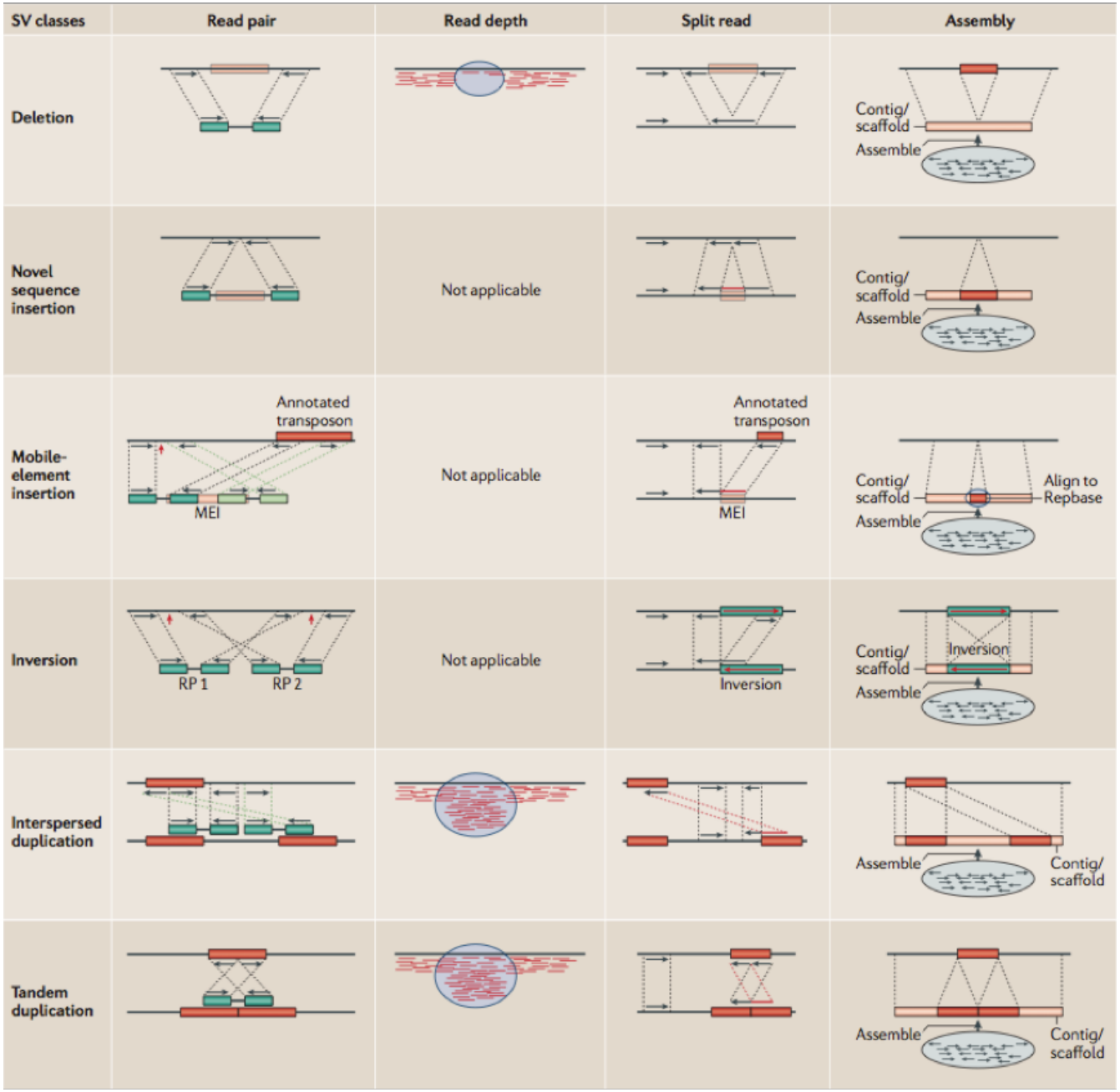
D. Assembly (AS)-based approach detects CNVs by **mapping contigs** to the reference genome.

E. Combinatorial approach combines **RD and PEM** information to detect CNVs.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3846878/

Zhao *et al. BMC Bioinformatics* 2013 **14**

Structural variation - what can we detect



Can et. al.,
Genome structural variation discovery and genotyping, Nature Rev. 2011

Structural variations - tools

Breakdancer

- Insertions, deletions, inversions, translocations
- Fast, simple to run

Pindel

- Insertions, deletions

GASVPro

- Combines read depth info along with discordant paired-read mappings
- Duplications, deletions, insertions, inversions and translocations

SVMerge

- Results from several different SV caller (Breakdancer, Pindel, SE Cluster, RDXplorer, RetroSeq)

Mavis

- post-processing of structural variant calls.
- http://mavis.bcgsc.ca/

Structural variations - tools

LUMPY

- Integrates different sequence alignment signals (read-pair, split-read and read-depth)
- https://github.com/arq5x/lumpy-sv

Manta

- Calling structural variants, medium-sized indels and large insertions
- Very fast

Delly

- Integrates short insert paired-ends, long-range mate-pairs and split-read alignments
- Detects CNVs, deletion, tandem duplication events, inversions or reciprocal translocations

tardis

- Rapid discovery of structural variants
- Available as Docker image

Structural variations – combination/evaluation

SURVIVOR

- Simulates SVs given a reference, number and size ranges for each SV
  insertions, deletions, duplications, inversions and translocations
  - bed file to report the locations of the simulated SVs

- Evaluates SV
  - VCF input
  - start & stop coordinates of the sim and ident SV within 1 kb (parameter)

- Filter and combine the calls from VCF files

https://www.nature.com/articles/ncomms14061

Meta SV-caller

Parliament2

- For WGS data
- Runs a combination of tools:
  - Breakdancer
  - Breakseq2
  - CNVnator
  - Delly2
  - Manta
  - Lumpy
- Merges calls with SURVIVOR

https://github.com/dnanexus/parliament2

SV – more tools

Abel *et al. Cancer Genetics* 2013 Pages 432–440

Review article

Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches

Haley J. Abel[a], Eric J. Duncavage[b,]

Show more

http://dx.doi.org/10.1016/j.cancergen.2013.11.002

Get rights and content

Next generation sequencing (NGS), or massiv
methods in which numerous sequencing reac
amounts of sequencing data for a small fraction

Table 1.
Software tools for evaluation of structural variation in NGS data

	Comment	Download link
Translocations and Inversions		
Discordant paired end		
BreakDancer	Fast, simple to run	http://breakdancer.sourceforge.net
Hydra	Considers multiple mappings of discordant pairs	https://code.google.com/p/hydra-sv/
VariationHunter	Considers multiple mappings of discordant pairs	http://variationhunter.sourceforge.net/Home
PEMer	Simulates structural	http://sv.gersteinlab.org/pemer/introduction.html

Structural variations - challenges

Often many false positives

- Short reads + heuristic alignment + rep. genome **= systematic alignment artifacts (false calls)**
- Ref. genome errors (e.g., gaps, misassemblies)
- **ALL** SV mapping studies use strict filters for above

The false negative rate is also typically high

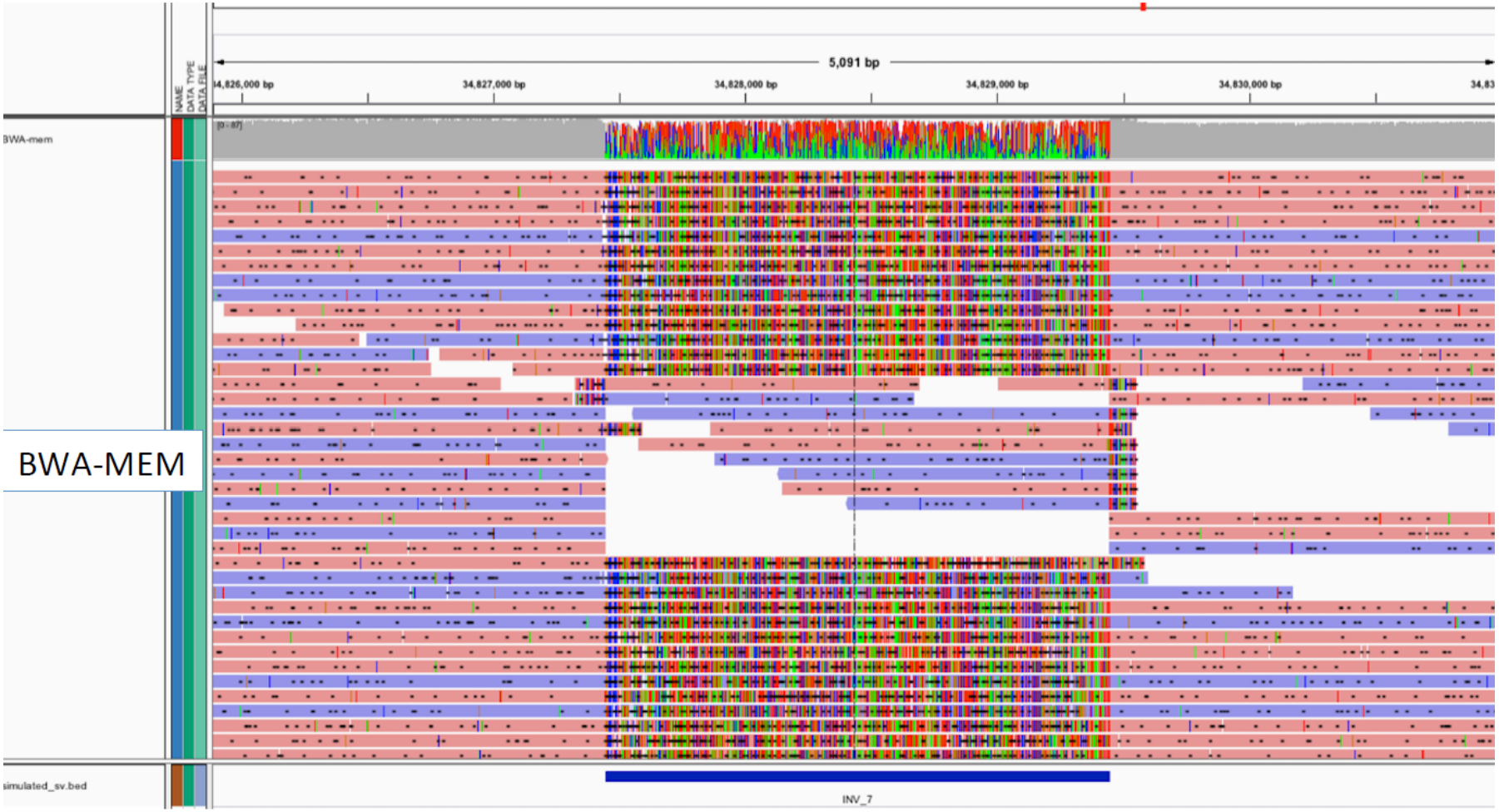
- Most current datasets have low to moderate ***physical*** coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another

Long Read Technologies

(+) SVs in repetitive regions
(+) Can identify nested SVs

(-) Higher error rate
(-) Hard to align

Hard to align



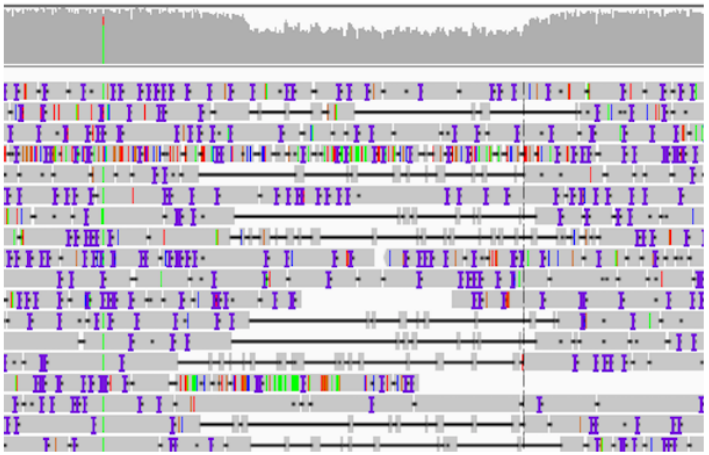
Human genome: 1kb Inversion

Improving long read alignment

- NGM – http://cibiv.github.io/NextGenMap/

1. Split the reads:
   - Translocations
   - Inversions
   - Duplications

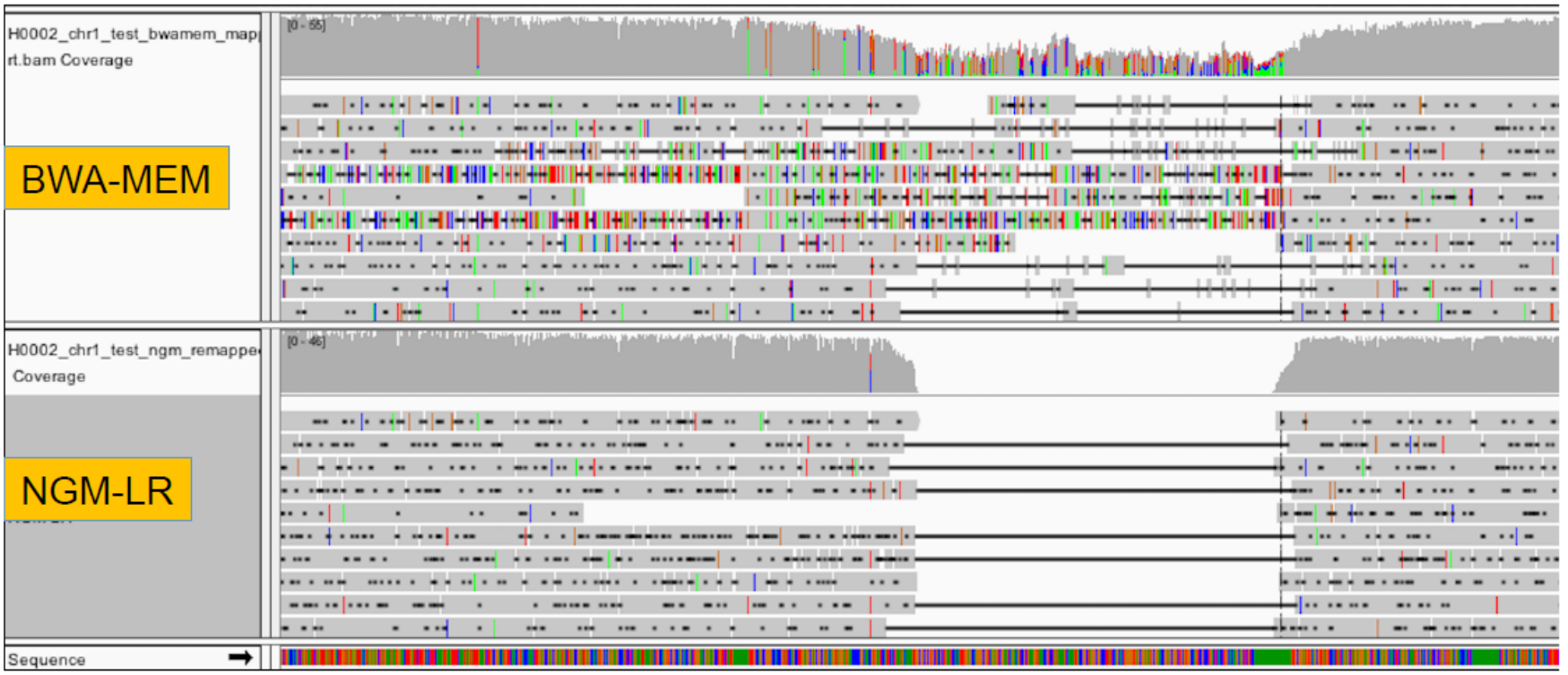
2. Improve alignment:
   - Insertions
   - Deletions



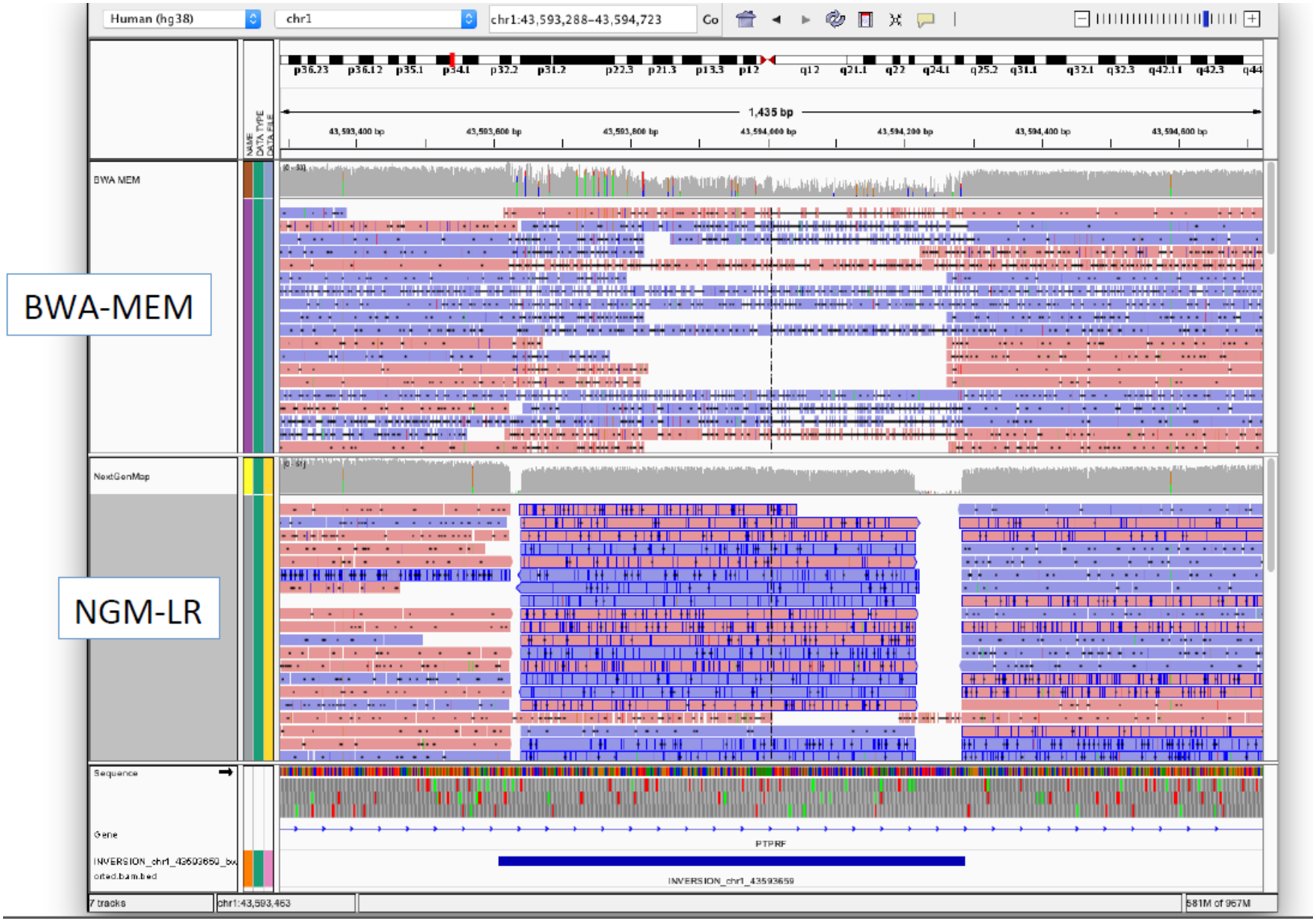
Structural variations with 3rd gen sequencing

NGM-LR + Sniffles: PacBio SV Analysis Tools

- ***1. NGM-LR***: Improve mapping of noisy long reads: improved seeding, convex gap scoring
- ***2. Sniffles***: Integrates evidence from split-reads, alignment fidelity, breakpoint concordance



NGM-LR complex SV



Benchmark for structural variant calling

12,745 isolated events

- 7,281 insertions
- 5,464 deletions

Used 19 sequence-resolved variant calling methods from diverse technologies

Already used to evaluate new tools (ClinSV, SVIM-asm, PopDel, ClipSV …)

Zook, J.M., Hansen, N.F., Olson, N.D. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38,** 1347–1355 (2020).