LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

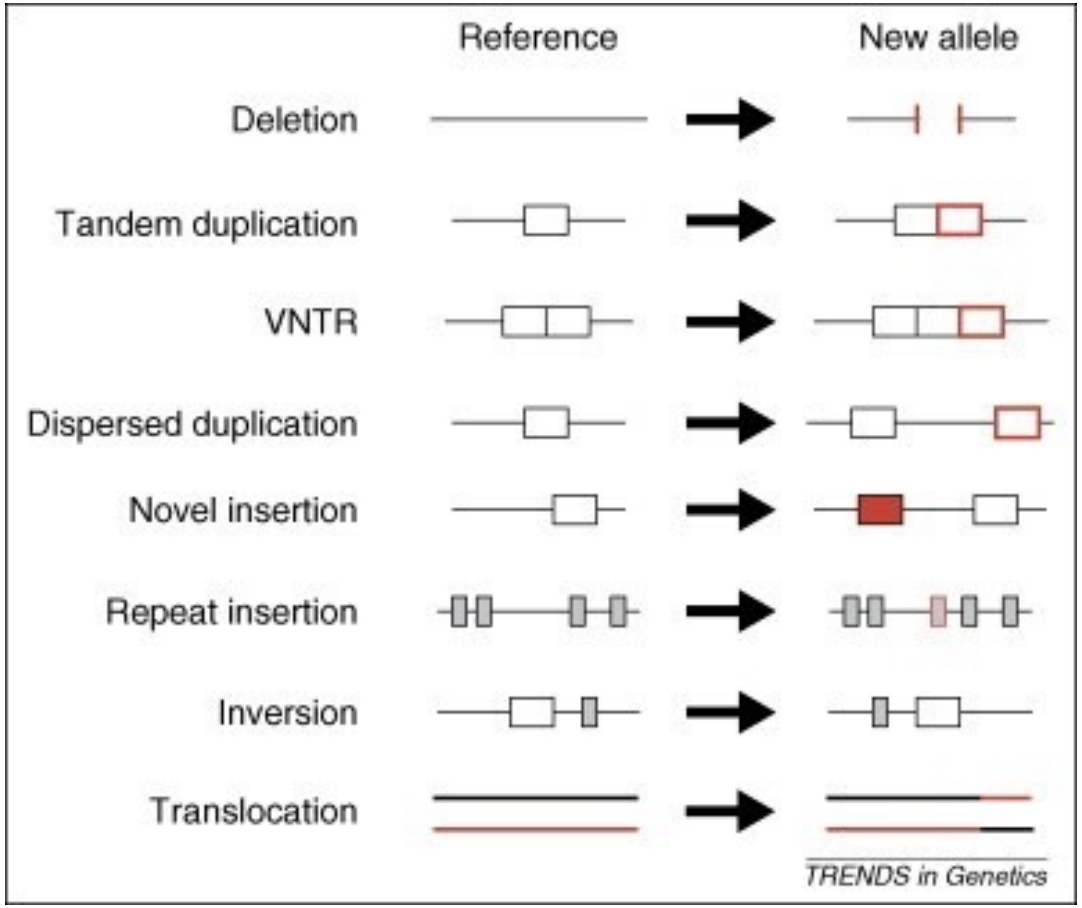
Variant Detection

London School of Hygiene and Tropical Medicine

Outline

- Types of variants
- Technologies
- Detecting variants using next generation sequencing data
- Practical

Different types of variants



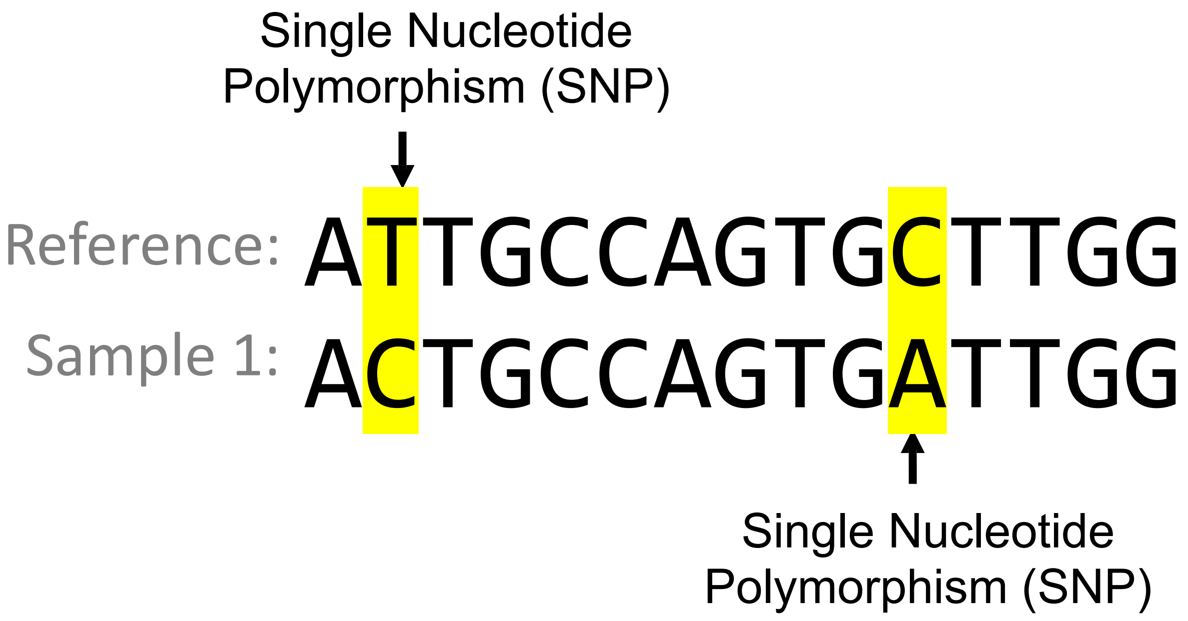
Structural Variants

- **Small scale variants**
  - Single nucleotide polymorphisms (SNPs)
  - Insertions and deletions (indels)
  - Variable tandem repeats (VNTRs)
- **Large scale variants (>1kb)**
  - Copy number variations (CNVs)
  - Insertions and deletions
  - Inversions

…and things in between small and large, and combinations of the above

Hurles et al., 2008. *Trends in Genetics*, *24*:238-245

Different types of variants

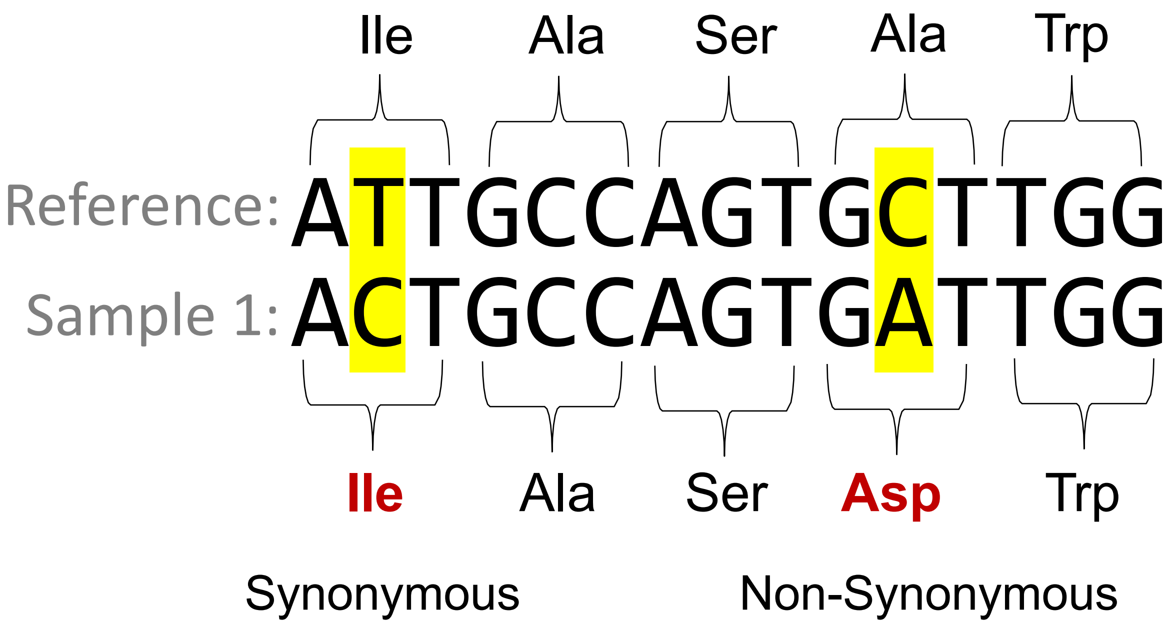


Single Nucleotide Polymorphism (SNP)

Reference: ATTGCCAGTGCTTGG

Sample 1: ACTGCCAGTGATTGG

Single Nucleotide Polymorphism (SNP)



Ile Ala Ser Ala Trp

Reference: ATTGCCAGTGCTTGG

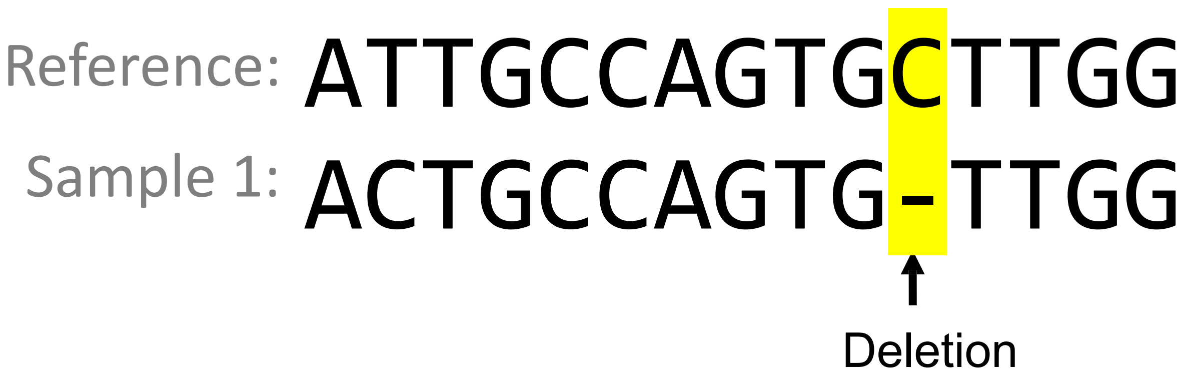
Sample 1: ACTGCCAGTGATTGG

Ile Ala Ser **Asp** Trp

Synonymous

Non-Synonymous

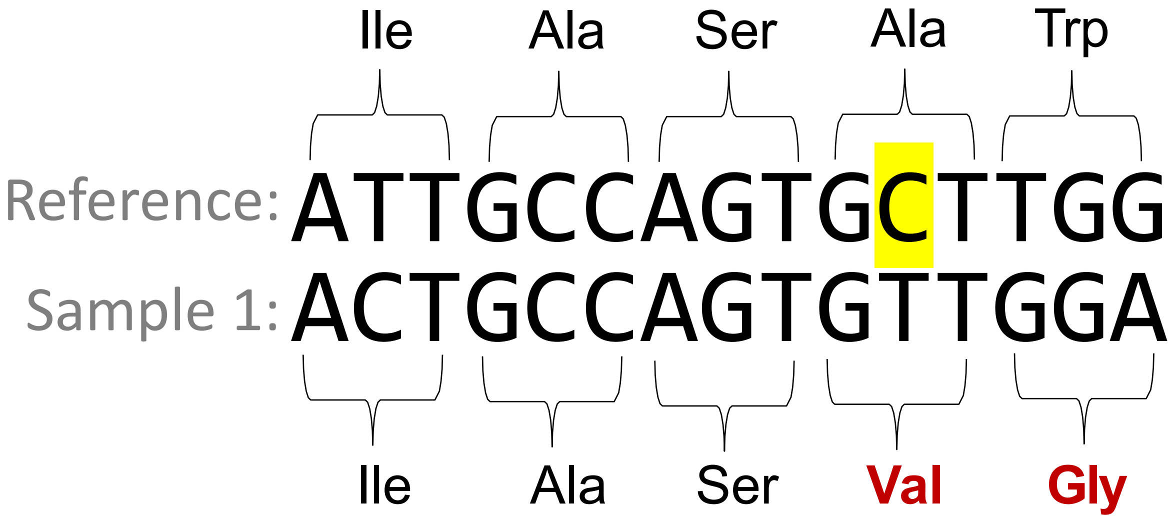
Different types of variants



Reference: ATTGCCAGTGCTTGG

Sample 1: ACTGCCAGTG-TTGG

Deletion



Ile Ala Ser Ala Trp

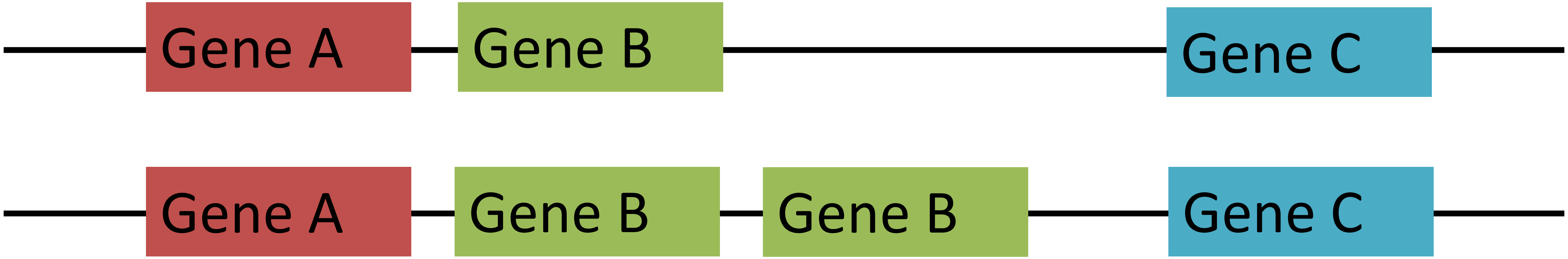
Reference: ATTGCCAGTGCTTGG

Sample 1: ACTGCCAGTGTTGGA

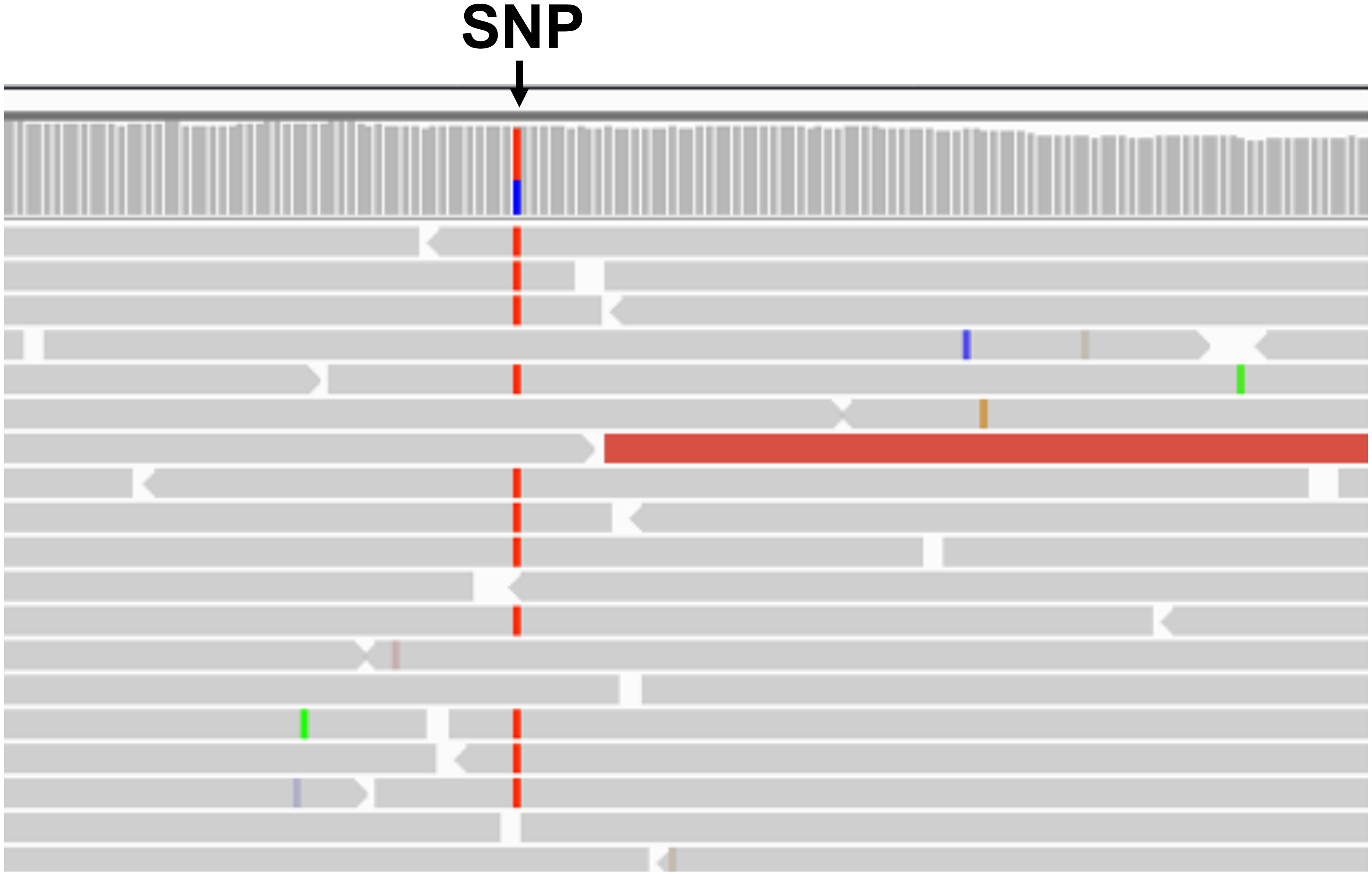
Ile Ala Ser **Val** **Gly**

Different types of variants

Copy Number Variation (CNV)



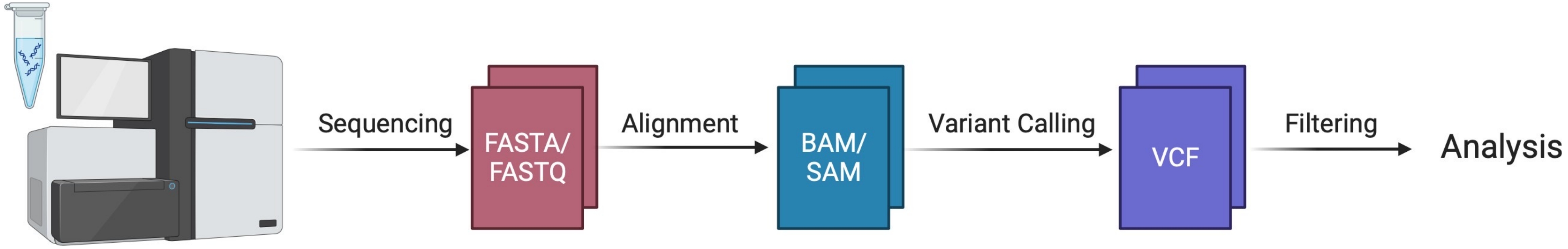
Detecting SNPs from alignments



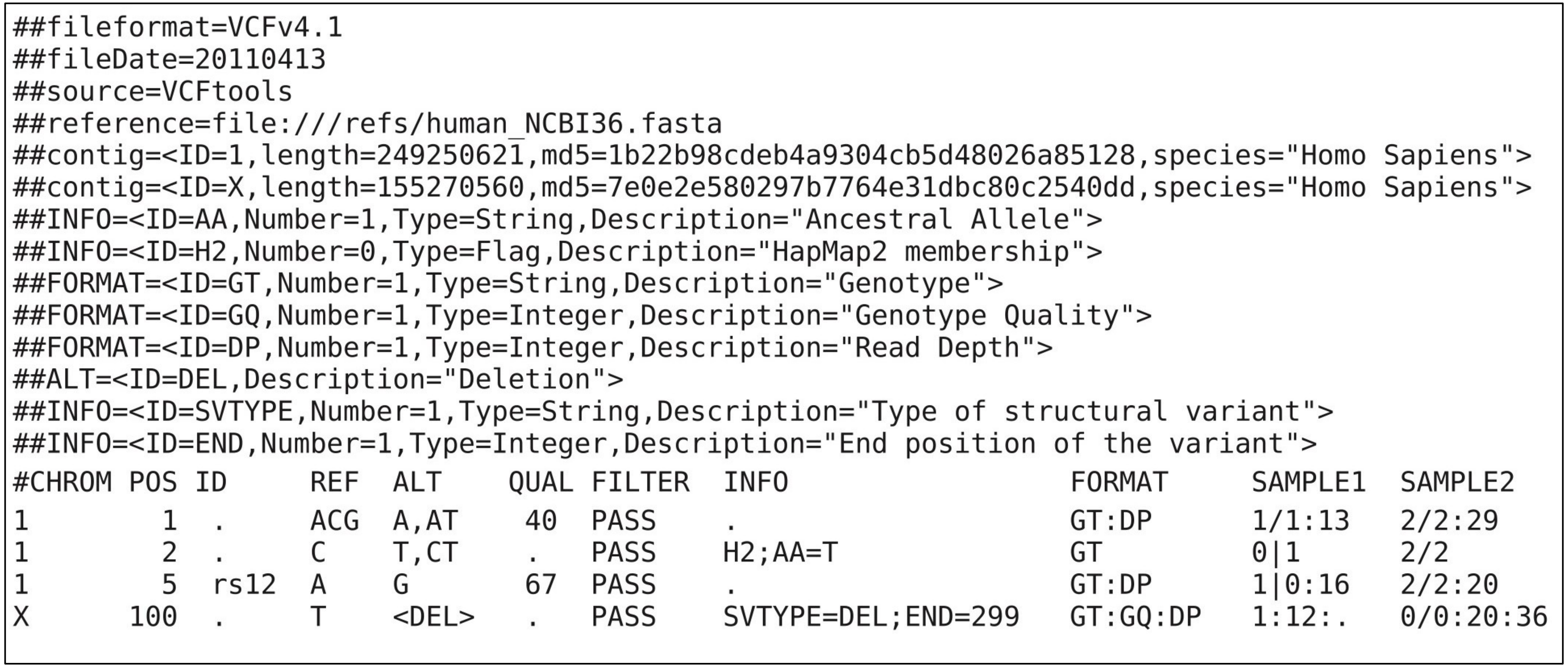
IGV View

Variant Discovery Pipeline for NGS data

- **Aim**:
  - Start with sequencing reads and perform a series of steps to determine the presence of genetic variants (e.g., SNPs, indels, CNVs)
- **Process**:
  - Creation of the variant call format (VCF) file…



Variant Call Format (VCF)

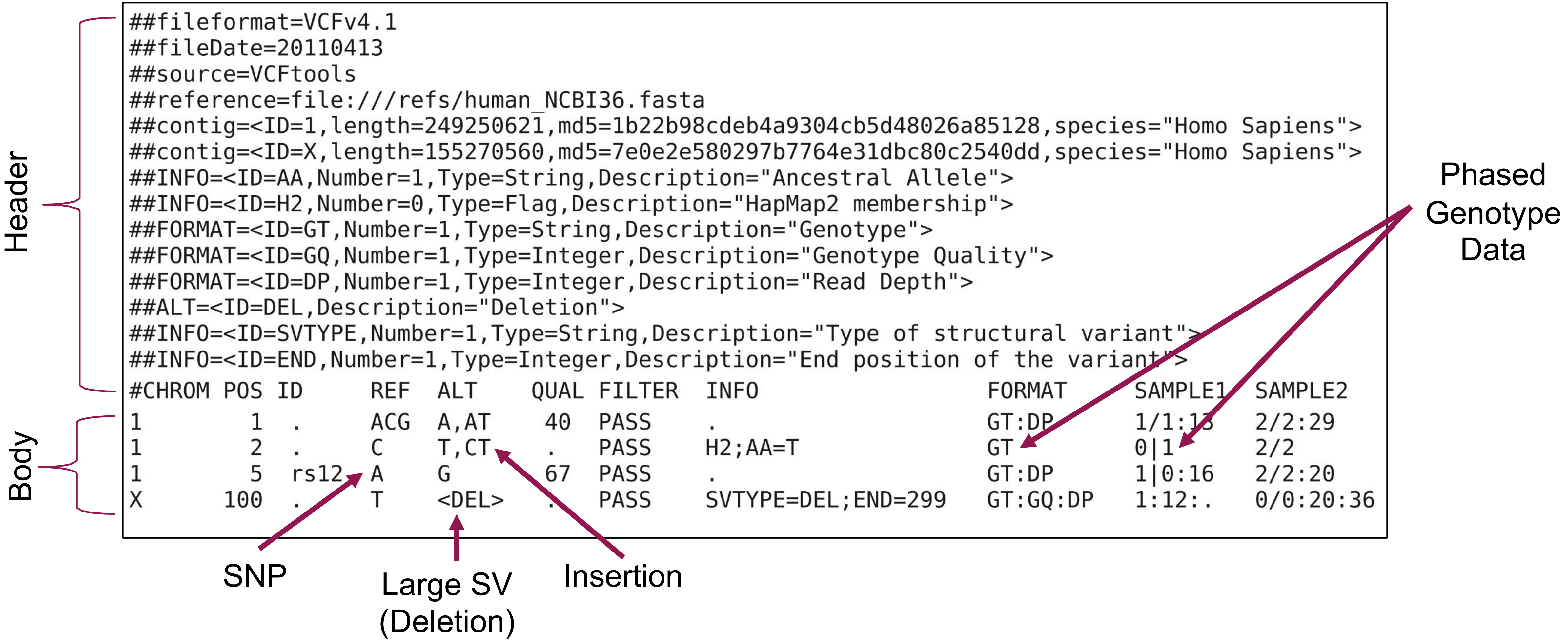


```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF ALT    QUAL FILTER INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1  .     ACG A,AT    40  PASS   .                    GT:DP      1/1:13   2/2:29
1        2  .     C   T,CT    .   PASS   H2;AA=T              GT         0|1      2/2
1        5  rs12  A   G       67  PASS   .                    GT:DP      1|0:16   2/2:20
X      100  .     T   <DEL>   .   PASS   SVTYPE=DEL;END=299   GT:GQ:DP   1:12:.   0/0:20:36
```

VCFs are unambiguous, scalable, and flexible

Danecek et al., 2011. *Bioinformatics, 27:2156-2158*

Variant Call Format (VCF)

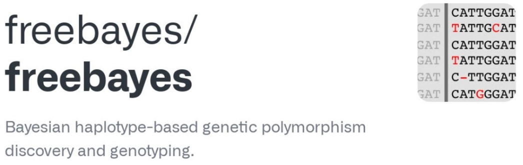


```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF ALT    QUAL FILTER INFO                 FORMAT     SAMPLE1 SAMPLE2
1        1 .      ACG A,AT    40  PASS   .                    GT:DP      1/1:13  2/2:29
1        2 .      C   T,CT     .  PASS   H2;AA=T              GT         0|1     2/2
1        5 rs12   A   G       67  PASS   .                    GT:DP      1|0:16  2/2:20
X      100 .      T   <DEL>    .  PASS   SVTYPE=DEL;END=299   GT:GQ:DP   1:12:.  0/0:20:36
```

Danecek et al., 2011. *Bioinformatics, 27:2156-2158*

SNP and short indel calling tools

- Genome Analysis Toolkit (GATK)*
- bcftools *mpileup* (MAQ SNP Caller)*
- FreeBayes*
- VarDict
- CASAVA SNP Caller (Illumina)
- Commercial Packages (CLC Bio, Genomatix)



freebayes/
freebayes

Bayesian haplotype-based genetic polymorphism discovery and genotyping.

samtools/**bcftools**

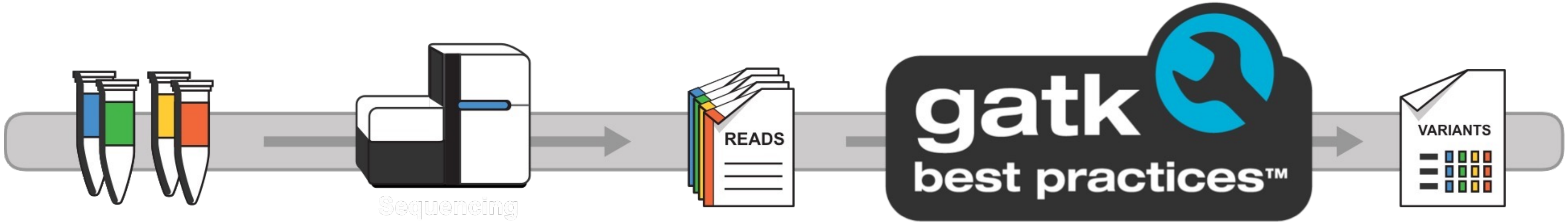
This is the official development repository for BCFtools. See installation instructions and other documentation here http://samtools.github.io/bcftools/howtos/install.html

…all essentially are performing the same thing (variant calling) but employ different models.

***Used by the open-source community**

Variant discovery with *GATK*

- GATK is a package of command-line tools written in Java
- GATK provides end-to-end workflows called *best practices*
- Supports most common file formats e.g., VCF
- Easily extendable and adaptable, but **slow**



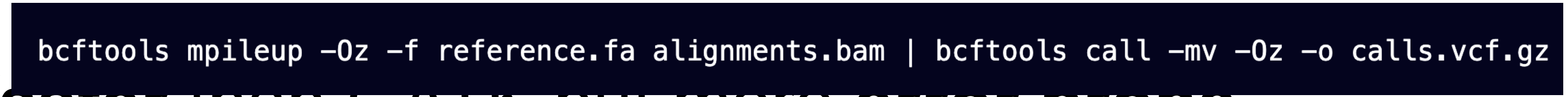
https://gatk.broadinstitute.org/hc/en-us

Variant discovery with *bcftools*

Two-step procedure where:

1. `bcftools mpileup` summarises the coverage of each base at each genomic position from a BAM file and their associated alleles

2. `bcftools call` with the `–v` flag applies the statistical model to make variant calls and generate the output as a VCF



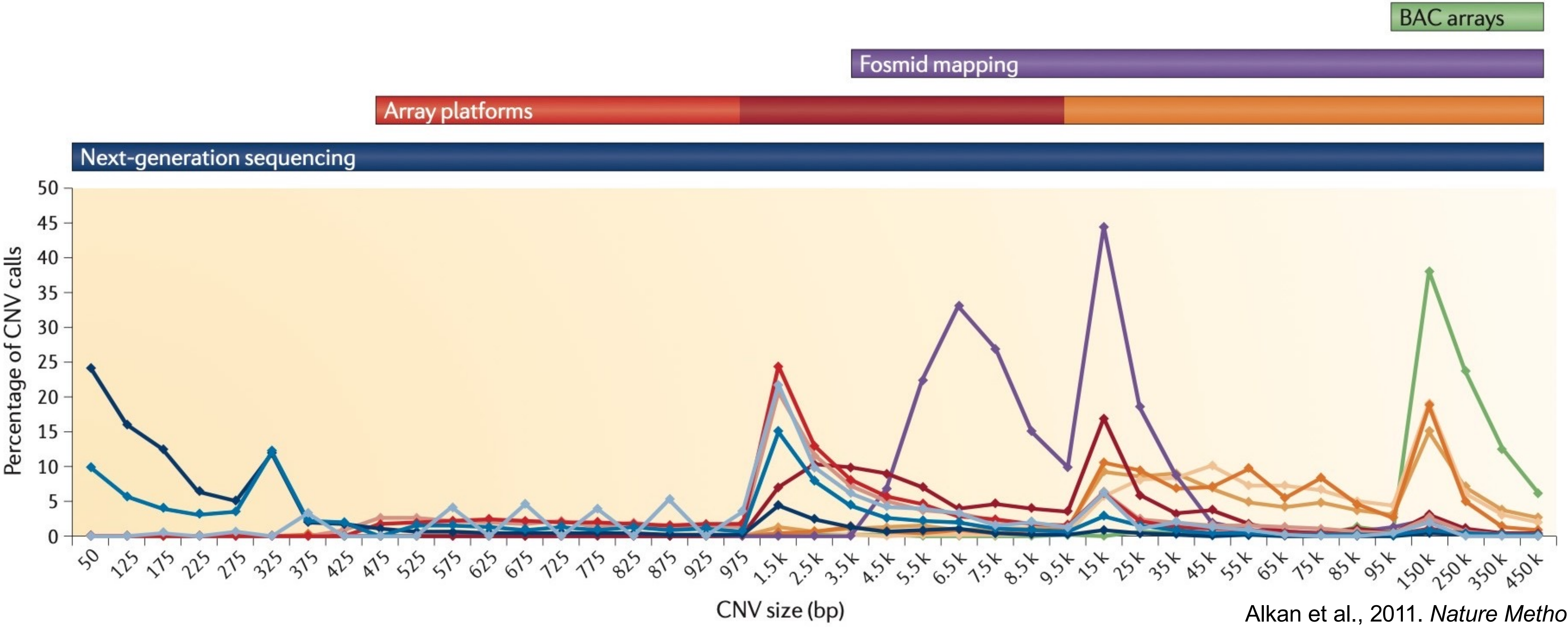
```
bcftools mpileup –Oz –f reference.fa alignments.bam | bcftools call –mv –Oz –o calls.vcf.gz
```

- Much **faster** than GATK but more **error prone**

https://samtools.github.io/bcftools/howtos/variant-calling.html

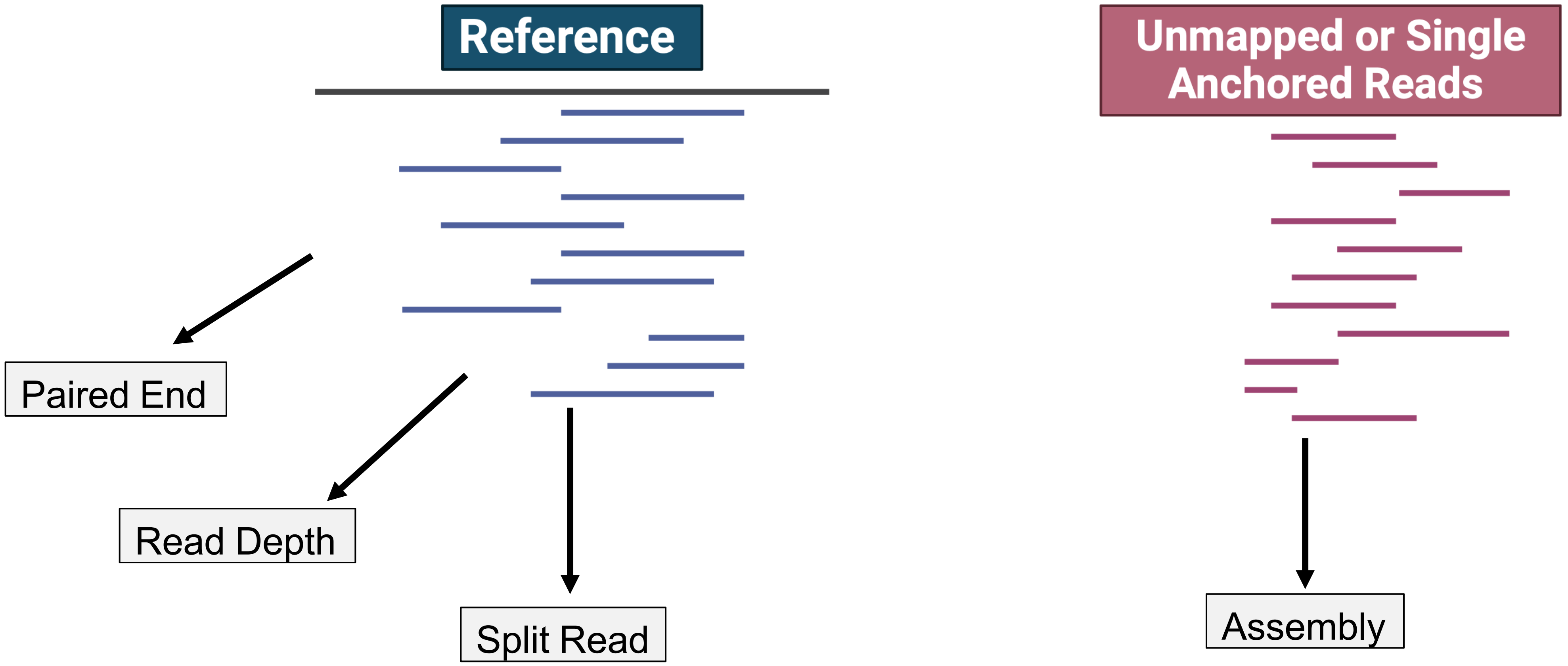
Discovery of large structural variants

Next generation and third generation sequencing offers the widest range of detection of CNVs.



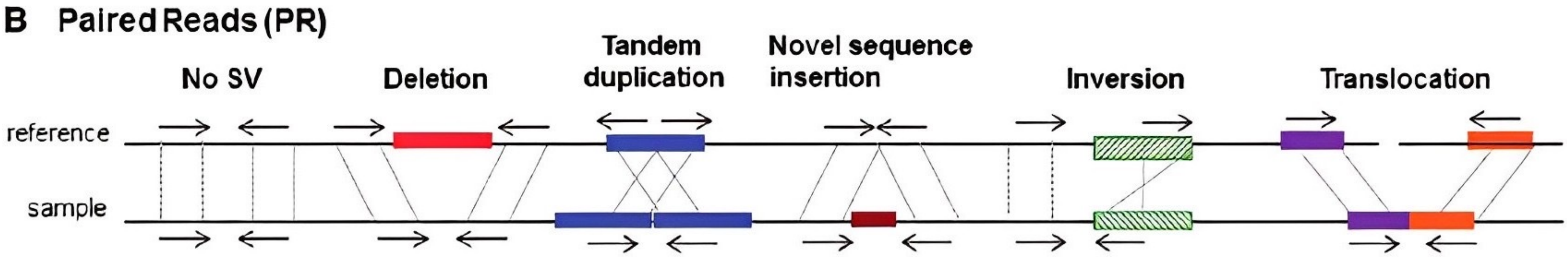
Alkan et al., 2011. *Nature Methods* 8:61–65

Approaches to identify structural variants



Paired End Approach

- Assesses span and orientation of paired end reads
  - If the inferred mapping span is ***greater*** than expected ⟶ deletion
  - If the inferred mapping span is ***less*** than expected ⟶ insertion

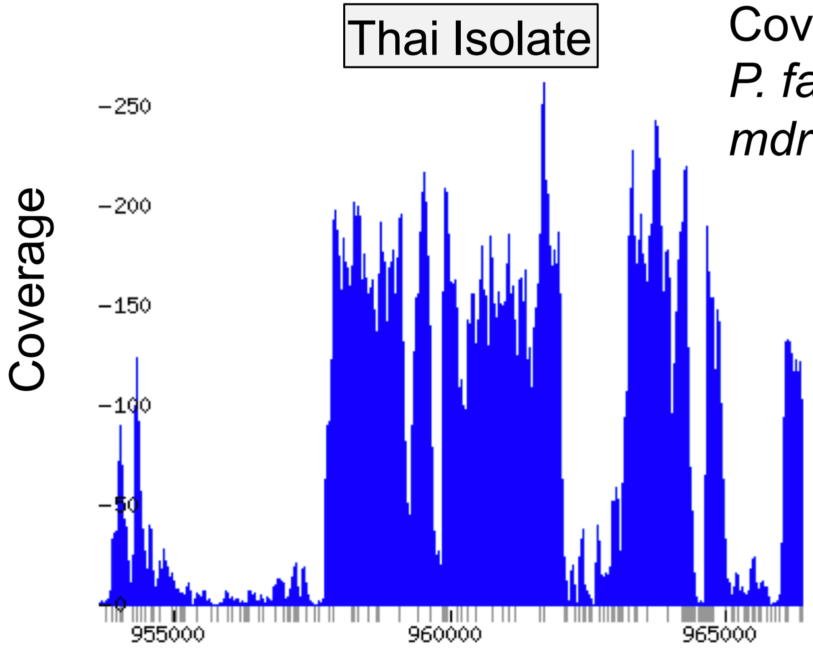


Escarimís et al., 2015. *Briefings in Functional Genomics, 14:305-314*

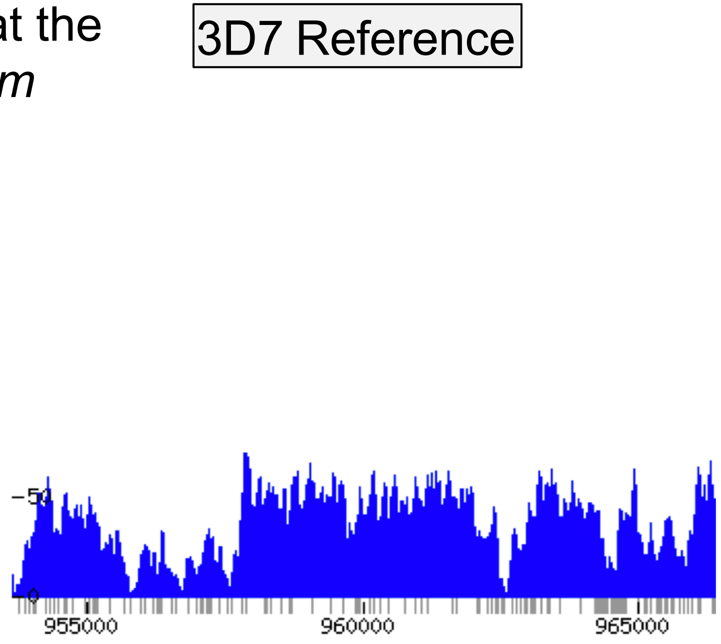
Read Depth Approach

Detects deletions or duplications based on divergences in mapping depth:

- Low or zero coverage suggests a deletion
- Excess coverage suggests a duplication

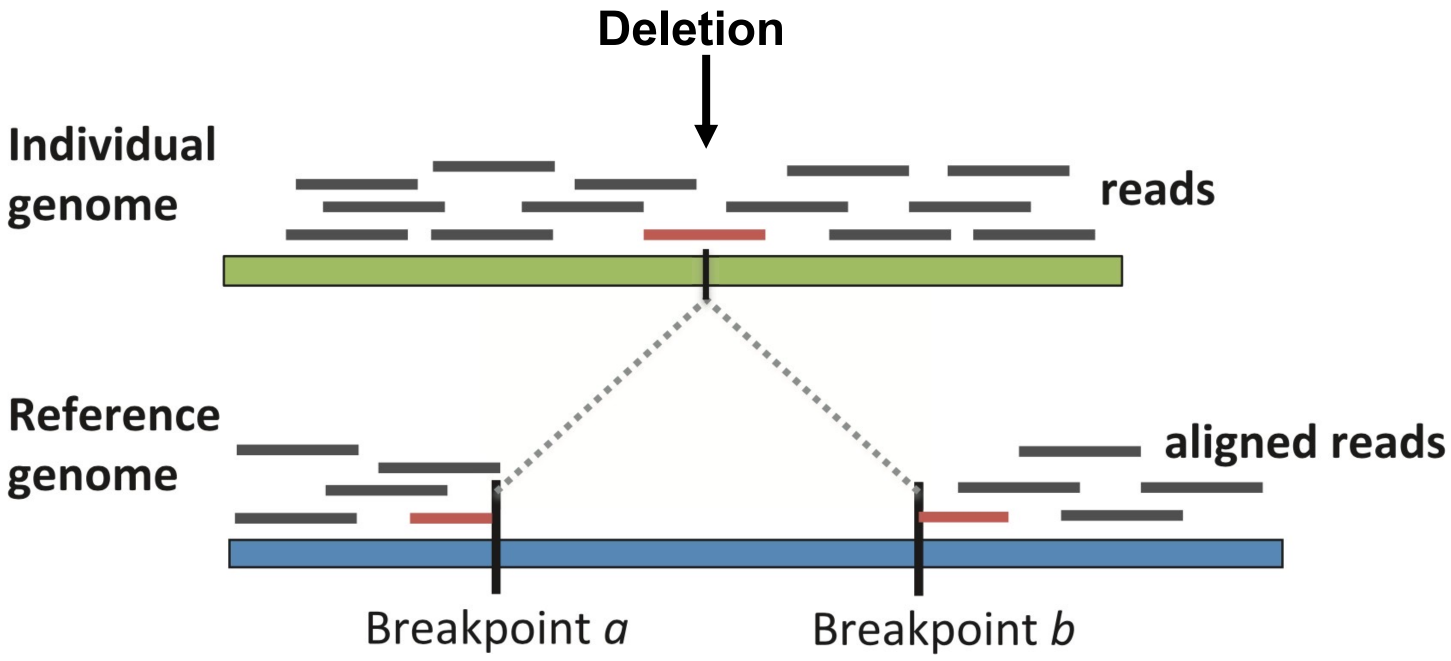


Coverage at the *P. falciparum mdr1* gene



Split Read Approach

- Defines a breakpoint of a structural variant based on a split/broken sequence read
- Can identify SVs at a single base pair resolution



Raphael, 2012. *PLoS Computational Biology,* 8:12

SV detection tools

- **Paired End Approach**
  - **Delly**, Breakdancer, Corona, HYDRA, MoDIL
  - MoGUL, PEMer, SPANNER

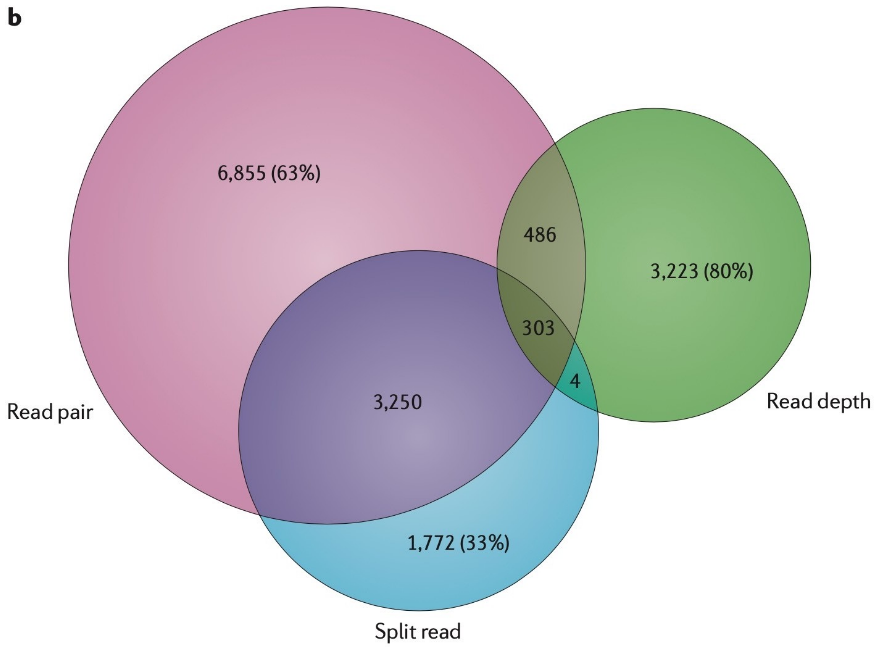
- **Read Depth Approach**
  - CNVer, CNVnator, FreeC

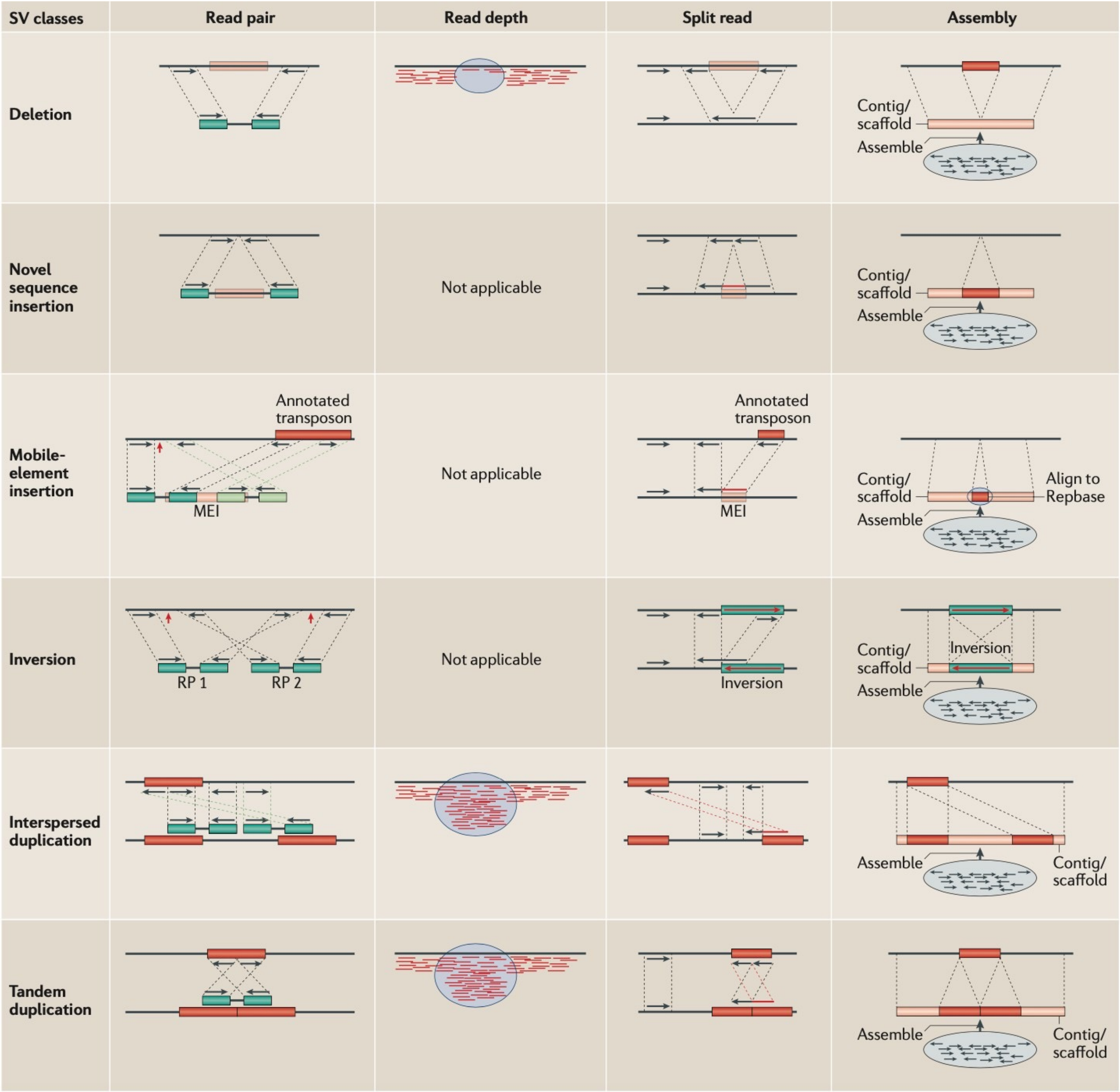
- **Split Read Approach**
  - **Delly**, Age, Pindel

dellytools/**delly**

DELLY2: Structural variant discovery by integrated paired-end and split-read analysis

Summary of SV detection approaches





SV classes	Read pair	Read depth	Split read	Assembly
Deletion				Contig/ scaffold Assemble
Novel sequence insertion		Not applicable		Contig/ scaffold Assemble
Mobile-element insertion	Annotated transposon MEI	Not applicable	Annotated transposon MEI	Contig/ scaffold Align to Repbase Assemble
Inversion	RP 1 RP 2	Not applicable	Inversion Inversion	Contig/ scaffold Inversion Assemble
Interspersed duplication				Assemble Contig/ scaffold
Tandem duplication				Assemble Contig/ scaffold

Alkan et al., 2011. *Nature Methods* 8:61–65

Conclusions

- Many types of variants

- VCF file stores data on genetic variants:
  - It is the output for several variant calling software (e.g., GATK)
  - It is the input for downstream filtering and analysis

- Different approaches (can vary for small vs. large variants)

- Each strategy has advantages and disadvantages

Practical