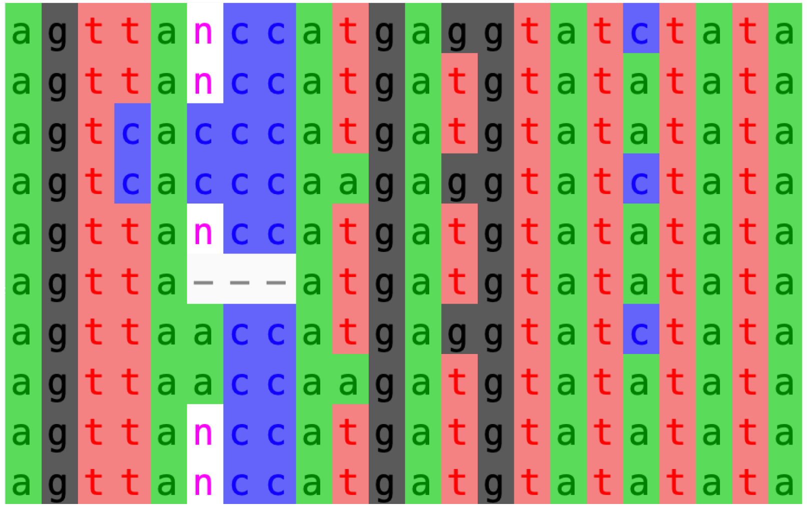
Variant Detection

Infectious Disease ‘Omics Short Course



```
a g t t a n c c a t g a g g t a t c t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t c a c c c a t g a t g t a t a t a t a
a g t c a c c c a a g a g g t a t c t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t t a - - - a t g a t g t a t a t a t a
a g t t a a c c a t g a g g t a t c t a t a
a g t t a a c c a a g a t g t a t a t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t t a n c c a t g a t g t a t a t a t a
```

4th-7th December 2023

Outline

Introduction

- What are the different types of small variants?
- What are the different types of structural (large) variants?

Variant calling from next generation sequencing data

- What are the major steps of a variant discovery pipeline?
- What are VCFs?

Discovery of small variants (SNPs and indels)

- What tools are used to detect small variants?
- Variant discovery with GATK and *bcftools*

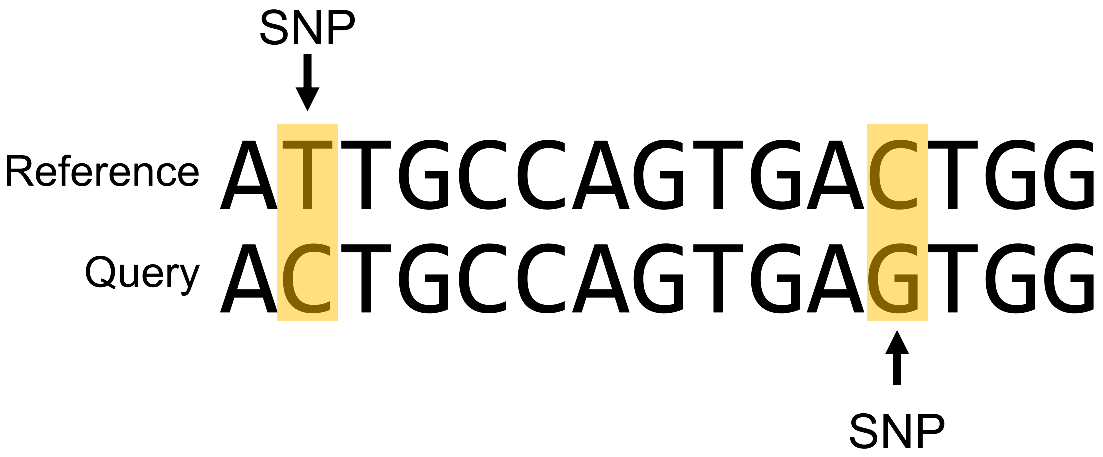
Discovery of large variants (structural variants)

- What are the 4 main approaches used to detect large structural variants?
- What tools can we use to detect large structural variants?

Conclusions

Practical

Different types of variants: SNPs



SNP

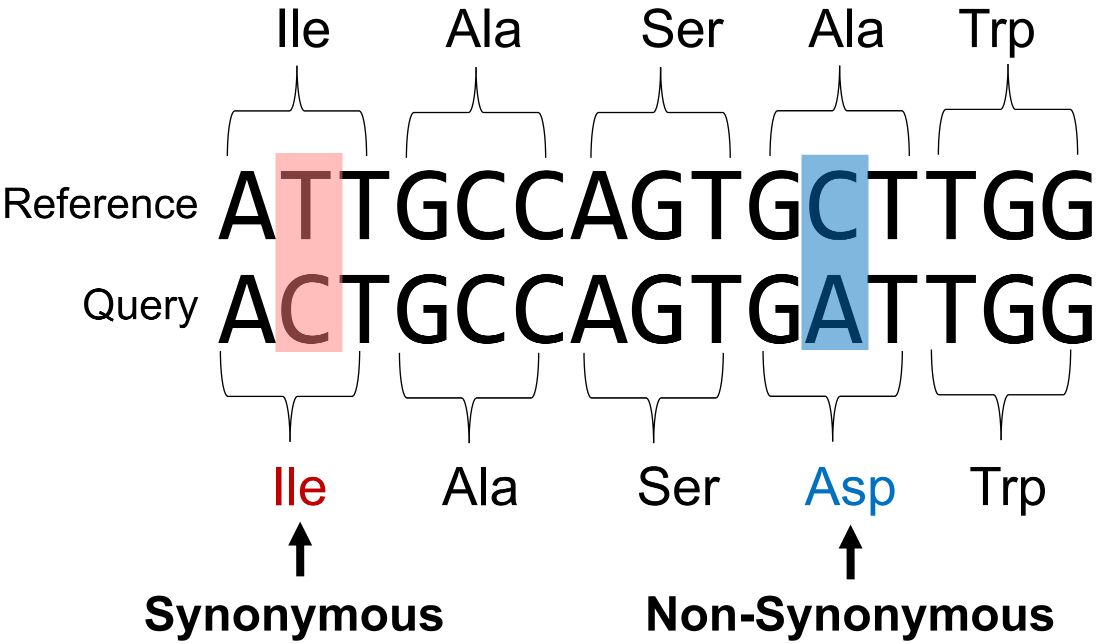
Reference ATTGCCAGTGACTGG

Query ACTGCCAGTGAGTGG

SNP

Single **N**ucleotide **P**olymorphisms (SNPs) are single base pair variations at specific locations in the genome, with respect to a reference genome.

Different types of variants: SNPs



Ile Ala Ser Ala Trp

Reference ATTGCCAGTGCTTGG

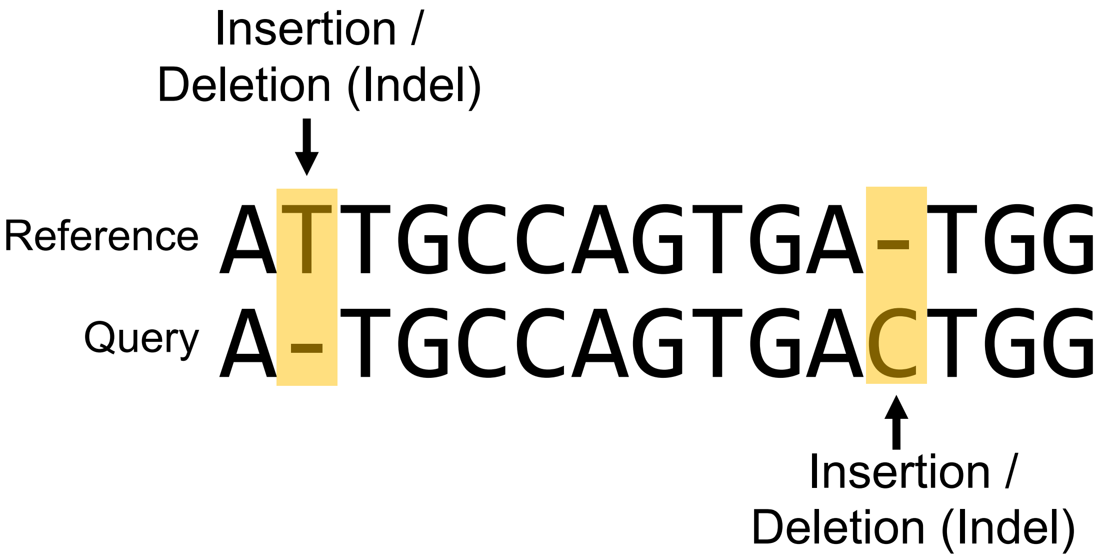
Query ACTGCCAGTGATTGG

Ile Ala Ser Asp Trp

Synonymous

Non-Synonymous

Different types of variants: Indels



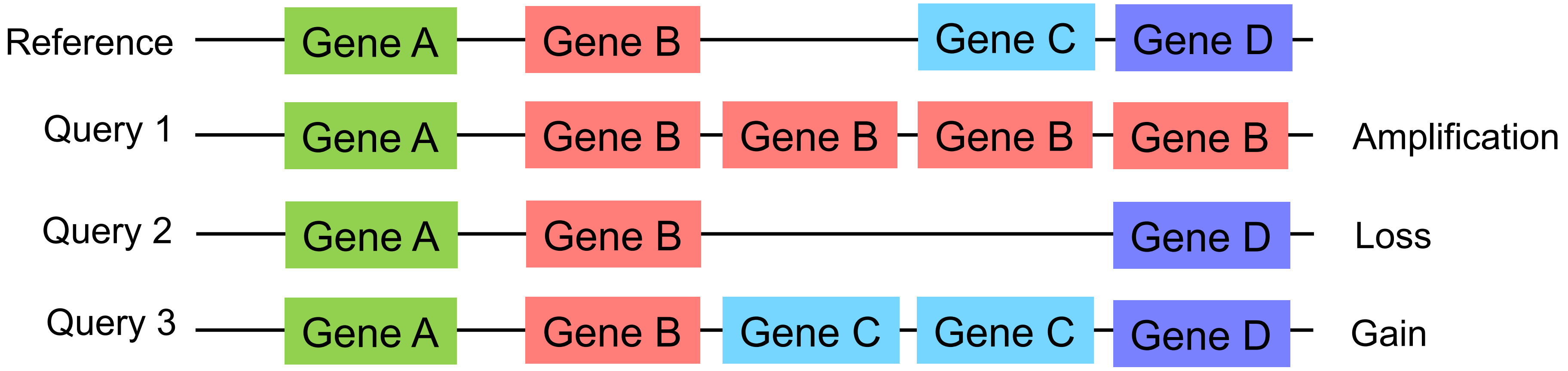
Insertion / Deletion (Indel)

```
Reference  ATTGCCAGTGA-TGG
Query      A-TGCCAGTGACTGG
```

Insertion / Deletion (Indel)

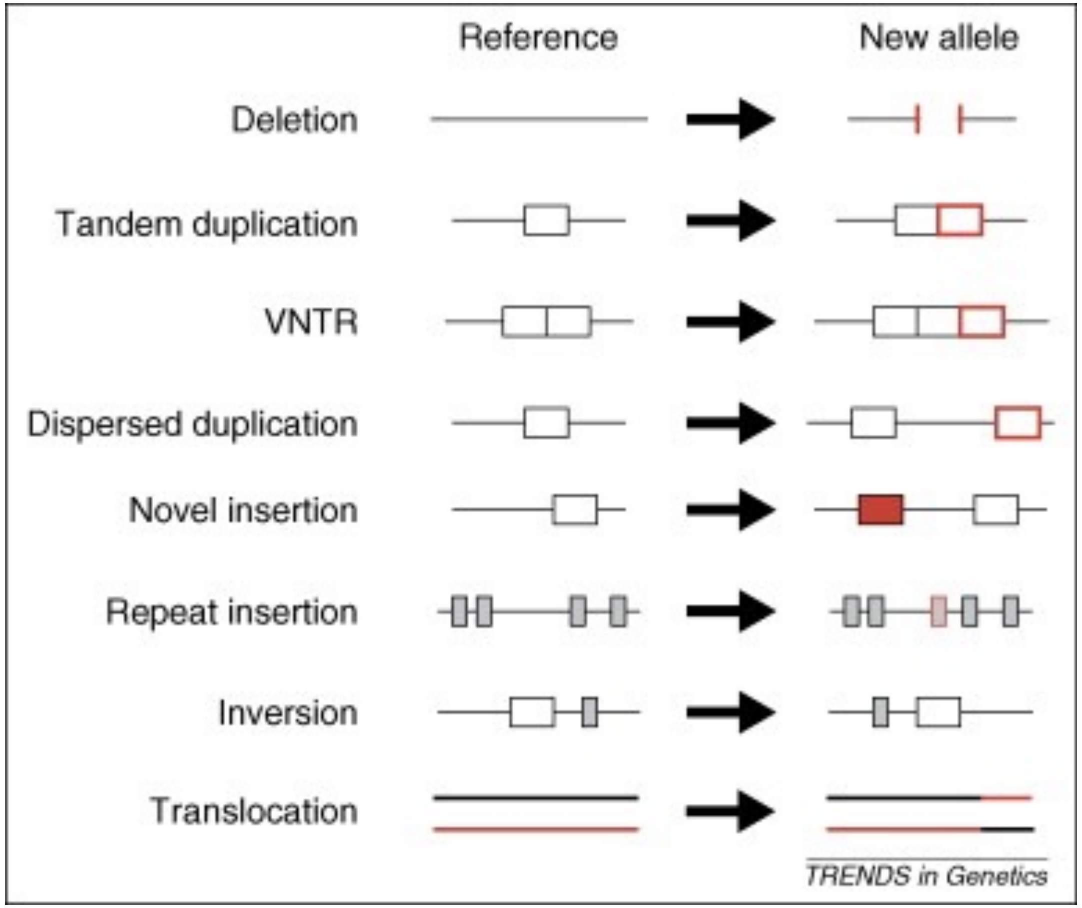
Insertion-**Del**etions (Indels) are insertions and/or deletions of nucleotides at specific locations in the genome, with respect to a reference, and are often events of less than 1kb.

Different types of variants: CNVs



Copy **N**umber **V**ariants **(CNVs)** are a type of structural variation where the number of copies of a specific segment of DNA varies among different individuals' genomes of the same species.

Other types of variants



Structural Variants

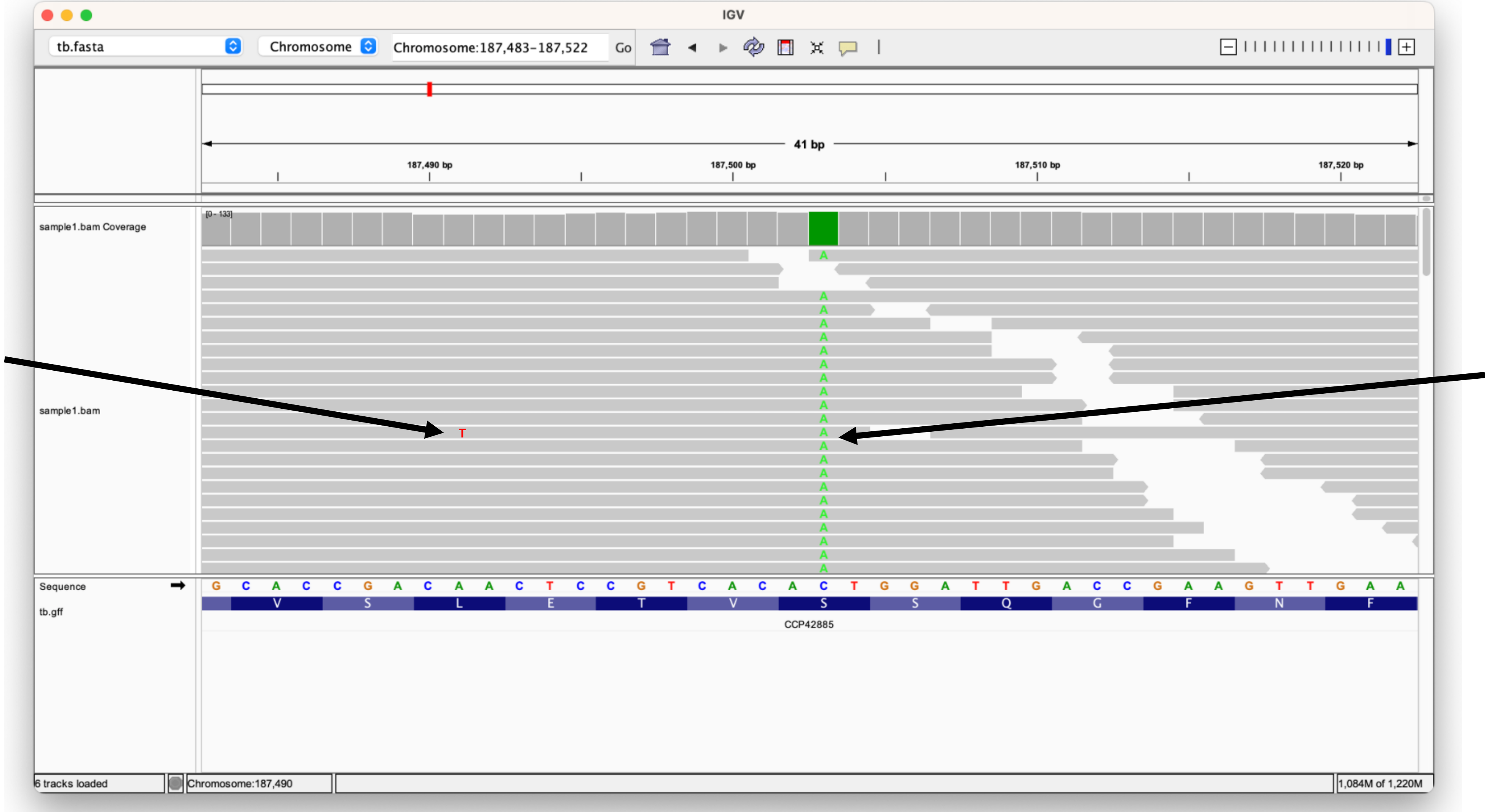
- **Small scale variants**
  - Single nucleotide polymorphisms (SNPs)
  - Insertions and deletions (indels)
  - Variable tandem repeats (VNTRs)
- **Large scale variants (>1kb)**
  - Copy number variations (CNVs)
  - Duplications, inversions
  - Translocations

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

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

Detecting SNPs from alignments

Sequencing or data processing artifact

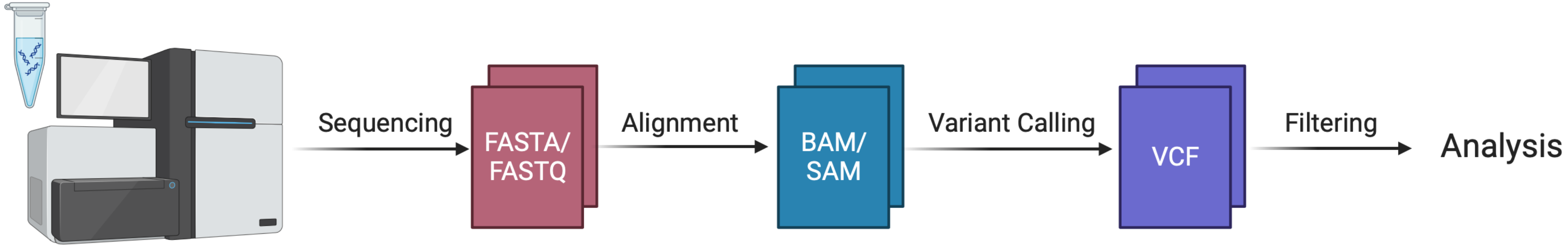


True SNP

Viewing *Mtb* data in IGV

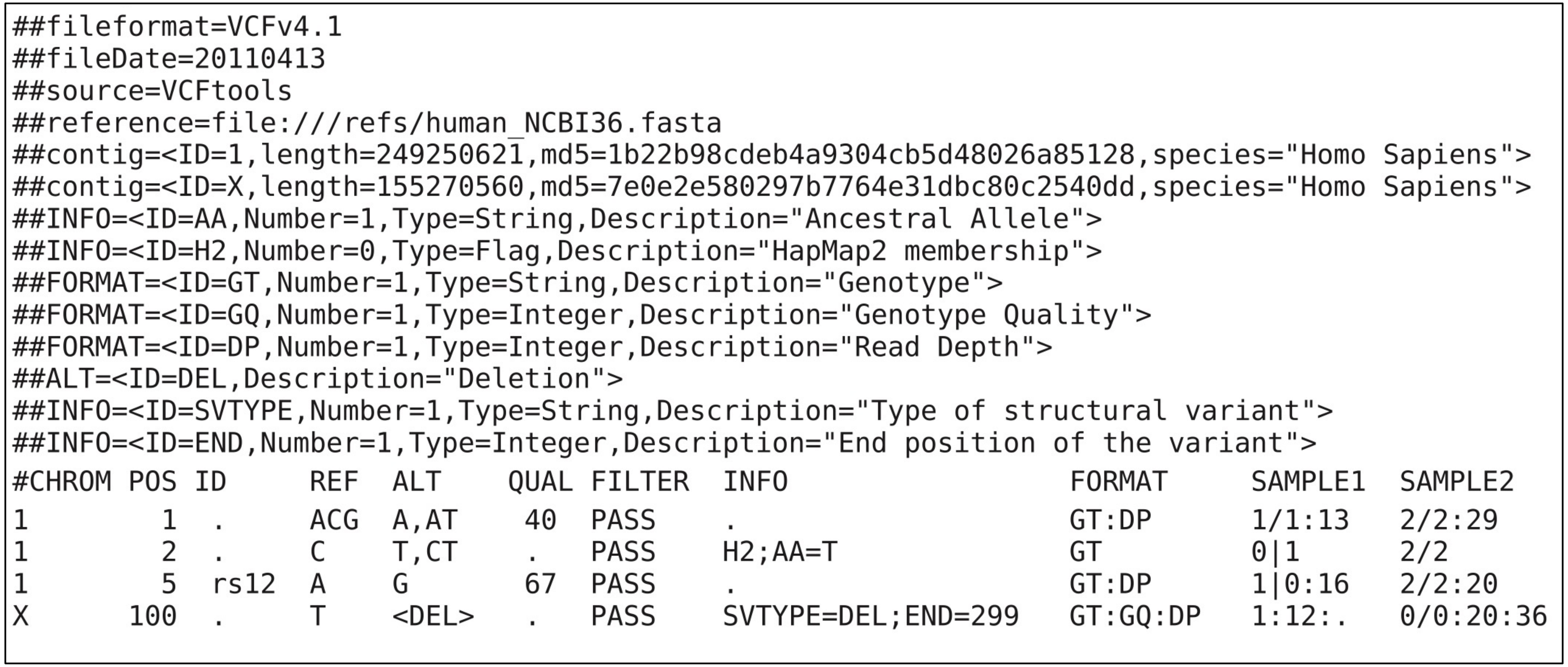
Variant Discovery Pipeline for NGS data

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



Variant Call Format (VCF)

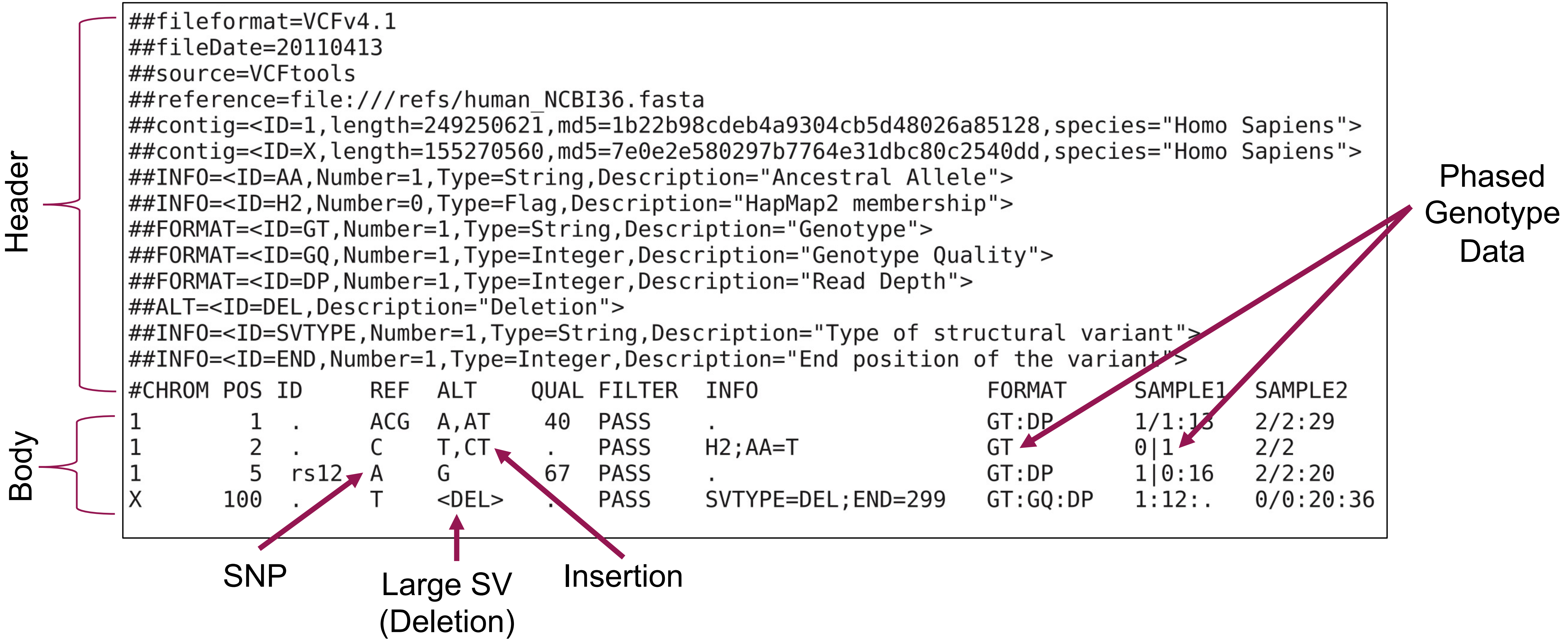
- A VCF is a text file format employed to store genetic variation with respect to a reference genome.



```
##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*

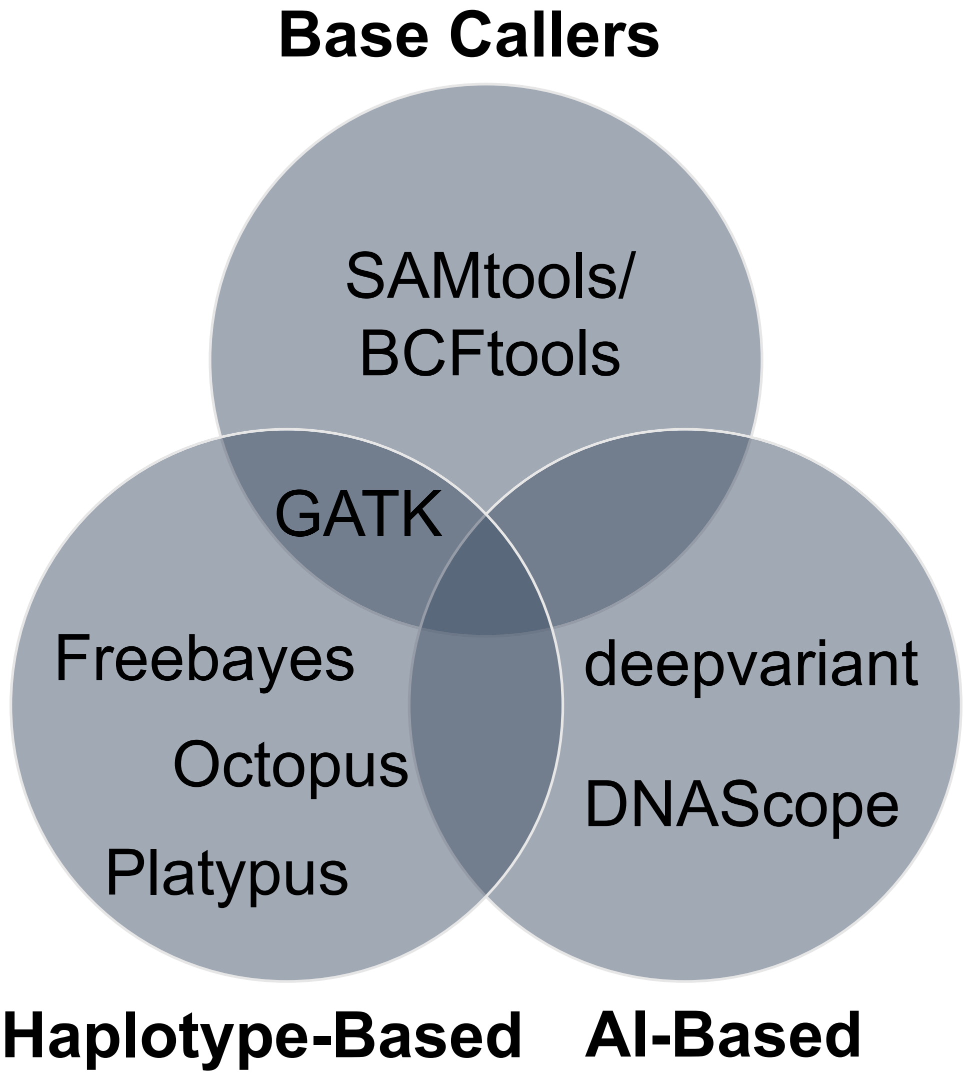
Variant Call Format (VCF)



```
##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*

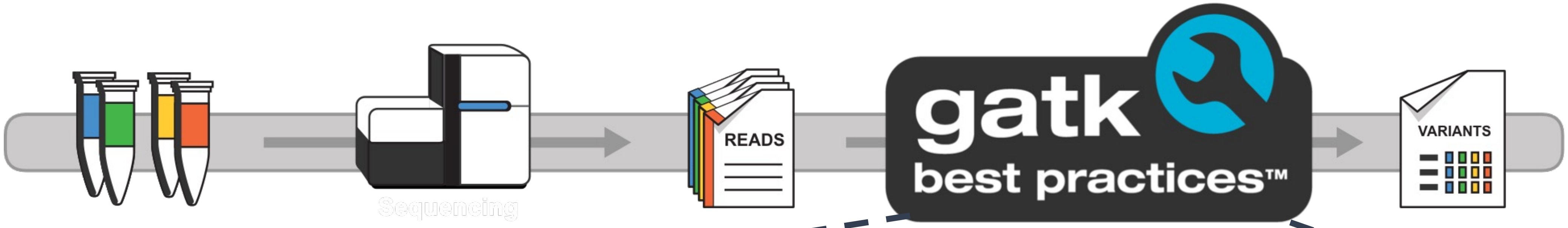
SNP and short indel calling tools



All (described here) are used by the open-source community.

All are performing variant calling but **employing different models to do so.**

Variant discovery with GATK



1. Data Pre-Processing

Correction of BAM files for systematic errors in base quality scores

`BaseRecalibrator`
`ApplyRecalibration`

2. Variant Discovery

SNP and indel calling

`HaplotypeCaller`
`CombineGVCFs`
`GenomicsDBImport`
`GenotypeGVCFs`

3. Evaluation & Refinement

Improving genotype accuracy by identifying false positives

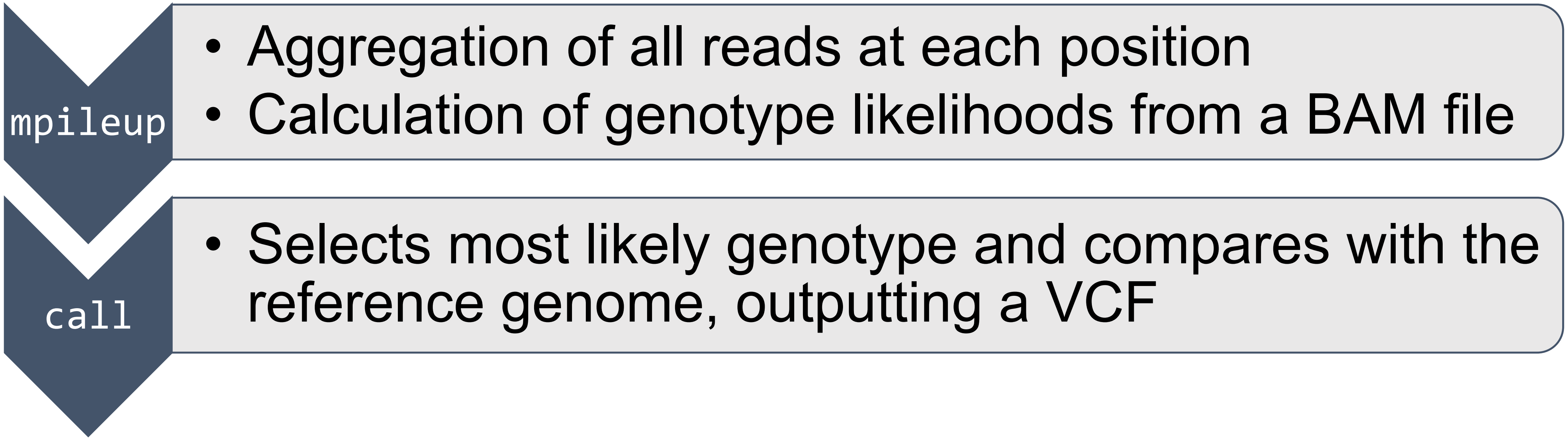
`VariantRecalibrator`
`ApplyVQSR`

GATK is a package of command-line tools written in Java and provides end-to-end workflows called best practices. It is easily parallelised and scalable, but run times are long!

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

Variant discovery with bcftools

A two-stage process using two algorithms:



`mpileup`

- Aggregation of all reads at each position
- Calculation of genotype likelihoods from a BAM file

`call`

- Selects most likely genotype and compares with the reference genome, outputting a VCF

Much **faster** than GATK!

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

Next generation sequencing and SVs

2000s

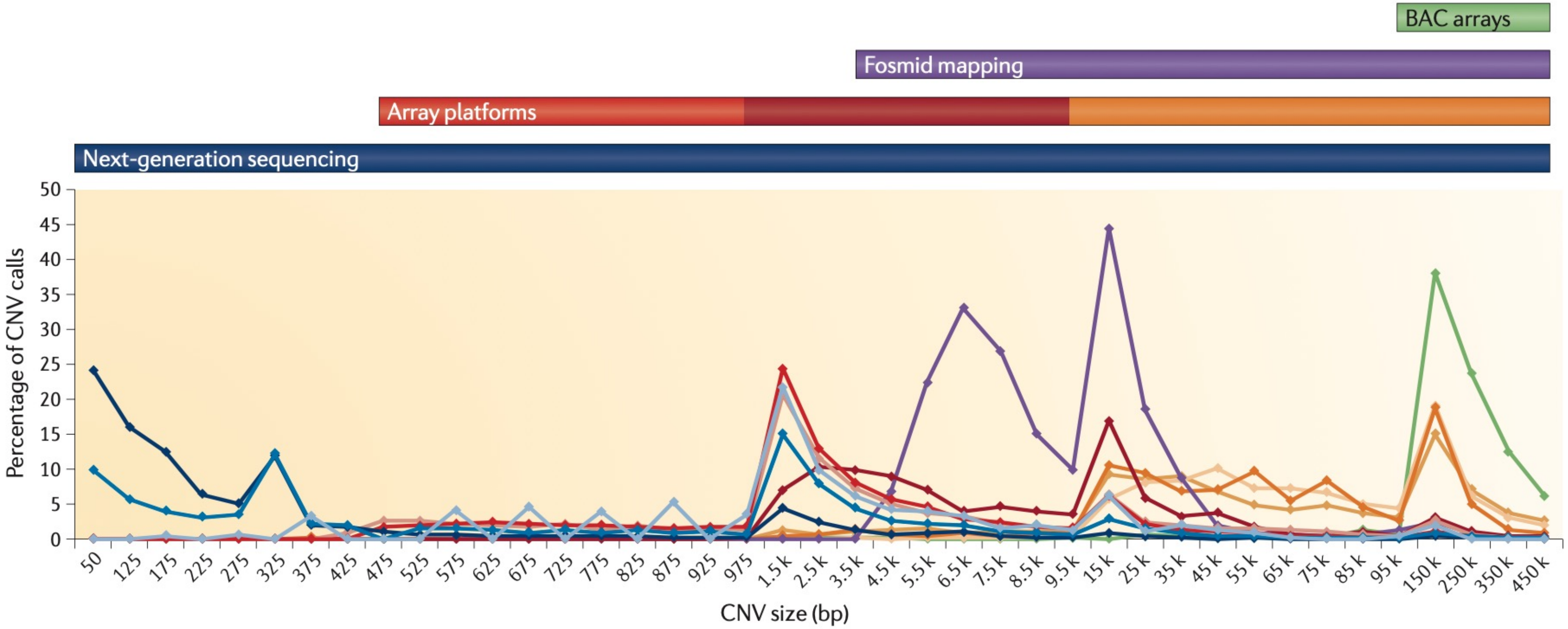
Microarrays

2010s

Short Reads

2020s

Long Reads

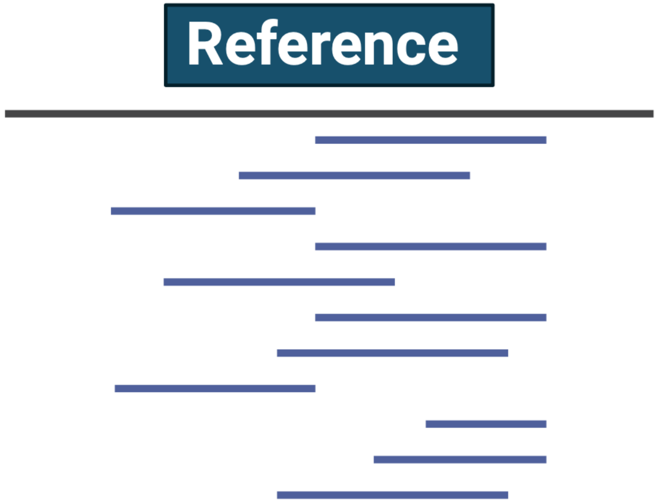


Next generation sequencing offers the widest range of detection of structural variants.

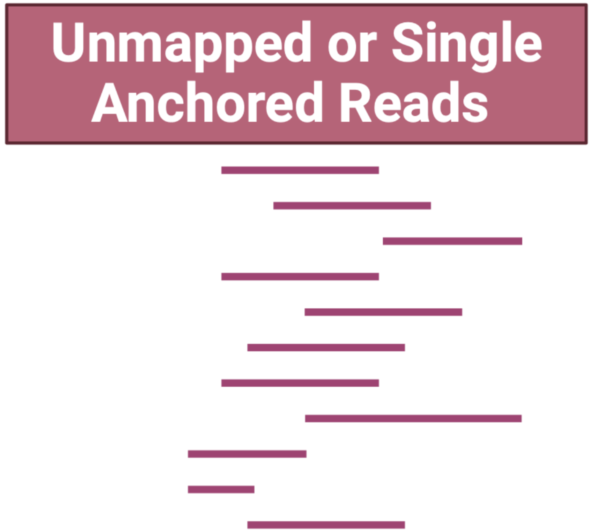
Alkan et al., 2011. *Nature Methods* 8:61–65

Discovery of structural variants

When short read data are mapped to a reference, structural variants can be identified by their unique signatures.



Approaches:
Paired End, Split Read, Read Depth



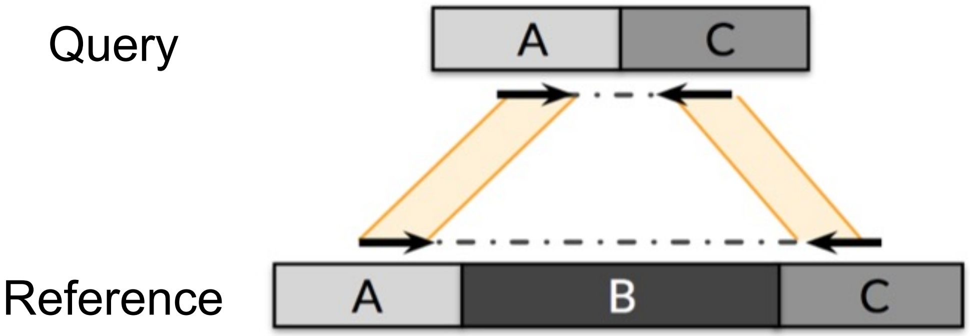
Approaches:
Assembly (Discussed on Day 2)

Quinlan and Hall, 2012. *Trends in Genetics*, *28*:43-53

Paired End (PE) Approach

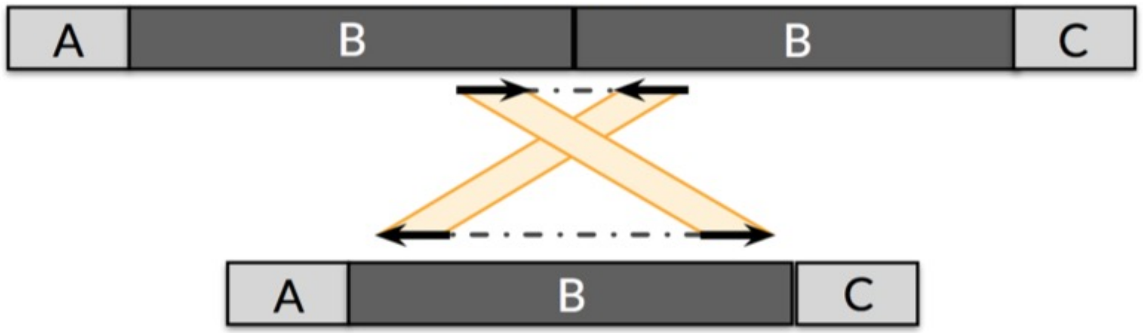
- Assesses the span and orientation of PE reads.
- If an SV is present, it will produce ‘discordant’ alignments.

Deletion



Reads mapping further apart than expected with respect to reference

Tandem Duplication

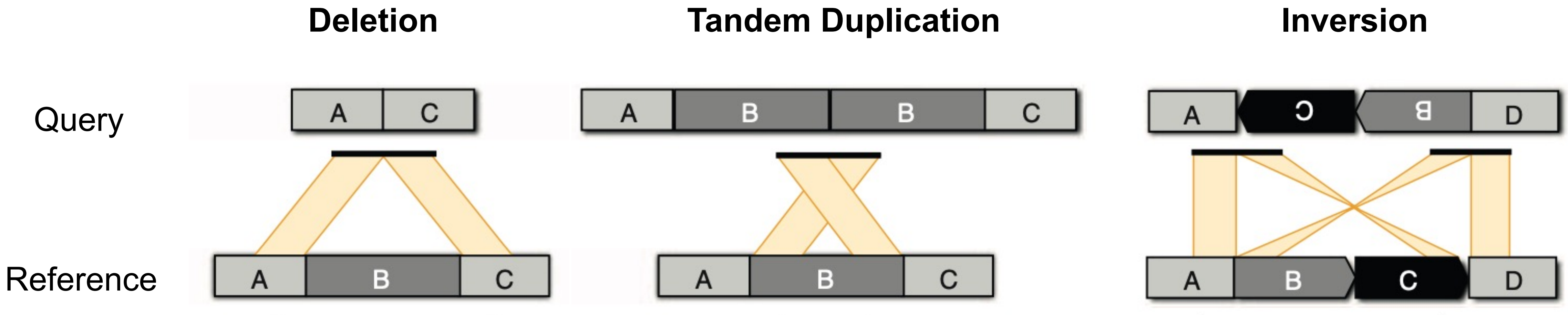


Reads mapping in the opposite orientation than expected with respect to the reference

Quinlan and Hall, 2012. *Trends in Genetics*, *28*:43-53

Split Read (SR) Approach

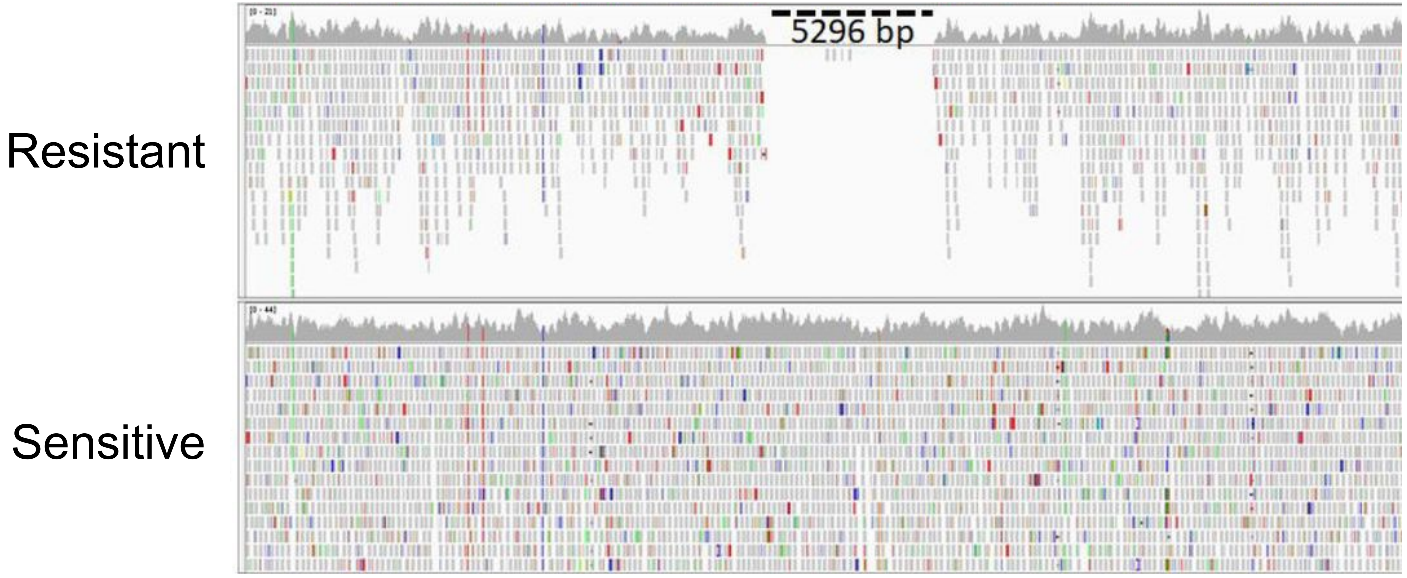
- Identifies sequences containing a breakpoint, mapping them to single base-pair resolution.



Quinlan and Hall, 2012. *Trends in Genetics*, *28*:43-53

Read Depth (RD) Approach

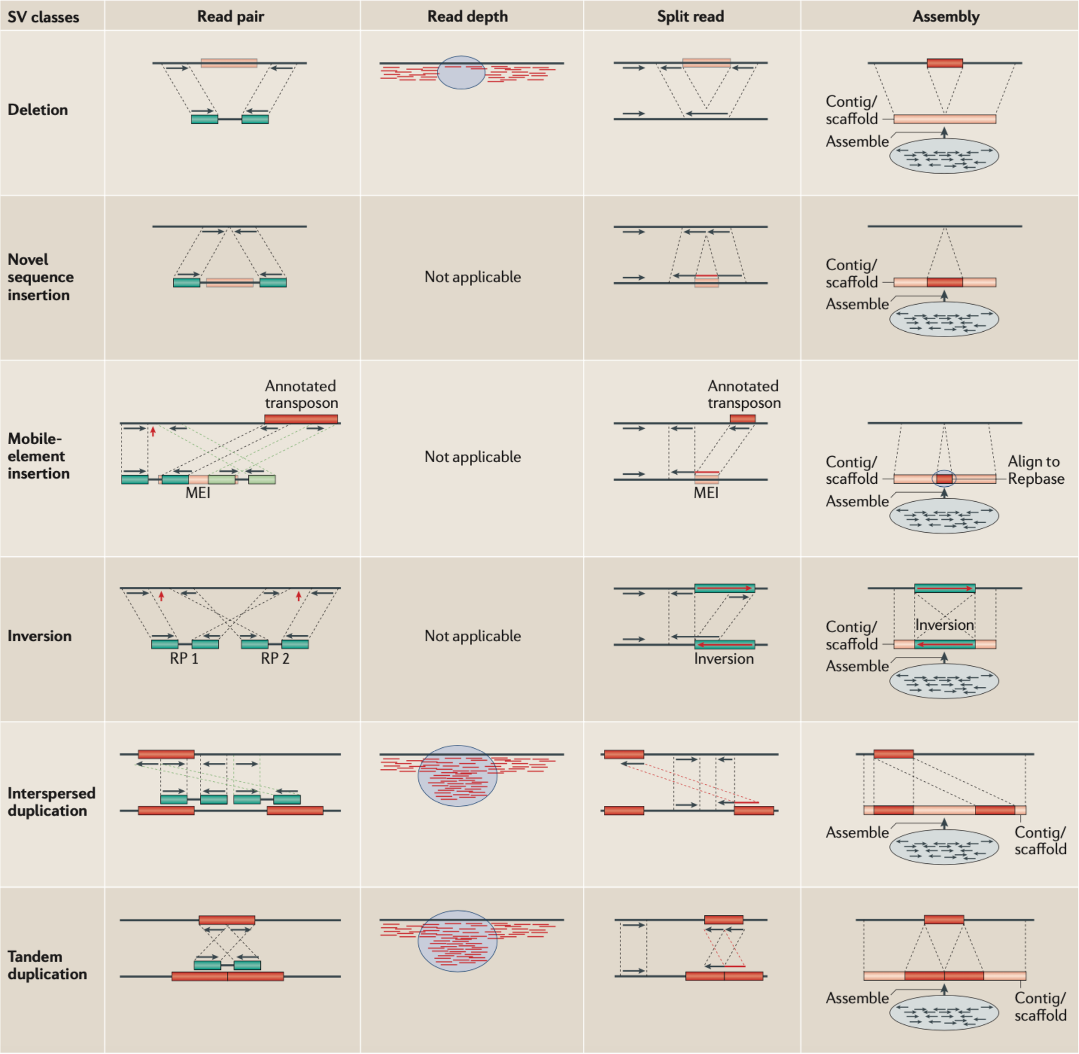
- Detects deletions or duplications based on divergences in mapping depth:
  - Low or zero coverage suggests a deletion
  - Excess coverage suggests a duplication



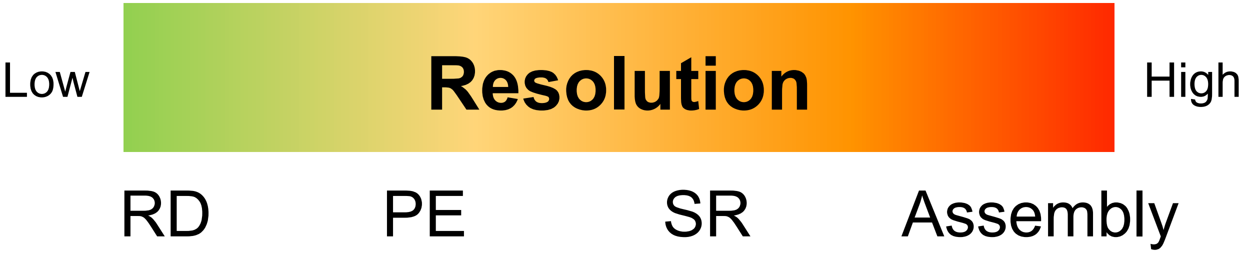
Deletion

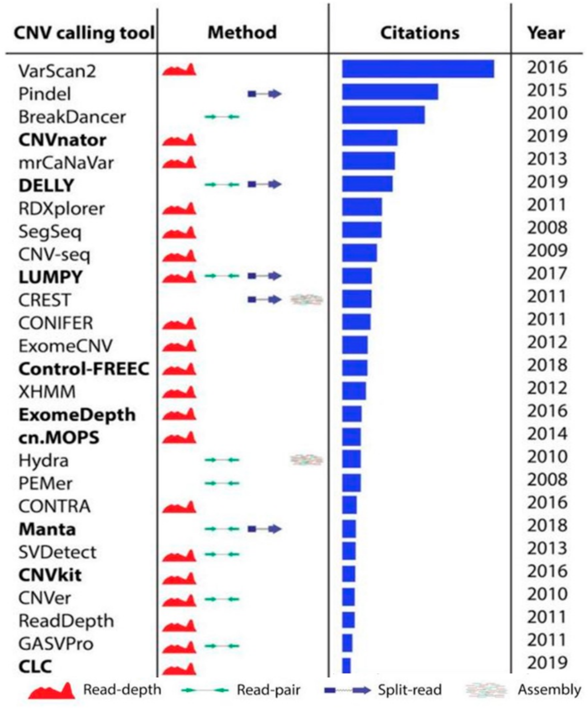
Jain *et al.*, 2021. *Plos one*, *16*:7

Summary of SV detection and tools



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 Assemble Align to Repbase
Inversion	RP 1 RP 2	Not applicable	Inversion	Contig/ scaffold Inversion Assemble
Interspersed duplication				Assemble Contig/ scaffold
Tandem duplication				Assemble Contig/ scaffold





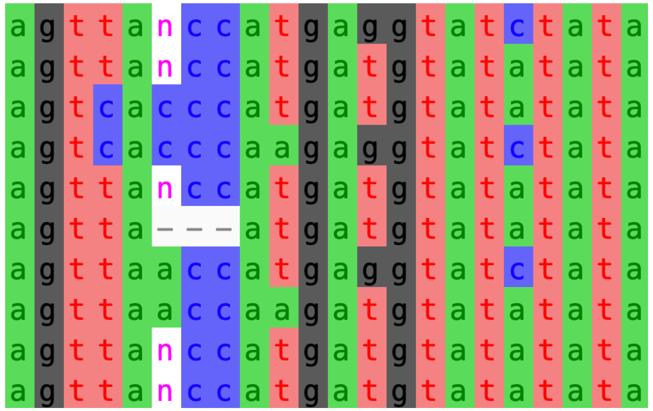
CNV calling tool	Method	Citations	Year
VarScan2	Read-depth		2016
Pindel	Split-read		2015
BreakDancer	Read-pair		2010
CNVnator	Read-depth		2019
mrCaNaVar	Read-depth		2013
DELLY	Read-pair, Split-read		2019
RDXplorer	Read-depth		2011
SegSeq	Read-depth		2008
CNV-seq	Read-depth		2009
LUMPY	Read-depth, Read-pair, Split-read		2017
CREST	Split-read, Assembly		2011
CONIFER	Read-depth		2011
ExomeCNV	Read-depth		2012
Control-FREEC	Read-depth		2018
XHMM	Read-depth		2012
ExomeDepth	Read-depth		2016
cn.MOPS	Read-depth		2014
Hydra	Read-pair, Assembly		2010
PEMer	Read-pair		2008
CONTRA	Read-depth		2016
Manta	Read-pair, Split-read		2018
SVDetect	Read-depth, Read-pair		2013
CNVkit	Read-depth		2016
CNVer	Read-depth, Read-pair		2010
ReadDepth	Read-depth		2011
GASVPro	Read-depth, Read-pair		2011
CLC	Read-depth		2019

...Again, all are essentially performing structural variant calling but employing different methods to do so.

Alkan et al., 2011. *Nature Methods* 8:61–65; Gabrielaite et al., 2021. *Cancers* 2021, 13:24

Conclusions

- Many different types and combinations of variants
- A VCF stores data on variants:
  - It is the output for several variant calling software (e.g., GATK)
  - It is the input for downstream filtering and analysis (e.g., population genetics)
- Detection of small vs. large variants requires different approaches.
  - Whichever strategy employed comes with its own advantages and disadvantages.
- **It is a combination of the choice of software tools for both alignment and variant calling that will influence the final result (i.e., variants called).**
  - Use whatever works best for your research question/project!



```
a g t t a n c c a t g a g g t a t c t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t c a c c c a t g a t g t a t a t a t a
a g t c a c c c a a g a g g t a t c t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t t a - - - a t g a t g t a t a t a t a
a g t t a a c c a t g a g g t a t c t a t a
a g t t a a c c a a g a t g t a t a t a t a
a g t t a n c c a t g a t g t a t a t a t a
a g t t a n c c a t g a t g t a t a t a t a
```

Practical