# *Next generation sequencing*

Calling SNPs and structural variants

Stig Uggerhøj Andersen, PhD

Department of Molecular Biology

University of Aarhus

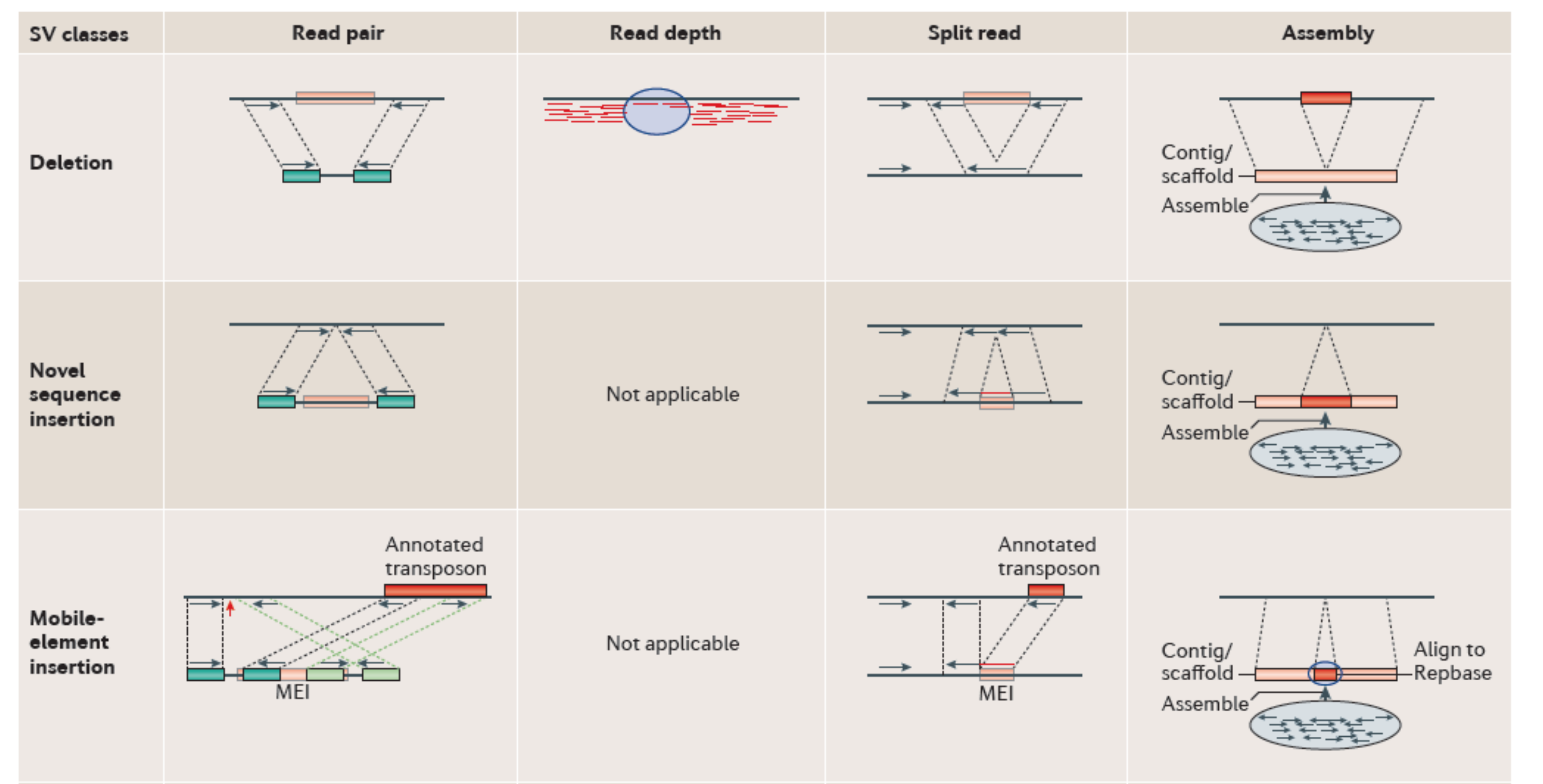
# Lecture overview

- Structural variants
- SNPs

# Structural variation

- Why are structural variants interesting?
  - Copy number variation (gene amplification)
  - Gene fusions by translocation
  - Natural variation underlying traits of interest
  - Major confounding effect in SNP calling

# NGS discovery of SVs with short reads

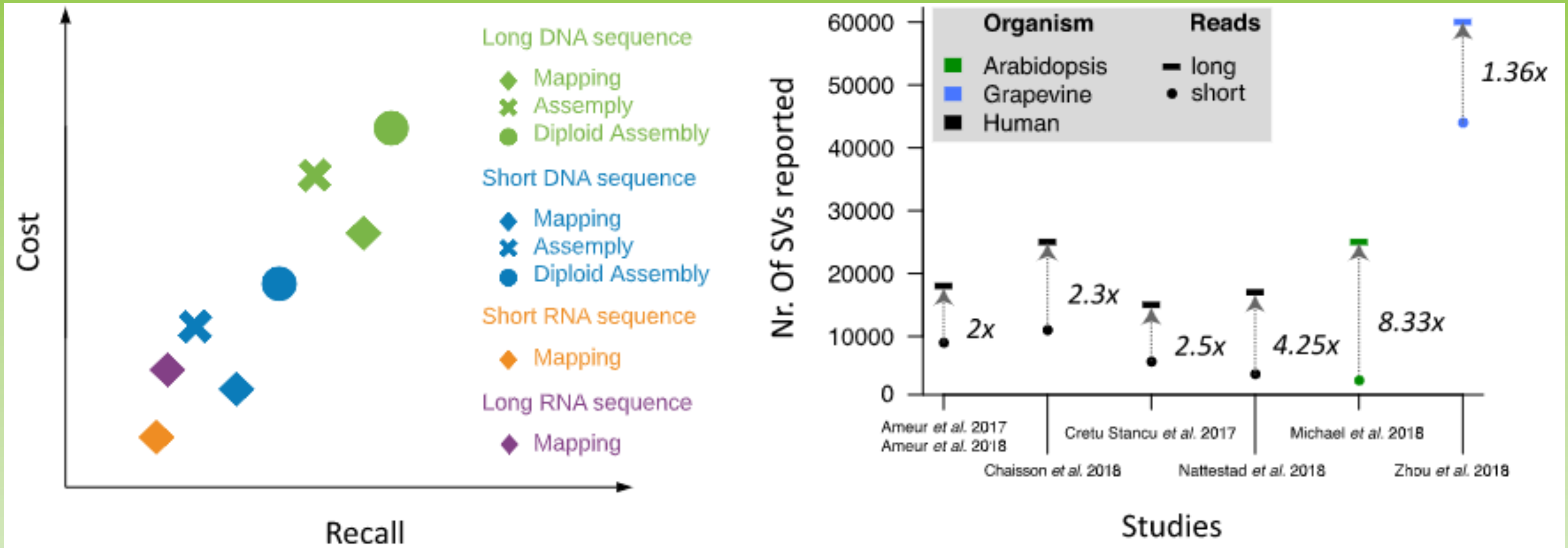


| SV classes | Read pair | Read depth | Split read | Assembly |
|---|---|---|---|---|
| Deletion | | | | Contig/<br>scaffold<br>Assemble |
| Novel sequence insertion | | Not applicable | | Contig/<br>scaffold<br>Assemble |
| Mobile-element insertion | Annotated transposon<br>MEI | Not applicable | Annotated transposon<br>MEI | Contig/<br>scaffold<br>Assemble<br>Align to Repbase |

SVs are very difficult to reliably detect using short reads

PMID: 21358748

# Structural variation software



| | | | |
|---|---|---|---|
| SVDetect: | PMID: 20639544 | PBHoney: | PMID: 24915764 |
| ParMap: | PMID: 20507604 | NanoSV: | PMID: 29109544 |
| Slope: | PMID: 20876606 | Picky: | PMID: 29713081 |
| SOAPindel | PMID: 22972939 | NGMLR+Sniffles: | PMID: 29713083 |
| BayesTyper | PMID: 29915429 | | |

Impact of PacBio Hifi reads!

Recent review: 10.1186/s13059-019-1828-7

Comparison of long read-based SV callers PMID: 32211024

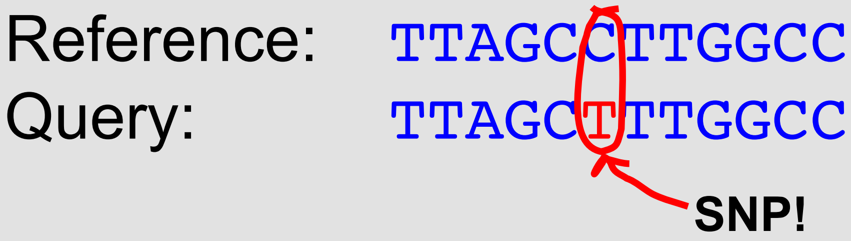
# Structural variation software

- Conclusions
  - Including *de novo* assembly is beneficial, especially with ever-improving assembly quality
  - Use long, rather than short, reads
  - Alignment of fully assembled genomes will be increasingly used to determine haplotype patterns
  - K-mer based approaches can be used for validation, e.g. BayesTyper

# Identifying Single Nucleotide Polymorphisms

- Why are SNPs interesting?
  - Common type of genetic variation
  - Disease associated SNPs
  - SNPs are easily scorable genetic markers

# Identifying Single Nucleotide Polymorphisms



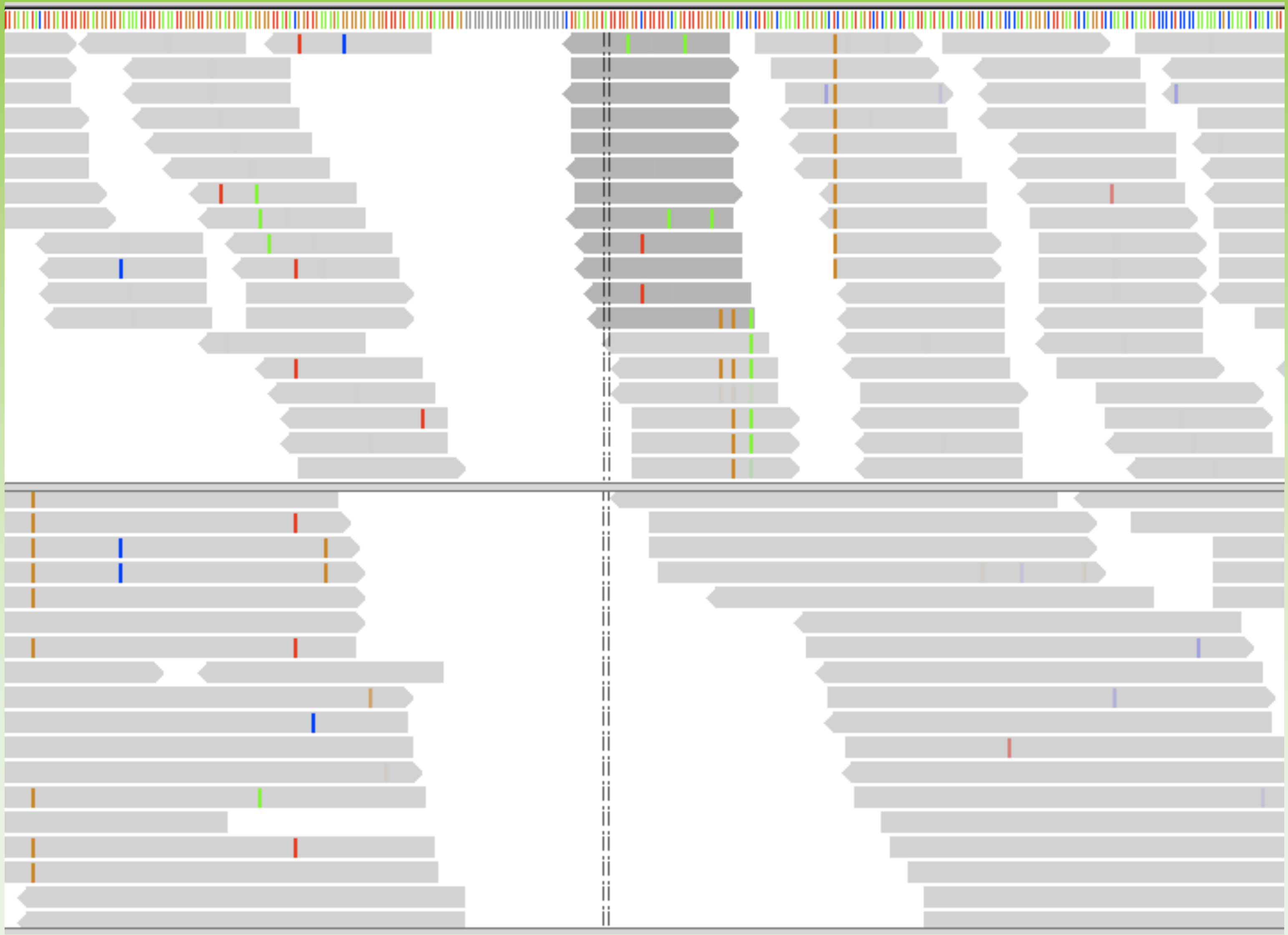
Reference: TTAGCCTTGGCC

Query: TTAGCTTTGGCC

SNP!

- SNP calling is conceptually simple
- … but in practice quite complicated

# Identifying Single Nucleotide Polymorphisms



SNP calling

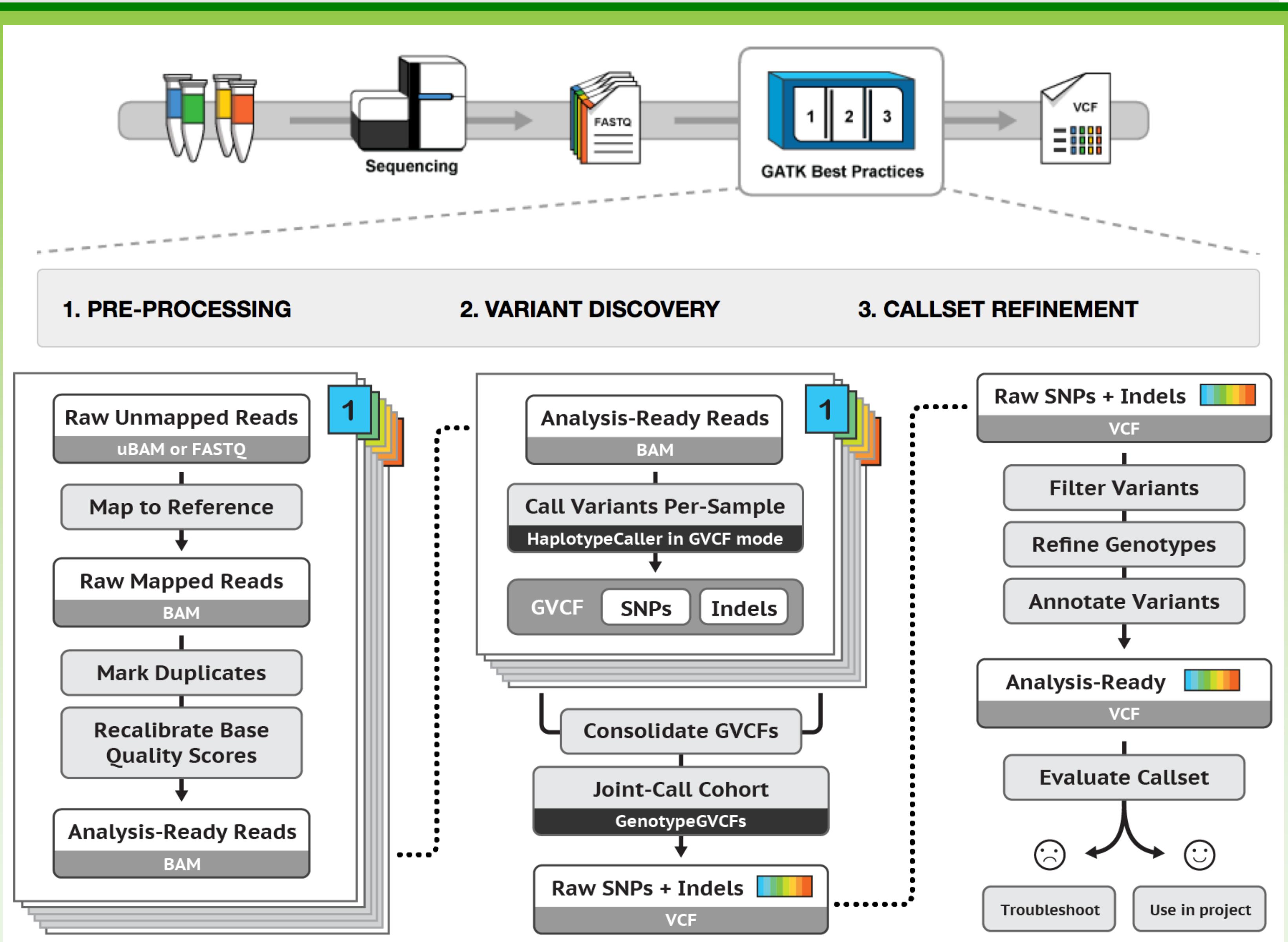
# Identifying Single Nucleotide Polymorphisms

- Confounding effects
  - INDELs
  - Mis-mapped reads

# Identifying Single Nucleotide Polymorphisms

- Different complexity levels in SNP calling
  - Identify only homozygous SNPs between inbred lines
  - Call genotypes in heterozygous diploid individuals
  - Call genotypes in heterozygous polyploid individuals
  - Identify rare variants in pooled samples

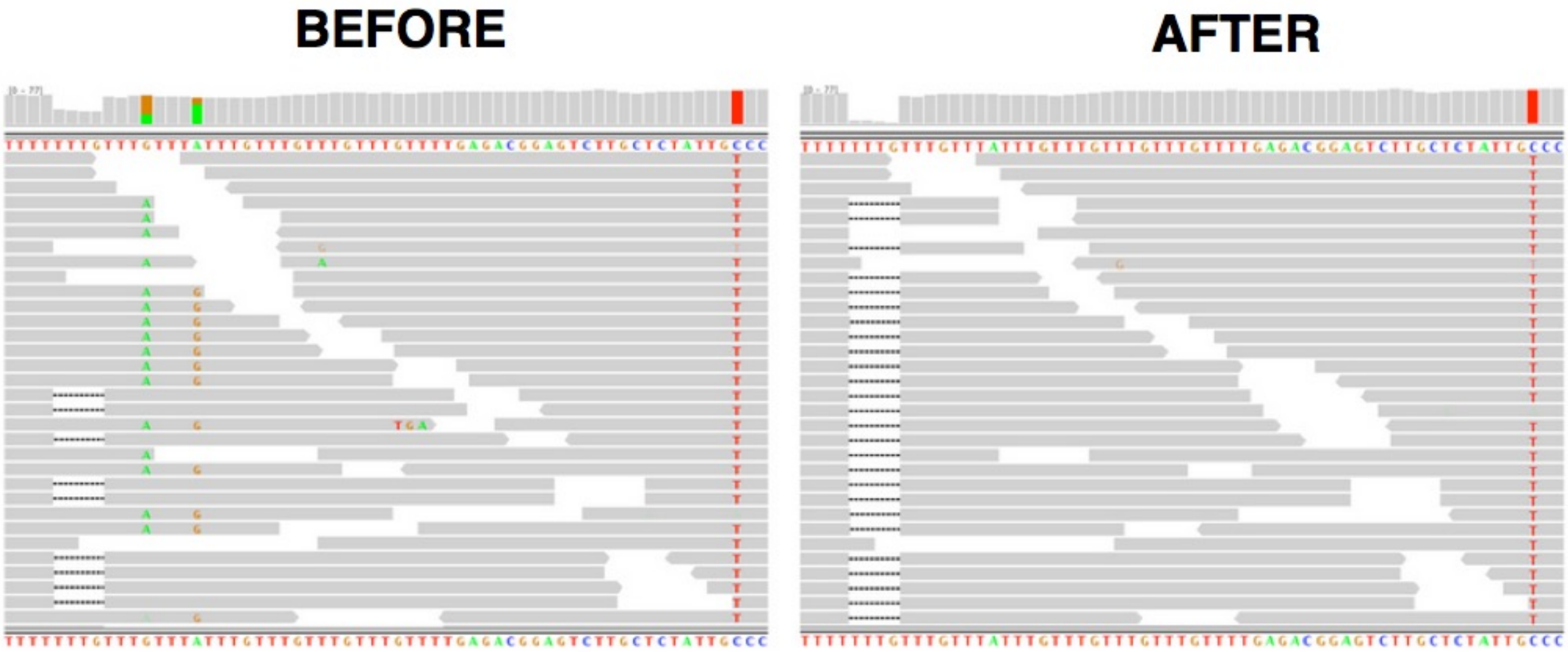
# Identifying Single Nucleotide Polymorphisms



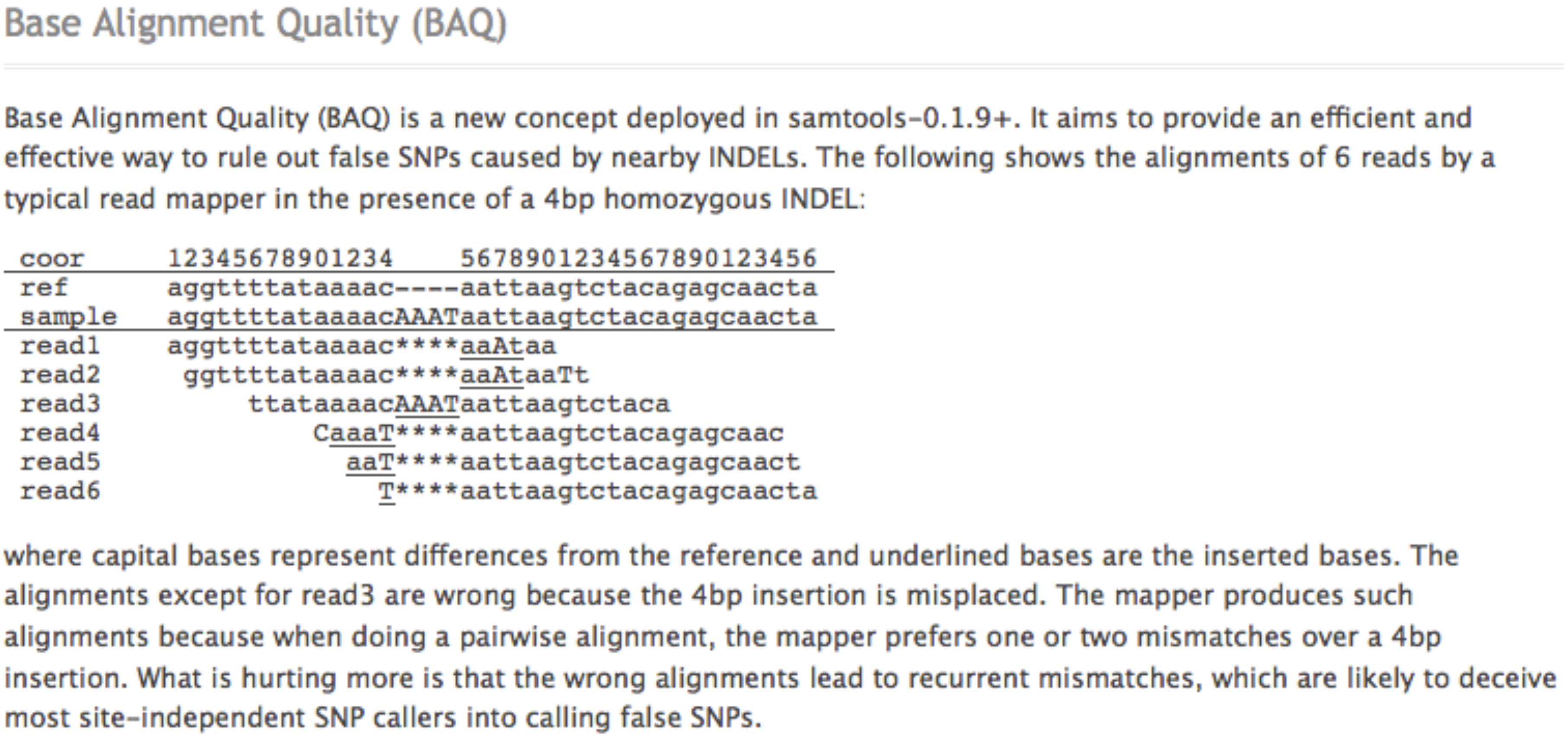
https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145

# Identifying Single Nucleotide Polymorphisms

- GATK indel realign



# Identifying Single Nucleotide Polymorphisms



## Base Alignment Quality (BAQ)

Base Alignment Quality (BAQ) is a new concept deployed in samtools-0.1.9+. It aims to provide an efficient and effective way to rule out false SNPs caused by nearby INDELs. The following shows the alignments of 6 reads by a typical read mapper in the presence of a 4bp homozygous INDEL:

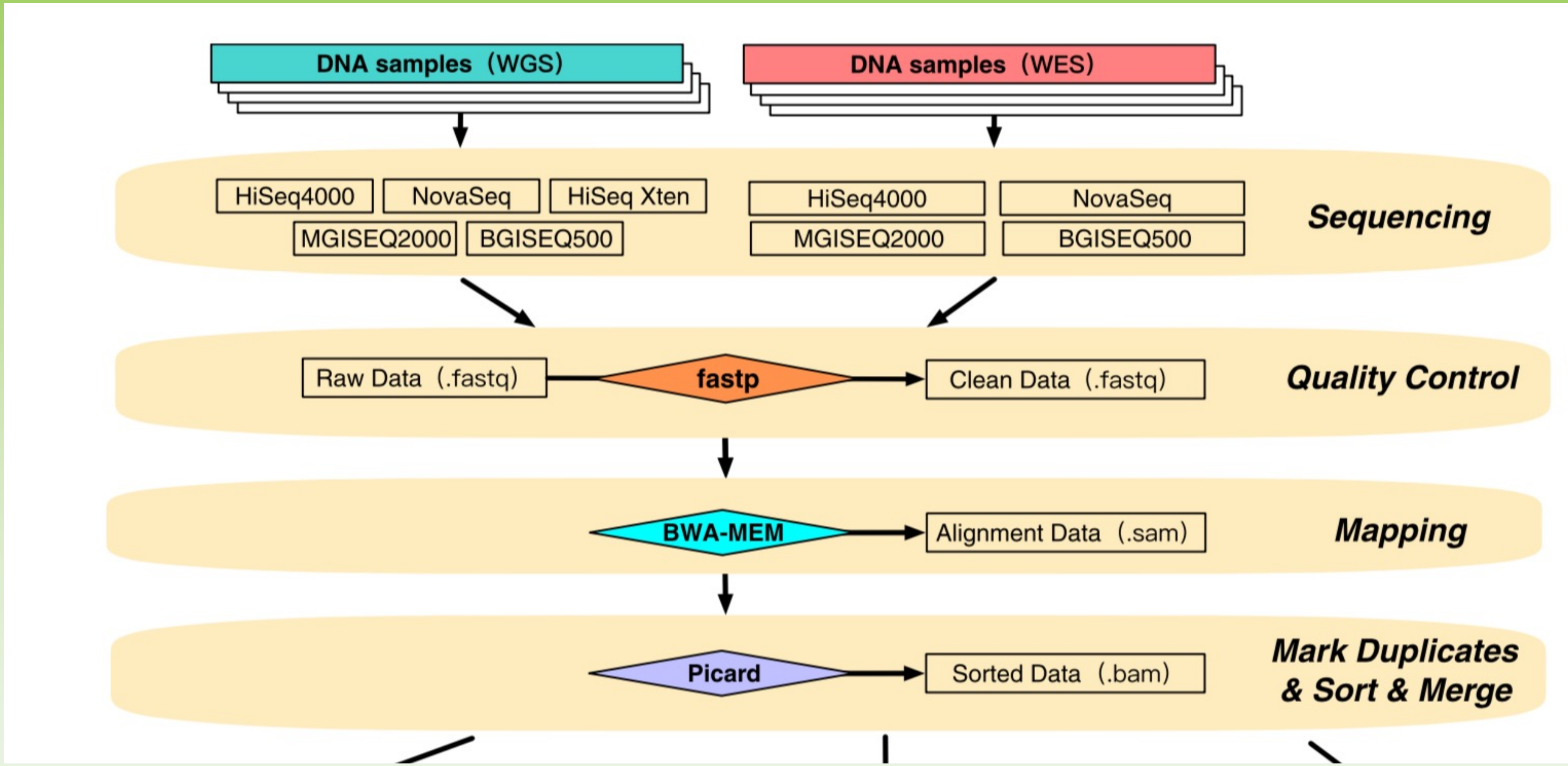
```
coor     12345678901234    5678901234567890123456
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac****aaAtaa
read2     ggttttataaaac****aaAtaaTt
read3         ttataaaacAAATaattaagtctaca
read4             CaaaT****aattaagtctacagagcaac
read5               aaT****aattaagtctacagagcaact
read6                 T****aattaagtctacagagcaacta
```

where capital bases represent differences from the reference and underlined bases are the inserted bases. The alignments except for read3 are wrong because the 4bp insertion is misplaced. The mapper produces such alignments because when doing a pairwise alignment, the mapper prefers one or two mismatches over a 4bp insertion. What is hurting more is that the wrong alignments lead to recurrent mismatches, which are likely to deceive most site-independent SNP callers into calling false SNPs.

http://www.htslib.org/

http://samtools.sourceforge.net/mpileup.shtml

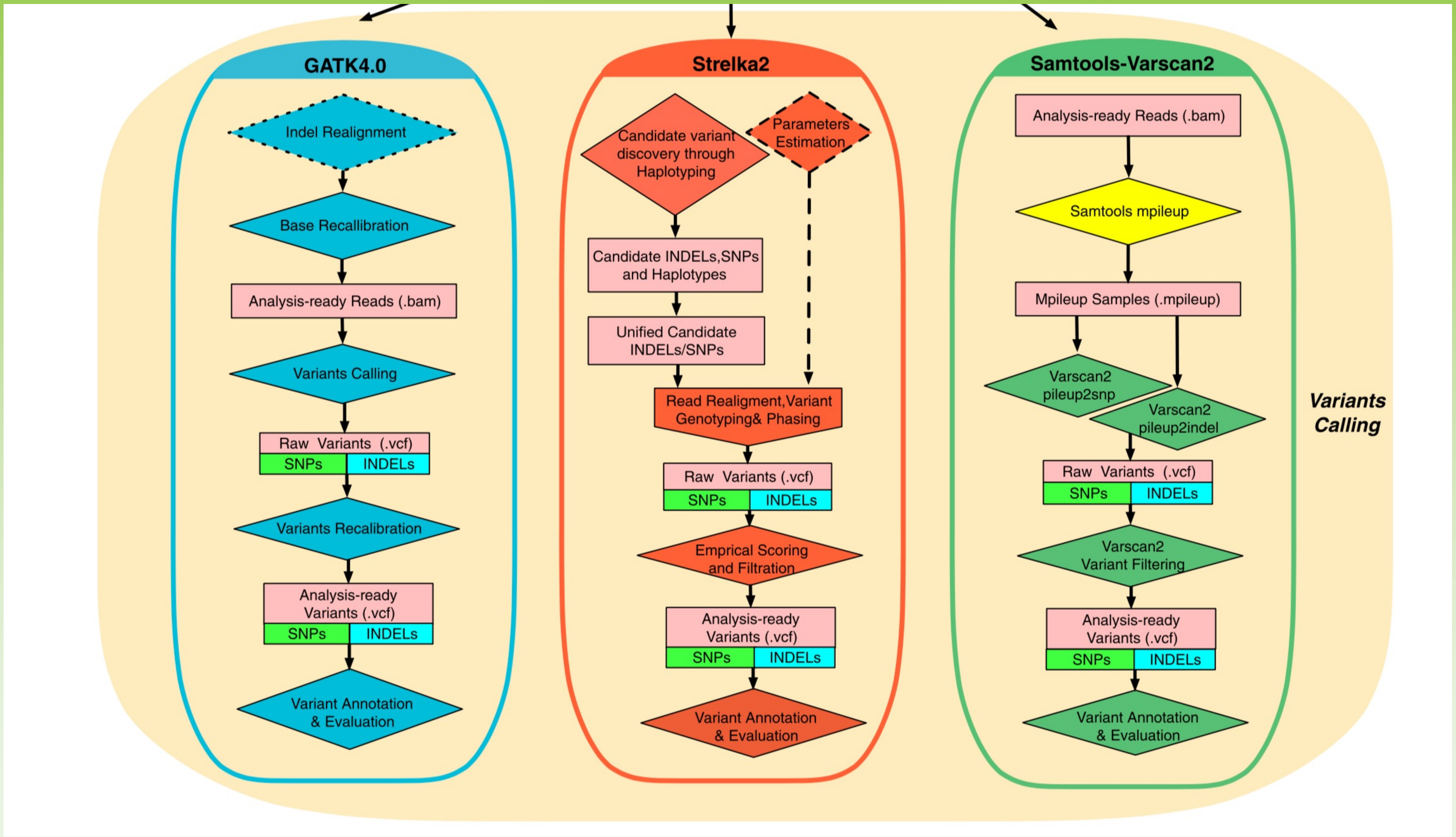
# Identifying Single Nucleotide Polymorphisms



PMID: 31249349

SNP calling

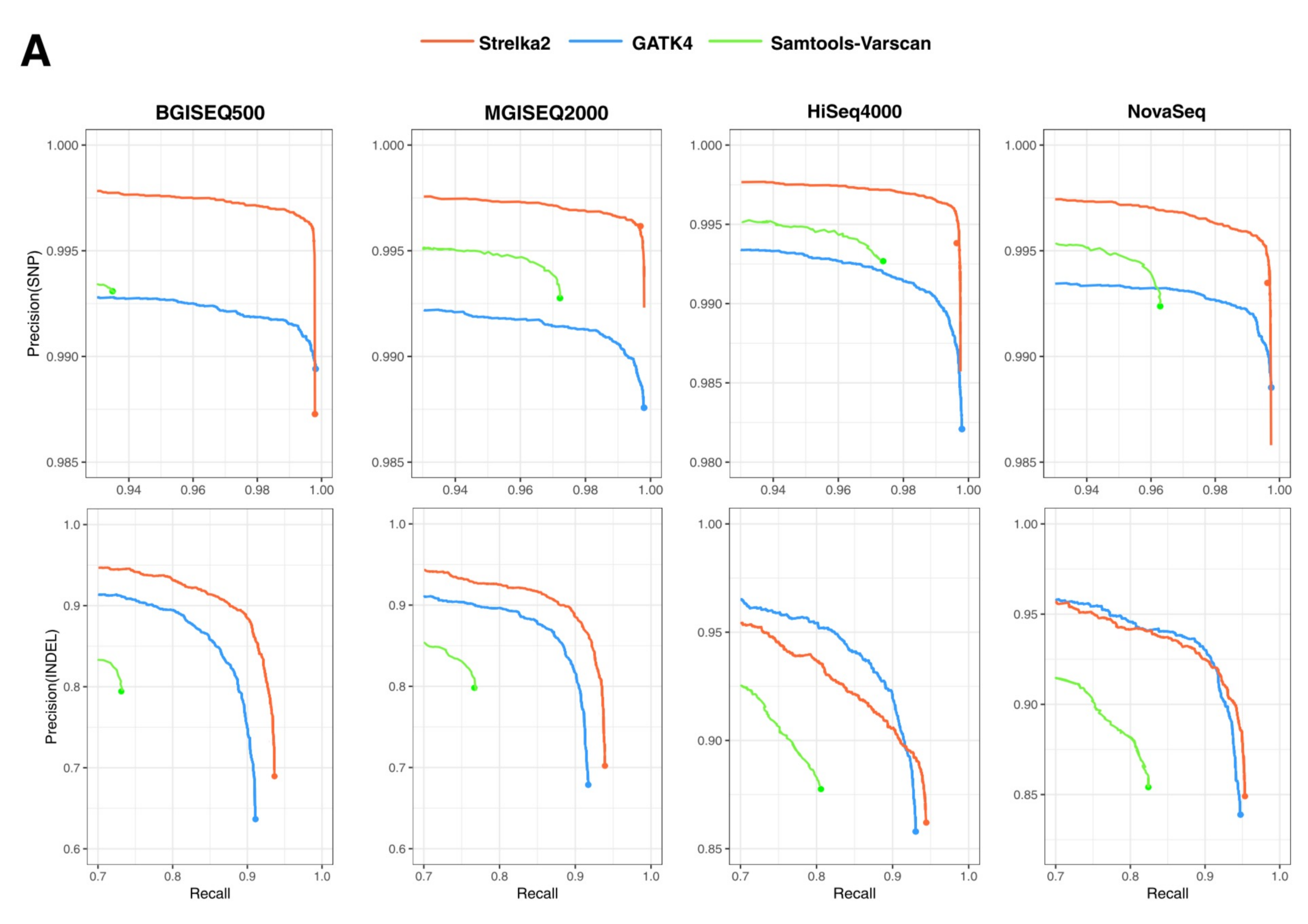
# Identifying Single Nucleotide Polymorphisms


GATK4.0
Indel Realignment
Base Recallibration
Analysis-ready Reads (.bam)
Variants Calling
Raw Variants (.vcf)
SNPs
INDELs
Variants Recalibration
Analysis-ready Variants (.vcf)
SNPs
INDELs
Variant Annotation & Evaluation
Strelka2
Candidate variant discovery through Haplotyping
Parameters Estimation
Candidate INDELs,SNPs and Haplotypes
Unified Candidate INDELs/SNPs
Read Realigment,Variant Genotyping& Phasing
Raw Variants (.vcf)
SNPs
INDELs
Emprical Scoring and Filtration
Analysis-ready Variants (.vcf)
SNPs
INDELs
Variant Annotation & Evaluation
Samtools-Varscan2
Analysis-ready Reads (.bam)
Samtools mpileup
Mpileup Samples (.mpileup)
Varscan2 pileup2snp
Varscan2 pileup2indel
Raw Variants (.vcf)
SNPs
INDELs
Varscan2 Variant Filtering
Analysis-ready Variants (.vcf)
SNPs
INDELs
Variant Annotation & Evaluation
Variants Calling

PMID: 31249349

SNP calling

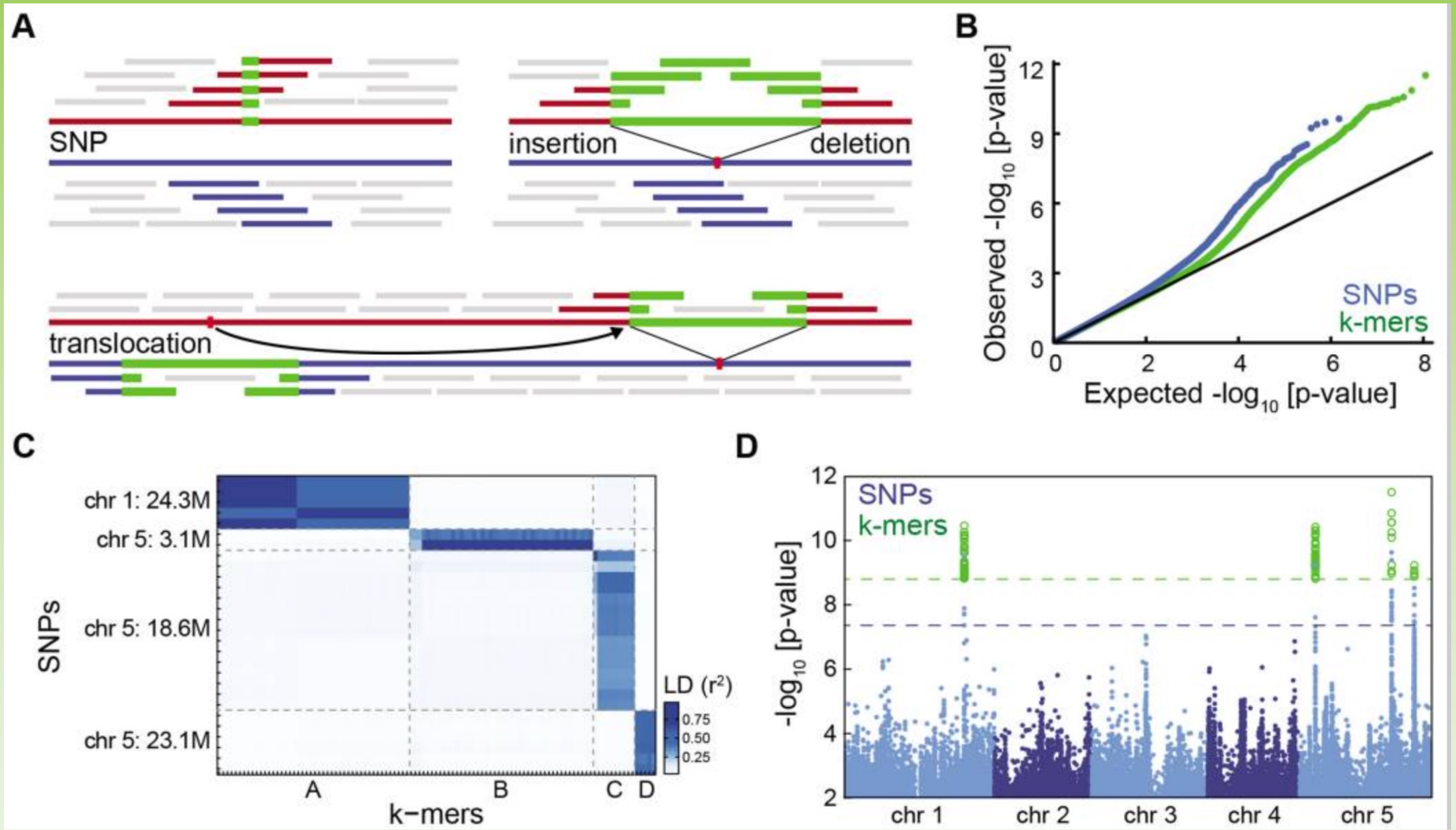
# Identifying Single Nucleotide Polymorphisms



PMID: 31249349

SNP calling

# Reference-free k-mer based approaches



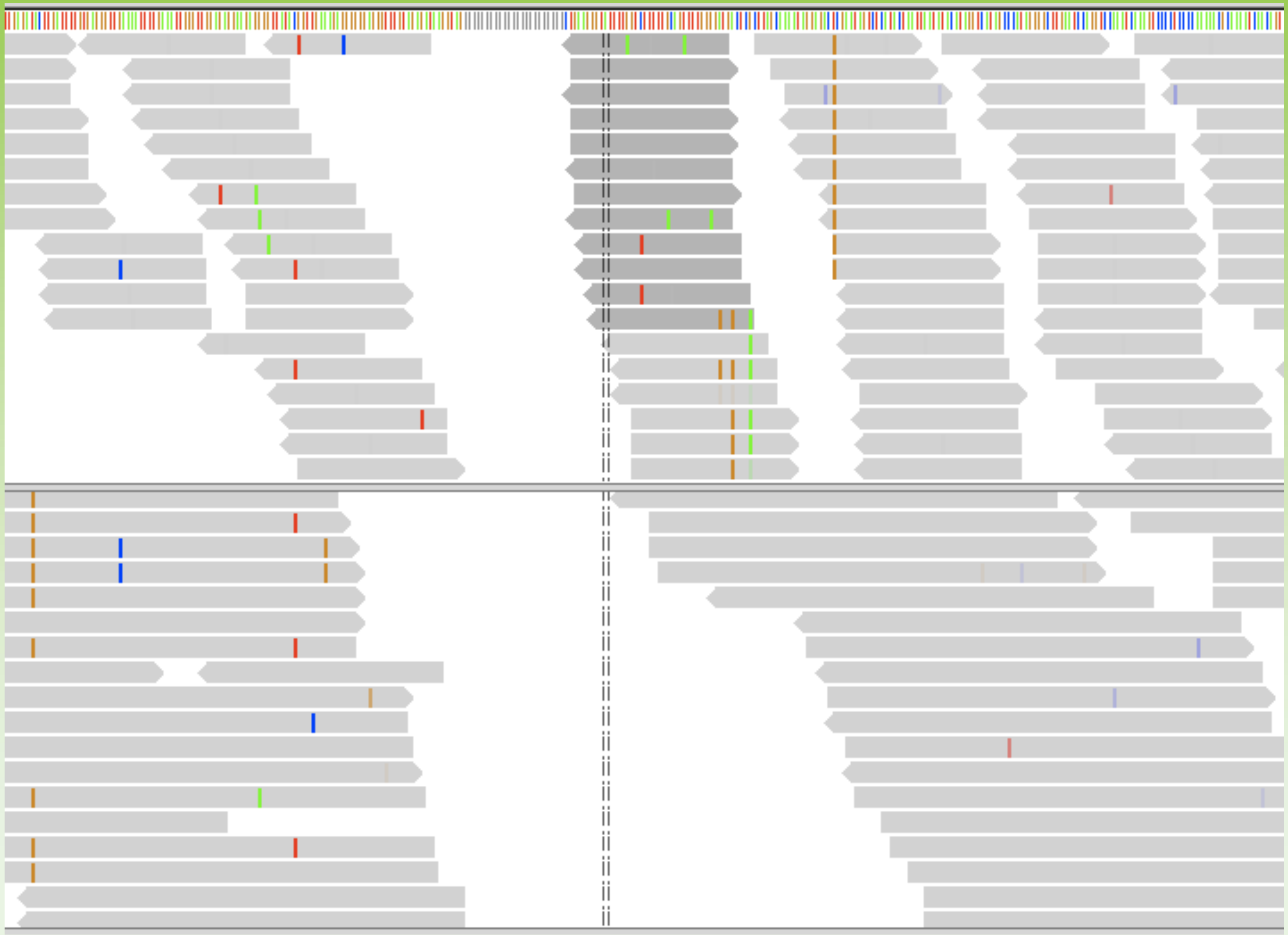
PMID: 32284578

SNP calling

# SAMtools for SNP calling

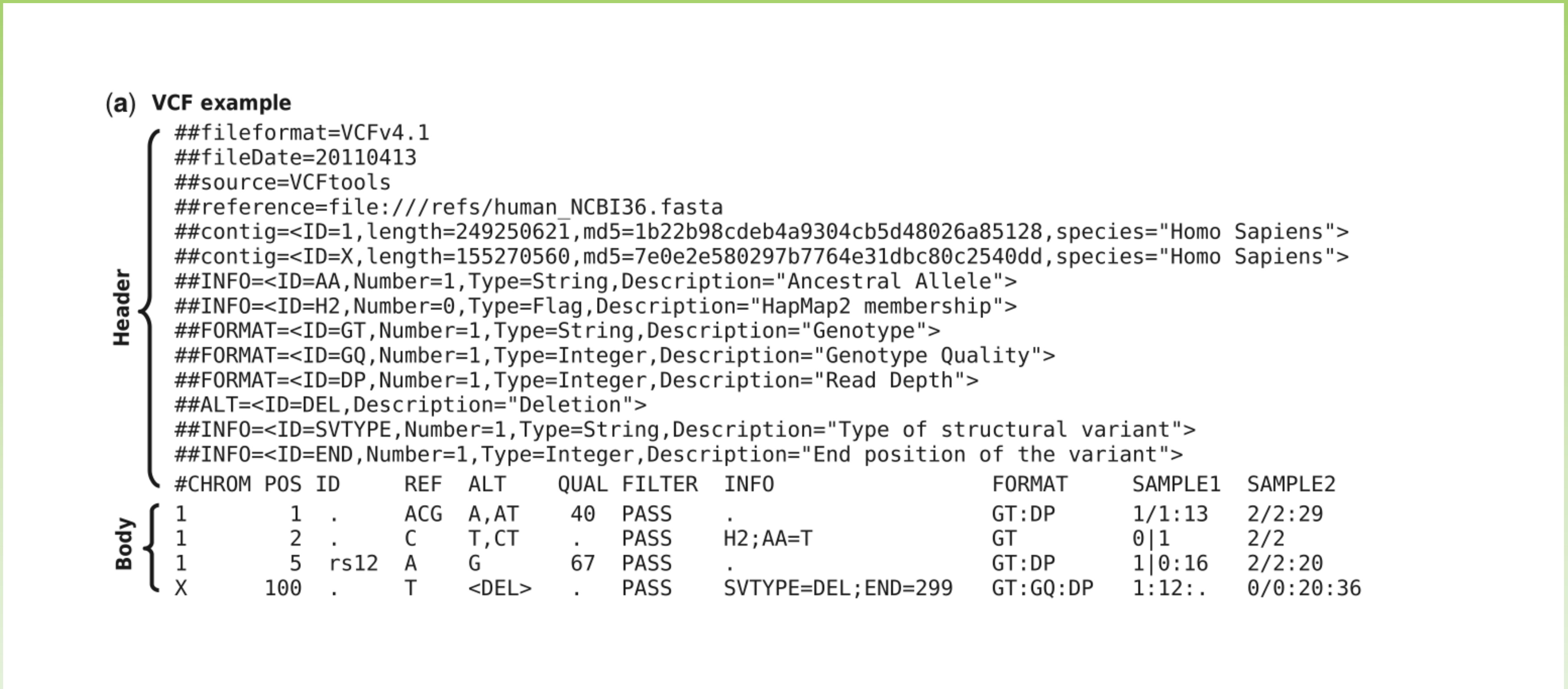
- SAM alignment format is read-based
- For SNP calling, we need a format that is position based
- SAMtools mpileup can produce variant call format (VCF) files based on SAM/BAM alignment files, where all read information is summarized by genomic posistion

# Identifying Single Nucleotide Polymorphisms



SNP calling

# VCF format

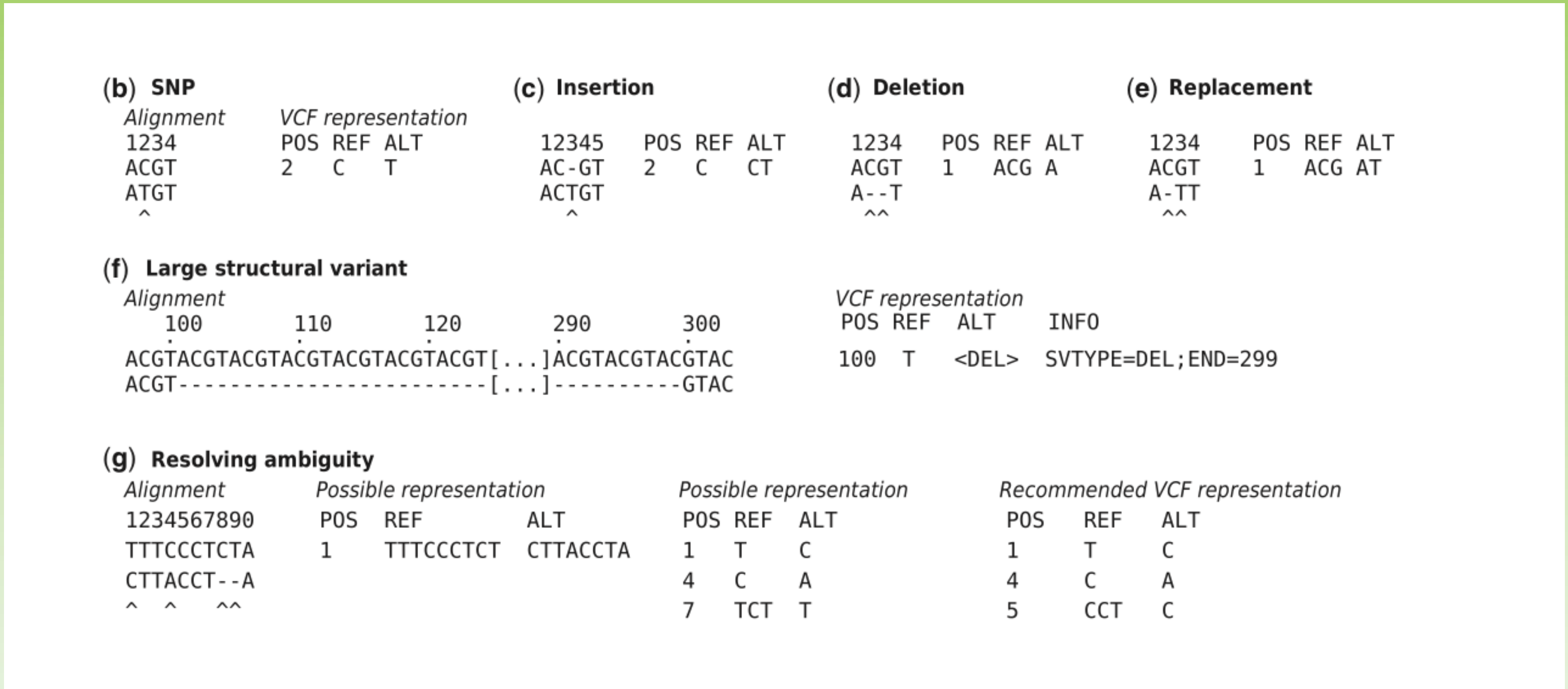

Header
Body

(a) **VCF example**

```
##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., Bioinformatics (2011) vol. 27 (15)

# VCF format



**(b) SNP**

```
Alignment    VCF representation
1234         POS REF ALT
ACGT         2   C   T
ATGT
 ^
```

**(c) Insertion**

```
12345   POS REF ALT
AC-GT   2   C   CT
ACTGT
  ^
```

**(d) Deletion**

```
1234   POS REF ALT
ACGT   1   ACG A
A--T
 ^^
```

**(e) Replacement**

```
1234   POS REF ALT
ACGT   1   ACG AT
A-TT
 ^^
```

**(f) Large structural variant**

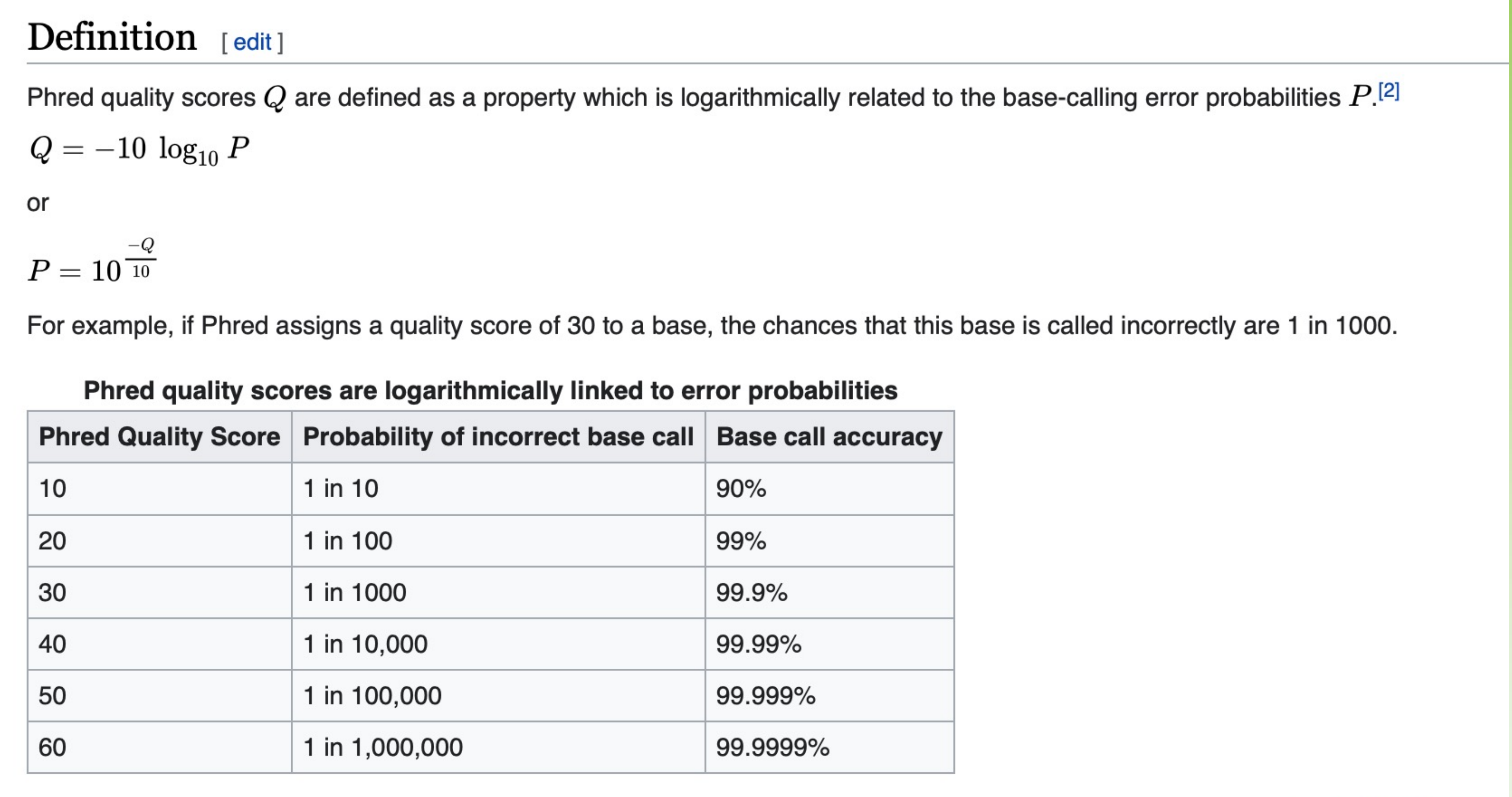
```
Alignment                                                VCF representation
   100       110       120       290       300            POS REF  ALT    INFO
   .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC           100  T   <DEL>  SVTYPE=DEL;END=299
ACGT------------------------[...]----------GTAC
```

**(g) Resolving ambiguity**

```
Alignment     Possible representation     Possible representation   Recommended VCF representation
1234567890    POS  REF        ALT         POS REF ALT               POS   REF   ALT
TTTCCCTCTA    1    TTTCCCTCT  CTTACCTA    1   T   C                 1     T     C
CTTACCT--A                                4   C   A                 4     C     A
^  ^   ^^                                 7   TCT T                 5     CCT   C
```

Danecek et al., Bioinformatics (2011) vol. 27 (15)

# The PHRED scale



## Definition [ edit ]

Phred quality scores $Q$ are defined as a property which is logarithmically related to the base-calling error probabilities $P$.[2]

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

**Phred quality scores are logarithmically linked to error probabilities**

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

https://en.wikipedia.org/wiki/Phred_quality_score

# VCF format

**Genotype calls for each sample**

**Format:** GT:PL:GQ

##FORMAT=<ID=**GT**,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=**PL**,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=**GQ**,Number=1,Type=Integer,Description="Genotype Quality”>

Examples:

**0/0:0,33,255:36**

**1/1:255,117,0:99**

**0/1:6,0,255:4**

Three possible genotypes: 0/0, 0/1, and 1/1. We set the most likely genotype PL to 0 for easy reading purpose. The other values are scaled relative to this most likely genotype. Keep in mind that when we say PL is the "Phred-scaled likelihood of the genotype", we mean it is "How much less likely that genotype is compared to the best one"

See also: https://www.broadinstitute.org/gatk/guide/article?id=1268