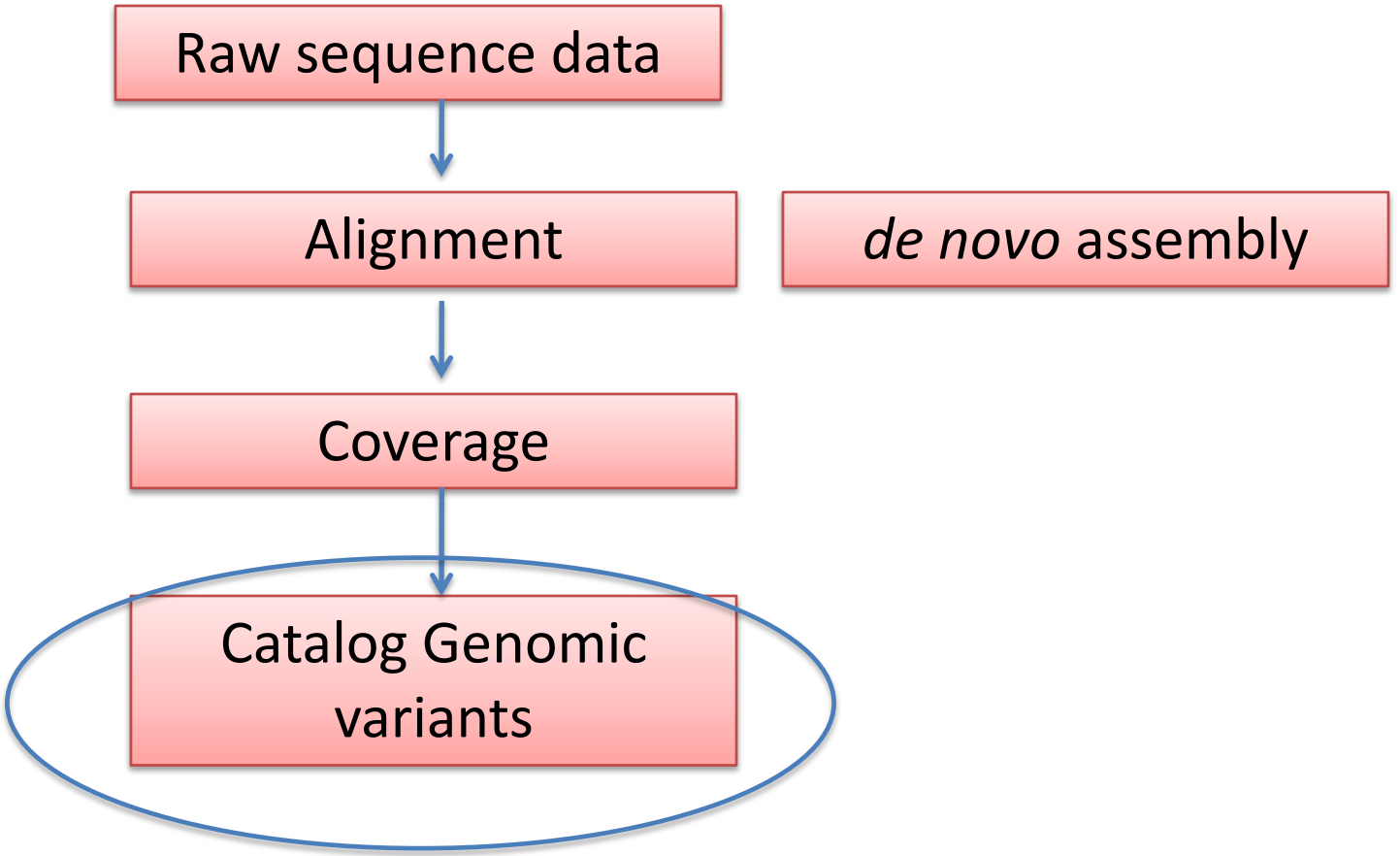
# Variant detection and calling

London School of Hygiene and
Tropical Medicine

# Some aspects of the course




Population genetics

Whole genome Association studies

# Outline

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

# Examples of variants

```
ACTCTACGATTTACGGTACTTAGGAGCATATGCTACT
ACTGTACGATTTACGGTACTTAG.AGCATATGCTACT
```

**SNP**
single nucleotide polymorphism

**indel**
insertion / deletion



**Copy number variation** involving a large chunk of DNA that includes the whole of gene B

# Many types of variants

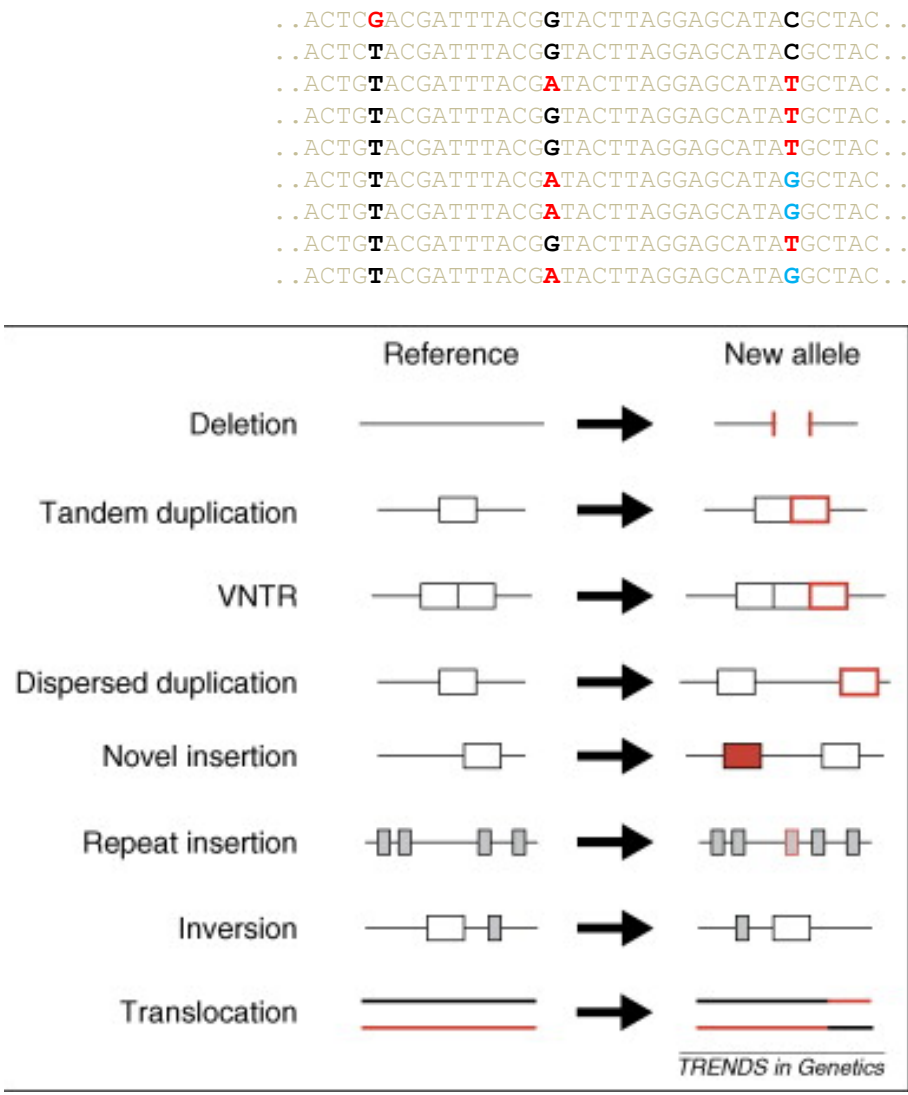
**From small scale ...**

- Single nucleotide polymorphisms (SNPs)
- Insertions and deletions (indels)
- Variable tandem repeats (VNTR)

**... to large scale (>1kb) ...**

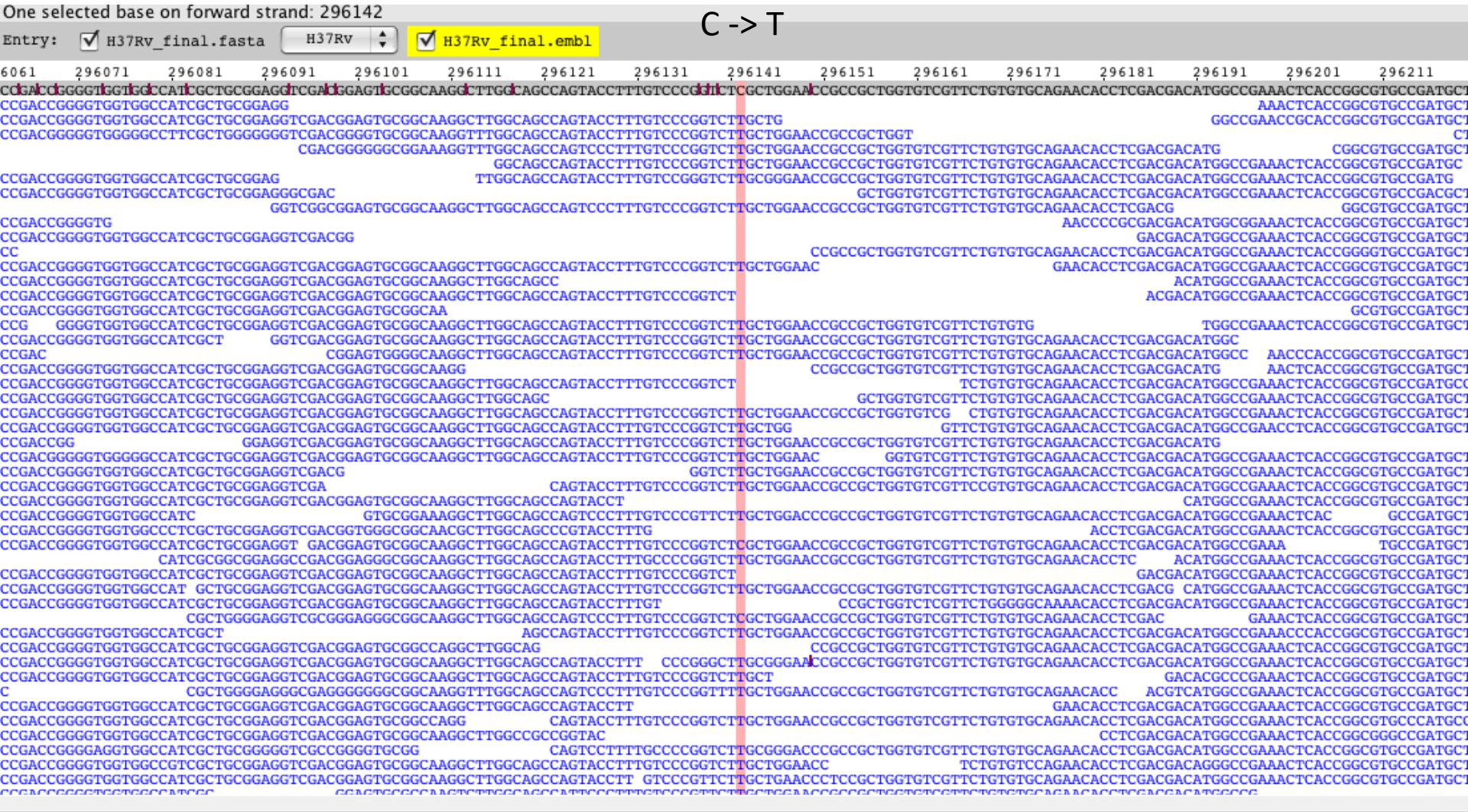
- Copy number variations (CNVs)
- Insertions and deletions
- Inversions

**... and things in between small and large, and combinations of above**




Aikan et al, 2011

# Detecting SNPs from alignments




Lisboa TB strain

~450 SNPs

Rv0245

Artemis view

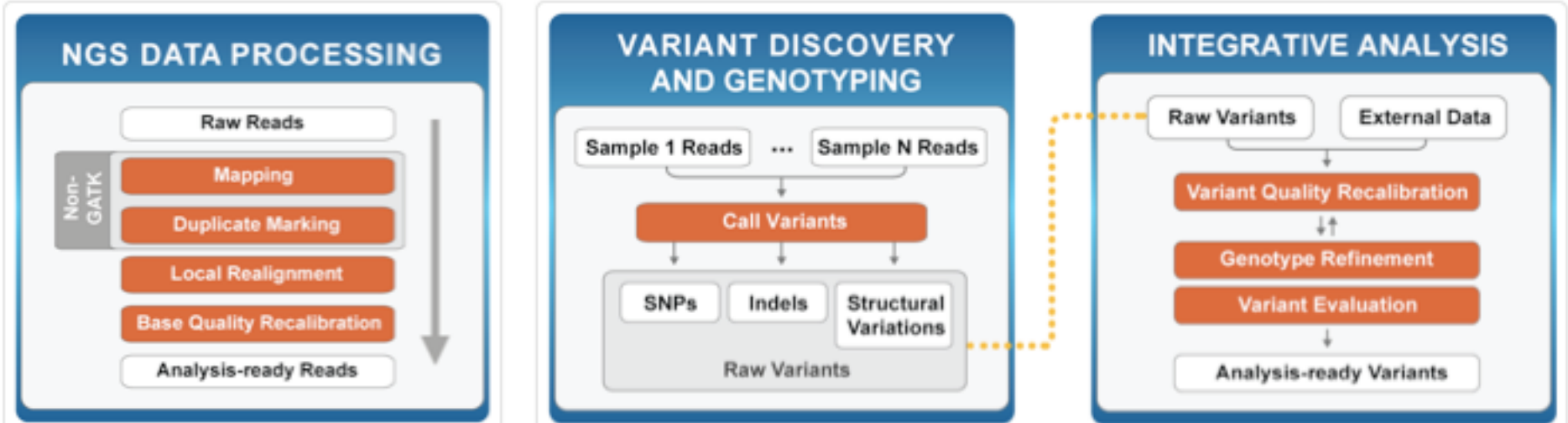
# SNP calling

- Tools include:
  - GATK (Genome Analysis Toolkit) *
  - **SAMtools mpileup (MAQ SNP Caller) ***
  - CASAVA SNP Caller
  - Commercial packages (CLC Bio, Genomatix, etc.)

* Used by the open-source community

# GATK

- Java library (http://www.broadinstitute.org/gatk)
- Easily extendable
- Ready-to-use tools are scarce
- SNP Calling + Annotation is difficult to get running, long runtime (chromosome-by-chromosome works)
- Supports most common file formats (e.g. VCF)



# SAMtools mpileup

- Call & Filter SNPs and INDELs
- Very fast compared to GATK
- SNP Calling (output in BCF format)

```
samtools mpileup –uf ref.fa aln1.bam aln2.bam|bcftools call –mvO z - > var.raw.vcf.gz
```

- SNP Filtering (output in Variant Calling format (VCF))

```
zcat var.raw.vcf.gz |vcfutils.pl varFilter  -D 100 > var.flt.vcf
```

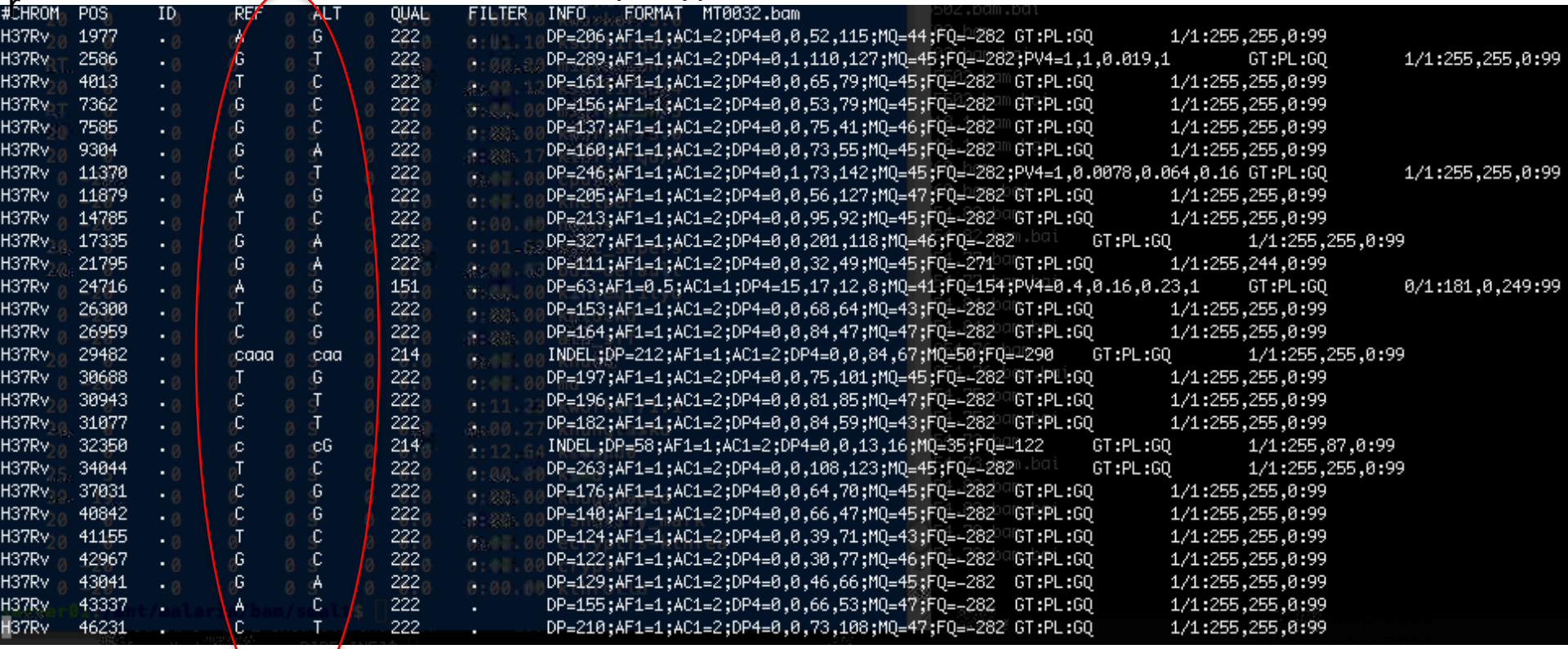
# SNPs and small indels in VCF format

Chr

ID

Alleles

Information concerning read depth and genotype / haplotype calls



| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | MT0032.bam |
|---|---|---|---|---|---|---|---|---|---|
| H37Rv | 1977 | . | A | G | 222 | . | DP=206;AF1=1;AC1=2;DP4=0,0,52,115;MQ=44;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 2586 | . | G | T | 222 | . | DP=289;AF1=1;AC1=2;DP4=0,1,110,127;MQ=45;FQ=-282;PV4=1,1,0.019,1 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 4013 | . | T | C | 222 | . | DP=161;AF1=1;AC1=2;DP4=0,0,65,79;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 7362 | . | G | C | 222 | . | DP=156;AF1=1;AC1=2;DP4=0,0,53,79;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 7585 | . | G | C | 222 | . | DP=137;AF1=1;AC1=2;DP4=0,0,75,41;MQ=46;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 9304 | . | G | A | 222 | . | DP=160;AF1=1;AC1=2;DP4=0,0,73,55;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 11370 | . | C | T | 222 | . | DP=246;AF1=1;AC1=2;DP4=0,1,73,142;MQ=45;FQ=-282;PV4=1,0.0078,0.064,0.16 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 11879 | . | A | G | 222 | . | DP=205;AF1=1;AC1=2;DP4=0,0,56,127;MQ=47;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 14785 | . | T | C | 222 | . | DP=213;AF1=1;AC1=2;DP4=0,0,95,92;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 17335 | . | G | A | 222 | . | DP=327;AF1=1;AC1=2;DP4=0,0,201,118;MQ=46;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 21795 | . | G | A | 222 | . | DP=111;AF1=1;AC1=2;DP4=0,0,32,49;MQ=45;FQ=-271 | GT:PL:GQ | 1/1:255,244,0:99 |
| H37Rv | 24716 | . | A | G | 151 | . | DP=63;AF1=0.5;AC1=1;DP4=15,17,12,8;MQ=41;FQ=154;PV4=0.4,0.16,0.23,1 | GT:PL:GQ | 0/1:181,0,249:99 |
| H37Rv | 26300 | . | T | C | 222 | . | DP=153;AF1=1;AC1=2;DP4=0,0,68,64;MQ=43;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 26959 | . | C | G | 222 | . | DP=164;AF1=1;AC1=2;DP4=0,0,84,47;MQ=47;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 29482 | . | caaa | caa | 214 | . | INDEL;DP=212;AF1=1;AC1=2;DP4=0,0,84,67;MQ=50;FQ=-290 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 30688 | . | T | G | 222 | . | DP=197;AF1=1;AC1=2;DP4=0,0,75,101;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 30943 | . | C | T | 222 | . | DP=196;AF1=1;AC1=2;DP4=0,0,81,85;MQ=47;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 31077 | . | C | T | 222 | . | DP=182;AF1=1;AC1=2;DP4=0,0,84,59;MQ=43;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 32350 | . | c | cG | 214 | . | INDEL;DP=58;AF1=1;AC1=2;DP4=0,0,13,16;MQ=35;FQ=-122 | GT:PL:GQ | 1/1:255,87,0:99 |
| H37Rv | 34044 | . | T | C | 222 | . | DP=263;AF1=1;AC1=2;DP4=0,0,108,123;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 37031 | . | C | G | 222 | . | DP=176;AF1=1;AC1=2;DP4=0,0,64,70;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 40842 | . | C | G | 222 | . | DP=140;AF1=1;AC1=2;DP4=0,0,66,47;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 41155 | . | T | C | 222 | . | DP=124;AF1=1;AC1=2;DP4=0,0,39,71;MQ=43;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 42967 | . | G | C | 222 | . | DP=122;AF1=1;AC1=2;DP4=0,0,30,77;MQ=46;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 43041 | . | G | A | 222 | . | DP=129;AF1=1;AC1=2;DP4=0,0,46,66;MQ=45;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 43337 | . | A | C | 222 | . | DP=155;AF1=1;AC1=2;DP4=0,0,66,53;MQ=47;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |
| H37Rv | 46231 | . | C | T | 222 | . | DP=210;AF1=1;AC1=2;DP4=0,0,73,108;MQ=47;FQ=-282 | GT:PL:GQ | 1/1:255,255,0:99 |

Position

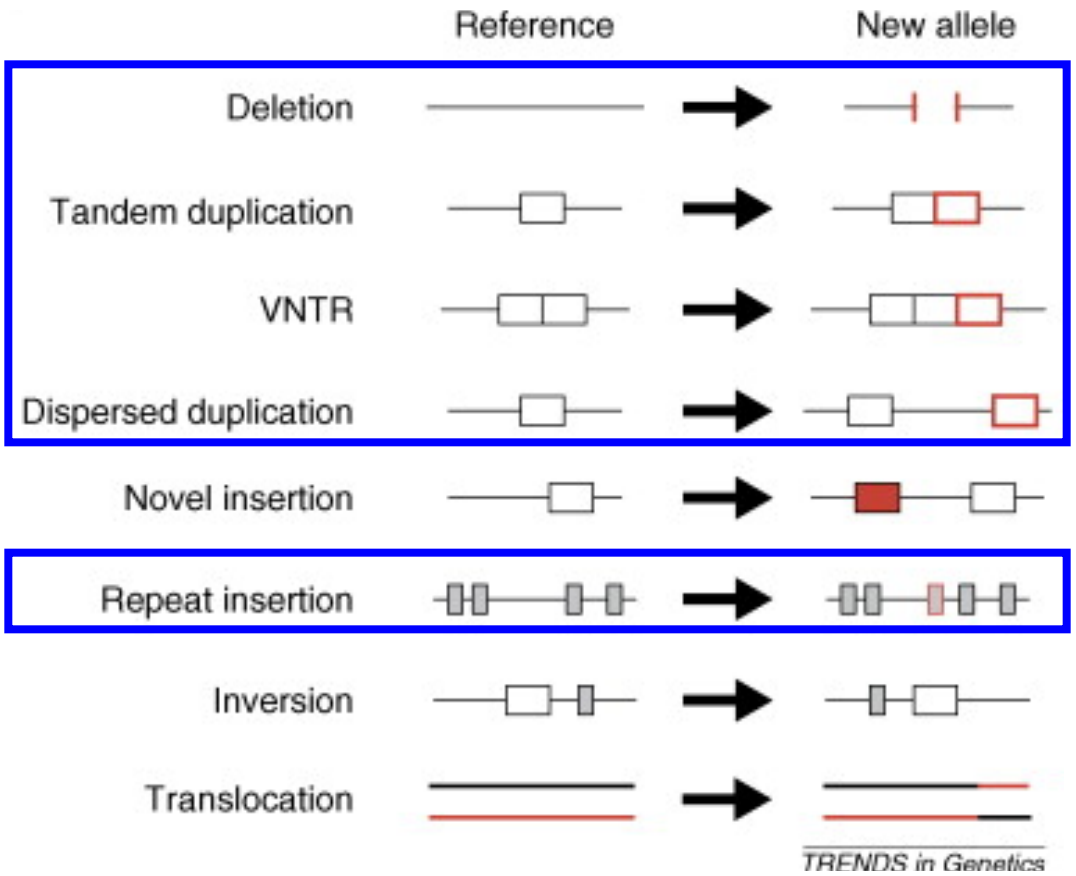
Quality

VCF files generated directly BAM files using Samtools & BCF/VCFtools

# Variant call format (VCF)

- 1000 genomes format, with meta-information in the header lines to describe elements
- Flexible format
  - Specify filtering (assigning PASS in “FILTER”)
  - Incorporate locations of SVs or RS numbers (“ID”)
  - Different ploidy genotypes, with phasing (“GT”)
  - Multi-sample files
  - Zipped for some viewers
- Viewers
  - Tablet, Artemis, Savant, VarB
- Interfaces with Plink for association analysis

# Structural Variation

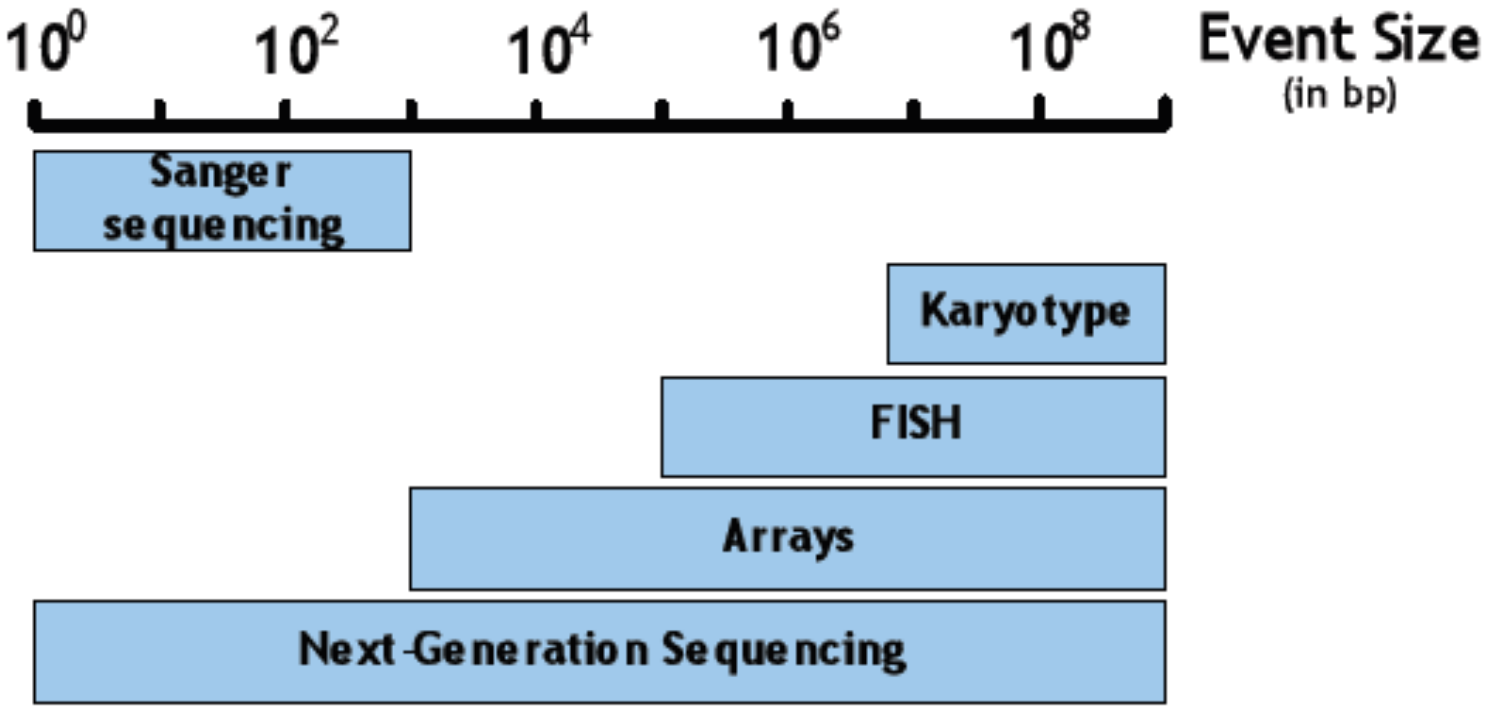



Detectable with arrays

All are accessible using sequencing, with breakpoint estimation

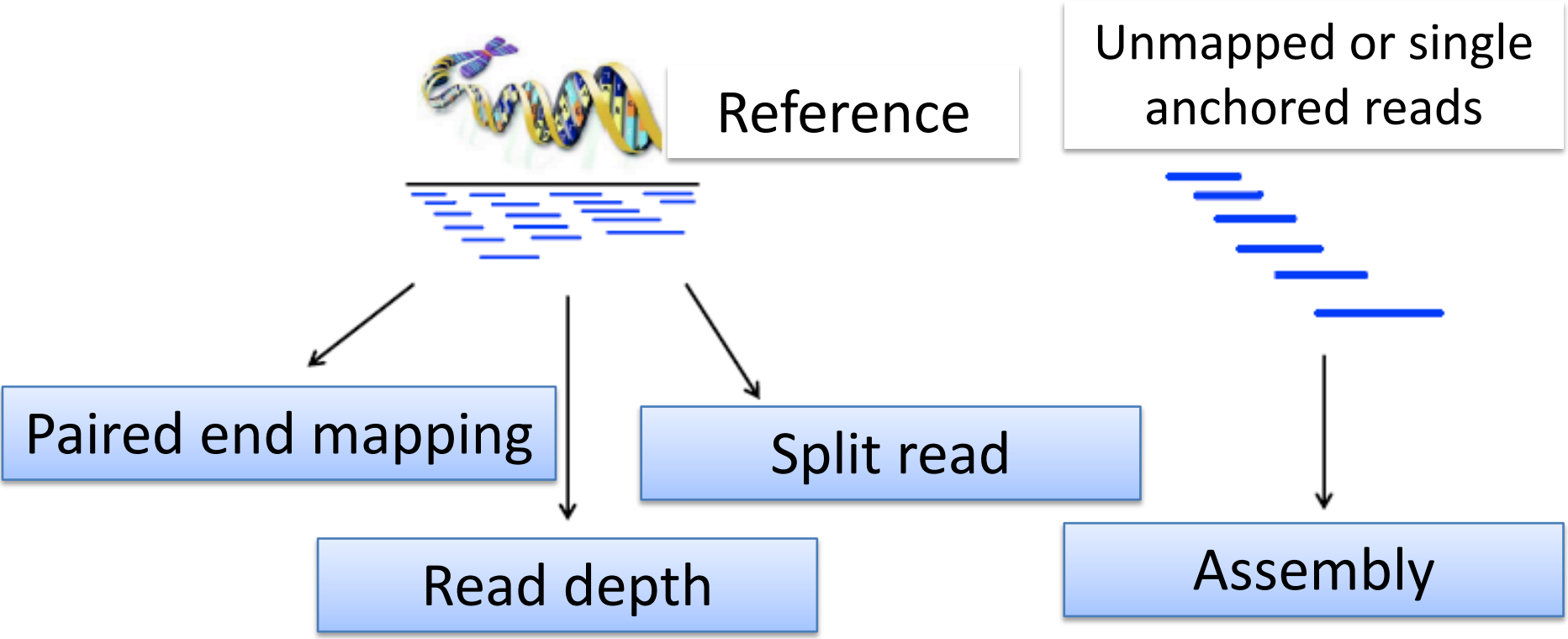
Citation: Hurles et, Trends in Genetics, 2007

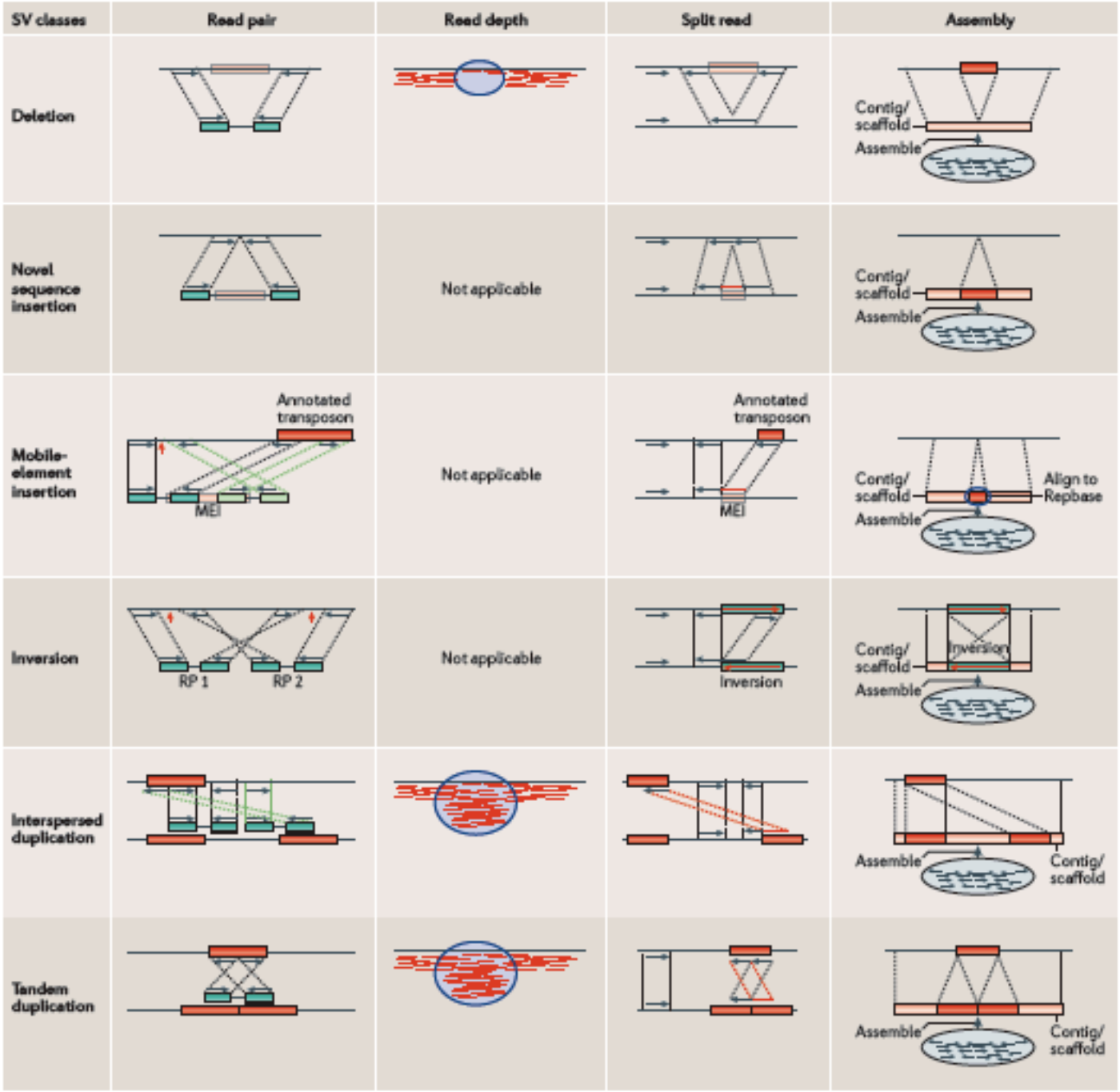
# SV Resolution




Rausch et al, 2012

# Approaches


Reference
Unmapped or single anchored reads
Paired end mapping
Split read
Read depth
Assembly

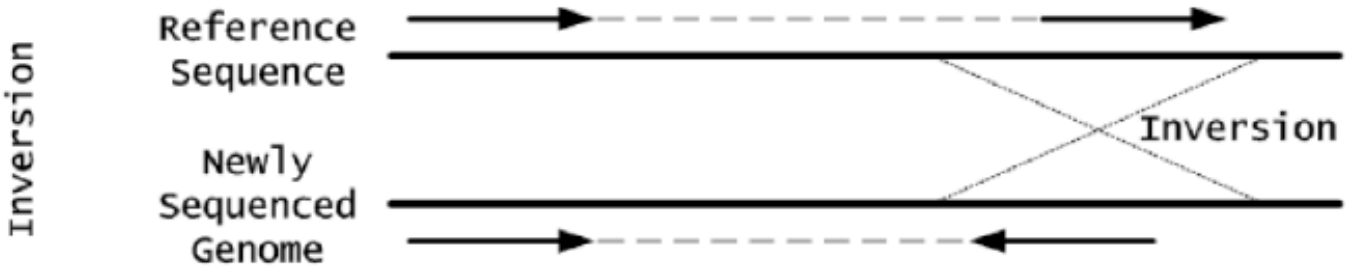



Alkan et al., 2011

# Paired end (read pair) mapping abnormalities

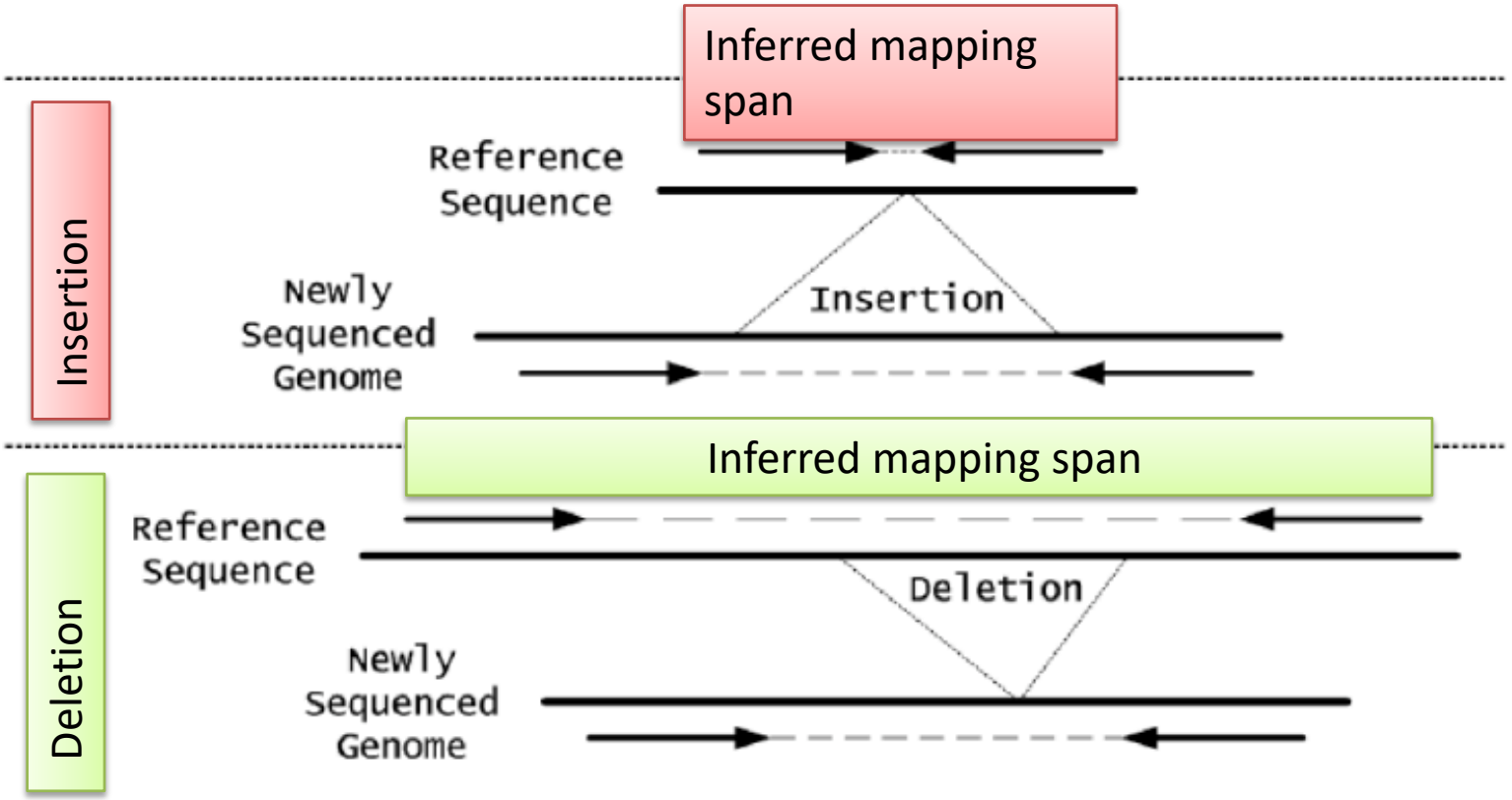
Detecting inversions (& some tandem duplications)

-where one read maps in the correct orientation, and the other maps in reverse to the reference




Lee et al. 2007

# Read pair approach



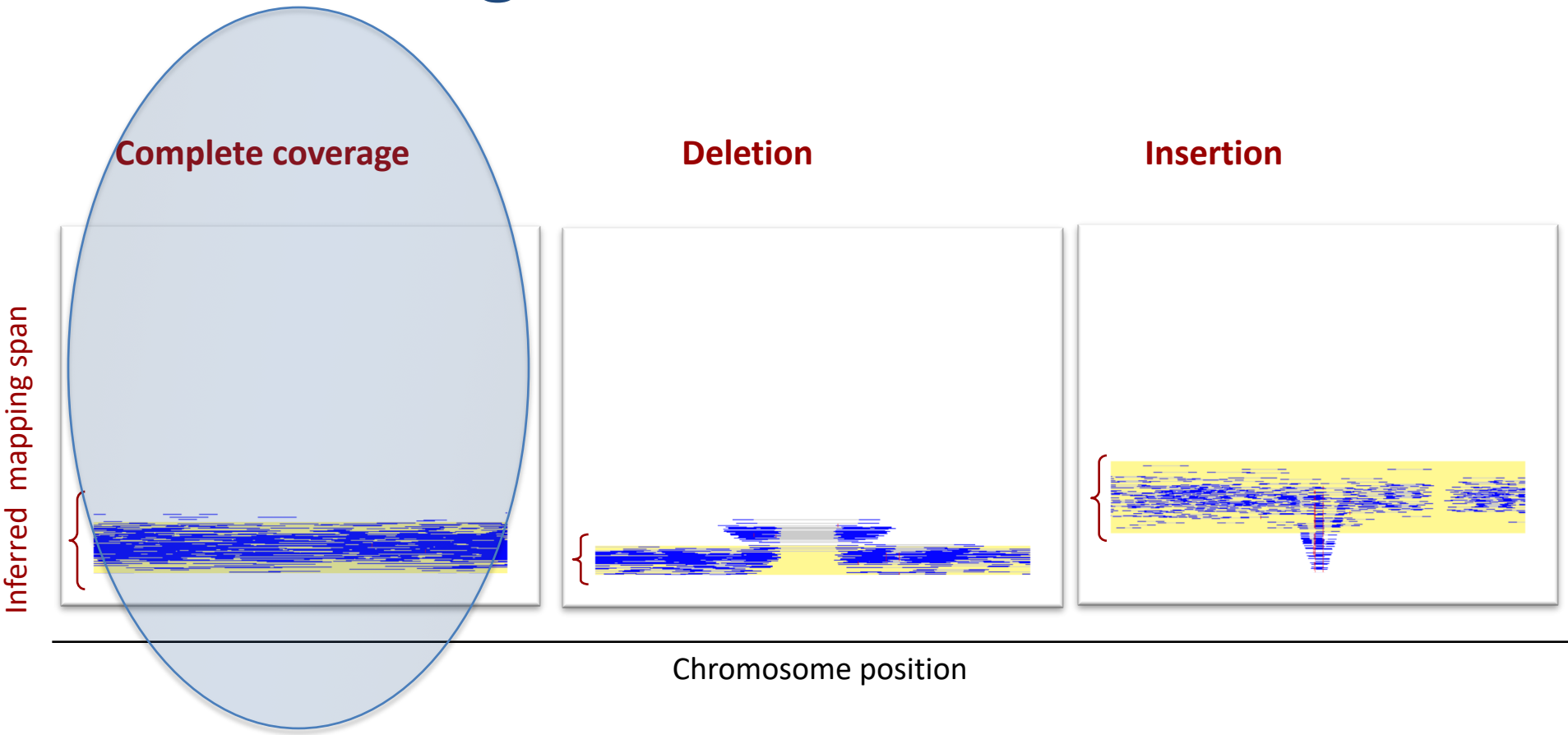

## Indels:

If the inferred mapping span is < expected → insertion

If the inferred mapping span is > expected → deletion

Korbel et al. 2007

# Signatures of indels

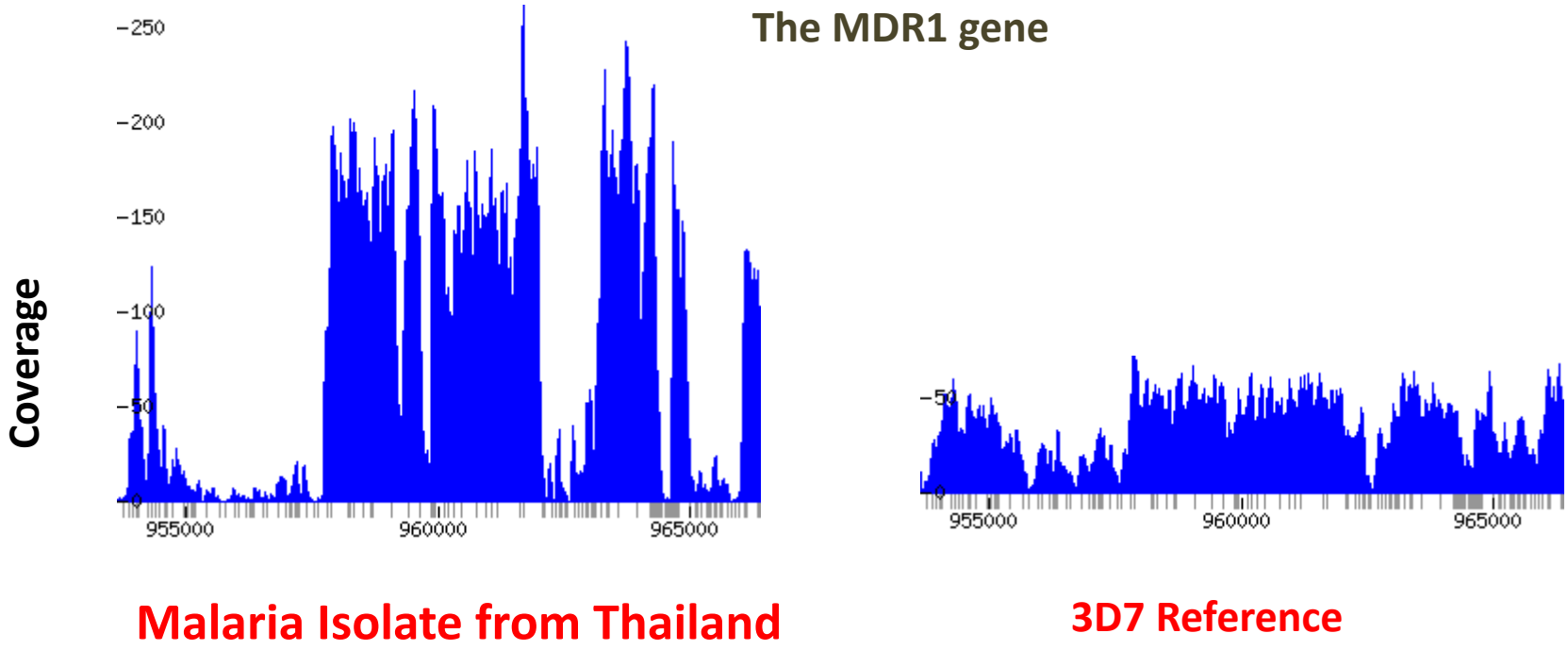



"If the inferred mapping span from the experiment is close to that expected (yellow), then we have complete coverage

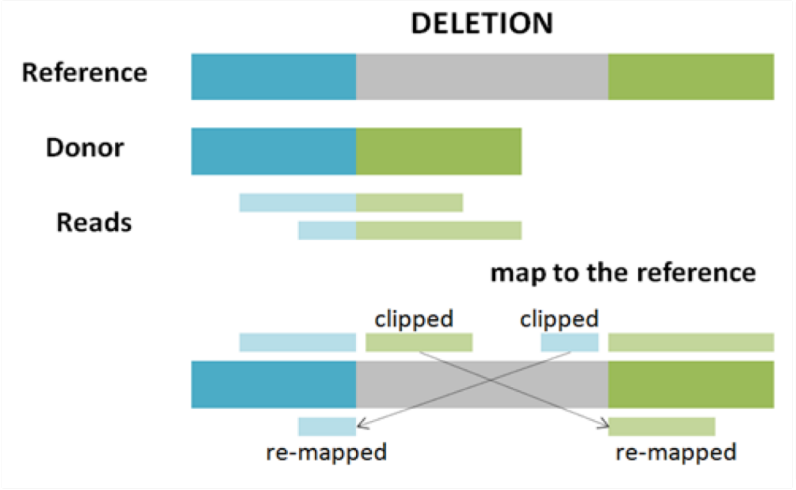
# Using Read depth

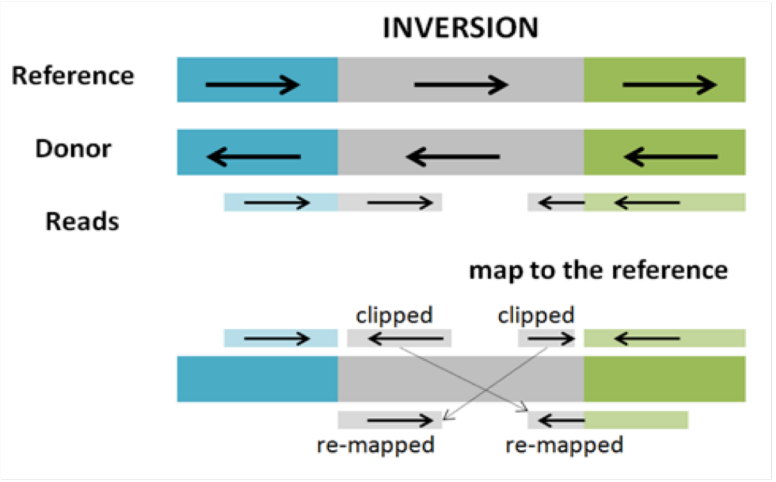
- Low or zero coverage may indicate the presence of a deletion
- Excess coverage may imply duplications

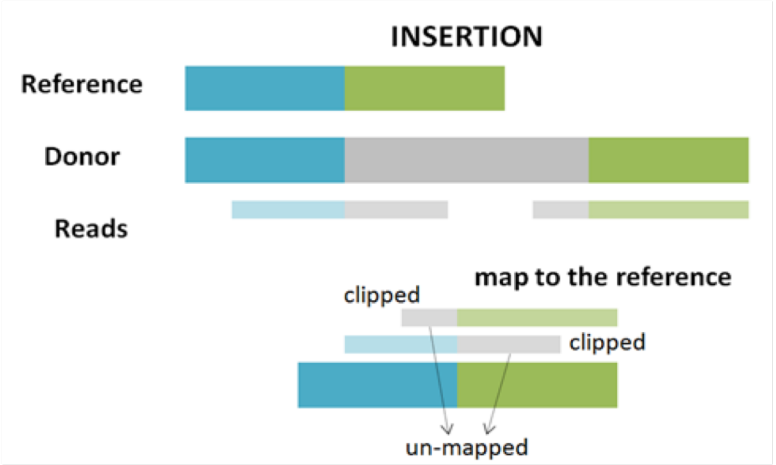
(Yoon et al, 2009; Boeva et al, 2011)



# Split read




- Methods modified from Sanger read approaches
- Defining SV breakpoints by looking at where alignments of reads are broken
- A continuous stretch of gaps in the read → deletion
  or in the reference → insertion
- Limited by read size, but can anchor reads using pairs (e.g. Ye et al., 2009)

# SV detection Tools

- **Read-depth tools**

  – CNVer, CNVnator, FreeC, Sepulveda et. al (2013), Yoon et al. (2009)

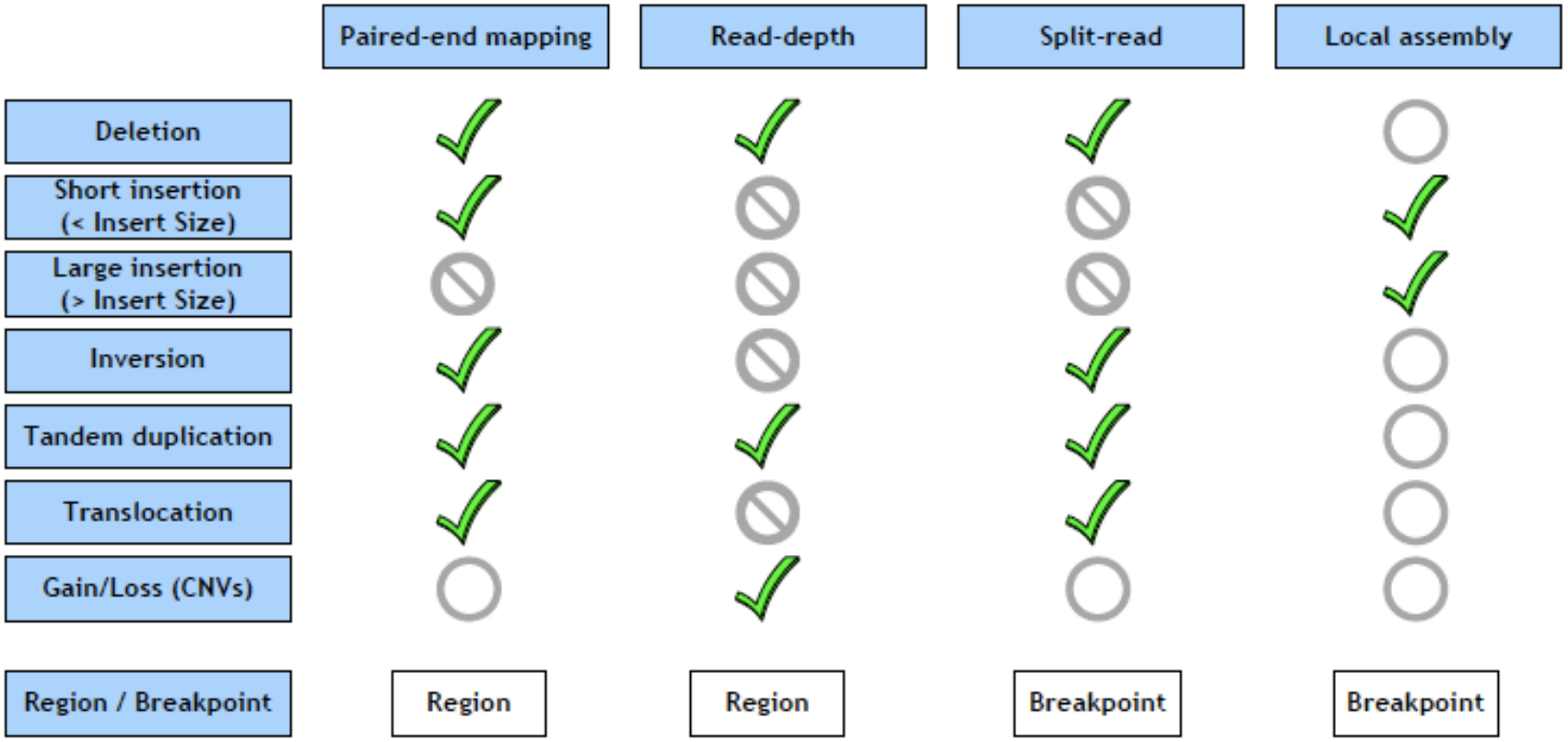
- **Paired-end mapping**

  –Delly, Breakdancer, Corona, HYDRA, MoDIL, MoGUL, PEMer, SPANNER

- **Split-read**

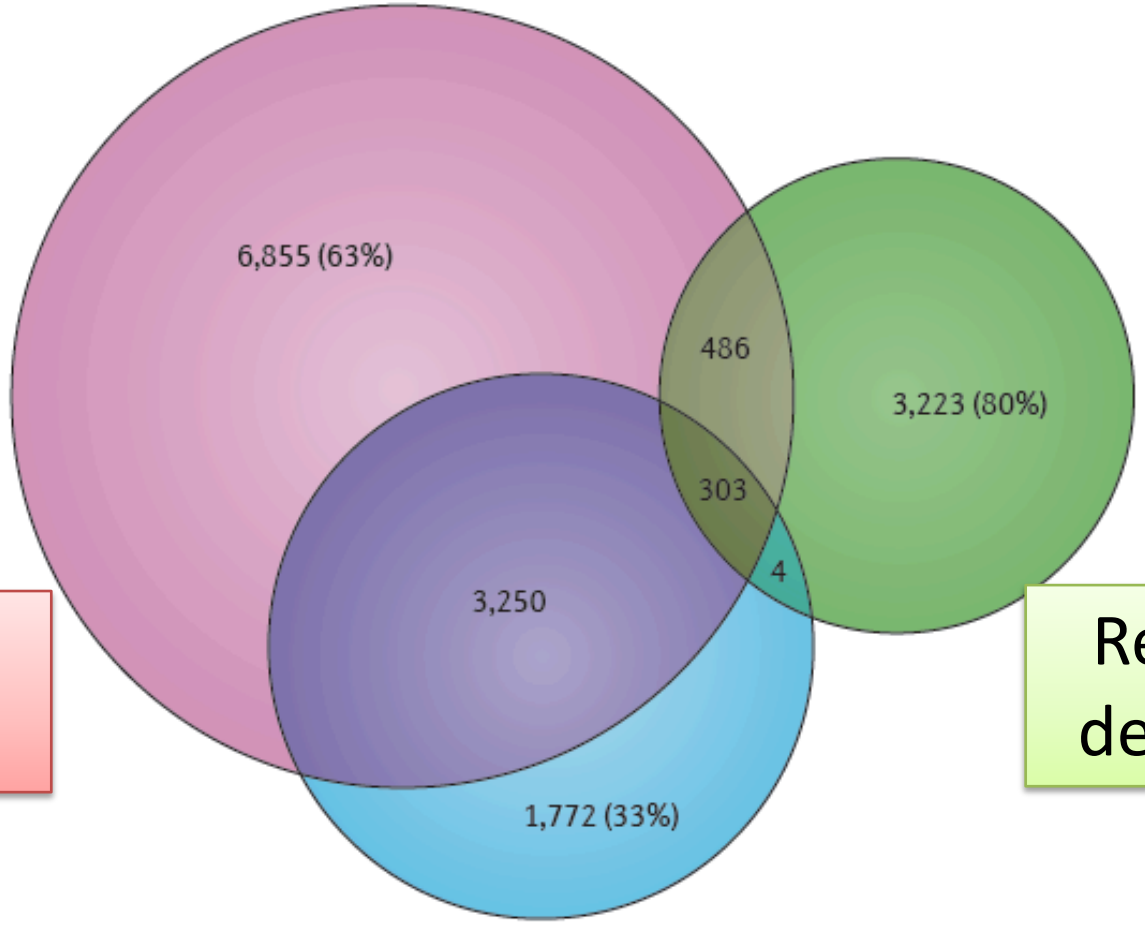
  – Delly, Age, Pindel

# Summary of approaches



# Overlap

1000 Genomes Project
14 distinct algorithms
Mill et al, 2011

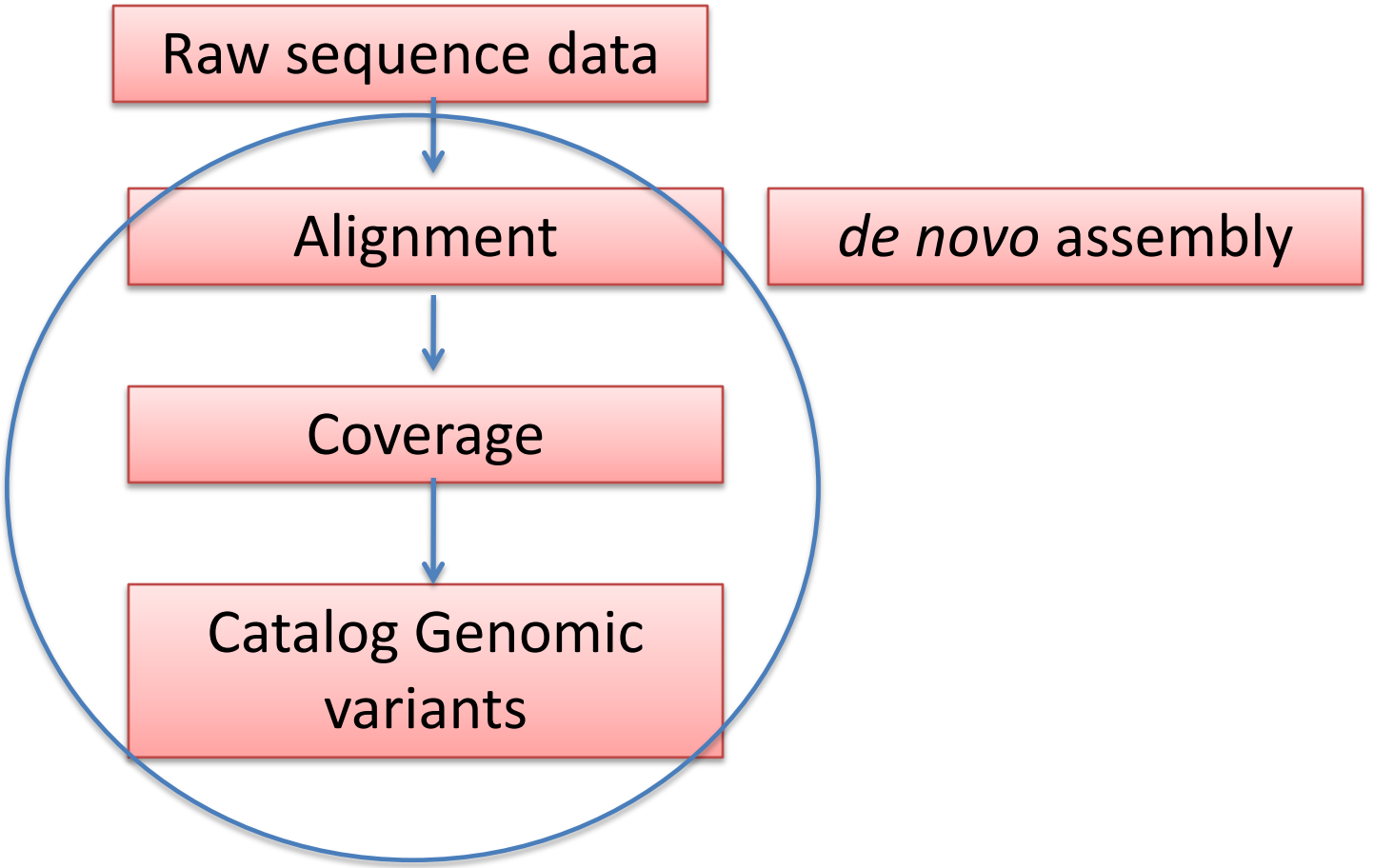



Paired end mapping

Read depth

Split read

# Finding genomic variants




Population genetics

Whole genome Association studies