Variant Calling

November 6, 2024

Professor Qasim Ayub

Director Monash University Malaysia Genomics Platform (MUMGP)

Deputy Head of School (Research)

qasim.ayub@monash.edu

Slides Modified From
Petr Danecek
Wellcome Sanger Institute
petr.danecek@sanger.ac.uk

wellcome connecting science

MONASH University | MALAYSIA

Genomics Platform

bsi ISO 9001 Quality Management Systems CERTIFIED

Learning Outcomes

- Understand various types of variation and how they are ascertained.
- Understand how variant calls are made.
- Assess variant quality and visualise variants.
- Annotate variants and assess consequences.

Outline

- DNA variations and how they arise.
- Genomic DNA variations.
- Practical applications.
- Ascertaining variation.
- Analyzing variant calls.
- Variation consequences.

DNA Variations

Any variation or change in the DNA base sequence is referred to as mutation.

How Do DNA Sequences Change?

DNA sequences change over time due to:

- DNA replication errors:
  - *De novo* errors in copying DNA during cell division.
- Recombination.
- Gene conversion.
- Transposition.
- Non-replicative DNA damage:
  - Chemically induced.
  - Radiation.

In sexually reproducing organisms they are only inherited if they are present in the male or female gametes.

Germline vs Somatic mutations

- Germline mutations:
  - Heritable variation in the germ cells.

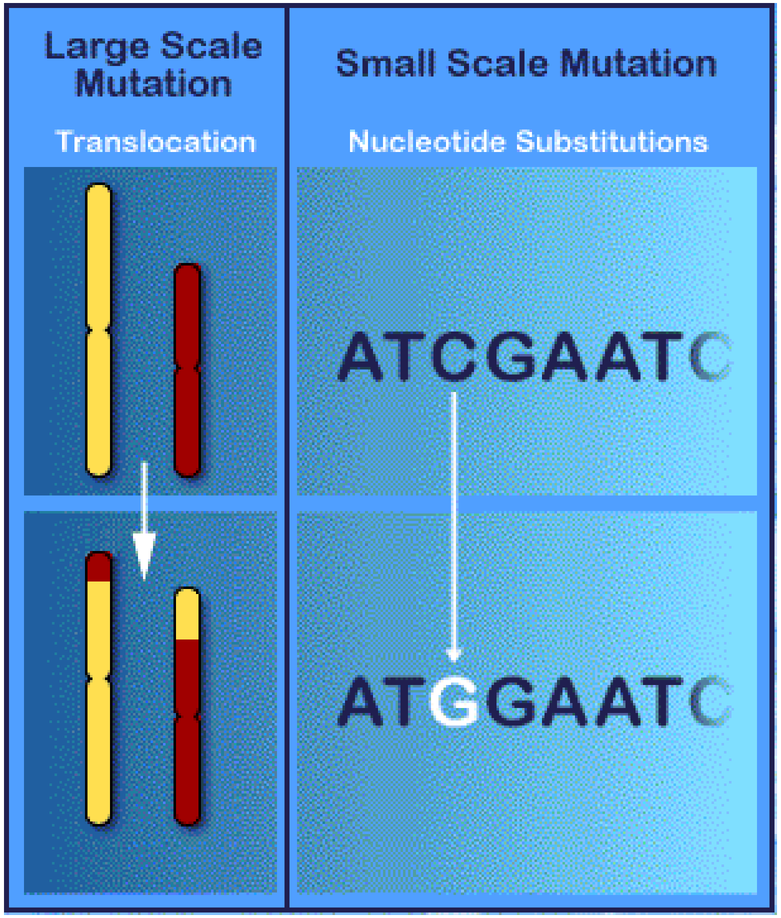
- Somatic mutations:
  - Variation in non-germline tissues.

Types of Genetic Variation

DNA Variations

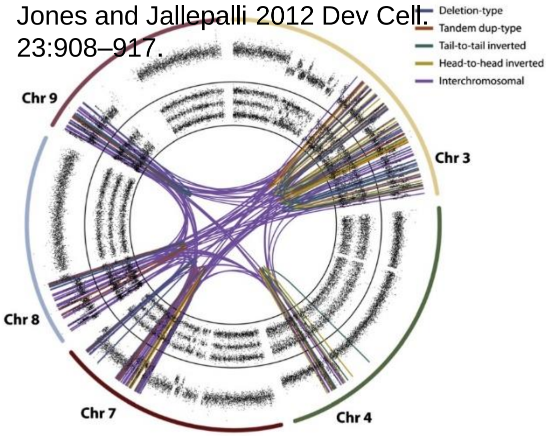
- **Large scale**
  500 > 10^6 bp

- **Small scale**
  <50 bp

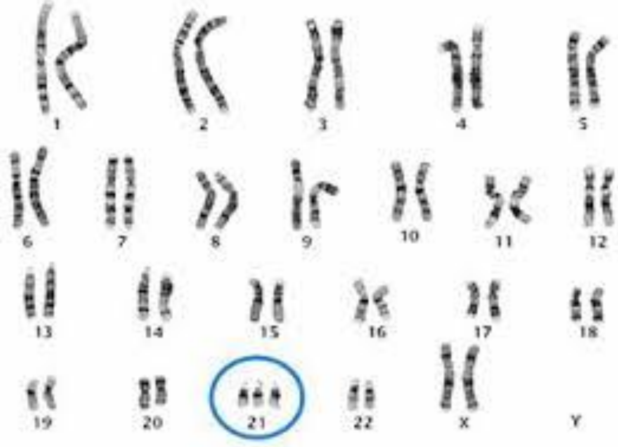


Large Scale Variations

- Gain/loss of chromosomes.
- Chromothripsis.
- Translocations.
- Copy Number Variants (CNVs).
- Structural Variants (SVs)



Jones and Jallepalli 2012 Dev Cell. 23:908–917.

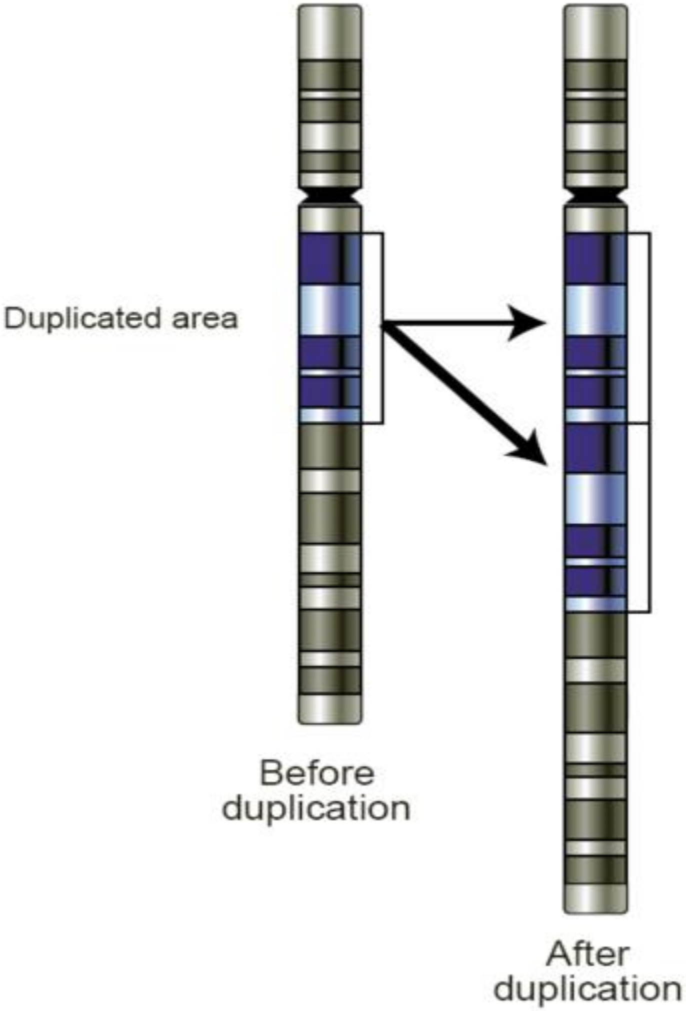


Copy Number Variants

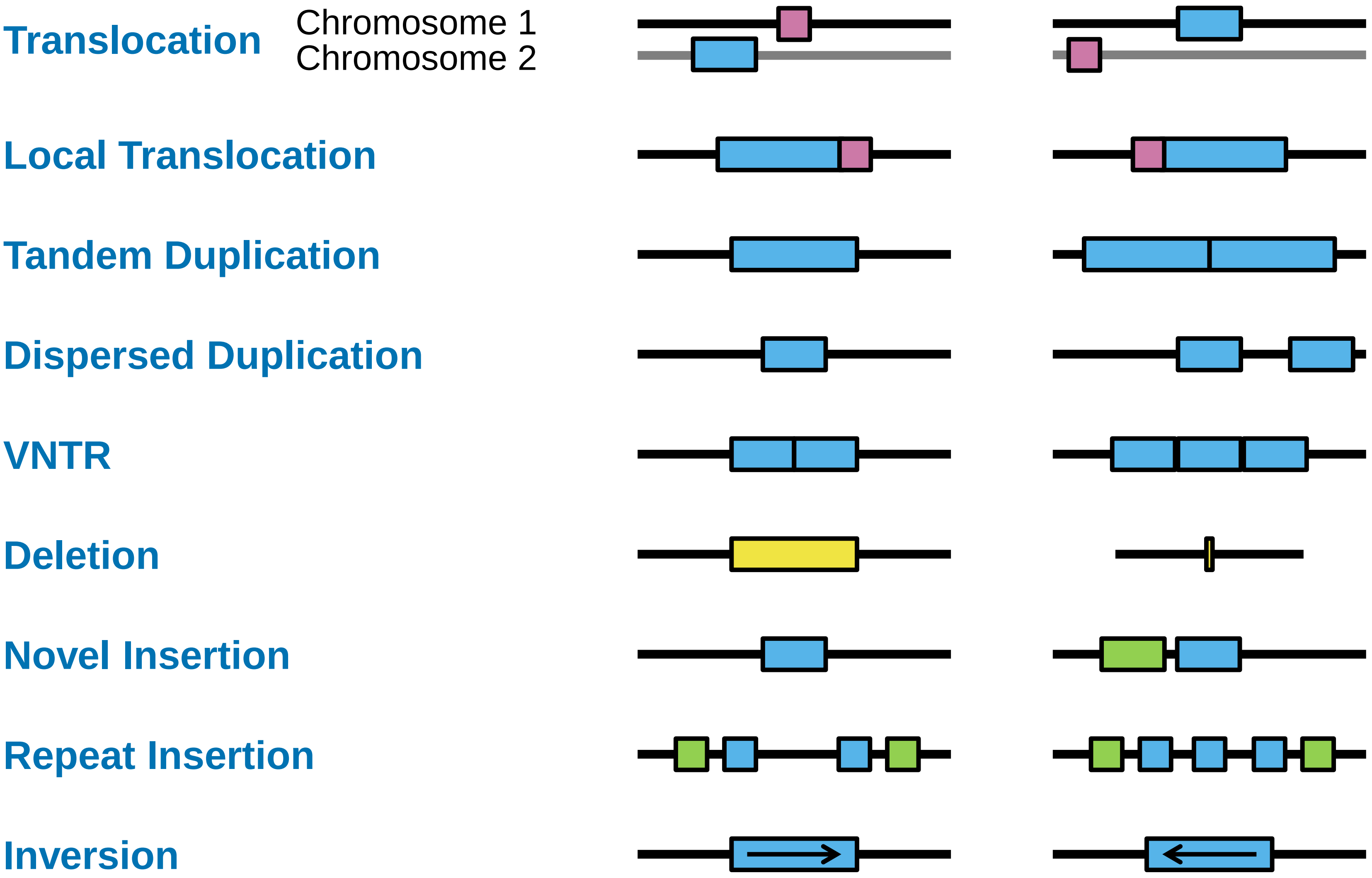
- CNV are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA.

- This variation accounts for roughly 12% of human genomic DNA and each variation may range from about one kilobase several megabases in size.

- A structural variant consists of a DNA sequence >50 bp, typically 1 kilobase, that deviates from a reference sequence in content, order and/or orientation.



Structural Variants



Translocation

Chromosome 1

Chromosome 2

Local Translocation

Tandem Duplication

Dispersed Duplication

VNTR

Deletion

Novel Insertion

Repeat Insertion

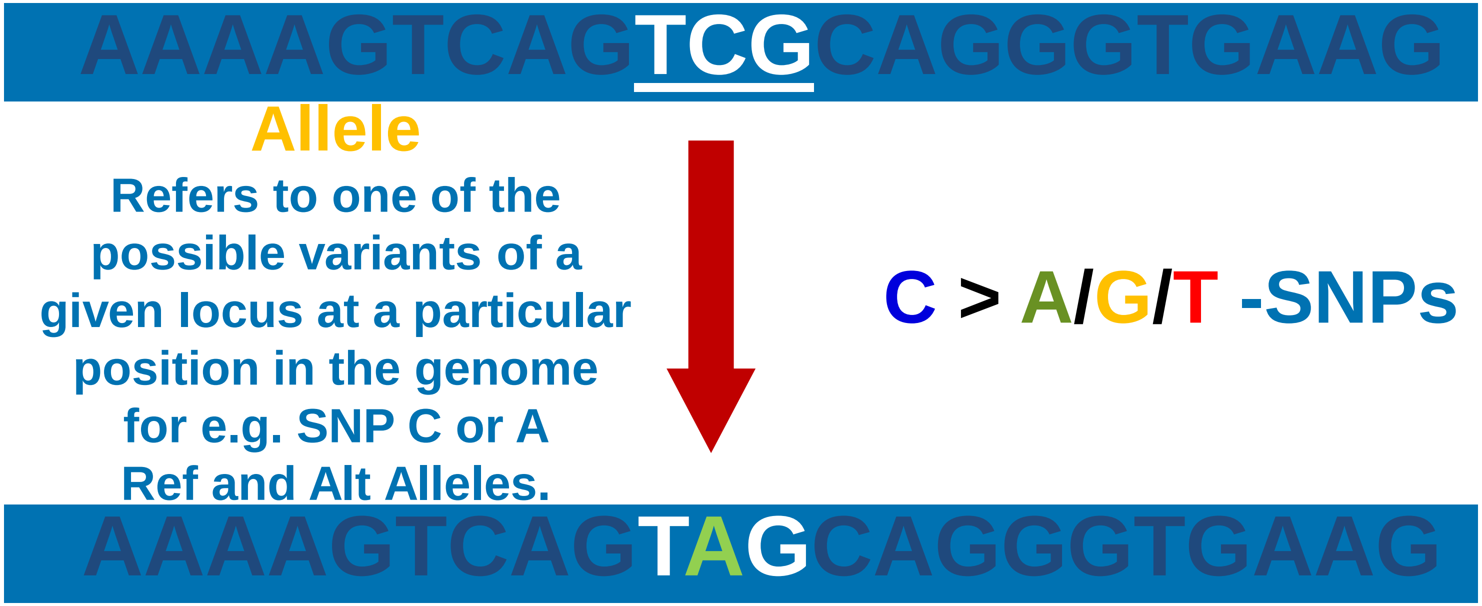
Inversion

Small Scale Variations

- Single base changes – Single Nucleotide Variants (SNVs)
  a. Substitutions (SNPs).
  b. Deletions } Indels
  c. Insertions

- Multiple base changes
  a. Multi-nucleotide polymorphisms.
  b. Insertions/deletions.
  c. Mini-satellites.
  d. Micro-satellites.

Small Nucleotide Variants (SNVs)



AAAAGTCAGTCGCAGGGTGAAG

Allele

Refers to one of the possible variants of a given locus at a particular position in the genome for e.g. SNP C or A Ref and Alt Alleles.

C > A/G/T -SNPs

AAAAGTCAGTAGCAGGGTGAAG

Homozygote = Both chromosomes have the same base at a particular position.

Heterozygote = Both chromosomes have a different base at a particular position.

Small Nucleotide Variants (SNVs)

AAAAGTCAG**TCG**CAGGGTGAAG

Allele

Refers to one of the possible variants of a given locus at a particular position in the genome.

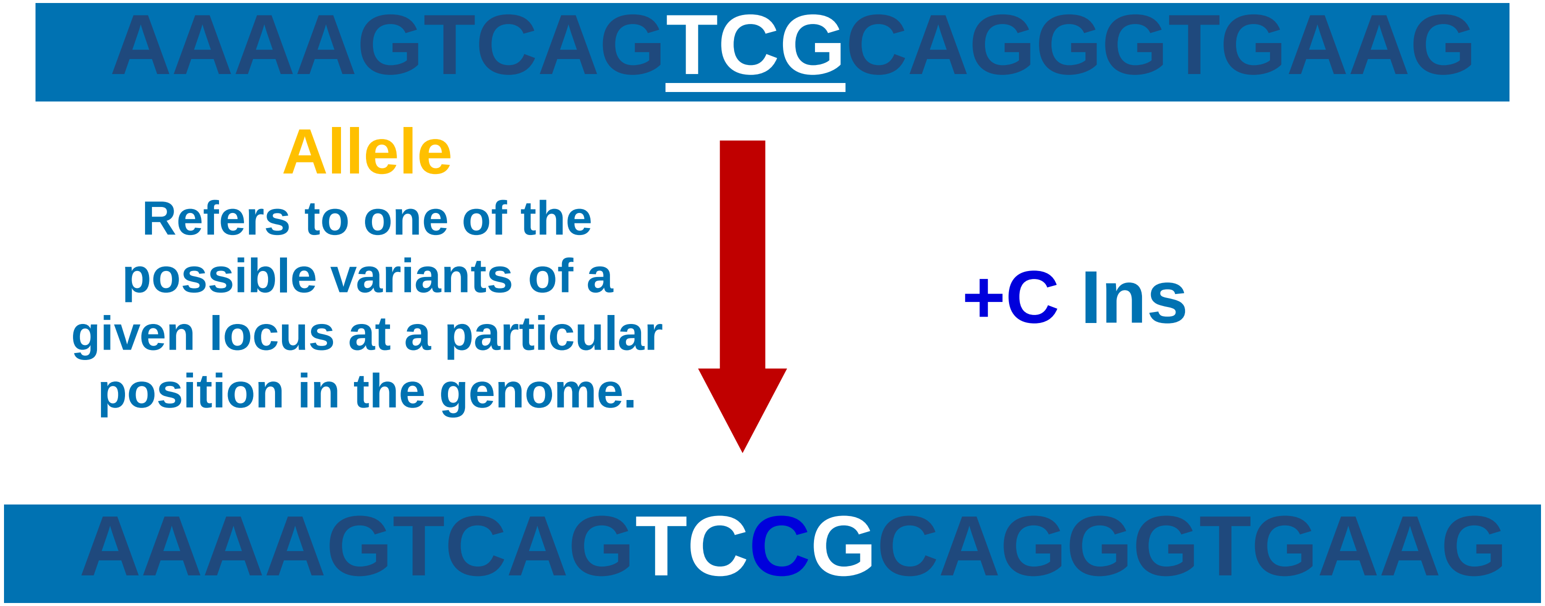
-C Del

AAAAGTCAG**TG**CAGGGTGAAG

Homozygote = Both chromosomes have the same base at a particular position.

Heterozygote = Both chromosomes have a different base at a particular position.

Small Nucleotide Variants (SNVs)



AAAAGTCAGTCGCAGGGTGAAG

Allele

Refers to one of the possible variants of a given locus at a particular position in the genome.

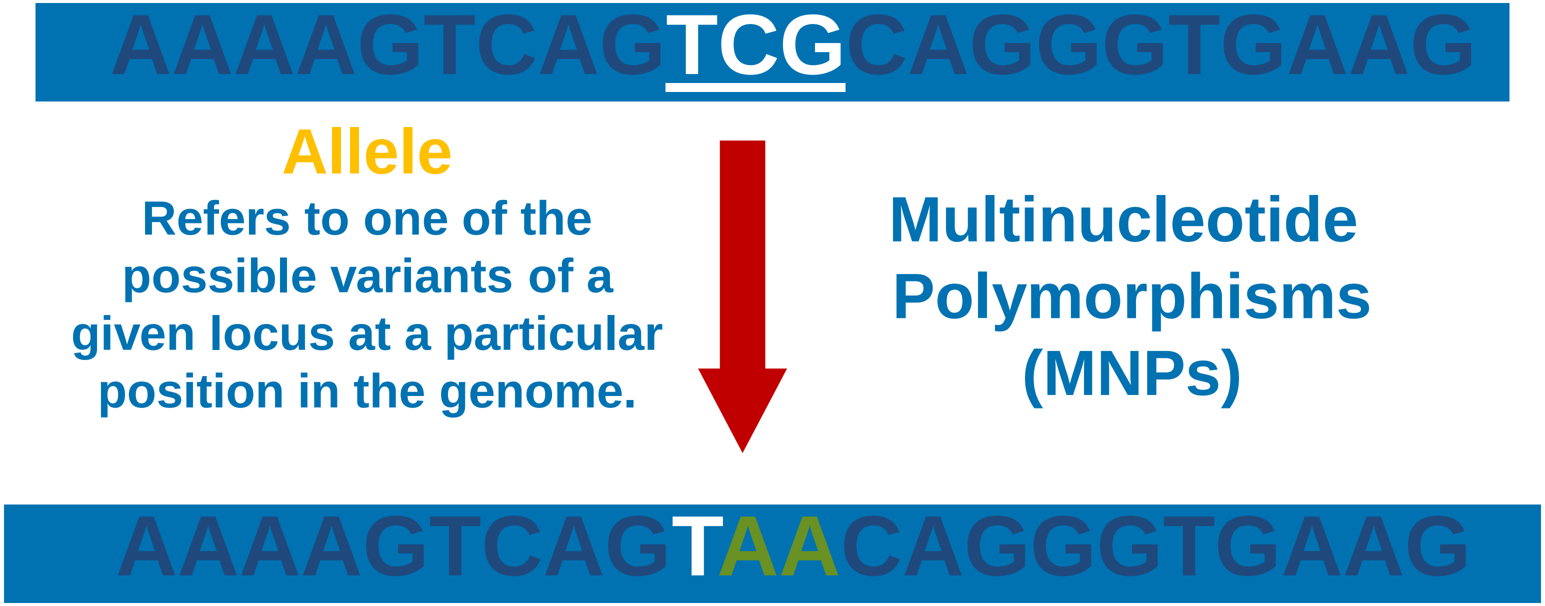
+C Ins

AAAAGTCAGTCCGCAGGGTGAAG

Homozygote = Both chromosomes have the same base at a particular position.

Heterozygote = Both chromosomes have a different base at a particular position.

Small Nucleotide Variants (SNVs)



AAAAGTCAGTCGCAGGGTGAAG

Allele

Refers to one of the possible variants of a given locus at a particular position in the genome.

Multinucleotide Polymorphisms (MNPs)

AAAAGTCAGTAACAGGGTGAAG

Homozygote = Both chromosomes have the same base at a particular position.

Heterozygote = Both chromosomes have a different base at a particular position.

Ancestral and Derived Alleles

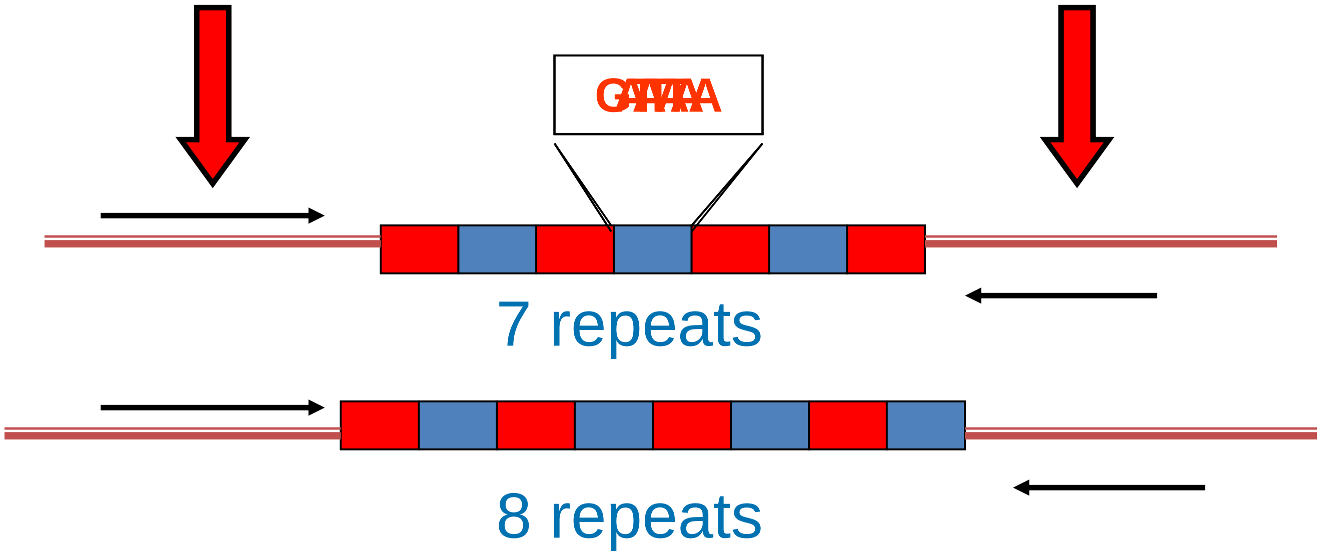
- The starting state of a variant is referred to as “ancestral”.
- The state after a mutation is “derived”.

rs80356779 G > A variant in *CPT1A*

Human	GGCCACGATCGGCGCATCTGC
Chimpanzee	GGCCACGATCGGCGCATCTGC
Gorilla	GGCCACGATCGGTGCATCTGC
Orangutan	GGCCATGATCGGCGCATCTGC
Vervet-AGM	GGCCATGATCGGCGCGTCTGC
Macaque	GGCCATGATCGGTGCGTCTGC
Olive baboon	GGCCATGATCGGTGCGTCTGC
Marmoset	GCCGACGATGGGCGCGTCTGC

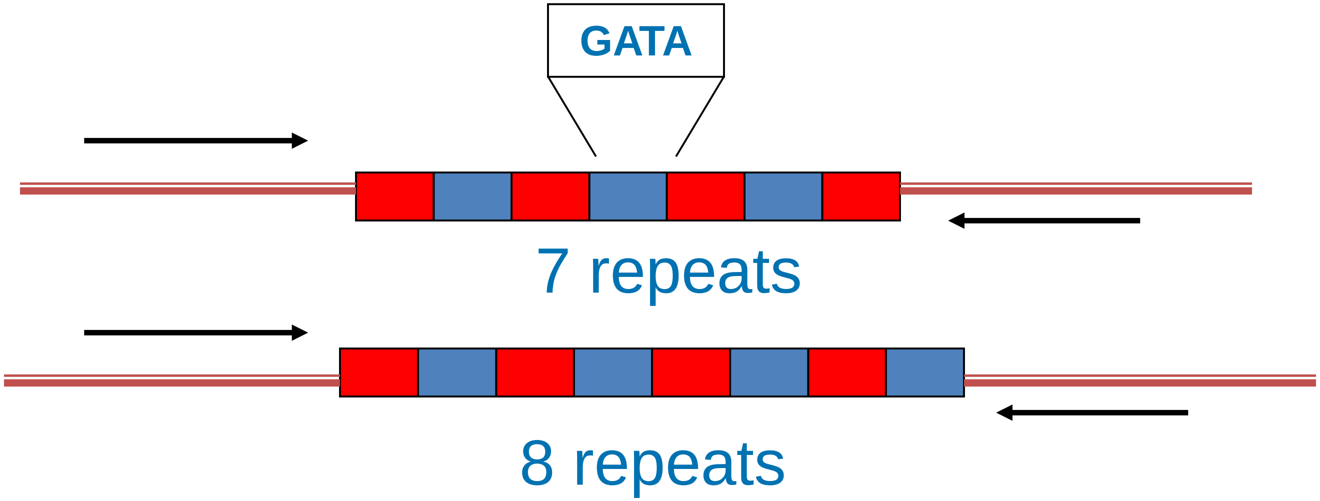
Ancestral Allele = **G**

Microsatellites or STRs



The repeat region is variable between samples while the flanking regions where PCR primers bind are constant

STRs



Tetranucleotide STR

atgccaaaatGATAGATAGATAGATAGATAGATAGATAGATAgggttttggacaatta

atgccaaaatGATAGATAGATAGATAGATAGATAGATAgggttttggaacaatta

Homozygote = Both alleles are the same length give a similar size PCR product

Heterozygote = Both alleles differ and can be resolved from one another

SNPs vs STRs

SNPs	STRs
Usually biallelic; Seldom recurrent.	Multiallelic.
Low mutation rate: 1-1.25 x 10^{-8} /bp/generation.	Higher mutation rate: 2 x 10^{-4} – 1.3 x 10^{-2} /marker/generation.
~ 3,000,000 in humans	~500,000 in humans
Ancestral state deduced from an out-group.	Difficult to deduce.

What we Ascertain?

Genetic Makeup of an Individual

- **The human genome is diploid.**

- **Genotype:**

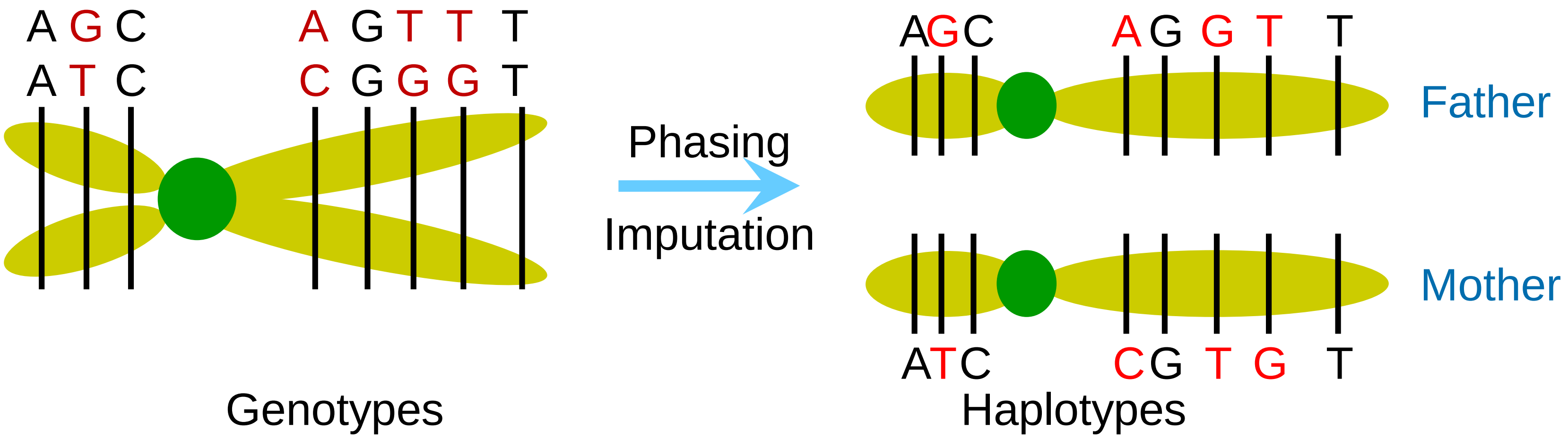
  Refers to the genetic constitution of an individual.

- **Haplotype:**

  Refers to the combination of alleles at a particular segment of a chromosome.

Genotypes and Haplotypes

- A haplotype stands for a set of linked variants on the same chromosome.
- It can be simply considered as a binary string since each SNP is binary.



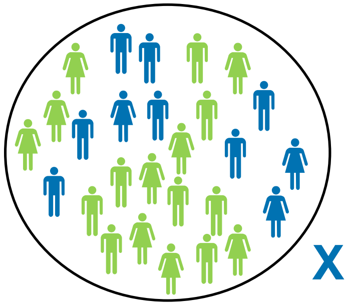
Practical Applications

- Catalog biological diversity.
- Disease diagnosis.
- Genotype-phenotype association studies.
- Pharmacogenomics.
- DNA forensics.
- Population genetics.
- Evolutionary studies.
- Marker-assisted selection.

How Do we Detect Variation?

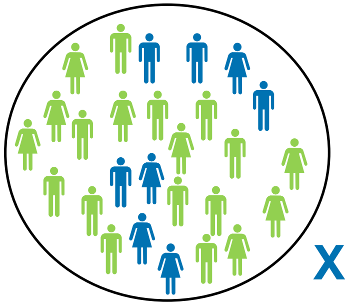
How Do We Detect Change?

Previously



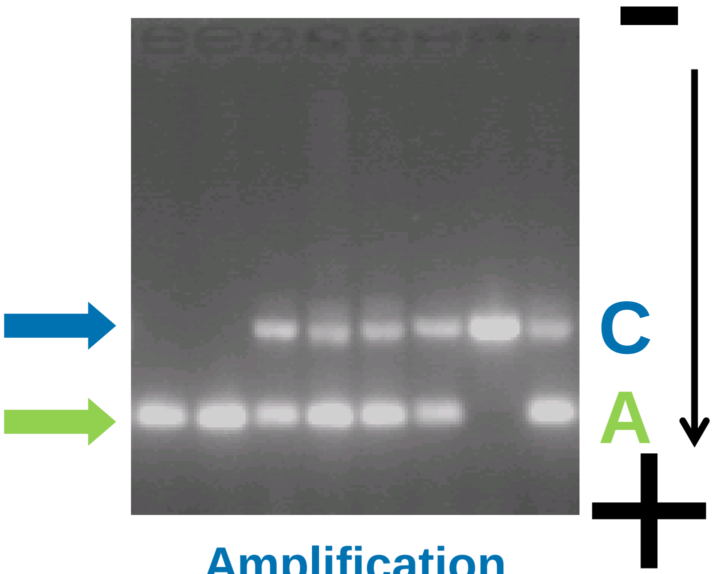
X 100's

Now



X 1,000's

Genotyping

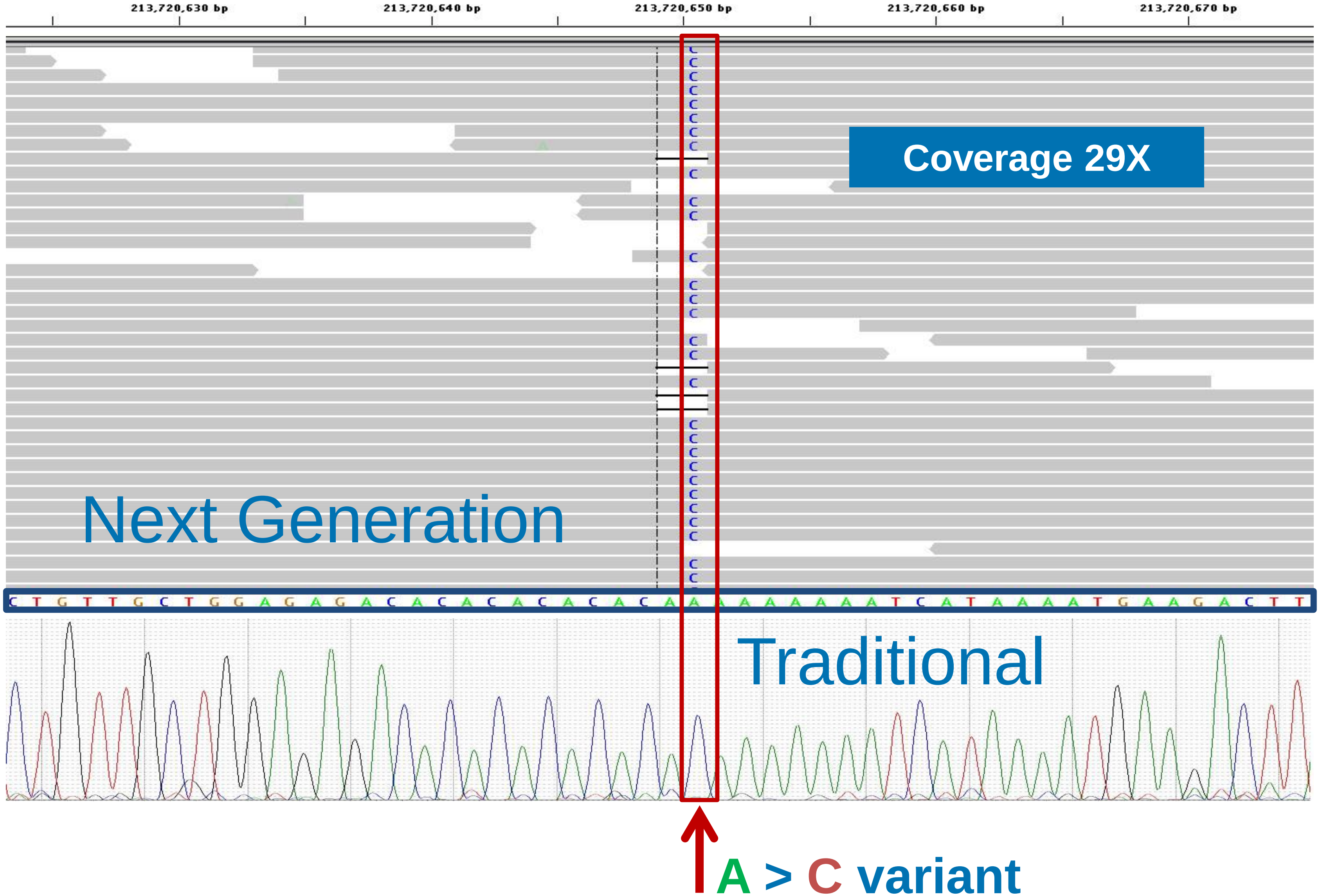


Amplification
&
Agarose Gel Electrophoresis

Affymetrix
Human SNP
Array 6.0

Omni 5M Quad
SNP_Chip

From Electropherograms to Pileups



Coverage 29X

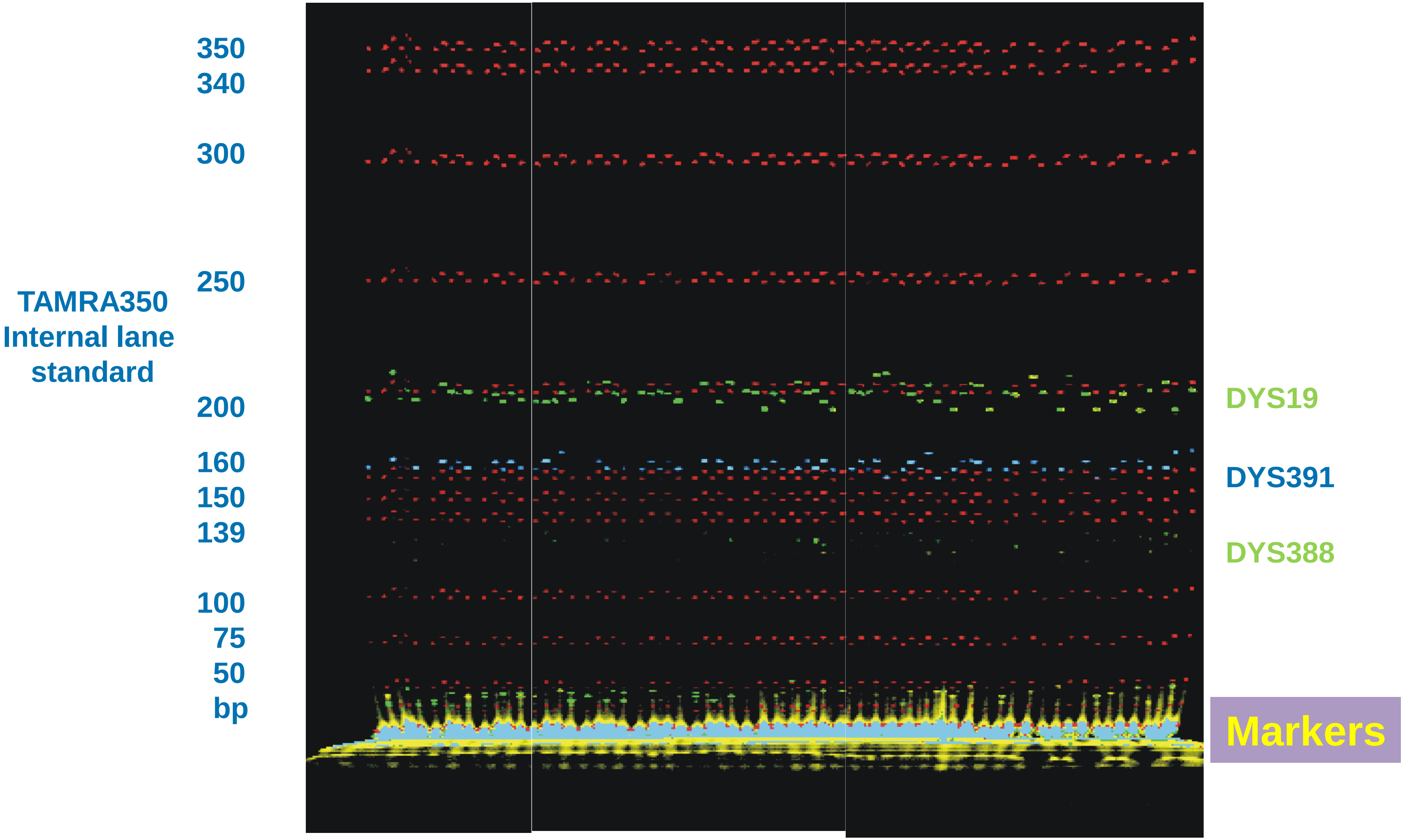
Next Generation

Traditional

A > C variant

STR Detection on Polyacrylamide Gel

ABI 377 4% Denaturing PAGE Gel Image



Variant Calling

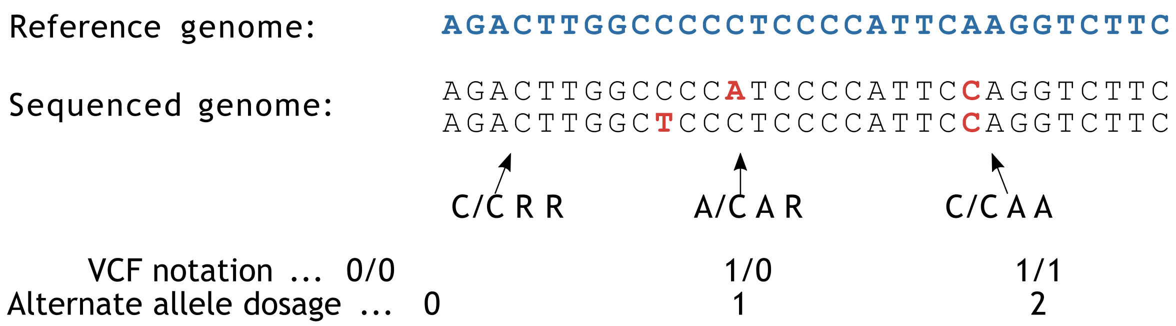
- Process of identifying changes in DNA sequences between a given, usually reference genome, and other sequenced samples.

- The goal of variant calling is to determine the genotype at each position in the genome.

- Genotype:
  - in the broad sense . . . genetic makeup of an organism.
  - in the narrow sense . . . the combination of alleles at a position.

Reference and Alternate Alleles

- Reference and alternate alleles – Ref (R) and Alt (A).
- In diploid organisms with two chromosomal copies, there are three possible genotypes:
  - RR .. homozygous reference genotype.
  - RA .. Heterozygous.
  - AA .. homozygous alternate



Reference genome: AGACTTGGCCCCCTCCCCATTCAAGGTCTTC

Sequenced genome: AGACTTGGCCCC**A**TCCCCATTC**C**AGGTCTTC
AGACTTGGC**T**CCCTCCCCATTC**C**AGGTCTTC

C/C R R

A/C A R

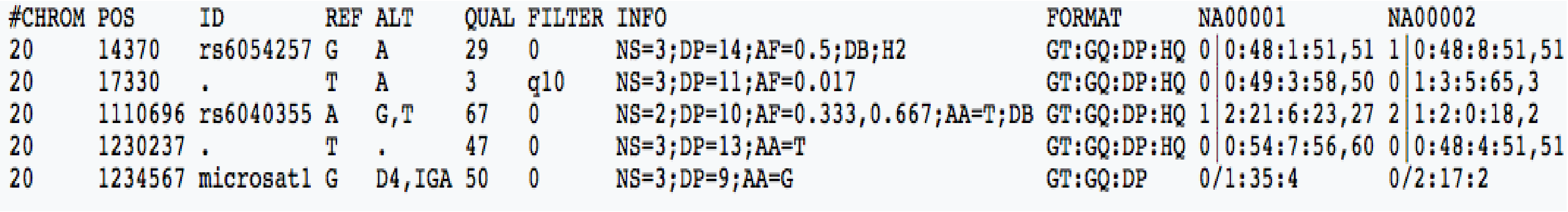
C/C A A

VCF notation ... 0/0 1/0 1/1

Alternate allele dosage ... 0 1 2

The Variant Call Format

Format: VCF



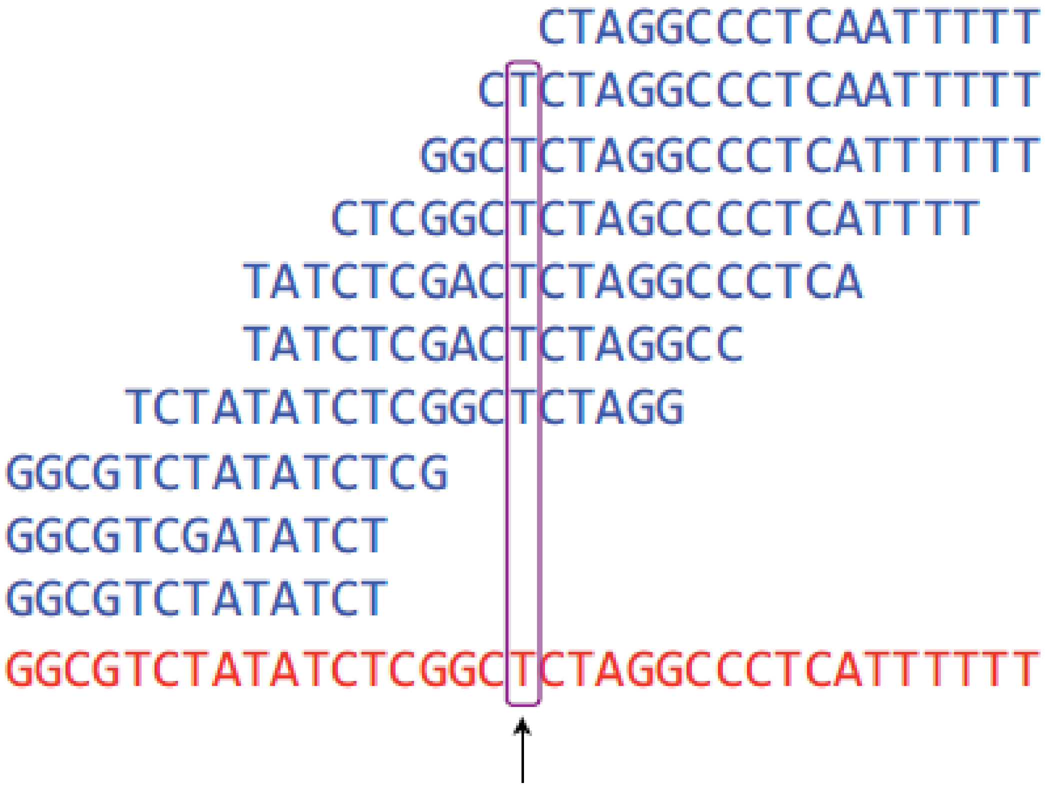
```
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002
20     14370   rs6054257 G   A      29   0      NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
20     1110696 rs6040355 A   G,T    67   0      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
20     1230237 .         T   .      47   0      NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
20     1234567 microsat1 G   D4,IGA 50   0      NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2
```

VCFTools: http://vcftools.sourceforge.net/

VcfCTools: https://github.com/AlistairNWard/vcfCTools

Allelic Depth

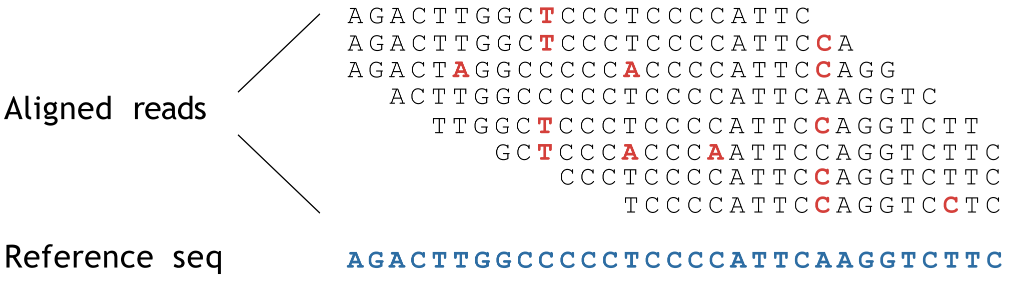
- Variant depth usually refers to average number of reads covering a particular position in the genome.



Coverage at this position = 6

Germline vs Somatic Mutations

- Germline variant calling:
  - Expect the following fractions of alternate alleles in the pileup:
    - 0.0 for RR genotype (plus sequencing errors)
    - 1.0 for AA (plus sequencing errors)
    - 0.5 for RA (random variation of binomial sampling)
- Somatic variant calling:
  - Expect any fraction of alt AF possible - subclonal variation, admixture of normal cells in the tumor samples.



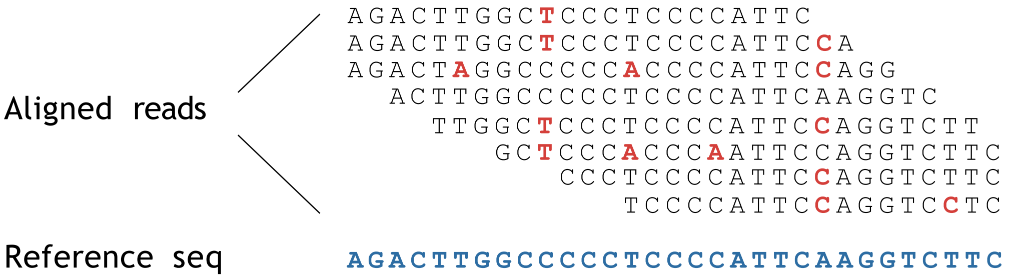
Aligned reads

```
AGACTTGGCTCCCTCCCCATTC
AGACTTGGCTCCCTCCCCATTCCA
AGACTAGGCCCCCACCCCATTCCAGG
  ACTTGGCCCCCTCCCCATTCAAGGTC
    TTGGCTCCCTCCCCATTCCAGGTCTT
       GCTCCCACCCAATTCCAGGTCTTC
          CCCTCCCCATTCCAGGTCTTC
             TCCCCATTCCAGGTCCTC
```

Reference seq

```
AGACTTGGCCCCCTCCCCATTCAAGGTCTTC
```

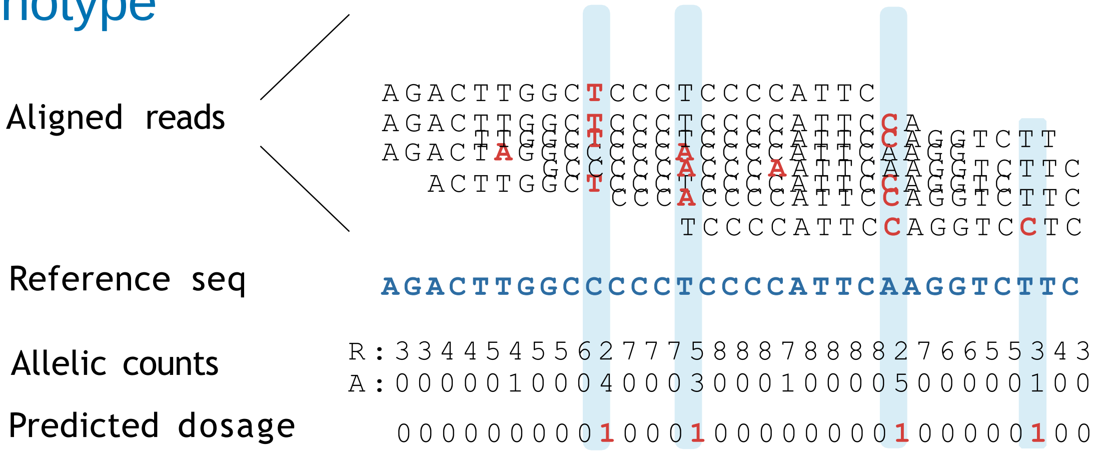
Naive Variant Calling



```
                 AGACTTGGCTCCCTCCCCATTC
                 AGACTTGGCTCCCTCCCCATTCCA
                 AGACTAGGCCCCCACCCCATTCCAGG
                   ACTTGGCCCCCTCCCCATTCAAGGTC
Aligned reads        TTGGCTCCCTCCCCATTCCAGGTCTT
                        GCTCCCACCCAATTCCAGGTCTTC
                           CCCTCCCCATTCCAGGTCTTC
                              TCCCCATTCCAGGTCCTC

Reference seq    AGACTTGGCCCCCTCCCCATTCAAGGTCTTC
```

Fixed Allele Thresholds

- Use fixed allele frequency threshold to determine the genotype



Aligned reads

```
AGACTTGGCTCCCTCCCCATTC
AGACTTGGCTCCCTCCCCATTCCA
     TTGGCTCCCTCCCCATTCCAGGTCTT
AGACTAGGCCCCCACCCCATTCAAGG
       GCCCCCACCCAATTCAAGGTCTTC
  ACTTGGCTCCCTCCCCATTCCAGGTC
          CCCACCCCATTCCAGGTCTTC
             TCCCCATTCCAGGTCCTC
```

Reference seq

```
AGACTTGGCCCCCTCCCCATTCAAGGTCTTC
```

Allelic counts

```
R:3344545562777588878888276655343
A:0000010004000300010000500000100
```

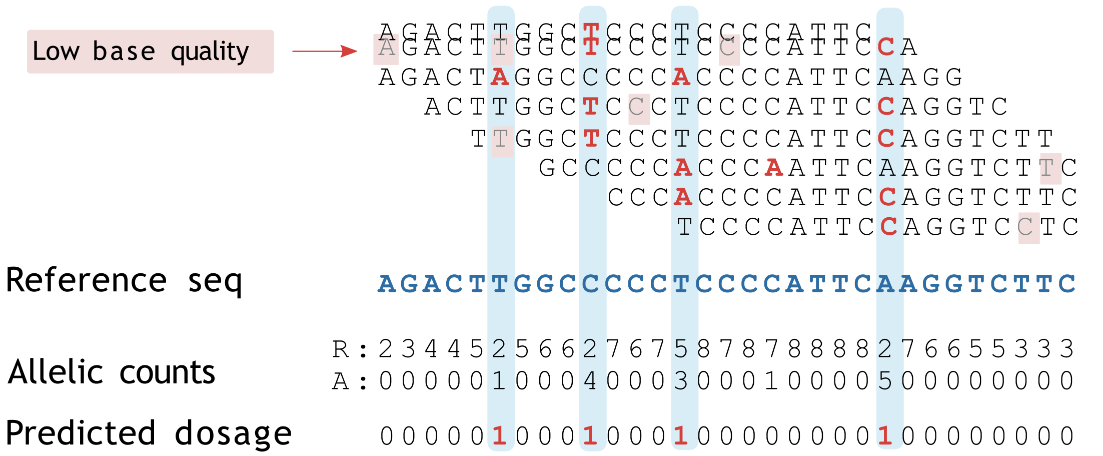
Predicted dosage

```
  0000000001000100000000100000100
```

alt AF	genotype
[0.0, 0.2)	RR .. homozygous reference
[0.2, 0.8]	RA .. herezogyous
(0.8, 1.0]	AA .. homozygous variant

Base Quality Filtering

- Filter out low quality bases before calling genotypes.



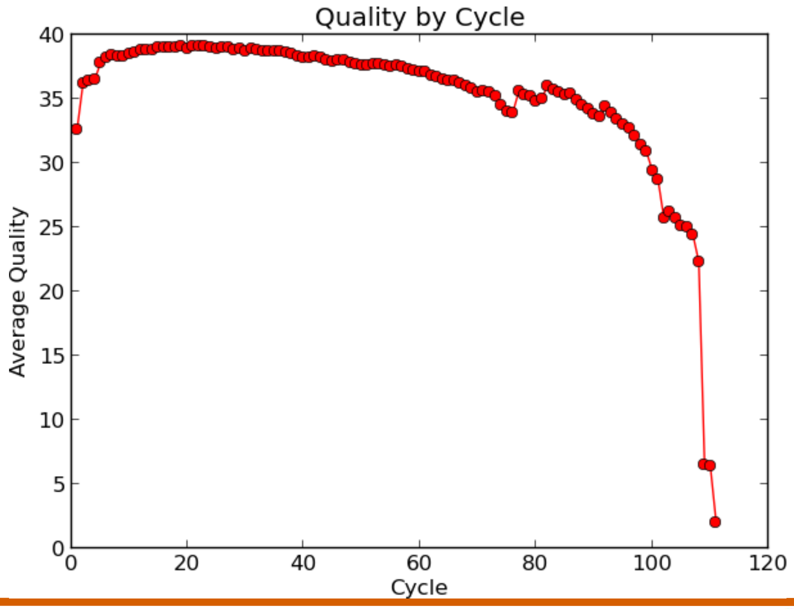
```
Low base quality  →    AGACTTGGCTCCCTCCCCATTC
                       AGACTTGGCTCCCTCCCCATTCCA
                       AGACTAGGCCCCCACCCCATTCAAGG
                         ACTTGGCTCCCTCCCCATTCCAGGTC
                           TTGGCTCCCTCCCCATTCCAGGTCTT
                              GCCCCCACCCAATTCAAGGTCTTC
                                 CCCACCCCATTCCAGGTCTTC
                                    TCCCCATTCCAGGTCCTC

Reference seq          AGACTTGGCCCCCTCCCCATTCAAGGTCTTC

                     R:2344525662767587878882766553333
Allelic counts       A:0000010004000300010000500000000

Predicted dosage       0000010001000100000000100000000
```

- Filter base calls by quality
  - Ignore bases Q<20

Phred quality score

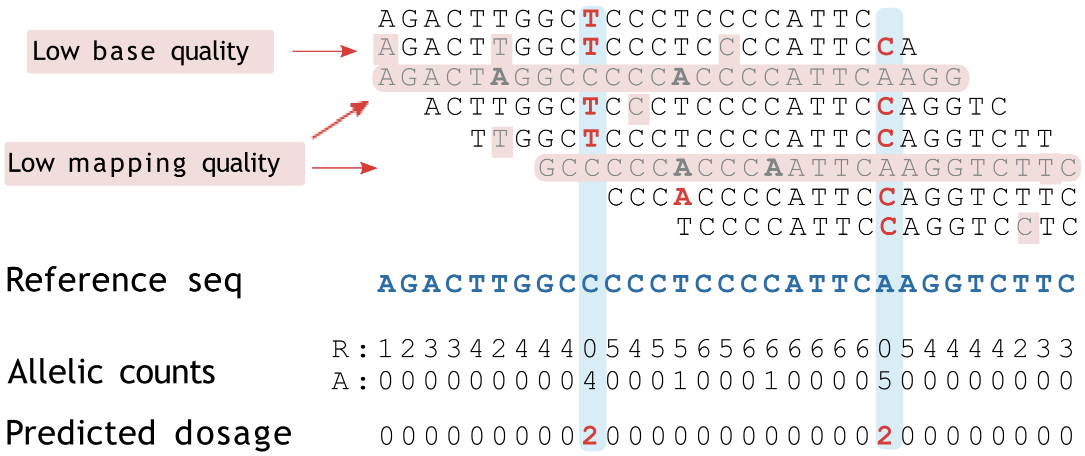
$Q = -10 \log_{10} P_{err}$

Quality	Error probability	Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%



Filtering Variants by Quality

Use fixed allele frequency threshold to determine the genotype



Low base quality

Low mapping quality

```
      AGACTTGGCTCCCTCCCCATTC
      AGACTTGGCTCCCTCCCCATTCCA
      AGACTAGGCCCCCACCCCATTCAAGG
        ACTTGGCTCCCTCCCCATTCCAGGTC
          TTGGCTCCCTCCCCATTCCAGGTCTT
             GCCCCCACCCAATTCAAGGTCTTC
                CCCACCCCATTCCAGGTCTTC
                   TCCCCATTCCAGGTCCTC
```

Reference seq

```
      AGACTTGGCCCCCTCCCCATTCAAGGTCTTC
```

Allelic counts

```
R:    1233424440545565666666054444233
A:    0000000004000100010000500000000
```

Predicted dosage

```
      0000000002000000000000200000000
```

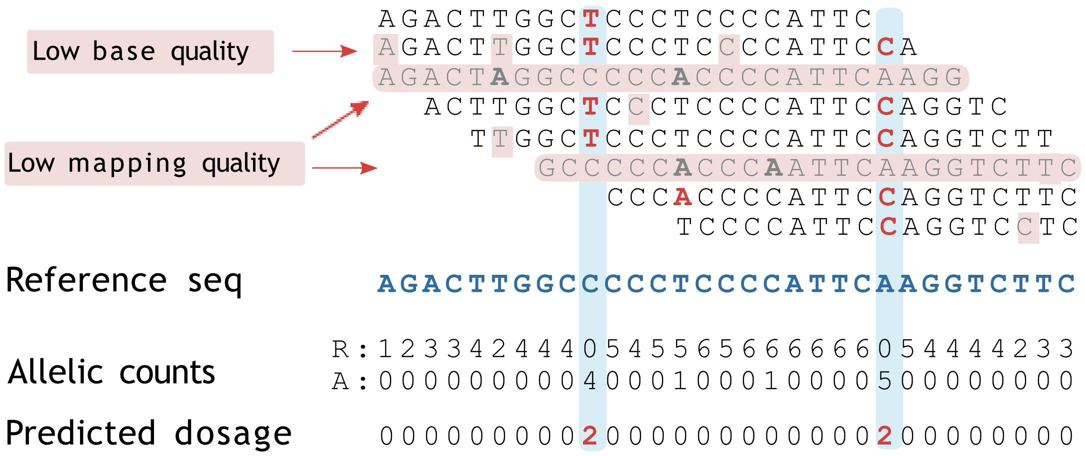
1) Filter base calls by quality.
For e.g. ignore bases Q<20.

alt AF	genotype
[0.0, 0.2)	RR .. homozygous reference
[0.2, 0.8]	RA .. herezogyous
(0.8, 1.0]	AA .. homozygous variant

2) Filter reads with low mapping quality.

Issues with Naïve Variant Calling

Use fixed allele frequency threshold to determine the genotype



1) Filter base calls by base quality
e.g. ignore bases Q<20

2) Filter reads with low mapping quality

alt AF	genotype
[0.0, 0.2)	RR .. homozygous reference
[0.2, 0.8]	RA .. herezogyous
(0.8, 1.0]	AA .. homozygous variant

Problems:

- Undercalls heterozygotes in low-coverage data.
- Throws away information due to hard quality thresholds.
- Gives no measure of confidence.

Variant Calling Models

More sophisticated models apply a statistical framework

Likelihood

Prior

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

Posterior

Normalization

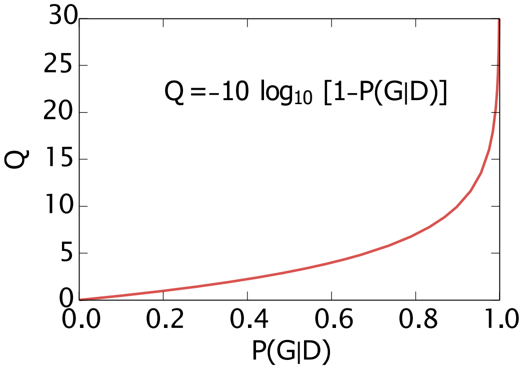
To determine:

1. the most likely genotype $g \in \{RR, RA, AA\}$ given the observed data D

$$g = \underset{G}{\operatorname{argmax}}\, P(G|D)$$

2. and the genotype quality

$$Q = -10 \log_{10}[1 - P(G|D)]$$



Genotype Likelihoods

Genotype likelihoods

- which of the three genotypes RR, RA, AA is the data most consistent with?
- calculated from the alignments, the basis for calling
- takes into account:
  - base calling errors
  - mapping errors
  - statistical fluctuations of random sampling
  - local indel realignment (base alignment quality, BAQ)

Prior probability

- how likely it is to encounter a variant base in the genome?
- some assumptions are made
  - allele frequencies are in Hardy-Weinberg equilibrium

    $P(RA) = 2f(1-f)$, $P(RR) = (1-f)^2$, $P(AA) = f^2$
- can take into account genetic diversity in a population

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

Variant Calling Example

Inputs

- alignment file
- reference sequence

Outputs

- VCF or BCF file

Example

```
bcftools mpileup -f ref.fa aln.bam | bcftools call -mv
```

Tips

`bcftools mpileup`

- increase/decrease the required number (`-m`) and the fraction (`-F`) of supporting reads for indel calling
- the `-Q` option controls the minimum required base quality (30)
- BAQ realignment is applied by default and can be disabled with `-B`
- streaming the uncompressed binary BCF (`-Ou`) is much faster than the default text VCF

`bcftools call`

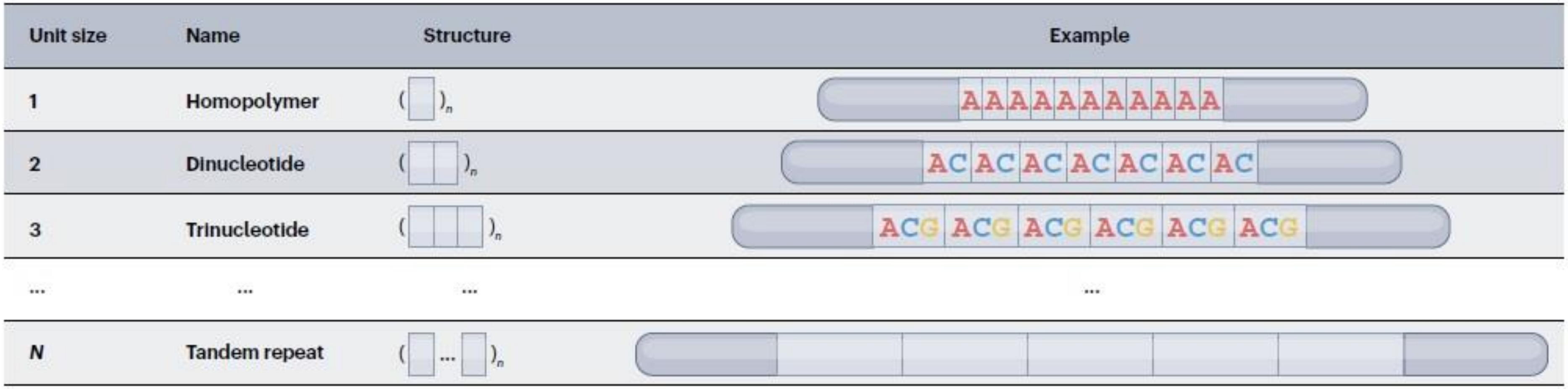
- decrease/increase the prior probability (`-P`) to decrease/increase sensitivity

General advice

- take time to understand the options
- play with the parameters, see how the calls change

Errors in Variant Calling

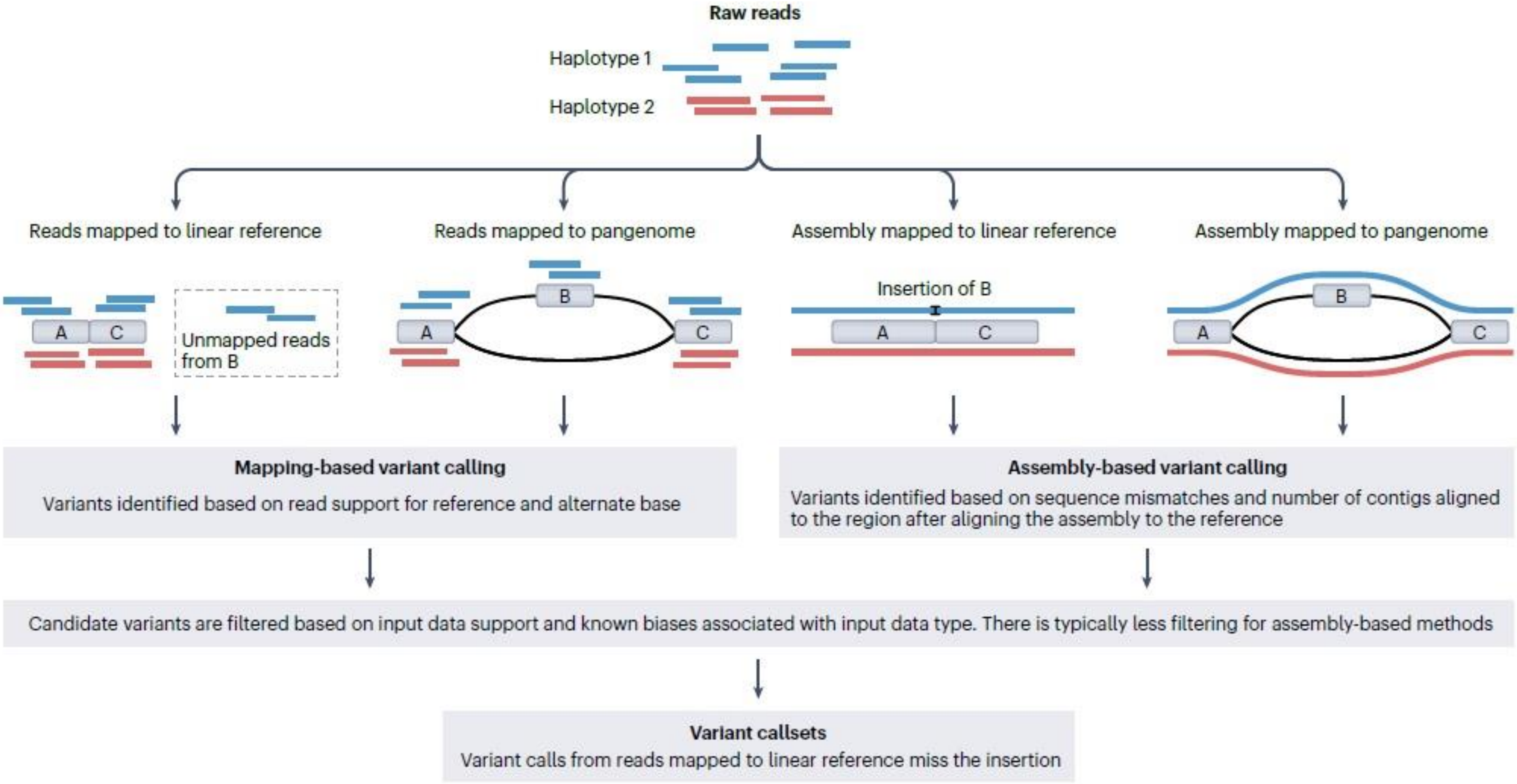
- Homopolymers and tandem repeats in sequences are common sources of systematic sequencing and mapping errors.



Unit size	Name	Structure	Example
1	Homopolymer	(□)$_n$	AAAAAAAAAAA
2	Dinucleotide	(□□)$_n$	AC AC AC AC AC AC AC
3	Trinucleotide	(□□□)$_n$	ACG ACG ACG ACG ACG ACG
...	...	...	...
N	Tandem repeat	(□ ... □)$_n$	

Olson *et. al.* 2023 Nature Reviews Genetics 24:464–483.

Variant Calling Workflows


Raw reads
Haplotype 1
Haplotype 2
Reads mapped to linear reference
Reads mapped to pangenome
Assembly mapped to linear reference
Assembly mapped to pangenome
A
C
Unmapped reads from B
A
B
C
Insertion of B
A
C
A
B
C

Mapping-based variant calling

Variants identified based on read support for reference and alternate base

Assembly-based variant calling

Variants identified based on sequence mismatches and number of contigs aligned to the region after aligning the assembly to the reference

Candidate variants are filtered based on input data support and known biases associated with input data type. There is typically less filtering for assembly-based methods

Variant callsets

Variant calls from reads mapped to linear reference miss the insertion

Olson *et. al.* 2023 Nature Reviews Genetics 24:464–483.

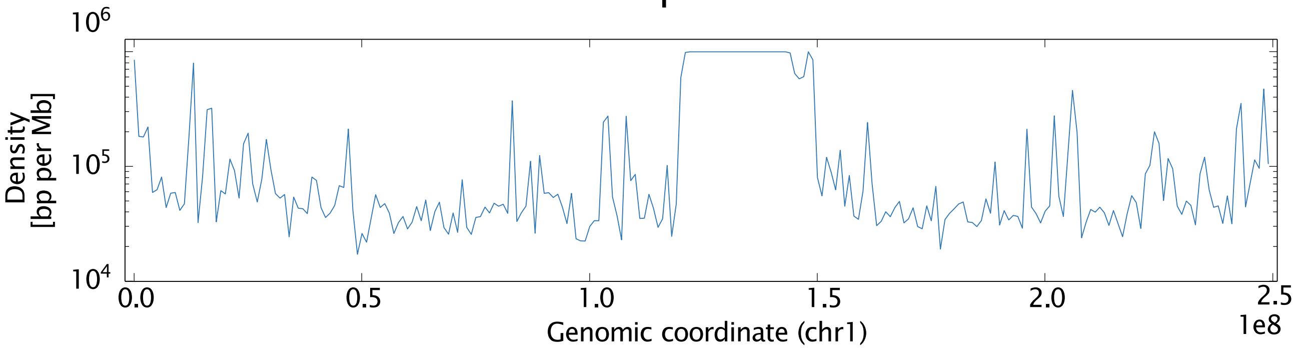
Factors to Consider in Variant Calling

- Many calls are not real, a filtering step is always necessary.
- False calls can have many causes:
  - Contamination.
  - PCR errors.
  - Sequencing errors:
    - homopolymer runs.
  - Mapping errors:
    - Repetitive sequence.
    - Structural variation.
  - Alignment errors:
    - False SNPs in proximity of indels.
    - Ambiguous indel alignment.

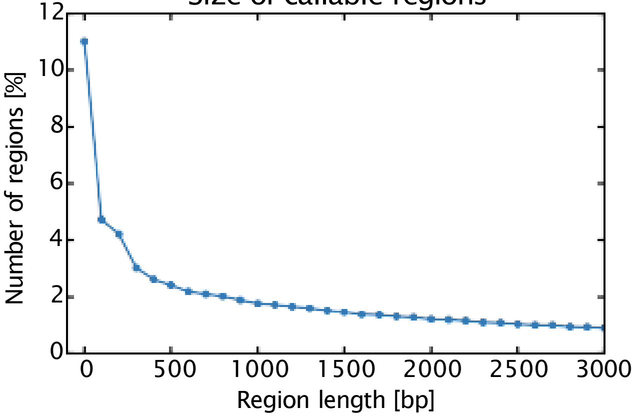
Callable Genome

- Large parts of the genome are still inaccessible.
- The Genome in a Bottle high-confidence regions:
  - Covers 89% of the reference genome.
  - Are short intervals scattered across the genome.

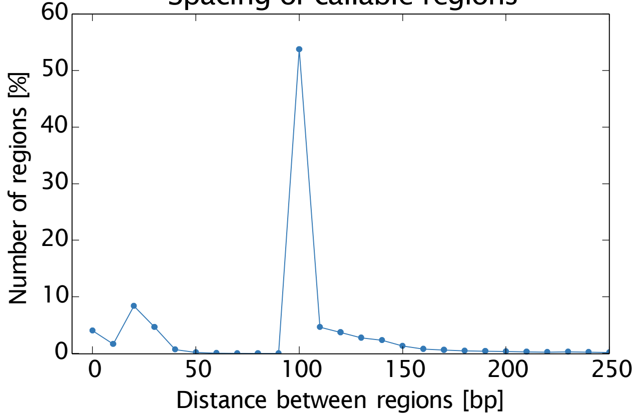
Uncallable positions



Size of callable regions



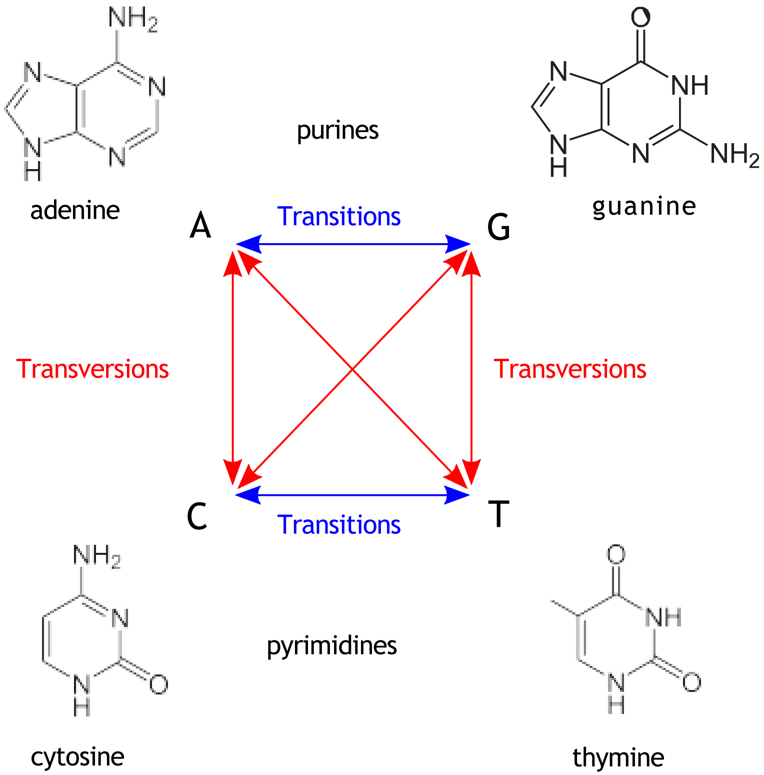
Spacing of callable regions

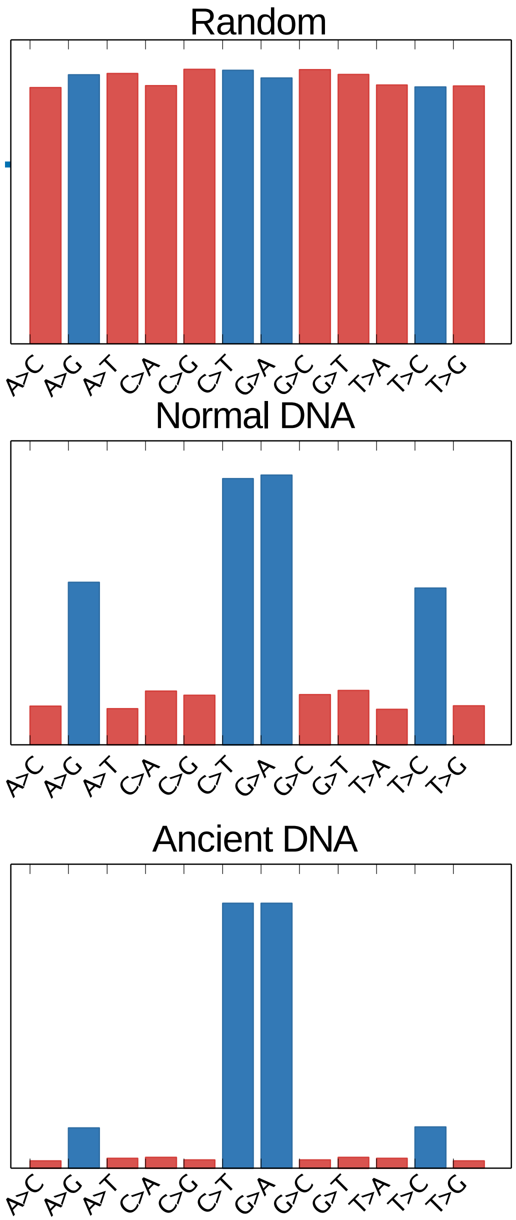


If possible, include only "nice" regions: for many analyses (e.g. population genetics studies) difficult regions can be ignored

Estimating the Quality of Called SNPs?

- Transitions vs transversions ratio, known as Ts/Tv.
- Transitions are 2-3 × more likely than transversions.





Indel Calling Challenges

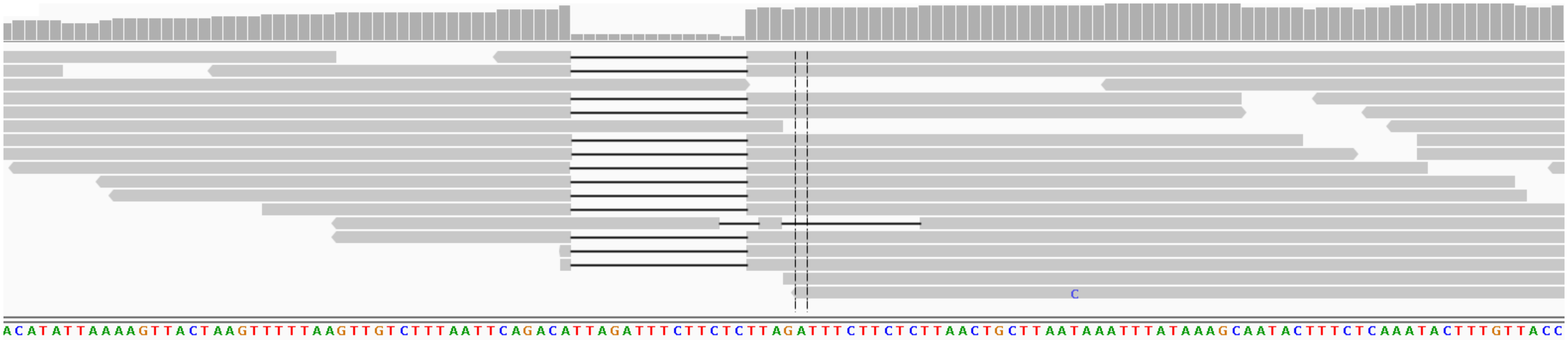
The sequencing error rate is elevated in microsatellites

Low reproducibility across callers

- 37.1% agreement between HapCaller, SOAPindel and Scalpel
  Narzisi et al. (2014) Nat Methods, 11(10):1033

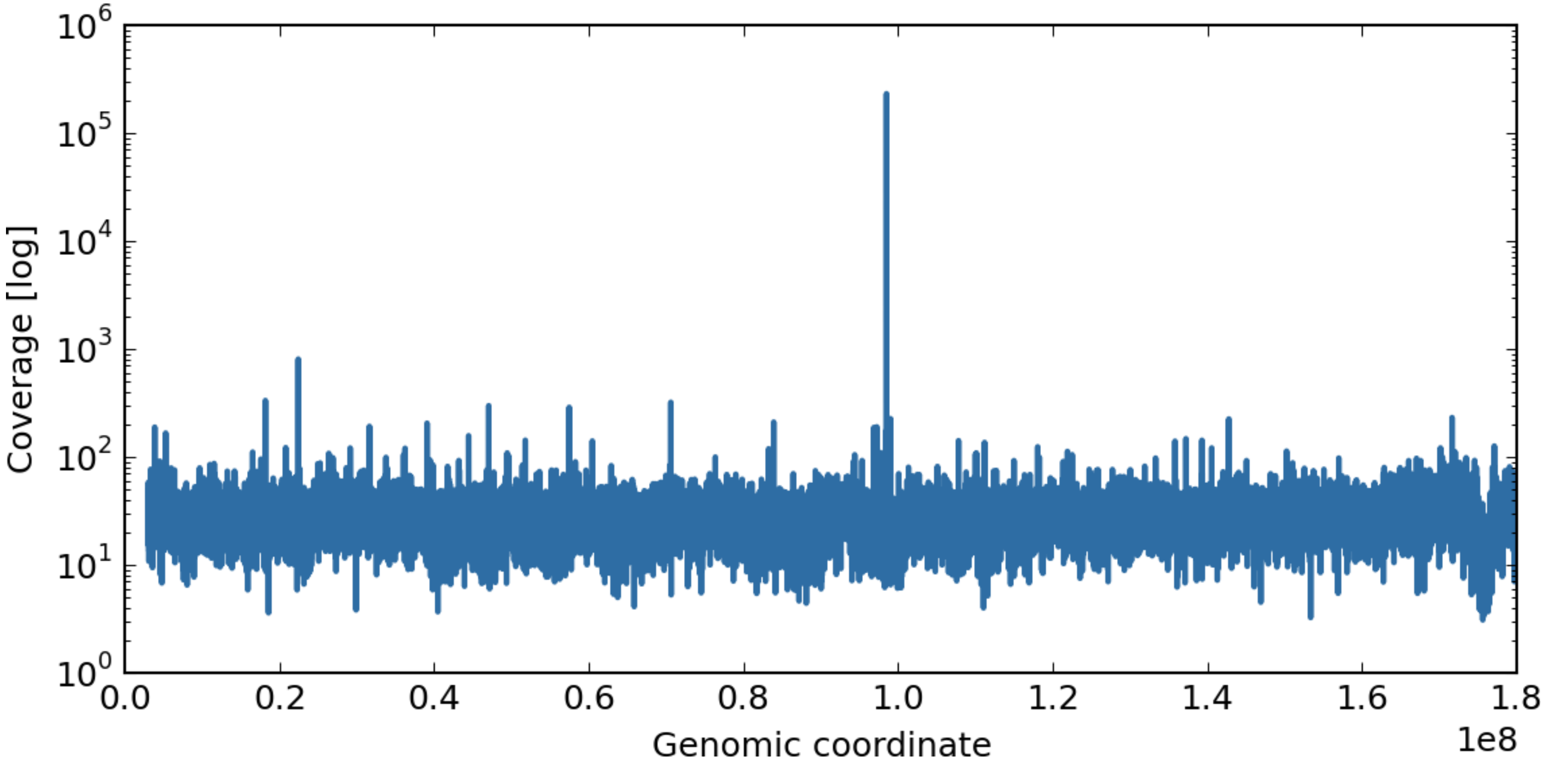
Reads with indels are more difficult to map and align

- the aligner can prefer multiple mismatches rather than a gap
- indel representation can be ambiguous


C
ACATATTAAAAGTTACTAAGTTTTTAAGTTGTCTTTAATTCAGACATTAGATTTCTTCTCTTAGATTTCTTCTCTTAACTGCTTAATAAATTTATAAAGCAATACTTTCTCAAATACTTTGTTACC

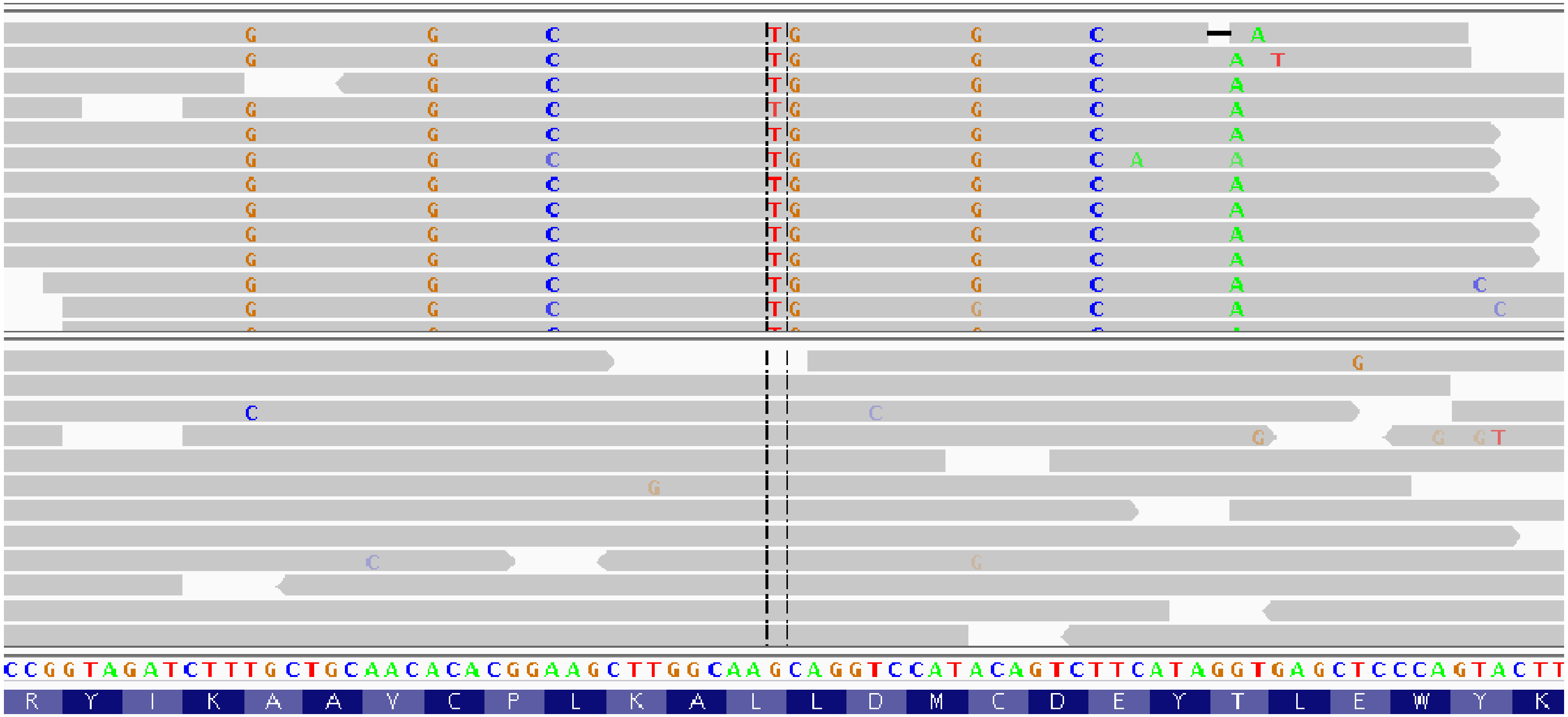
```
CTTTAATTCAGACATTAGATTTCTTCTC
CTTTAATTCAGACATTAGATTTCTTCTCTTA
CTTTAATTCAGACA--------------TTAGATTTCTTCTCTTAACTGCTT
CTTTAATTCAGACATTAGATTTCTTC---TA-----------TTAACTGCTT

CTTTAATTCAGACATTAGATTTCTTCTCTTAGATTTCTTCTCTTAACTGCTT
```

Maximum Depth



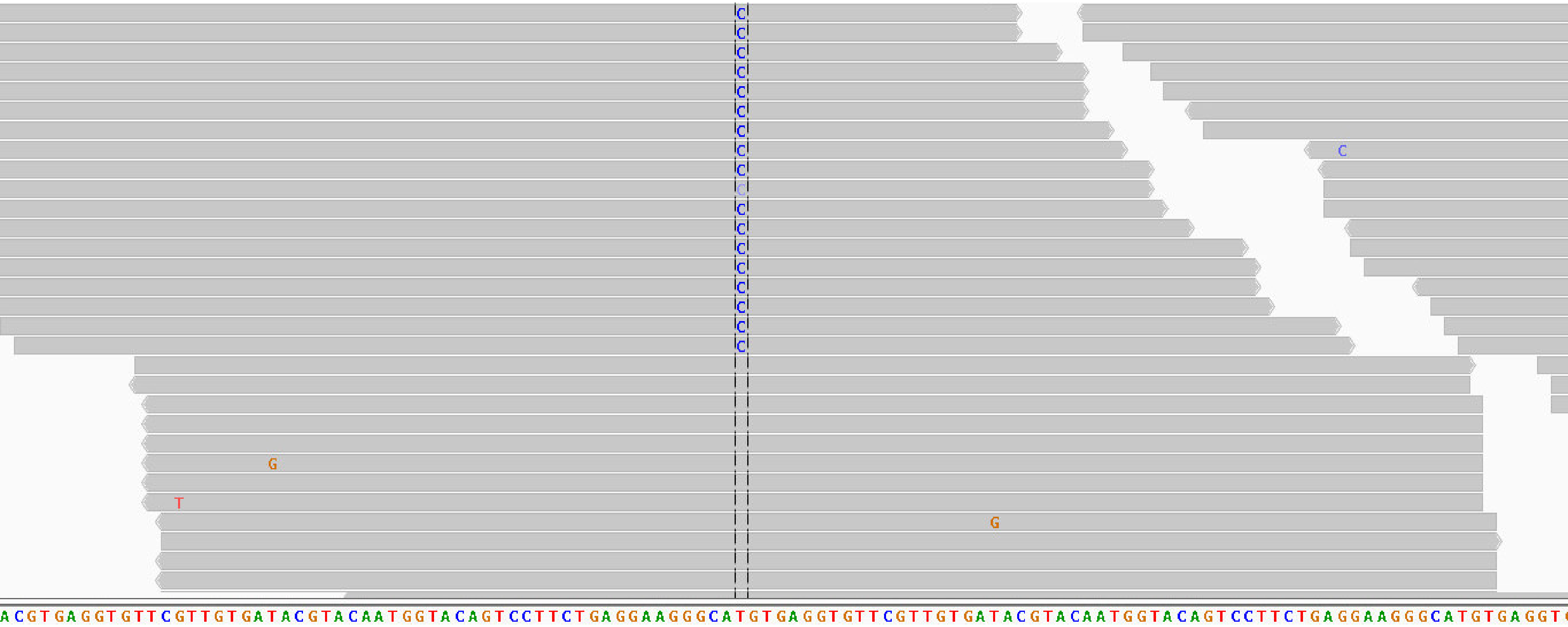
Q: Why is the sequencing depth thousandfold the average in some regions?

Mapping Errors


CCGGTAGATCTTTGCTGCAACACACGGAAGCTTGGCAAGCAGGTCCATACAGTCTTCATAGGTGAGCTCCCAGTACTT
R Y I K A A V C P L K A L L D M C D E Y T L E W Y K

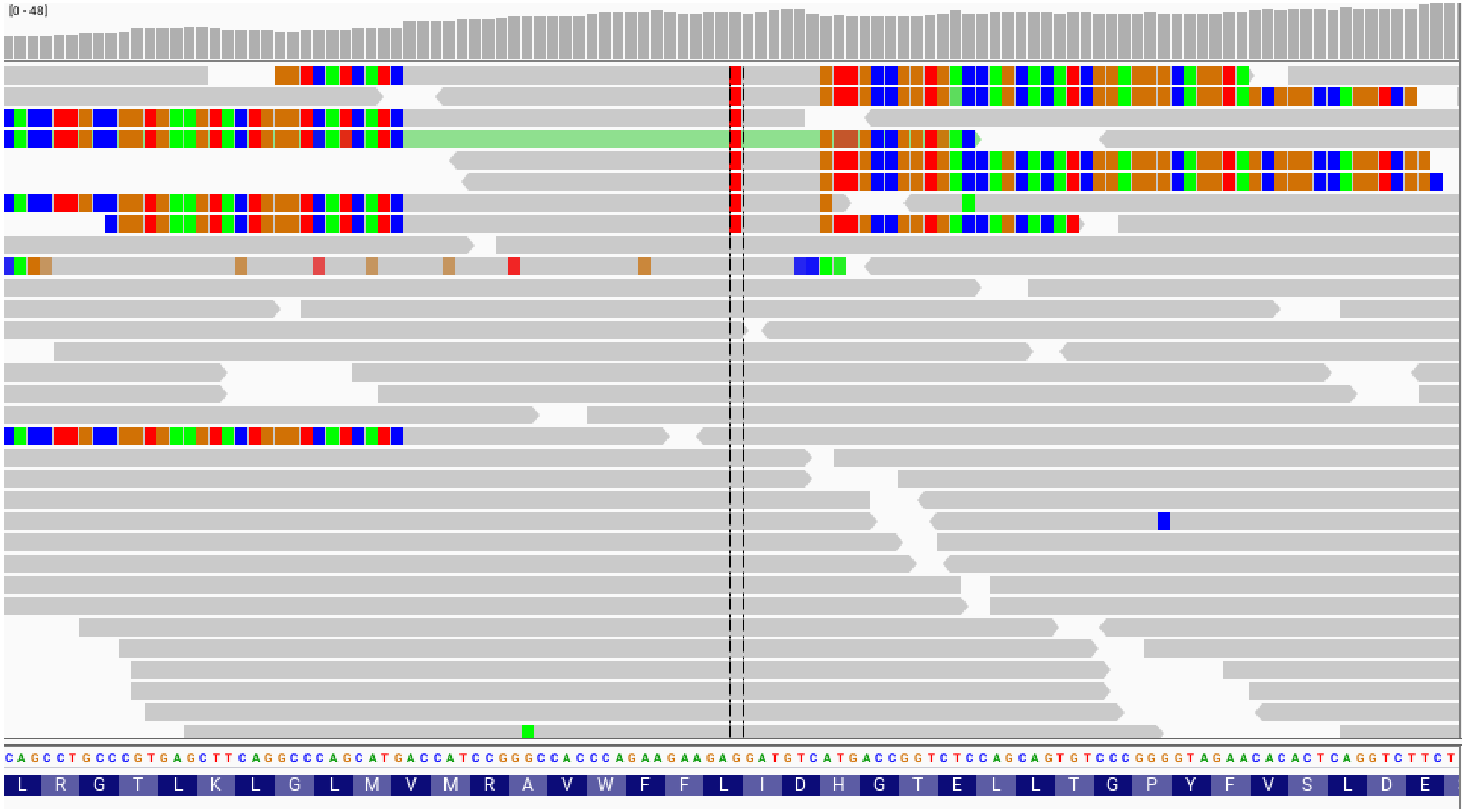
Q: RNA-Seq (top panel) and DNA sequencing data (bottom panel) from the same sample has been mapped onto the reference genome. Can you explain the novel SNVs?

Strand Bias


C
C
C
C
C
C
C
C
C
C
C
C
C
C
C
C
C
C
C
G
T
G
ACGTGAGGTGTTCGTTGTGATACGTACAATGGTACAGTCCTTCTGAGGAAGGGCATGTGAGGTGTTCGTTGTGATACGTACAATGGTACAGTCCTTCTGAGGAAGGGCATGTGAGGT

Q: Is this a valid call?

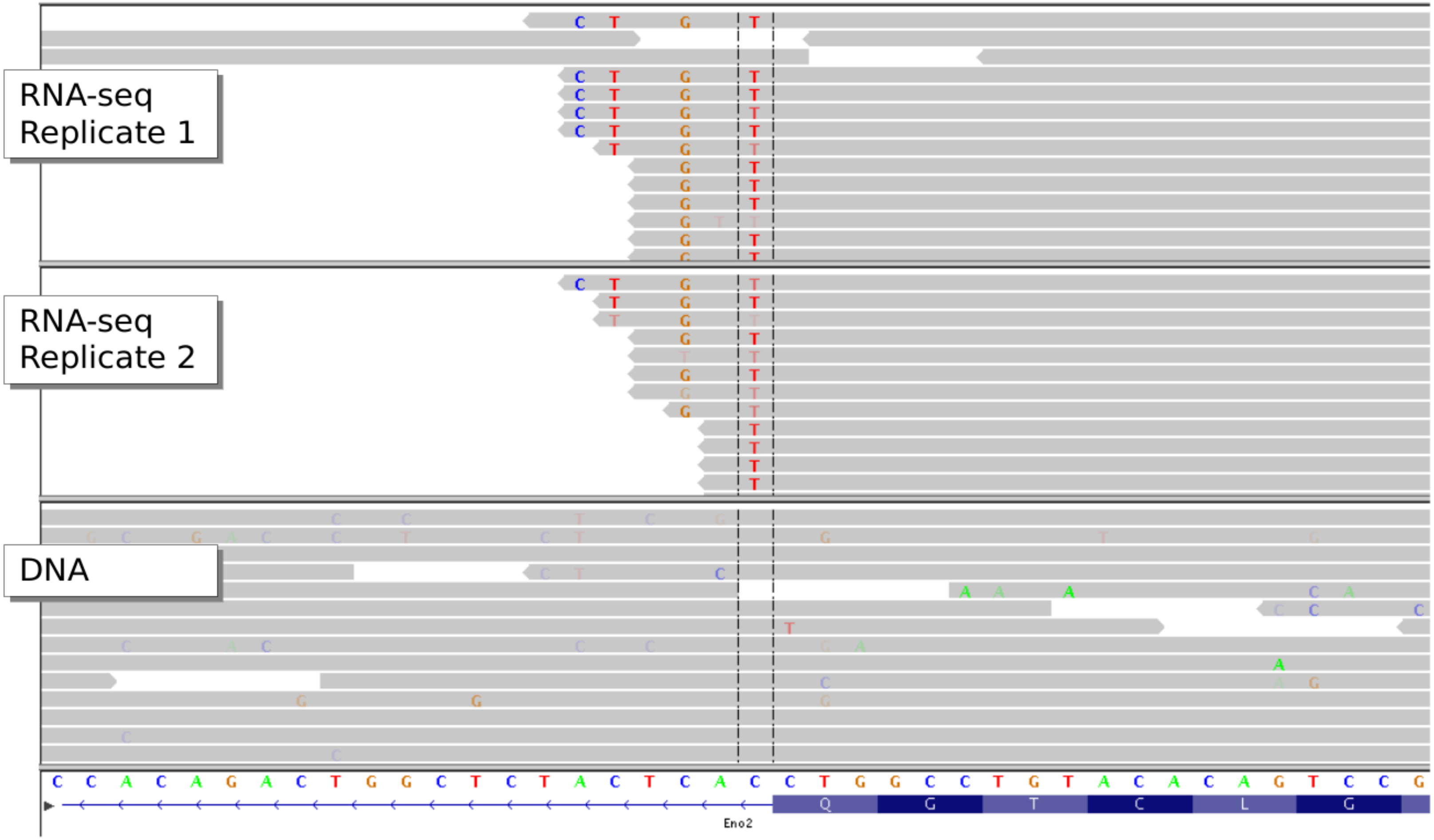
View in IGV to Reveal Artefacts


[0 - 48]
CAGCCTGCCCGTGAGCTTCAGGCCCAGCATGACCATCCGGGCCACCCAGAAGAAGAGGATGTCATGACCGGTCTCCAGCAGTGTCCCGGGGTAGAACACACTCAGGTCTTCT
L R G T L K L G L M V M R A V W F F L I D H G T E L L T G P Y F V S L D E

Display soft-clipped bases...

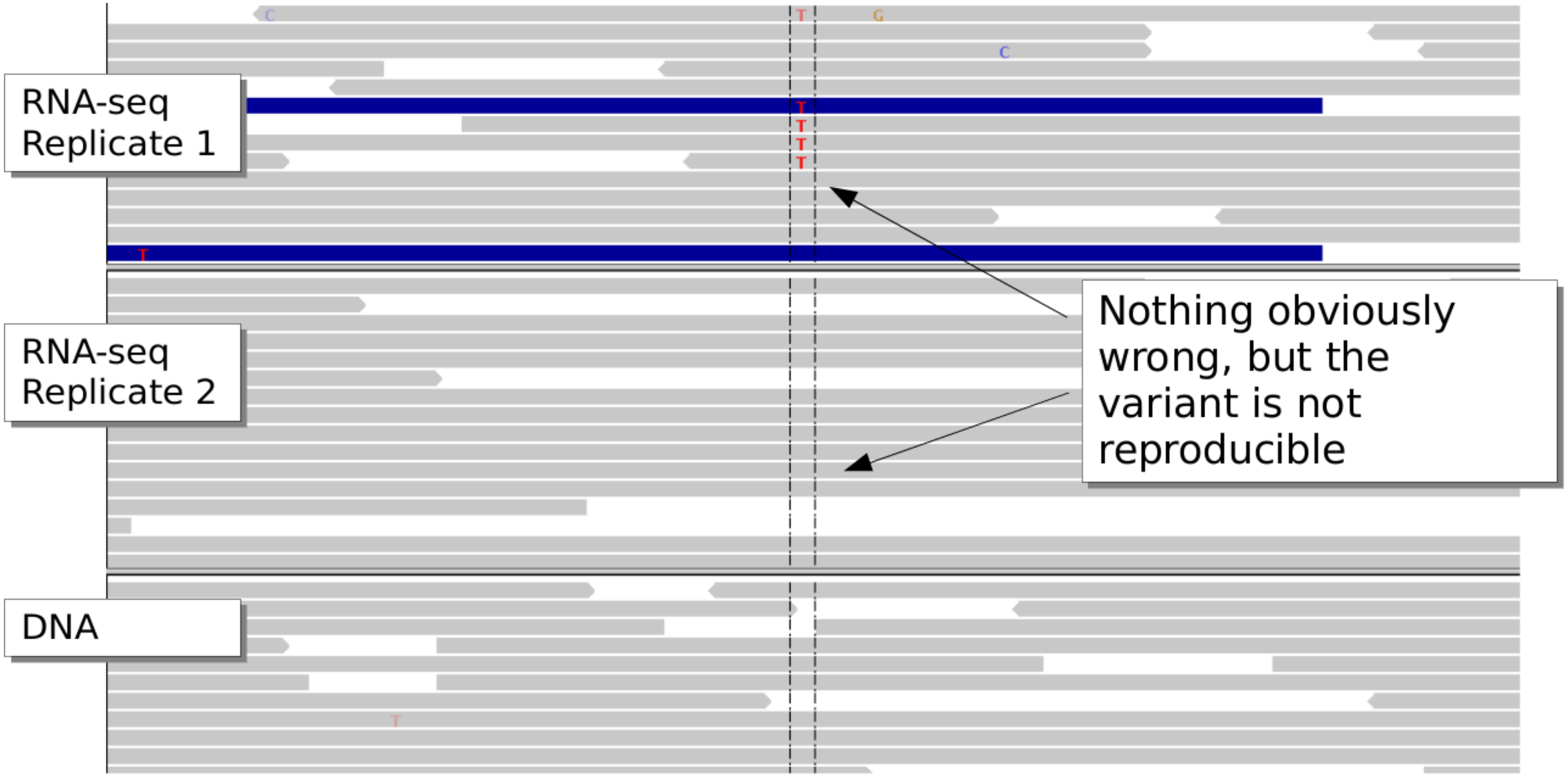
Too many soft-clipped reads in a region suggest mapping errors, be- ware!

Variant Distance Bias


RNA-seq Replicate 1
RNA-seq Replicate 2
DNA
C C A C A G A C T G G C T C T A C T C A C C T G G C C T G T A C A C A G T C C G
Q G T C L G
Eno2

Q: Can you explain what happened here?

Reproducibility

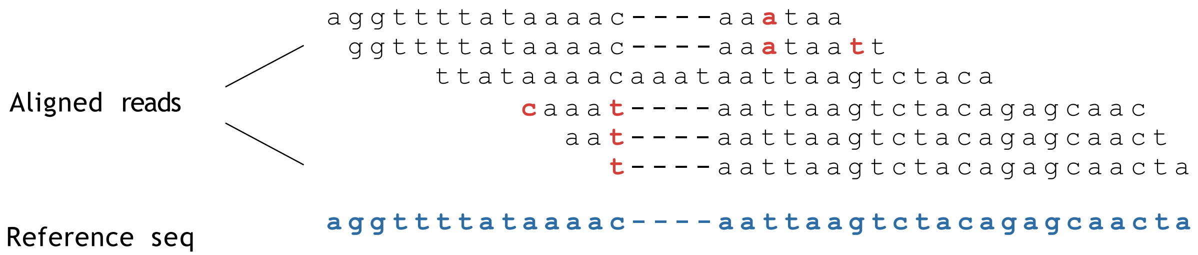

RNA-seq Replicate 1
RNA-seq Replicate 2
DNA
Nothing obviously wrong, but the variant is not reproducible

Mind the biological variability. If possible, validate and replicate.

False SNPs Caused by Incorrect Alignment

Pairwise alignemnt artefacts can lead to false SNPs

- multiple sequence alignment is better, but very expensive
- instead: base alignment quality (BAQ) to lower quality of misaligned bases



Aligned reads

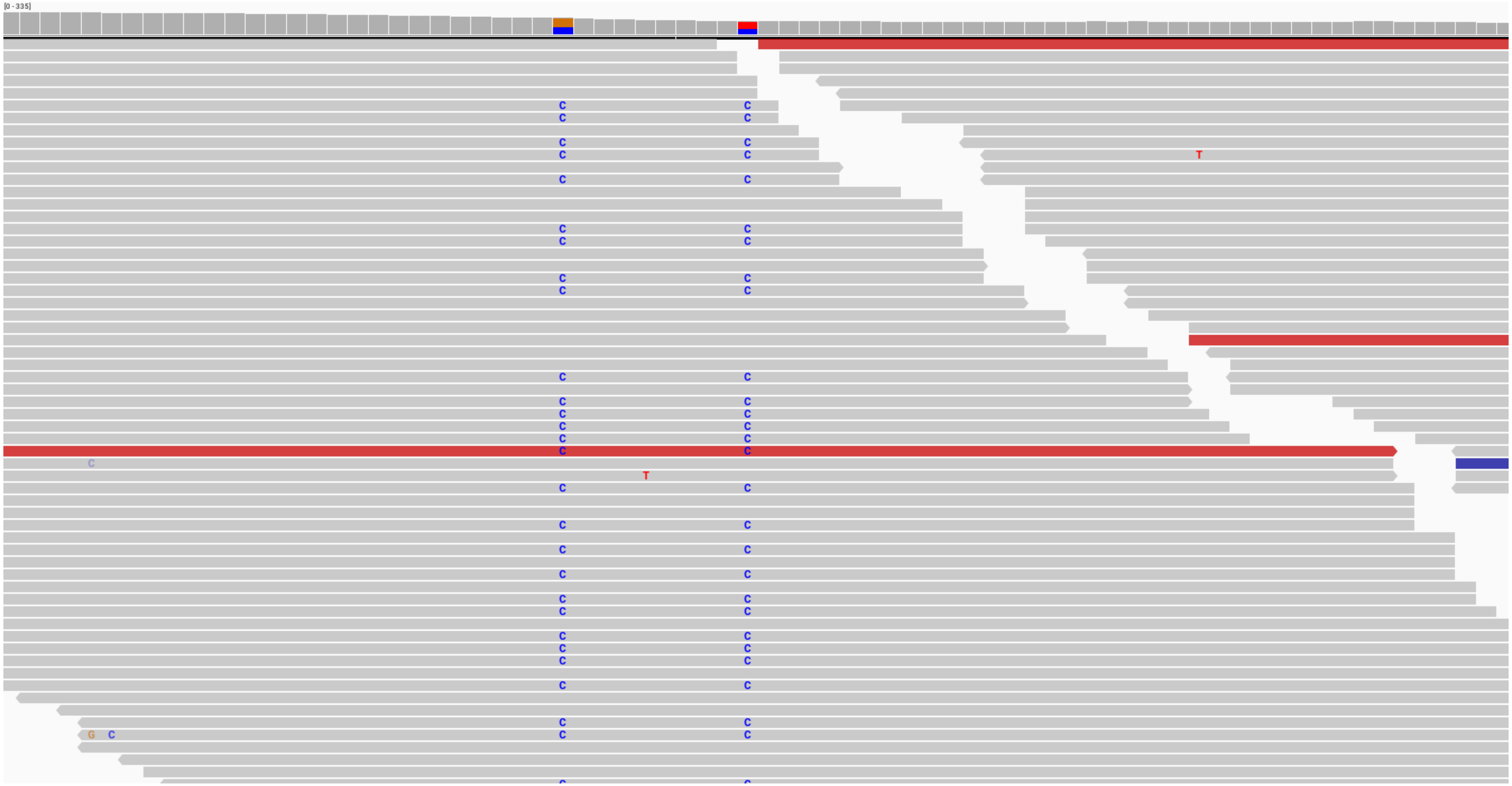
```
aggttttataaaac----aaataa
 ggttttataaaac----aaataatt
     ttataaaacaaataattaagtctaca
         caaat----aattaagtctacagagcaac
           aat----aattaagtctacagagcaact
             t----aattaagtctacagagcaacta
```

Reference seq

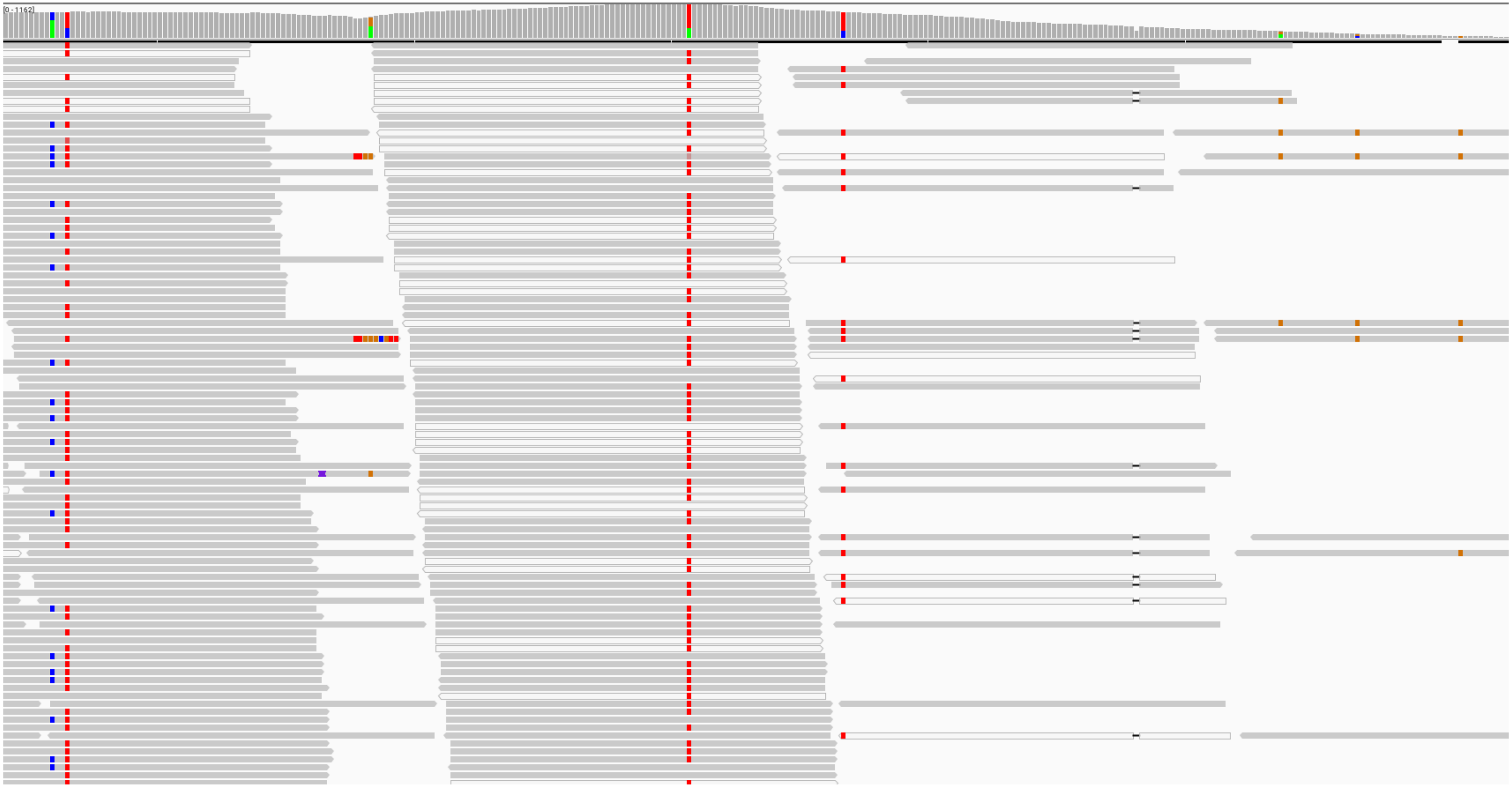
```
aggttttataaaac----aattaagtctacagagcaacta
```

Q: How many SNPs are real?

What Good SNPs Look Like?



What good SNPs look like?



Q: Is this call real? There are many reads with MQ=0.

Future of Variant Calling

Current approaches

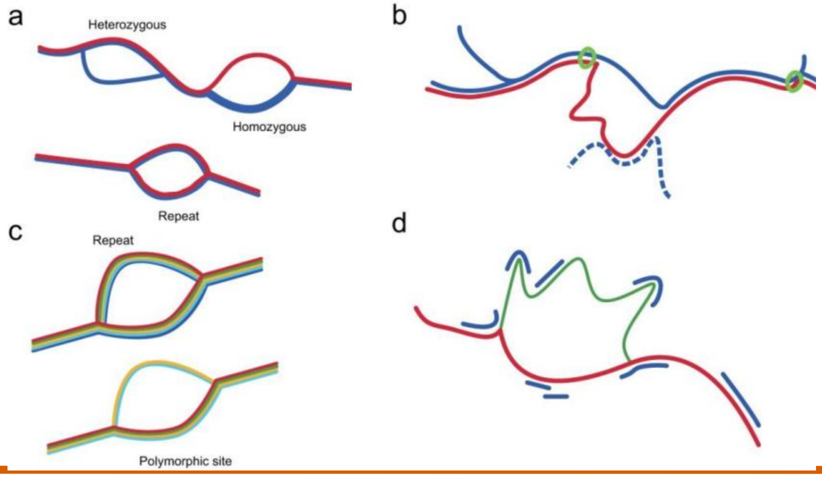
- rely heavily on the supplied alignment, but aligners see one read at a time
- largely site based, do not examine local haplotype and linked sites

Local *de novo* assembly based variant callers

- call SNPs, indels, MNPs and small SV simultaneously
- can remove alignment artefacts
- eg GATK haplotype caller, Scalpel, Octopus

Variation graphs

- align to a graph rather than a linear sequence



Iqbal et al. (2012) Nat Gen 44(2):226

Functional Annotation

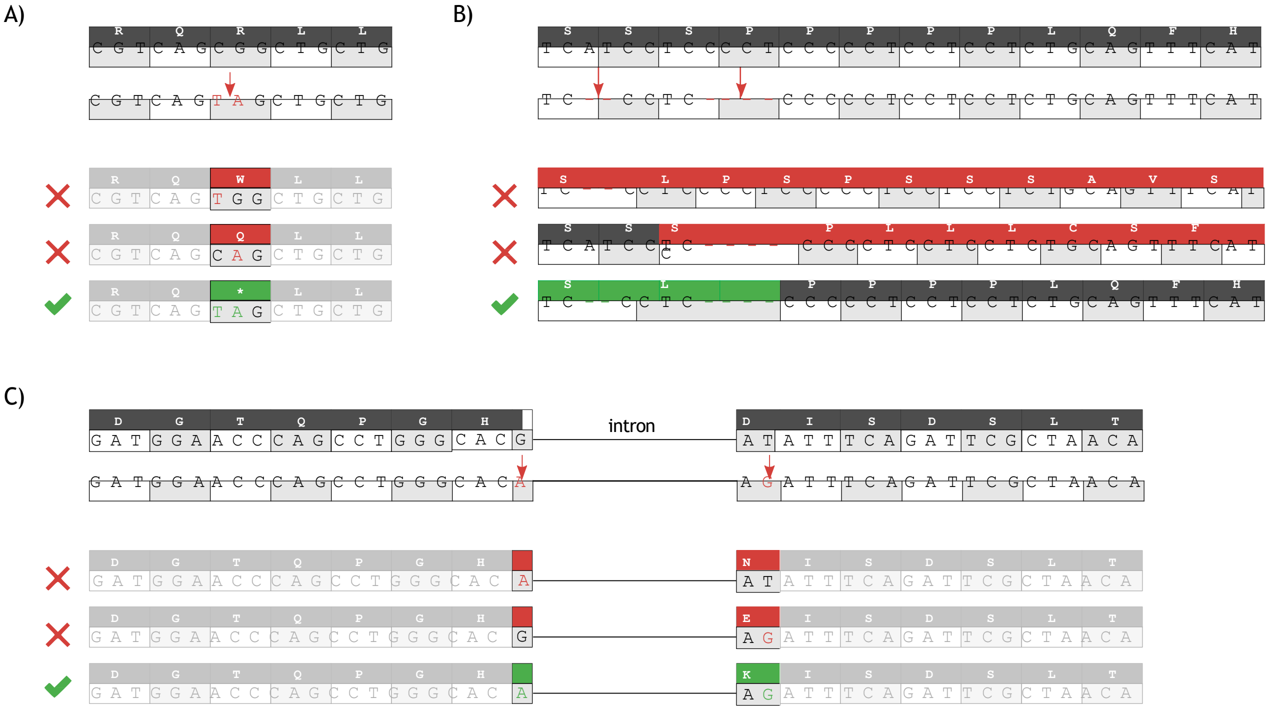
VCF can store arbitrary INFO tags (per site) and FORMAT tags (per sample)

- describe genomic context of the variant (e.g. coding, intronic, UTR)
- predict functional consequence (e.g. synonymous, missense, start lost)

Several tools for annotating a VCF, only few are haplotype-aware

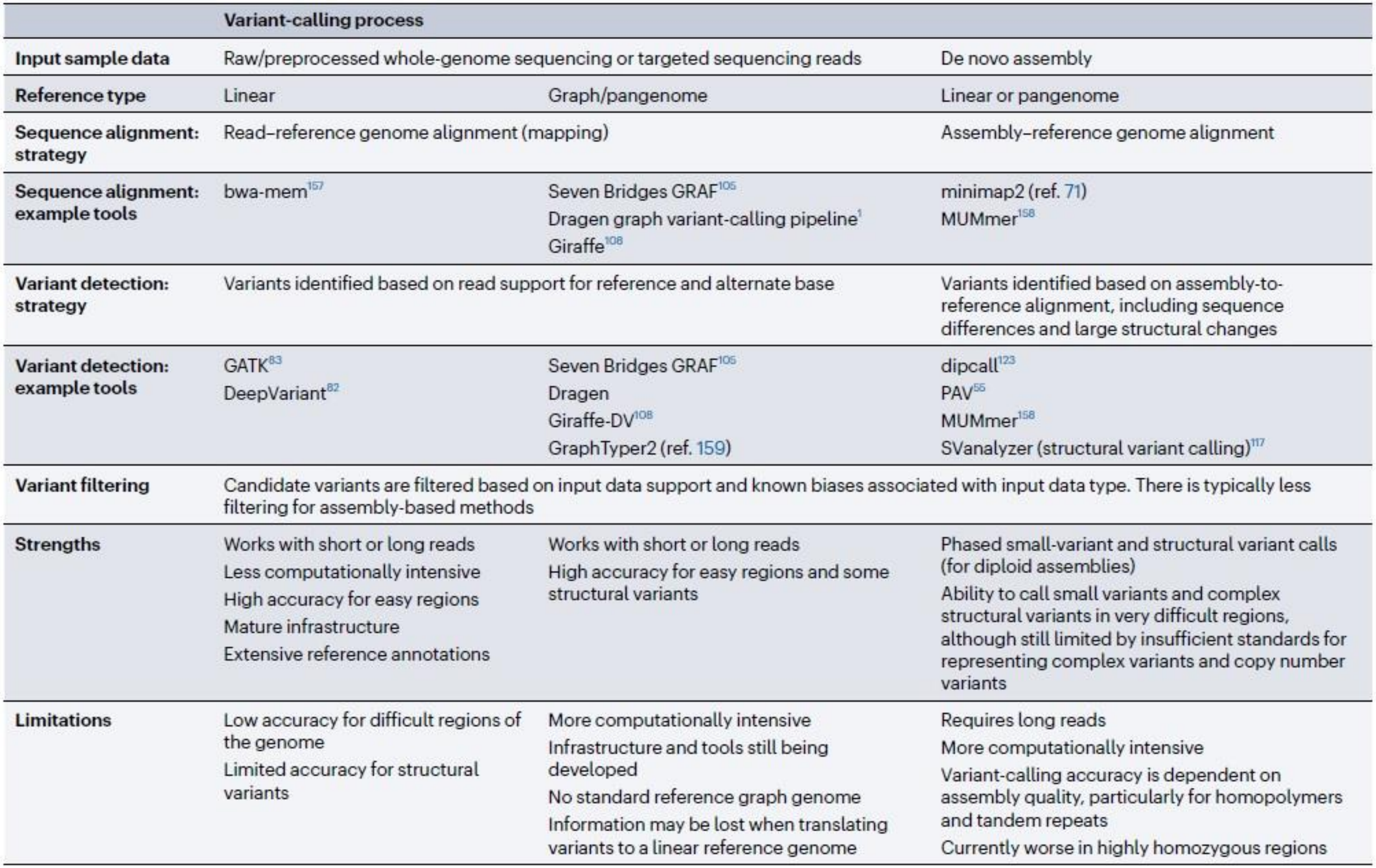
BCFtools/csq http://github.com/samtools/bcftools

VEP Haplosaurus http://github.com/willmclaren/ensembl-vep



A)
R Q R L L
C G T C A G C G G C T G C T G
C G T C A G T A G C T G C T G
R Q W L L
C G T C A G T G G C T G C T G
R Q Q L L
C G T C A G C A G C T G C T G
R Q * L L
C G T C A G T A G C T G C T G
✗
✗
✓
B)
C)
D G T Q P G H
G A T G G A A C C C A G C C T G G G C A C G
intron
D I S D S L T
A T A T T T C A G A T T C G C T A A C A
N
E
K

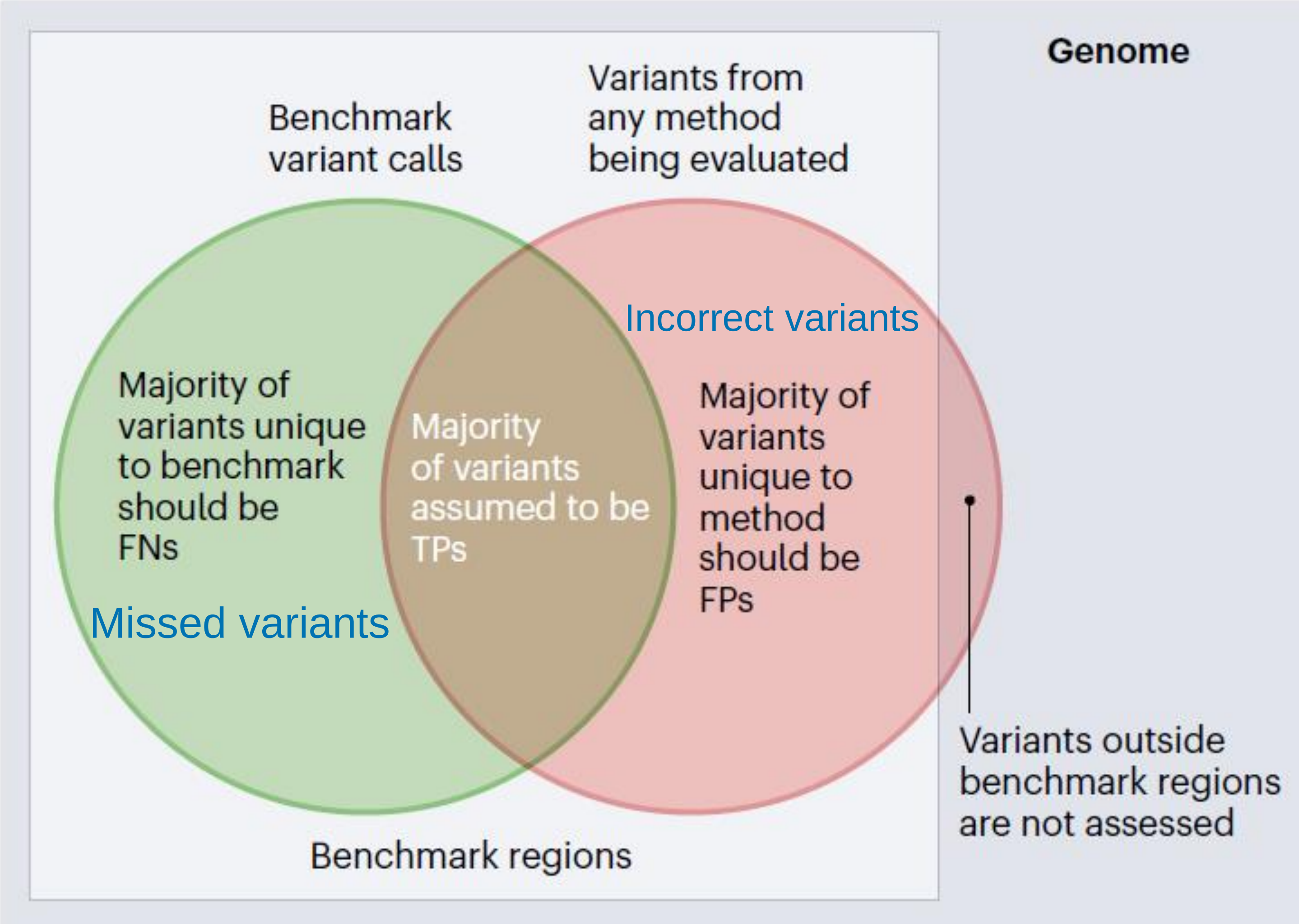
Typical Variant Calling Process



	Variant-calling process		
Input sample data	Raw/preprocessed whole-genome sequencing or targeted sequencing reads		De novo assembly
Reference type	Linear	Graph/pangenome	Linear or pangenome
Sequence alignment: strategy	Read–reference genome alignment (mapping)		Assembly–reference genome alignment
Sequence alignment: example tools	bwa-mem[157]	Seven Bridges GRAF[105] Dragen graph variant-calling pipeline[1] Giraffe[108]	minimap2 (ref. 71) MUMmer[158]
Variant detection: strategy	Variants identified based on read support for reference and alternate base		Variants identified based on assembly-to-reference alignment, including sequence differences and large structural changes
Variant detection: example tools	GATK[83] DeepVariant[82]	Seven Bridges GRAF[105] Dragen Giraffe-DV[108] GraphTyper2 (ref. 159)	dipcall[123] PAV[55] MUMmer[158] SVanalyzer (structural variant calling)[117]
Variant filtering	Candidate variants are filtered based on input data support and known biases associated with input data type. There is typically less filtering for assembly-based methods		
Strengths	Works with short or long reads Less computationally intensive High accuracy for easy regions Mature infrastructure Extensive reference annotations	Works with short or long reads High accuracy for easy regions and some structural variants	Phased small-variant and structural variant calls (for diploid assemblies) Ability to call small variants and complex structural variants in very difficult regions, although still limited by insufficient standards for representing complex variants and copy number variants
Limitations	Low accuracy for difficult regions of the genome Limited accuracy for structural variants	More computationally intensive Infrastructure and tools still being developed No standard reference graph genome Information may be lost when translating variants to a linear reference genome	Requires long reads More computationally intensive Variant-calling accuracy is dependent on assembly quality, particularly for homopolymers and tandem repeats Currently worse in highly homozygous regions

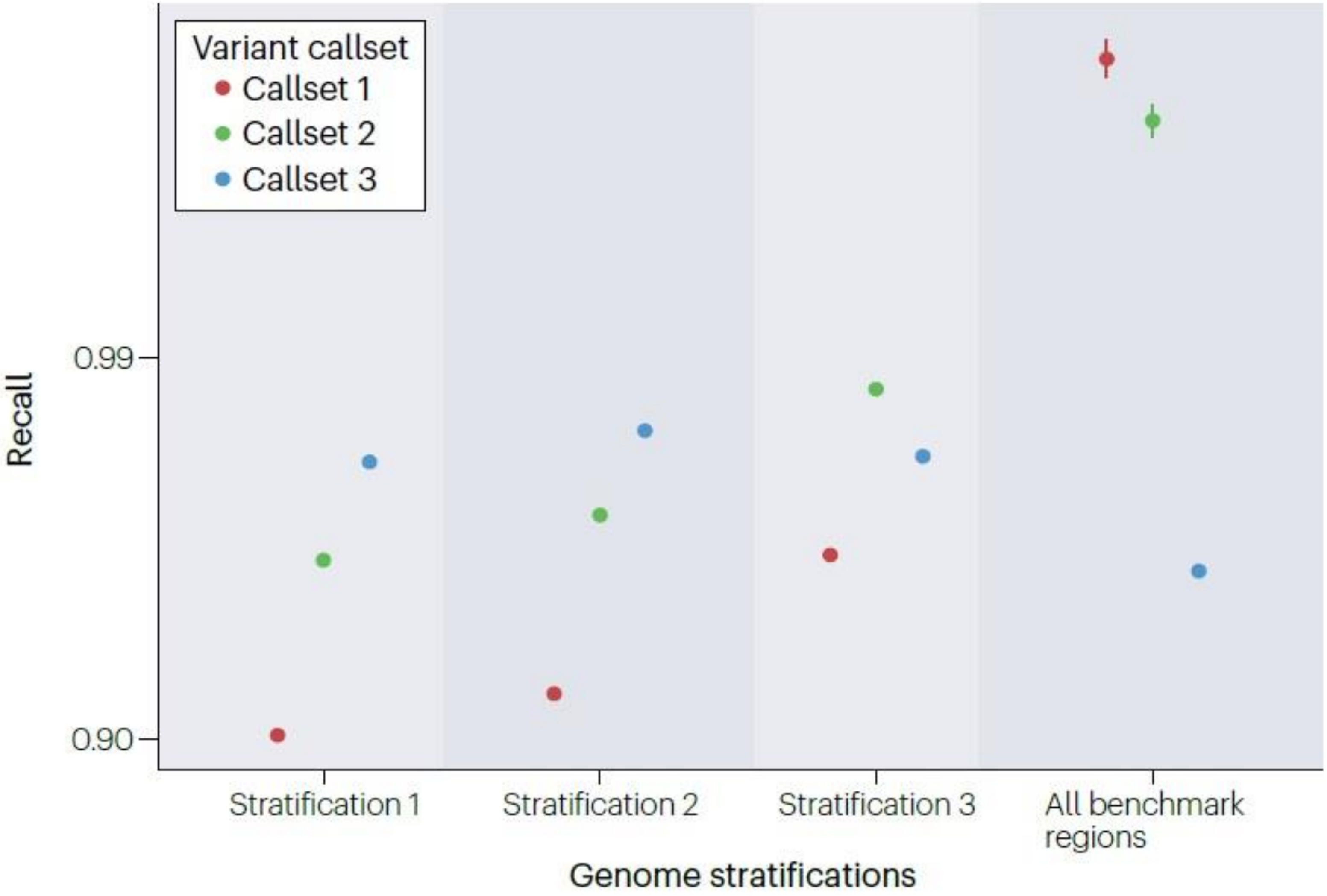
Olson *et. al.* 2023 Nature Reviews Genetics 24:464–483.

Benchmarking Variant Calls


Genome
Benchmark variant calls
Variants from any method being evaluated
Incorrect variants
Majority of variants unique to benchmark should be FNs
Majority of variants assumed to be TPs
Majority of variants unique to method should be FPs
Missed variants
Variants outside benchmark regions are not assessed
Benchmark regions

Olson *et. al.* 2023 Nature Reviews Genetics 24:464–483.

Which is A Good Callset?

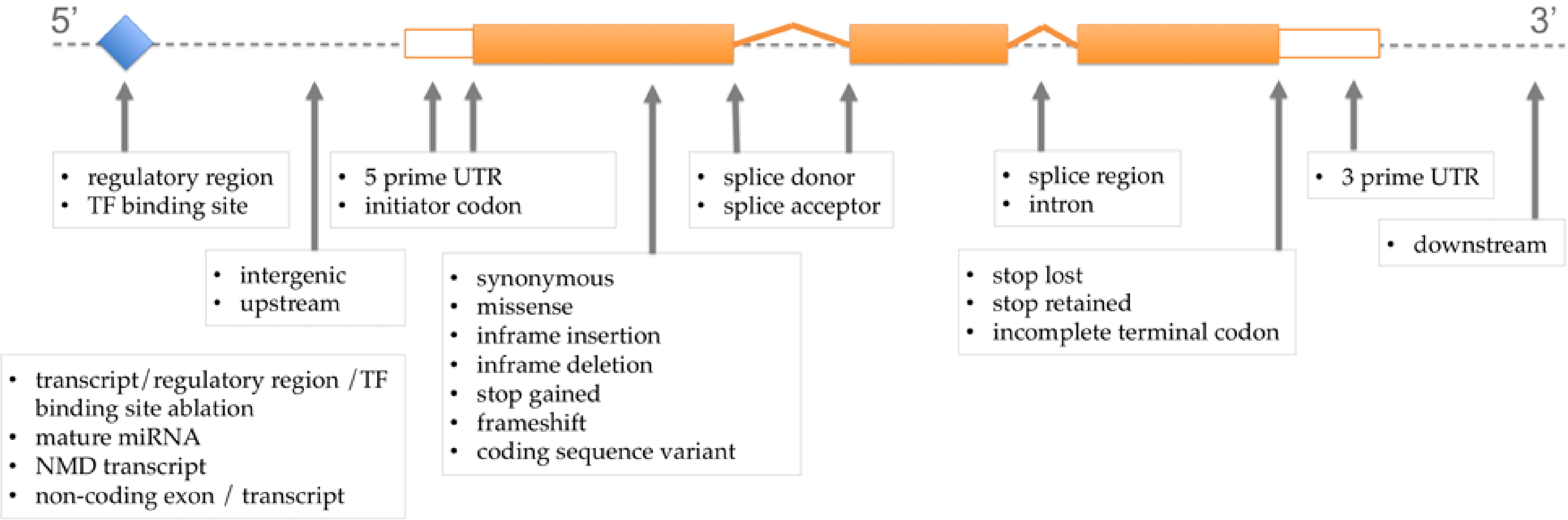


Olson *et. al.* 2023 Nature Reviews Genetics 24:464–483.

Variation Consequences

Variant Consequences

Most variation has no effect



- regulatory region
- TF binding site

- intergenic
- upstream

- 5 prime UTR
- initiator codon

- synonymous
- missense
- inframe insertion
- inframe deletion
- stop gained
- frameshift
- coding sequence variant

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

- 3 prime UTR

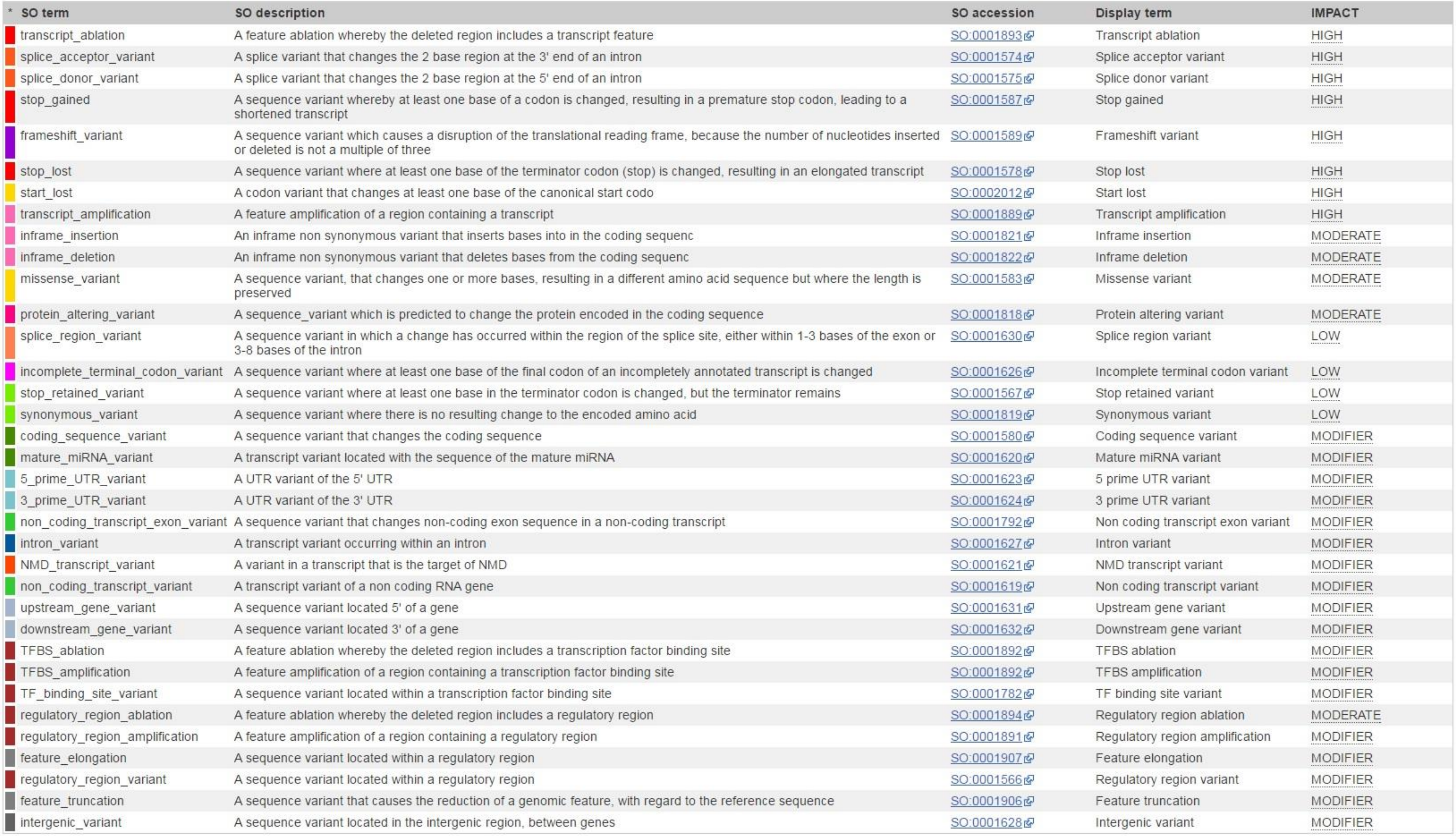
- downstream

- transcript/ regulatory region / TF binding site ablation
- mature miRNA
- NMD transcript
- non-coding exon / transcript

www.ensembl.org/info/docs/variation/

On average, every person carries mutations that inactivate at least one copy of 200 or so genes and both copies of around 20 genes

SNPs in Ensembl - Types



*	SO term	SO description	SO accession	Display term	IMPACT
🟥	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
🟧	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
🟧	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
🟥	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
🟪	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
🟥	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
🟨	start_lost	A codon variant that changes at least one base of the canonical start codo	SO:0002012	Start lost	HIGH
🟪	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
🟪	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequenc	SO:0001821	Inframe insertion	MODERATE
🟪	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	SO:0001822	Inframe deletion	MODERATE
🟨	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
🟥	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
🟧	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
🟪	incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
🟩	stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
🟩	synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW
🟩	coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding sequence variant	MODIFIER
🟩	mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Mature miRNA variant	MODIFIER
🟦	5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5 prime UTR variant	MODIFIER
🟦	3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3 prime UTR variant	MODIFIER
🟩	non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	SO:0001792	Non coding transcript exon variant	MODIFIER
🟦	intron_variant	A transcript variant occurring within an intron	SO:0001627	Intron variant	MODIFIER
🟧	NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript variant	MODIFIER
🟩	non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO:0001619	Non coding transcript variant	MODIFIER
🟦	upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream gene variant	MODIFIER
🟦	downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream gene variant	MODIFIER
🟫	TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO:0001892	TFBS ablation	MODIFIER
🟫	TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	SO:0001892	TFBS amplification	MODIFIER
🟫	TF_binding_site_variant	A sequence variant located within a transcription factor binding site	SO:0001782	TF binding site variant	MODIFIER
🟫	regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	SO:0001894	Regulatory region ablation	MODERATE
🟫	regulatory_region_amplification	A feature amplification of a region containing a regulatory region	SO:0001891	Regulatory region amplification	MODIFIER
⬛	feature_elongation	A sequence variant located within a regulatory region	SO:0001907	Feature elongation	MODIFIER
🟫	regulatory_region_variant	A sequence variant located within a regulatory region	SO:0001566	Regulatory region variant	MODIFIER
⬛	feature_truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	SO:0001906	Feature truncation	MODIFIER
⬛	intergenic_variant	A sequence variant located in the intergenic region, between genes	SO:0001628	Intergenic variant	MODIFIER

* Corresponding colours for the Ensembl web displays.

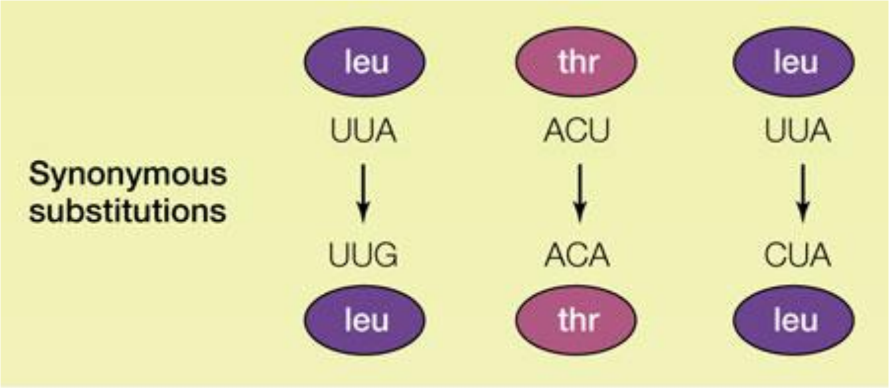
MONASH University MALAYSIA

wellcome connecting science

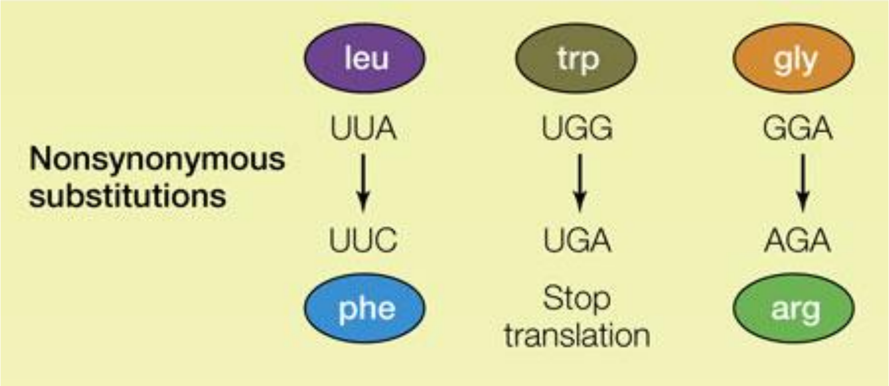
Genomics Platform

Types of Protein Coding Mutations

- Synonymous substitutions are those that do not change the amino acid sequence.



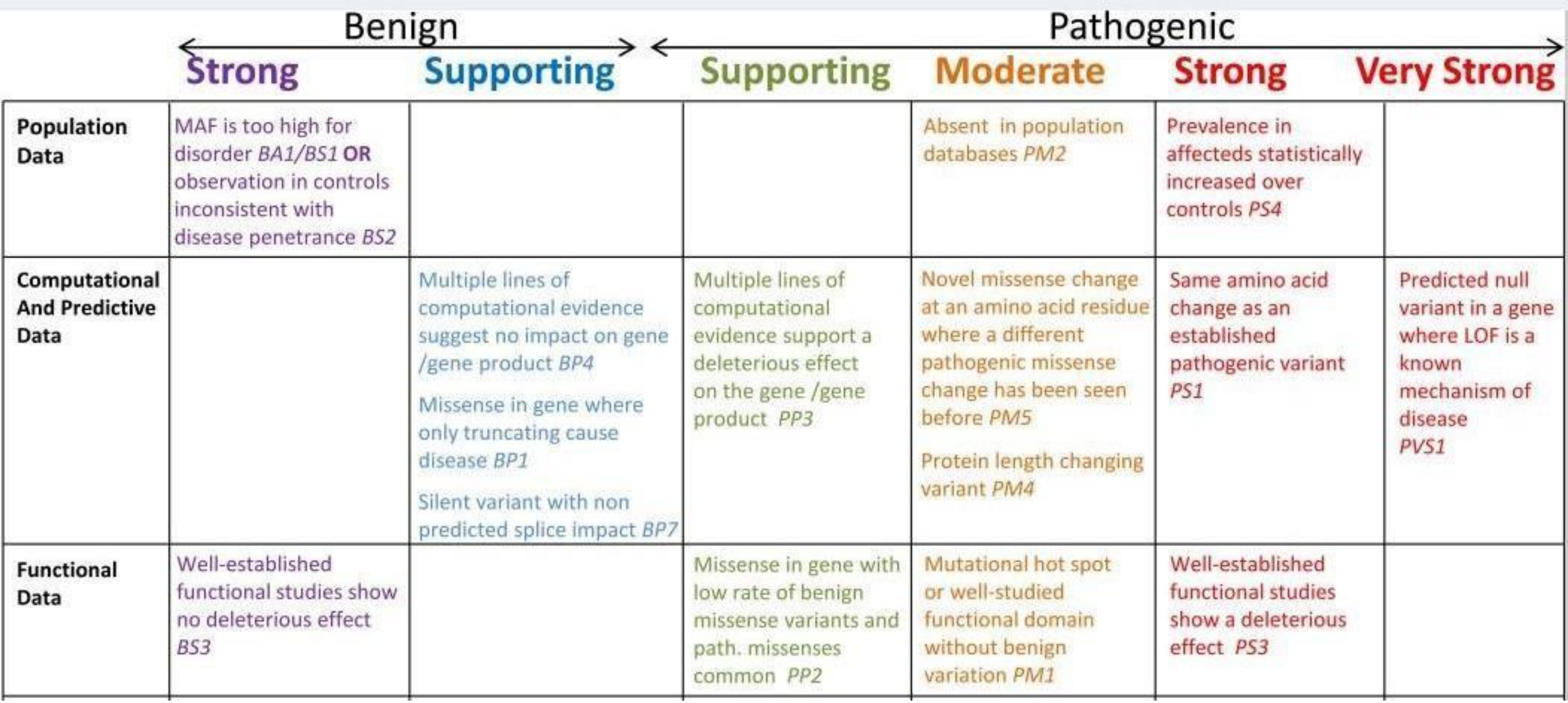
- Non-synonymous or missense substitutions are those that change the amino acid sequence.



Variant Pathogenicity

- Pathogenic:
  - Disease causing.
- Likely Pathogenic:
  - Might be disease causing.
- Likely Benign:
  - Most likely does not cause disease.
- Benign:
  - Non-disease causing.
- Variants of Uncertain Significance (VUS):
  - Do not meet any of the above criteria or the criteria for benign and pathogenic are contradictory.

Classifying Disease Variants



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	

Human Population Specific Variation

International HapMap Project

Home | About the Project | Data | Publications | Tutorial

http://hapmap.ncbi.nlm.nih.gov/

A recent computer security audit revealed security flaws in the legacy HapMap site and NCBI has took it down in June 2016.

http://www.internationalgenome.org/

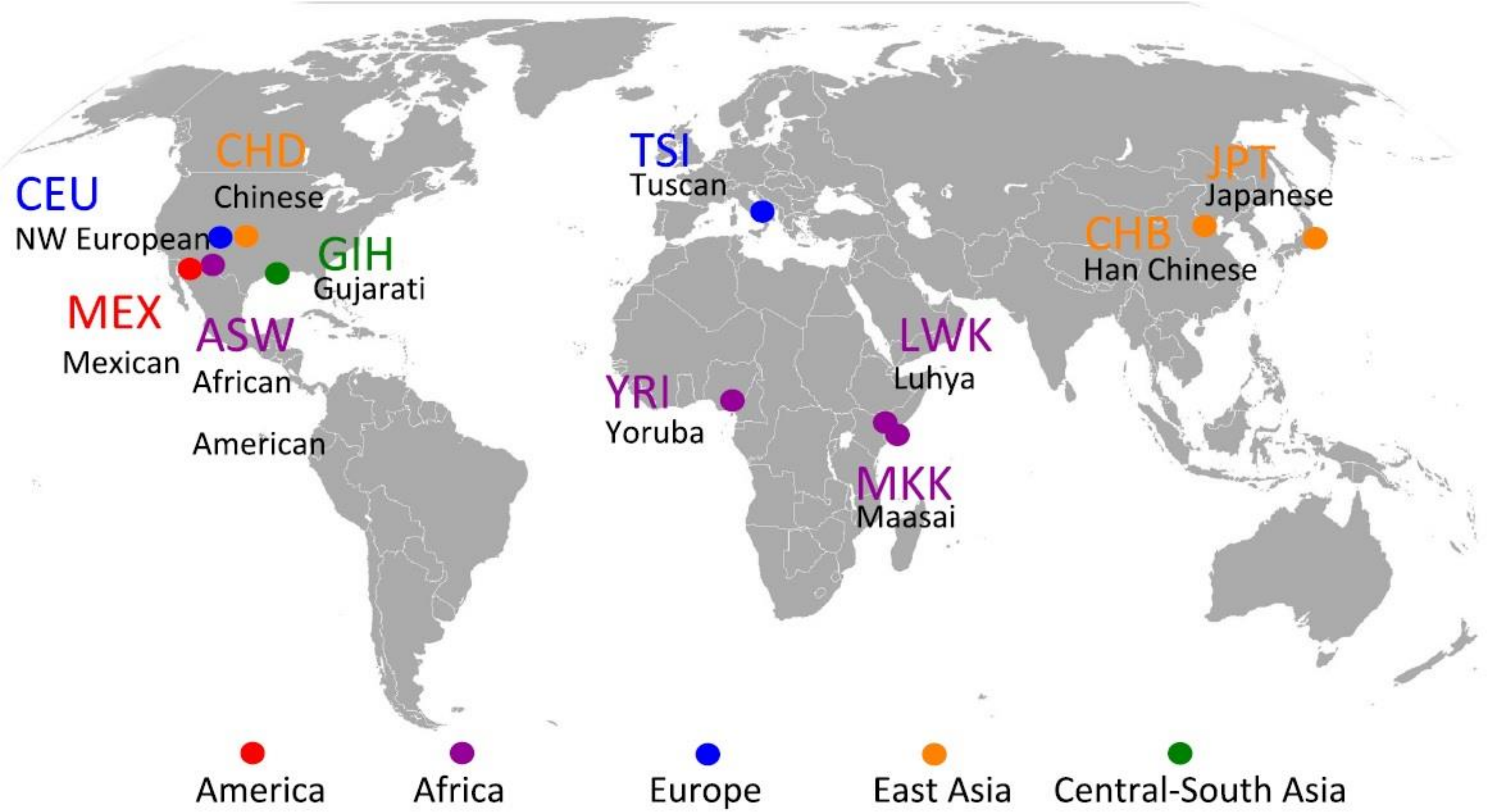
IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

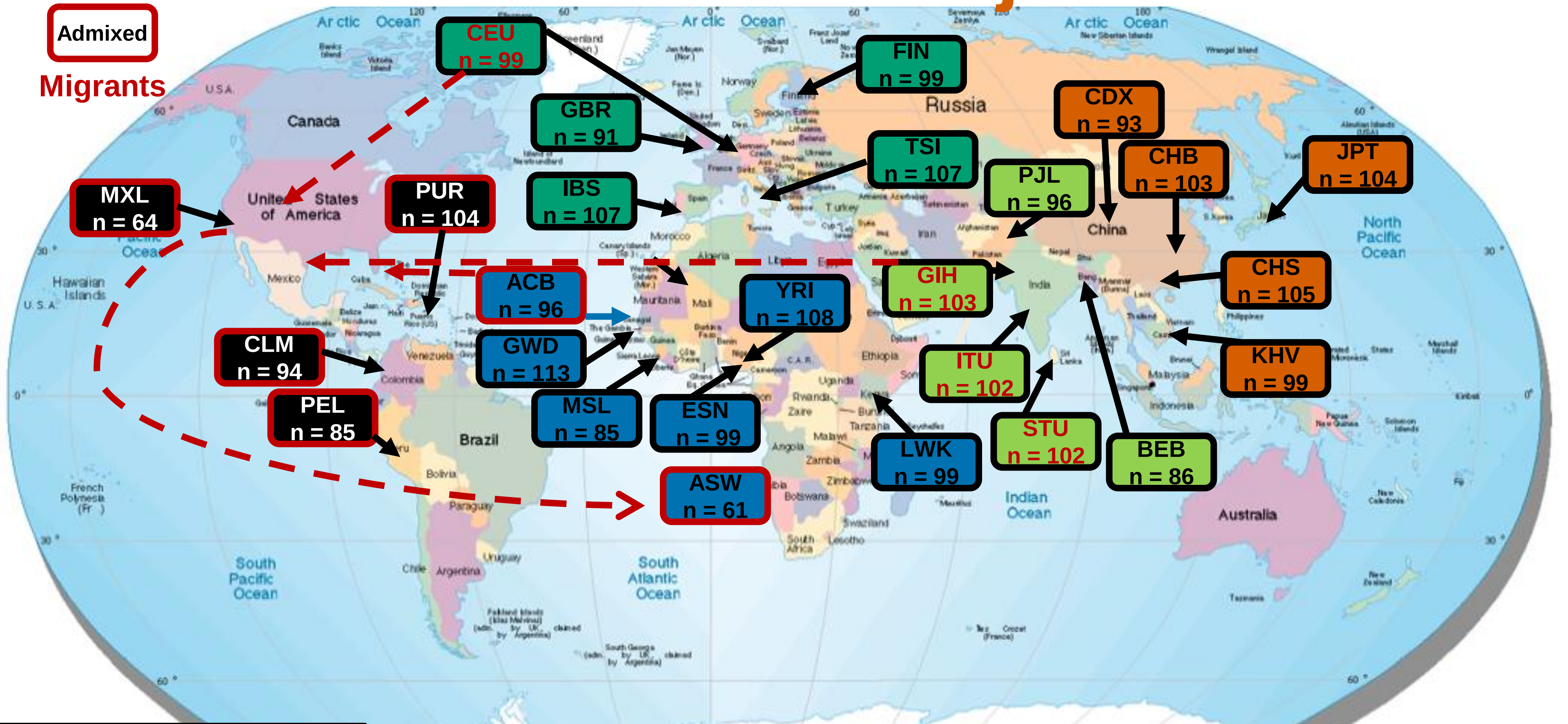
http://exac.broadinstitute.org/

Exome Aggregation Consortium (ExAC) Browser

International HapMap Project



The 1000 Genomes Project Dataset



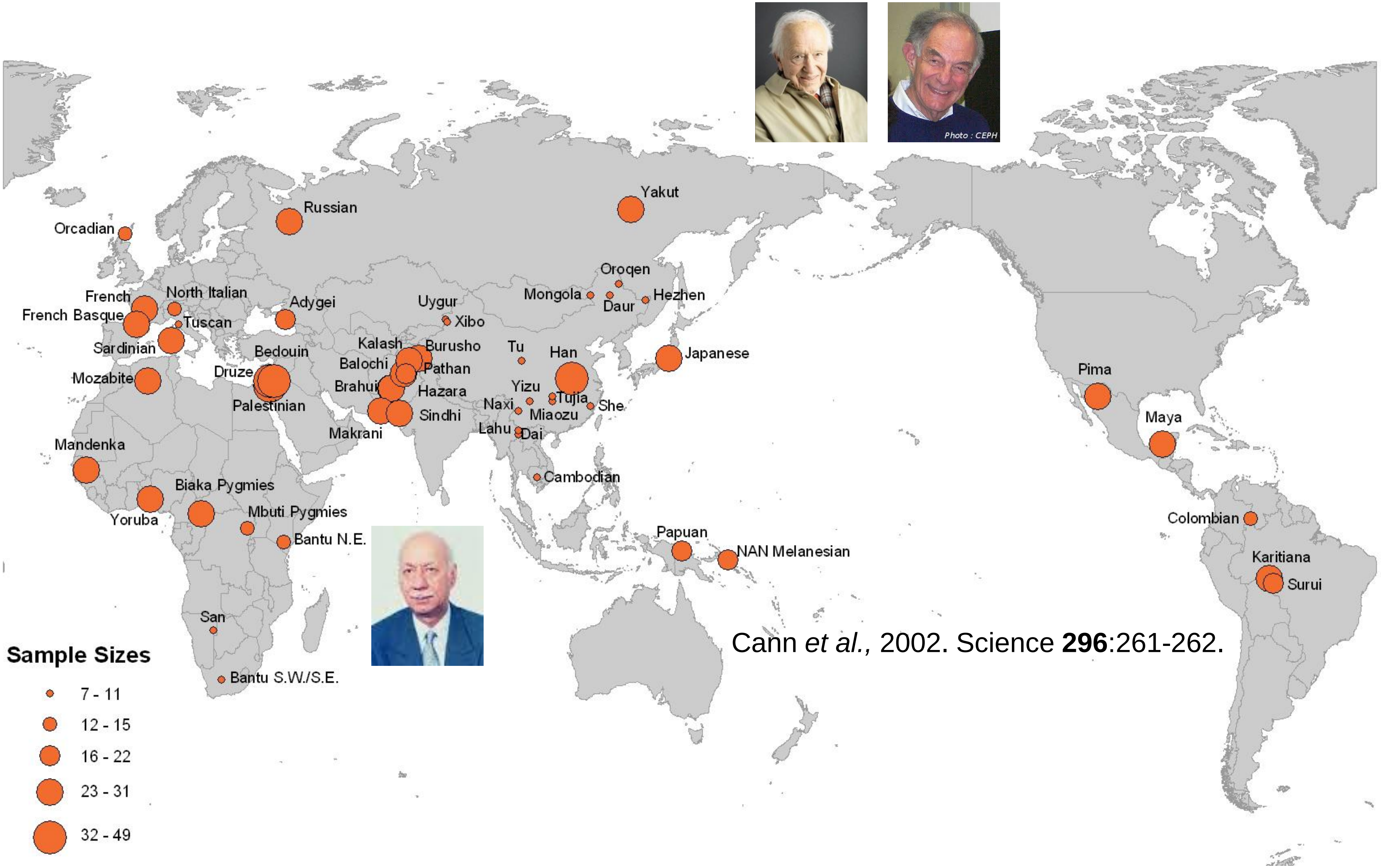
🟦	African	(n = 661)
🟧	East Asian	(n = 504)
🟩	South Asian	(n = 489)
🟩	European	(n = 503)
⬛	American	(n = 347)

The 1000 Genomes Project Consortium Nature (2015) 526:68-74.

http://www.1000genomes.org/page.php

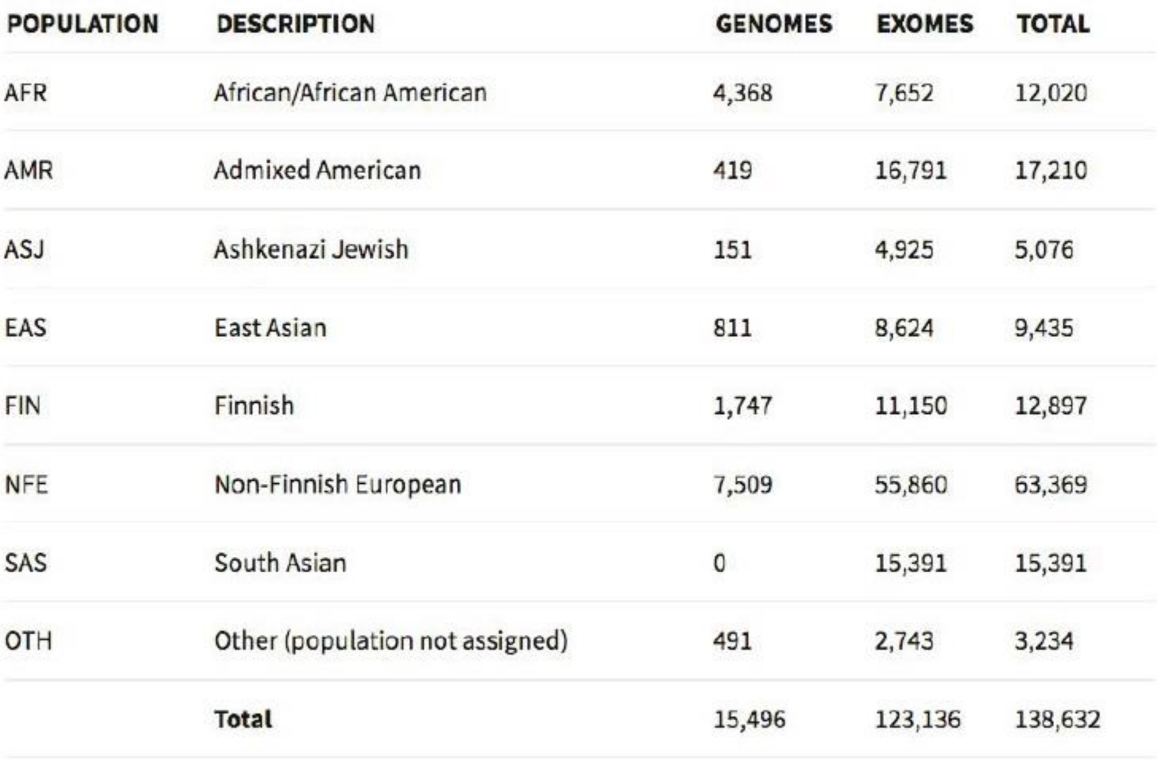
Samples	Populations	Mean Coverage	SNPs
2,504	26	7.4 X	84.7 M

The HGDP-CEPH Cell Line Panel

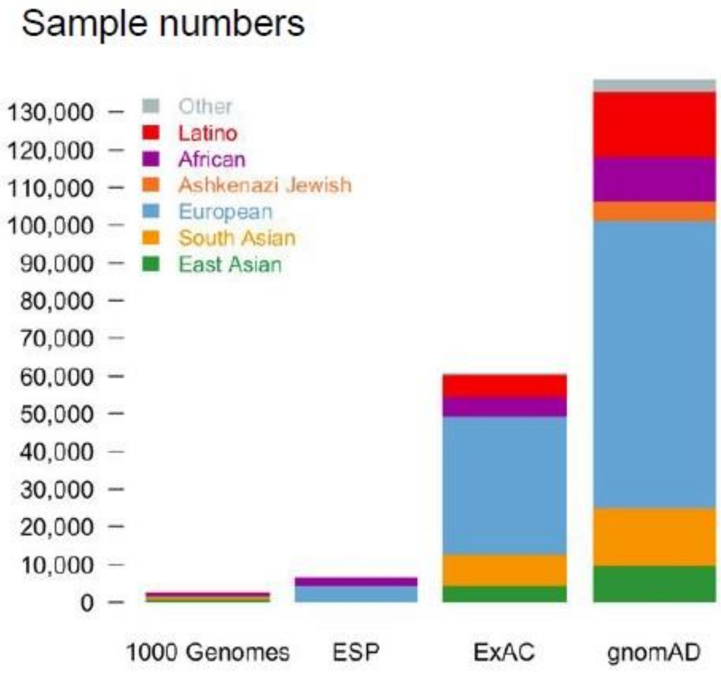


Cann *et al.*, 2002. Science **296**:261-262.

Genome Aggregation Database (gnomAD)

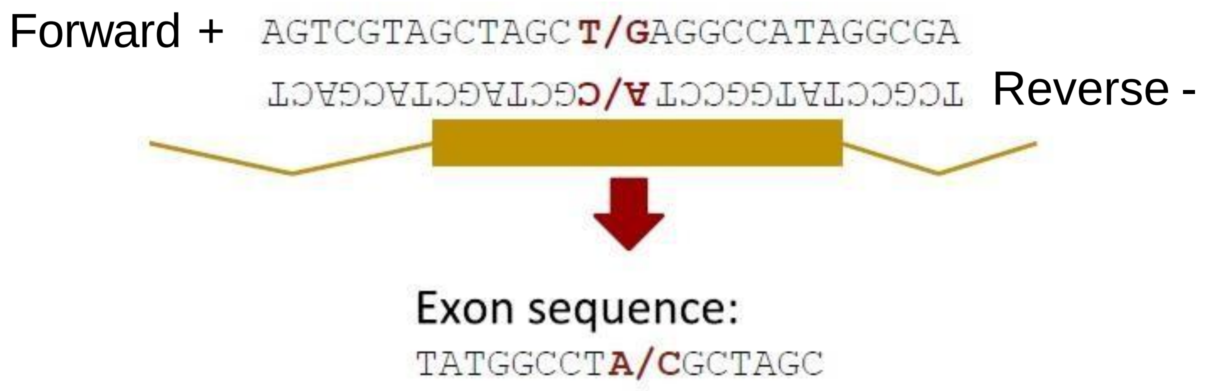


POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632



https://gnomad.broadinstitute.org/about

Ensembl Variation



Forward + AGTCGTAGCTAGC**T/G**AGGCCATAGGCGA

TCGCCTATGGCCT**A/C**GCTAGCTACGACT Reverse -

Exon sequence:

TATGGCCT**A/C**GCTAGC

Ensembl	Alleles in database = T/G Alleles in gene = A/C
dbSNP **Short Genetic Variations**	Alleles = A/C -ve strand or T/G +ve strand
PubMed	Alleles = A/C or T/G Often lack further info

Questions

qasim.ayub@monash.edu