Variant Calling - SNPs and short indels

petr.danecek@sanger.ac.uk

wellcome trust sanger institute

HTS workflow

Library preparation

- DNA extraction
- fragmentation
- adapter ligation
- amplification

Sequencing

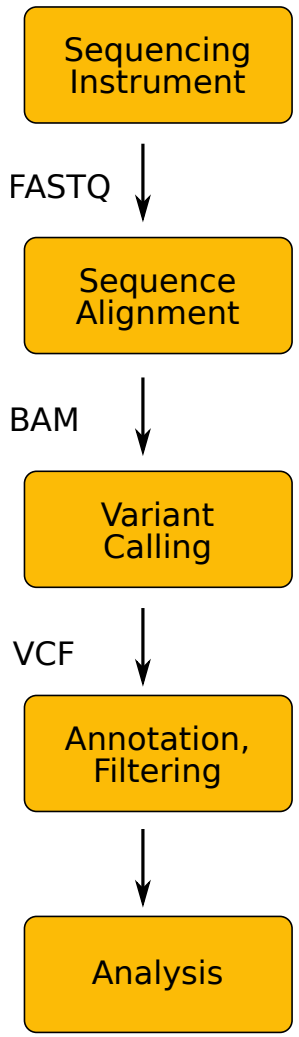
- base calling
- de-multiplexing

Data processing

- read mapping
- variant calling
- variant filtering

Analysis

- Variant annotation
- ...



Variant types

SNPs/SNVs . . . Single Nucleotide Polymorphism/Variation

```
ACGTTTAGCAT
ACGTTCAGCAT
```

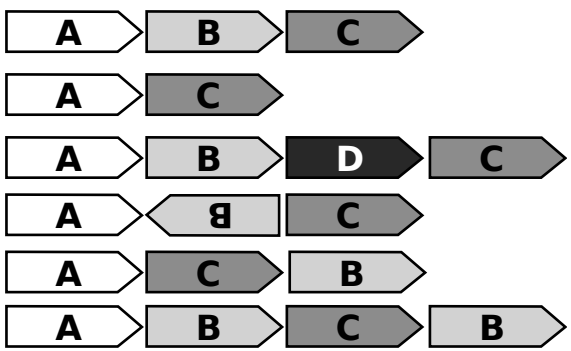
MNPs . . . Multi-Nucleotide Polymorphism

```
ACGTCCAGCAT
ACGTTTAGCAT
```

Indels . . . short insertions and deletions

```
ACGTTTAGCA-TT
ACGTT-AGCAGTT
```

SVs . . . Structural Variation



Some terminology

The goal is to determine the genotype at each position in the genome

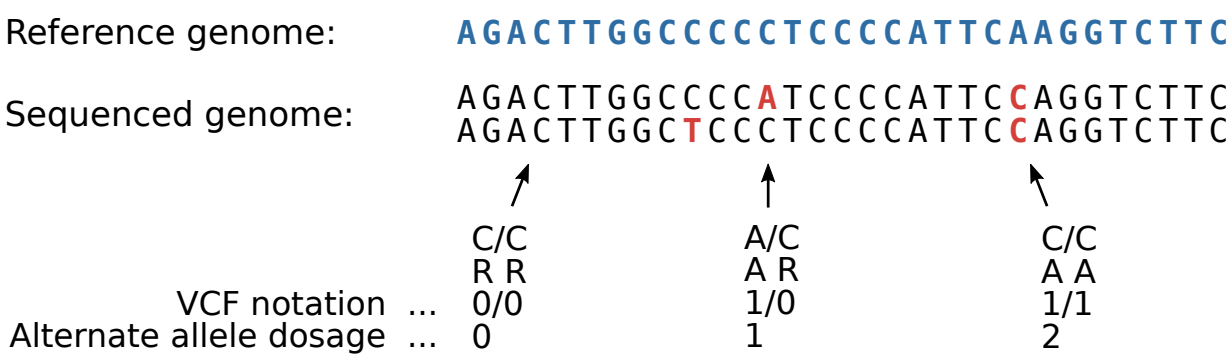
Genotype

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

Reference and alternate alleles - R and A

Diploid organism

- two chromosomal copies, three possible genotypes
  - RR .. homozygous reference genotype
  - RA .. heterozygous
  - AA .. homozygous alternate



Germline vs somatic mutation

Germline mutation

- heritable variation in the germ cells

Somatic mutation

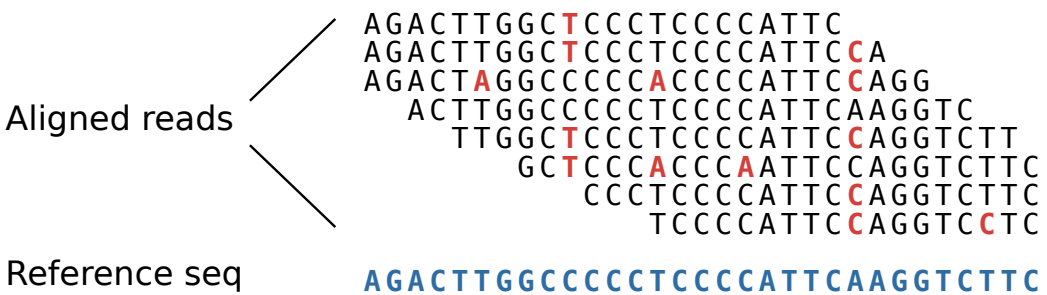
- variation in non-germline tissue, tumors. . .

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)

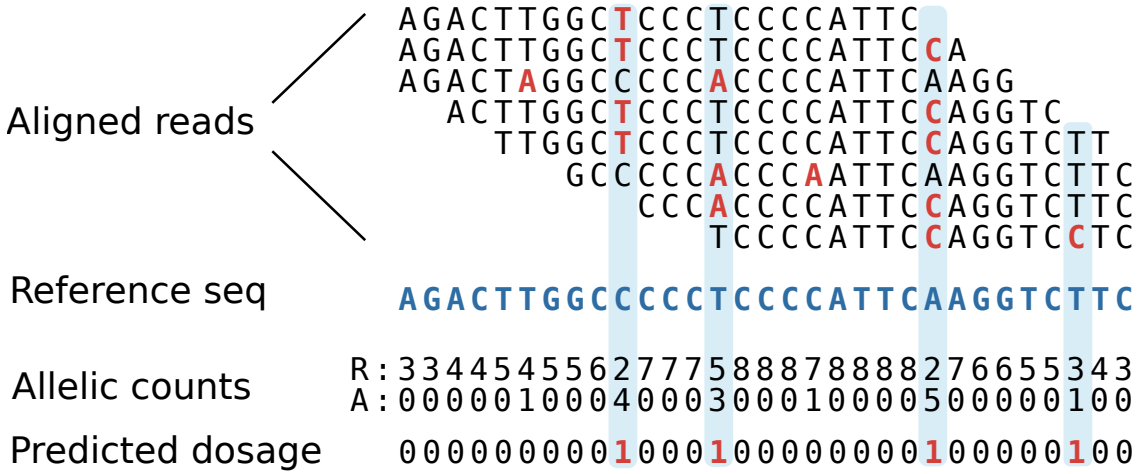
Somatic

- any fraction of alt AF possible - subclonal variation, admixture of normal cells in tumor sample



Naive variant calling

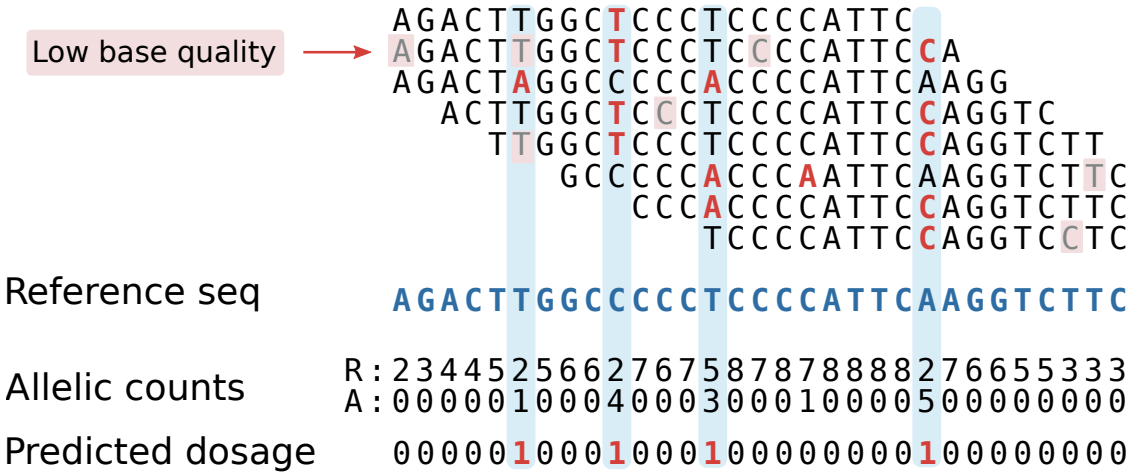
Use fixed allele frequency threshold to determine the genotype



alt AF	genotype
$[0, 0.2)$	RR .. homozygous reference
$[0.2, 0.8]$	RA .. herezogyous
$(0.8, 1]$	AA .. homozygous variant

Naive variant calling

Use fixed allele frequency threshold to determine the genotype

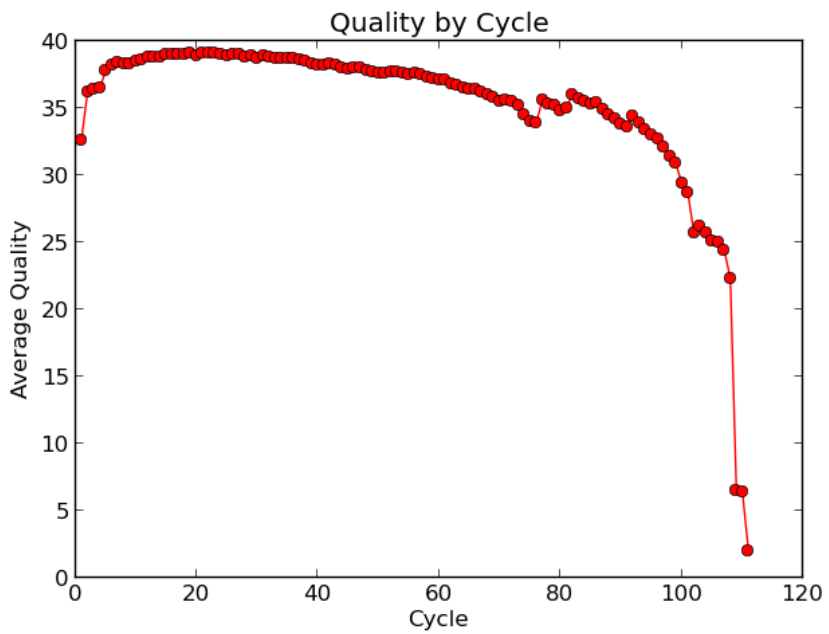


1) Filter base calls by quality
   e.g. ignore bases Q<20

Phred quality score

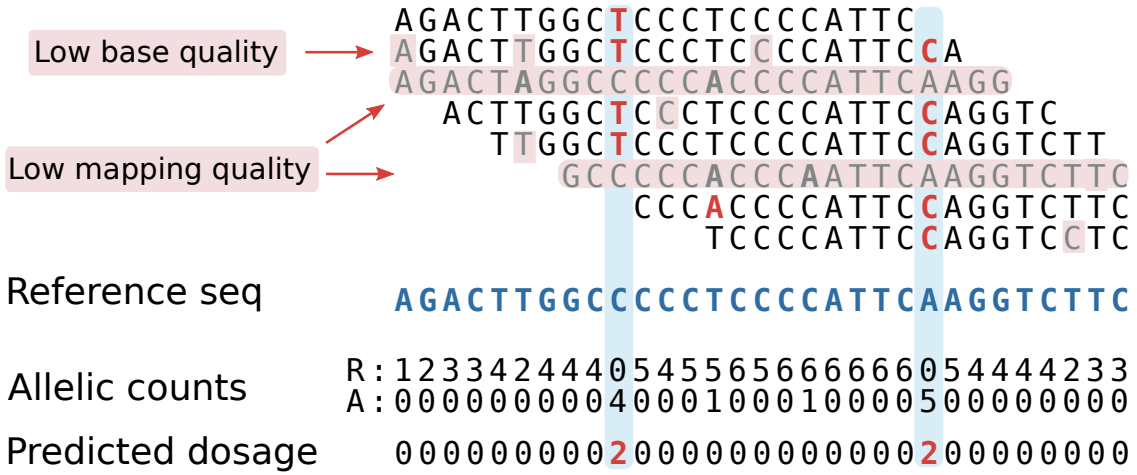
$$Q = -10 \log_{10} P_{\text{err}}$$

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%



Naive variant calling

Use fixed allele frequency threshold to determine the genotype



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

2) Filter reads with low mapping quality

alt AF	genotype
$[0, 0.2)$	RR .. homozygous reference
$[0.2, 0.8]$	RA .. herezogyous
$(0.8, 1]$	AA .. homozygous variant

Problems:

- undercalls hets in low-coverage data
- throws away information due to hard quality thresholds
- gives no measure of confidence

Real life calling models

More sophisticated models apply a statistical framework

Likelihood Prior

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

Posterior Normalization

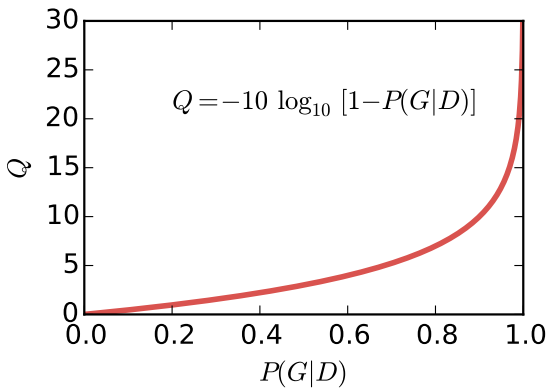
to determine:

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

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

2. and the genotype quality

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



Important terms you may encounter

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

    $P(\text{RA}) = 2f(1-f)$, $P(\text{RR}) = (1-f)^2$, $P(\text{AA}) = f^2$
- 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

Factors to consider in calling

Many calls are not real, a **filtering** step **is necessary**

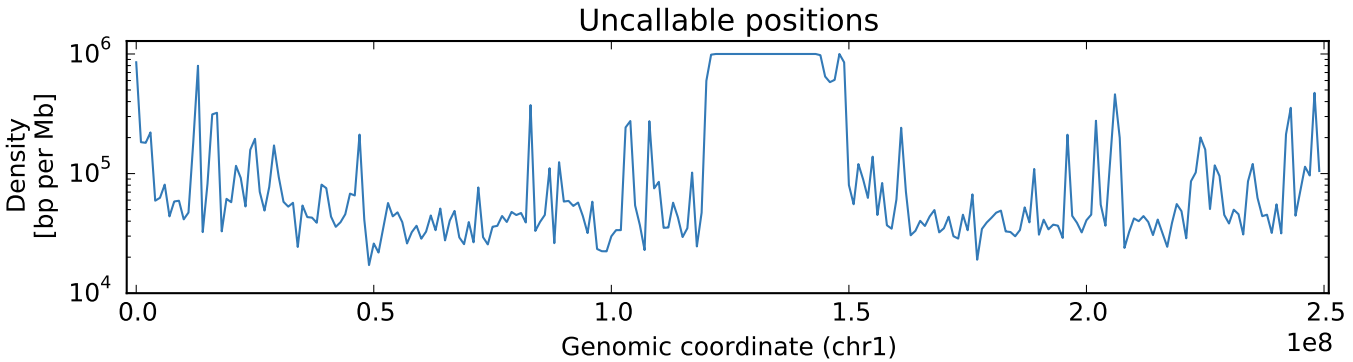
False calls can have many causes

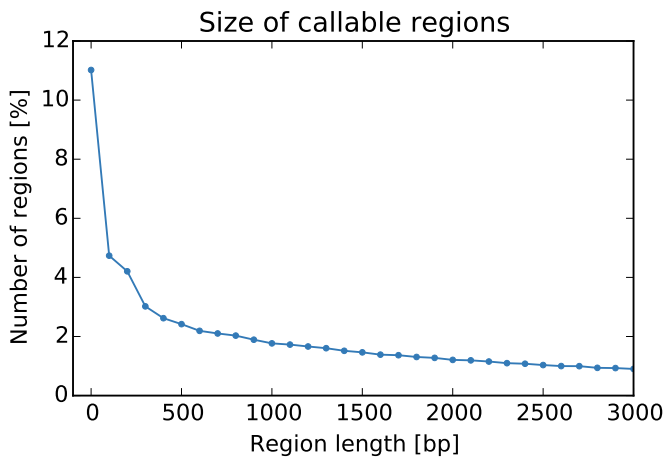
- contamination
- PCR errors
- sequencing errors
    - homopolymer runs
- mapping errors
    - repetitive sequence
    - structural variation
- alignment errors
    - false SNPs in proximity of indels
    - ambiguous indel alignment

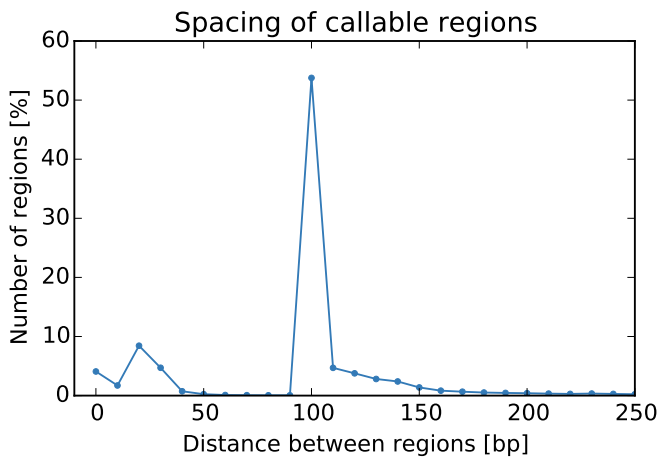
Callable genome

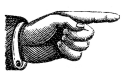
Large parts of the genome are still inaccessible

- the Genome in a Bottle high-confidence regions:
  - cover 89% of the reference genome
  - are short intervals scattered across the genome



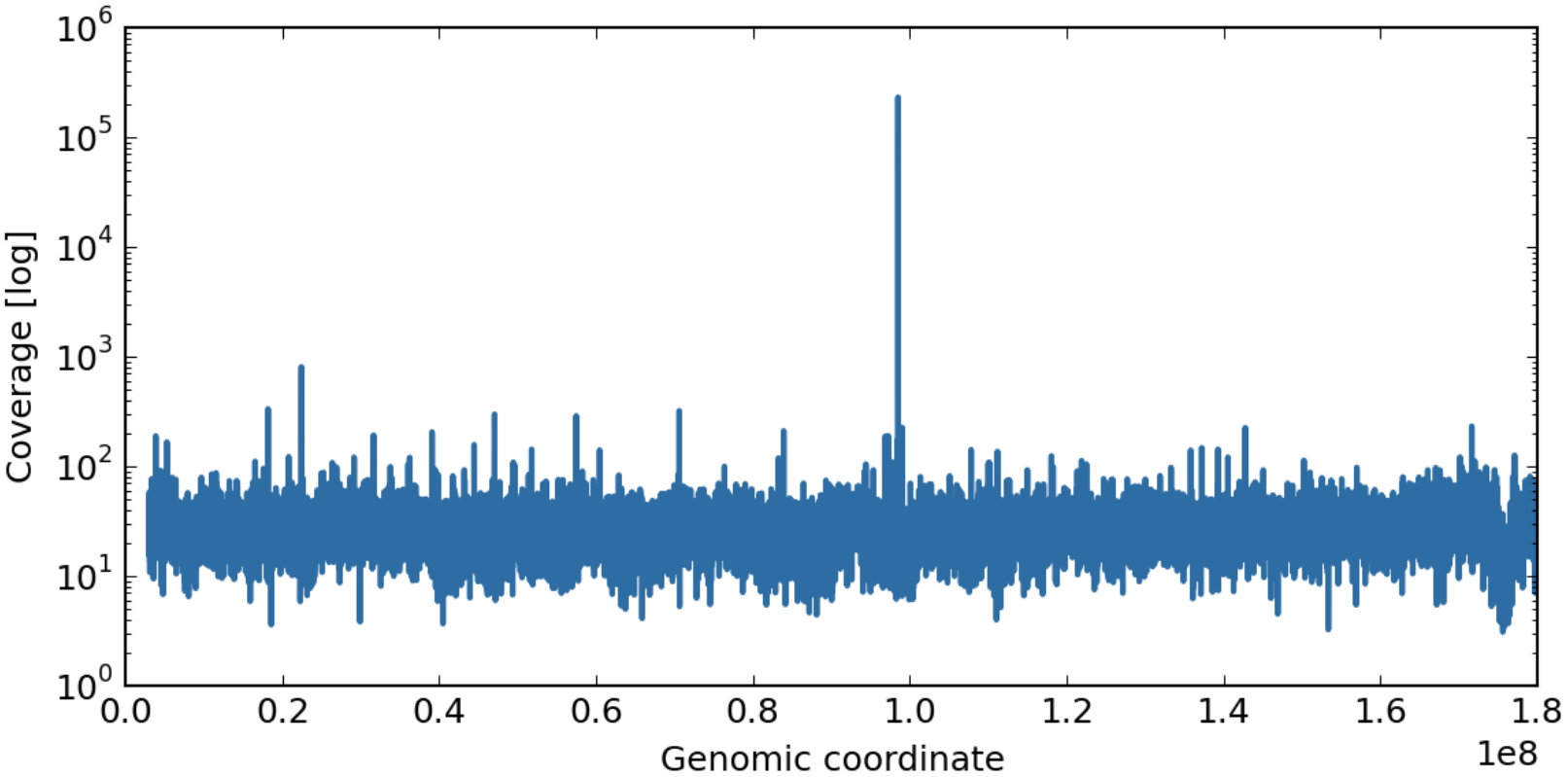






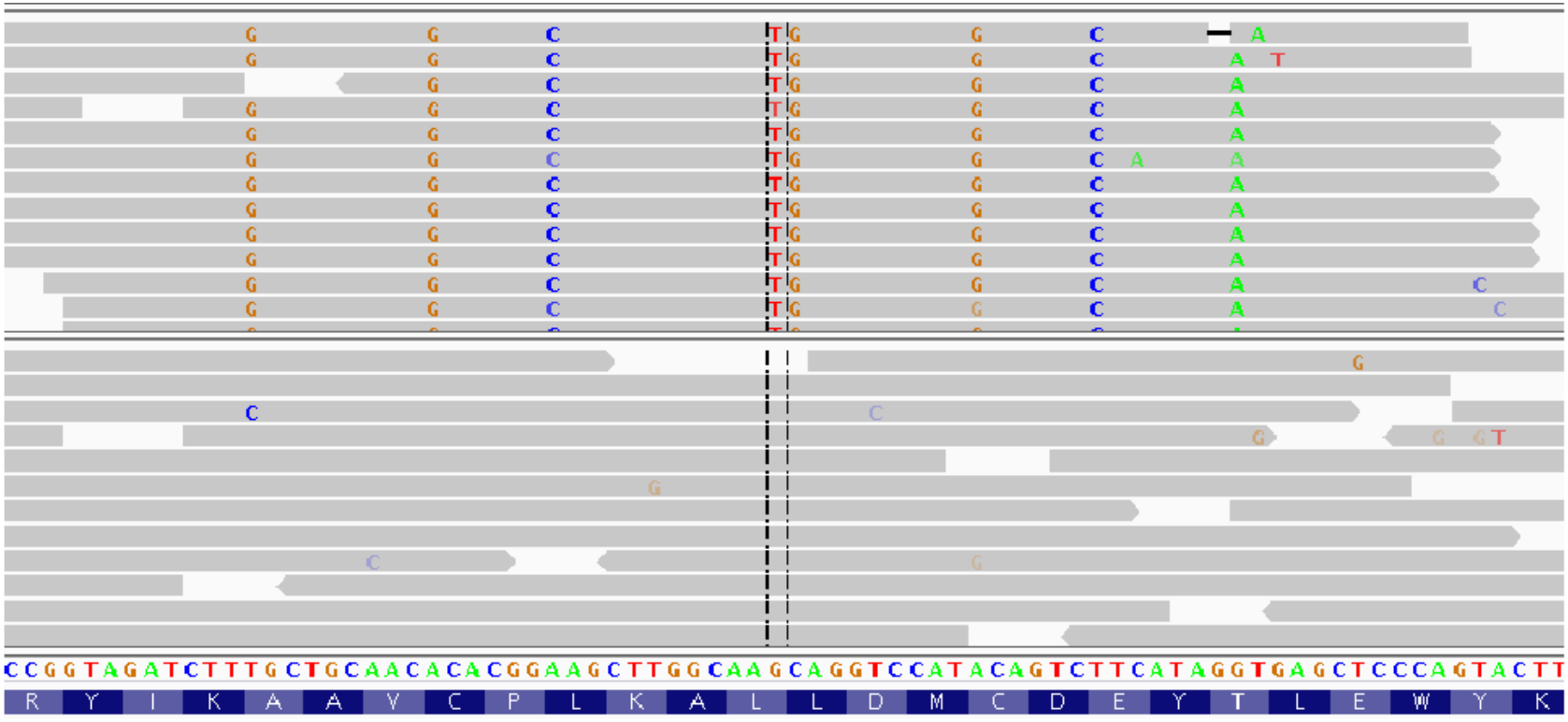
If possible, include only "nice" regions: for many analyses (e.g. population genetics studies) difficult regions can be ignored

Maximum depth



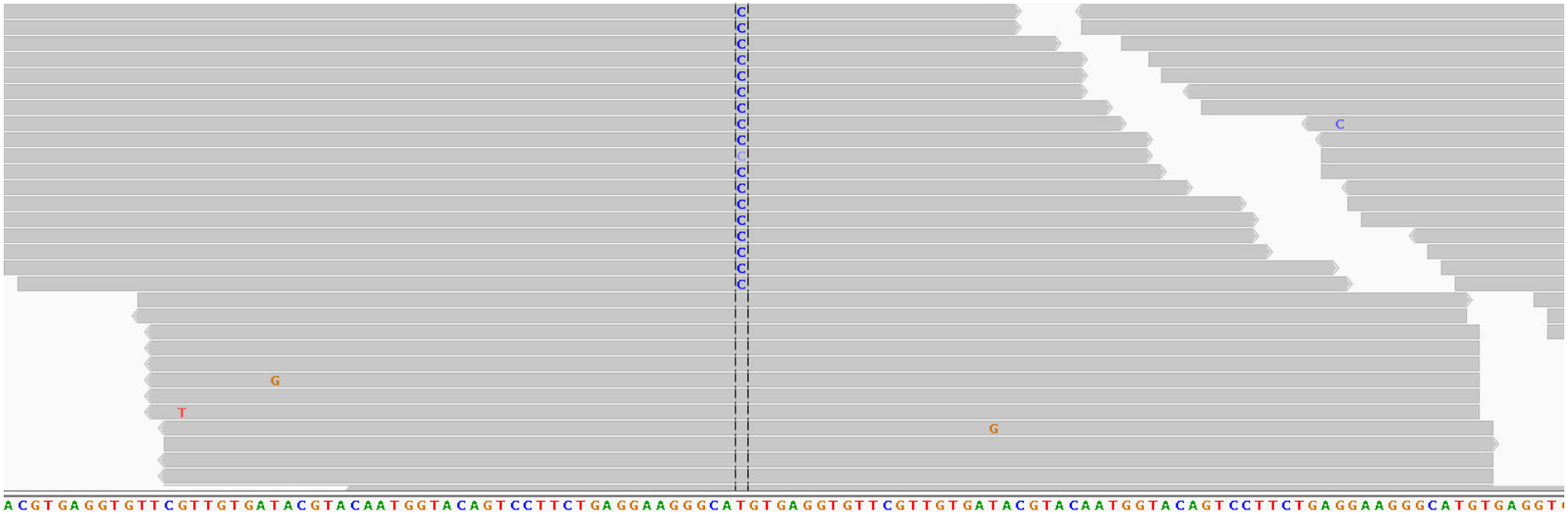
Q: Why is the sequencing depth thousandfold the average in some regions?

Mapping errors



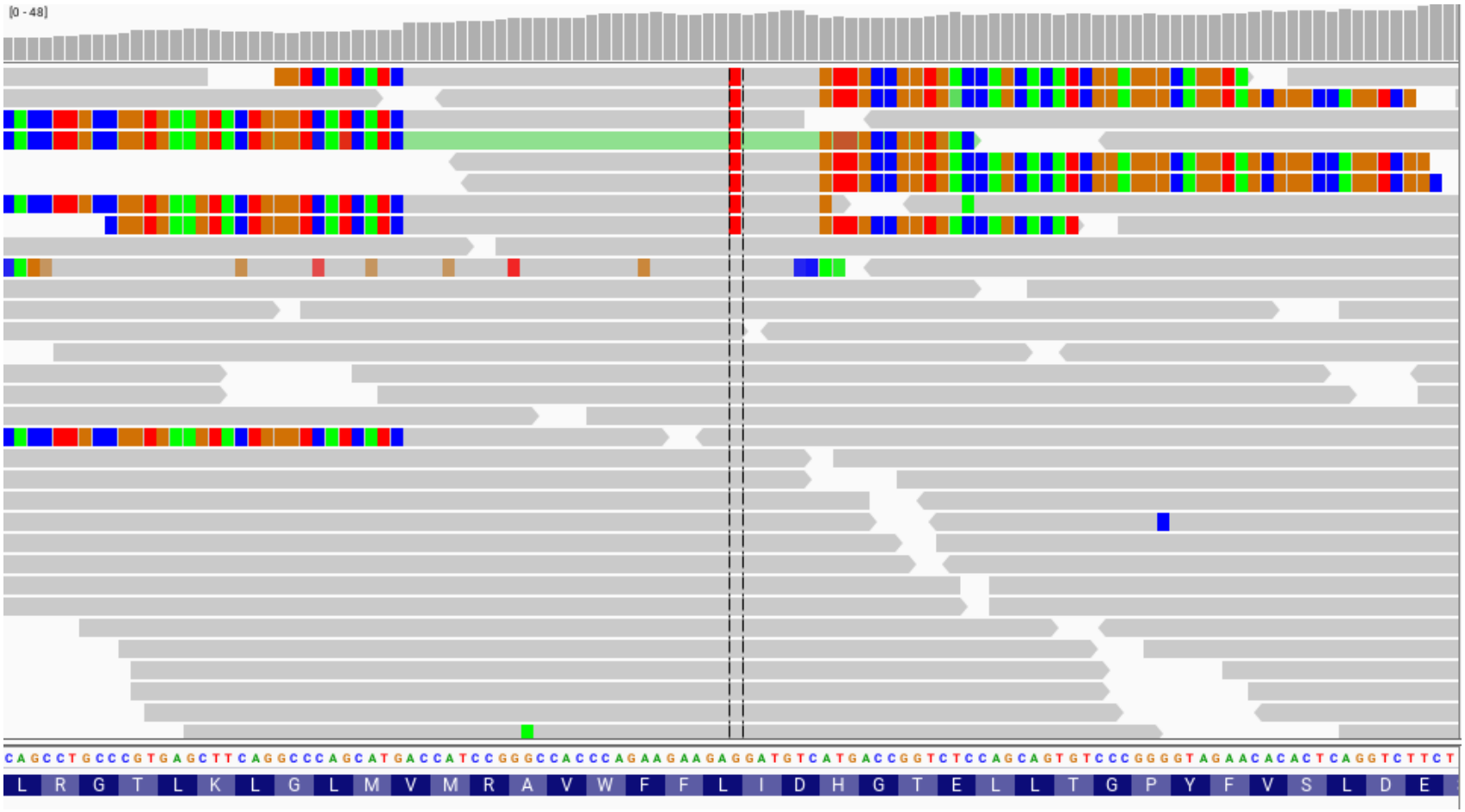
Q: RNA-seq (top) and DNA data (bottom) from the same sample has been mapped onto the reference genome. Can you explain the novel SNVs?

Strand bias



Q: Is this a valid call?

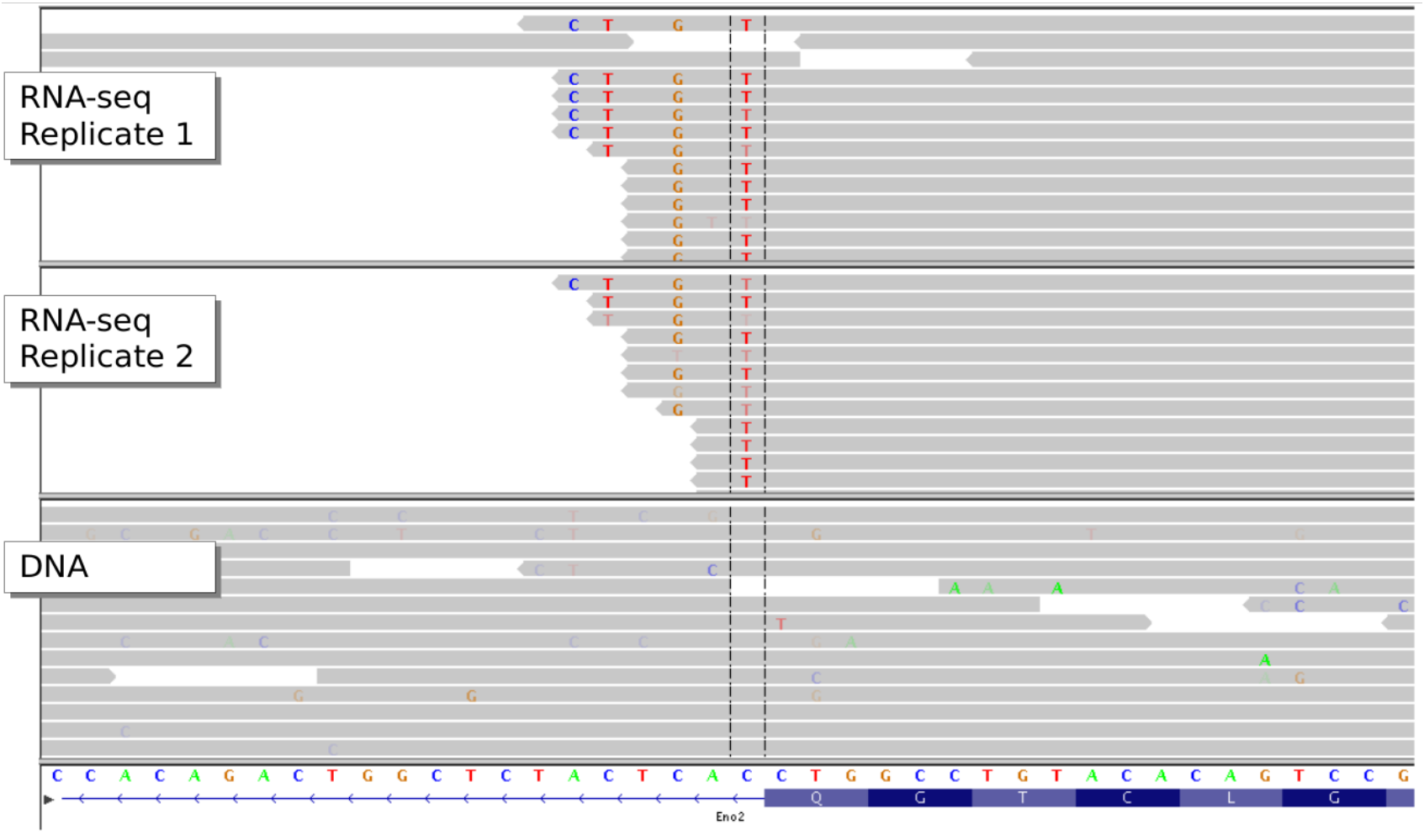
Change the display in IGV to reveal artefacts



Display soft-clipped bases...

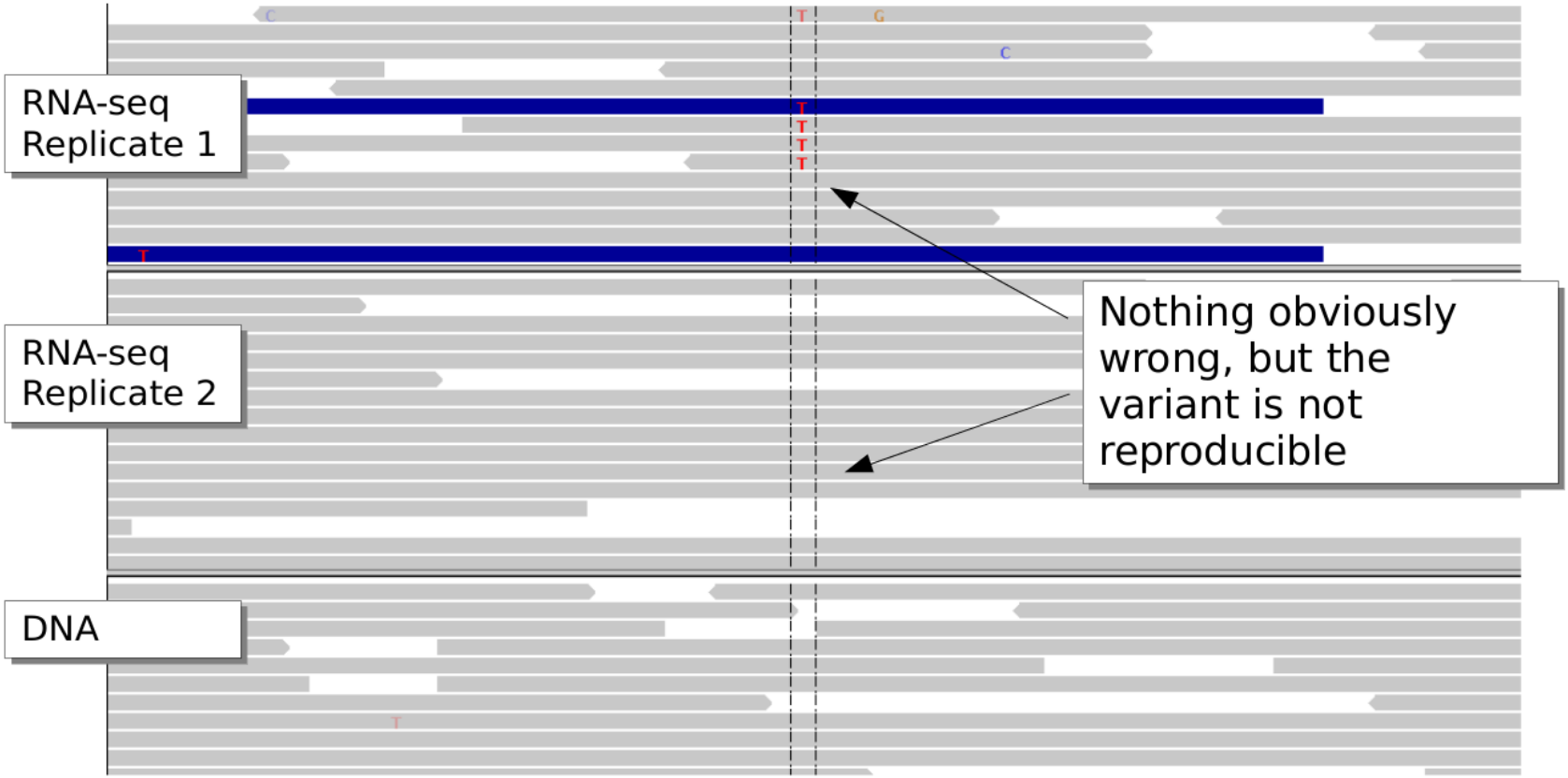
Too many soft-clipped reads in a region suggest mapping errors, beware!

Variant distance bias



Q: Can you explain what happened here?

Reproducibility



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

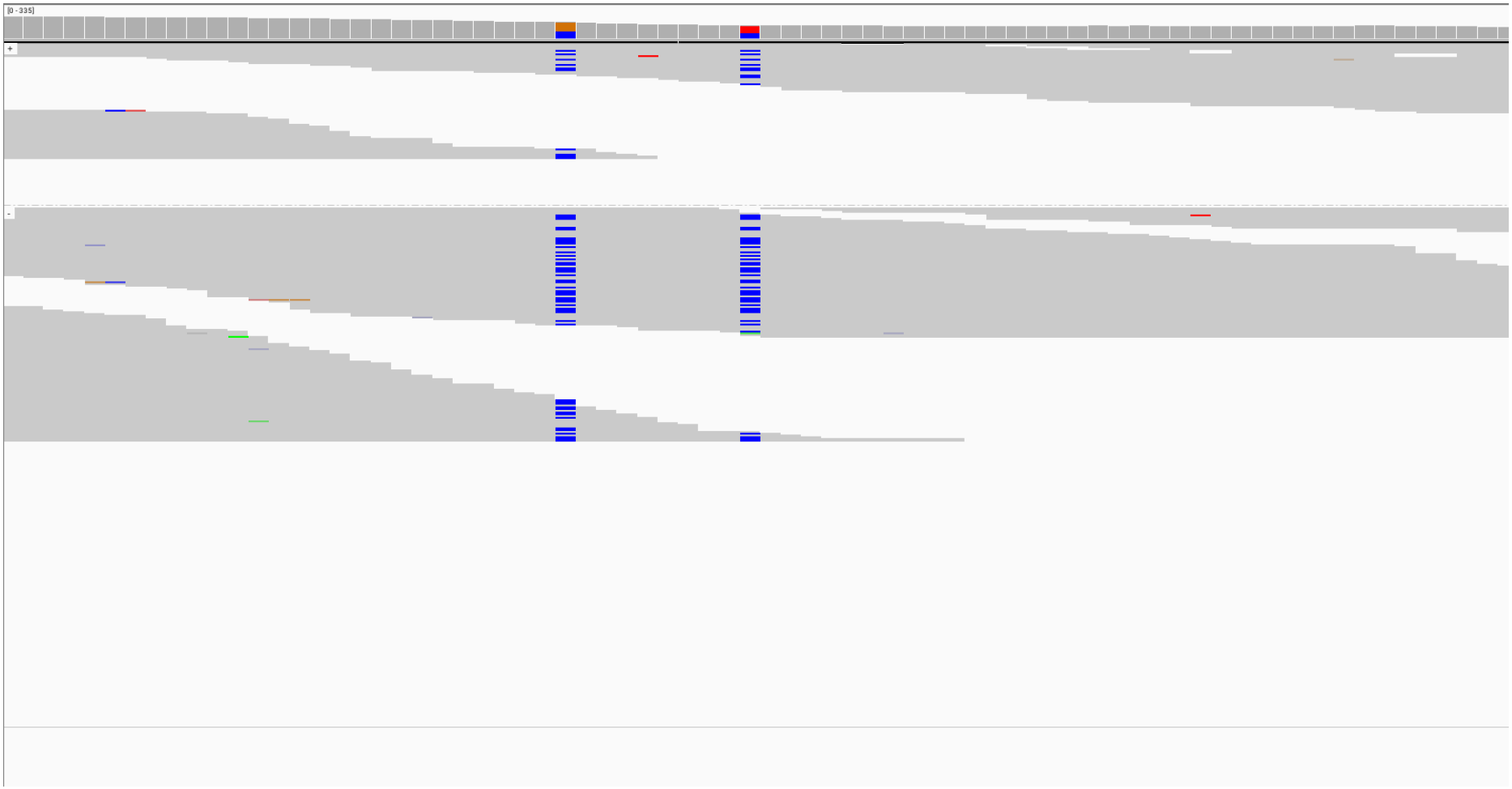
False SNPs caused by incorrect alignment

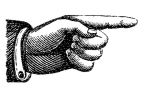
Pairwise alignemnt artefacts can lead to false SNPs

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

Q: How many SNPs are real?

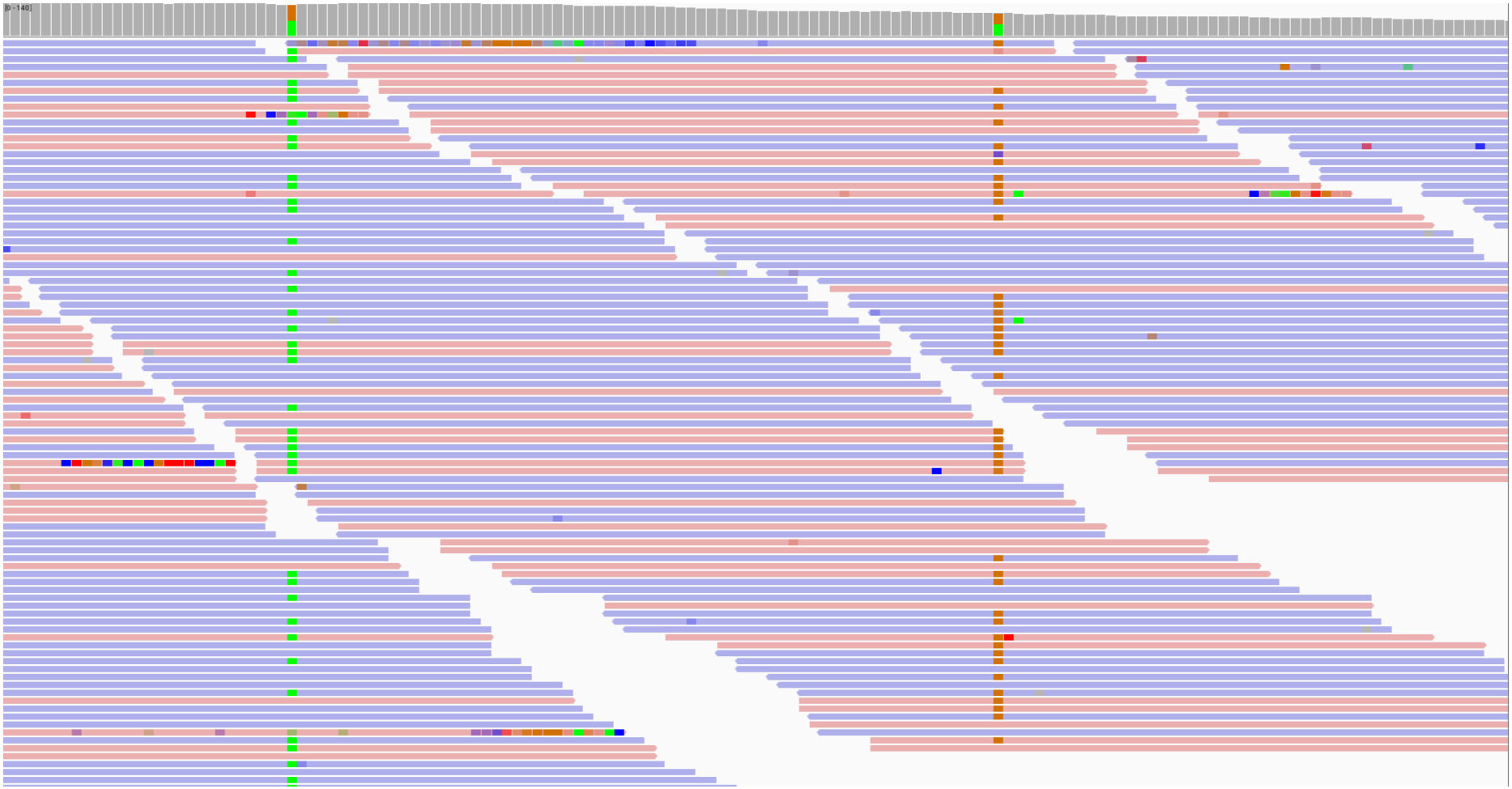
What good SNPs look like?

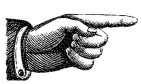




Change the view IGV to inspect possible biases. Here the reads were squished and grouped by read strand to confirm two clean unbiased calls.

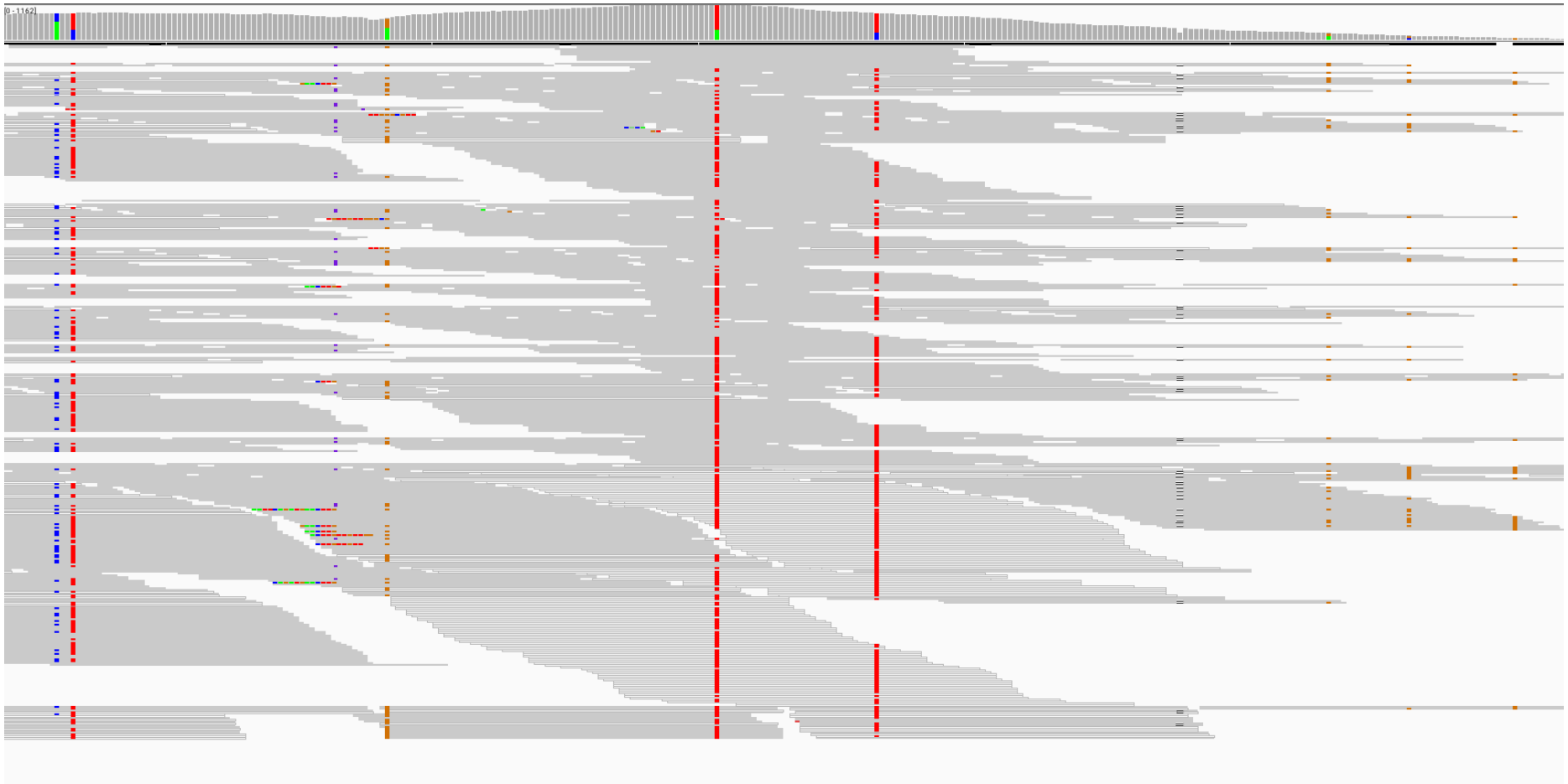
What good SNPs look like?



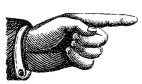


Change the view IGV to inspect possible biases. Here the reads were colord by read strand to confirm another two clean unbiased calls.

What good SNPs look like?



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

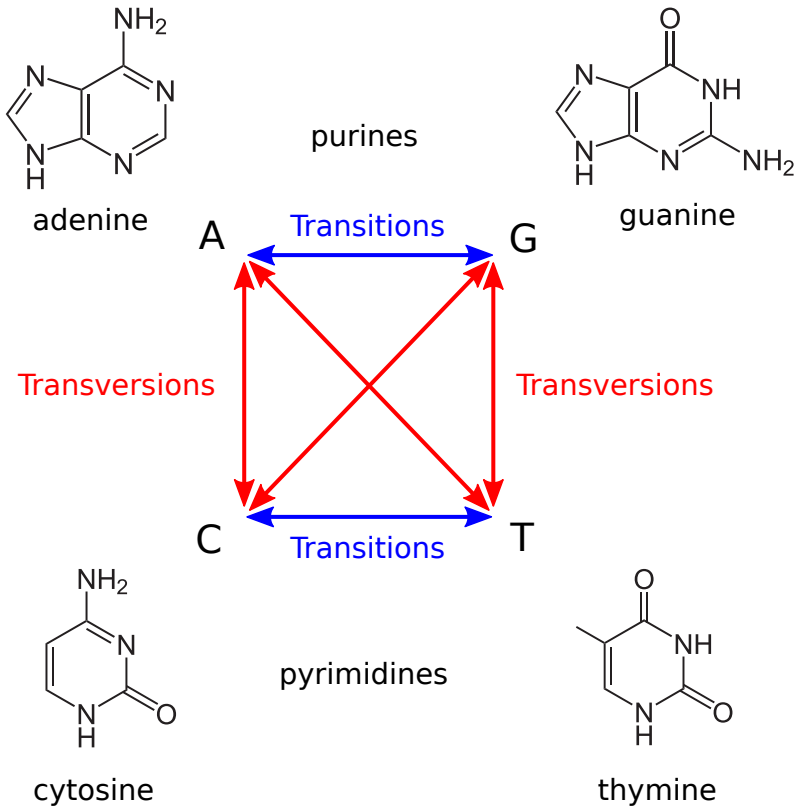


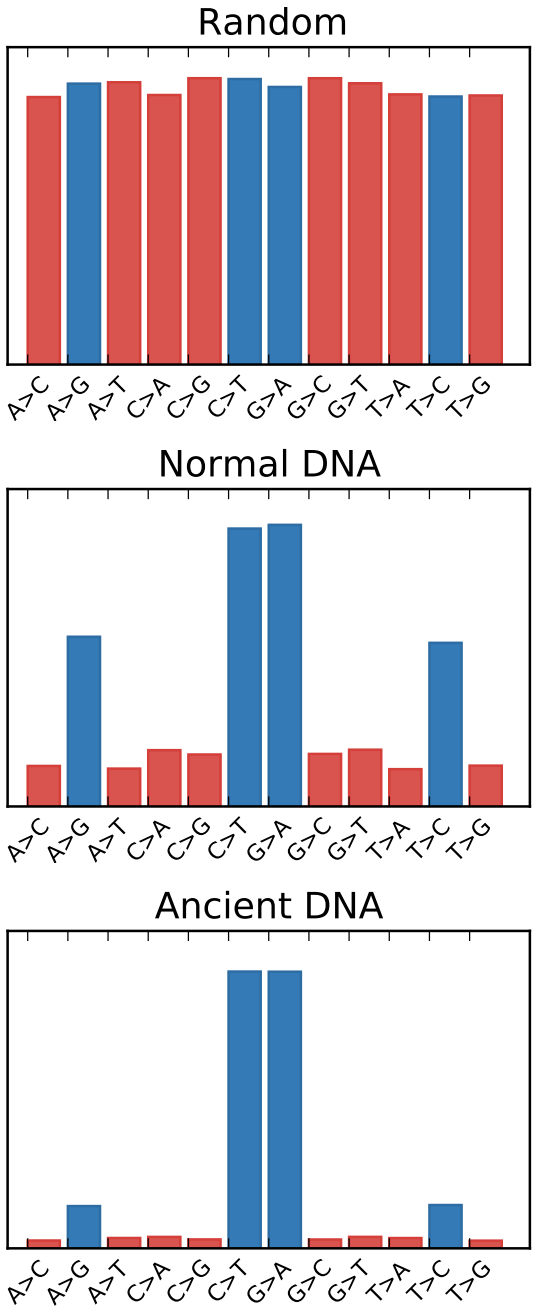
Sorting the reads by MQ reveals the variant is also supported by many high-quality reads.

How to estimate the quality of called SNPs?

Transitions vs transversions ratio, known as ts/tv

- transitions are 2-3× more likely than transversions





Indel calling challenges

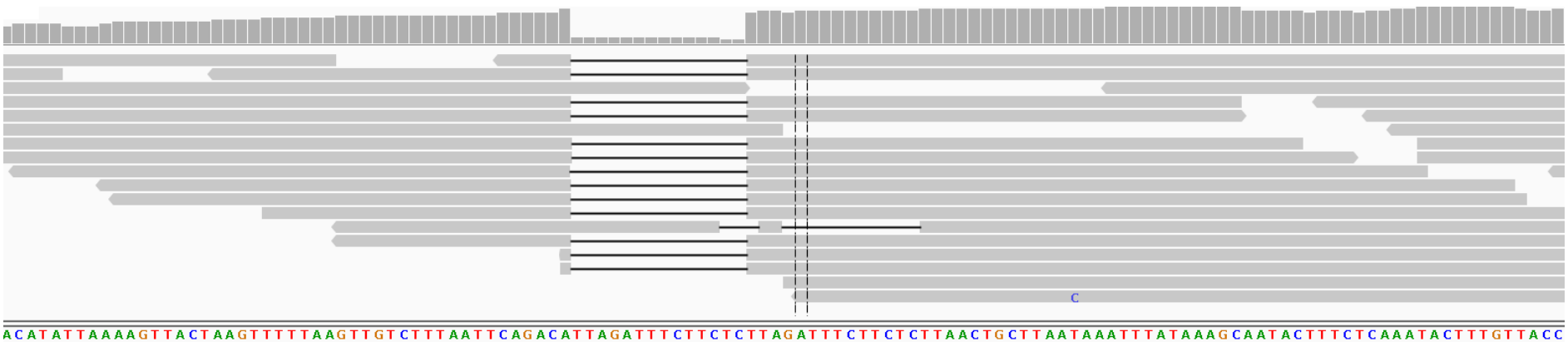
The sequencing error rate is elevated in microsatellites

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




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

CTTTAATTCAGACATTAGATTTCTTCTCTTAGATTTCTTCTCTTAACTGCTT
```

Future of variant calling

Current approaches

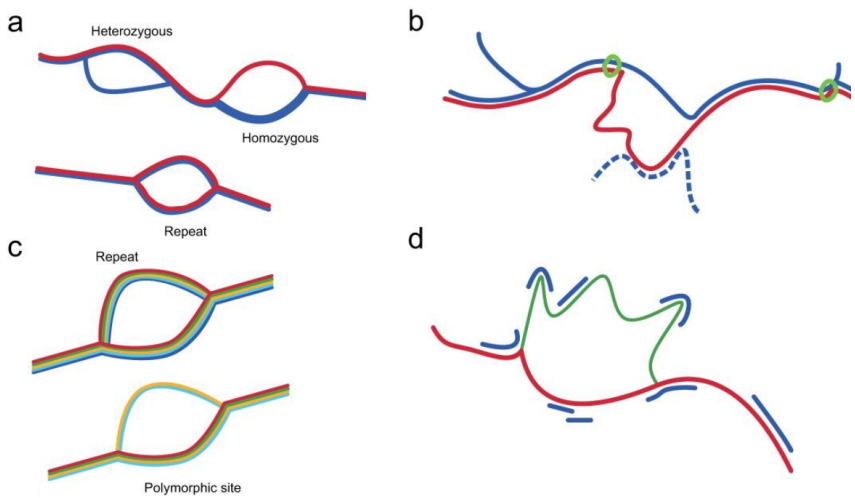
- rely heavily on the supplied alignment, but aligners see one read at a time
- largely site based, do not examine local haplotype and linked sites

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

Single vs multi-sample and gVCF calling

VCF files can be **very** big, therefore we often store only variant sites[1]

- however, variant-only VCFs are difficult to compare - was a site dropped because of a reference call or because of low coverage?
- we need evidence for both variant and non-variant positions in the genome

gVCF

- represents blocks of reference-only calls in a single record using the END tag
- symbolic allele in raw "callable" gVCFs allows to calculate genotype likelihoods only once (an expensive step), then do calling repeatedly as more samples come in

```
#CHROM POS  ID REF ALT  QUAL FILTER    INFO              FORMAT Sample
19    9902 .  G   <*>   .    .      END=9915;MinDP=0    PL:DP  0,0,0:0
19    9916 .  C   <*>   .    .      END=9922;MinDP=5    PL:DP  0,15,137:5
19    9923 .  G   <*>   .    .      END=9948;MinDP=10   PL:DP  0,30,214:10
19    9949 .  G   A,<*> .    .      DP=28               PL:DP  0,60,255,78,255,255:27
19    9950 .  C   <*>   .    .      END=9958;MinDP=28   PL:DP  0,84,255:28
19    9959 .  G   T,<*> .    .      DP=34               PL:DP  0,82,255,99,255,255:34
19    9960 .  C   <*>   .    .      END=9969;MinDP=34   PL:DP  0,102,255:34
```

Symbolic "unobserved" allele
Represents any other possible alternate allele

A block of 10 sites with
at least 34 reference reads

Genotype likelihoods
for CC, C*, **

[1] Annotated VCF with 3,781 samples, variant sites only, UK10k project . . . 680GB

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