Variant Calling - SNPs and short indels

petr.danecek@sanger.ac.uk

wellcome trust sanger institute

Variant types

SNPs/SNVs ... Single Nucleotide Polymorphism/Variation

```
ACGTTTAGCAT
ACGTTCAGCAT
```

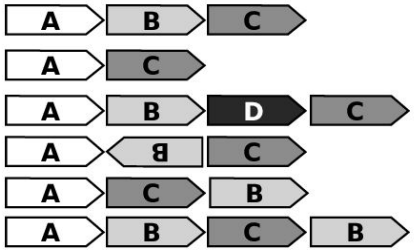
MNPs ... Multi-Nucleotide Polymorphism

```
ACGTCCAGCAT
ACGTTTAGCAT
```

Indels ... short insertions and deletions

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

SVs ... Structural Variation



Germline vs somatic mutation

Germline mutation

- heritable variation in the germ cells

Somatic mutation

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

Germline vs somatic mutation

Germline mutation

- heritable variation in the germ cells

Somatic mutation

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

Germline variant calling

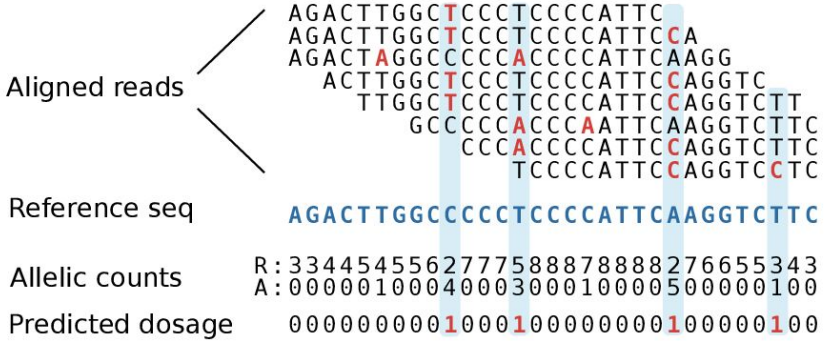
- expect the following fractions of alternate alleles in the pileup:
  - 0.0 for RR genotype (plus sequencing errors)
  - 1.0 for AA (plus sequencing errors)
  - 0.5 for RA (random variation of binomial sampling)

Somatic

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

Naive variant calling

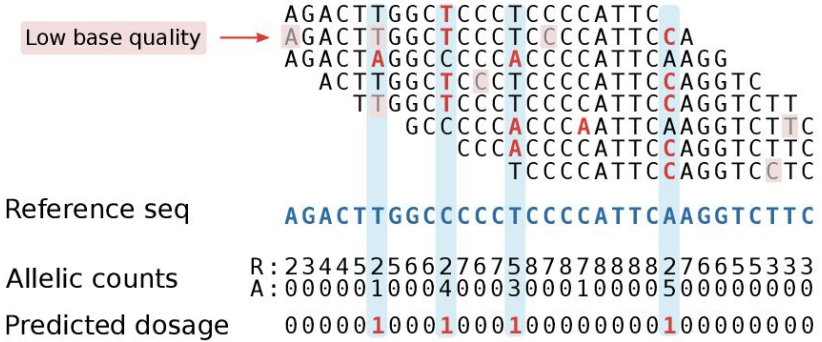
Use fixed allele frequency threshold to determine the genotype



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

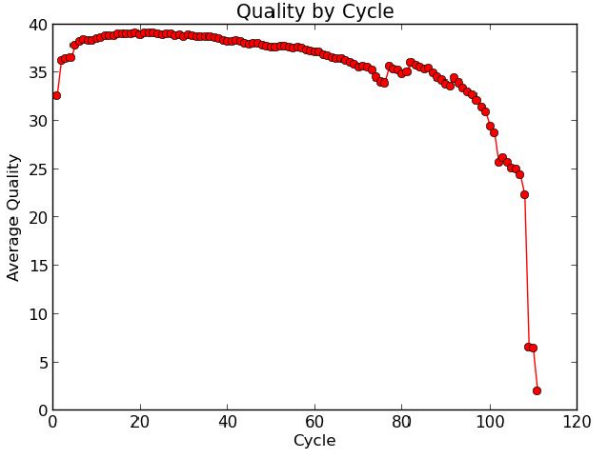


1) Filter base calls by quality
   e.g. ignore bases Q<20

Phred quality score

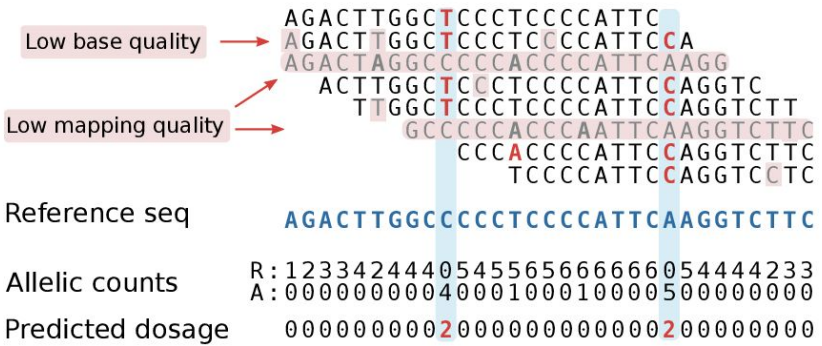
$$Q = -10 \log_{10} P_{\text{err}}$$

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



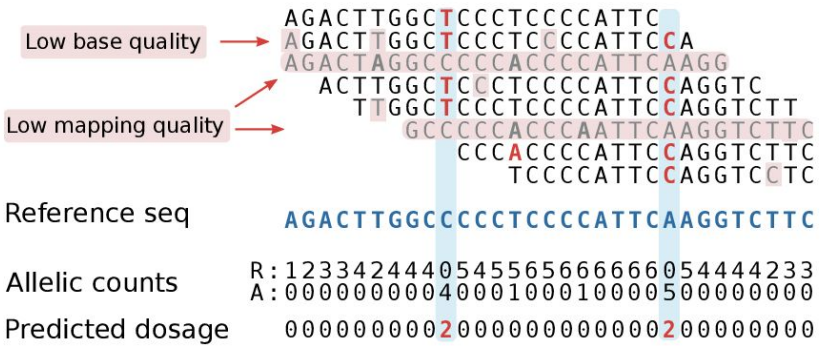
1) Filter base calls by quality
   e.g. ignore bases Q<20

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

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

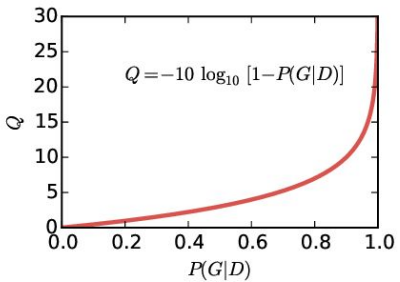
to determine:

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

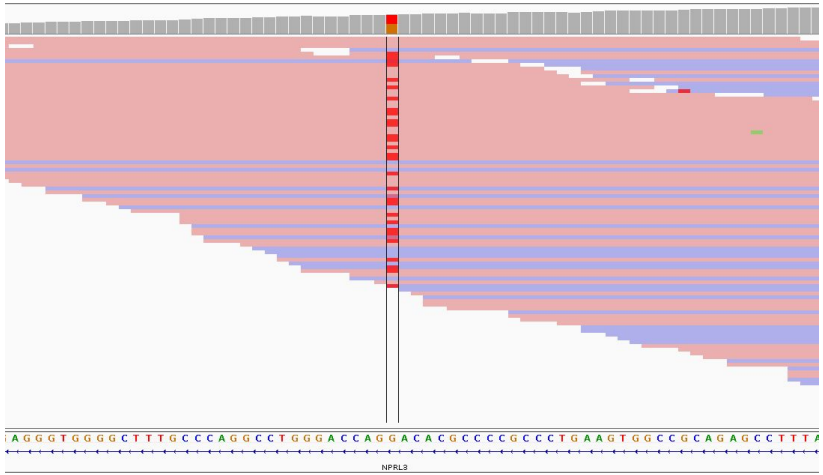
Factors to consider in calling

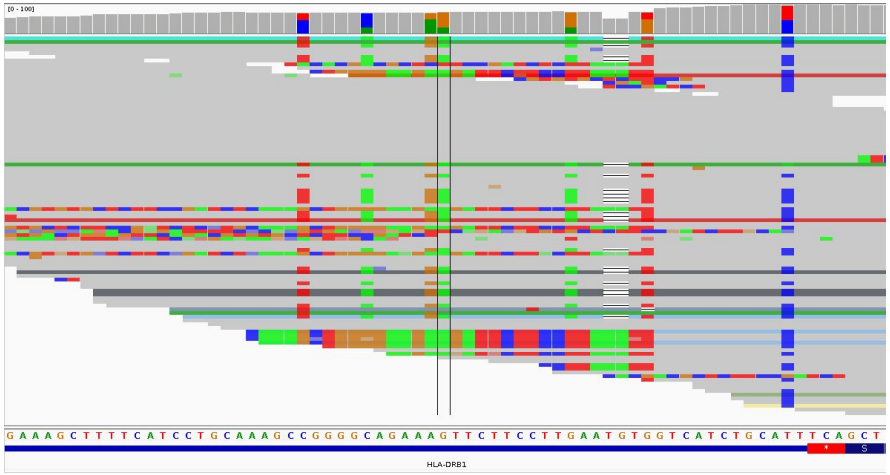
Many calls are not real, a **filtering** step **is necessary**

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

The good, the bad, and the ugly





HLA-DRB1 is a gene with 13 paralogs

Finding a needle in a needle stack

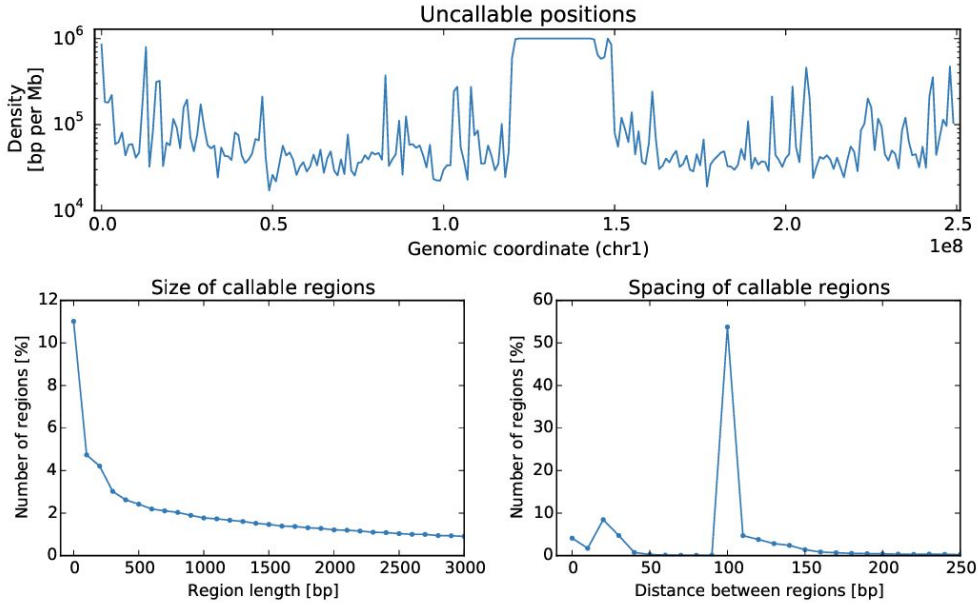
3,494,429 SNVs and short indels

- which ones are real?
- which ones are causal?

Callable genome

Large parts of the genome are still inaccessible

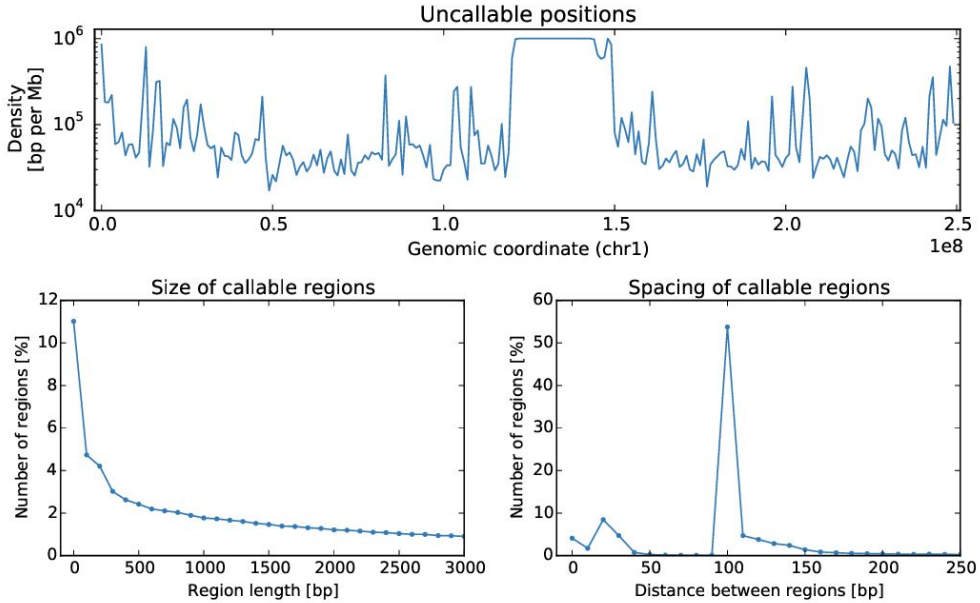
- the Genome in a Bottle high-confidence regions:
  - cover 89% of the reference genome
  - are short intervals scattered across the genome



Callable genome

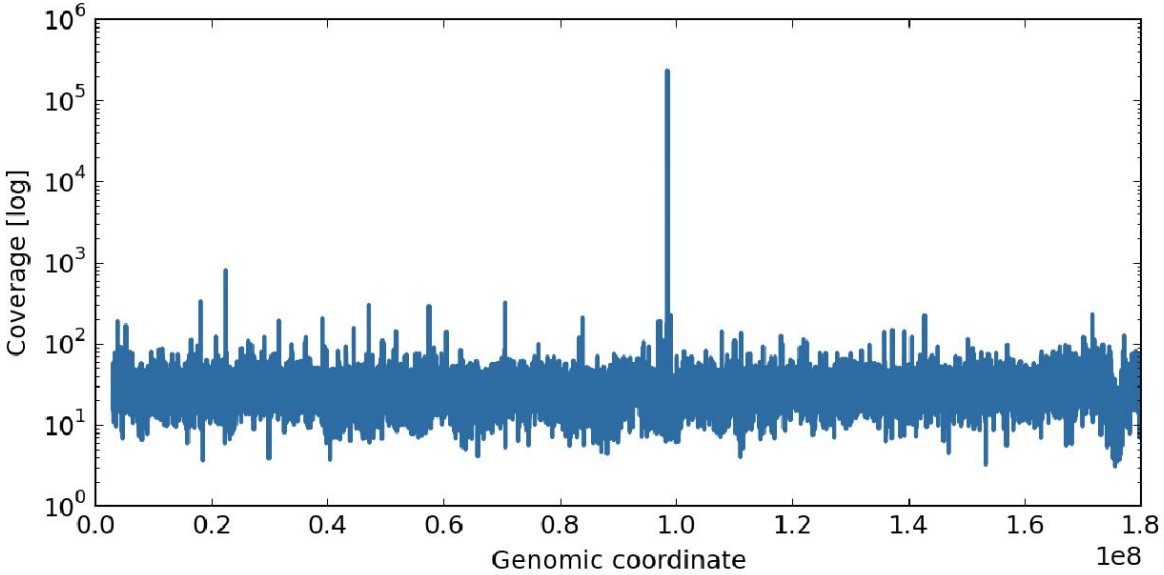
Large parts of the genome are still inaccessible

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



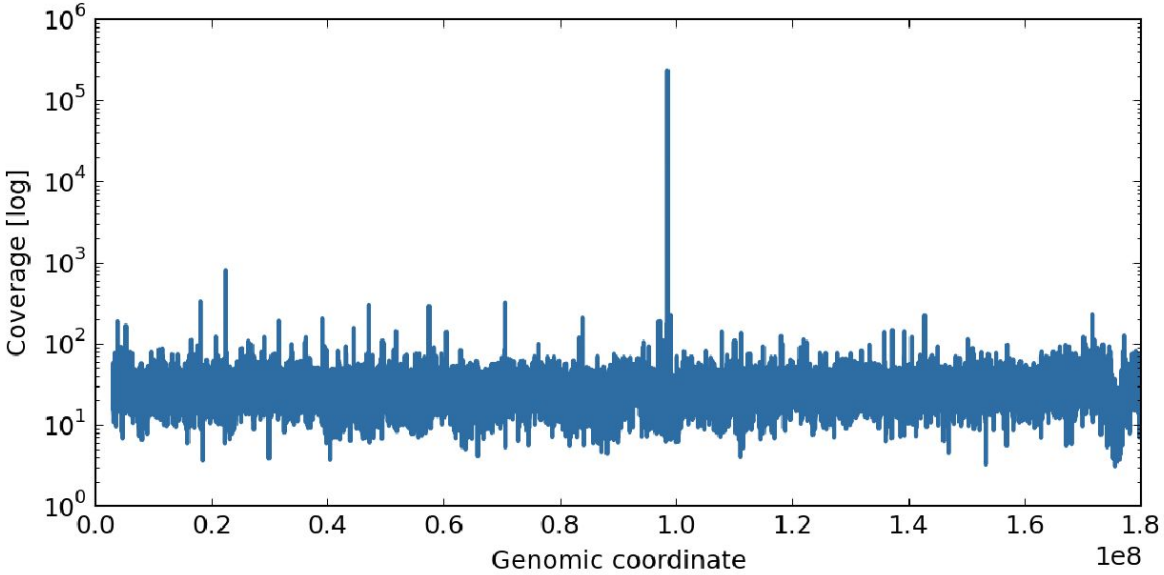
If possible, include only "nice" regions: for many analyses (e.g. population genetics studies) difficult regions can be ignored

Maximum depth



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

Maximum depth



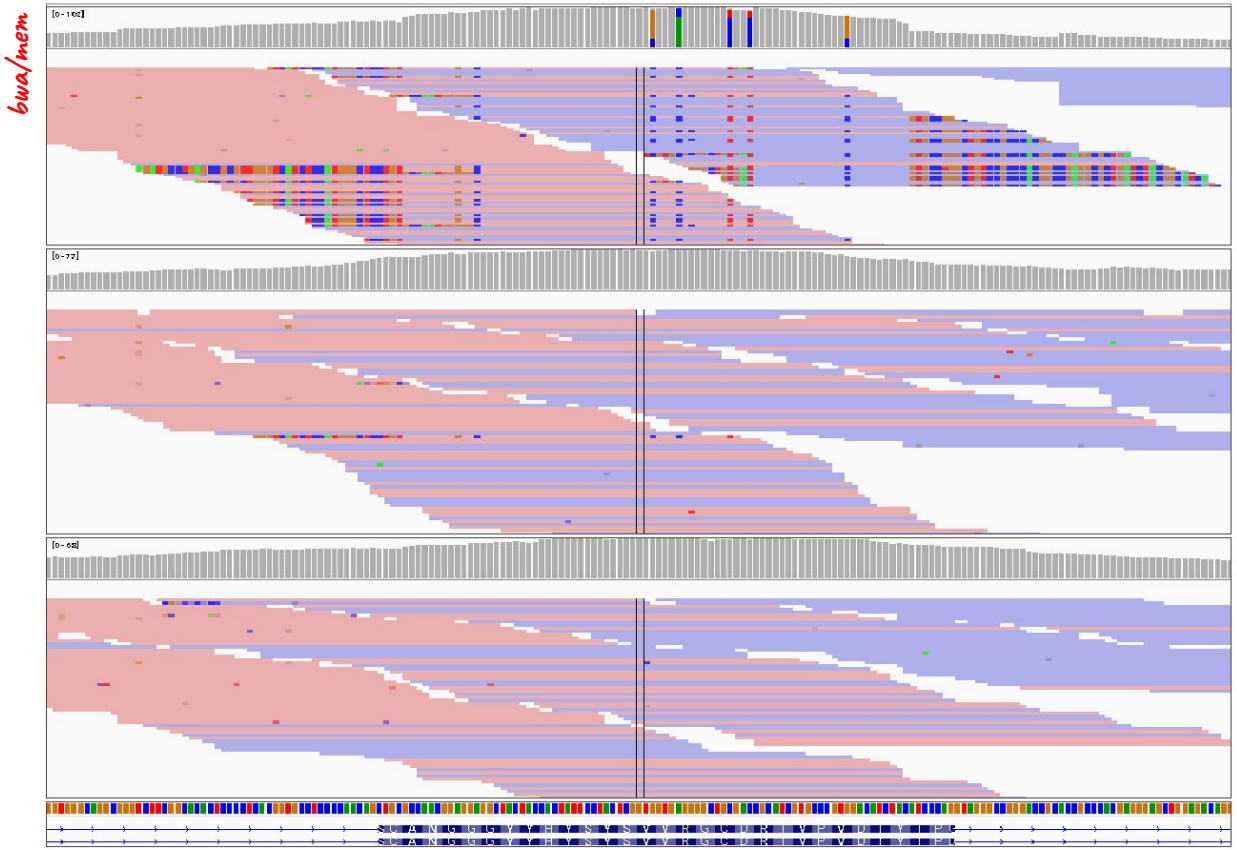
Q: Why is the sequencing depth thousandfold the average in some regions?

A: The reference genome is not complete. This sample was sequenced to 30x coverage, we can infer it has $\sim$ 30 copies of this region.

Filter calls with a too high depth (for example, 2x the average in WGS)

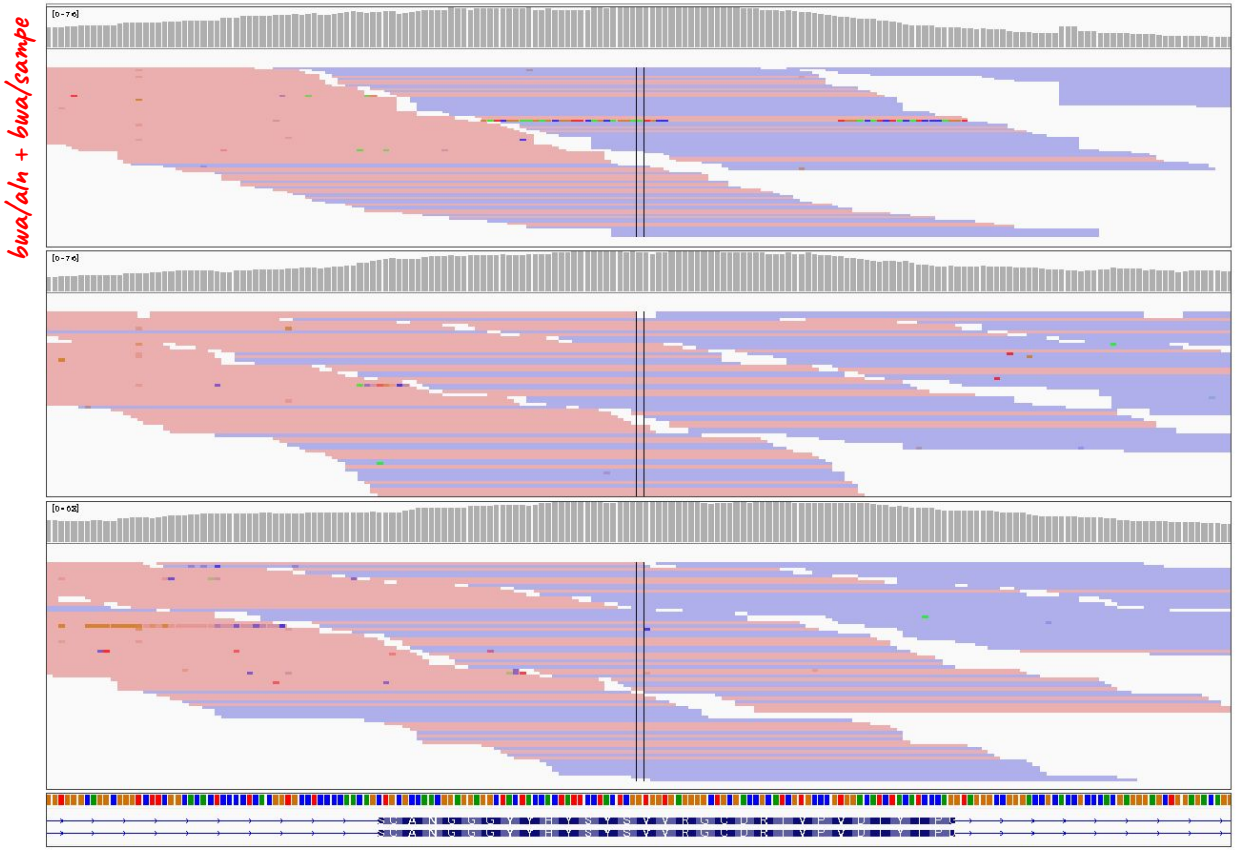
Mapping errors

Different mapping algorithm can lead to different results

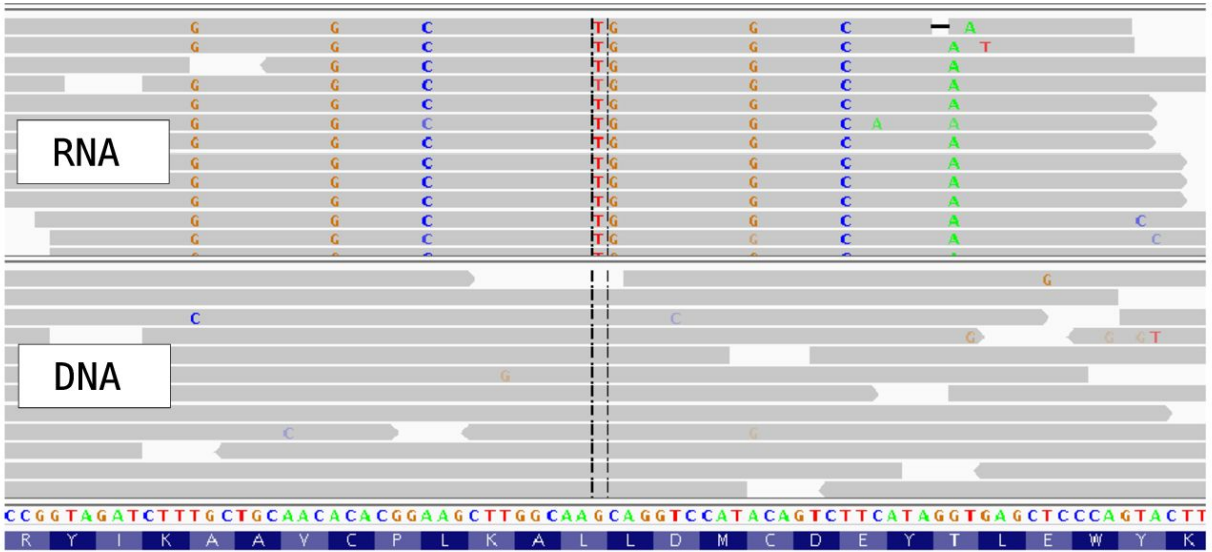


Mapping errors

Different mapping algorithm can lead to different results

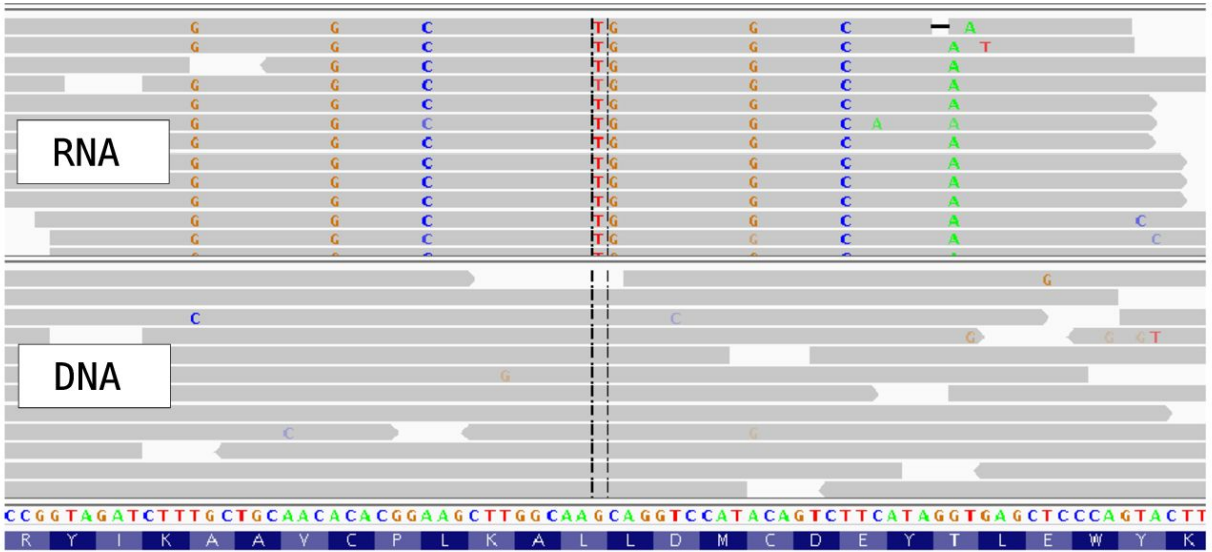


Mapping errors



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?

Mapping errors

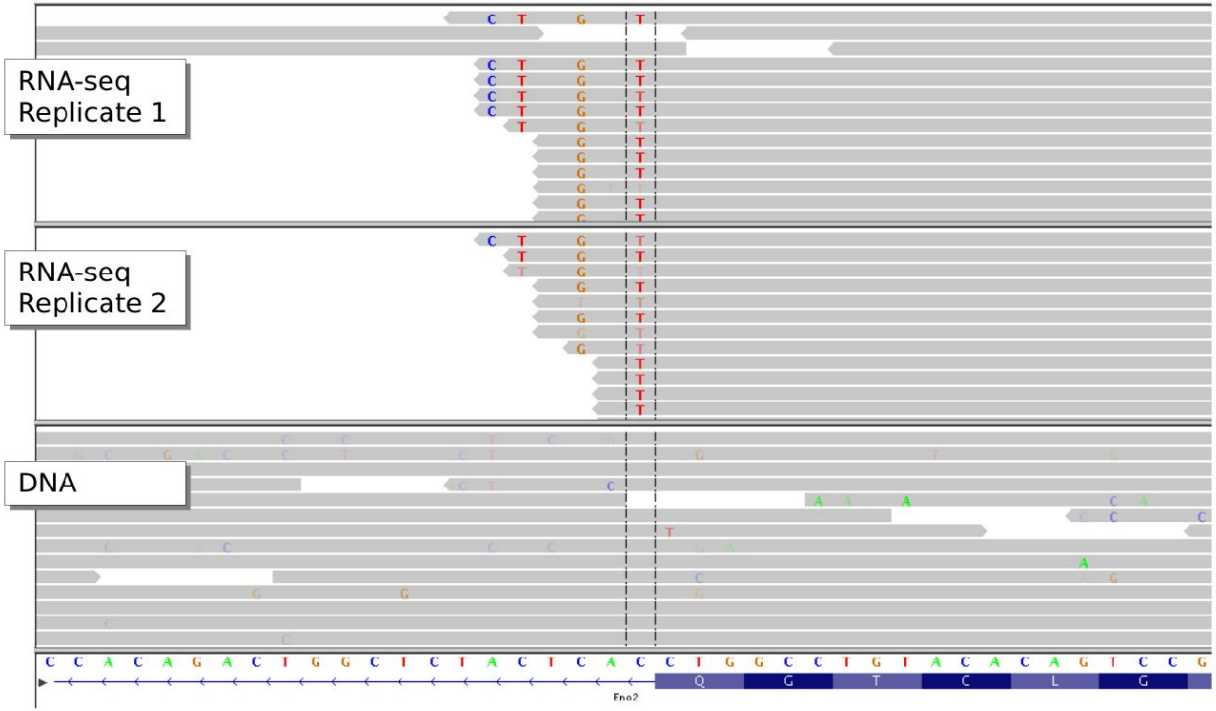


Q: RNA-seq (top) and DNA data (bottom) from the same sample has been mapped onto the reference genome. Can you explain the novel SNVs?

A: The reads were mapped to a pseudogene and originate in a paralog with 92% identity.

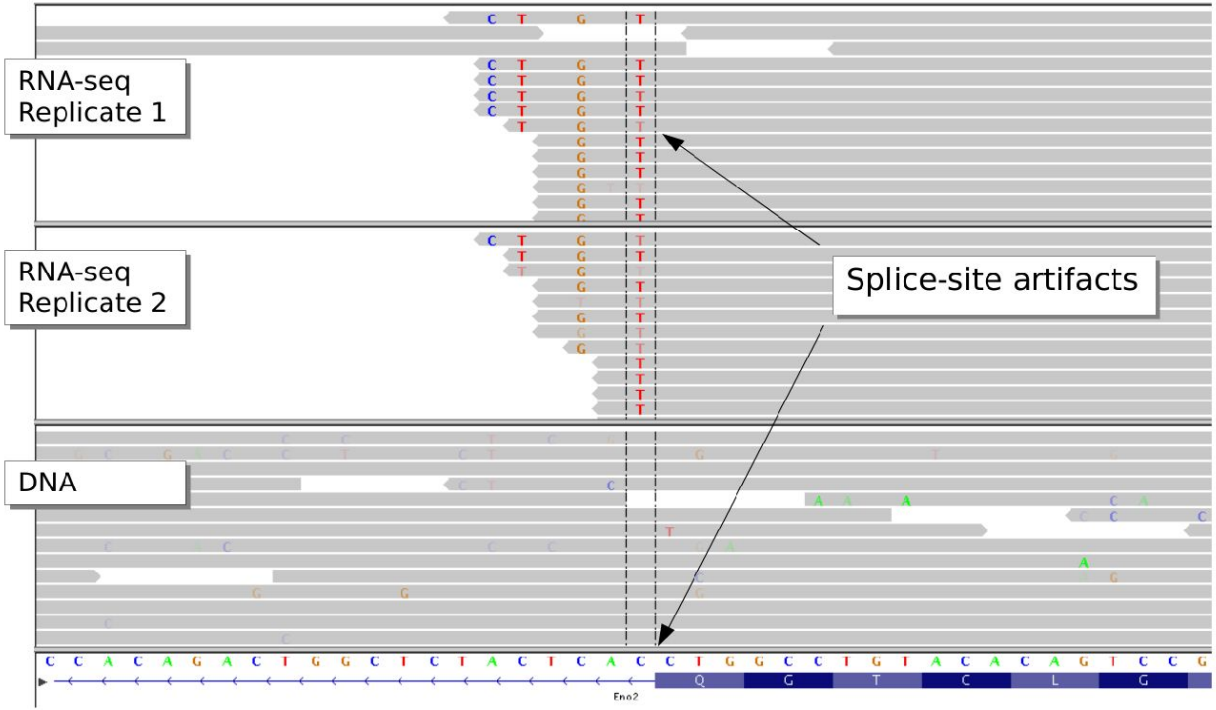
Beware of mapping errors, especially when aligning RNA-seq data on the genome.

Variant distance bias



Q: Can you explain what happened here?

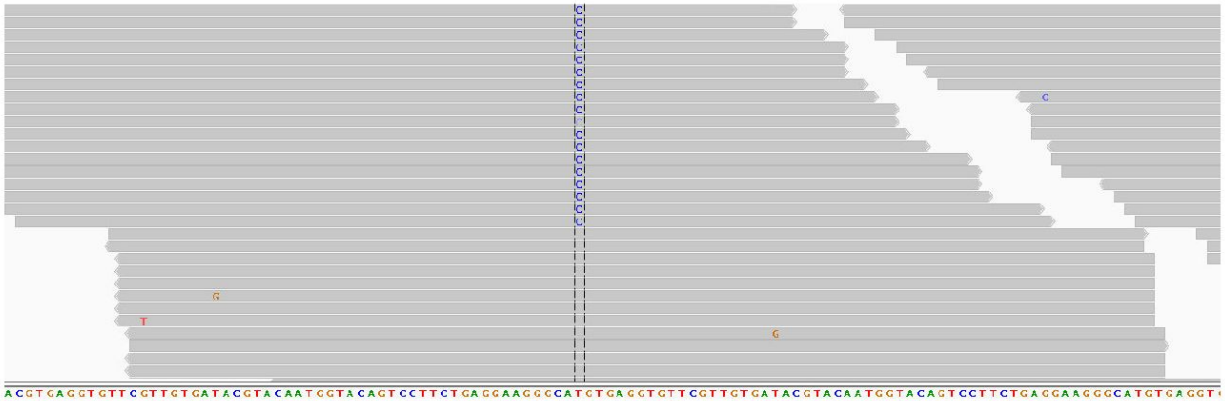
Variant distance bias



Q: Can you explain what happened here?
A: Processed transcript with introns spliced out.

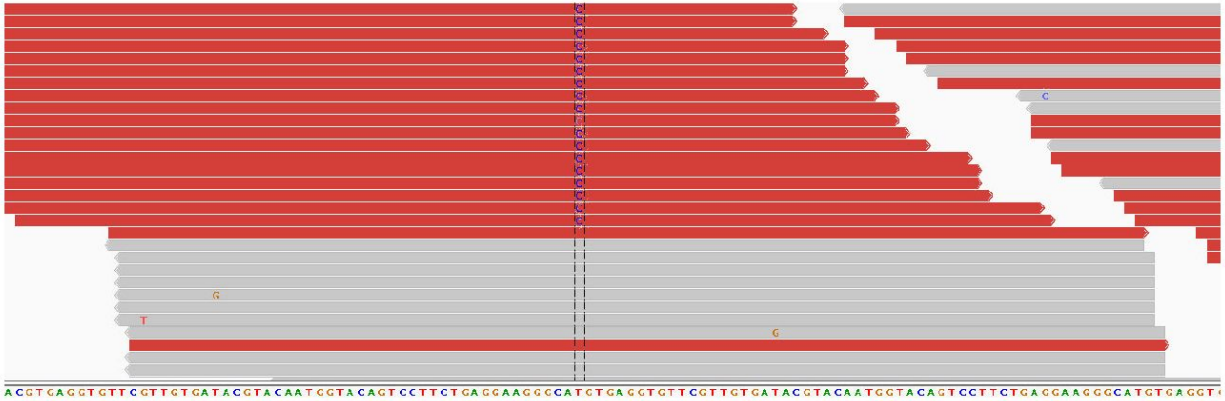
Better to use a splice-aware mapper when working with RNA-seq data, or filter most extreme cases using annotations such as VDB

Strand bias



Q: Is this a valid call?

Strand bias

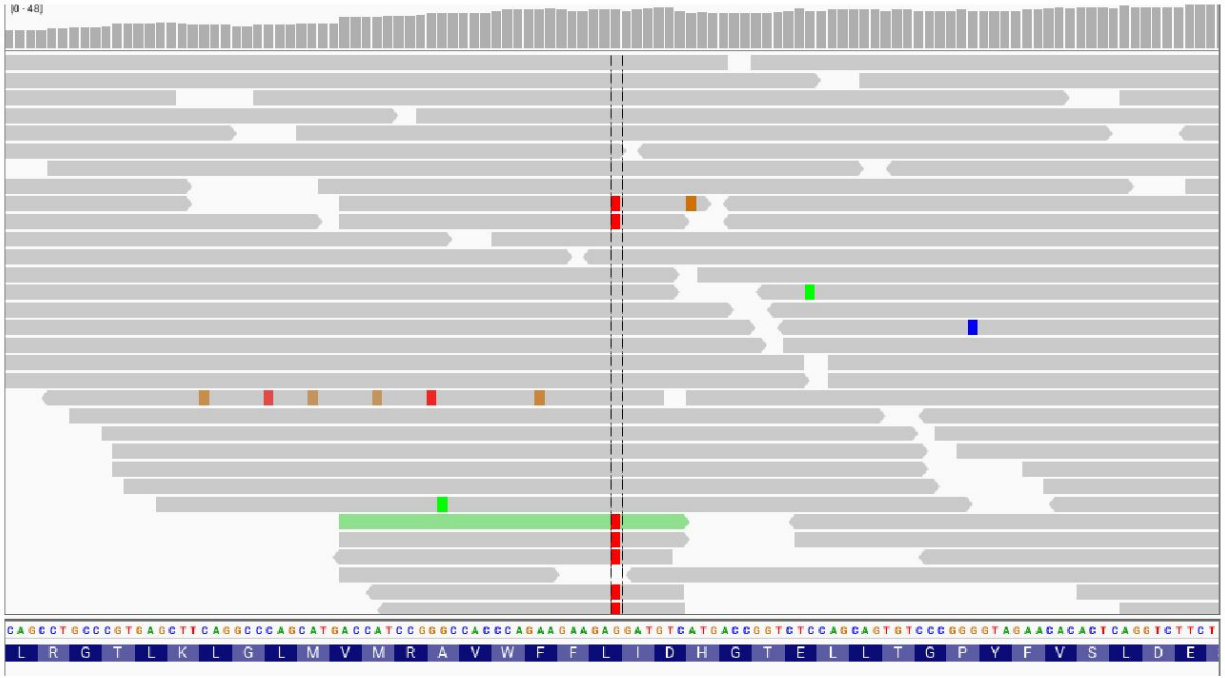


Q: Is this a valid call?

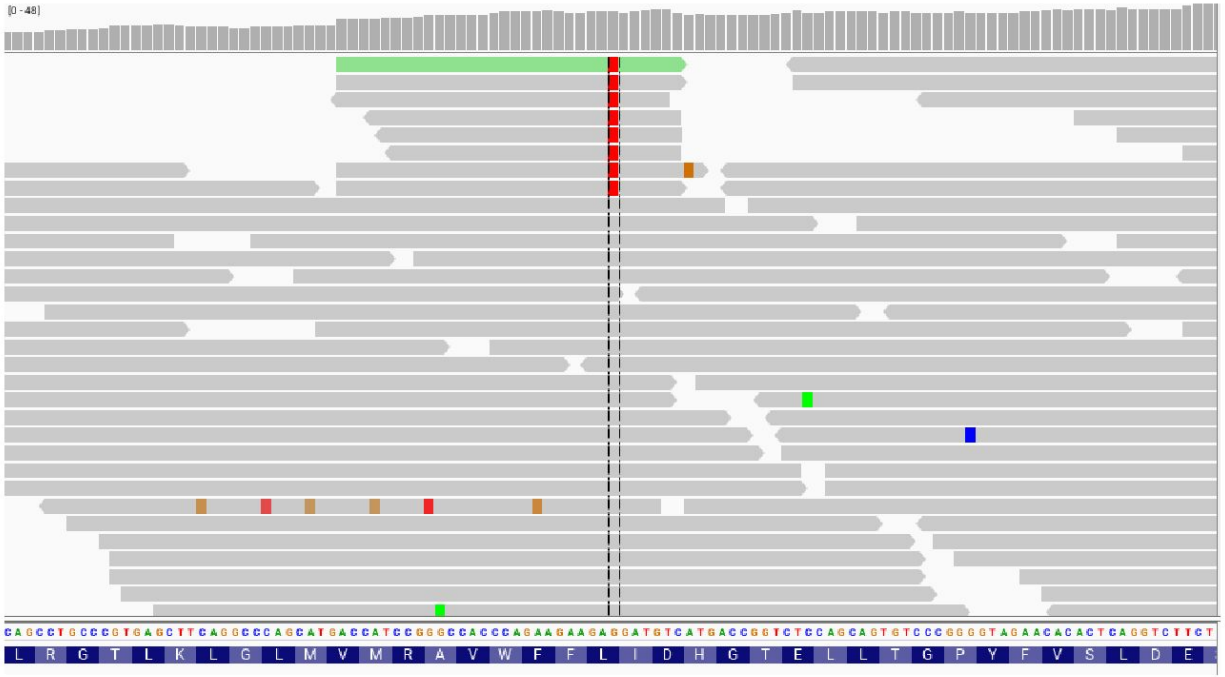
A: No, it is a mapping artefact, the call is supported by forward reads only.

Filter extremely biased calls using annotations generated by your caller (e.g. Fisher or rank-sum test)

Change the display in IGV to reveal artefacts

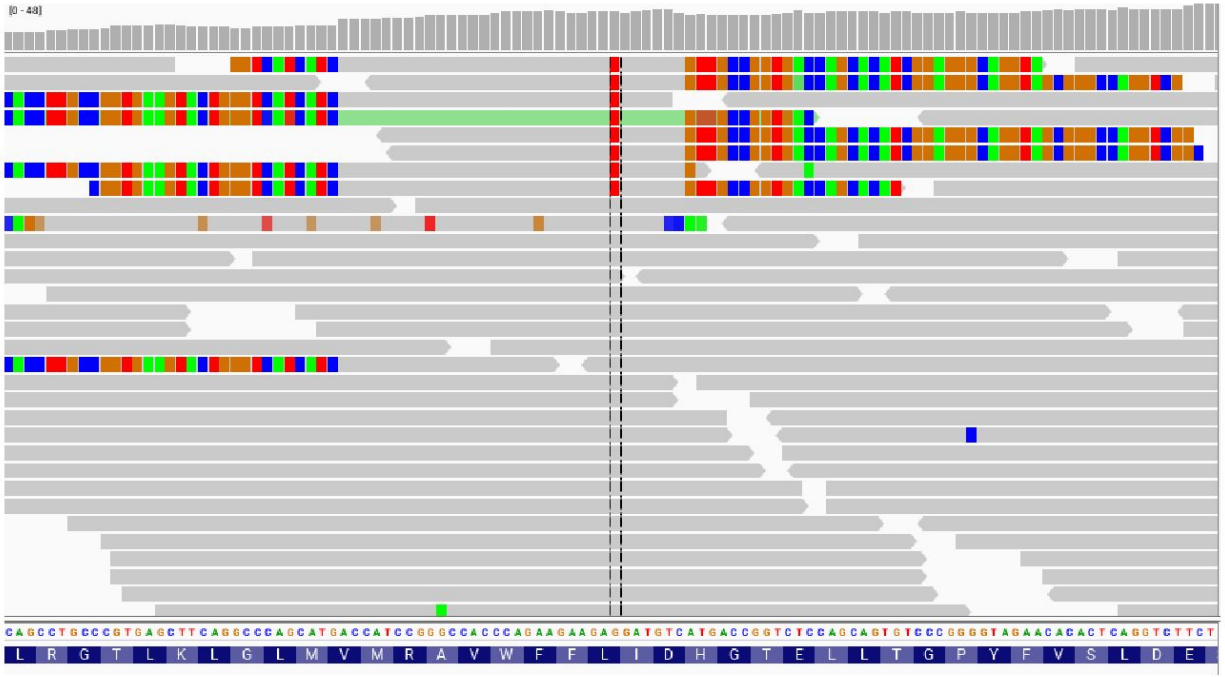


Change the display in IGV to reveal artefacts



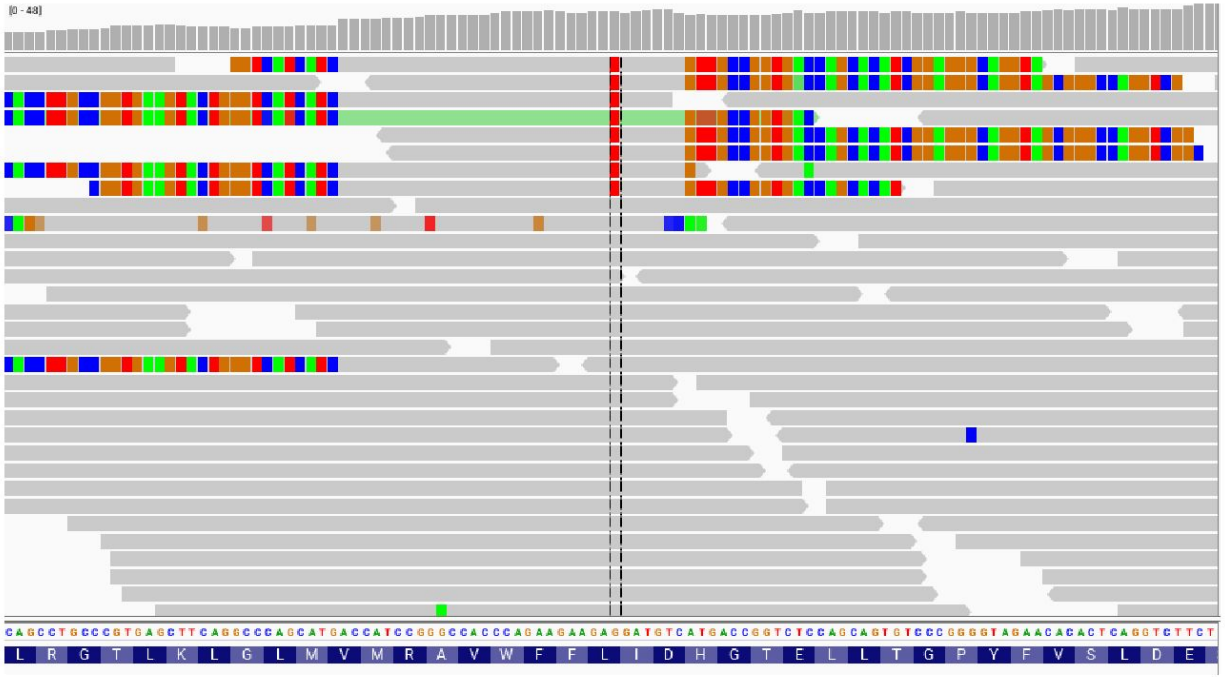
Sort by base...

Change the display in IGV to reveal artefacts



Display soft-clipped bases...

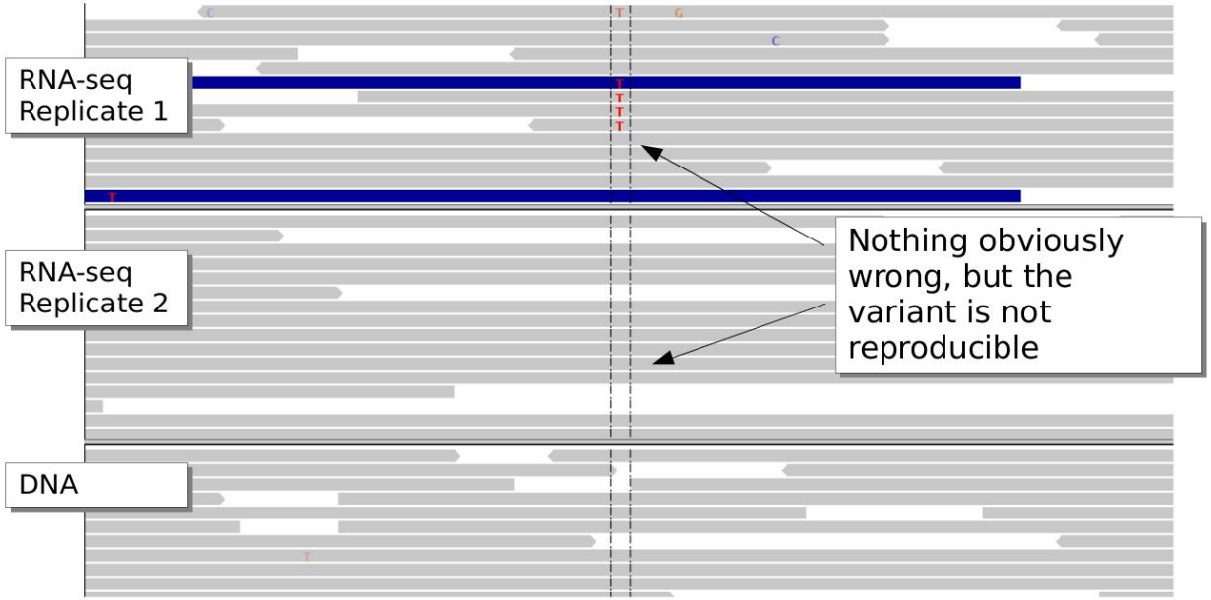
Change the display in IGV to reveal artefacts



Display soft-clipped bases...

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

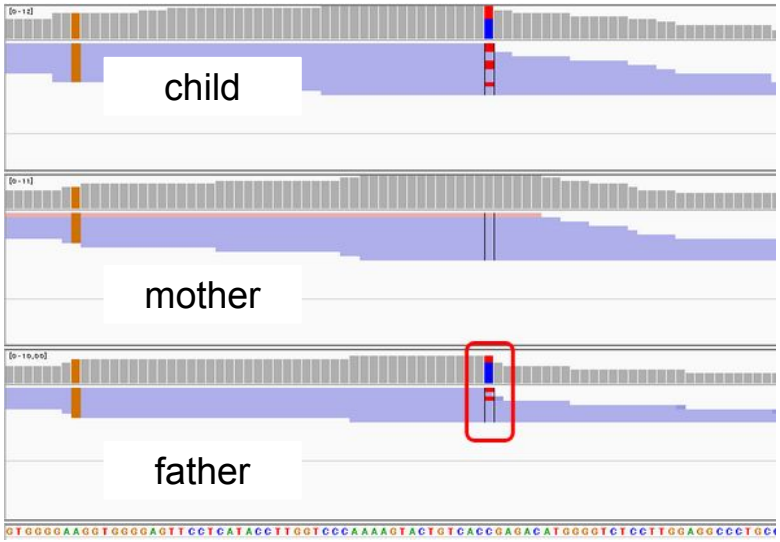
Reproducibility



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

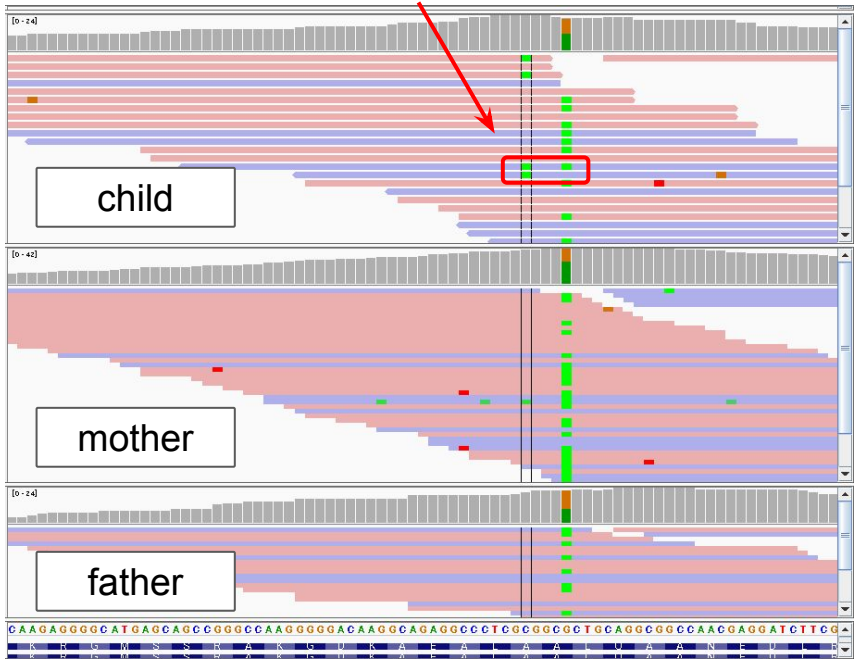
De novo mutations

Not a real DNM



Haplotype consistency

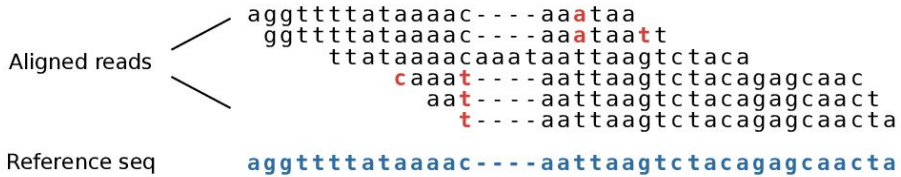
Both chromosomal copies affected, very likely a false positive!



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

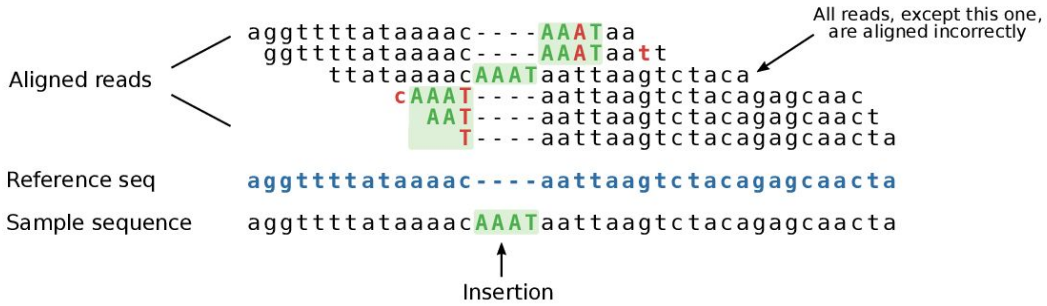


Q: How many SNPs are real?

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?

A: None.

Be careful when looking at SNPs close to indels.

Indel calling challenges

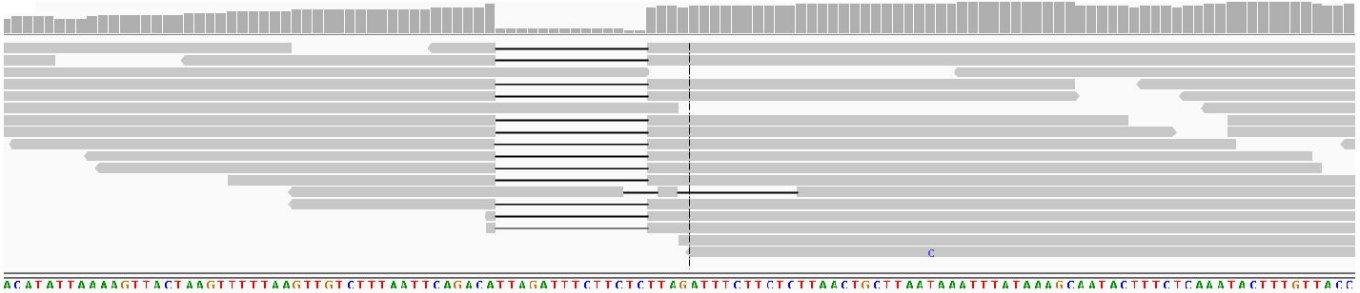
The sequencing error rate is elevated in microsatellites

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
```

Indel calling challenges

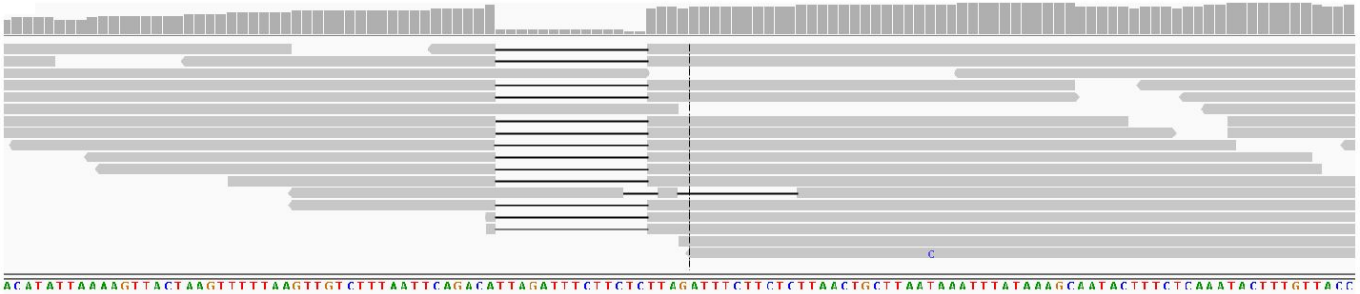
The sequencing error rate is elevated in microsatellites

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

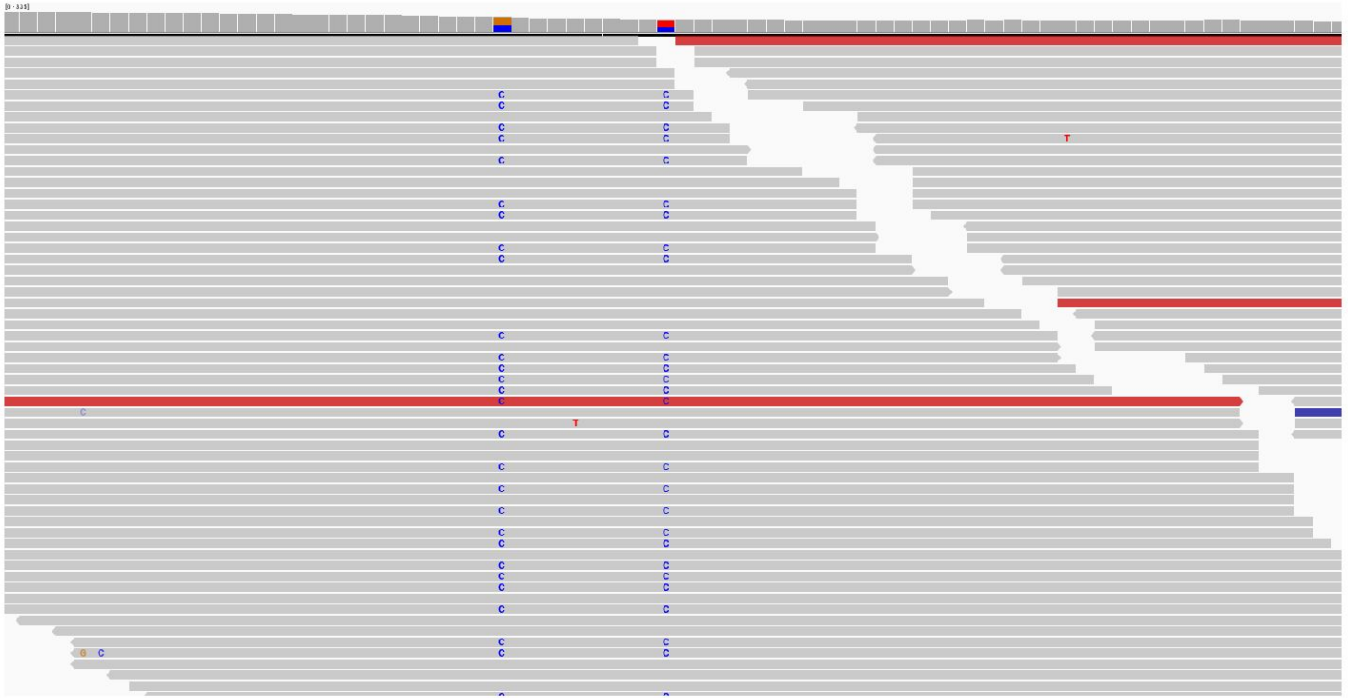
- the aligner can prefer multiple mismatches rather than a gap
- indel representation can be ambiguous



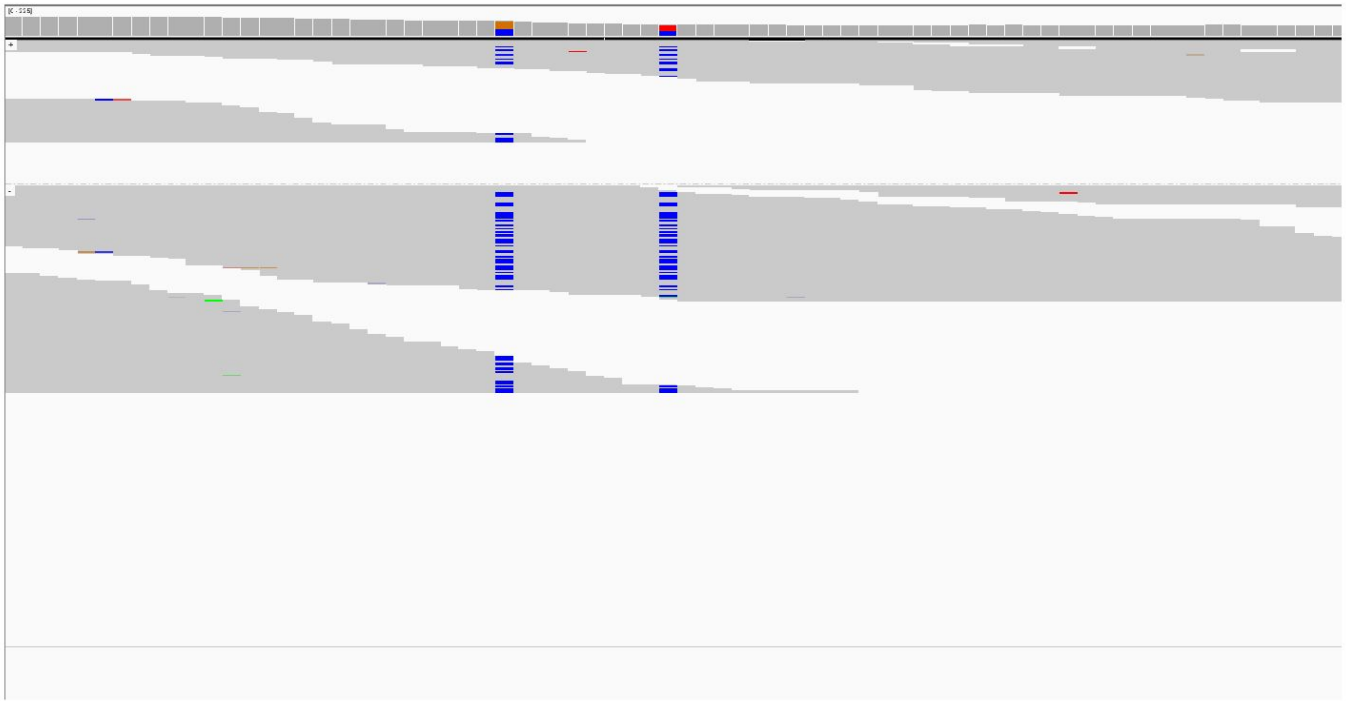
```
CTTTAATTCAGACA~~~~~~~~~~~~~~~TTAGATTTCTTCTC
CTTTAATTCAGACA~~~~~~~~~~~~~~~TTAGATTTCTTCTCTTA
CTTTAATTCAGACA---------------TTAGATTTCTTCTCTTAACTGCTT
CTTTAATTCAGACA~~~~~~~~~~~~~~~TTAGATTTCTTCTATTAACTGCTT

CTTTAATTCAGACATTAGATTTCTTCTCTTAGATTTCTTCTCTTAACTGCTT
```

What good SNPs look like?

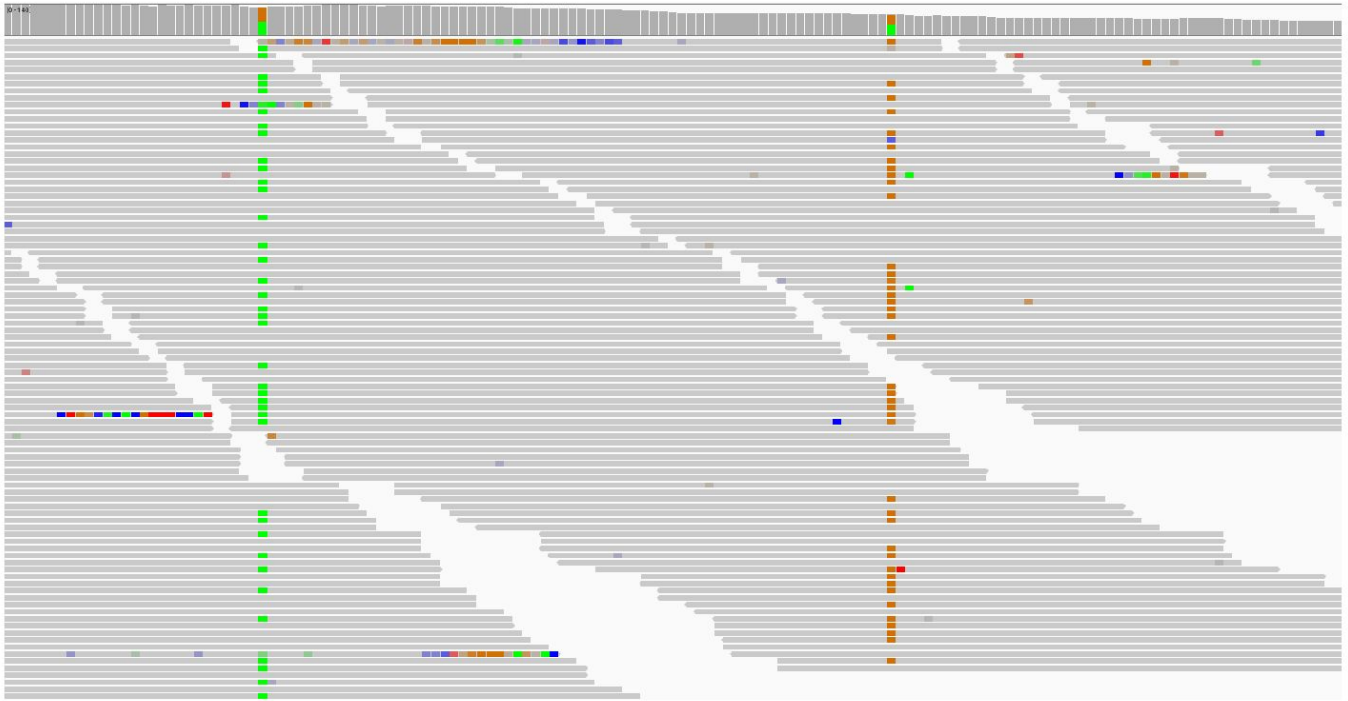


What good SNPs look like?

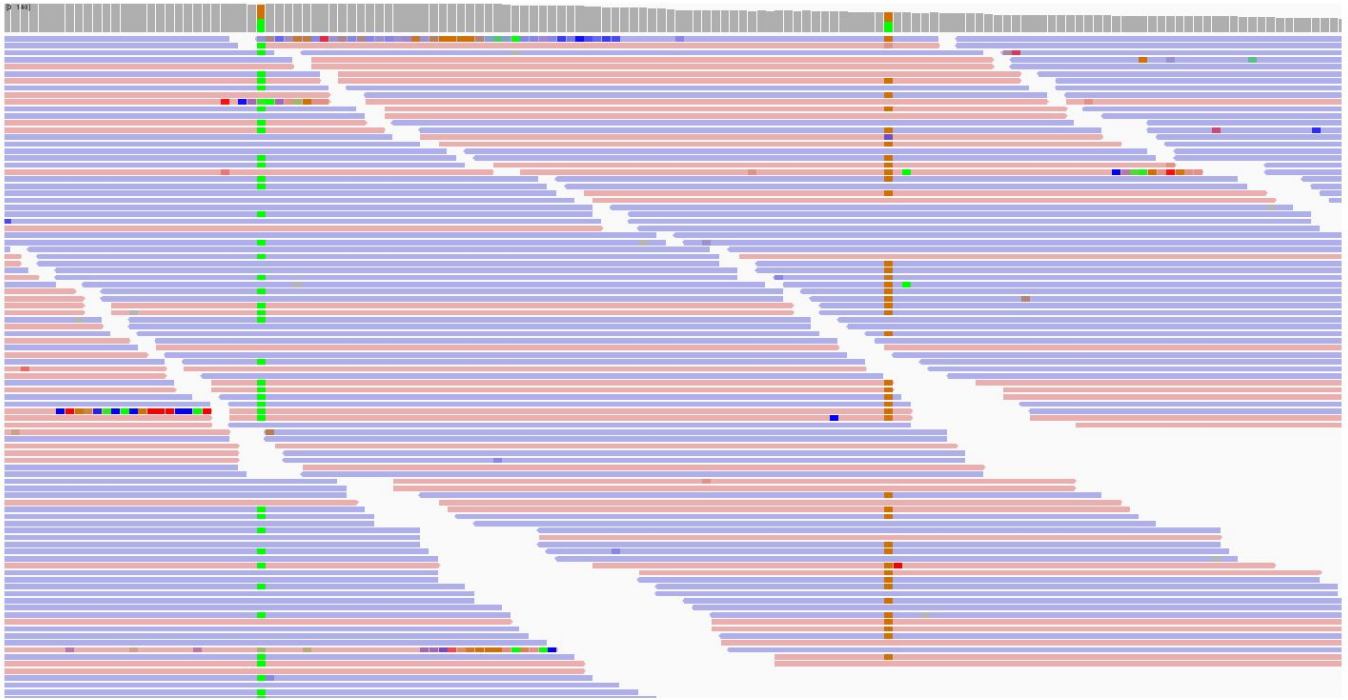


Change the view IGV to inspect possible biases. Here the reads were squished and grouped by read strand to confirm two clean unbiased calls.

What good SNPs look like?



What good SNPs look like?

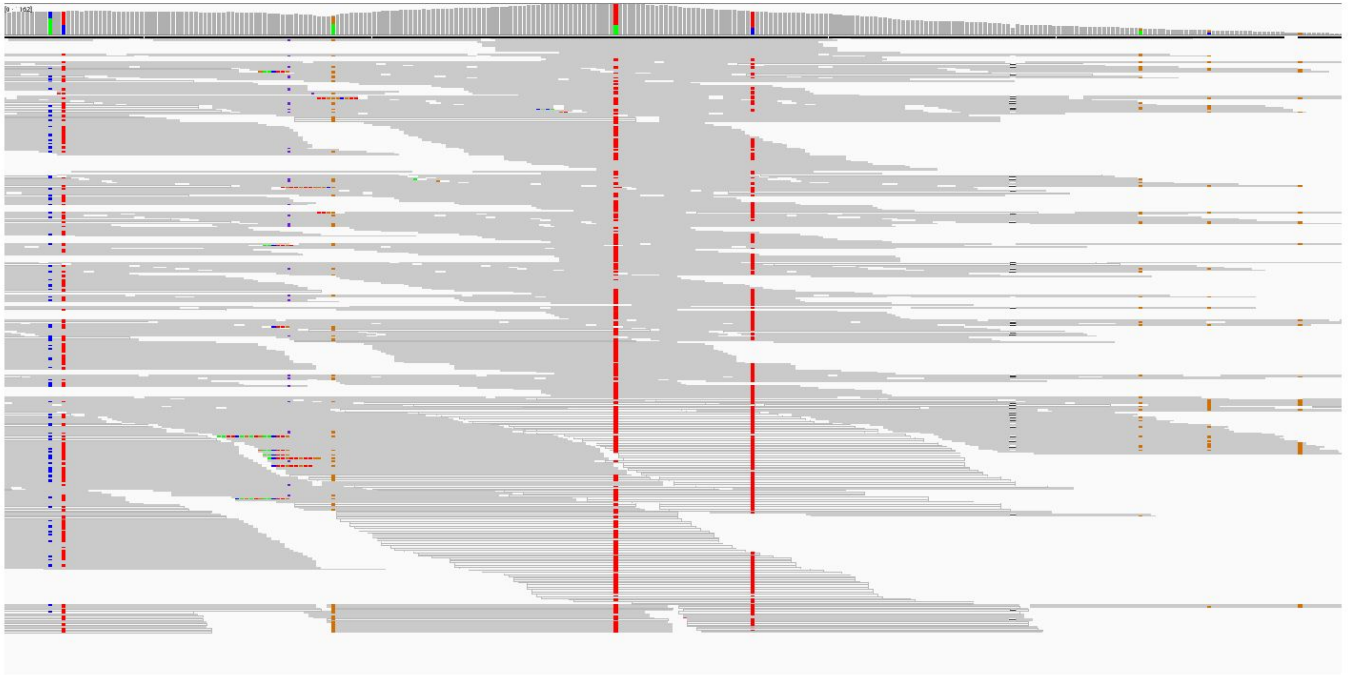


Change the view IGV to inspect possible biases. Here the reads were colord by read strand to confirm another two clean unbiased calls.

What good SNPs look like?

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

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 overall callset quality?

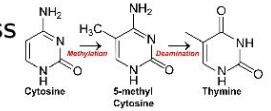
- ti/tv .. proportion of transitions vs transversions
- VAFxx .. proportion of calls with small VAF (variant allele frequency)
- het/hom .. proportion of heterozygous vs homozygous calls

In trios

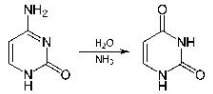
- transmission rate .. ~50% of parental singletons should be transmitted to the child

Detour: Some causes of SNPs

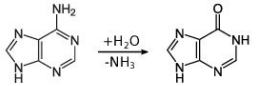
Spontaneous chemical processes which lead to base modification or loss



- Deamination
  - methylated CpG dinucleotides: 5-methylcytosine → T
  - hydrolytic deamination of C → U (400 cytosines daily in each cell)

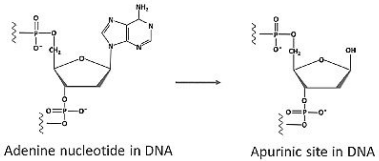


  - A → hypoxanthine (pairs with C, A-to-G mutation)
- Depurination (loss of A or G)
  - purines are cleaved from the backbone (10^2-10^3 daily in each cell)
  - if base excision repair fails, random base is inserted



DNA damage by mutagens

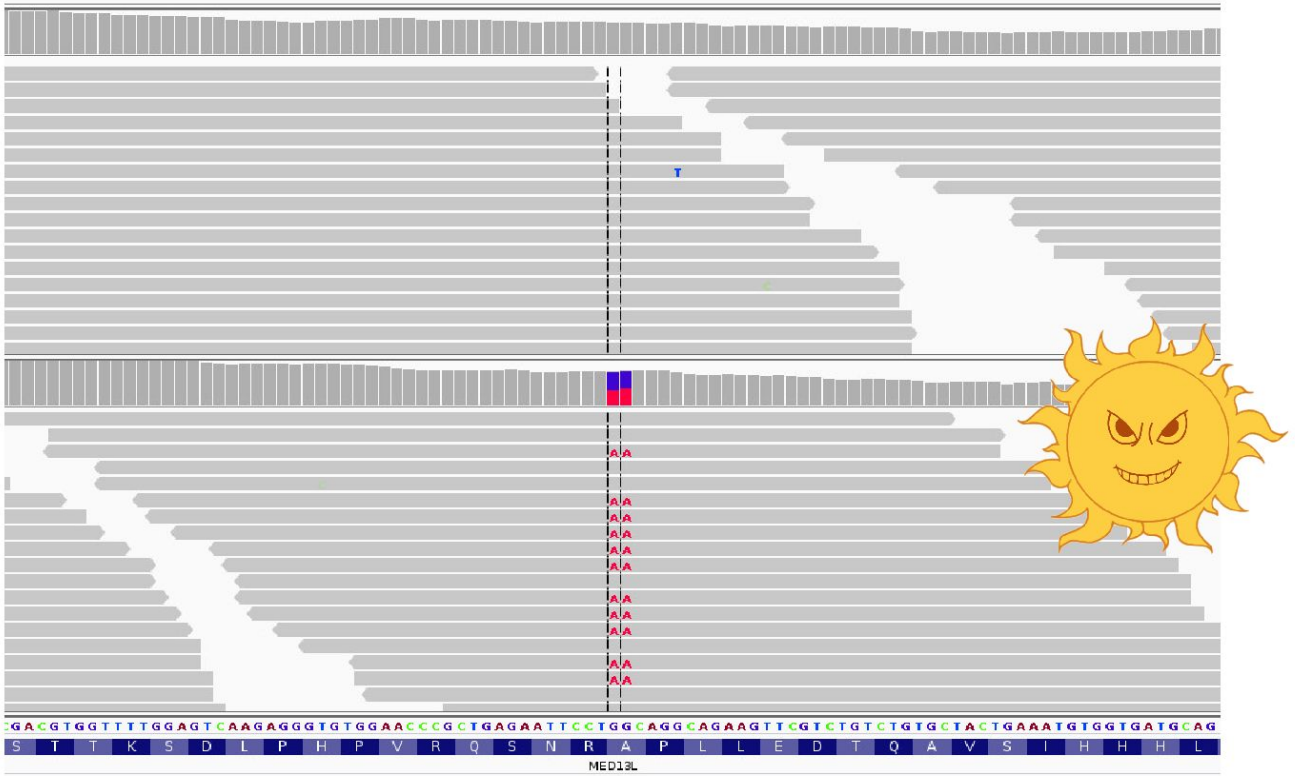
- base analogs
  - incorporation of chemicals with different properties
- base-modifying agents



Radiation

Some causes of MNPs

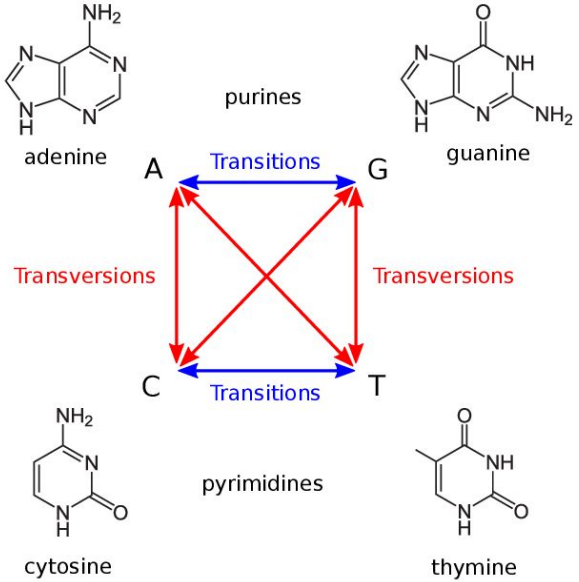
UV-induced mutations (CC → TT in skin cells)

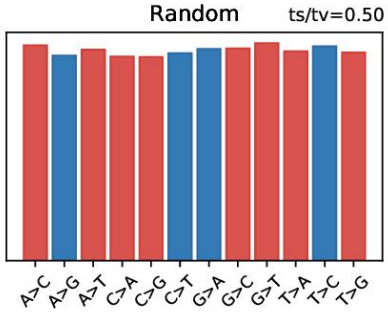


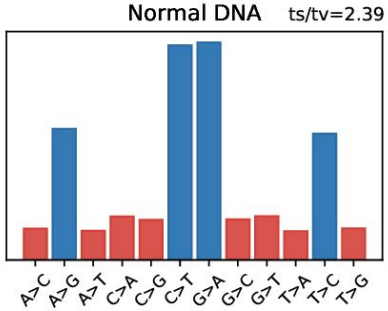
Also known as ti/tv

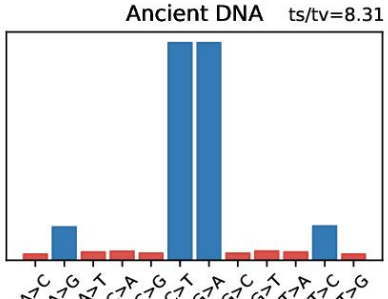
Transitions vs transversions ratio, known as ts/tv

- transitions are 2-3× more likely than transversions



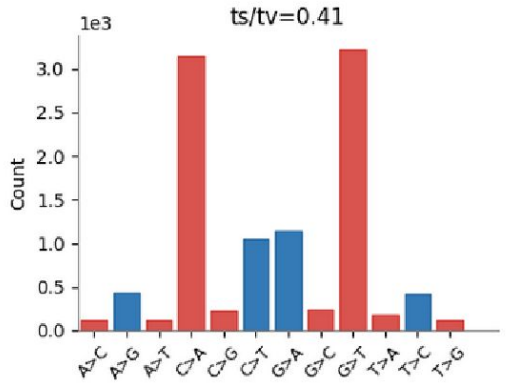


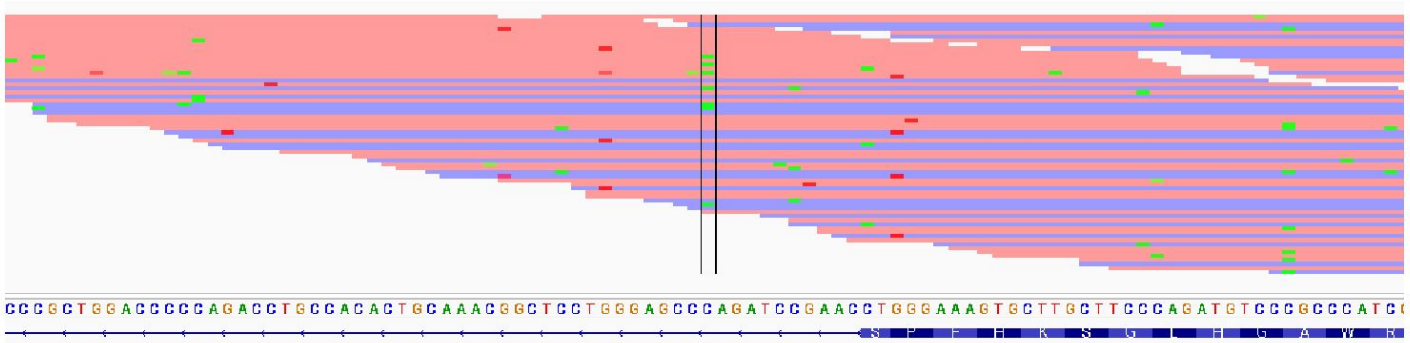




Example: false C>A transversions due to a failed library prep

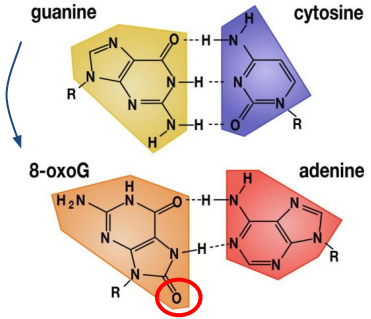
Cause unknown, likely oxidative damage





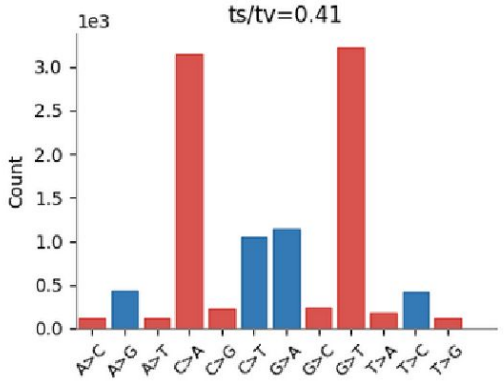
Example: false C>A transversions due to a failed library prep

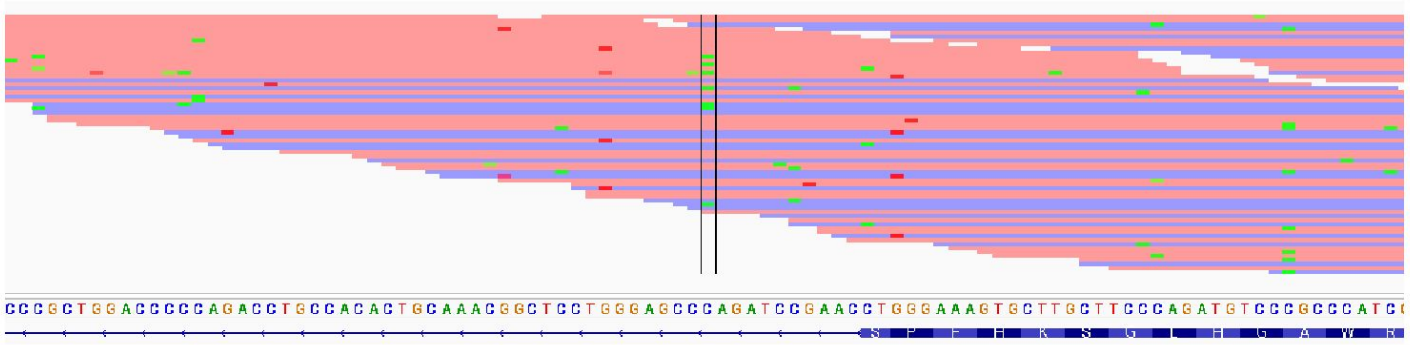
Cause unknown, likely oxidative damage



8-oxo-7,8-dihydroguanine likes to pair with adenine

doi: 10.1016/j.csbj.2019.12.013

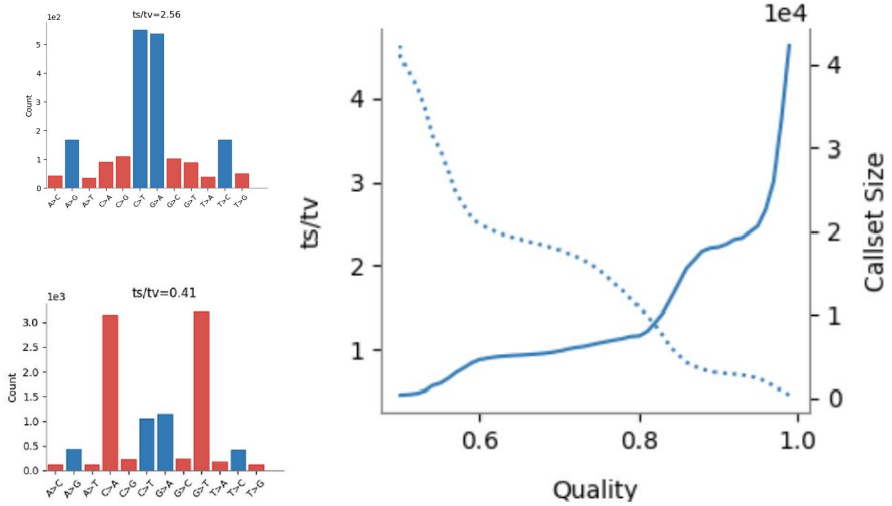




ts/tv by size

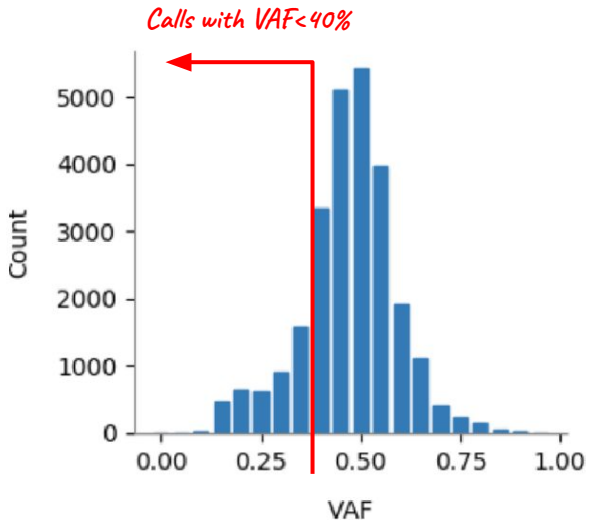
Ts/tv is a convenient metric to compare callsets

- sort calls by a quality metric
- calculate ts/tv at various thresholds
- bigger ts/tv indicates fewer false positives in the callset

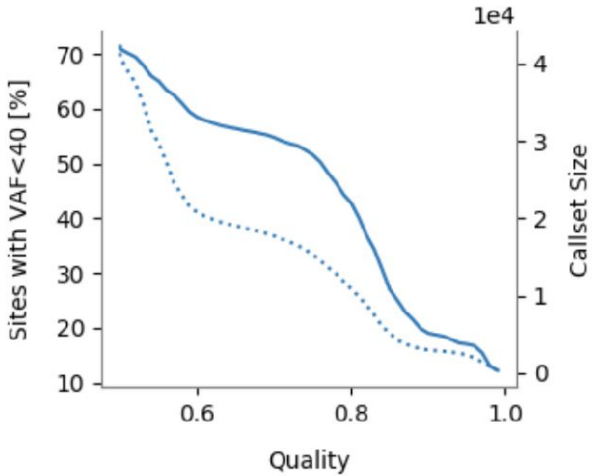


VAF by size

VAF = variant allele frequency (fraction of reads with the alternate allele)



VAF distribution of the final callset



Fraction of sites with VAF<40% at various quality cutoffs

Sensitivity vs Specificity

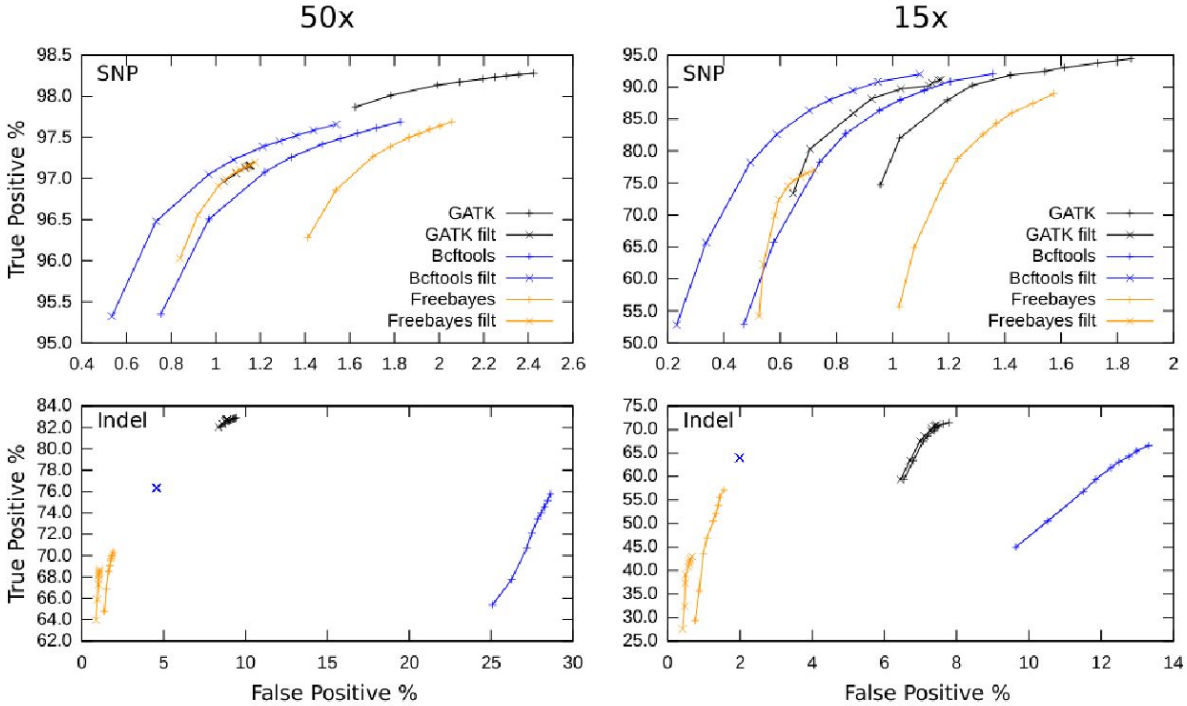


Figure 4: *A summary of True Positive vs False Negative rates of GATK HaplotypeCaller, Bcftools and Freebayes at multiple quality thresholds, with and without filtering.*

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

Genome VCF (gVCF)

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
19	9902	.	G	.	.	.	DP=0
19	9903	.	T	.	.	.	DP=0
19	9904	.	A	.	.	.	DP=0
19	9905	.	C	.	.	.	DP=0
19	9906	.	G	.	.	.	DP=5
19	9907	.	T	.	.	.	DP=7
19	9908	.	A	.	.	.	DP=10
19	9909	.	C	.	.	.	DP=13
19	9910	.	G	A	.	.	DP=15
19	9911	.	T	.	.	.	DP=14
19	9912	.	A	.	.	.	DP=19
19	9913	.	C	.	.	.	DP=23
19	9914	.	G	.	.	.	DP=22
19	9915	.	T	.	.	.	DP=17
19	9916	.	G	T	.	.	DP=18
19	9917	.	A	.	.	.	DP=19
19	9918	.	C	.	.	.	DP=16
19	9919	.	G	.	.	.	DP=25
19	9920	.	T	.	.	.	DP=23

→

gVCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
19	9902	.	G	.	.	.	MinDP=0;END=9905
19	9906	.	G	.	.	.	MinDP=5;END=9909
19	9910	.	G	A	.	.	DP=15
19	9911	.	T	.	.	.	MinDP=14;END=9915
19	9916	.	G	T	.	.	DP=18
19	9917	.	A	.	.	.	MinDP=16;END=9920

Often it is not sufficient to keep only *variant* sites:

- is there **no alternate allele** or is there **no coverage**???
- need evidence for both variant and non-variant positions in the genome

VCF vs BCF

VCFs can be very big

- compressed VCF with 3781 samples, human data:
  - 54 GB for chromosome 1
  - 680 GB whole genome

VCFs can be slow to parse

- text conversion is slow
- main bottleneck: FORMAT fields

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##ALT=<ID=DEL,Description="Deletion">
##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  3  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  4  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
1  5  .  C  T  .  PASS  AC=20;AN=6701;DP=5234  GT:PL:DP:GQ  1/0:255,0,75:32:15  0/0:0,2,170:14:90   1/1:0,9,73:13:31  0/0:0,6,50:13:80    0/0:0,2,80:14:90
1  6  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  7  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
```

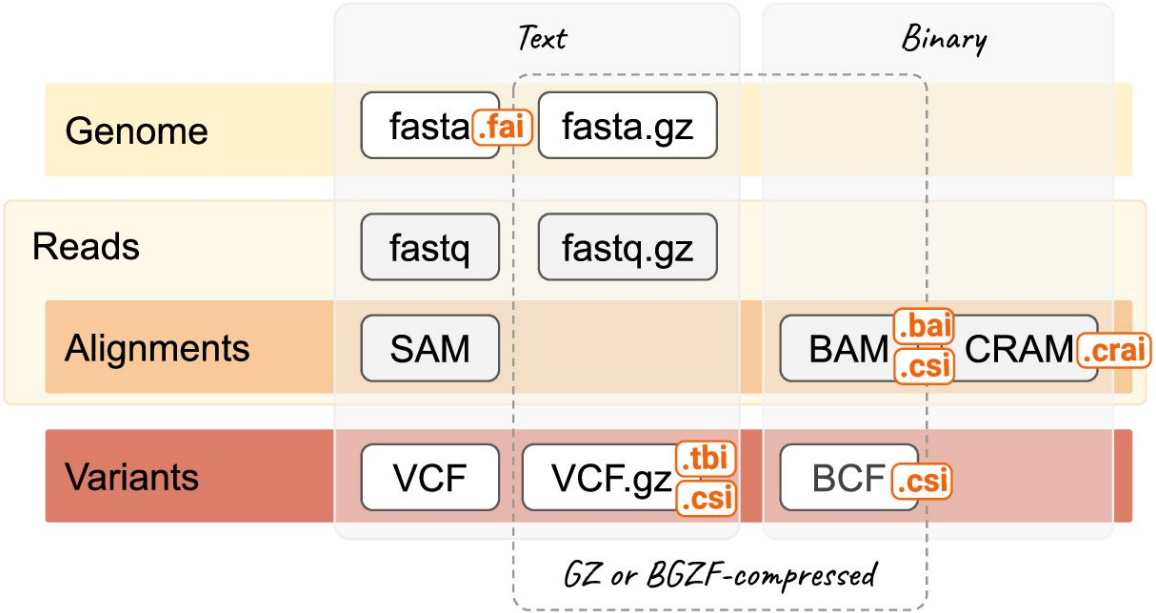
BCF

- binary representation of VCF
- fields rearranged for fast access

```
#CHROM POS ID REF ALT QUAL FILTER INFO          FORMAT      SAMPLE1          SAMPLE2          SAMPLE3          SAMPLE4          SAMPLE5
1      6   .  A   G   .    PASS   AC=67;AN=540  GT:PL:DP:GQ  1/1:0,9,73:26:22  0/0:0,9,73:13:31  0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15

1      6   .  A   G   .    PASS   AC=67;AN=540  GT:1/1:0/0:0/0:1/0:1/0  PL:0,9,73:0,9,73:0,9,73:255,0,75:255,0,75  DP:26:13:48:32:32  GQ:22:31:99:15:15
```

File formats summary



Note: BCF can be compressed (`bcftools view -Ob`) or uncompressed (`bcftools view -Ou`). Use the latter for streaming, it is much faster!