SNPs

Niklaus Zemp
29 June 2021

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Sequence alignment

http://hurlab.med.und.edu/cgi-bin/SimpleSeqAlign/SimpleSeqAlign.cgi

AATTTCCC
AATATCCC

AATTTCCC
AATTCCC

AATTTCCC
AATTCCCAAT

Sequence alignment

http://hurlab.med.und.edu/cgi-bin/SimpleSeqAlign/SimpleSeqAlign.cgi

AATTTCCC
AATATCCC

AATTTCCC
AATT CCC

AATTTCCC
AATT CCCAAT

Sequence alignment - global

http://hurlab.med.und.edu/cgi-bin/SimpleSeqAlign/SimpleSeqAlign.cgi

AATTTCCC
AATATCCC

```
seq1: 1 AATTTCCC
        |||*||||
seq2: 1 AATATCCC
```

AATTTCCC
AATTCCC

```
seq1: 1 AATTTCCC
          |||| |||
seq2: 1 AATT-CCC

seq1: 1 AATTTCCC
          ||| ||||
seq2: 1 AAT-TCCC

seq1: 1 AATTTCCC
          || |||||
seq2: 1 AA-TTCCC
```

AATTTCCC
AATTCCCAAT

```
seq1:   1 AATTTCCC--
            ||||*||*
seq2:   1 AATTCCCAAT

seq1:   1 AATTTCC-C-
            ||||*||  *
seq2:   1 AATTCCCAAT

seq1:   1 AATTTCC--C
            ||||*||   *
seq2:   1 AATTCCCAAT
```

Sequence alignment - local

AATTTCCC
AATTCCCAAT

```
seq1: 1 AATTTCCC
        |||| |||
seq2: 1 AATT-CCC

seq1: 1 AATTTCC
        ||||*||
seq2: 1 AATTCCC
```

Alignments

https://www.ndsu.edu/pubweb/~mcclean/plsc411/Blast-explanation-lecture-and-overhead.pdf

Local

Smith-Waterman (algorithm)

Uses a dynamic programming approach

Fast because only small part to work on but works only locally

Global

Needleman-Wunsch (algorithm)

Slow because large sequences to align, therefore CPU-”expensive”

Task: Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi and blast the sequences below

BLAST-local alignment

TruSeqUniversalAdapter

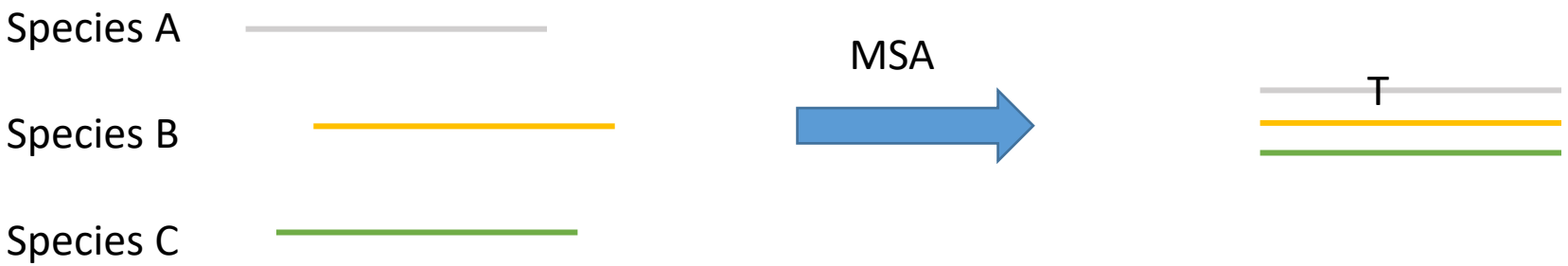
5’

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

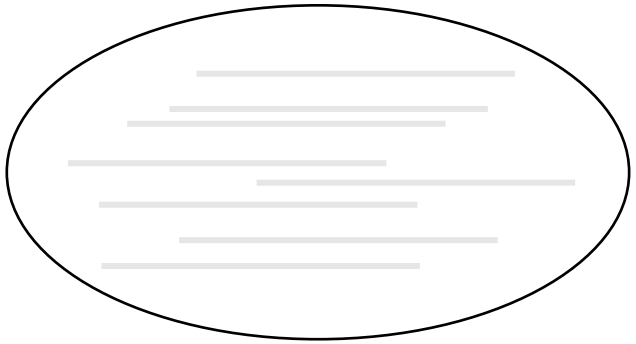
Transcript

TTGTTAAAAAATTTTTTTTAAGTTTTTTTCTCTTTTTTTCATTTAAATATATTTTATAAATTTCTATGAA
ATAGTTAACATTGAATAAGCGAATTTAAAAAAAATGTTCATGATCTTAGATAGACTAATAACGACCTGAT
TATATTCGAGCTGTAGTATTTTTATATTTCACTATTATGTATGAAATTTTTAACATCACAGCCAAGTTAA
TATAACCTCGCTCCAAACCTGAACATTCAAACACTAACTATACTTAAAACGCTAGTTTTGTTAAGTCTAT
CTAAGACCATGATGTAGTTGTATAGCTCGGATCATTTTGAAAATAATAATTGGACTAAACTATAAAAAAA
AAAACATTGGAACATTGTATTATGTAAGTTCATCCAGTTAACTTGGAAAAATTAACTTGGAATGGAAACG
TAAGCTGAACTAAACTTTTCATTCACTTCAAAGCATCCGTATATTCTTGTCGGTGTATGGACTTGTTATG
TAGGATAATTCCATGTTGTGGATTGTTGATTGCGGACAATTGTCGTTTGTTTTAACATGACAATGTTTAT
GACATTTTATTAAACAATCTCTGCATTCGTAACCTTGTTTTCCTAATCTTCGAGCTATGCTTTTACTACA
AACTTGGCACACTGTTCCACCATTTAAGTGCTTGGCAATAAATGTATGATCATTAAAAATGTGCAATTTT
GTGCCTTTTTTACGCCATCTTGATTTTTGTGCTAATGATAATGGTACCAAATAATGTTTTTTAATACCAT
TTTCAAGTGTTTCAAGTACTAATGTGCTTGCTTCATTTAAGTATGTTCGAGTTGAAGCACCGCTTATACC

Multiple sequence alignments-global alignment



Mapping



Raw reads

Alignment/mapping

Reference

Mappers

Problem

The fast and exact algorithms for local alignments do not scale to large genomes. Do not handle high sequence errors well.

New approaches needed Solution

First apply very fast algorithms that match short local regions exactly. Then extend the short regions to larger regions.

Global mappers

k-mer based alignment -> RNA-seq

can be fast and quite accurate AGCTTTAGAC ->3-mers: AGC, GCT, CTT, TTT, TTA, TAG, AGA, GAC

when k-mers are redundant, i.e. appear often in sequences/genome

suffix-tree

a tree-like structure that contains all suffixes of the sequences (genome).

Subsequences (reads) can be looked-up very quickly.

needs a lot of memory

compressed suffix-tree

a compressed form, e.g. **Burrows-Wheeler transform** very fast, very memory efficient.

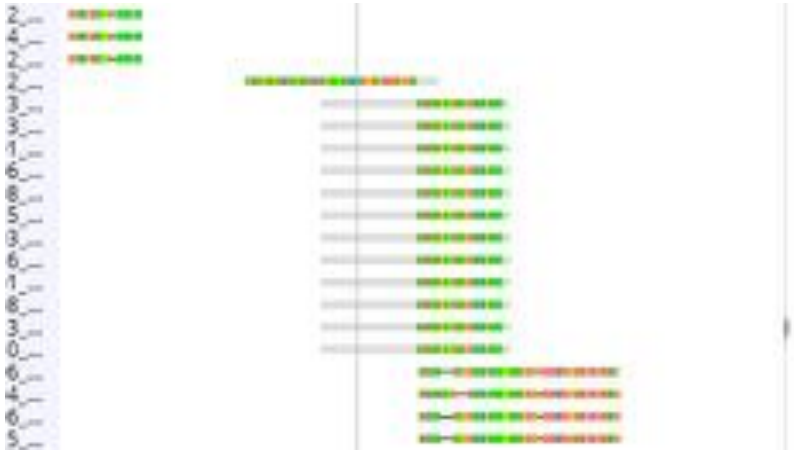
gets rather slow and inaccurate with high sequence error rates or long reads

MEM-mapping

maximal exact match

cannot be extended

Soft clipping during read mapping (bwa)



Sam/bam format

https://samtools.github.io/hts-specs/SAMv1.pdf


```
@HD          VN:1.3            SO:coordinate

@SQ          SN:scaffold3_8_ref0000002_ref0000084           LN:96419

@RG          ID:ERR3418961    SM:ERR3418961  PL:Illumina

@PG          ID:bwa            PN:bwa            VN:0.7.12-r1039  CL:bwa mem -R @RG\tID:ERR3418961\tSM:ERR3418961\tPL:Illumina Ref_C_bombi.fasta ERR3418961_1.fq.gz ERR3418961_2.fq.gz

ERR3418961.5587914       163          scaffold3_8_ref0000002_ref0000084        1238        60        151M        =        1565        478
             GTGGGGAGATAAAACCGAACTGGTGTGTTAACATCAACGTCTGTCCCTTCACGGTGGGGACCGCGTGCACGACGCGCATCATCCGTCTGCGTGGTGGGGGGTGGTAGTGTAGAATTTCGTTGGTGCTCCTGCCGTCGCAGGACC
GCCTCCA      AA7FFFAJFFJJJJJJFJJJJFJFJJJAJFJFJJJJJFJJJF7FJJJFJFJJJJJJJJJ<FJ7F<FJFJJJJJJ<JJJFAJFJJJJJFJJJJJJJJJJJFF-AJJJJJJJJJJJ-FFJJJJFFJFJJJFFJJ<-A<7FF-<AAF<<7<FF)        NM:i:0        MD:Z:151
             AS:i:151          XS:i:0            RG:Z:ERR3418961

ERR3418961.5587914       83           scaffold3_8_ref0000002_ref0000084        1565        60        151M        =        1238        -478
             CCGACGTCCAGGCGGACGCTCGAGCTCGTGACGGGGAGGTAATCGATGTAGGGCACGACAGTGGTGCCGGAGAAGCCGACGTCTACGACGATCGCAGTCCCACCCGCCTCACACGTCAATTCTCTGTCAATAACATTAGTACCA
ACACTAG      JJJFJJJFJFJJJJJJJJFFJJAFJFAJFFFFJFFJJJJJJJJJJJJJFJJJJFJJJJAFJJJJJJJJJJJJFFJJJFJJJJJJFJJFJJFJJJJJJFFFJJJJJJFFFJJJFJFJFJJFFJF<FJJJJJJJFJJJJJJJJJJJJJFFFAA        NM:i:0        MD:Z:151        AS:i:151        XS:i:0
             RG:Z:ERR3418961

ERR3418961.19317         185          scaffold3_8_ref0000002_ref0000084        4042        16        101S36M14S        =        4042
             0GCCGTGCATACGCGCAGCGCTCCACGTGACTGCAATTCGTCACATGCTCACTAGTACGTATTATCTCGCTGAACTGCGCTGGCGCTATATGTATATATATATATATATATATATACATATATATATATATATATATGCAAAGTGCAC
GCGT         )JJJJFJF7<FJJJJJJJJJJJJJJJJFAFFJAJJJF7JJJJJJJJFAFFJFFAJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA        NM:i:1        MD:Z:12T23        AS:i:31
             XS:i:27           RG:Z:ERR3418961              XA:Z:scaffold3_8_ref0000002_ref0000084,-56115,97S27M27S,0;scaffold3_8_ref0000002_ref0000084,+51496,34S26M91S,0;
```

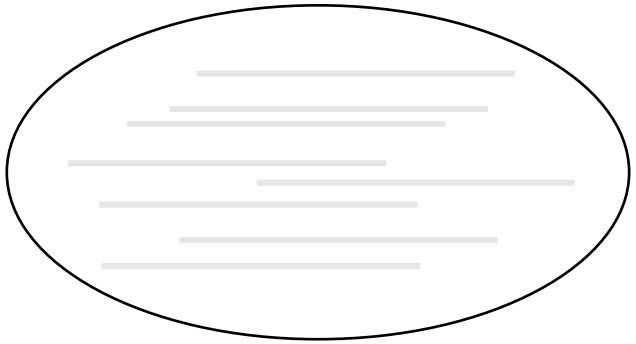
Read columns: 1) read name; 2) flag(binary); 3) contig/chromosome; 4) position; 5) mapping quality; 6) CIGAR string;7) RNEXT; 8) Position of next read in alignment (pair); 9) observed template length; 10) sequence; 11) quality per base; 12) optional information

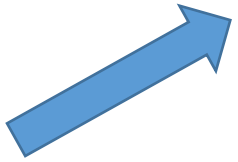
SAM flags: https://broadinstitute.github.io/picard/explain-flags.html

Genetic Diversity Centre Zurich

SNPs
caller

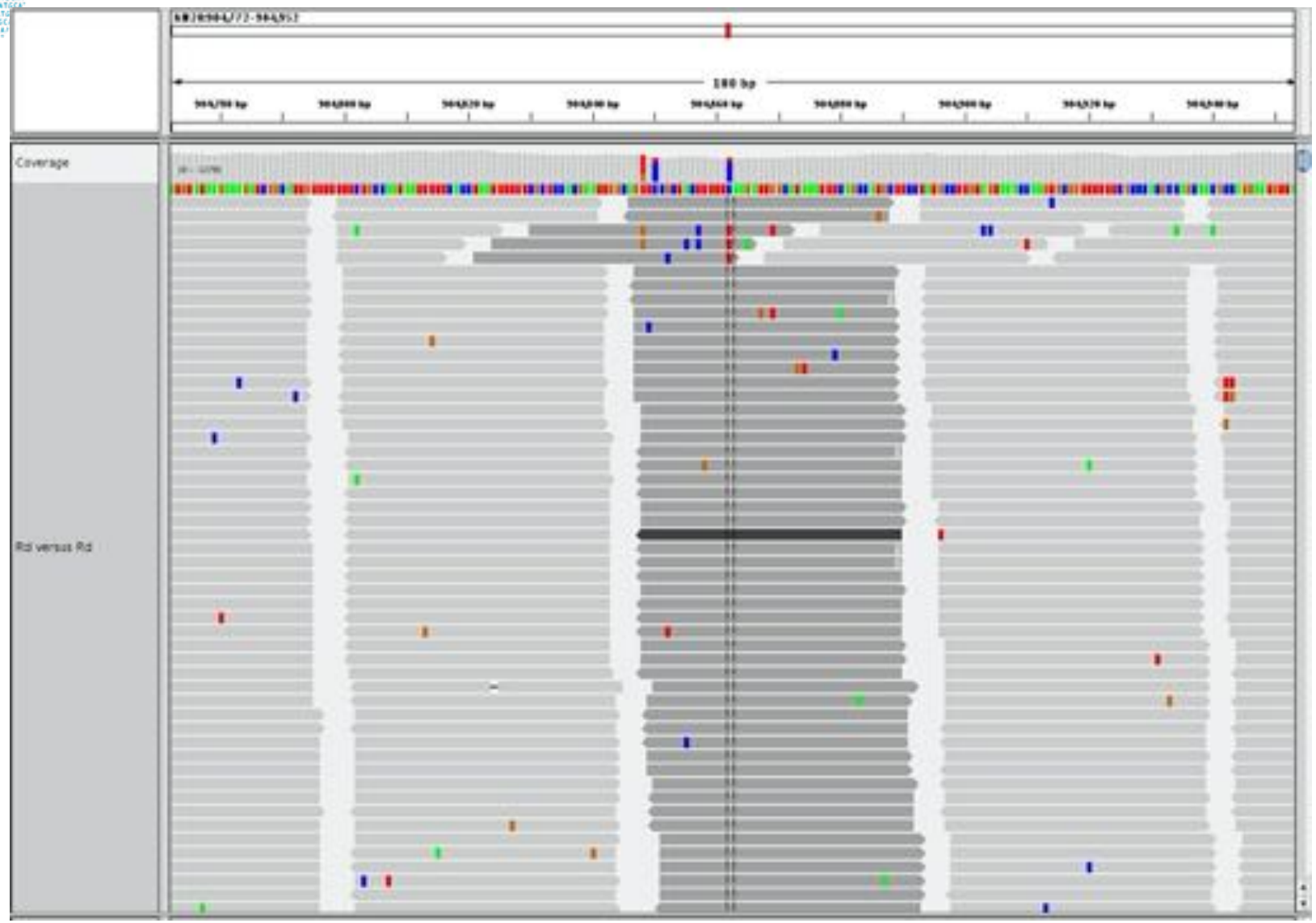
SNP caller



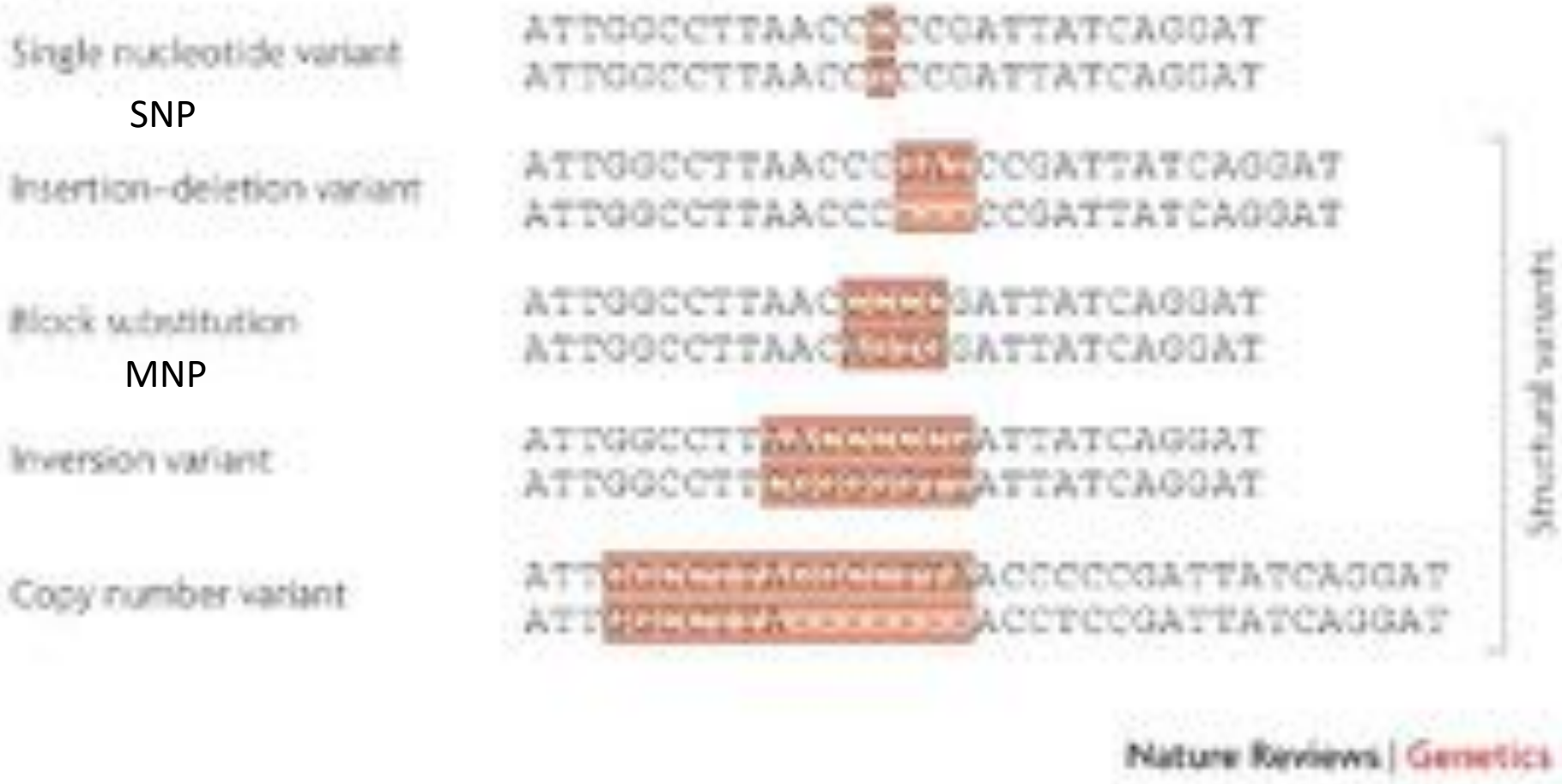


Alignment/mapping

	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT


180 bp
Coverage
Rd versus Rd

Structural variants

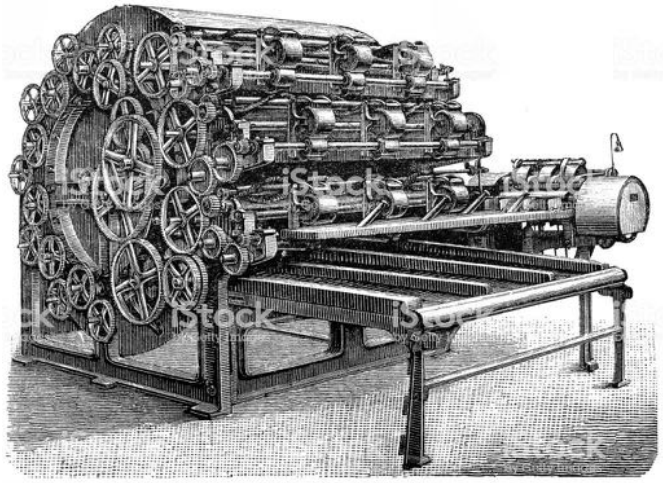


Variant types

- SNP (single nucleotide polymorphism)
- INDEL (insertion/deletion)
- MNP (multi-nucleotide polymorphism, e.g. a dinucleotide substitution, haplotypes)
- CLUMPED (A clumping of nearby SNPs, MNPs or Indel, haplotypes)

SNP caller

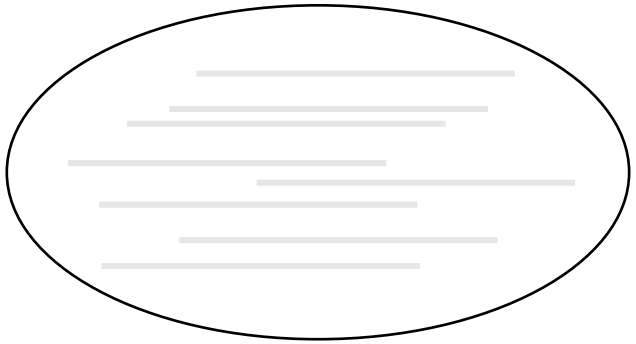
Alignment/mapping

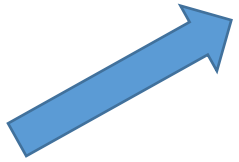


	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT

PCR duplicates

SNP caller

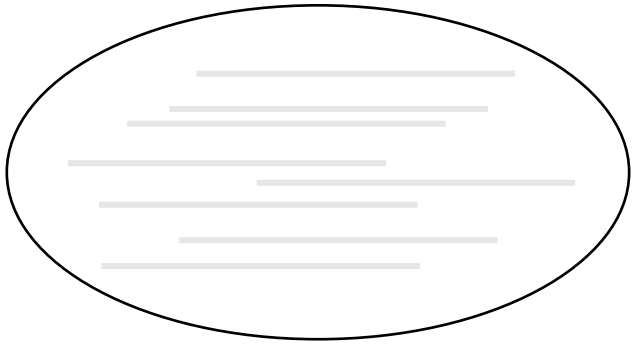


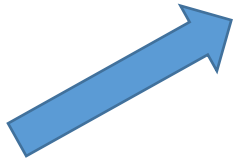


Local realignment

	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT

SNP caller





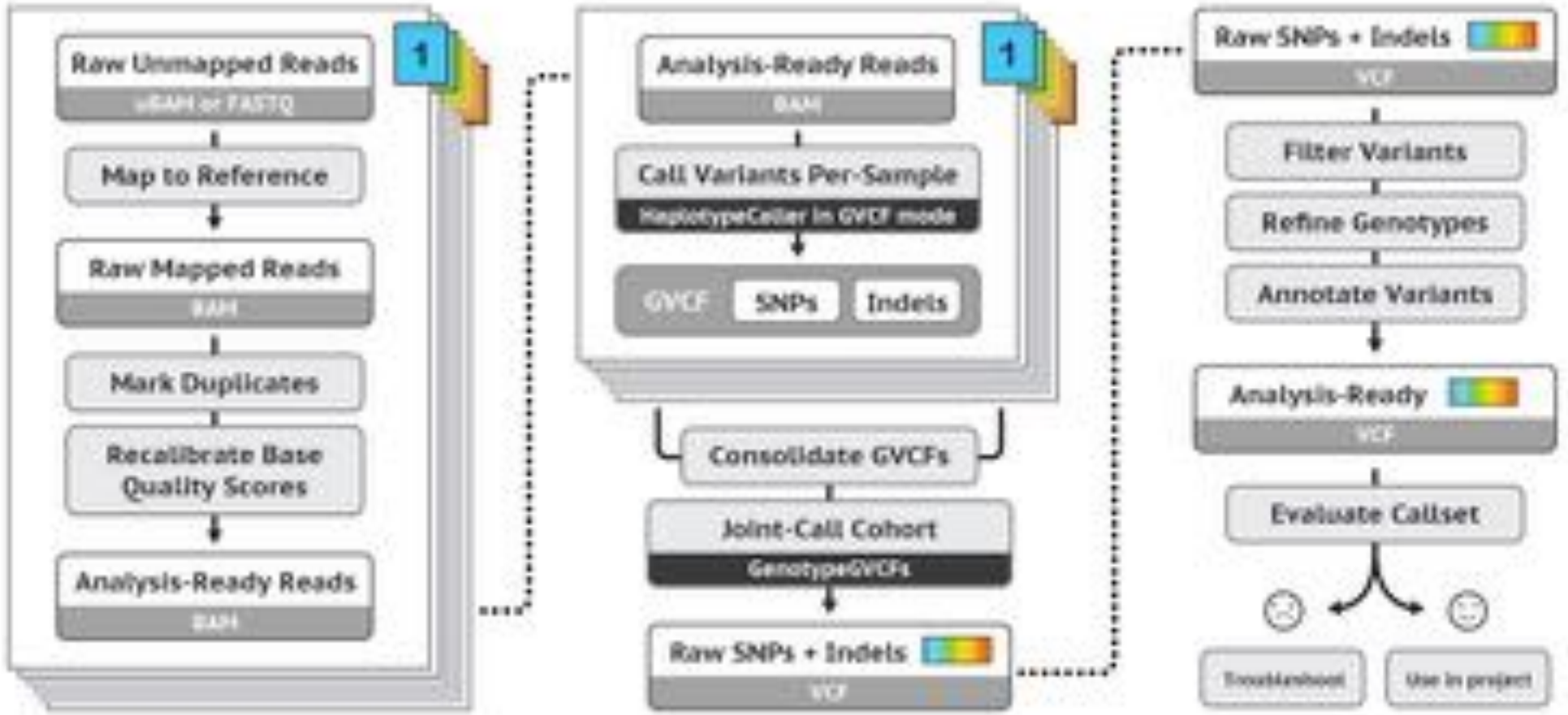
Joint calling

	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT

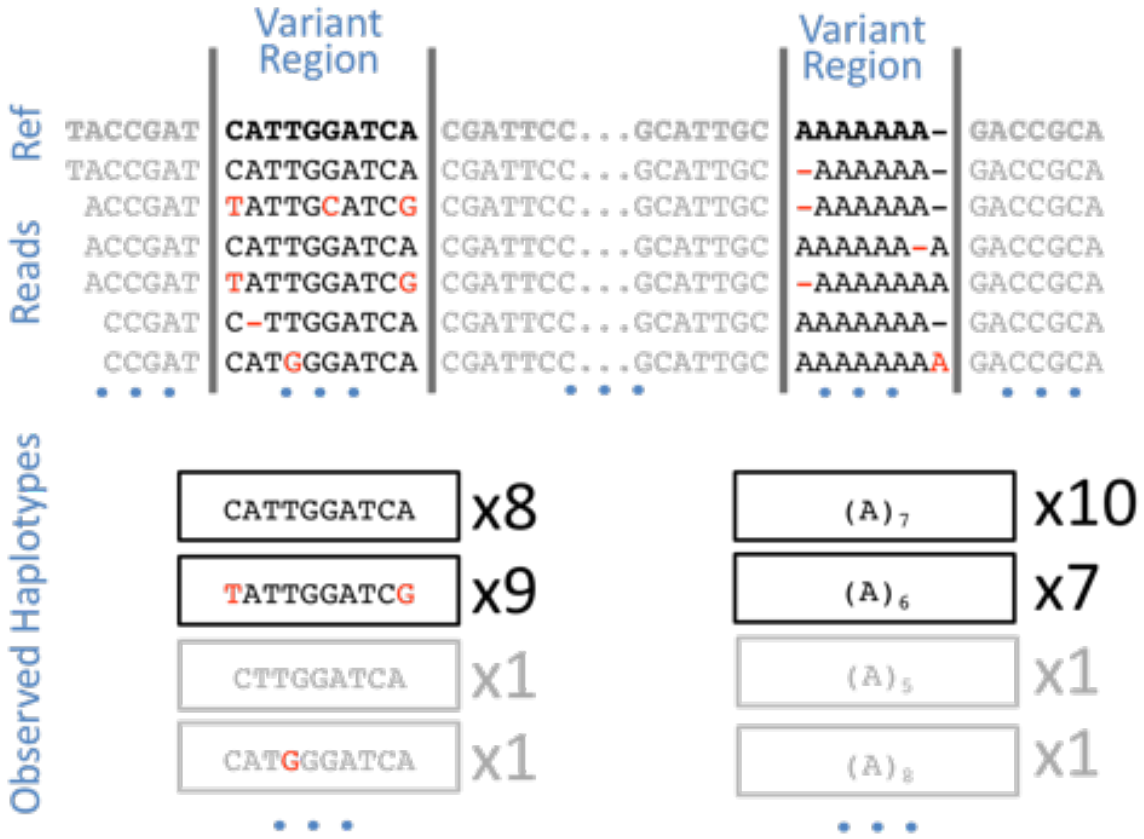
SNP callers

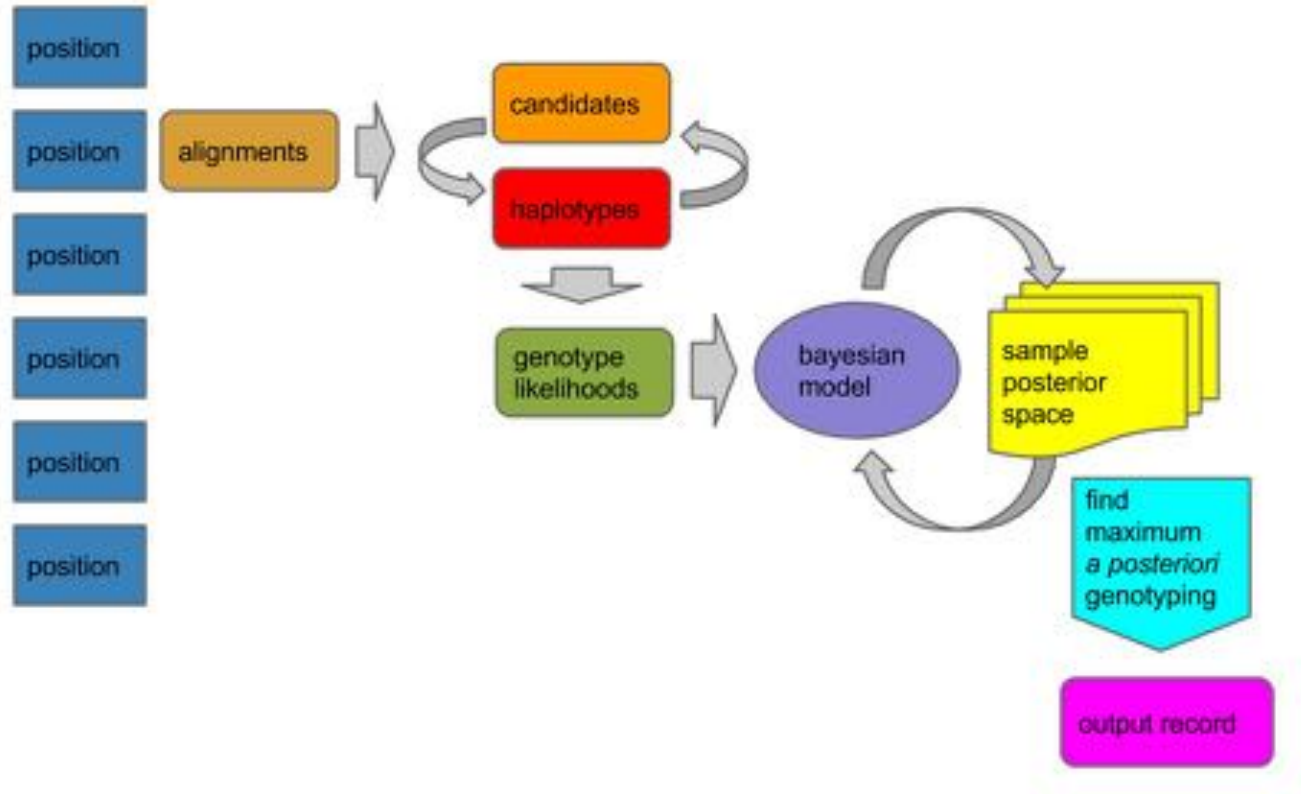
- Samtools
- GATK
- FreeBayes
- Platypus
- Popoolation

GATK


Raw Unmapped Reads
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset

FreeBayes


Variant Region
Variant Region
Ref
Reads
TACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
TACCGAT CATTGGATCA CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT TATTGCATCG CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAA-A GACCGCA
ACCGAT TATTGGATCG CGATTCC...GCATTGC -AAAAAAA GACCGCA
CCGAT C-TTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
CCGAT CATGGGATCA CGATTCC...GCATTGC AAAAAAAA GACCGCA
Observed Haplotypes
CATTGGATCA x8
TATTGGATCG x9
CTTGGATCA x1
CATGGGATCA x1
(A)7 x10
(A)6 x7
(A)5 x1
(A)8 x1


position
position
position
position
position
position
alignments
candidates
haplotypes
genotype likelihoods
bayesian model
sample posterior space
find maximum a posteriori genotyping
output record

Vcf format

##fileformat=VCFv4.3 ##fileDate=20090805 ##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x> ##phasing=partial ##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data"> ##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003

20 14370 . G A 29 . 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 1/1:43:5:.,.

20 17330 . T A 3 . NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 .:41:3

20 1110696 . A G,T 67 . 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 2/2:35:4

Vcf format

20 14370 . G A 29 . 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 1/1:43:5:.,.
20 17330 . T A 3 . NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 .:41:3
20 1110696 . A G,T 67 . 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 2/2:35:4

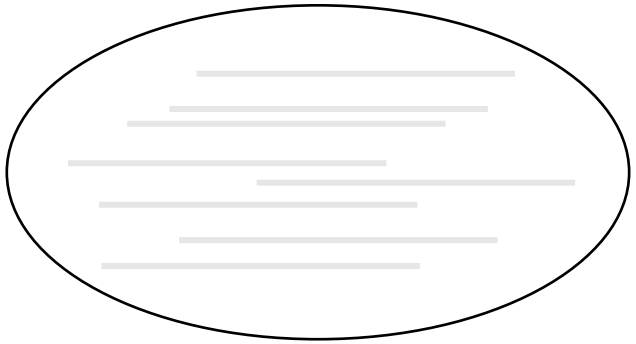
20 14370 G/G G/A A/A
20 17330 T/T T/A NA
20 1110696 G/T G/T T/T

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a '.'. Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: <key>=<data>[,data].
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.

Genetic Diversity Centre Zurich

SNPs
filtering

SNP filtering



hard filter

multiple samples

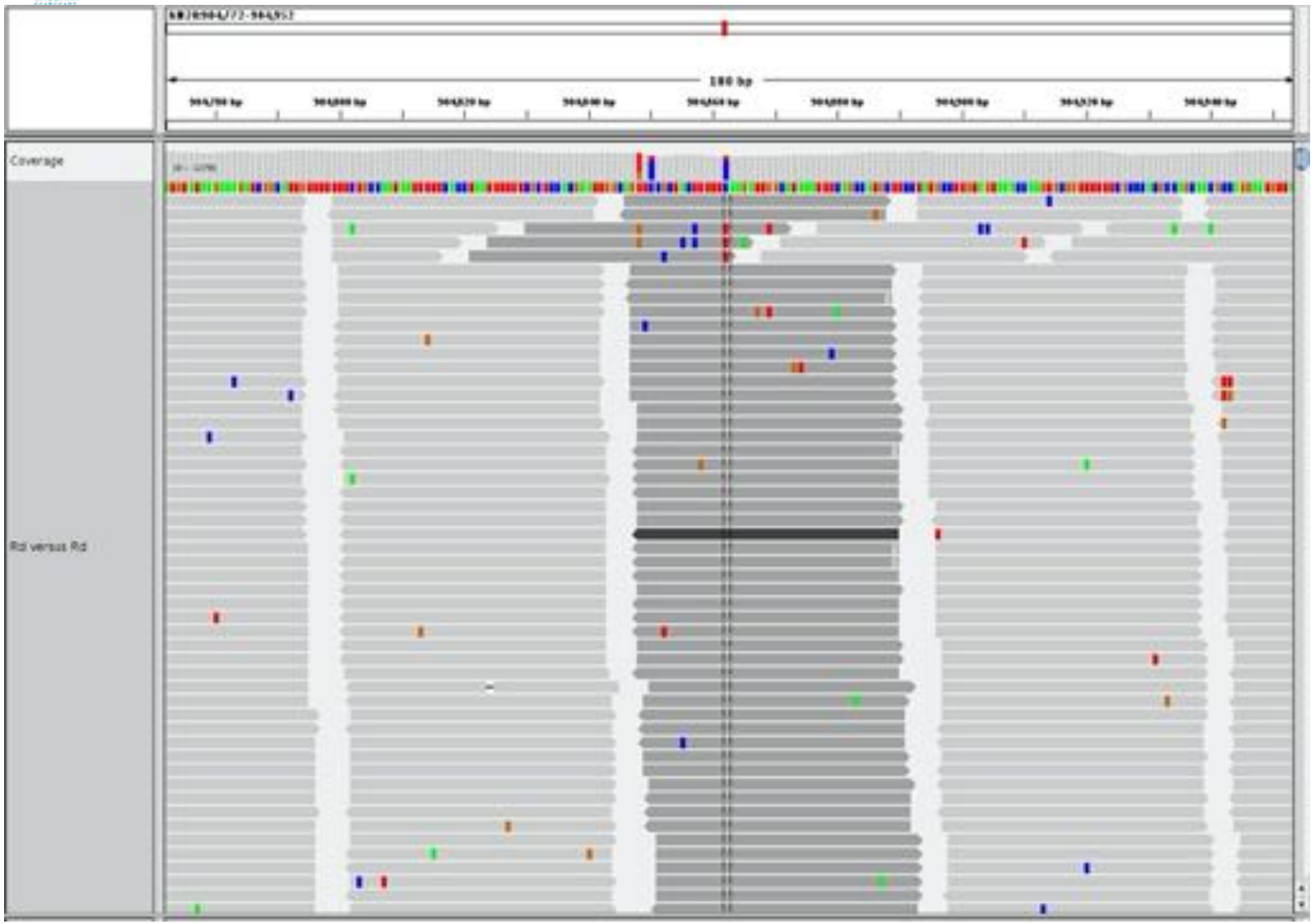
	Sample 1	Sample 2
Pos 1	AA	AT
Pos 3	AT	TT

	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT

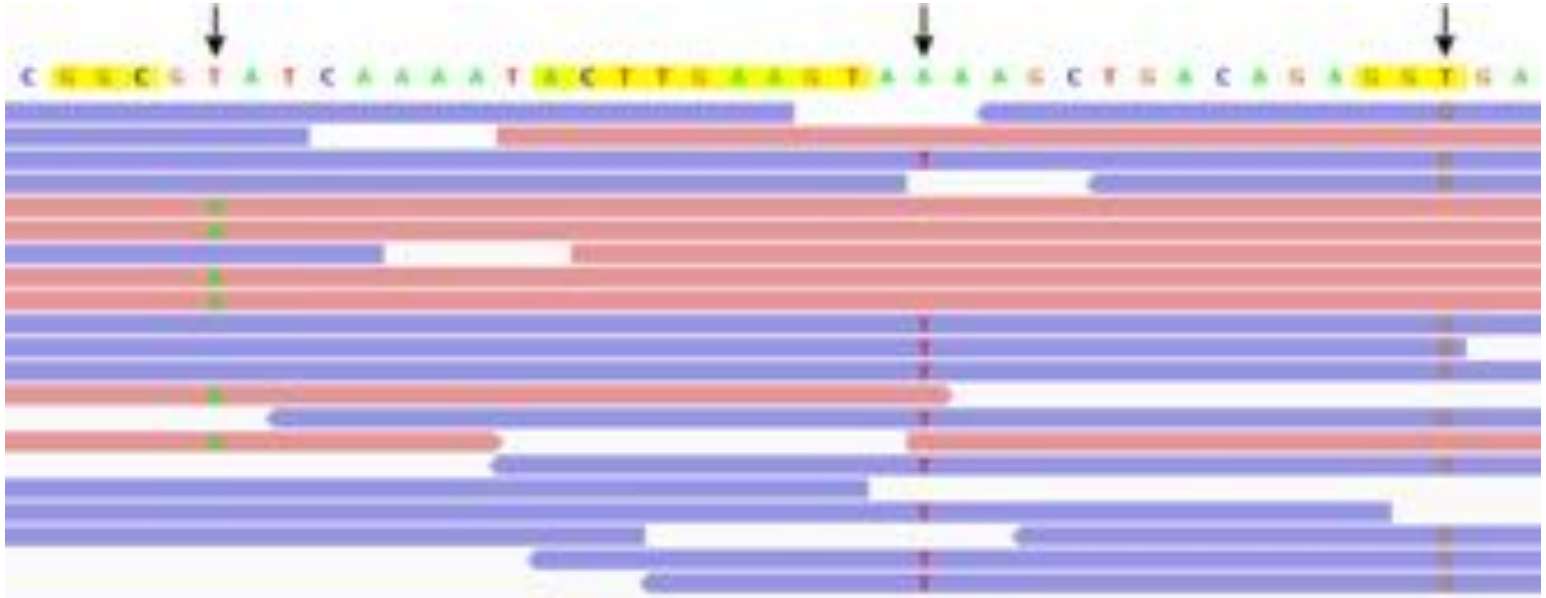
Tools for filtering

- Vcftools
- Vcflib
- Bcftools
- Rvcf

Sequencing errors



Context dependant errors



Rank	Context	FER [%]	RER [%]	ERD [%]
1	ACGGCGGT	26.1	0.5	25.6
2	GTGGCGGT	25.1	0.7	24.4
3	GCGGCGGT	22.9	0.7	22.2
4	GTGGCTGT	22.4	0.6	21.8
5	ATGGCGGT	21.2	1.0	20.3
6	NCGGCGGT	20.0	0.7	19.3
7	GTGGCTTG	20.2	1.2	19.0
8	GNGGCGGT	19.2	0.7	18.5
9	GCGGCTGT	18.8	0.7	18.1
10	ACGGCTGT	18.6	0.8	17.7

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622629/

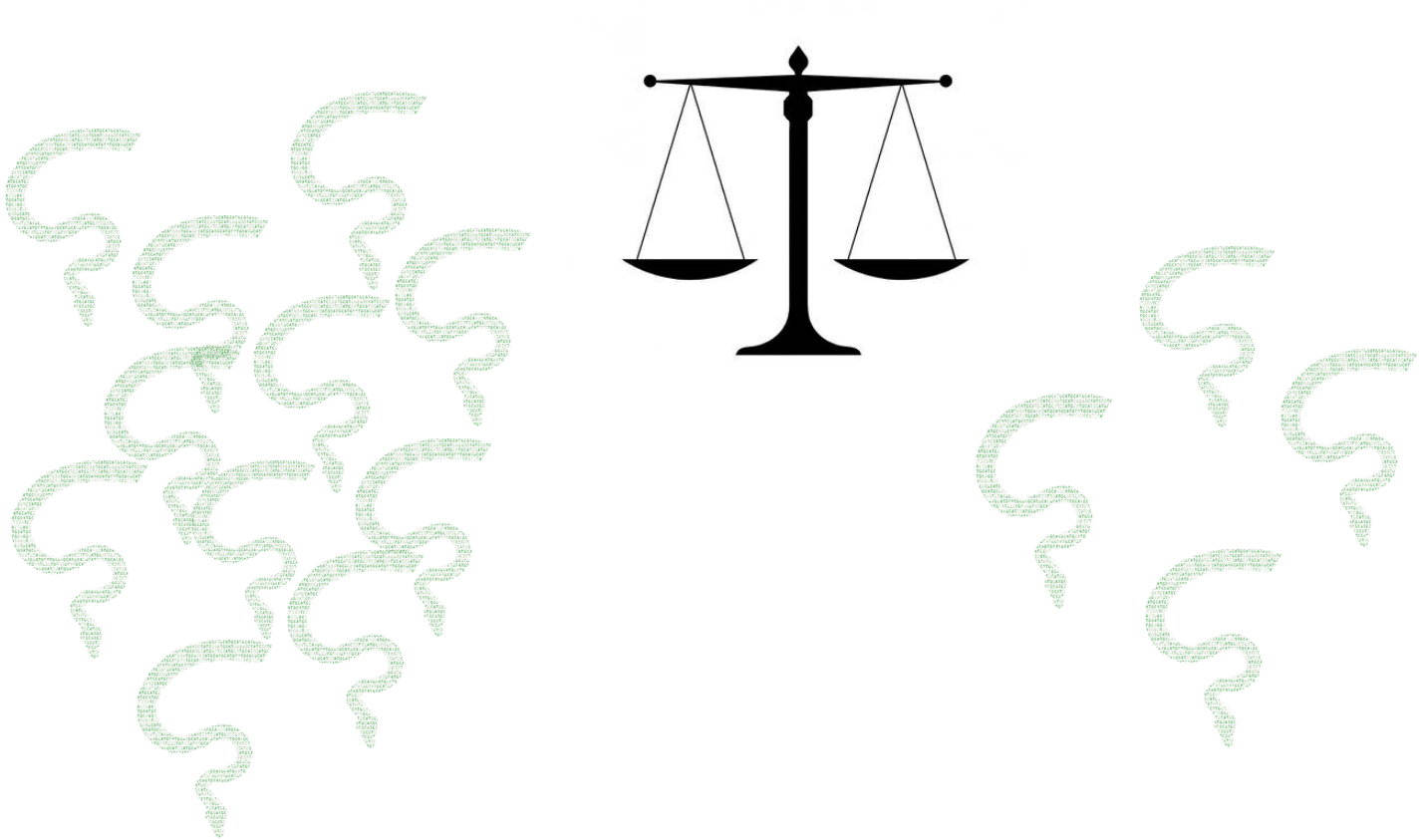
Filter Criteria

- low quality of mapping (and bases)
- read depth (coverage)
- missing sites
- remove samples with a lot of missign sites
- "quality / coverage"(higher coverage sites should have higher quality)
- keep SNPs only (= no indels)
- keep biallelic SNPs only (= no multiallelic SNPs)
- alleles that are found only on one strand (or reads)
- remove SNPs with large discrepancy between reference and alternative allele
- filter for minimum minor allele frequency (MAF), e.g. 5%
- filter for quality ratio of forward and reverse reads

Genotype likelihoods

scaffold3/99|ref0000050|ref0000027 285 . A G 539.871 . AB=0;ABP=0;AC=12;AF=0.146341;AN=82;AO=83;CIGAR=1X;DP=4295;DPB=4295;DPRA=0.359959;EPP=28.1523;EPPR=23.8636;GTI=6;LEN=1;MEANALT=1.16667;MQM=30.4096;MQMR=59.131;NS=41;NUMALT=1;ODDS=2.47263;PAIRED=1;PAIREDR=0.941051;PAO=0;PQA=0;PQR=0;PRO=0;QA=3137;QR=150052;RO=4207;RPL=2;RPP=166.289;RPPR=13.8625;RPR=81;RUN=1;SAF=59;SAP=35.0591;SAR=24;SRF=2074;SRP=4.80704;SRR=2133;TYPE=snp;technology.Illumina=1 GT:DP:DPR:RO:QR:AO:QA:GL 0/0:151:151,0:151:5286:0:0:0,-45.4555,-473.625 0/0:43:43,1:42:1397:1:34:0,-10.3232,-123.369 0/0:53:53,1:52:1826:1:34:0,-12.5878,-160.796 1/1:20:20,20:0:0:20:807:-55.4598,-6.0206,0 0/1:61:61,10:51:1695:10:354:-8.41372,0,-131.988

Ultra low coverage sequencing



Angsd (http://www.popgen.dk/angsd/index.php/ANGSD)

Take home massage

Mappings are normally be full of noise

SNP calling is computational intensive

Raw SNP tables needs to be filtered