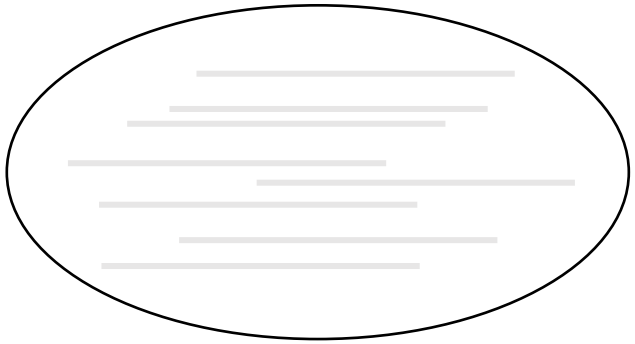
SNPs
caller

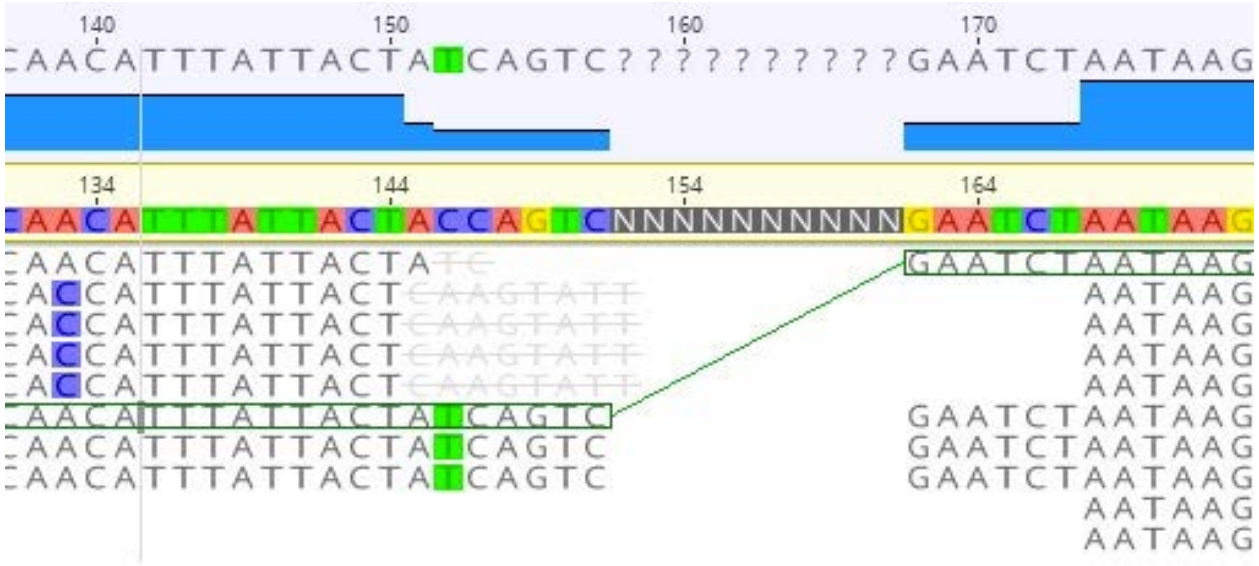
Niklaus Zemp
23 June 2020

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

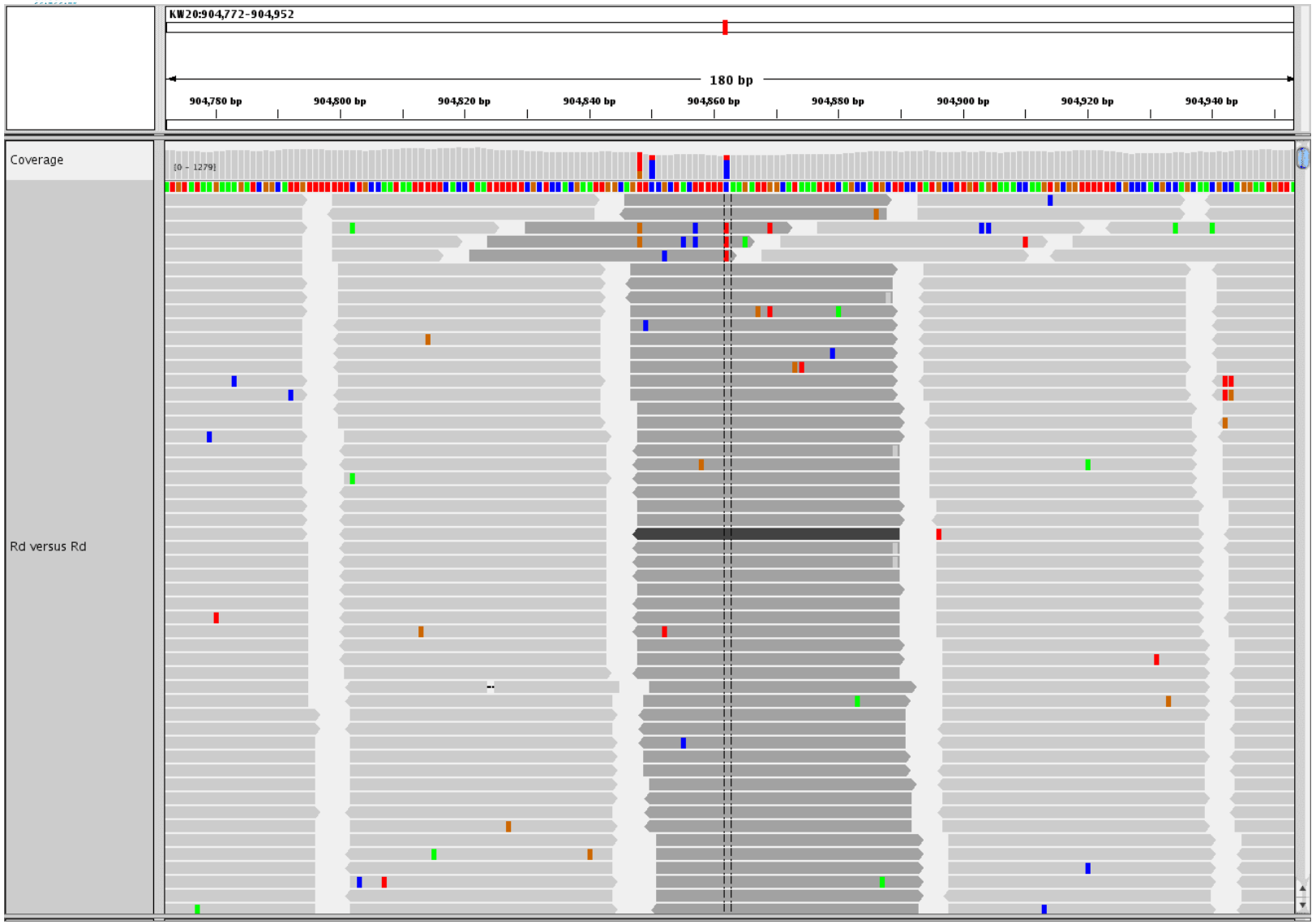
SNP caller



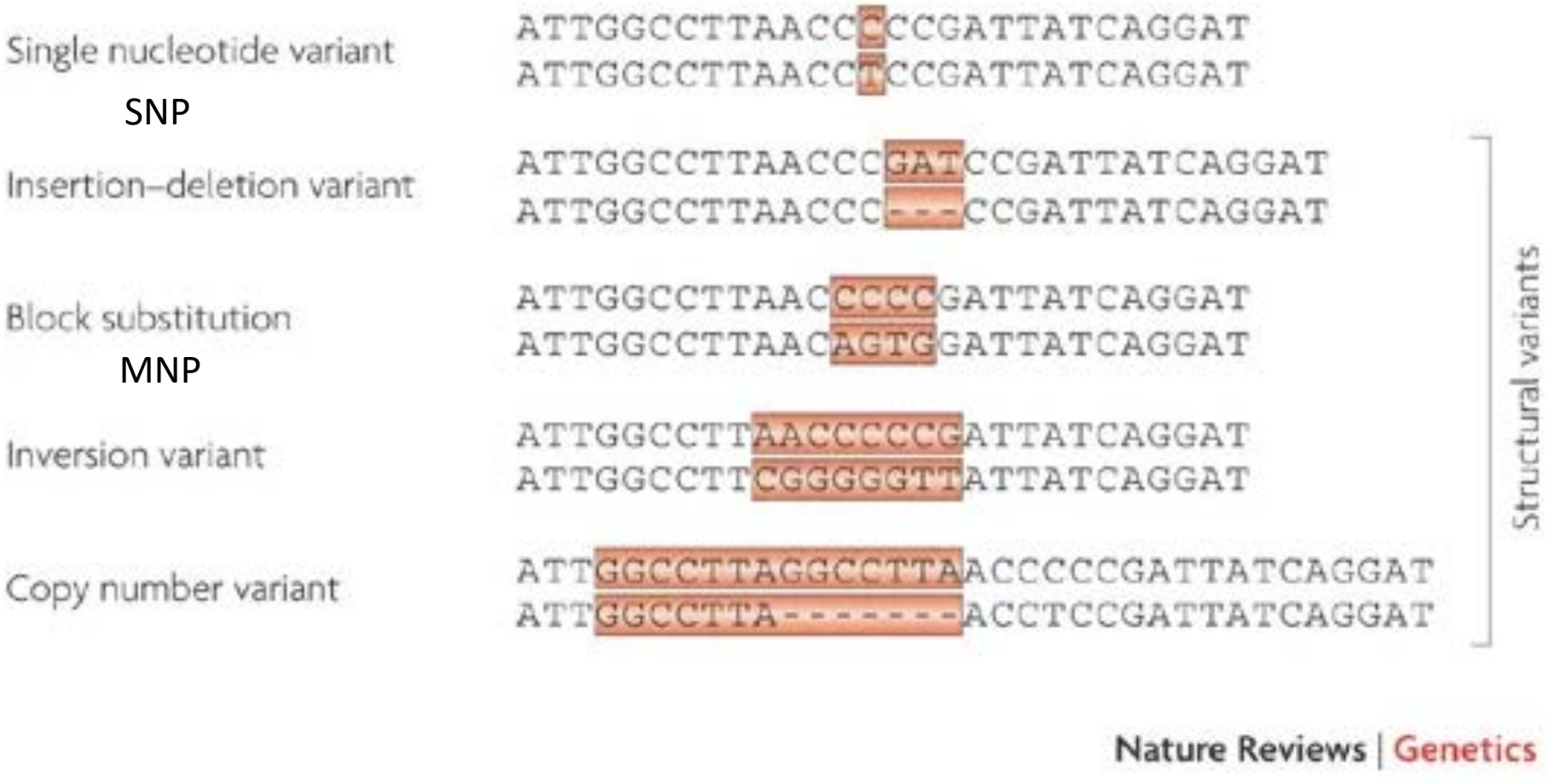
Alignment/mapping



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


KW20:904,772-904,952
180 bp
904,780 bp
904,800 bp
904,820 bp
904,840 bp
904,860 bp
904,880 bp
904,900 bp
904,920 bp
904,940 bp
Coverage
[0 - 1279]
Rd versus Rd

Structural variants

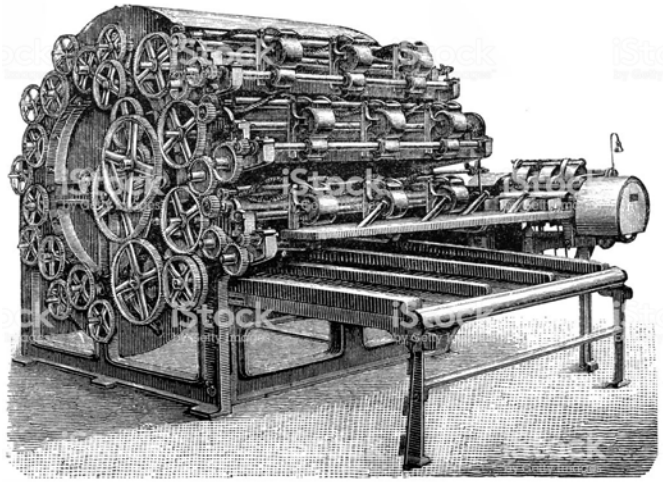


Variant types

- 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)

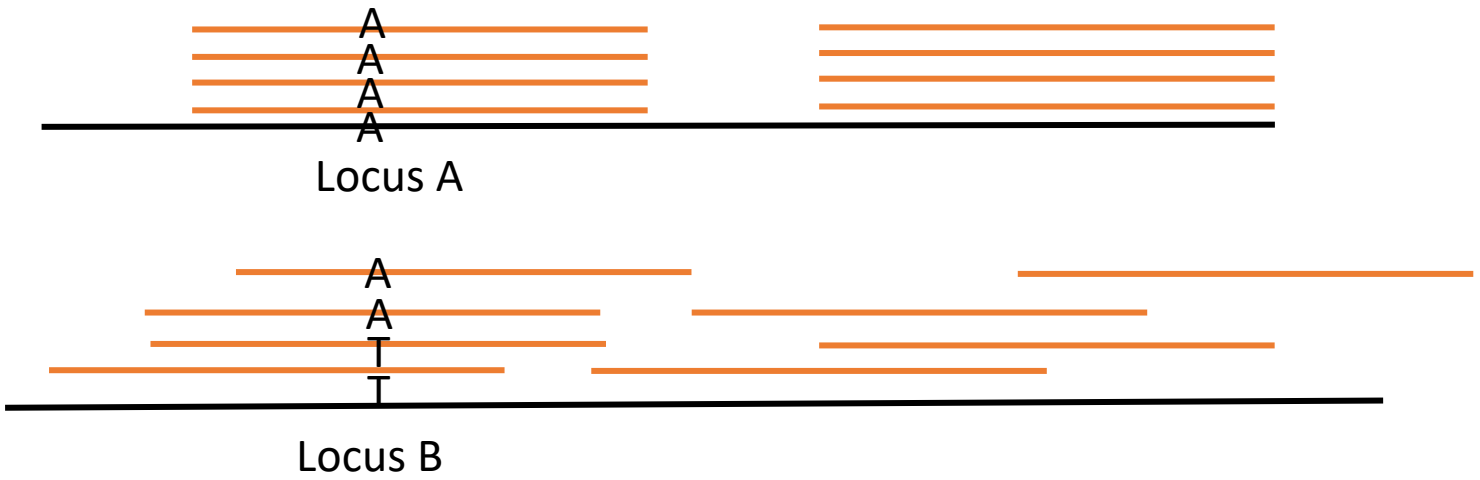
SNP caller

Alignment/mapping

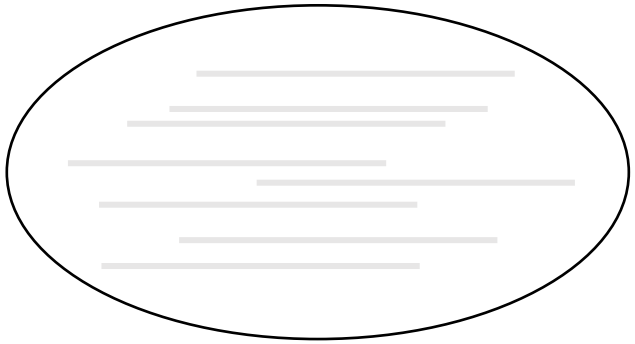


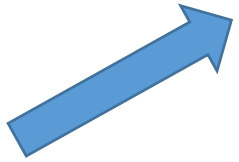
	Sample 1	Sample 2
Pos 1	AA	AT
Pos 2	TT	AA
Pos 3	AT	TT

PCR duplicates



SNP caller

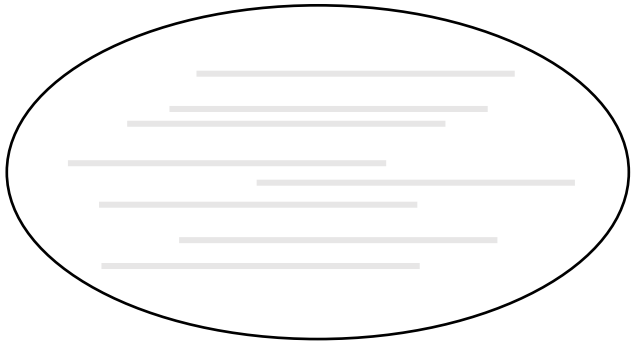


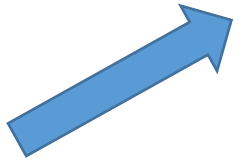


Local realignment

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

SNP caller





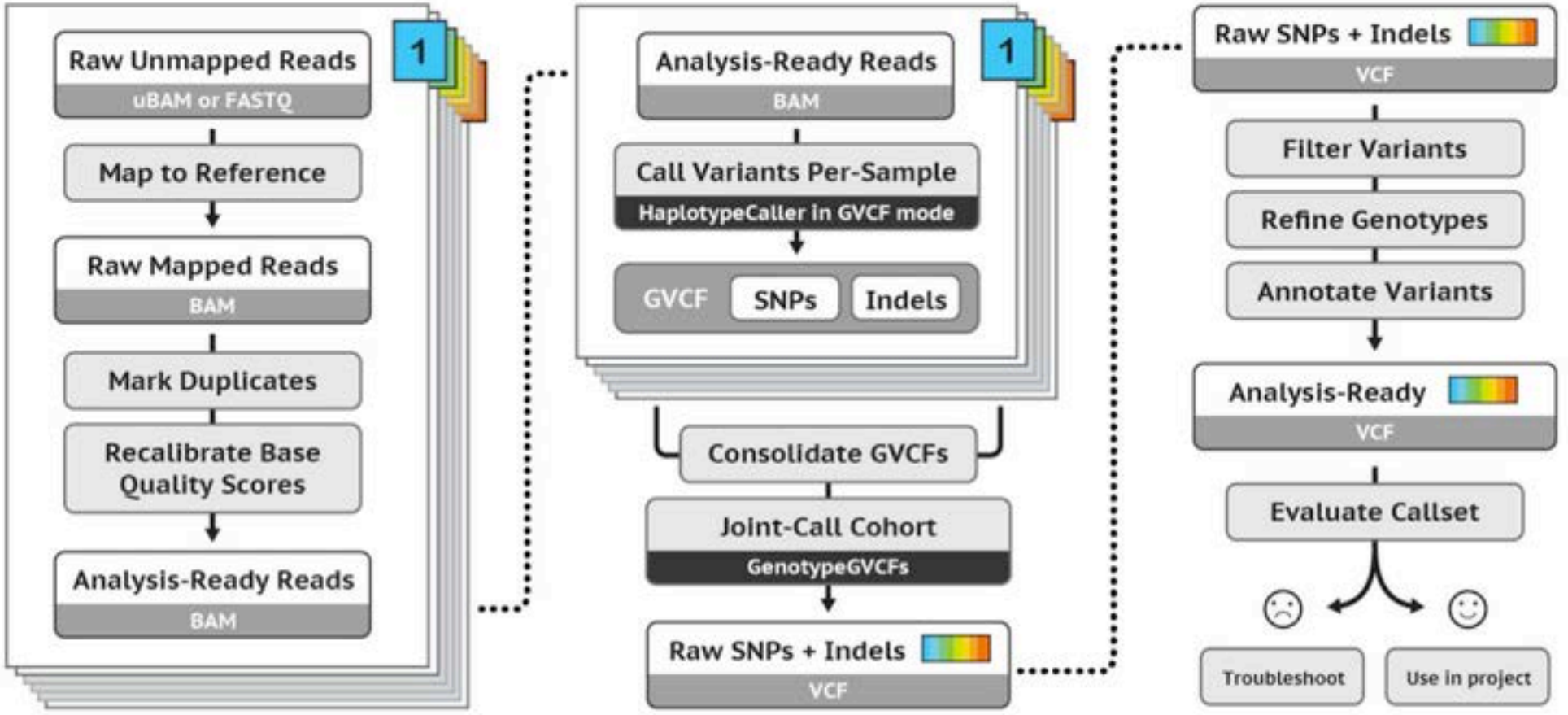
Joint calling

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

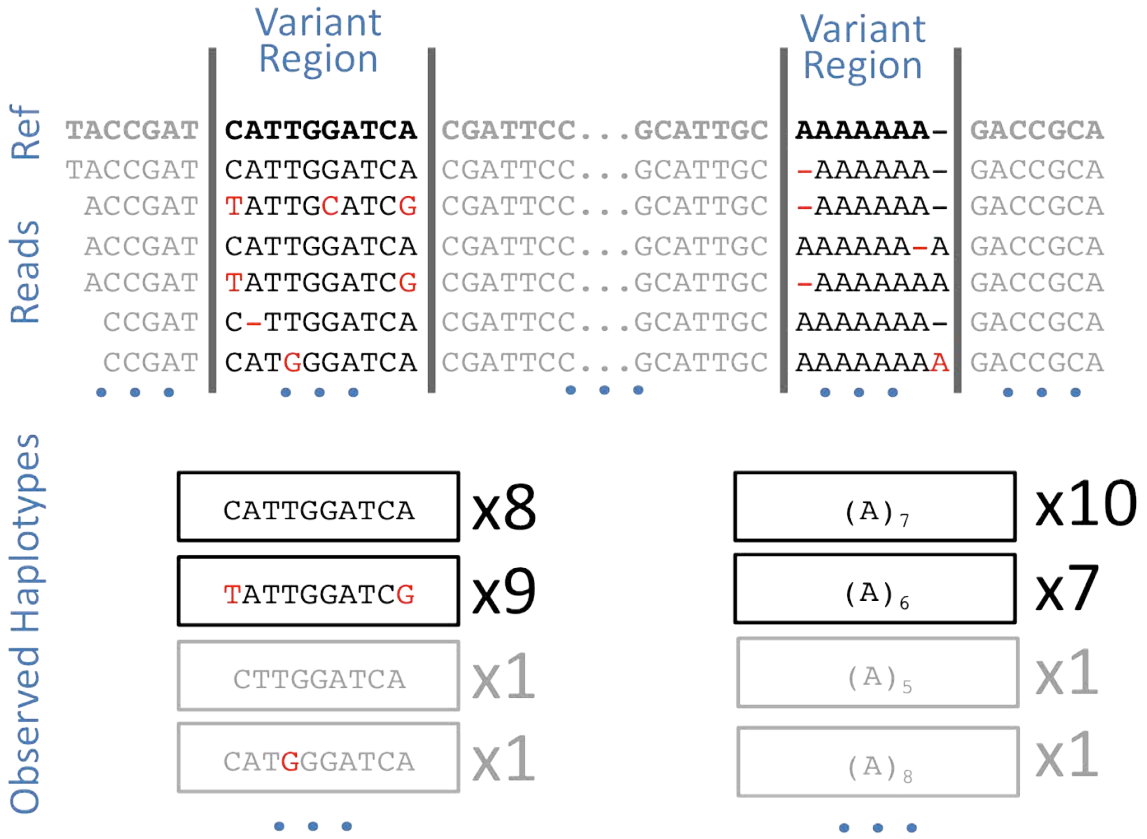
SNP callers

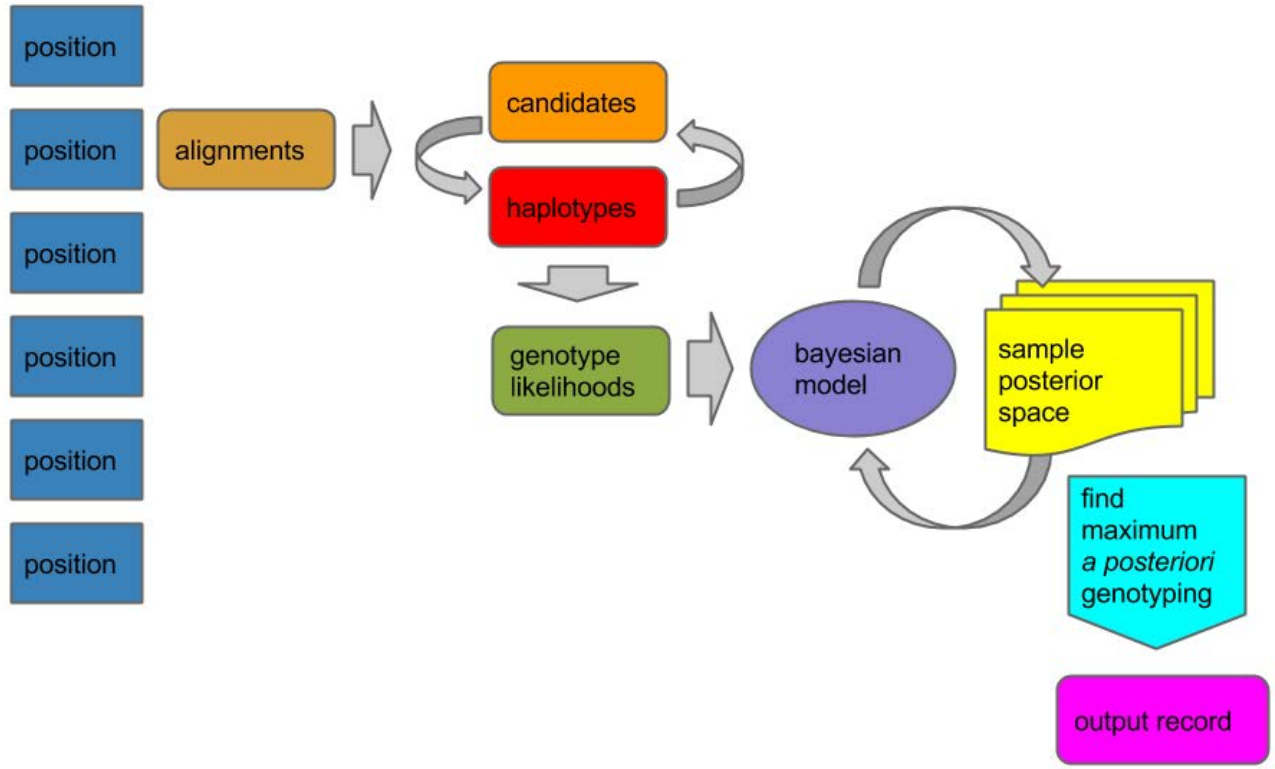
- Samtools
- GATK
- FreeBayes
- Platypus
- Popoolation

GATK


Raw Unmapped Reads
uBAM or FASTQ
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
Troubleshoot
Use in project

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.

Take home massage

- There are many SNPs callers available
- SNP calling is computational intensive
- Raw SNPs are full of false positives