Variant Calling with GATK

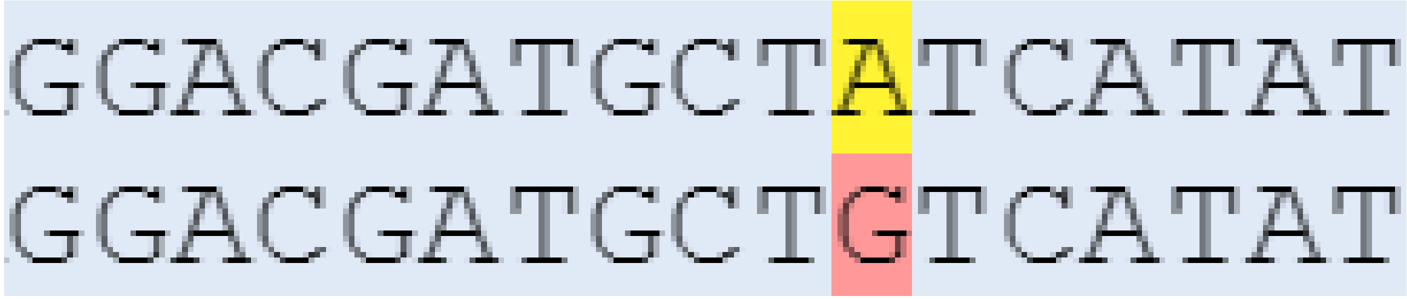
Compiled notes from
Mohammed Khalfan and
Jonathan Flowers

What is Variant Calling?

Identifying single nucleotide polymorphisms (SNPs) and small insertions and deletion (indels) from next generation sequencing data.

Plays an important role in scientific discovery.

Conceptually simple:



```
GGACGATGCTATCATAT
GGACGATGCTGTCATAT
```

Whole genome resequencing: SNP-calling

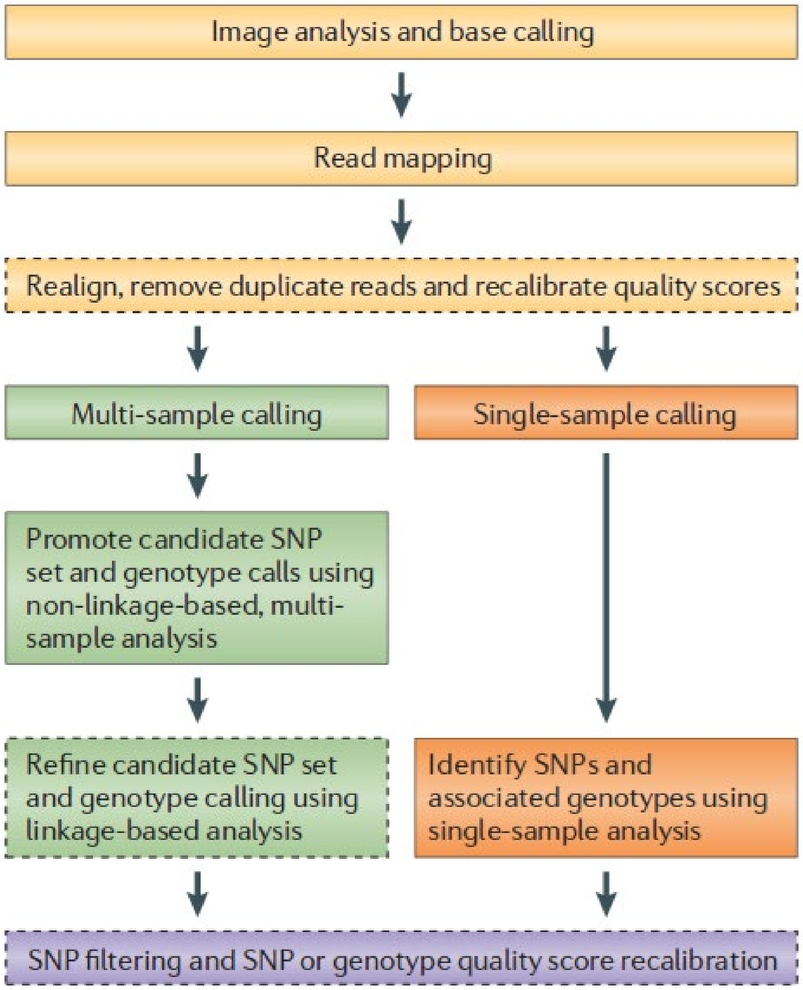


Image analysis and base calling
Read mapping
Realign, remove duplicate reads and recalibrate quality scores
Multi-sample calling
Single-sample calling
Promote candidate SNP set and genotype calls using non-linkage-based, multi-sample analysis
Refine candidate SNP set and genotype calling using linkage-based analysis
Identify SNPs and associated genotypes using single-sample analysis
SNP filtering and SNP or genotype quality score recalibration

← Requires assembled genome

← Steps to reduce the "false discovery rate" (FDR)

← Multi-sample algorithms improve SNP-calling in population studies

Mapped read (UnifiedGenotyepr) and de novo assembly-based (HaplotypeCaller) approaches to multi-sample snp calling

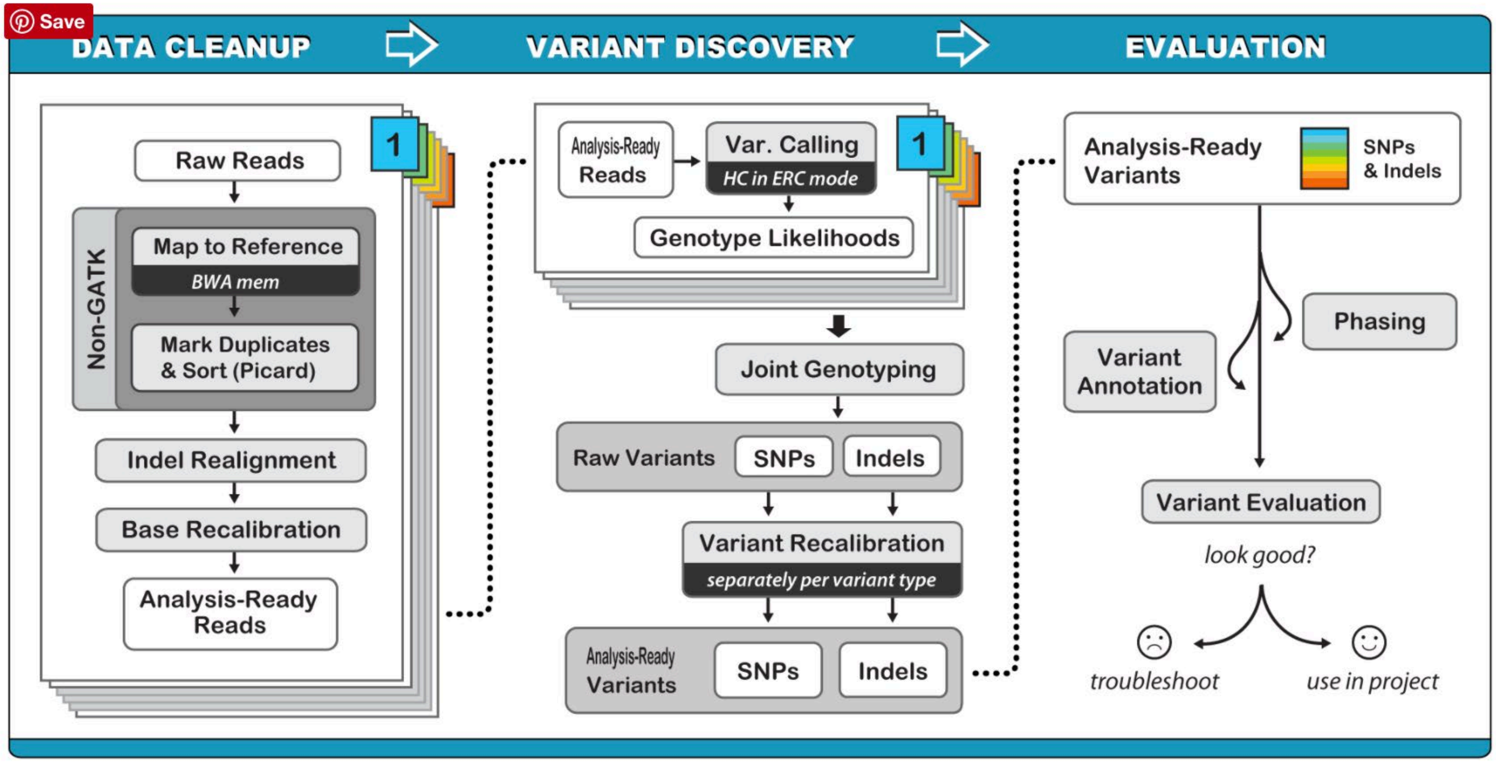
← Hard-filtering of suspect SNPs is required in most snp-calling and genotyping applications

Nat Rev Genet . 2011 June ; 12(6): 443–451. doi:10.1038/nrg2986

Genome Analysis Toolkit (GATK)

- Developed by the Broad Institute
- Industry Standard for identifying SNPs and indels in germline DNA and RNAseq data
- In addition to the variant callers themselves, GATK also includes many utilities to perform related tasks such as processing and quality control of high-throughput sequencing data.

Resequencing work flow (GATK v3.X best practices)**



**With some exceptions for non-human work flows

https://software.broadinstitute.org/gatk/documentation/article?id=3893

Resequencing work flow

- Prepare reference genome (e.g., index files)
- Process reads
- Align reads
- Coordinate sort reads
- Mark duplicate reads
- Re-alignment around insertions/deletions
- Base quality recalibration (human data only)
- SNP-calling
- Variant recalibration (human data only)
- Filtering / Quality control

Preparing the reference genome

- bwa index (see BWA)
- FASTA index (samtools faidx)
- GATK index (GATK CreateSequenceDictionary)

Read quality assessment

- Base qualities decay with advancing sequencing cycle in reads generated sequencing-by-synthesis (Illumina)

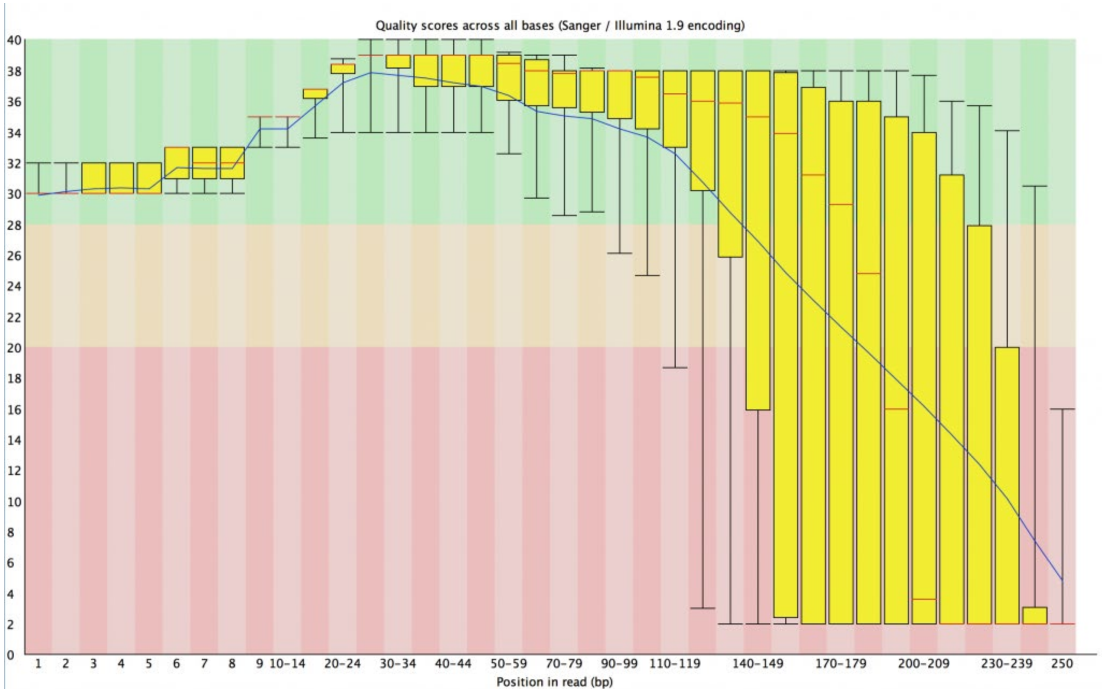
- Errors increase the edit or hamming distance between a read and reference)

- Low quality bases (e.g., PHRED < 20) can lead to lower mapping rates or artefacts

- Many work flows trim using various sliding window, or fixed length methods

- Example:

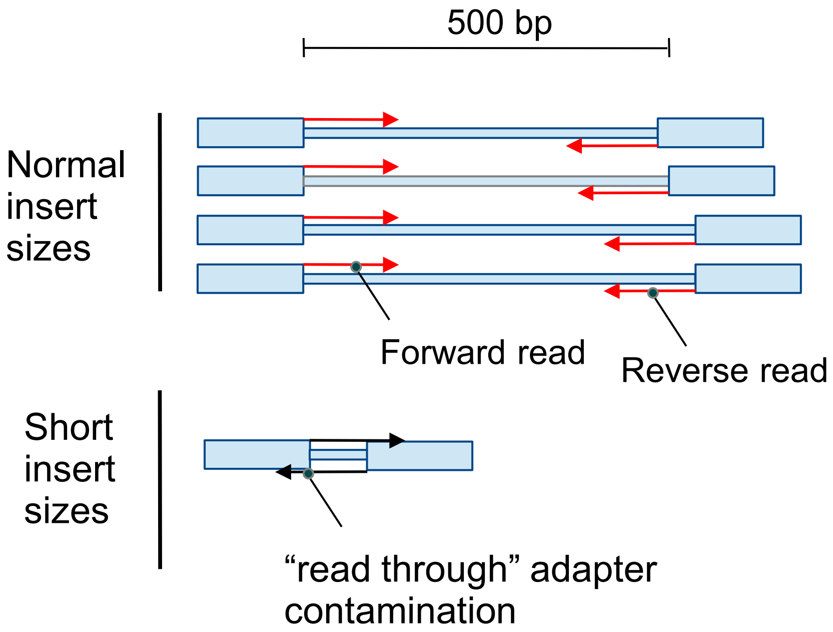
Trimmomatic sliding window quality filtering + removal of leading/trailing bases below some PHRED QUAL threshold



Adapter cleaning

- During library preparation, genomic DNA is fragmented and flanking adapter sequences attached
- short read library insert sizes are typically at least 400-600 bp in length (or greater for longer read libraries)
- Atypically short inserts can lead to “read through” contamination
- Adapters should be removed prior to read-mapping in most applications (esp. de novo assembly)
- Removal adapters is technically challenging because of sequencing errors and partial sequences

How? Trimmomatic for Illumina HiSeq data



Short read alignment (BWA)

- BWA MEM

  Recommended for read lengths > 75 bp

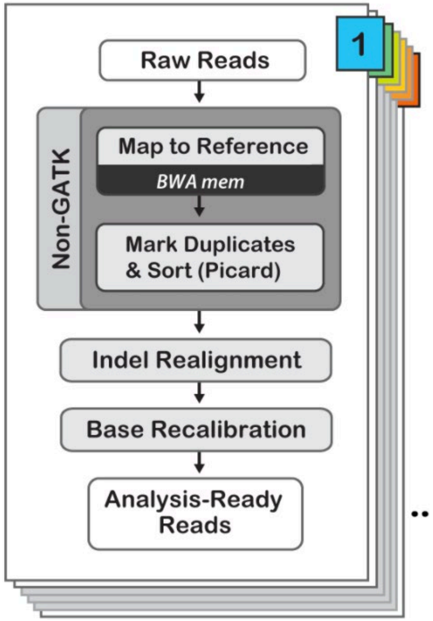
  Produces chimaeric alignments

  A "promiscuous" mapper

  Multi-reads are randomly assigned to targets

  Secondary/Supplementary segments from chimaeric alignments (marked in BAM) can be excluded in SNP-calling steps with GATK

  Add read groups with the -R option



Read Groups

- Multi-sample/multi-library BAMs are common in resequencing projects
- Need means of tracking which sample/library a read came from
- Meta-information for each read group stored in @RG header
- Each read group must have a unique identifier
- Each read is assigned an RG tag with a read group identifier
- Example:

@RG Header line (one for each sample/library)

```
@RG     ID:CR2342       PL:Illumina     LB:CR2342
    DS:/scratch/jmf11/chlamy/fastqs/CR2342/CR2342-
    1_I07CAGATC_CAGATC_L006_R1_001.fastq.gz_/scratch/jmf11/chlamy/fastqs/CR2342/
    CR2342-1_I07CAGATC_CAGATC_L006_R2_001.fastq.gz_       SM:CR2342
```

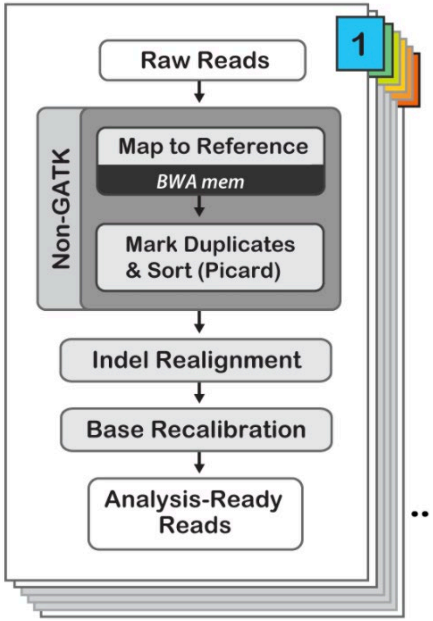
```
HWI-ST911:113:C0MK8ACXX:4:2105:4811:106636      163     chromosome_1     4       0
   101M    =       247     344  …  ...  X0:i:70 MD:Z:101       RG:Z:CR2342      XG:i:0
   AM:i:0  NM:i:0  SM:i:0  XM:i:0  XO:i:0  MQ:i:0  XT:A:R
```

Coordinate sorting SAM/BAM

- BAM alignments may either be unsorted, coordinate-sorted, or sorted on read name (fastq identifier)
- Sorting operations are best performed using Picard-tools SortSam
- Memory intensive
- Creates many temporary files
- How do you know if a BAM is coordinate sorted?

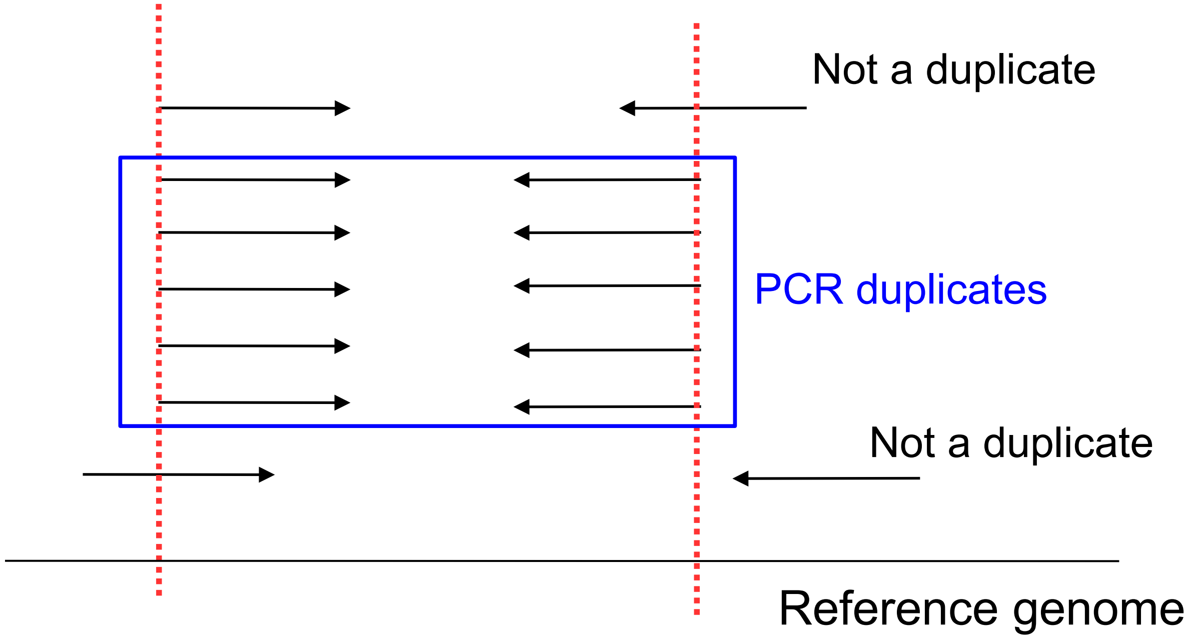
Check the BAM header @HD (frequently first line of SAM)

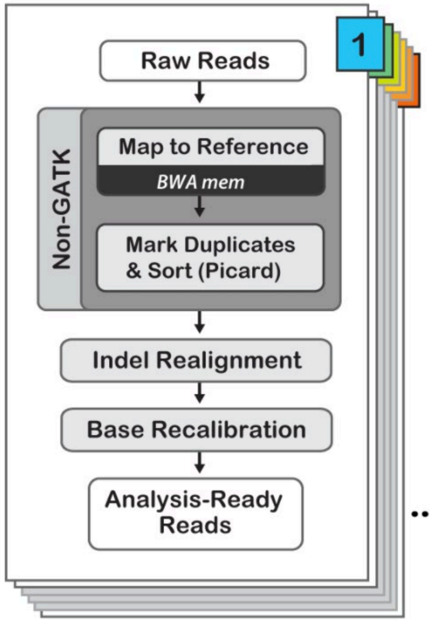
@HD SO:coordinate



Mark Duplicates: Handling PCR duplicates

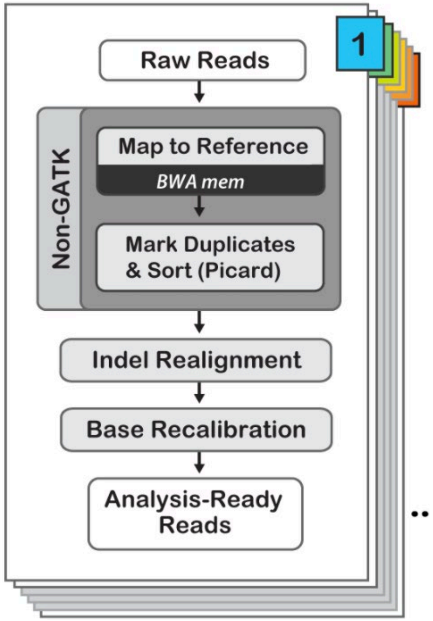
- What is a PCR duplicate?

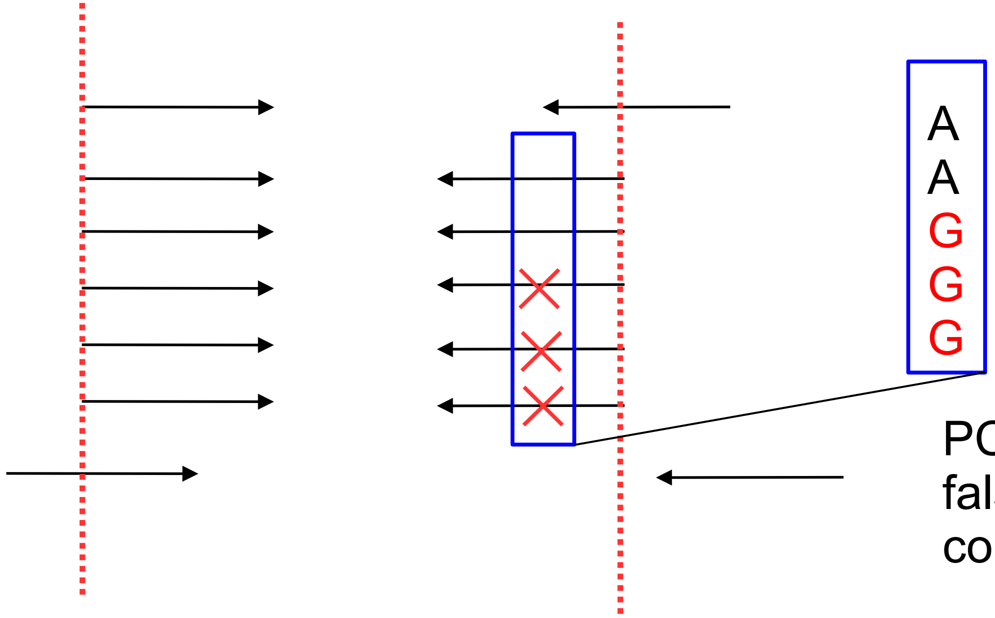




Mark Duplicates: Handling PCR duplicates

- What is a PCR duplicate?
- Why are duplicates a problem?



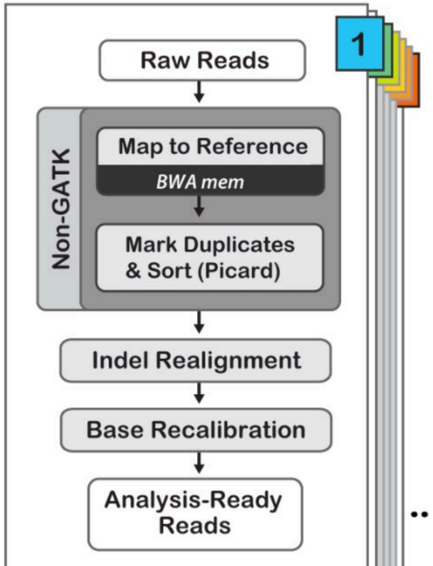


PCR duplicates can introduce false positive SNP and contribute to genotyping error

✕ PCR error

Indel Realignment

- Why is it necessary to re-align reads?
- Realignment refines insertion-deletion positioning and improves base quality recalibration (and reduction in false positive SNPs in UnifiedGenotyper work flow)
- GATK IndelRealignerTargetCreator / IndelRealigner



1

Raw Reads

Non-GATK

Map to Reference

BWA mem

Mark Duplicates & Sort (Picard)

Indel Realignment

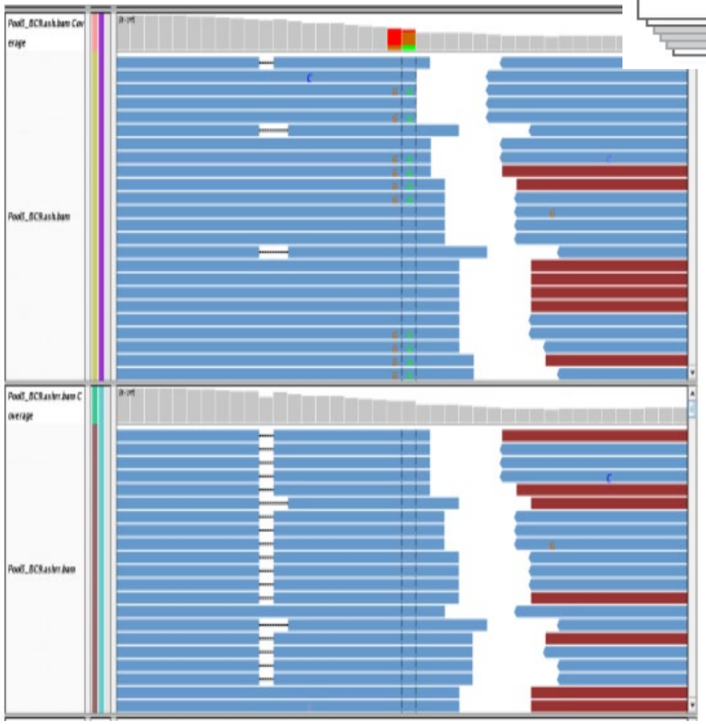
Base Recalibration

Analysis-Ready Reads

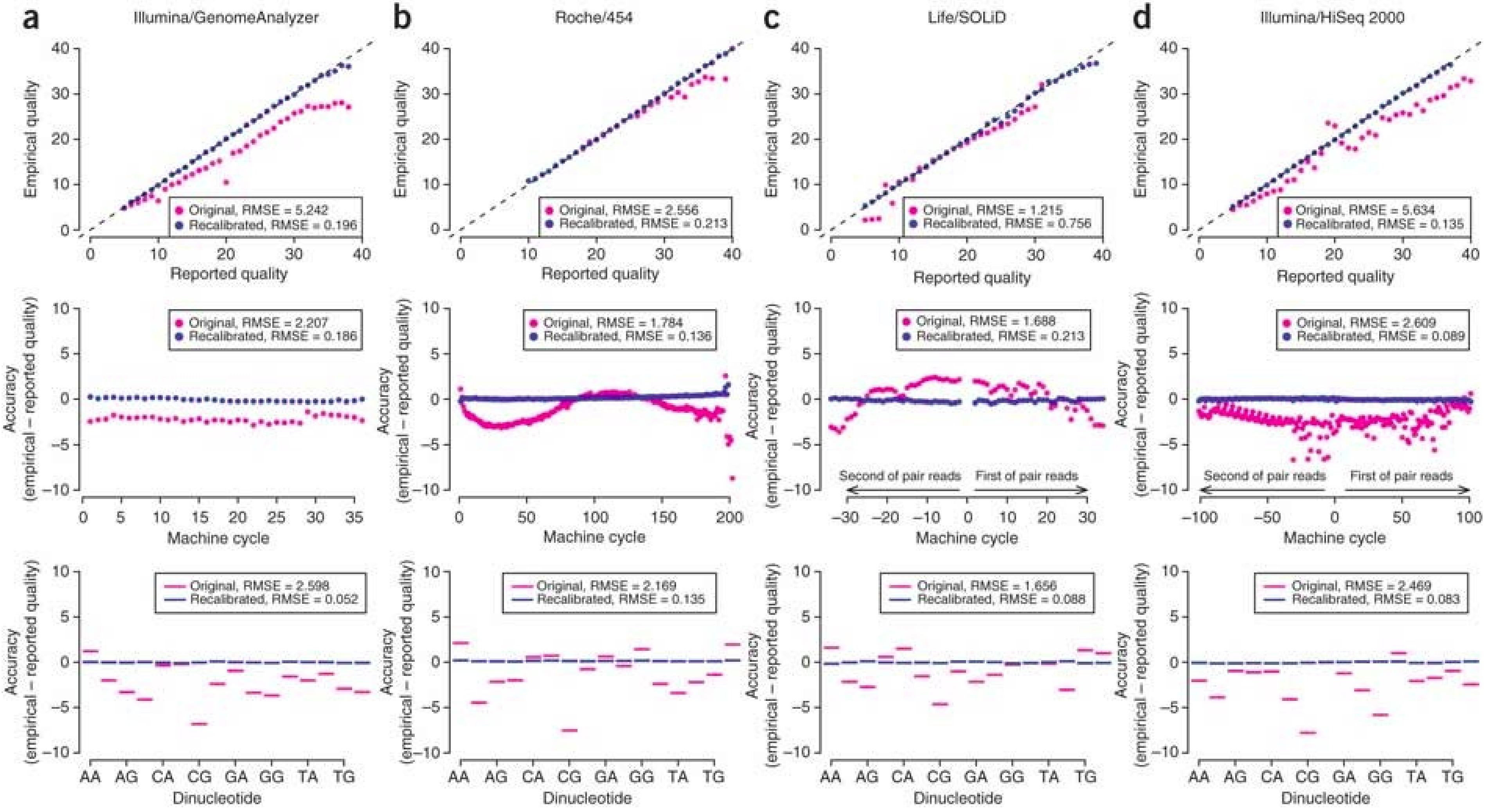
..

Before:

After:



Base Quality Score Recalibration (BQSR)



DePristo et al. (2011)

Base Quality Score Recalibration (BQSR)

- Why base quality recalibration?

  Base qualities in fastqs are not calibrated (i.e, they are inaccurate)

  SNP and Genotype likelihood models use base error probabilities from individual reads to determine SNPs/genotypes

- How is recalibration performed?

  Recalibration requires large database of high quality SNPs

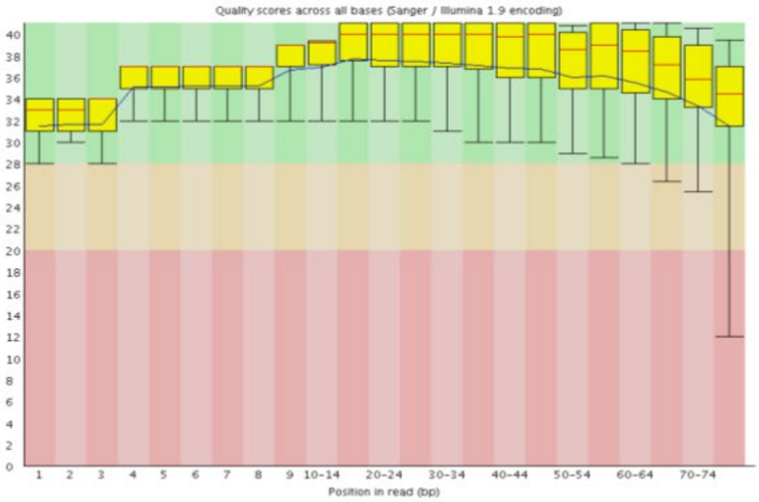
DePristo et al. (2011)

How are base qualities recalibrated?

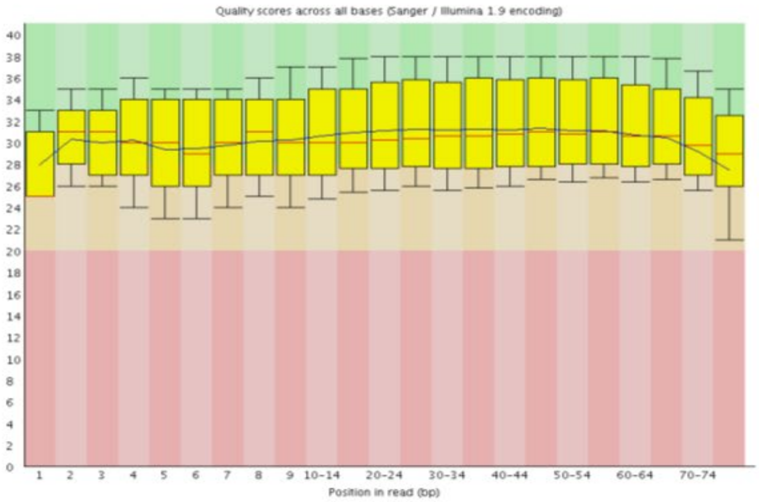
- Start with dbsnp (NCBI), hapmap or other high quality snp database
- For each read in SAM/BAM, identify mismatches with reference
- Determine if mismatch in read is in snp database
- If mismatch is not in dbsnp, then mismatch is considered an error
- Update empirical error rate for sequencing cycle, dinucleotide context etc.
- Use empirical estimates of the error rate to build model and adjust base quality scores in SAM/BAM. (Implemented with GATK BaseRecalibrator tool)

Base qualities before and after BQSR

Before



After

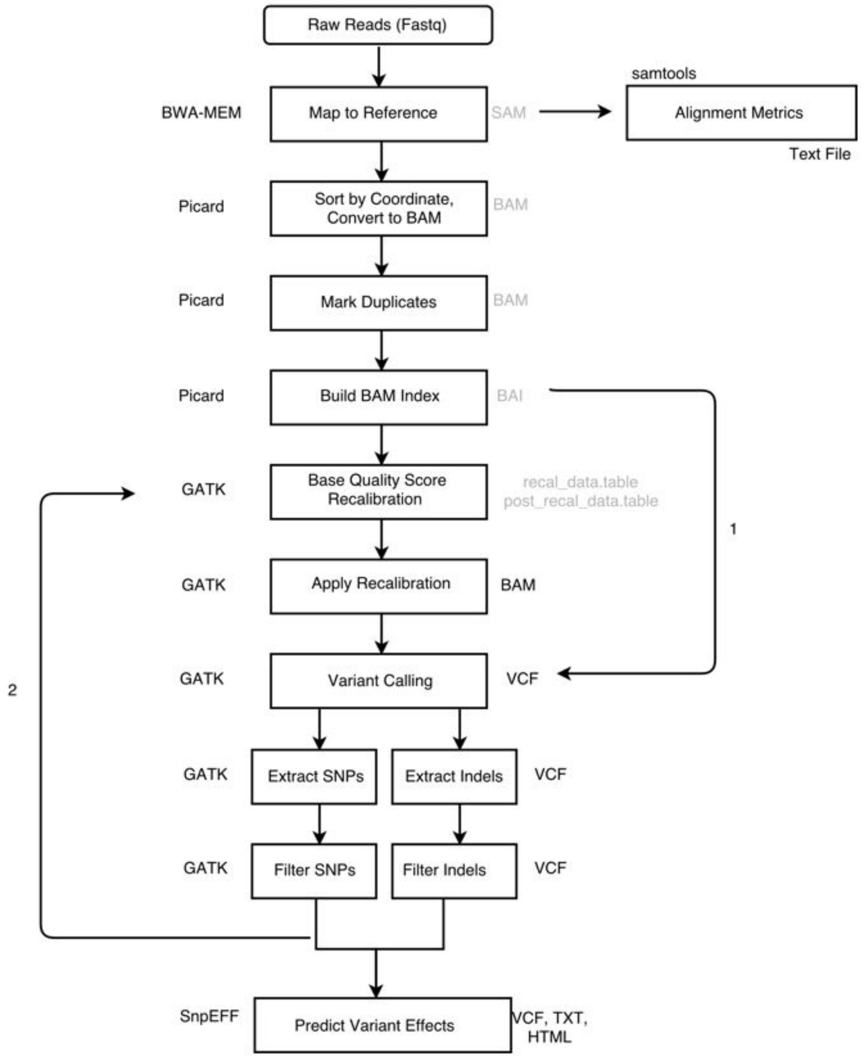


Variant Quality Score Recalibration (VQSR) (human only)

- Why variant quality recalibration?
- Errors due to systematic machine artifacts, library prep, SNP-calling, alignment
- Uses database of SNPs to train the recalibration model
- Allows rigorous assessment of specificity and sensitivity in a call set
- Alleviates strong dependence on hard-filtering, eliminates arbitrariness of thresholds
- Yields a log odds for each SNP that it is true

DePristo et al. 2011

What if you don’t have a set of known variants?



SNP-calling methods

- Consensus methods

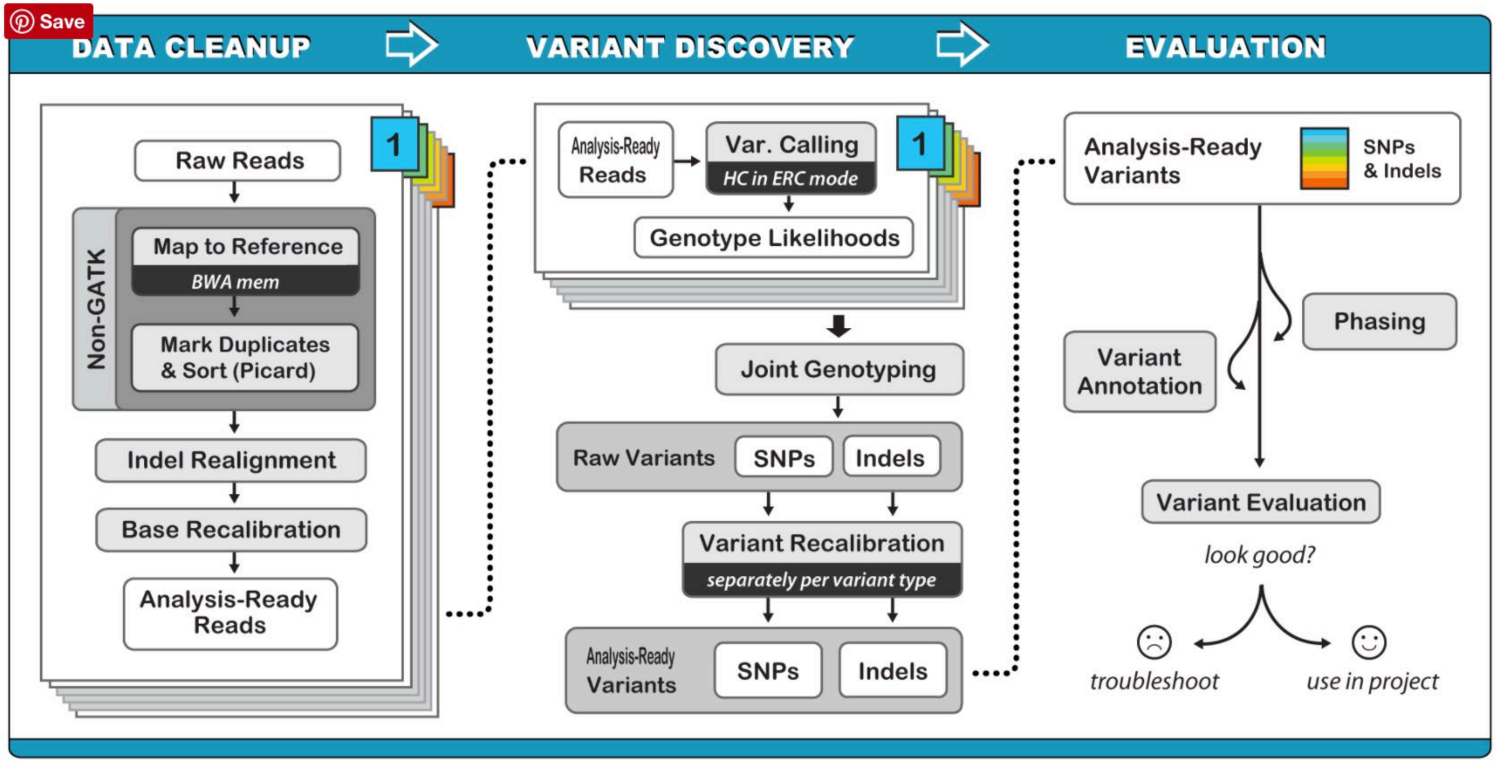
Example:

1. drop reads with base quality at focal position < Q20

2. if 20% to 80% of Q20+ reads support alternate allele then genotypes is heterozygous otherwise homozygous

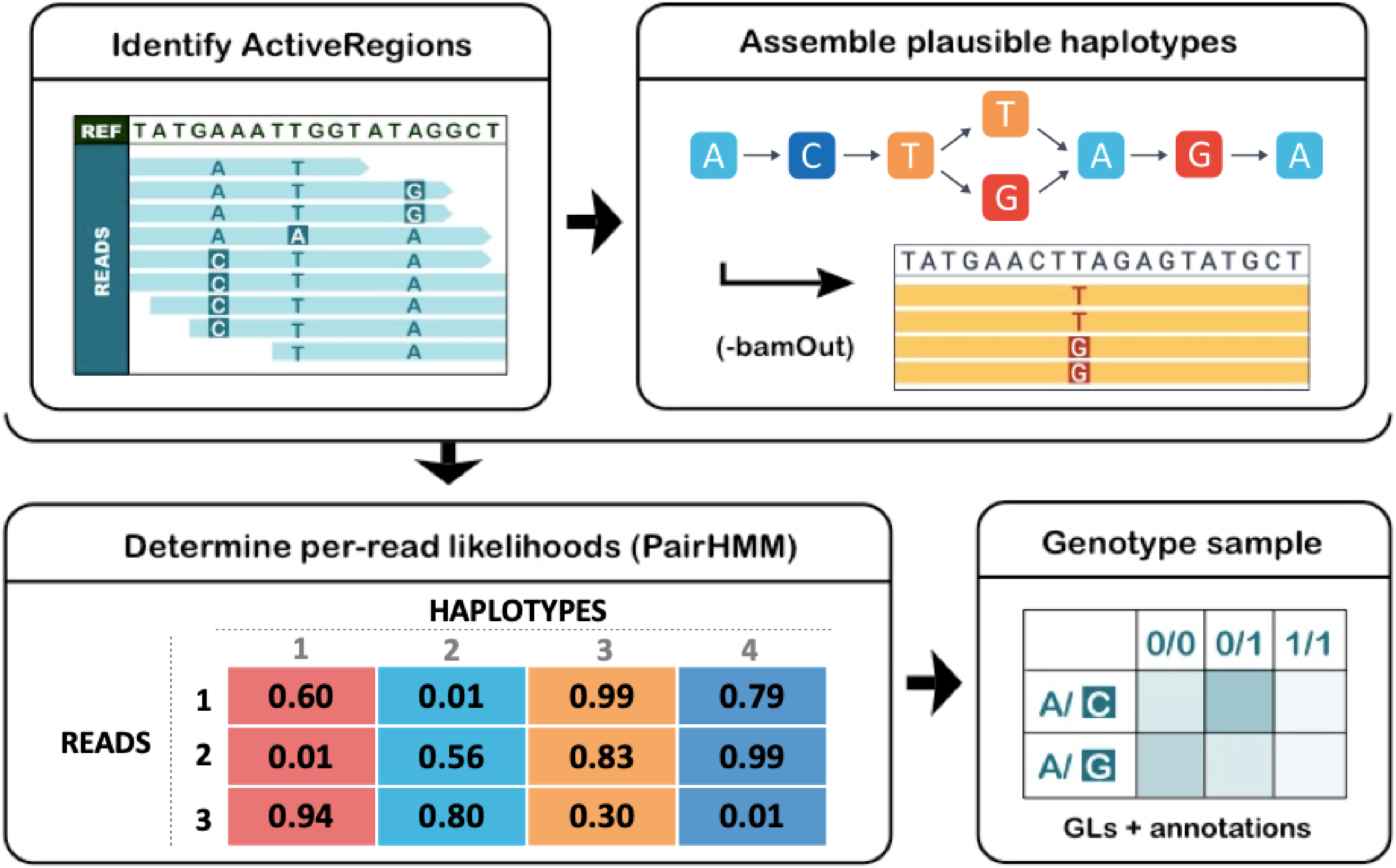
Single calling, joint genotyping with the Haplotype Caller

- Multi-sample SNP calling (Unified Genotyper, UG) does not scale well.
- Invariants sites poorly modeled by UG
- UG from poor indel calls and snp-calling errors around indels
- UG Suffers from the N+1 problem



https://software.broadinstitute.org/gatk/documentation/article?id=3893

How does the HaplotypeCaller work?



Identify ActiveRegions

Assemble plausible haplotypes

(-bamOut)

Determine per-read likelihoods (PairHMM)

		HAPLOTYPES			
		1	2	3	4
READS	1	0.60	0.01	0.99	0.79
	2	0.01	0.56	0.83	0.99
	3	0.94	0.80	0.30	0.01

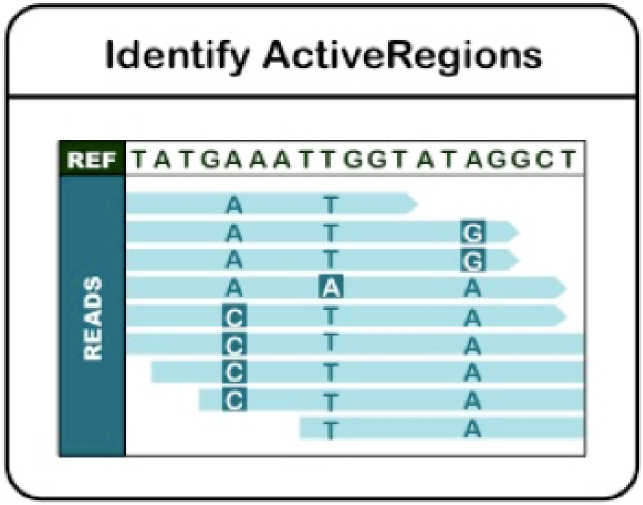
Genotype sample

	0/0	0/1	1/1
A/ C			
A/ G			

GLs + annotations

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

How does the HaplotypeCaller work?

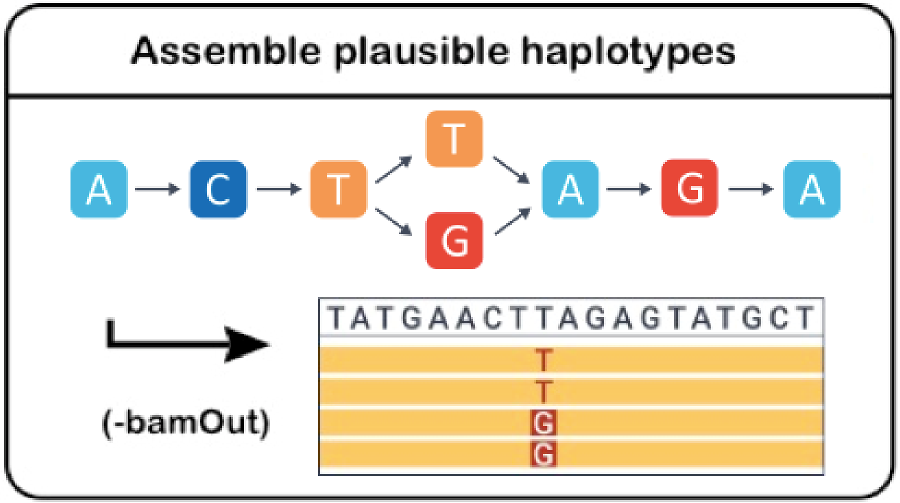


Identify “active” regions.

An active region is defined as an area that contains variation based on sequence alignment.

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

How does the HaplotypeCaller work?



Assemble haplotypes in the active regions.

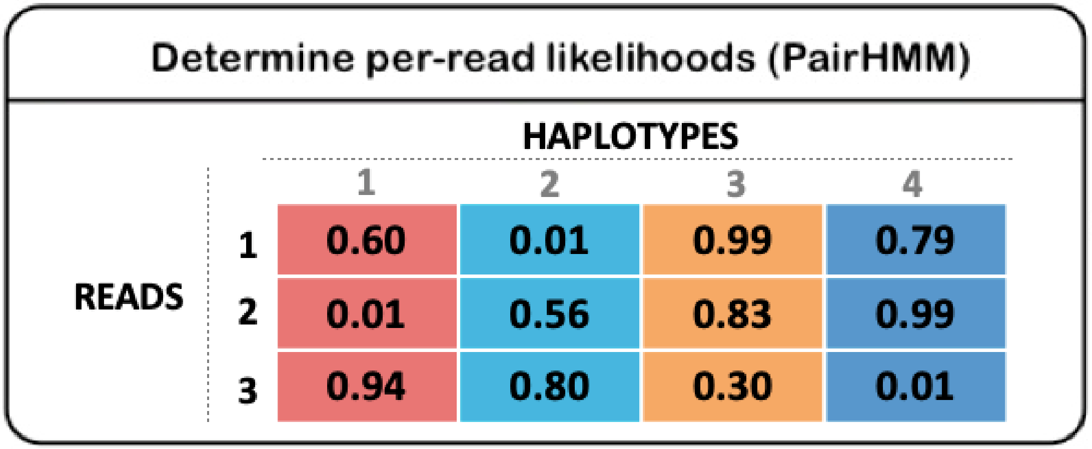
Assemble the reads to create the different possible haplotypes

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

How does the HaplotypeCaller work?

- Determine likelihoods of each haplotype given the reads

Performs a pairwise alignment, using PairHMM, of each read against the different haplotypes to create a matrix that represents the likely of each read to each haplotype.



Determine per-read likelihoods (PairHMM)

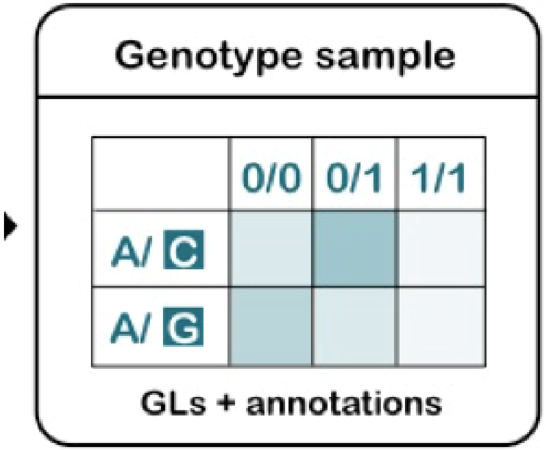
		HAPLOTYPES			
		1	2	3	4
READS	1	0.60	0.01	0.99	0.79
	2	0.01	0.56	0.83	0.99
	3	0.94	0.80	0.30	0.01

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

How does the HaplotypeCaller work?

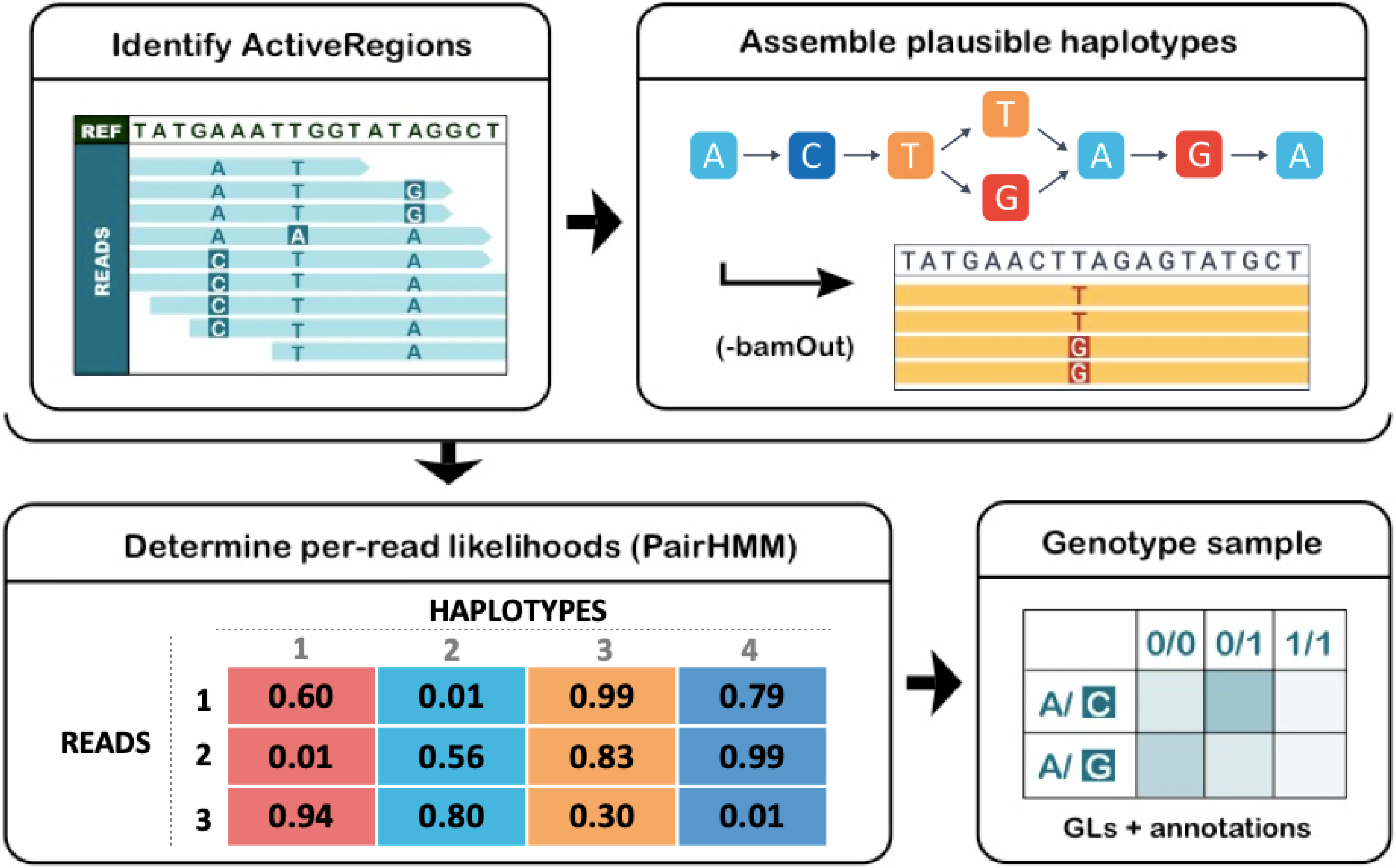
- Each variant gets assigned a genotype

Bayes rule is applied to calculate the likelihoods of each genotype based on the likelihood of the alleles



https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

How does the HaplotypeCaller work?

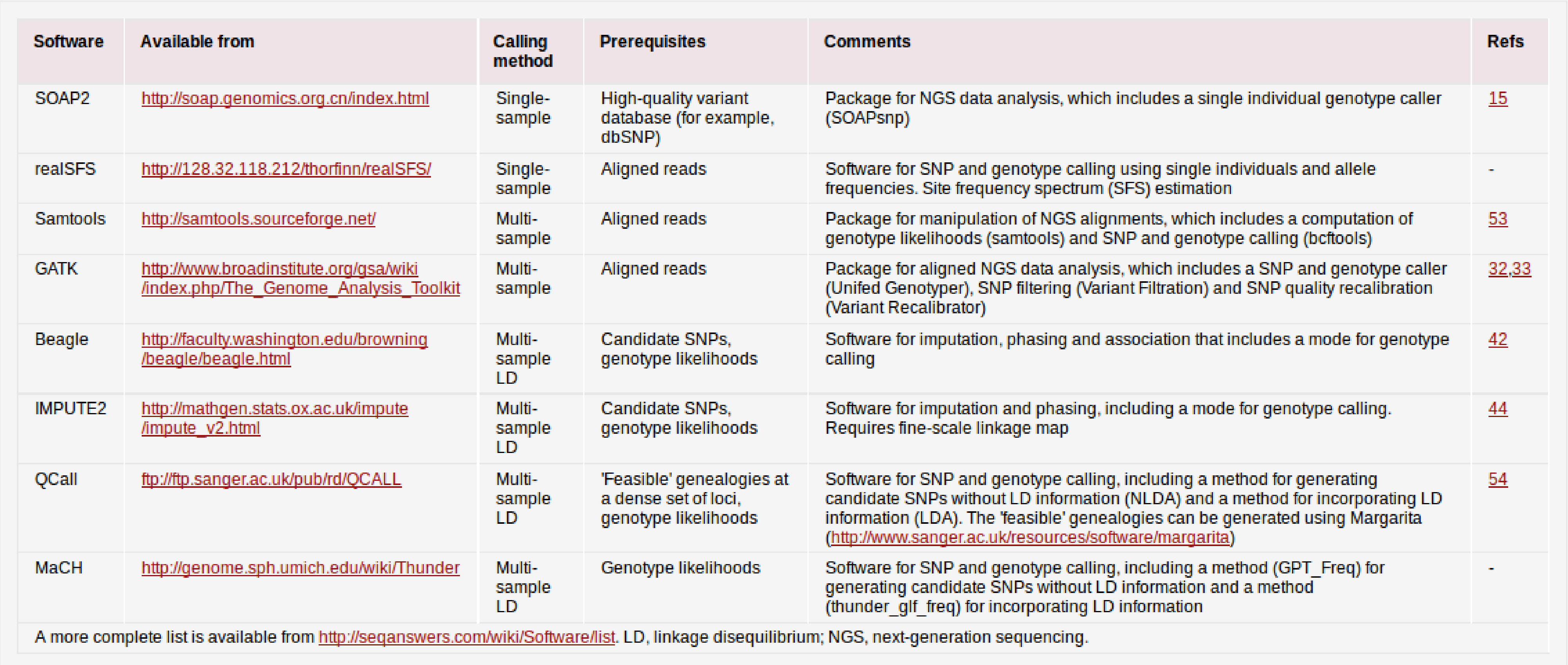


		HAPLOTYPES			
		1	2	3	4
READS	1	0.60	0.01	0.99	0.79
	2	0.01	0.56	0.83	0.99
	3	0.94	0.80	0.30	0.01

	0/0	0/1	1/1
A/ C			
A/ G			

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412?id=11068

SNP-calling methods



Software	Available from	Calling method	Prerequisites	Comments	Refs
SOAP2	http://soap.genomics.org.cn/index.html	Single-sample	High-quality variant database (for example, dbSNP)	Package for NGS data analysis, which includes a single individual genotype caller (SOAPsnp)	15
realSFS	http://128.32.118.212/thorfinn/realSFS/	Single-sample	Aligned reads	Software for SNP and genotype calling using single individuals and allele frequencies. Site frequency spectrum (SFS) estimation	-
Samtools	http://samtools.sourceforge.net/	Multi-sample	Aligned reads	Package for manipulation of NGS alignments, which includes a computation of genotype likelihoods (samtools) and SNP and genotype calling (bcftools)	53
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Multi-sample	Aligned reads	Package for aligned NGS data analysis, which includes a SNP and genotype caller (Unifed Genotyper), SNP filtering (Variant Filtration) and SNP quality recalibration (Variant Recalibrator)	32,33
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation, phasing and association that includes a mode for genotype calling	42
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation and phasing, including a mode for genotype calling. Requires fine-scale linkage map	44
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	Multi-sample LD	'Feasible' genealogies at a dense set of loci, genotype likelihoods	Software for SNP and genotype calling, including a method for generating candidate SNPs without LD information (NLDA) and a method for incorporating LD information (LDA). The 'feasible' genealogies can be generated using Margarita (http://www.sanger.ac.uk/resources/software/margarita)	54
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Multi-sample LD	Genotype likelihoods	Software for SNP and genotype calling, including a method (GPT_Freq) for generating candidate SNPs without LD information and a method (thunder_glf_freq) for incorporating LD information	-

A more complete list is available from http://seqanswers.com/wiki/Software/list. LD, linkage disequilibrium; NGS, next-generation sequencing.

Newer methods include UnifiedGenotyper, HaplotypeCaller, Platypus

Nielsen et al. 2011

Primary sources of error in snp-calling and genotyping

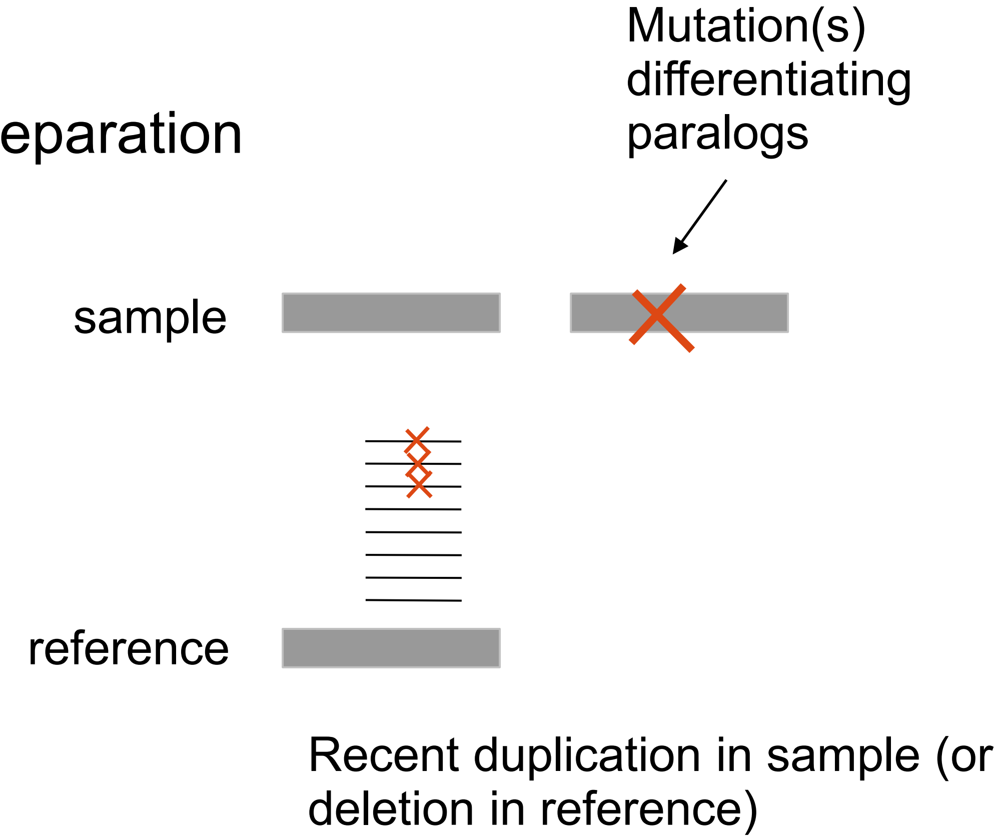
- Errors introduced during library preparation
- Base-calling errors
- Read mapping errors and biases

a) paralogous regions

b) mis-alignment around indels

c) multiple-mapping

d) mapping bias favoring reference alleles over non-reference alleles


Mutation(s) differentiating paralogs
sample
reference
Recent duplication in sample (or deletion in reference)

How to avoid errors in snp-calling and genotyping

- Re-alignment around short insertions / deletions (e.g., GATK IndelRealignmentTargetCreator/IndelRealigner)
- Use base alignment quality (BAQ) to cap base qualities at sites close to indel polymorphisms
- Ignore multiply-mapped reads during SNP-calling
- Filter SNPs in low complexity regions
- Filter SNPs with anomalously low or high read depth
- Apply filters to remove problematic reads

Filtering protocols adopted in non-human studies

Rice, 15X coverage (Xu et al. 2013)

Type	Filter	Threshold
variant	variant quality	PHRED quality > 20
variant	depth	Covered by ≥ 1 uniquely mapped read in each sample
variant	Hardy-Weinberg	Must be in HWE
genotype	genotype quality	PHRED quality > 20
genotype	Rank sum test	P > .05

Filtering protocols adopted in non-human studies

Plasmodium falciparum >16X (Manske et al. 2012)

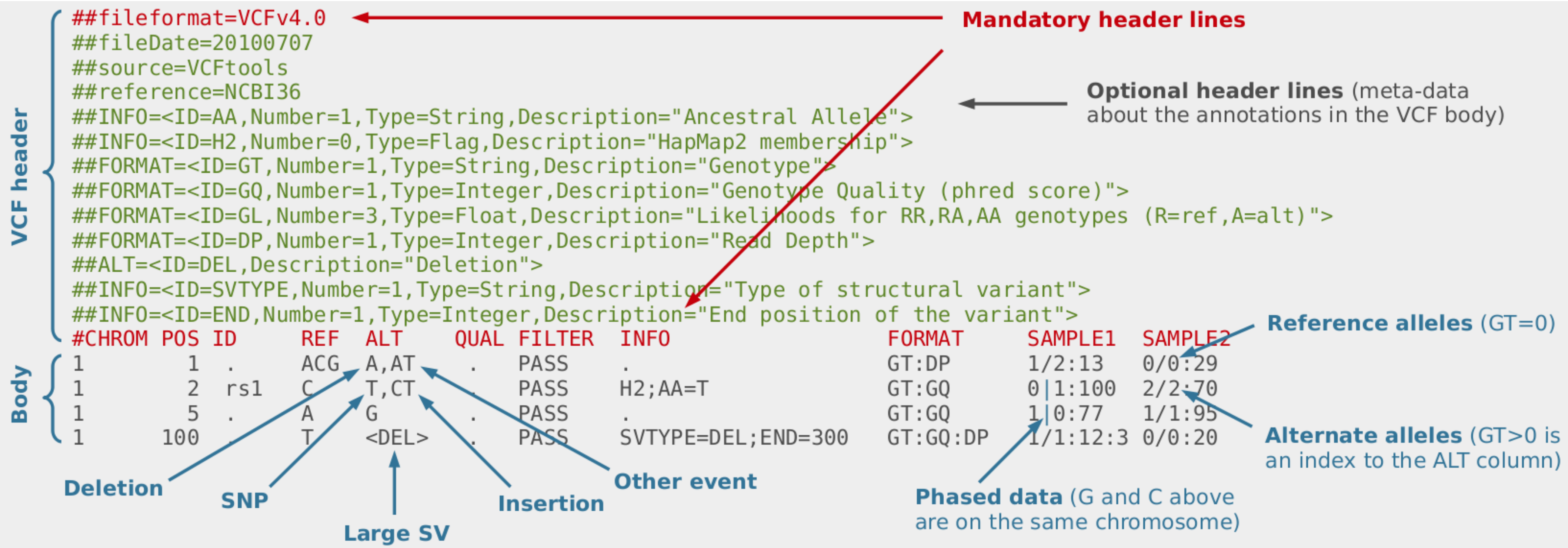
Type	Filter	Threshold
variant	coding/noncoding	All non-coding removed
variant	low population frequency	1% of reads across all samples must contain minor allele
variant	depth	Minor allele must be found at depth > 10 in at least 1 sample
variant	biallelic	SNPs must have 2 alleles only
variant	uniqueness score of flanking bases in ref	SNPs filtered if uniqueness score ≥ 26
variant	missingness	SNPs with fewer than 220 samples at 5X were filtered
variant	hyper-heterozygosity	Population specific cutoffs for variants exceeding HWE-based heterozygosity

Filtering protocols adopted in non-human studies

Sorghum 16-45X (Mace et al. 2013)

Type	Filter	Threshold
variant	variant quality	PHRED quality > 20
variant	SNP clusters	Maximum of 1 SNP per 5 bp*
variant	depth	100 < depth < 1300
variant	copy number	< 1.5
genotype	genotype quality	PHRED quality > 20
genotype	depth	4 < depth < 100
genotype	copy number of flanking	< 1.5
genotype	Rank sum test	P > .05

Variant filtering with VCF files

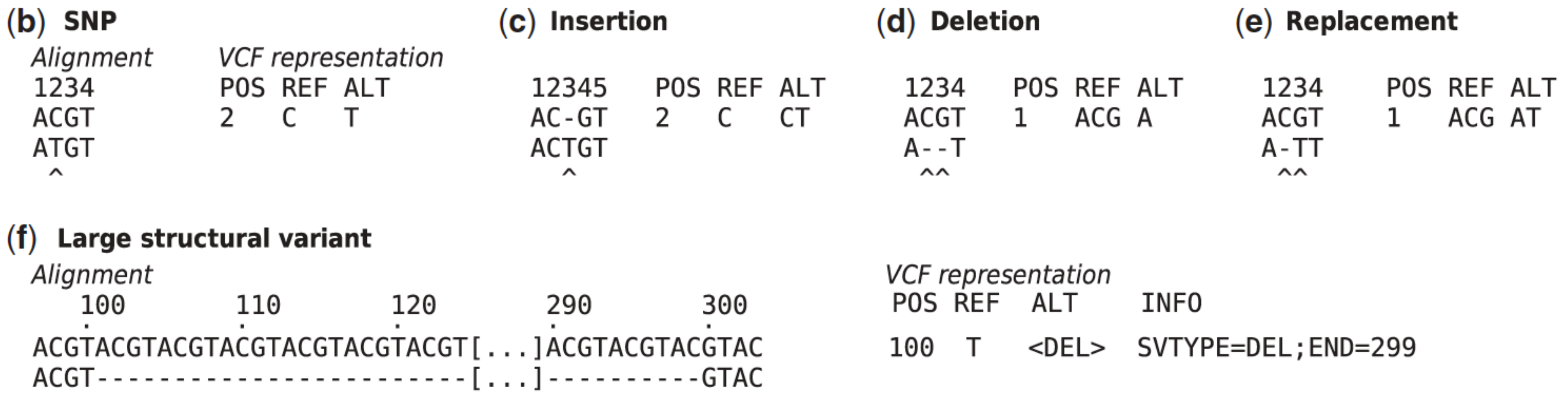


```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##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        1 .      ACG  A,AT   .    PASS   .                   GT:DP     1/2:13    0/0:29
1        2 rs1    C    T,CT   .    PASS   H2;AA=T             GT:GQ     0|1:100   2/2:70
1        5 .      A    G      .    PASS   .                   GT:GQ     1|0:77    1/1:95
1      100 .      T    <DEL>  .    PASS   SVTYPE=DEL;END=300  GT:GQ:DP  1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

- Filter column by default will have “.” (i.e., filters not applied)
- Apply filters by adding “tags” to filter column for variants that do not pass the filter
- Filters indicate whether values in QUAL or INFO fields meet a specific condition
- Can “hard-filter” (ie., remove from the VCF) or rely on downstream tools that that are “filter-aware” (e.g., VCFtools)

VCF examples



(b) SNP

```
Alignment      VCF representation
1234           POS REF ALT
ACGT           2   C   T
ATGT
 ^
```

(c) Insertion

```
12345   POS REF ALT
AC-GT   2   C   CT
ACTGT
  ^
```

(d) Deletion

```
1234   POS REF ALT
ACGT   1   ACG A
A--T
 ^^
```

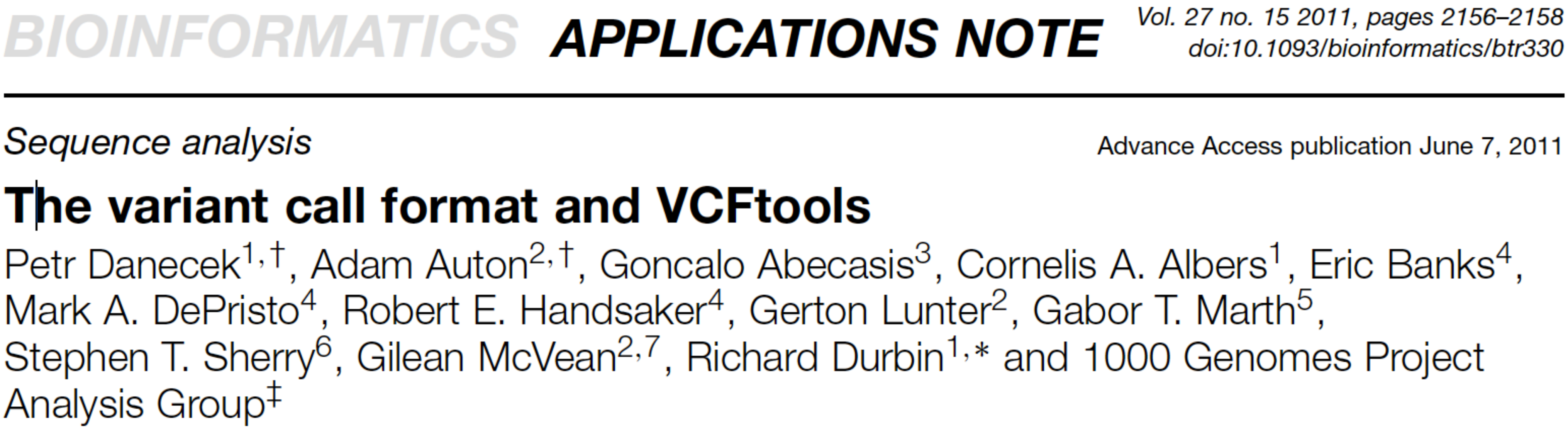
(e) Replacement

```
1234   POS REF ALT
ACGT   1   ACG AT
A-TT
 ^^
```

(f) Large structural variant

```
Alignment
   100       110       120       290       300
   .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

```
VCF representation
POS REF ALT   INFO
100 T   <DEL> SVTYPE=DEL;END=299
```



BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

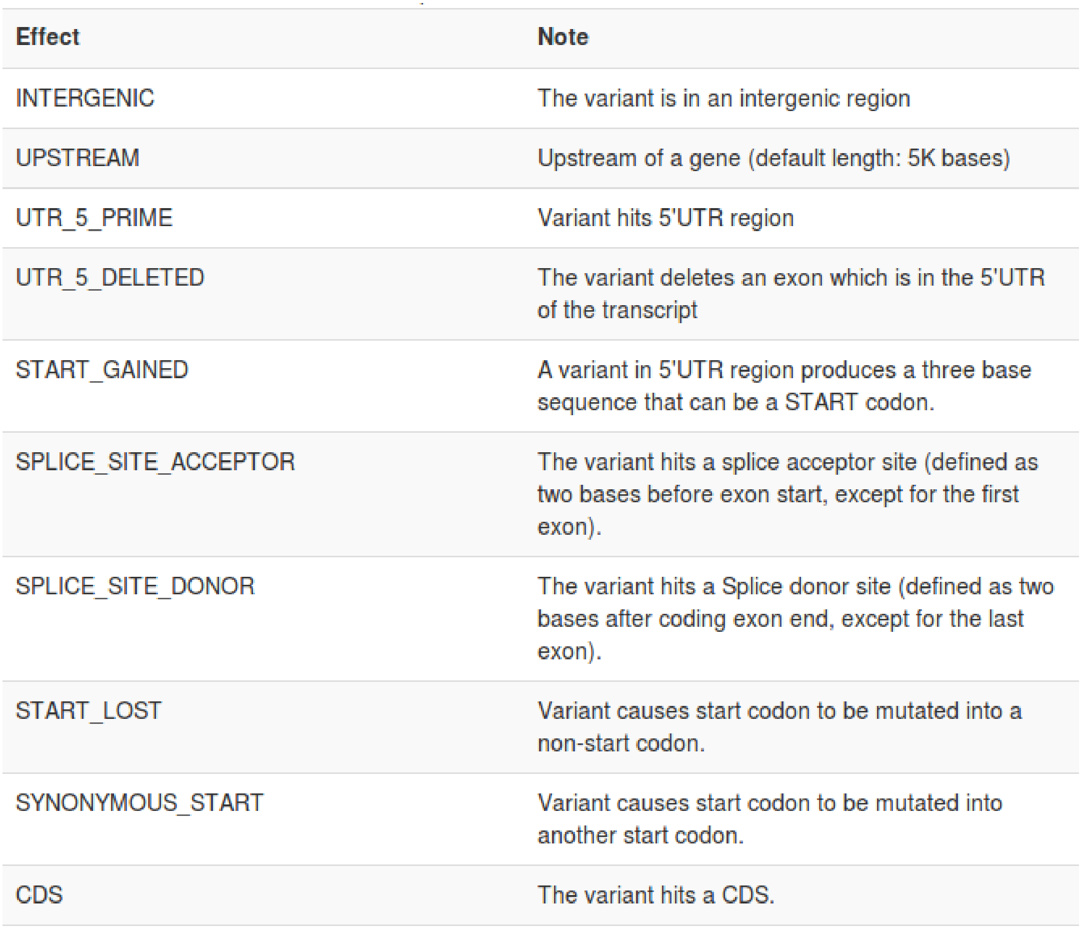
Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

Annotating SnpEff

- We use SnpEff
- Annotates and predicts the effects of variants on genes
  - Codon changes
  - Amino acid changes
  - Genomic region
  - Functional effect (silent, missense)
- SnpEff has pre-built databases for thousands of genomes

Functional Annotation of SNPs using snpEff

- What is snpEff?
- Why snpEff?
- What type of annotation does snpEff produce?
- SnpEff predictions can be integrated into VCF for ease of downstream analysis



Effect	Note
INTERGENIC	The variant is in an intergenic region
UPSTREAM	Upstream of a gene (default length: 5K bases)
UTR_5_PRIME	Variant hits 5'UTR region
UTR_5_DELETED	The variant deletes an exon which is in the 5'UTR of the transcript
START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon.
SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).
SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).
START_LOST	Variant causes start codon to be mutated into a non-start codon.
SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon.
CDS	The variant hits a CDS.

The “ANN” tag in VCF: A controlled vocabulary for annotating variants

- A new (2015) specification for the ANN tag in the vcf INFO column

- Adopts mutation vocabulary from Human Genome Variation Society (HGVS)

http://www.hgvs.org/mutnomen/

- VCFannotationformat_v1.0.pdf

- Example:

```
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO'">
```

```
AC=34;AF=0.279;AN=122;ANN=C|missense_variant|MODERATE|LOC_Os01g05960|LOC_Os01g05960|transcript|LOC_Os01g05960.1|Coding|1/2|c.40G>C|p.Val14Leu|85/3430|40/3156|14/1051||
```