Variant calling

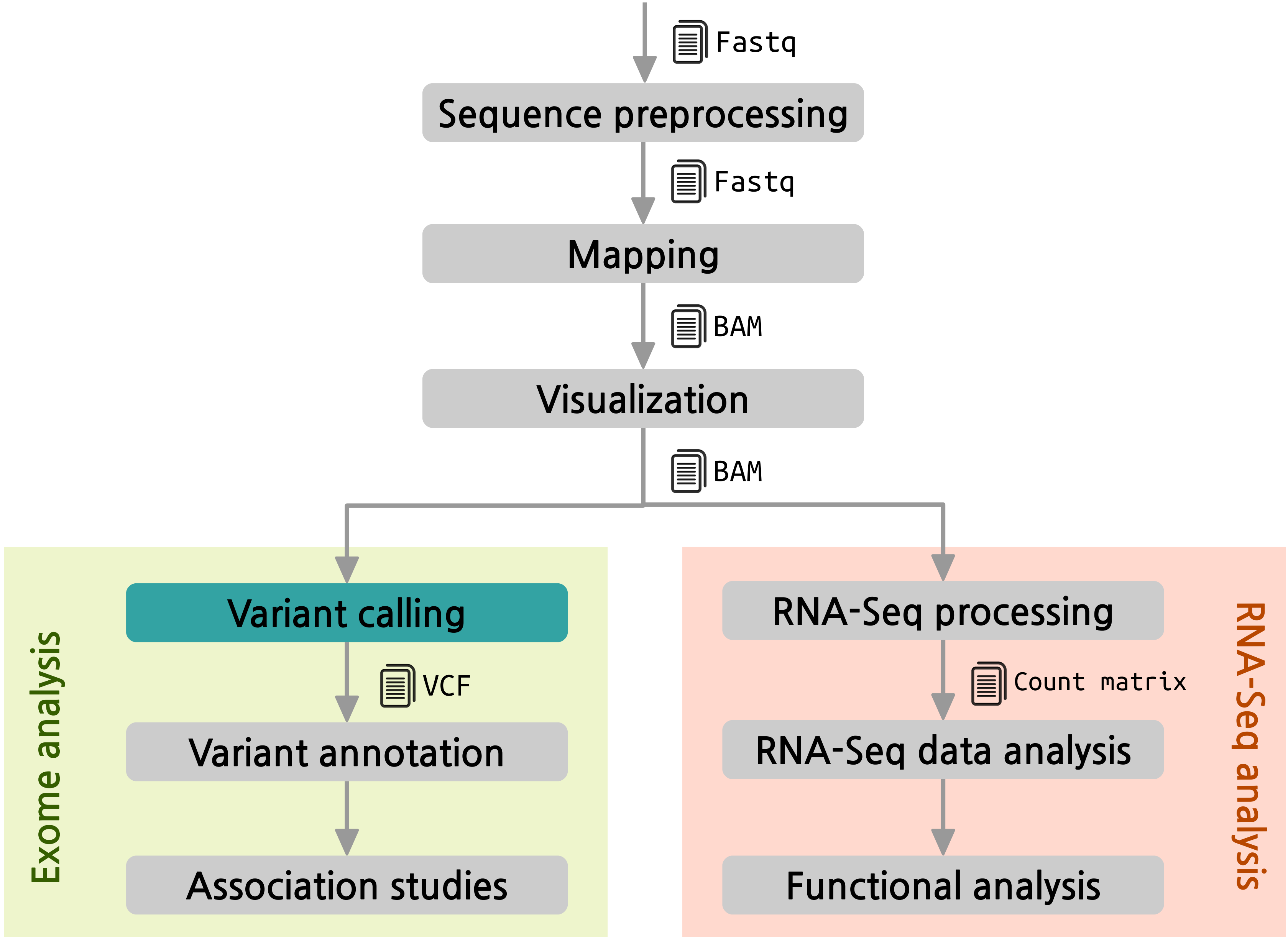
Detecting variants in NGS data

The Genome Analysis ToolKit (GATK)

University of Cambridge
Cambridge, UK
10th June 2014

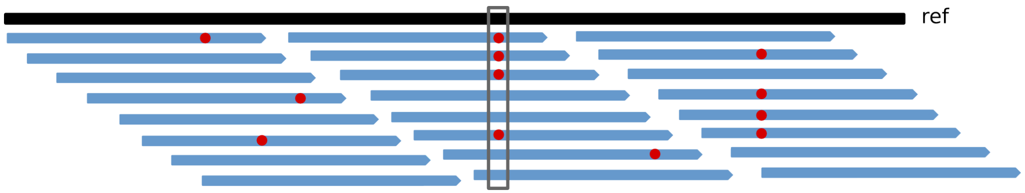
Marta Bleda Latorre
mbleda@cipf.es
PhD Student at the Computational Genomics Institute
Centro de Investigación Príncipe Felipe (CIPF)
Valencia, Spain

The pipeline



Objective

Assign a genotype to each position



Problems

Some variation observed in BAM files is caused by mapping and sequencing artifacts:

- **PCR artifacts**:
    - Mismatches due to errors in early PCR rounds
    - PCR duplicates
- **Sequencing errors**: erroneous call, either for physical reasons or to properties of the sequenced DNA
- **Mapping errors**: often happens around repeats or other low-complexity regions

Separate **true variation** from machine artifacts

Variant calling process pipeline

1. **Mark duplicates**

   Duplicates should not be counted as additional evidence

2. **Local realignment around INDELS**

   Reads mapping on the edges of INDELS often get mapped with mismatching bases introducing false positives

3. **Base quality score recalibration (BQSR)**

   Quality scores provided by sequencing machines are generally inaccurate and biased

4. **Variant calling**

   Discover variants and their genotypes

1. Mark duplicates

- All NGS **sequencing platforms are NOT single molecule sequencing** → the same DNA molecule can be sequenced several times
- **PCR** → duplicate DNA fragments in the final library
- If there is a base variation it will have **high depth support**
- Can result in **false variant calls**

Tools

- **Samtools**: samtools rmdup or samtools rmdupse
- **Picard**: MarkDuplicates

1. Mark duplicates

- All NGS **sequencing platforms are NOT single molecule sequencing** → the same DNA molecule can be sequenced several times
- **PCR** → duplicate DNA fragments in the final library
- If there is a base variation it will have **high depth support**
- Can result in **false variant calls**

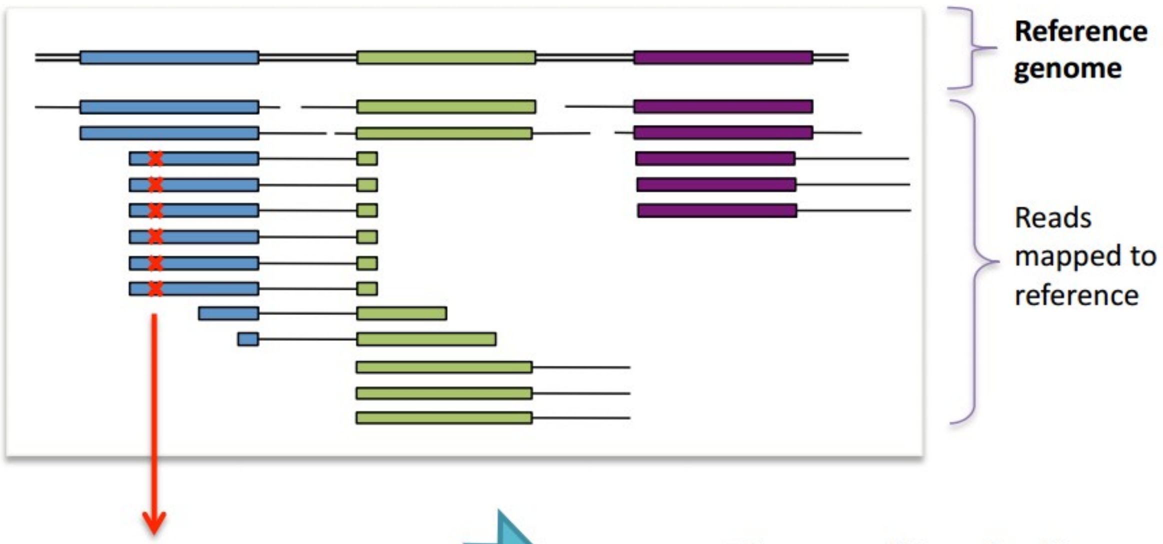
Tools

- **Samtools**: samtools rmdup or samtools rmdupse
- **Picard**: MarkDuplicates

1. Mark duplicates

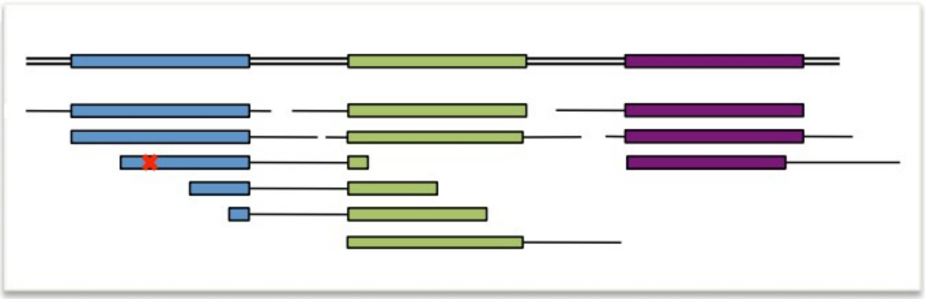
The reason why duplicates are bad

✖ = sequencing error propagated in duplicates



FP variant call (bad)

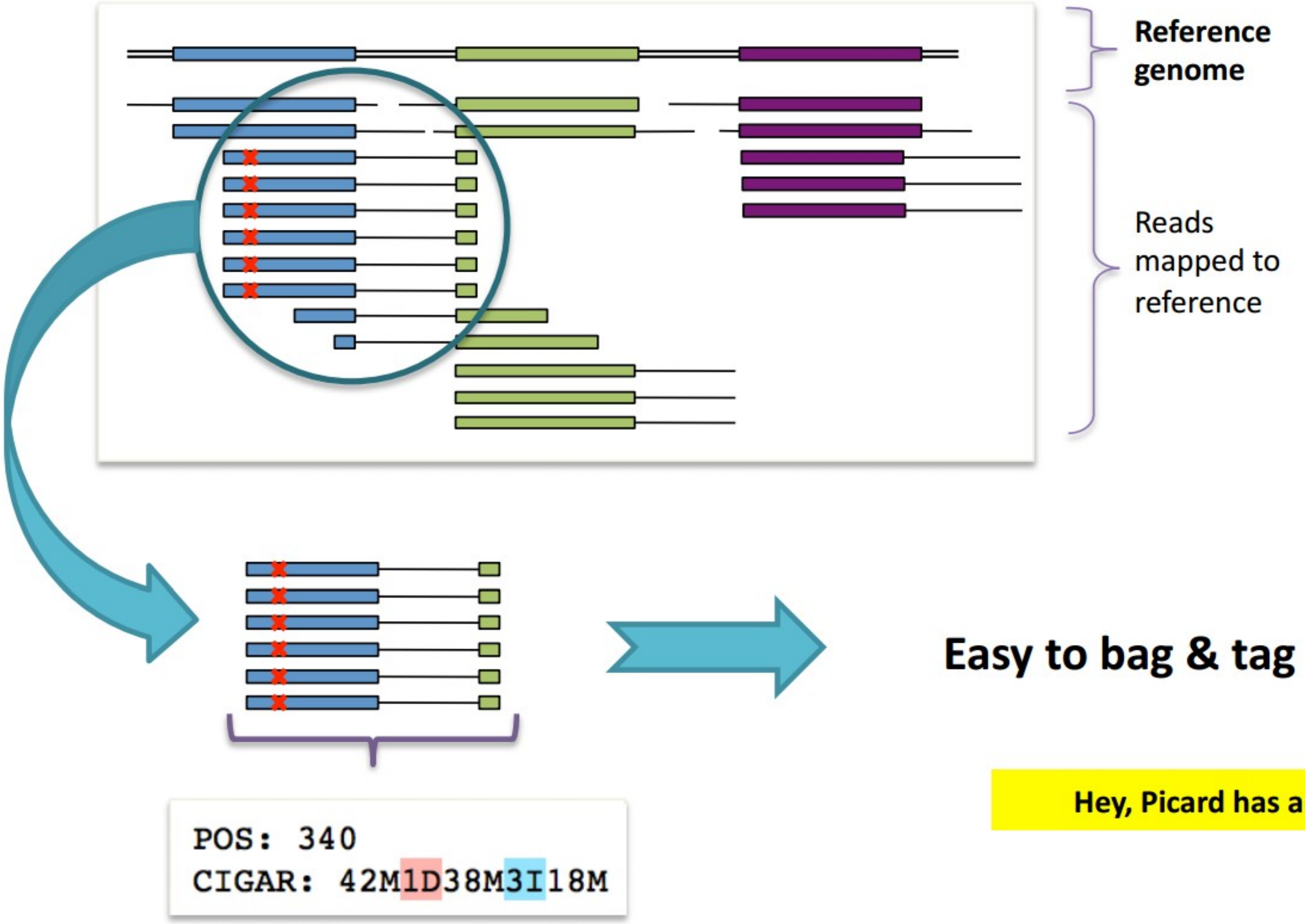
After marking duplicates, the GATK will only see :



... and thus be more likely to make the right call

1. Mark duplicates
Duplicate identification

Duplicates have the **same starting position** and the **same CIGAR** string



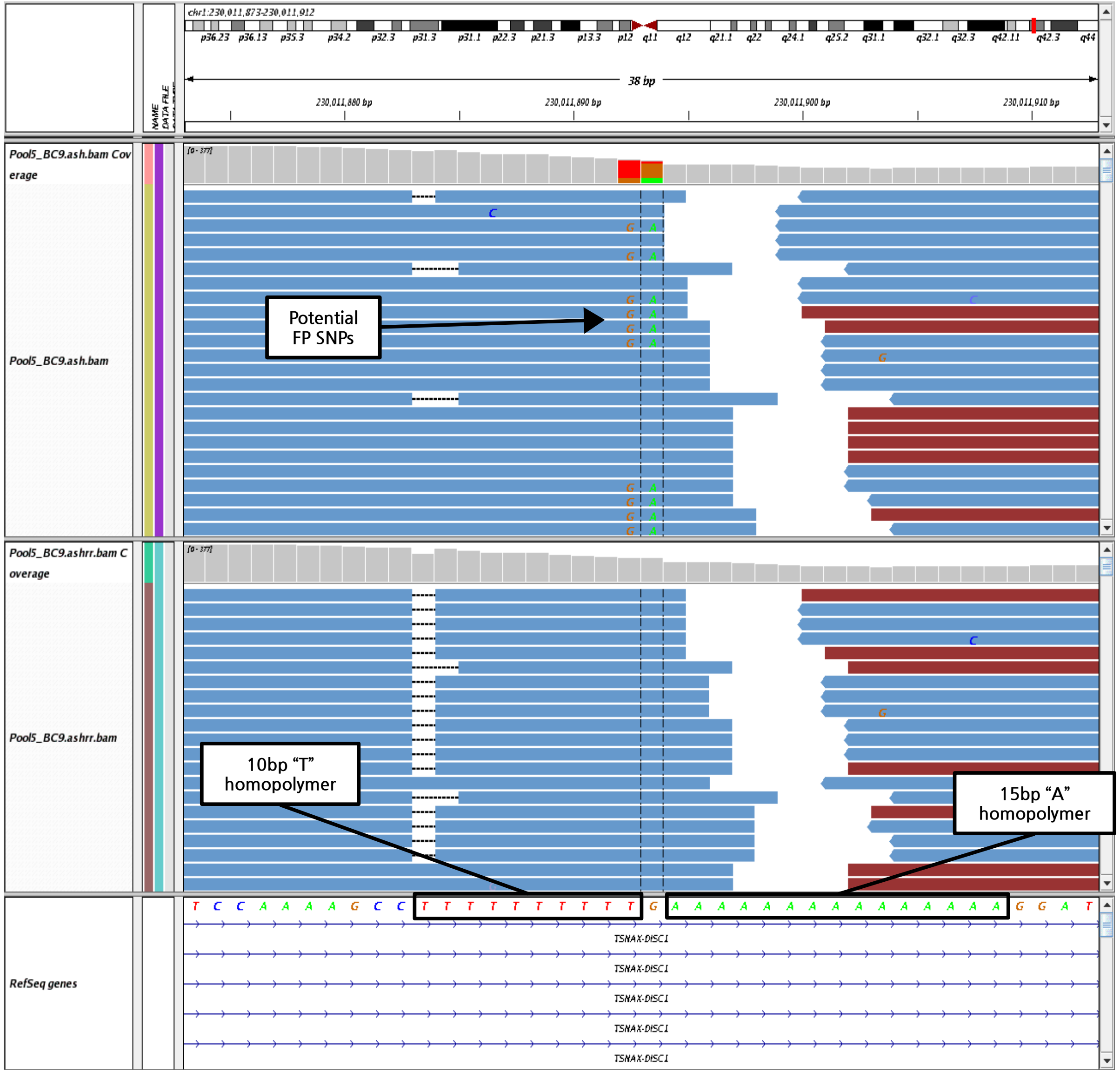
```
POS: 340
CIGAR: 42M1D38M3I18M
```

Easy to bag & tag

Hey, Picard has an app for that!

2. Local realignment around INDELS

- Reads **near INDELS** are mapped with mismatches
- **Realignment** can identify the most consistent placement for these reads
  1. **Identify** problematic regions
  2. **Determine the optimal** consensus sequence
- **Minimizes mismatches** with the reference sequence
- **Refines** location of **INDELS**



Potential FP SNPs

10bp "T" homopolymer

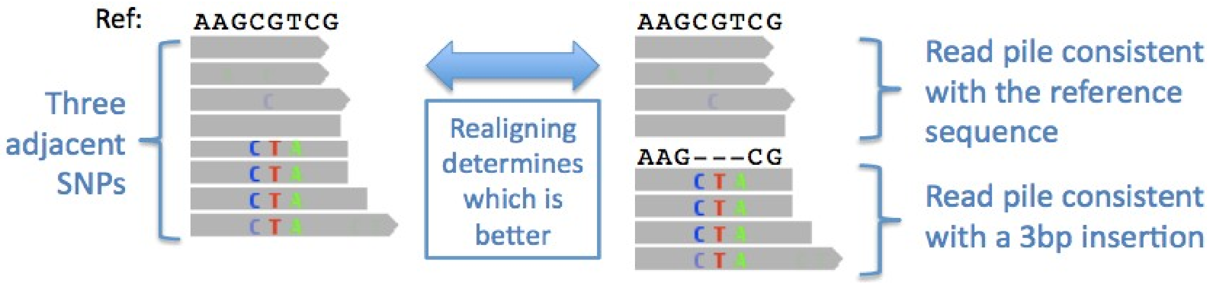
15bp "A" homopolymer

chr1:230,011,873-230,011,912
38 bp
230,011,880 bp
230,011,890 bp
230,011,900 bp
230,011,910 bp
Pool5_BC9.ash.bam Coverage
Pool5_BC9.ash.bam
Pool5_BC9.ashrr.bam Coverage
Pool5_BC9.ashrr.bam
RefSeq genes
TSNAX-DISC1
T C C A A A A G C C T T T T T T T T T T G A A A A A A A A A A A A A A A G G A T

DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. PMID: 21478889

2. Local realignment around INDELS

- Reads **near INDELS** are mapped with mismatches
- **Realignment** can identify the most consistent placement for these reads
  1. **Identify** problematic regions
  2. **Determine the optimal** consensus sequence
- **Minimizes mismatches** with the reference sequence
- **Refines** location of **INDELS**


Ref:
AAGCGTCG
Three adjacent SNPs
CTA
Realigning determines which is better
AAGCGTCG
AAG---CG
Read pile consistent with the reference sequence
Read pile consistent with a 3bp insertion

DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. PMID: 21478889

3. Base quality score recalibration

- **Calling algorithms rely** heavily on the **quality scores** assigned to the individual base calls in each sequence read
- Unfortunately, the scores produced by the machines are subject to various sources of **systematic error**, leading to over- or under-estimated base quality scores in the data

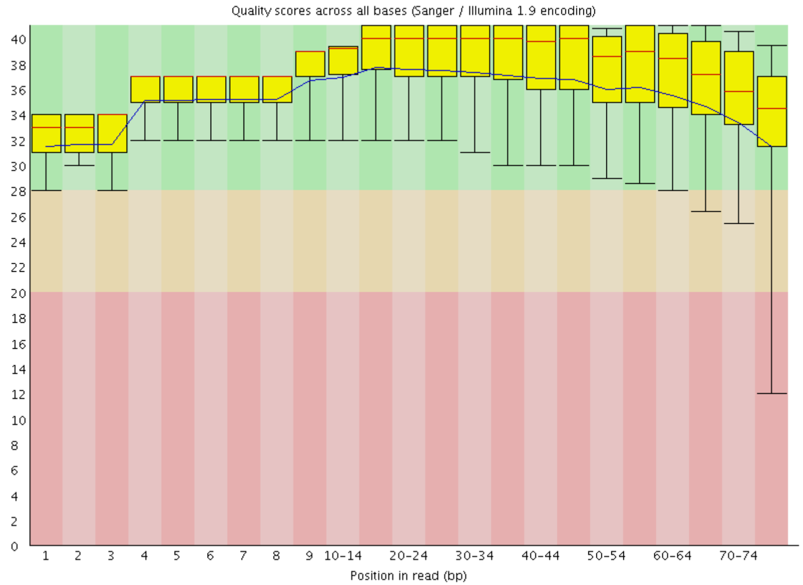
How?

1. **Analyze covariation** among several features of a base:
   - Reported quality score
   - Position within the read
   - Preceding and current nucleotide
2. Use a set of **known variants** (i.e.: dbSNP) to model error properties of real polymorphism and determine the **probability that novel sites are real**
3. **Adjust** the quality scores of all reads in a BAM file

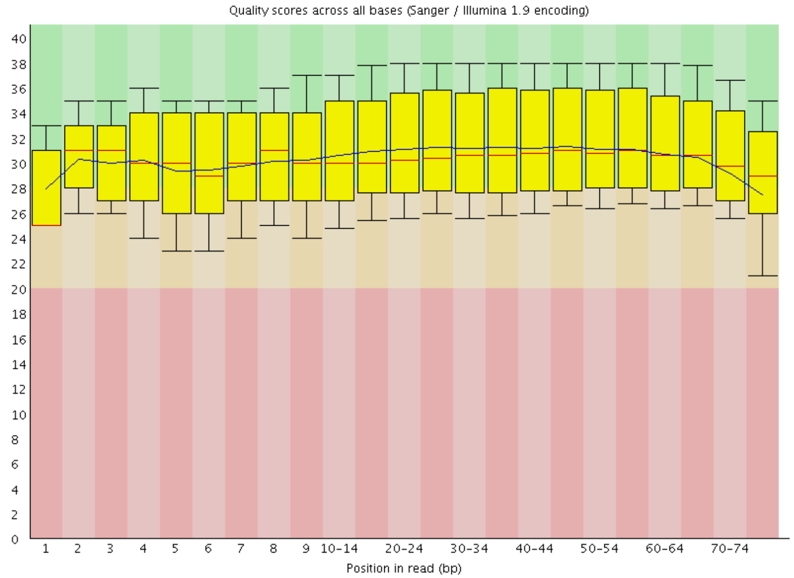
-

3. Base quality score recalibration

Before



After



Phred Quality score:

$$Q_{\text{Phred}} = -10 \log_{10} P(\text{error}).$$

A score of 20 corresponds to 1% error rate in base calling

4. Variant calling

Variant discovery process

Steps

1. **Variant calling:** Identify the positions that differ from the reference
2. **Genotype calling**: calculate the genotypes for each sample at these sites

Initial approach

Independent base assumption

Counting the number of times each allele is observed

Evolved approach

Bayesian inference → Compute genotype likelihood

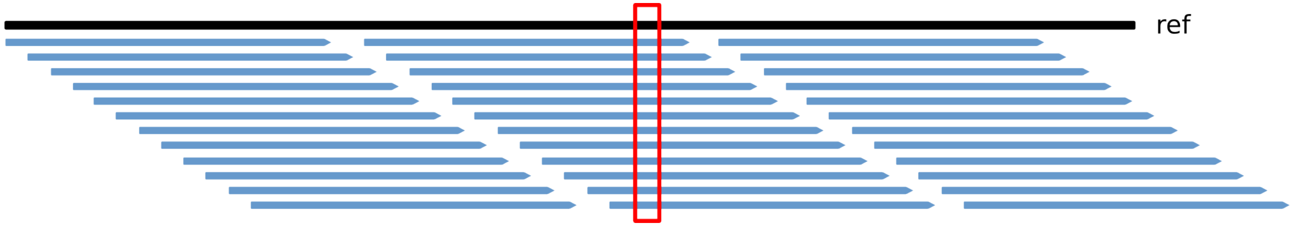
Advantages:

Provide statistical measure of **uncertainty**

Lead to **higher accuracy** of genotype calling

4. Variant calling

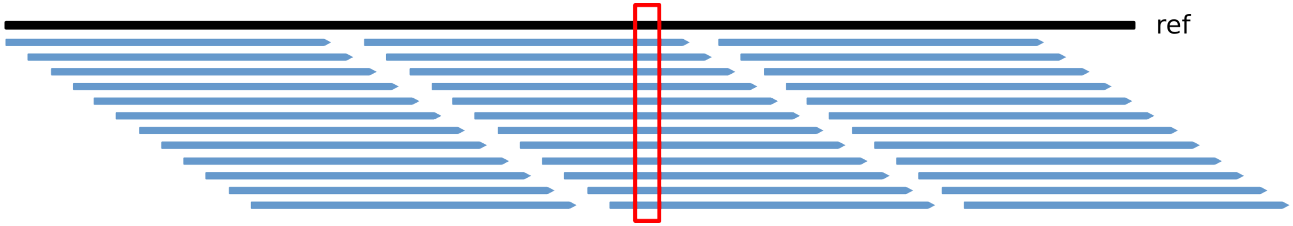
Variant discovery process



Reference = **A**

4. Variant calling

Variant discovery process



Reference = A

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

AAAAAAAAAAAAAAAGGGGGGGGGGGGGGG

AAAAAAAAAAAAAAAAGGGGGGGGGGGGCT

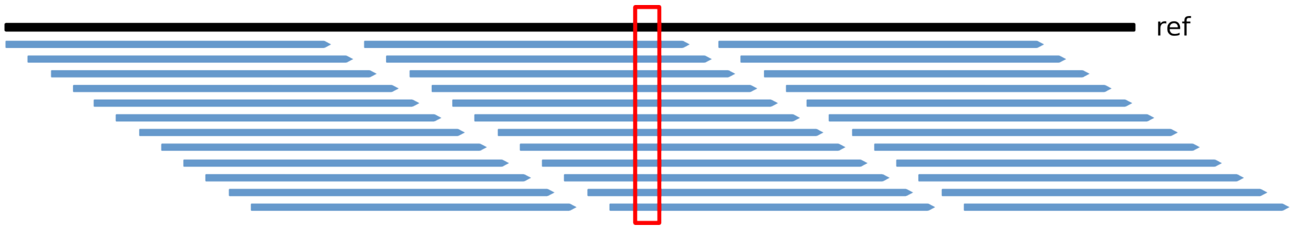
AAAGGGCCTT

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

4. Variant calling

Variant discovery process



Reference = A

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA N=30, X=0

GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG N=30, X=30

AAAAAAAAAAAAAAAGGGGGGGGGGGGGGG N=30, X=15

AAAAAAAAAAAAAAAAGGGGGGGGGGGGCT N=30, X=12

AAAGGGCCTT N=10, X=3

Cutoff for X → value or proportion

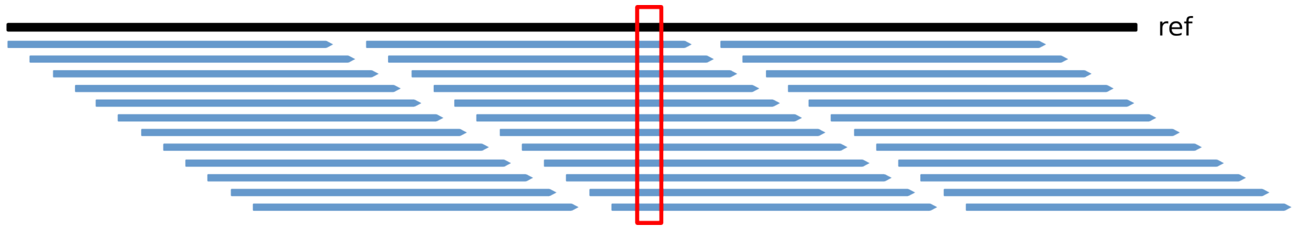
- $c = 30\%$ $X \leq c$ → **RR**, $X > c$ → **RV**
- $c_1 = 10\%$, $c_2 = 30\%$ $X \leq c_1$ → **RR**
  $c_1 < X < c_2$ → **RV**
  $X \geq c_2$ → **RR**

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

4. Variant calling

Variant discovery process



Reference = A

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	N=30,	X=0	→ **RR**
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	N=30,	X=30	→ **VV**
AAAAAAAAAAAAAAAGGGGGGGGGGGGGGG	N=30,	X=15	→ **RV**
AAAAAAAAAAAAAAAAAAGGGGGGGGGGCT	N=30,	X=12	→ **RV**
AAAGGGCCTT	N=10,	X=3	→ **RV?**

Cutoff for X → value or proportion

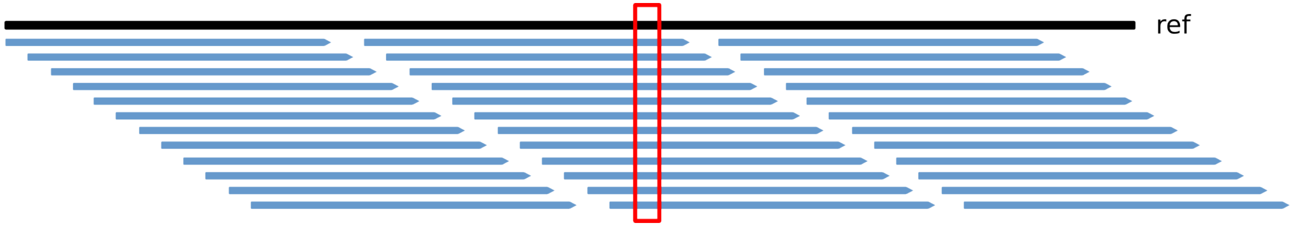
- $c = 30\%$ $X \leq c$ → **RR**, $X > c$ → **RV**
- $c_1 = 10\%$, $c_2 = 30\%$
  - $X \leq c_1$ → **RR**
  - $c_1 < X < c_2$ → **RV**
  - $X \geq c_2$ → **RR**

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

4. Variant calling

Variant discovery process



Bayesian approximation

α = nucleotide-base error rate

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

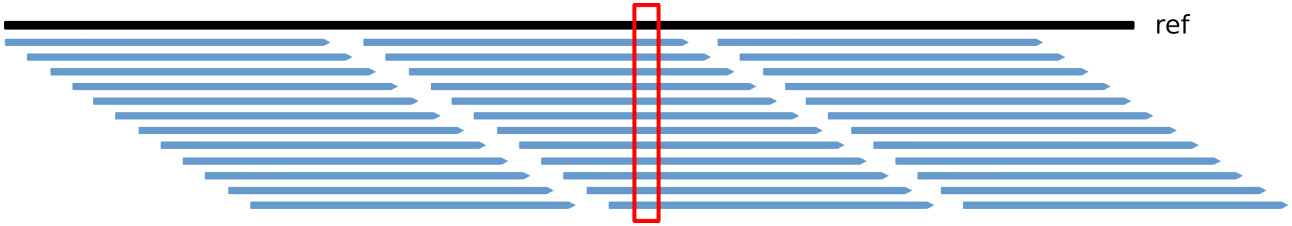
$P(G=RR, X|N, \alpha)$ = P of all R calls being correct and all V calls being wrong

$P(G=VV, X|N, \alpha)$ = P of all V calls being correct and all R calls being wrong

$P(G=RV, X|N, \alpha)$ = P of all R and V calls being correct

4. Variant calling

Variant discovery process



Bayesian approximation

α = nucleotide-base error rate

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

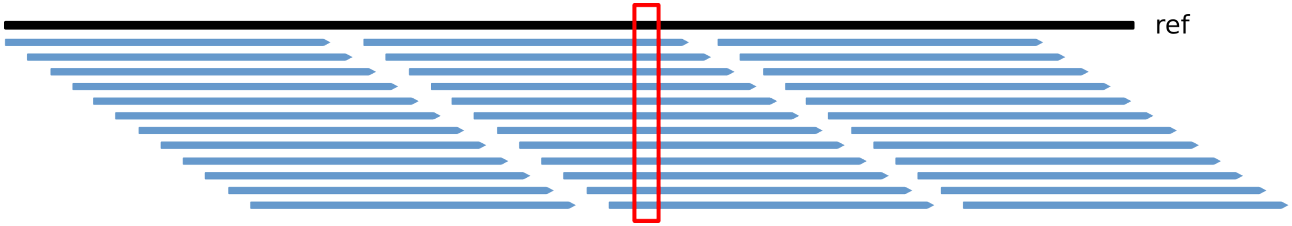
$$P(G=RR, X|N, \alpha) = \binom{N}{X}\alpha^{X}(1-\alpha)^{N-X}$$

$$P(G=VV, X|N, \alpha) = \binom{N}{X}(1-\alpha)^{X}\alpha^{N-X}$$

$$P(G=RV, X|N, \alpha) = \binom{N}{X}\left(\frac{1}{2}\right)^{N}$$

4. Variant calling

Variant discovery process



Bayesian approximation

α = nucleotide-base error rate

p_{VV}
p_{VR} } Prior probabilities

N = nucleotides
G = true genotype
R = reference base
V = variant base
X = variant nucleotides

Outcomes:
RR RV VV

$$P(G=RR, X|N,\alpha) = \binom{N}{X}\alpha^{X}(1-\alpha)^{N-X}(1-p_{VV}-p_{RV})$$

$$P(G=VV, X|N,\alpha) = \binom{N}{X}(1-\alpha)^{X}\alpha^{N-X}p_{VV}$$

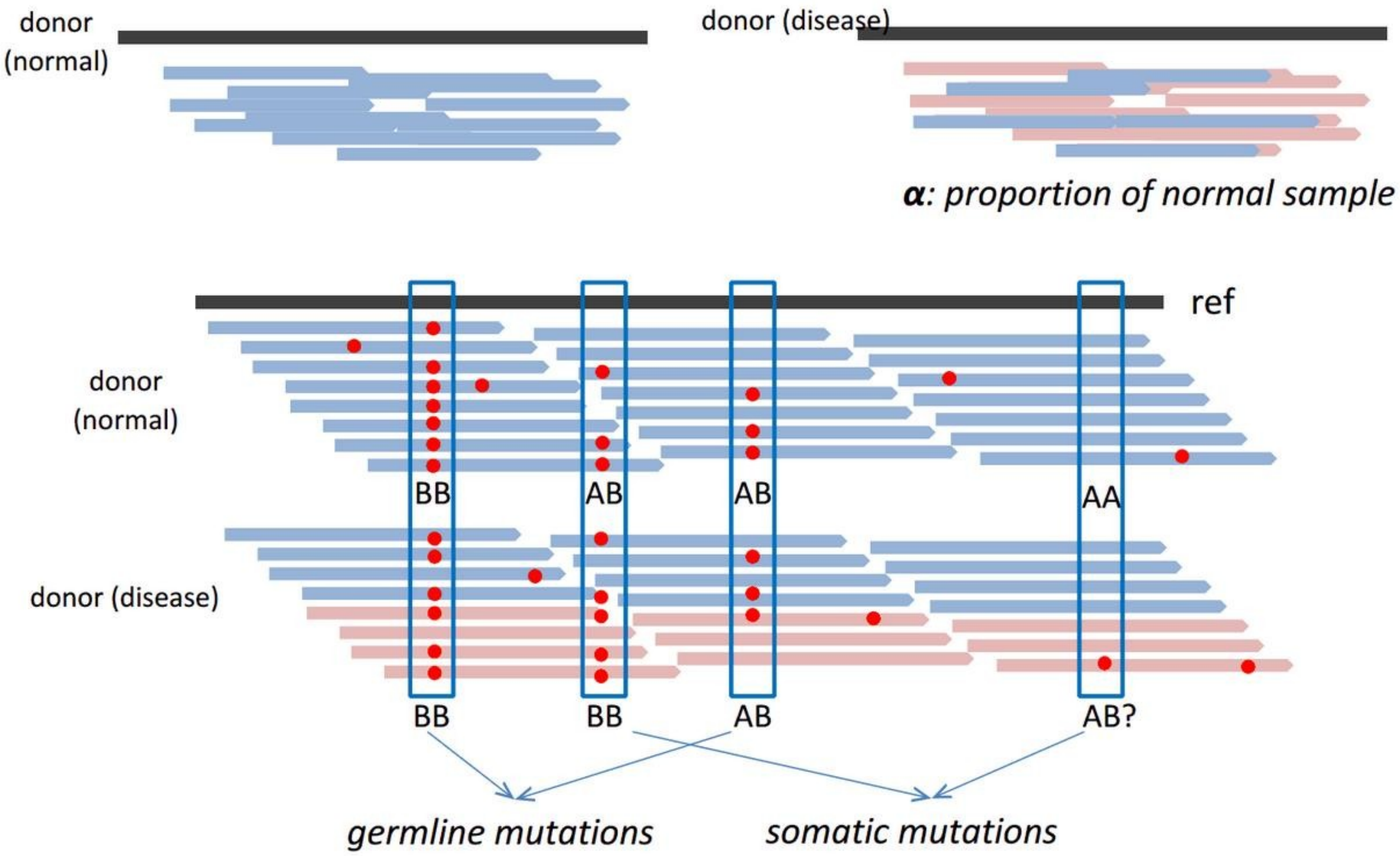
$$P(G=RV, X|N,\alpha) = \binom{N}{X}\left(\frac{1}{2}\right)^{N}p_{RV}$$

Somatic calling

Detecting somatic SNVs in cancer

Challenges:

- Somatic variants occur at low frequency in genome
- Most tumors are impure and heterogeneous

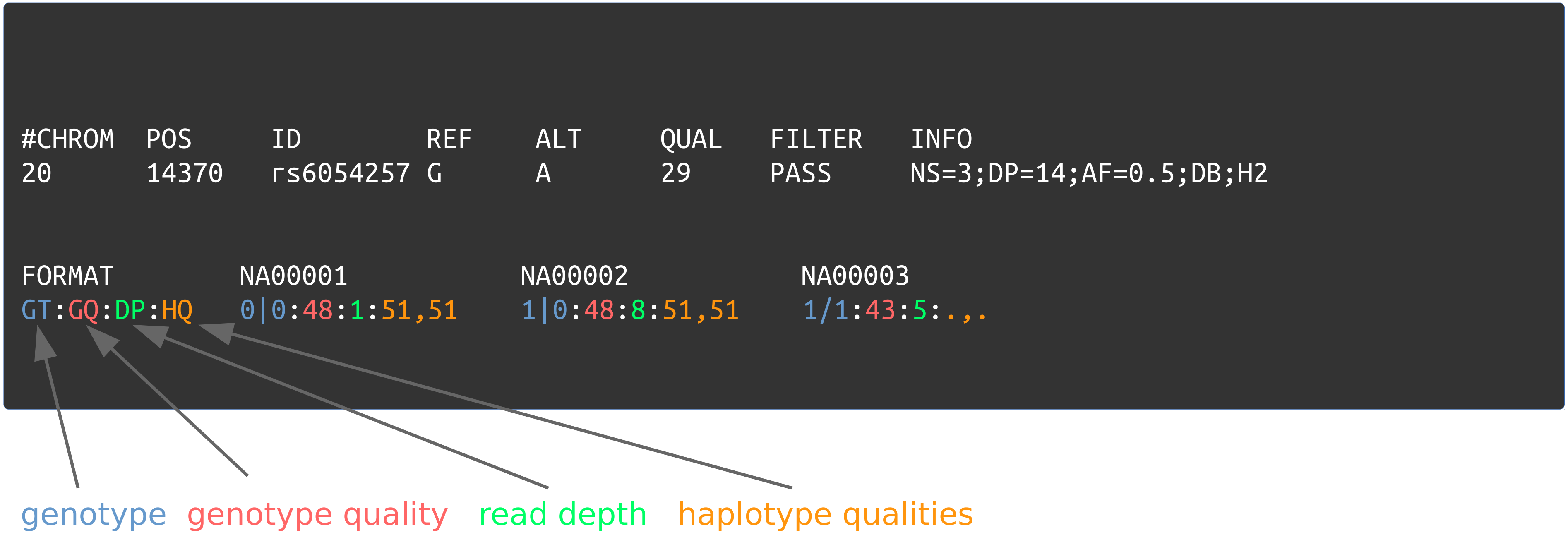


VCF file format

- Specification defined by the 1000 genomes (current version **4.2**): http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
- Commonly **compressed and indexed** with bgzip/tabix
- Single-sample or multi-sample VCF

```
##fileformat=VCFv4.1
##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   rs6054257 G      A       29   PASS   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    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   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     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF file format



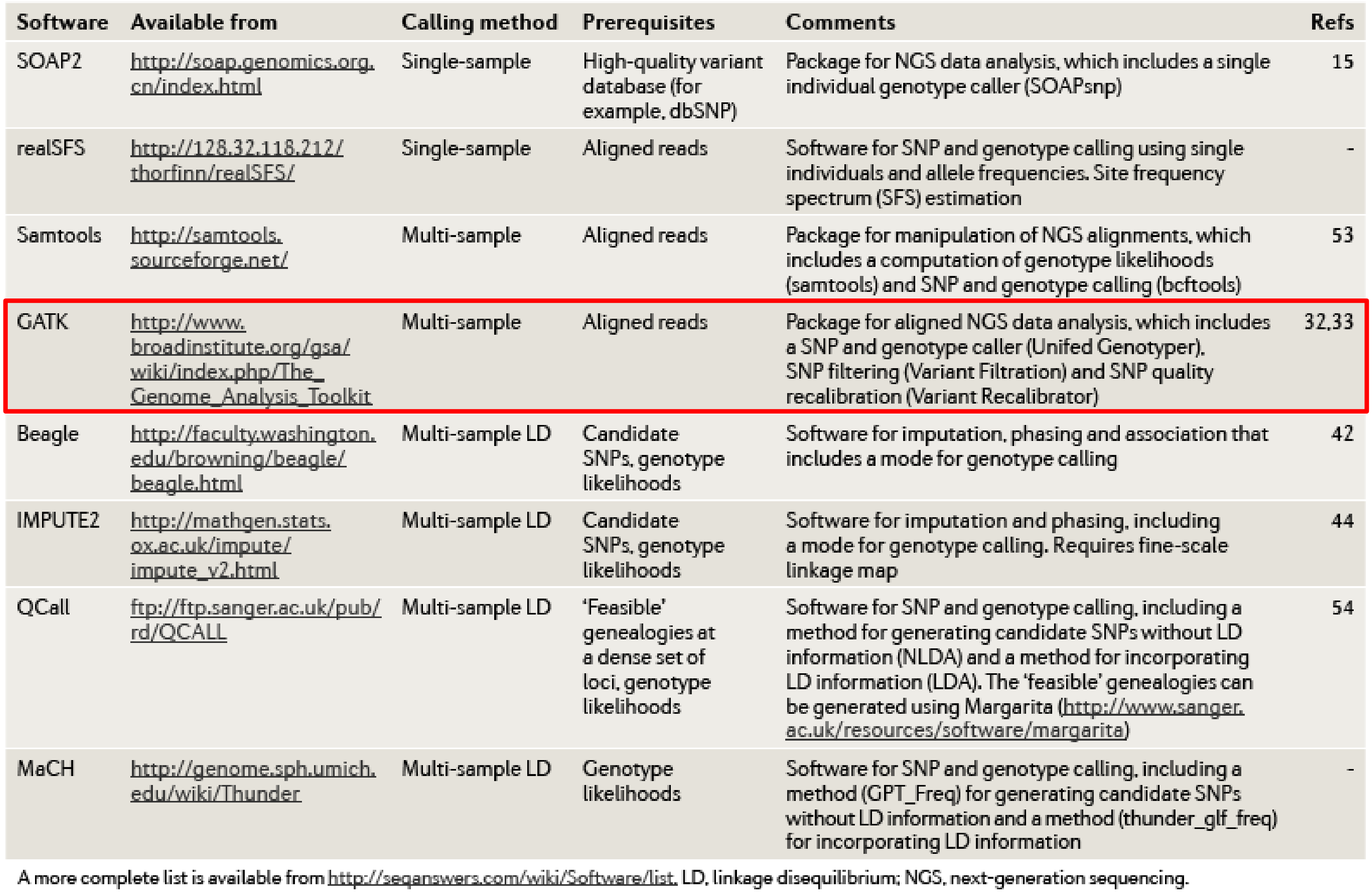
```
#CHROM  POS     ID         REF    ALT     QUAL   FILTER   INFO
20      14370   rs6054257  G      A       29     PASS     NS=3;DP=14;AF=0.5;DB;H2

FORMAT         NA00001           NA00002           NA00003
GT:GQ:DP:HQ    0|0:48:1:51,51    1|0:48:8:51,51    1/1:43:5:.,.
```

genotype genotype quality read depth haplotype qualities

- **CHROM**: chromosome
- **POS**: position
- **ID**: identifier
- **REF**: reference base(s)
- **ALT**: non-reference allele(s)
- **QUAL**: quality score of the calls (phed scale)
- **FILTER**: “PASS” or a filtering tag
- **INFO**: additional information
- **FORMAT**: describes the information given by sample

Software



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.

GATK (Genome Analysis ToolKit)

http://www.broadinstitute.org/gatk/

- Probabilistic method: **Bayesian estimation** of the most likely genotype
- Calculates many **parameters** for each position of the genome
- INDEL realignment
- Base quality recalibration
- SNP and INDEL calling
- **Multi-sample** calling
- Uses standard input and output files
- Used in **many NGS projects**, including the 1000 Genomes Project, The Cancer Genome Atlas, etc.

GATK prerequisites

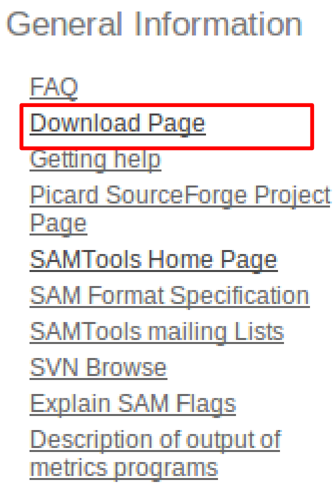
- **Requires Java** (http://www.oracle.com/technetwork/java/javase/downloads/index.html)
  - Check your java version

```
java -version
```

GATK ≥ 2.6 → Requires Java version 1.7

- **Picard**
  - Website: **http://picard.sourceforge.net/**
  - Go to Download page and select

Download picard-tools-1.114.zip (48.0 MB)



General Information

FAQ
Download Page
Getting help
Picard SourceForge Project Page
SAMTools Home Page
SAM Format Specification
SAMTools mailing Lists
SVN Browse
Explain SAM Flags
Description of output of metrics programs

- Testing:

```
java -jar AddOrReplaceReadGroups.jar -h
```

- Usage

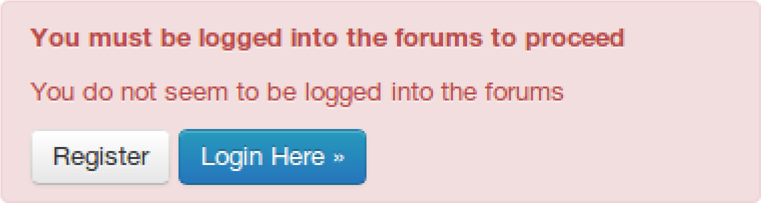
```
java -jar <ToolName> [options]
```

GATK installation

- **GATK 3.1 download**

  http://www.broadinstitute.org/gatk/

  - We need to register before download
  - Go to Downloads and click GATK 2
  - Accept the license agreement
  - Extract the file in the application folder



You must be logged into the forums to proceed

You do not seem to be logged into the forums

- **Check if GATK is working**

Show GATK help

```
java -jar GenomeAnalysisTK.jar -h
```

- **Usage**

```
java -jar GenomeAnalysisTK.jar -T <ToolName> [arguments]
```

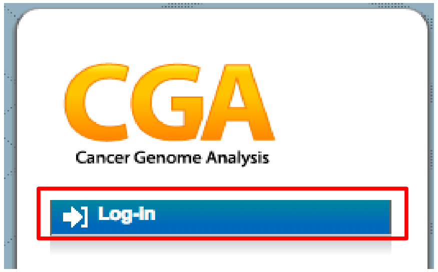
MuTect installation

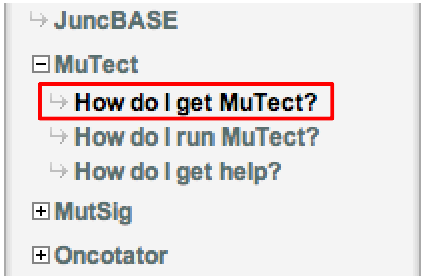
- **MuTect download**

  http://www.broadinstitute.org/cancer/cga/mutect

  - Click ***Log-in*** and go to the ***Create new account*** tab
  - Fill the form
  - Go to ***How do I get mutect*** and accept the license agreement
  - Download the latest version

    muTect-1.1.4-bin.zip
  - Extract the file in the applications folder





- **Check if MuTect is working**

```
java -jar muTect-1.1.4.jar -h
```

- **Usage**

```
java -jar muTect-1.1.4.jar --analysis_type MuTect [arguments]
```

THANK YOU.