Session– Variant Calling and Consensus Generation

- **BU-ISCIII**
- **Unidades Comunes Científico Técnicas – SGSAFI-ISCIII**

14-18 Noviembre 2023, 3ª Edición
Programa Formación Continua, ISCIII

Index

Mapping against reference genome and Variant Calling :

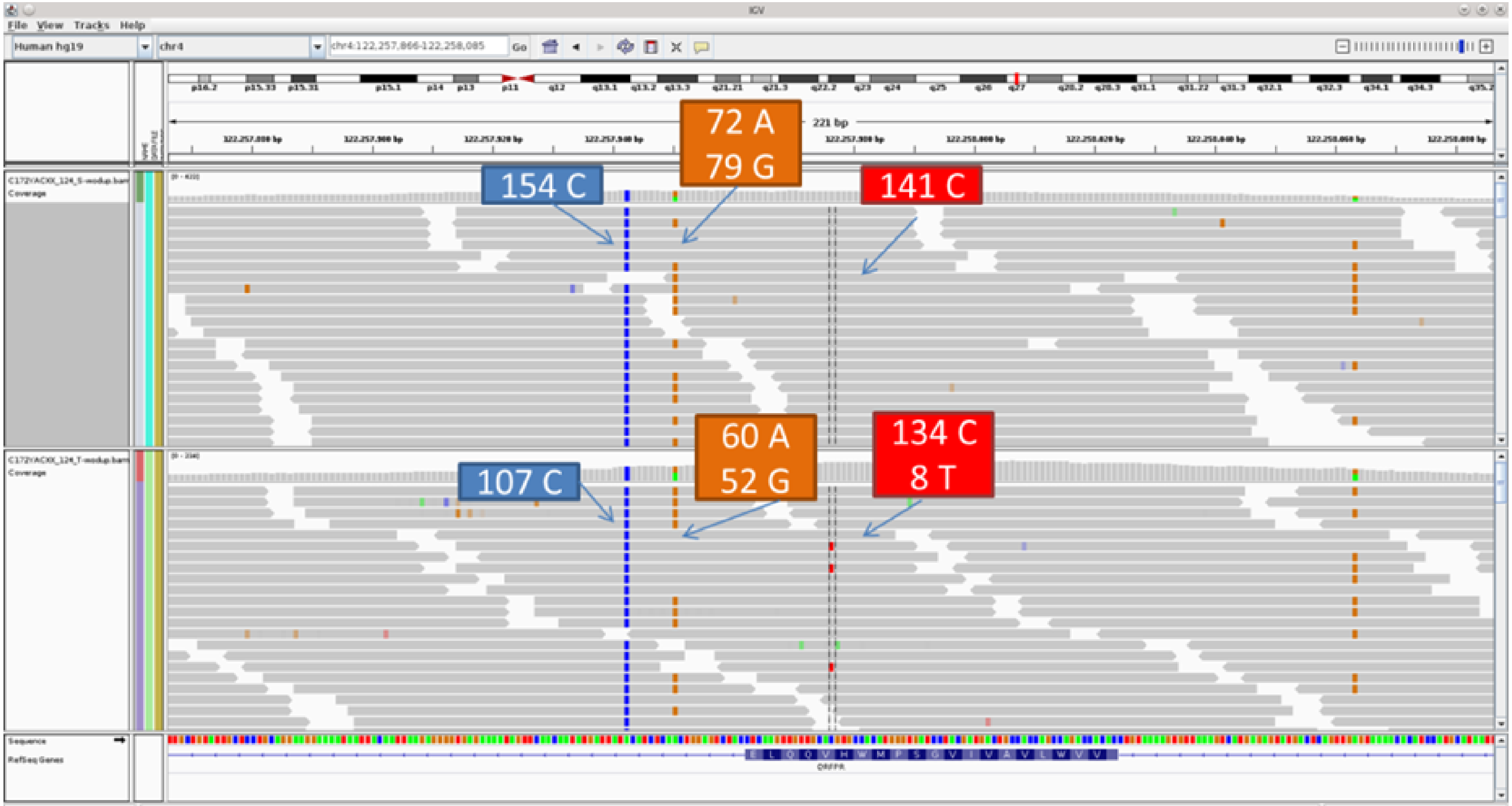
- Variant Calling
- Source of error and mitigation strategies
- VCF and bed format
- IVAR, LOFREQ, VARSCAN2
- Consensus generation: aproximations

Variant Calling

- **Variant calling concept is simple:**

Find positions in our reads different from the reference.

- We start with our secuences mapped against our reference genome, and we walk trough every column of the alignment counting the number of alleles found and comparing them against the reference.



154 C

72 A
79 G

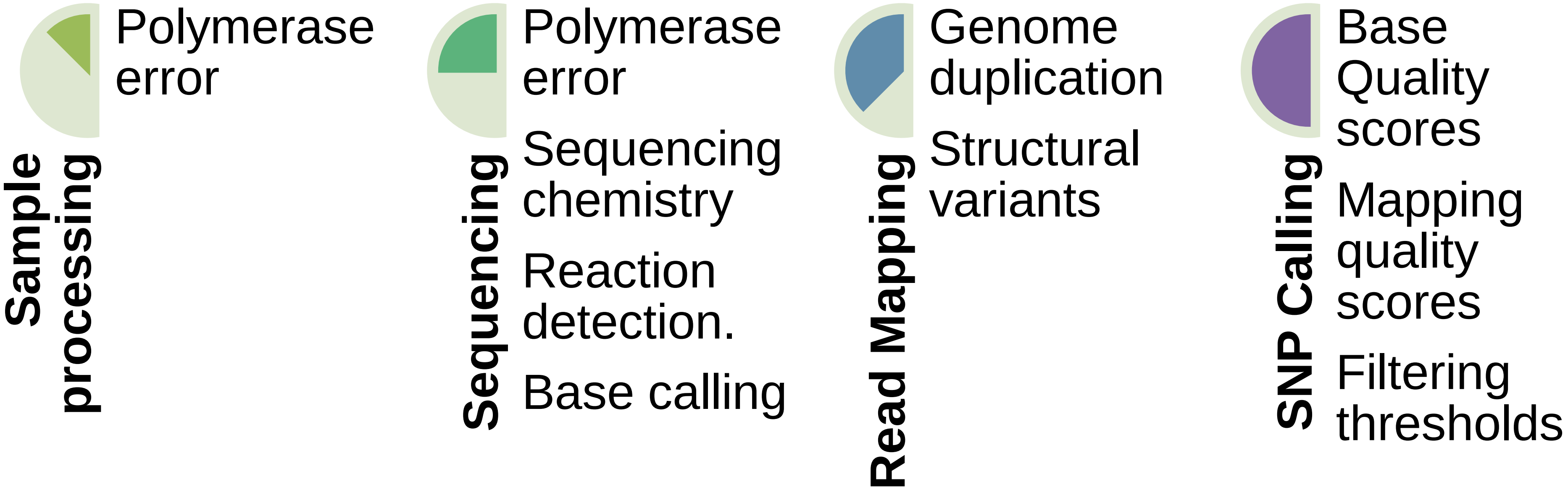
141 C

107 C

60 A
52 G

134 C
8 T

Sources of error and mitigation strategies



Sample processing

- Polymerase error

Sequencing

- Polymerase error
- Sequencing chemistry
- Reaction detection.
- Base calling

Read Mapping

- Genome duplication
- Structural variants

SNP Calling

- Base Quality scores
- Mapping quality scores
- Filtering thresholds

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

- **Sample processing errors.**
  - Random errors.
  - Associated with polymerase errors . (1 in 10^{2-3} bases)
  - Homopolymers and tandem repeats experience higher indel error rates.
- **Solutions:**
  - Paired-end libraries.
  - Minimization of PCR cycles.

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

- **Sequencing:**
  - Dependent on the platform.
  - Can be random and systematic.
  - 6% Illumina, 50% Roche (Ross et al.2013)
  - P.e Illumina commits error in the G/T channels.

- **Solutions:**
  - Strand bias.

Adapted from Olson et al. Frontiers in Genetics. 2015

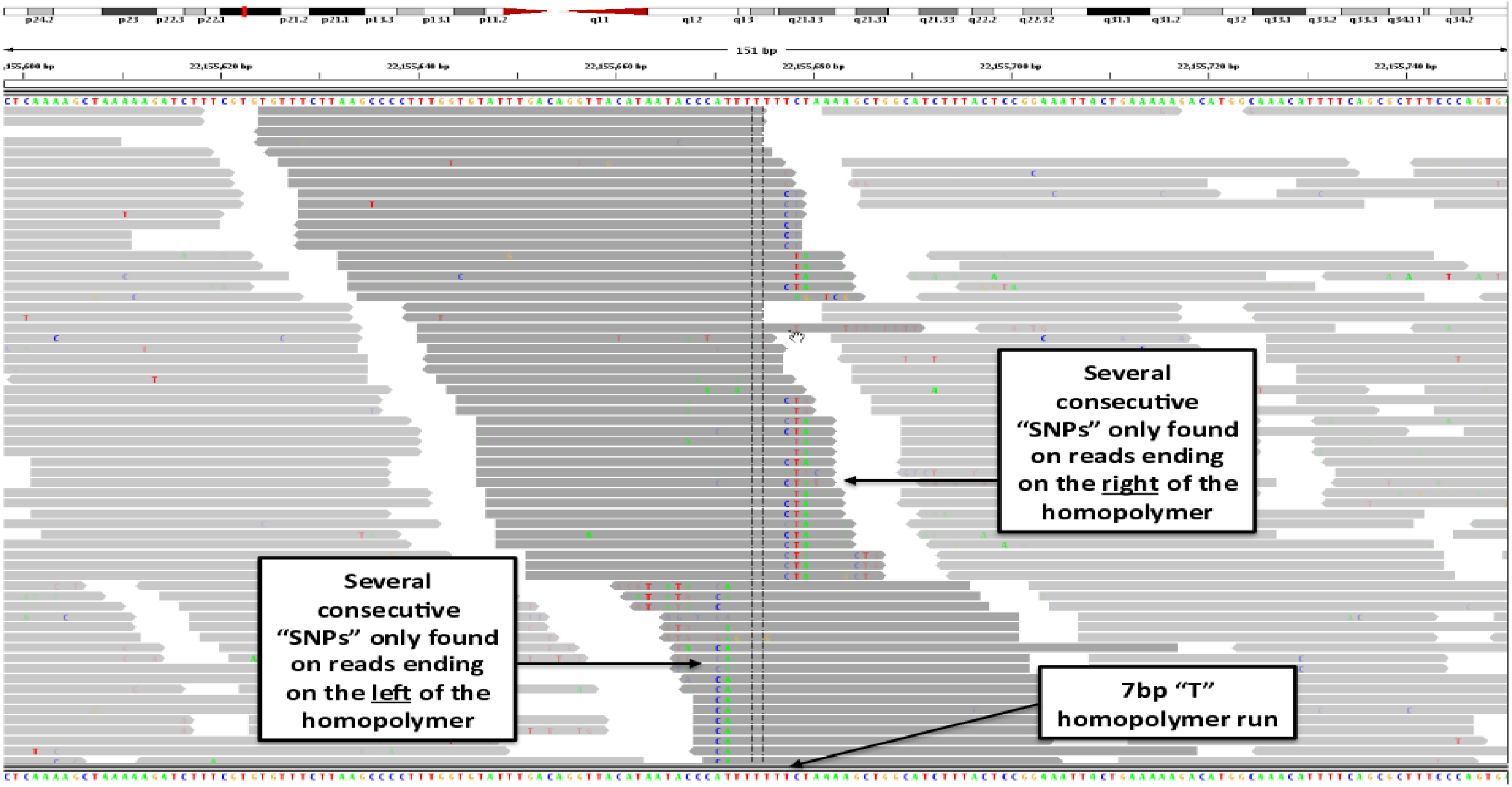
Sources of error and mitigation strategies

- **Mapping errors:**
  - Genomic duplication and structural variation.
  - High diverse areas.

- **Solutions**
  - Paired-end libraries.
  - Long reads / fragments.
  - MAPQ
  - Realignment around indels.

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

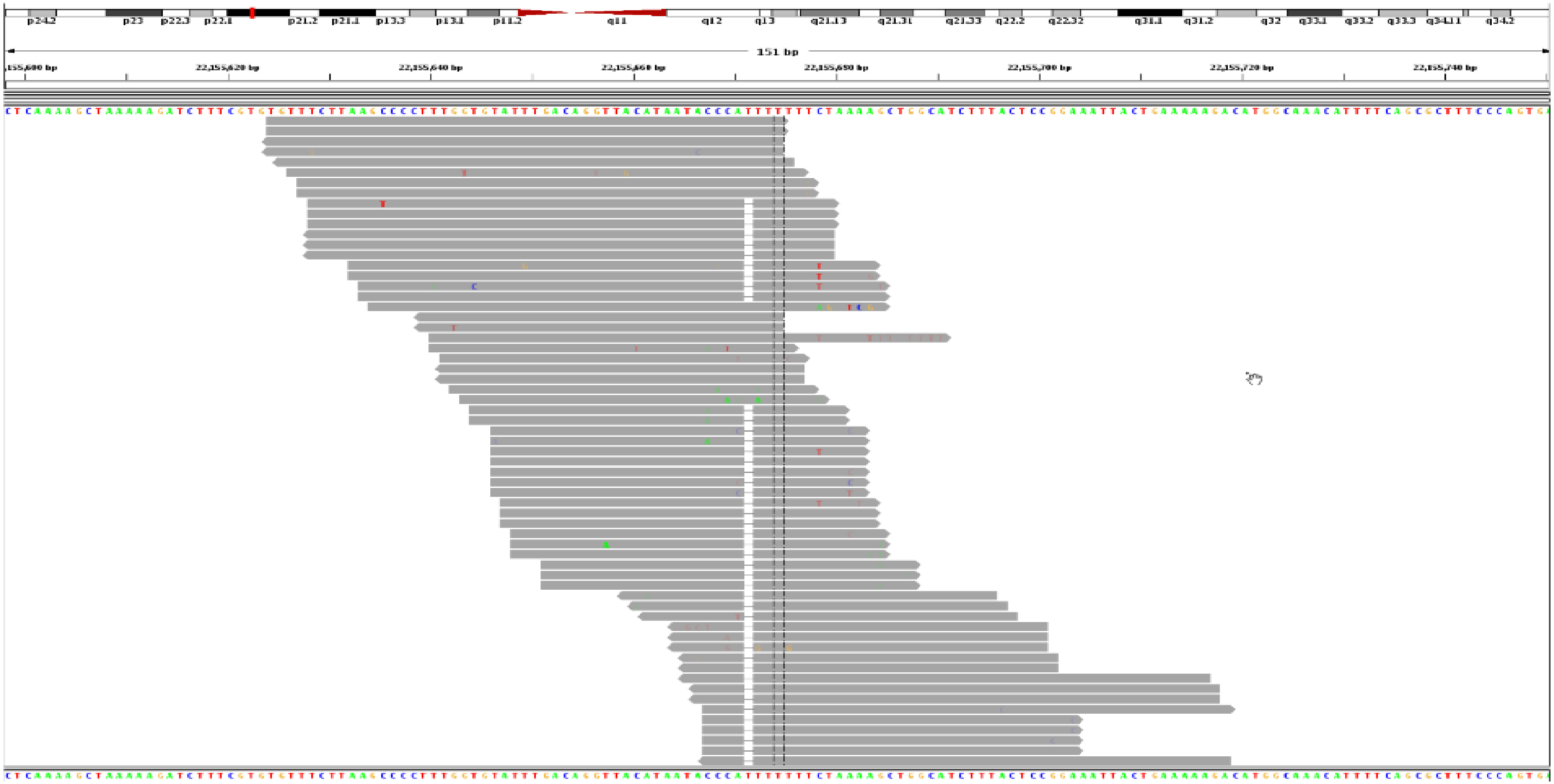


Several consecutive “SNPs” only found on reads ending on the right of the homopolymer

Several consecutive “SNPs” only found on reads ending on the left of the homopolymer

7bp “T” homopolymer run

Sources of error and mitigation strategies



Sources of error and mitigation strategies

- **SNP calling step**
  - Errors may result in base calling errors.
  - FP and FN calls.
- **Solutions**
  - Strand bias
  - Base quality rank sum
  - MAPQ
  - Hard filters:
    - Depth of coverage
    - Minimun base call frequency.

Adapted from Olson et al. Frontiers in Genetics. 2015

Reference selection

- Critial step <- Bias which SNPs are called.
- SNPs in genes not present in the reference **WON’T** be called.
- Less effect in clonal bacteria.
- Number of SNPs called vary **A LOT!**

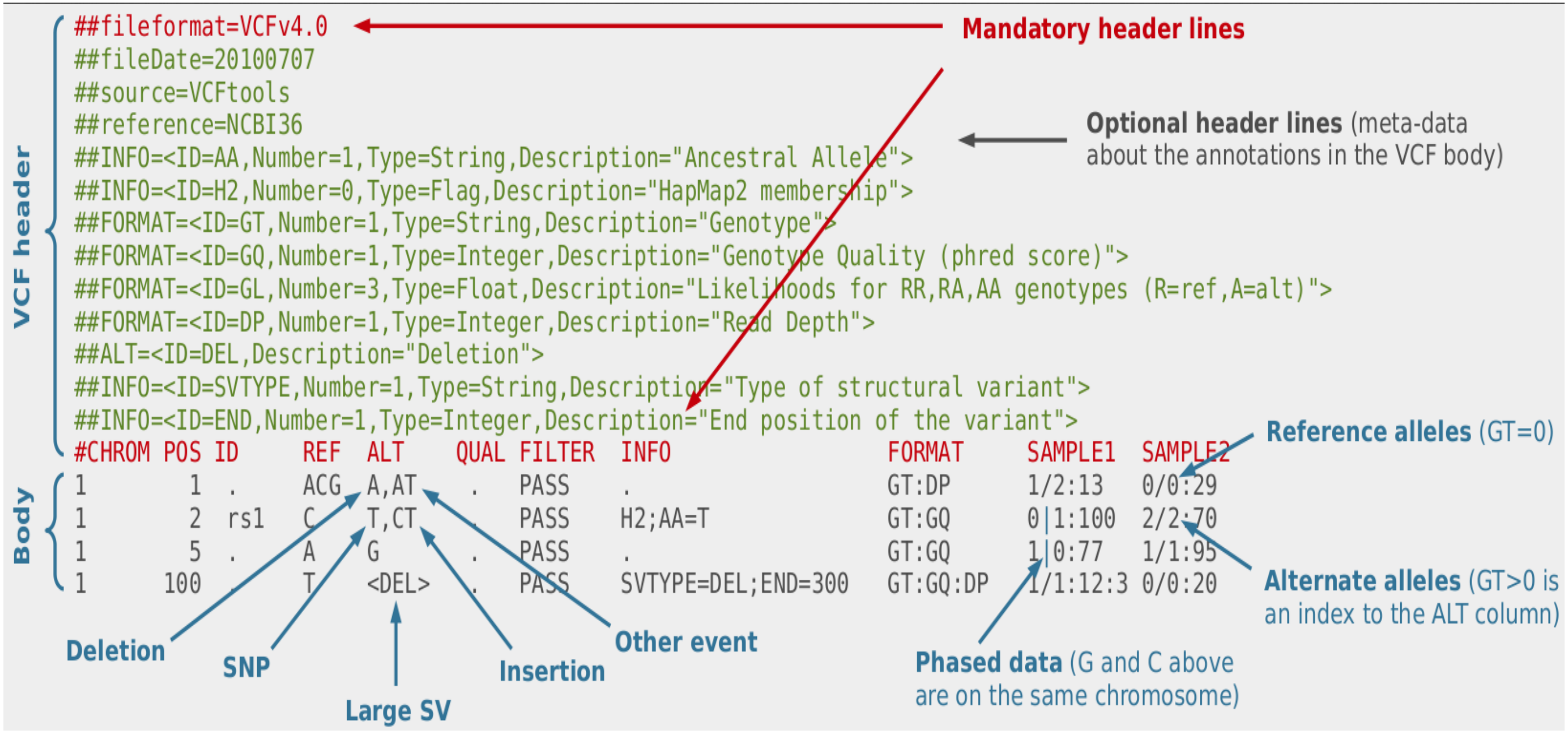
- **Solutions:**
  - Kmerfinder

Repetitive/Phage regions filtering

- **PHASTER**

- We can remove/mask phague/repetitive regions where reads won’t map.
- This way those areas will be out of analysis.
- Problem: those areas could be important!

VCF format



```
##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
```

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

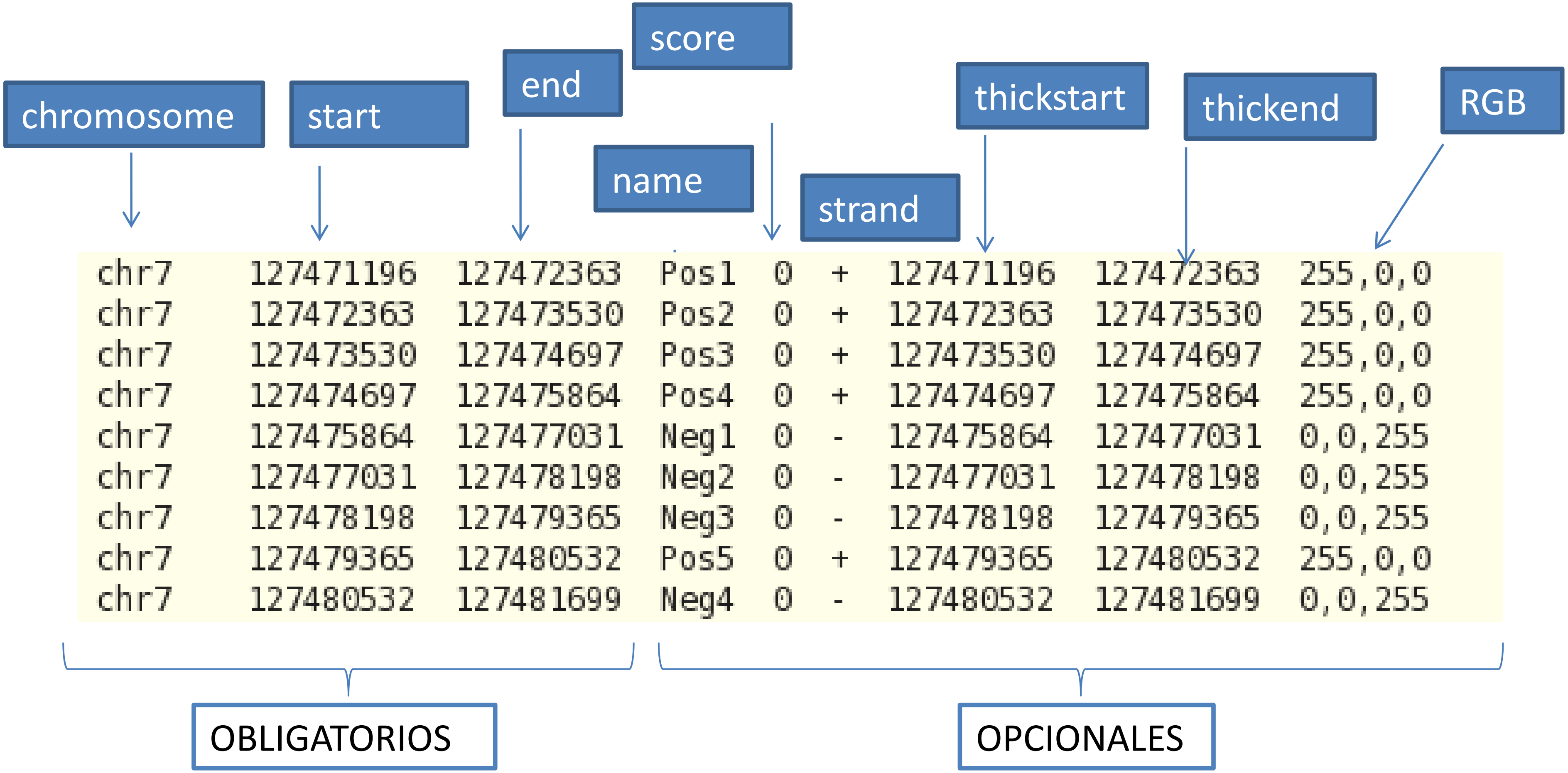
Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Bed format



chromosome

start

end

score

name

strand

thickstart

thickend

RGB

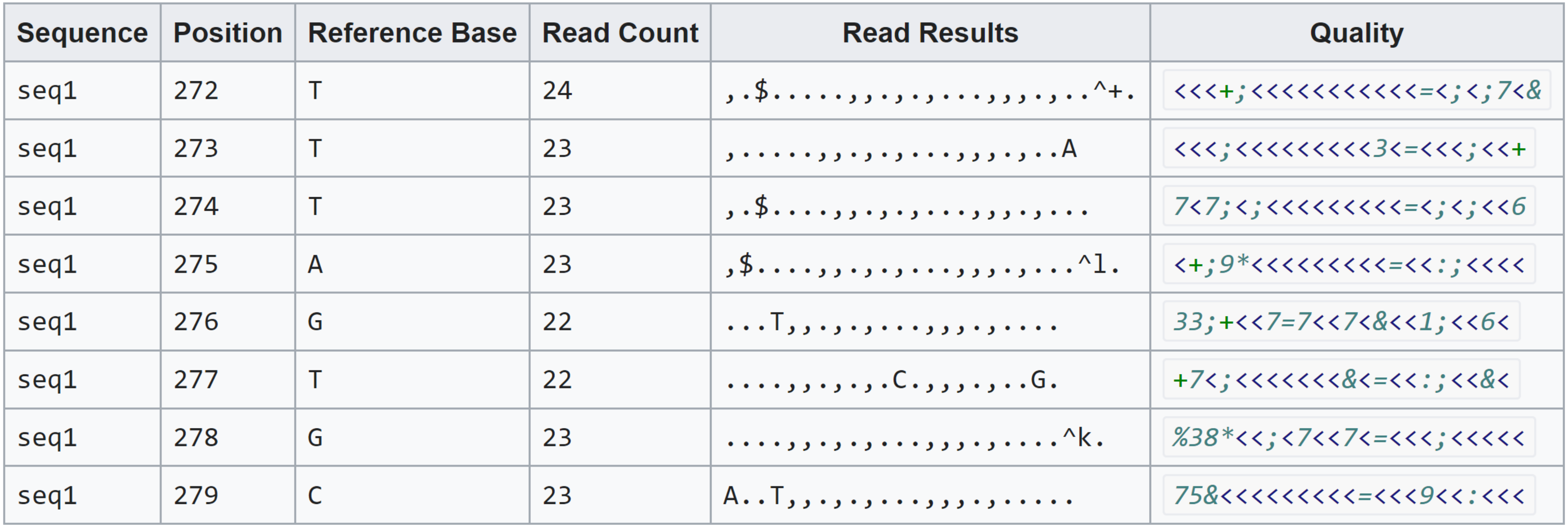
```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

OBLIGATORIOS

OPCIONALES

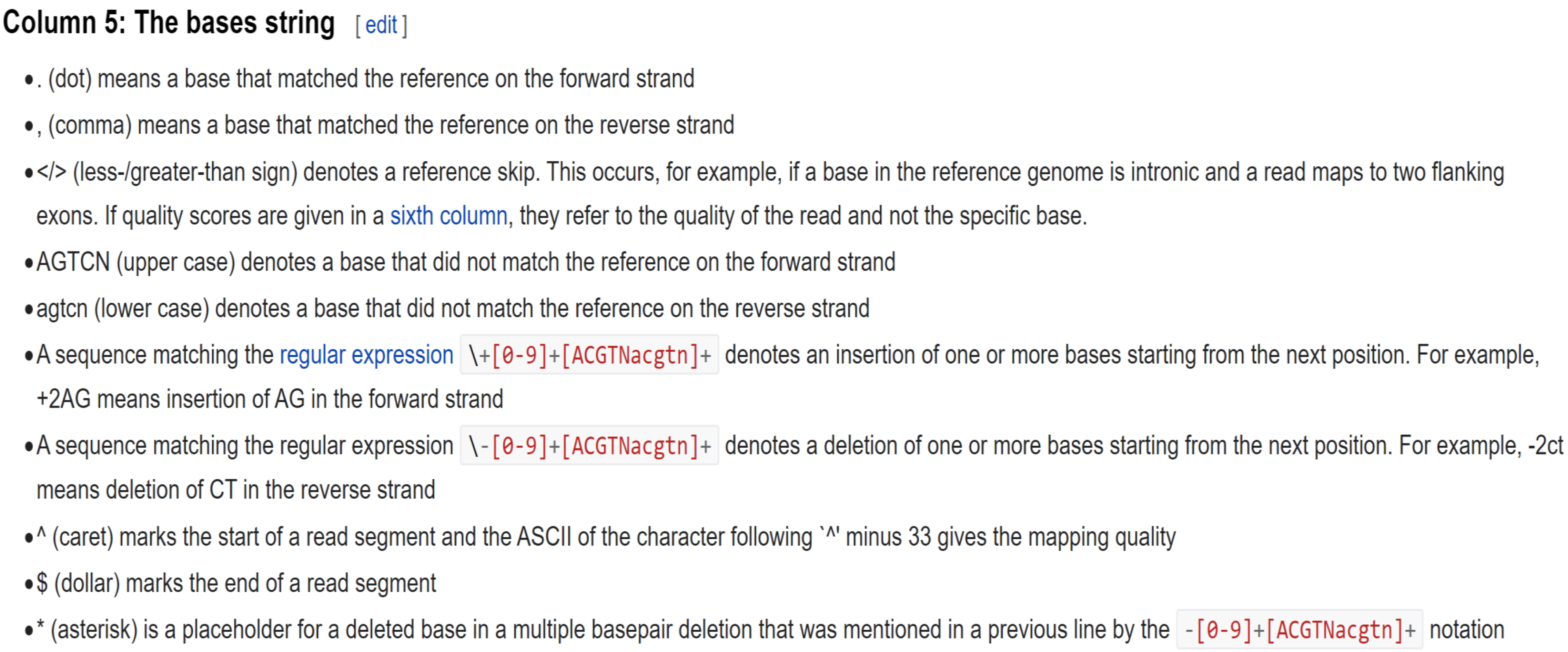
Análisis de genomas virales con galaxy

Mpileup format



Sequence	Position	Reference Base	Read Count	Read Results	Quality
seq1	272	T	24	,.$.....,,.,.,...,,,.,..^+.	<<<+;<<<<<<<<<<<=<;<;7<&
seq1	273	T	23	,.....,,.,.,...,,,.,..A	<<<;<<<<<<<<<3<=<<<;<<+
seq1	274	T	23	,.$....,,.,.,...,,,.,...	7<7;<;<<<<<<<<<=<;<;<<6
seq1	275	A	23	,$....,,.,.,...,,,.,...^l.	<+;9*<<<<<<<<<=<<:;<<<<
seq1	276	G	22	...T,,.,.,...,,,.,....	33;+<<7=7<<7<&<<1;<<6<
seq1	277	T	22	....,,.,.,.C.,,,.,..G.	+7<;<<<<<<<&<=<<:;<<&<
seq1	278	G	23	....,,.,.,...,,,.,....^k.	%38*<<;<7<<7<=<<<;<<<<<
seq1	279	C	23	A..T,,.,.,...,,,.,.....	75&<<<<<<<<<=<<<9<<:<<<

Mpileup format

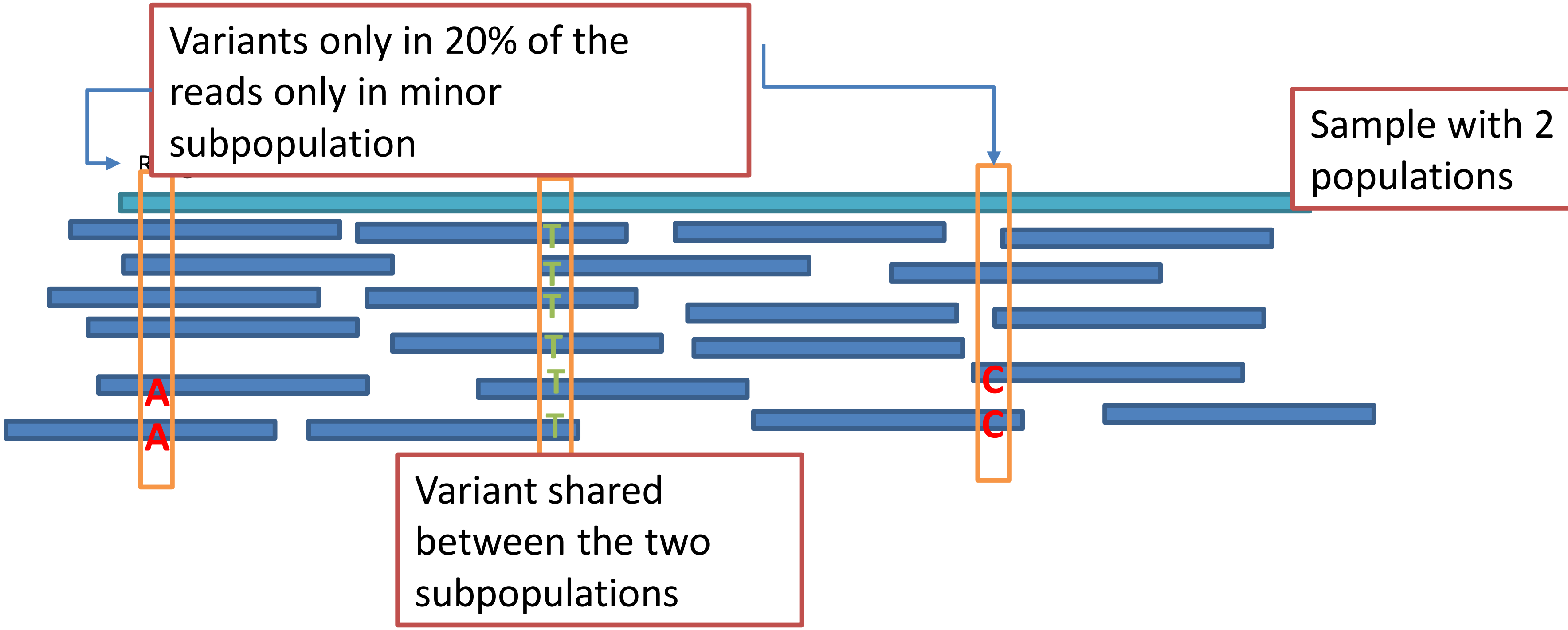


Column 5: The bases string [edit]

- . (dot) means a base that matched the reference on the forward strand
- , (comma) means a base that matched the reference on the reverse strand
- </> (less-/greater-than sign) denotes a reference skip. This occurs, for example, if a base in the reference genome is intronic and a read maps to two flanking exons. If quality scores are given in a sixth column, they refer to the quality of the read and not the specific base.
- AGTCN (upper case) denotes a base that did not match the reference on the forward strand
- agtcn (lower case) denotes a base that did not match the reference on the reverse strand
- A sequence matching the regular expression `\+[0-9]+[ACGTNacgtn]+` denotes an insertion of one or more bases starting from the next position. For example, +2AG means insertion of AG in the forward strand
- A sequence matching the regular expression `\-[0-9]+[ACGTNacgtn]+` denotes a deletion of one or more bases starting from the next position. For example, -2ct means deletion of CT in the reverse strand
- ^ (caret) marks the start of a read segment and the ASCII of the character following `^' minus 33 gives the mapping quality
- $ (dollar) marks the end of a read segment
- * (asterisk) is a placeholder for a deleted base in a multiple basepair deletion that was mentioned in a previous line by the `-[0-9]+[ACGTNacgtn]+` notation

Viral subpopulation - Quasispecies

- Just as in clonal subpopulations in tumor samples, we can have viral subpopulations called quasispecies in viral samples.
- We detect them using the alternative allele frequency.



Variants only in 20% of the reads only in minor subpopulation

Sample with 2 populations

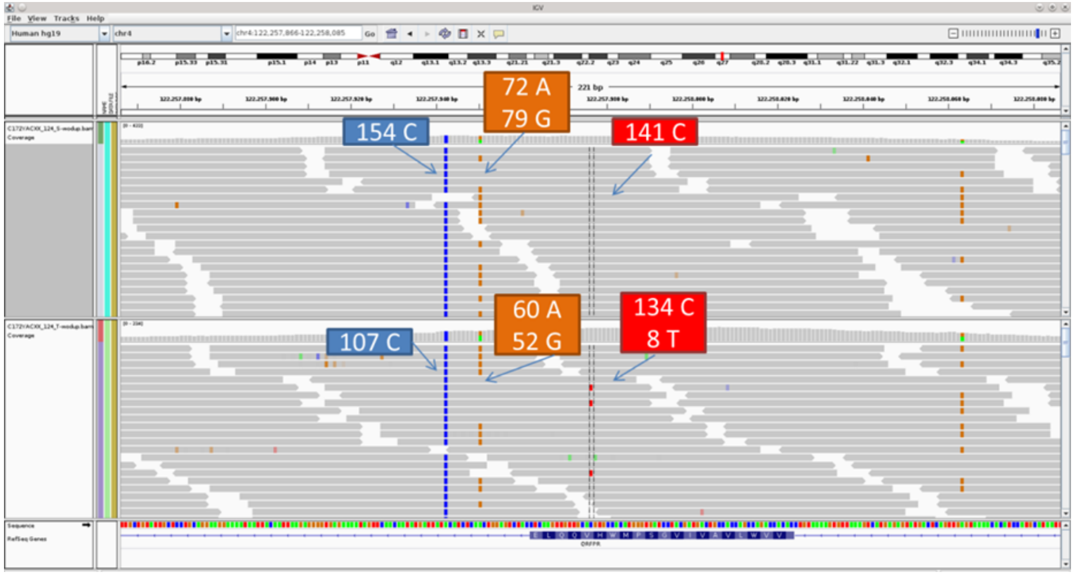
Variant shared between the two subpopulations

Population Allele frequency vs Sample Allele frequency

- **Population allele frequency:** probability of finding an allele in the population. Number of individuals carrying an allele vs total of individuals in the population.

Population Allele frequency vs Sample Allele frequency

- **Alternate/Base allele frequency**: number of reads supporting the alternate allele vs total of reads.



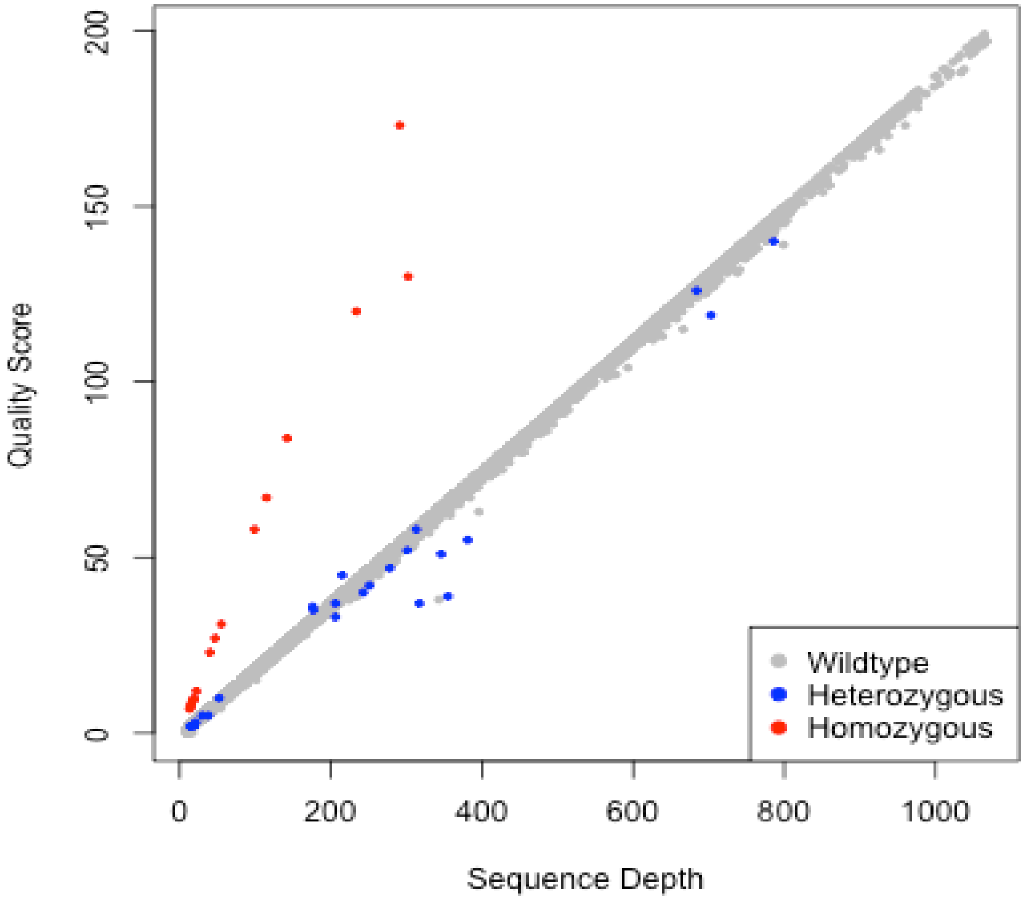
VARSCAN2

- Uses a heuristic/statistic method instead of bayesian.
- Allows more flexibility and hard filters.

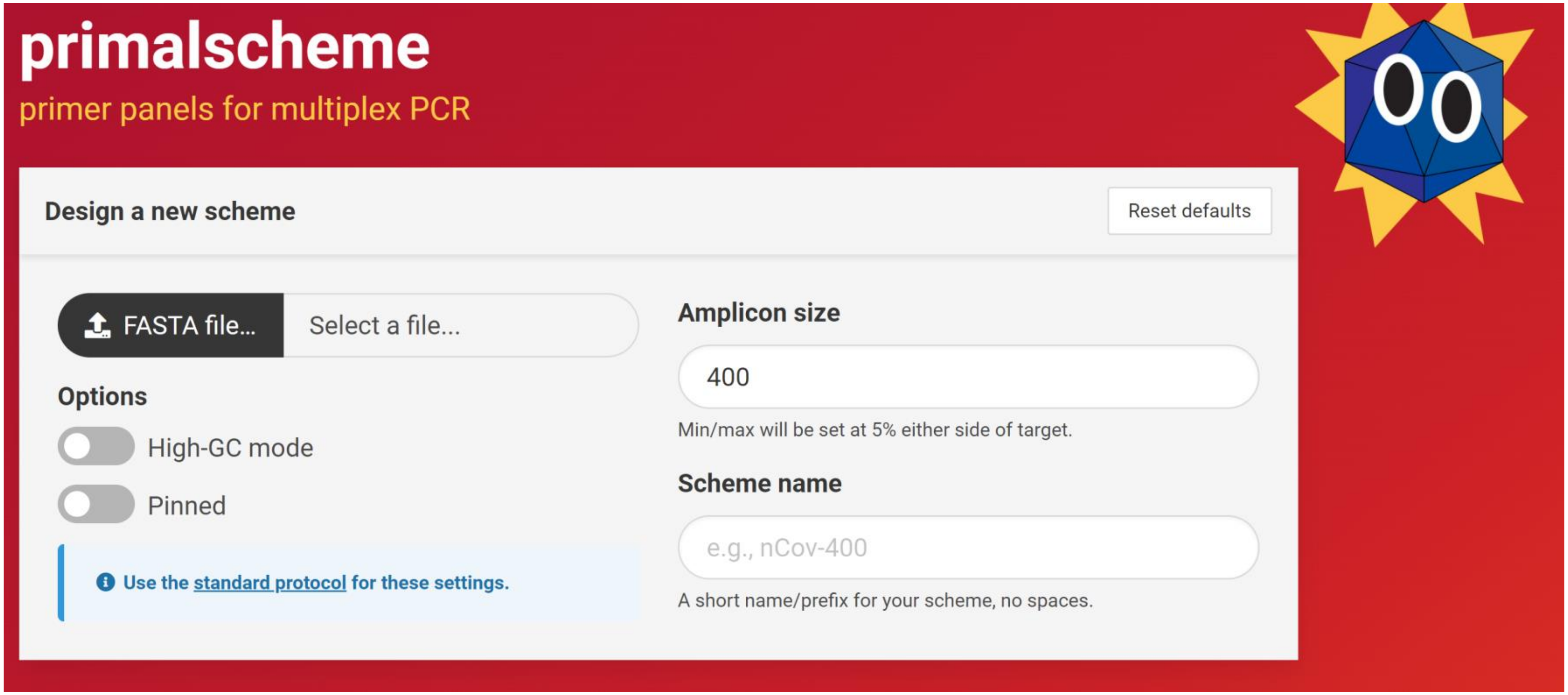
```
samtools pileup -f reference.fasta myData.bam | java -jar VarScan.v2.2.jar pileup2snp
```

http://varscan.sourceforge.net/

VARSCAN2


Quality Score
0
50
100
150
200
0
200
400
600
800
1000
Sequence Depth
● Wildtype
● Heterozygous
● Homozygous

IVAR



primalscheme

primer panels for multiplex PCR

Design a new scheme

Options

○ High-GC mode

○ Pinned

Use the standard protocol for these settings.

Amplicon size

400

Min/max will be set at 5% either side of target.

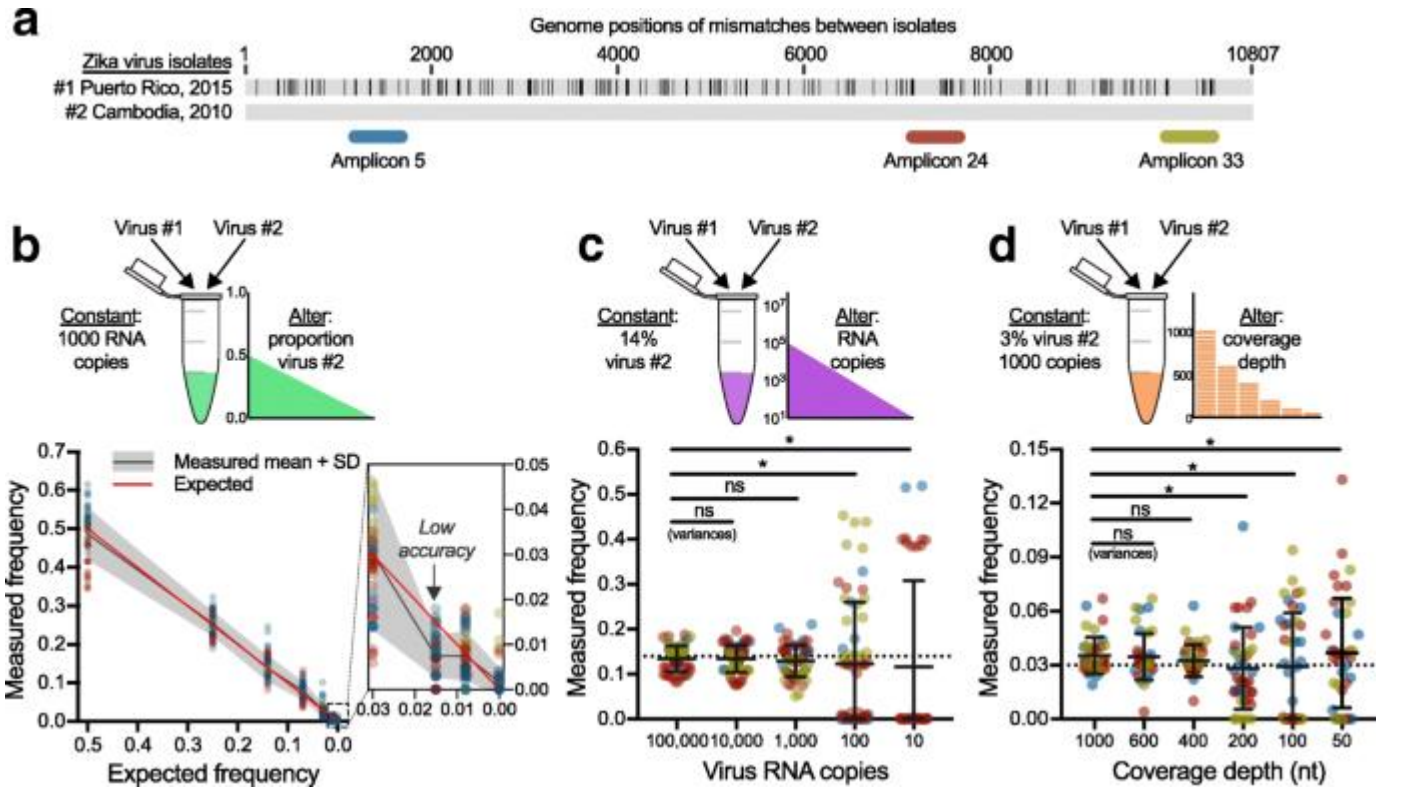
Scheme name

e.g., nCov-400

A short name/prefix for your scheme, no spaces.

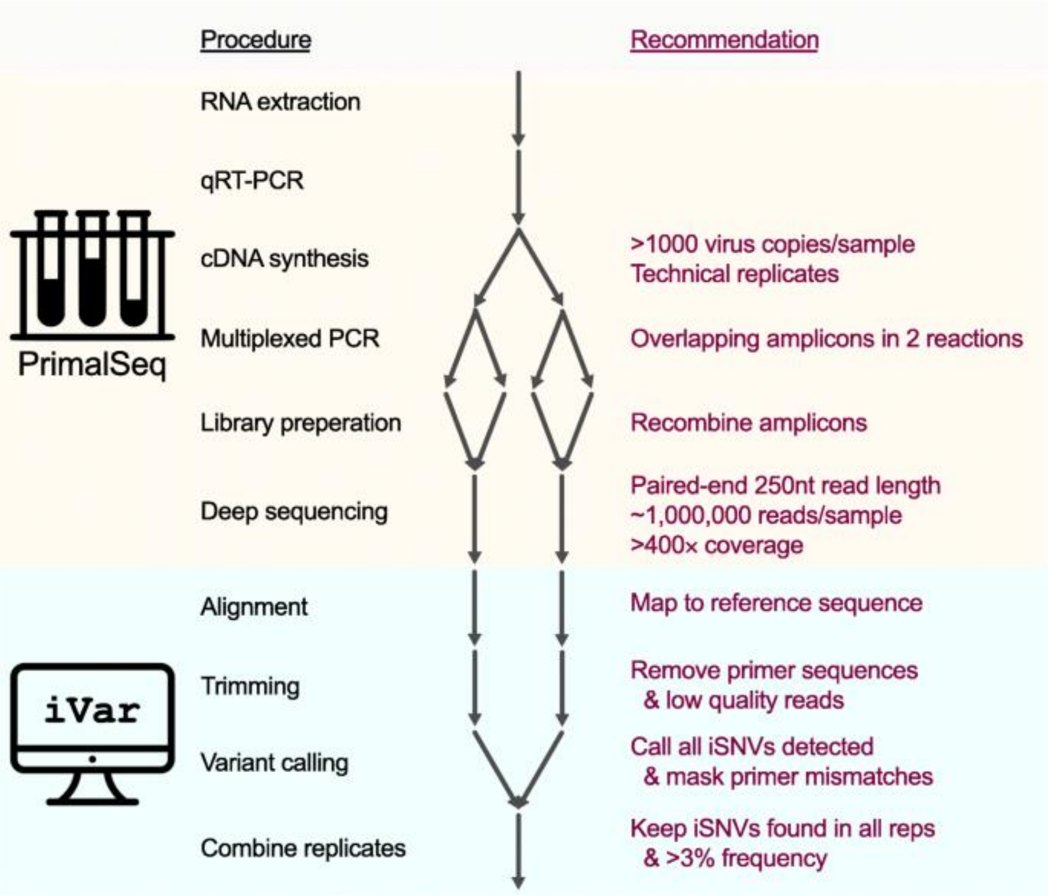
Quick et al. Nature Protoc. 2017

IVAR



Grubaugh et al. Genome Biology. 2019

IVAR


Procedure

Recommendation

RNA extraction

qRT-PCR

cDNA synthesis — >1000 virus copies/sample
Technical replicates

Multiplexed PCR — Overlapping amplicons in 2 reactions

Library preperation — Recombine amplicons

Deep sequencing — Paired-end 250nt read length
~1,000,000 reads/sample
>400x coverage

Alignment — Map to reference sequence

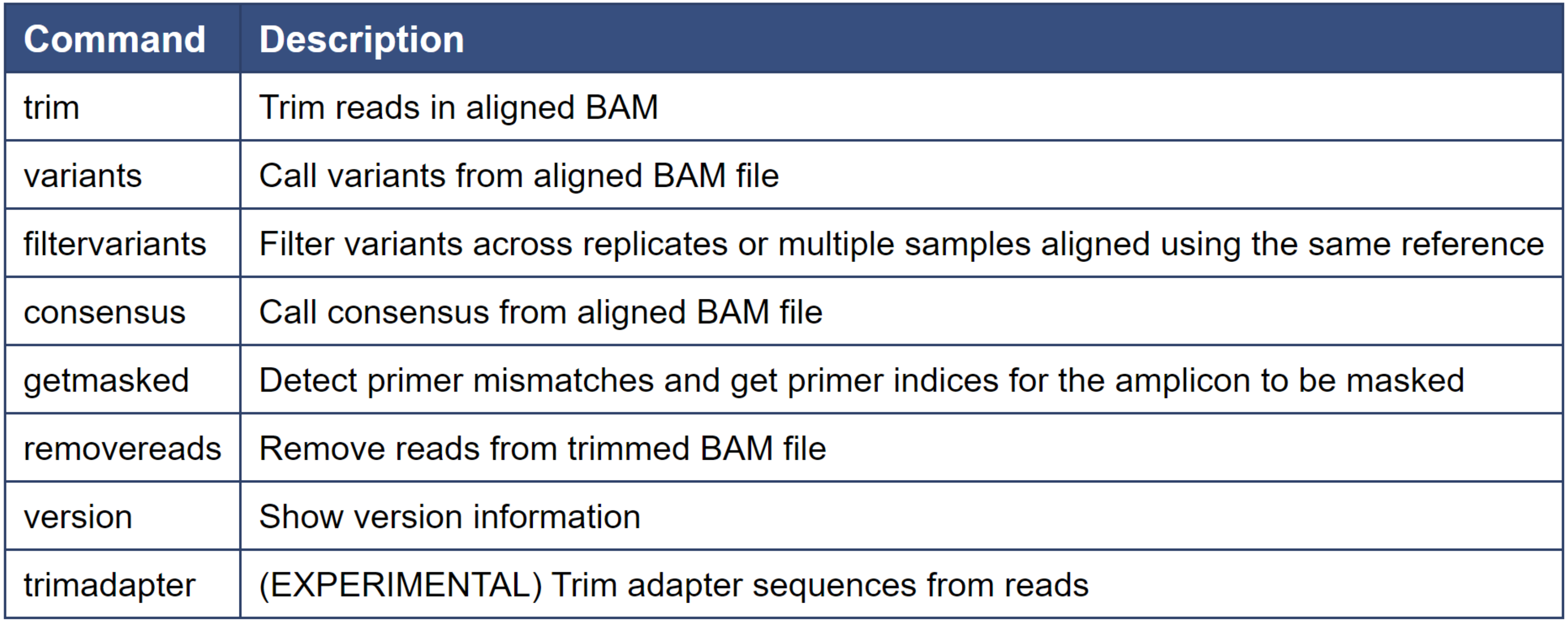
Trimming — Remove primer sequences & low quality reads

Variant calling — Call all iSNVs detected & mask primer mismatches

Combine replicates — Keep iSNVs found in all reps & >3% frequency

Grubaugh et al. Genome Biology. 2019

IVAR



Command	Description
trim	Trim reads in aligned BAM
variants	Call variants from aligned BAM file
filtervariants	Filter variants across replicates or multiple samples aligned using the same reference
consensus	Call consensus from aligned BAM file
getmasked	Detect primer mismatches and get primer indices for the amplicon to be masked
removereads	Remove reads from trimmed BAM file
version	Show version information
trimadapter	(EXPERIMENTAL) Trim adapter sequences from reads

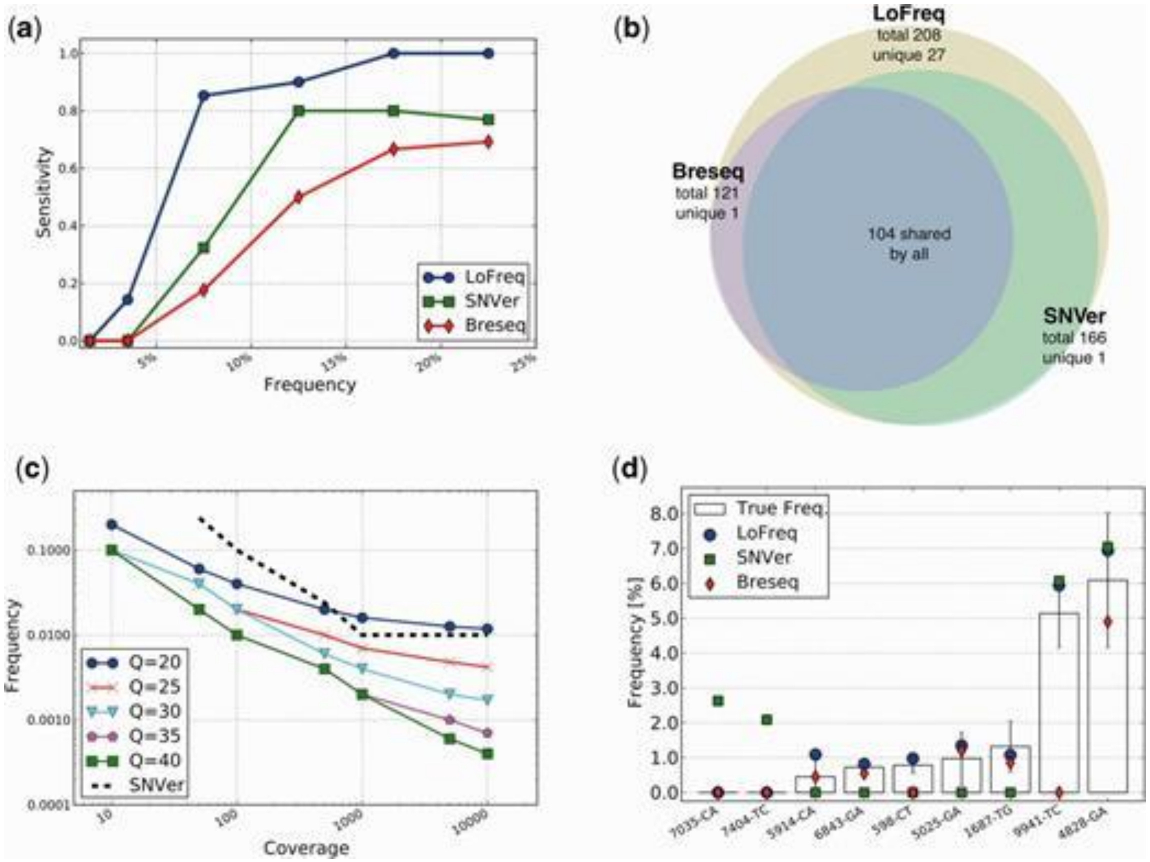
Grubaugh et al. Genome Biology. 2019

IVAR

- Input Options Description
  - -q Minimum quality score threshold to count base (Default: 20)
  - -t Minimum frequency threshold(0 - 1) to call variants (Default: 0.03)
  - -m Minimum read depth to call variants (Default: 0)
  - -r Reference file used for alignment. This is used to translate the nucleotide sequences and identify intra host single nucleotide variants
  - -g A GFF file in the GFF3 format can be supplied to specify coordinates of open reading frames (ORFs). In absence of GFF file, amino acid translation will not be done.
- Output Options Description
  - -p (Required) Prefix for the output tsv variant file

Grubaugh et al. Genome Biology. 2019

LOFREQ



Wilm et al. Nucleic Acids Res. 2012

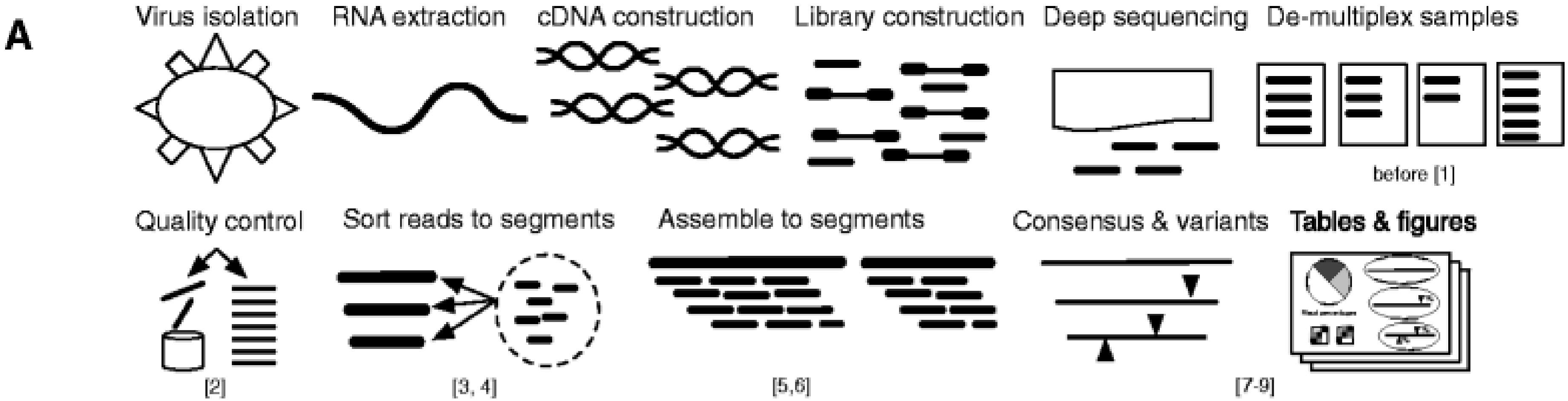
LOFREQ

- **lofreq viterbi:** realignment algorithm
- **lofreq call**: Warning! Only SNPs are called by default.
- **lofreq filter:** vcf filtering.

http://csb5.github.io/lofreq/

IRMA

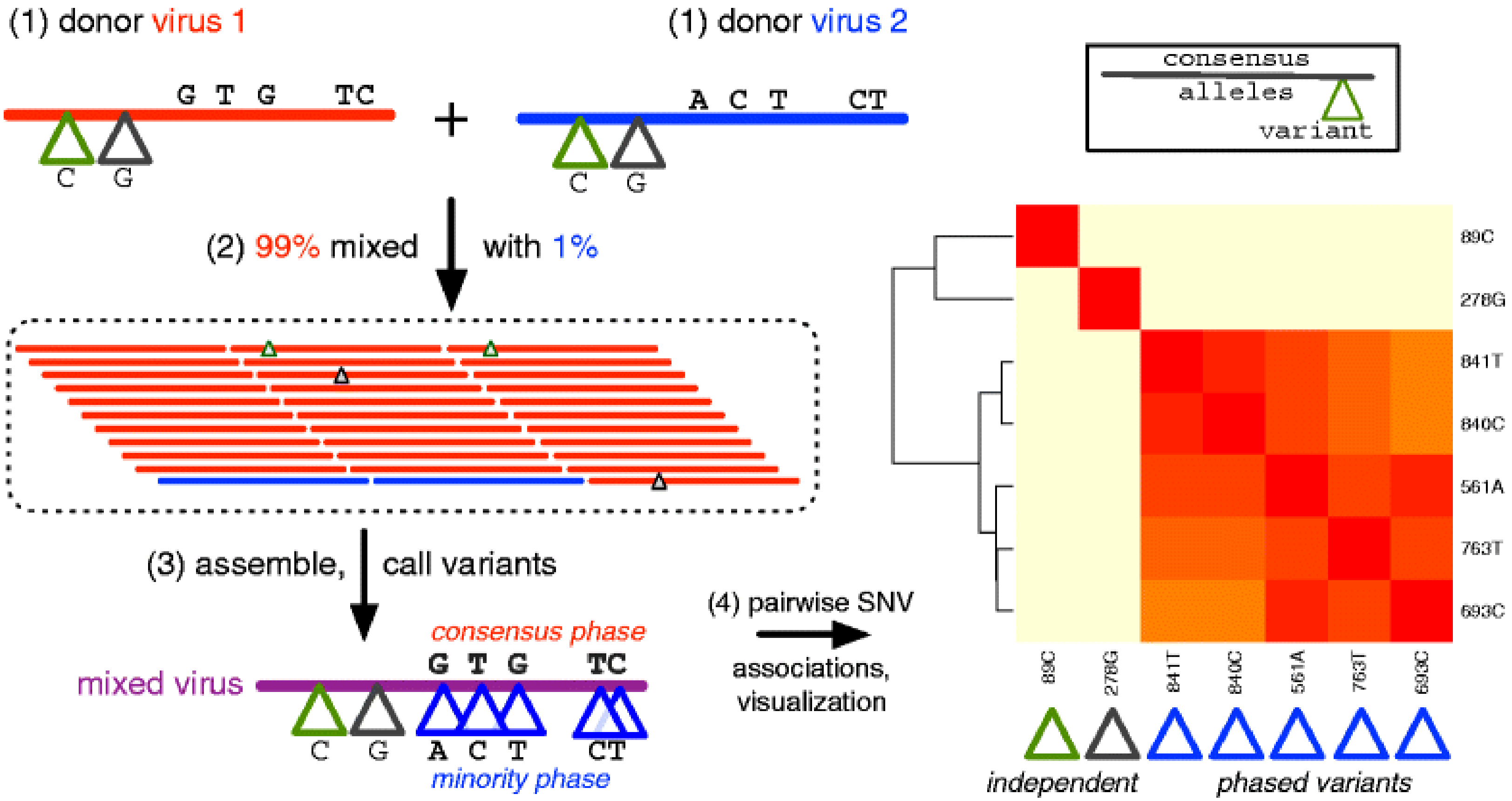
IRMA: Iterative Refinement Meta-Assembler



Shepard et al BMC Genomics 2016, **17**:708

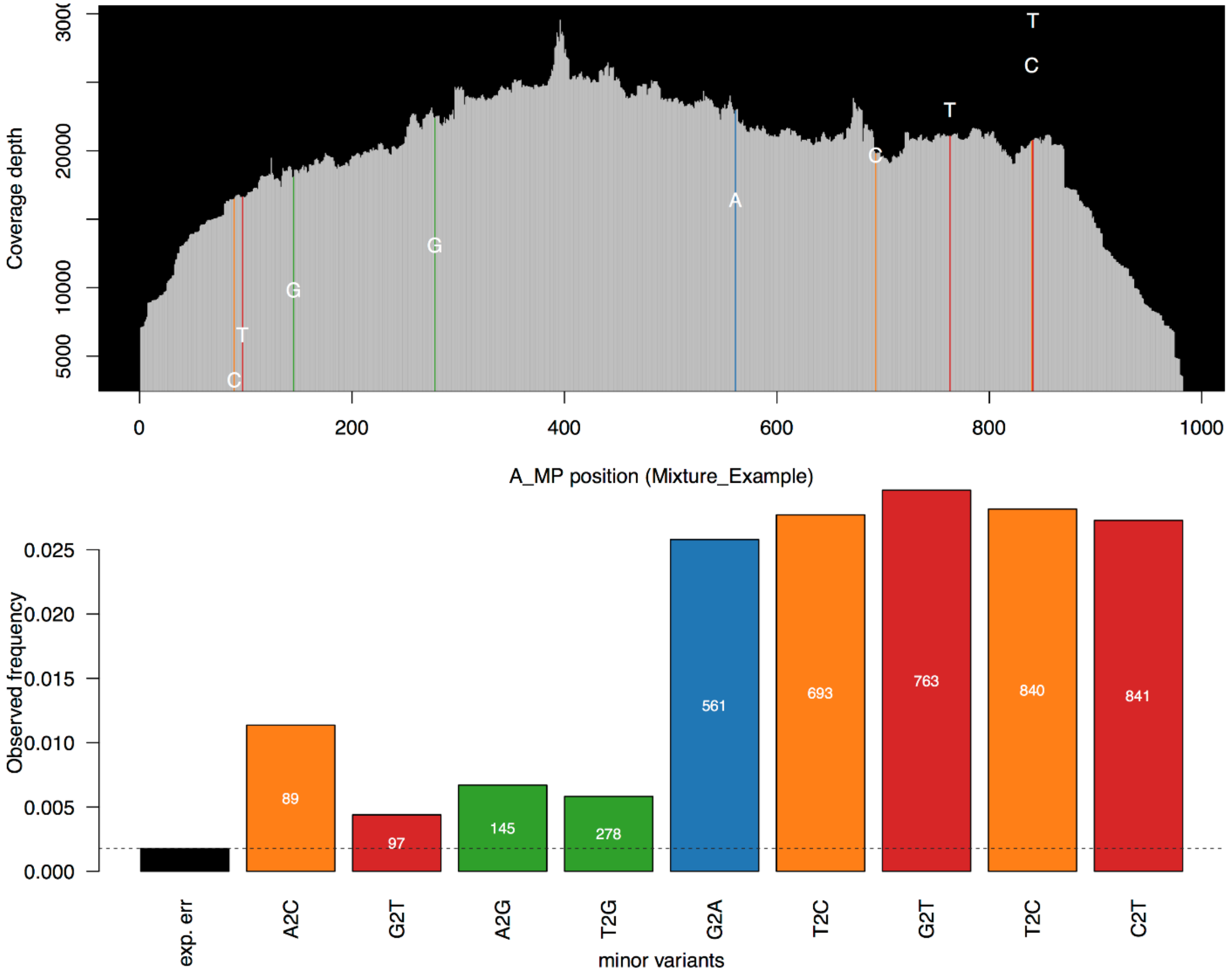
Available for **LU**, ***FLU_AD**, **EBOLA**, & **‡CoV & Sars-cov-2**

IRMA

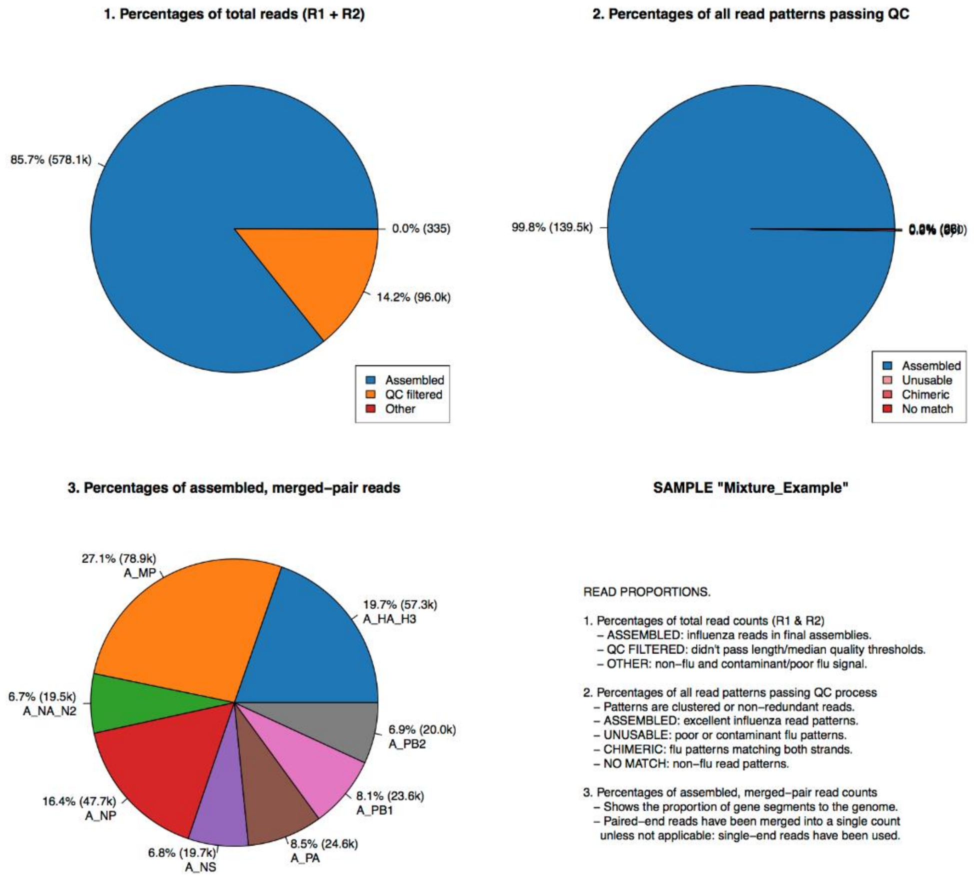


Shepard et al BMC Genomics 2016, **17**:708

IRMA


Coverage depth
300(
20000
10000
5000
0
200
400
600
800
1000
T
C
T
C
A
G
G
T
C
A_MP position (Mixture_Example)
Observed frequency
0.025
0.020
0.015
0.010
0.005
0.000
89
97
145
278
561
693
763
840
841
exp. err
A2C
G2T
A2G
T2G
G2A
T2C
G2T
T2C
C2T
minor variants

IRMA



SAMPLE "Mixture_Example"

READ PROPORTIONS.

1. Percentages of total read counts (R1 & R2)
   - ASSEMBLED: influenza reads in final assemblies.
   - QC FILTERED: didn't pass length/median quality thresholds.
   - OTHER: non−flu and contaminant/poor flu signal.

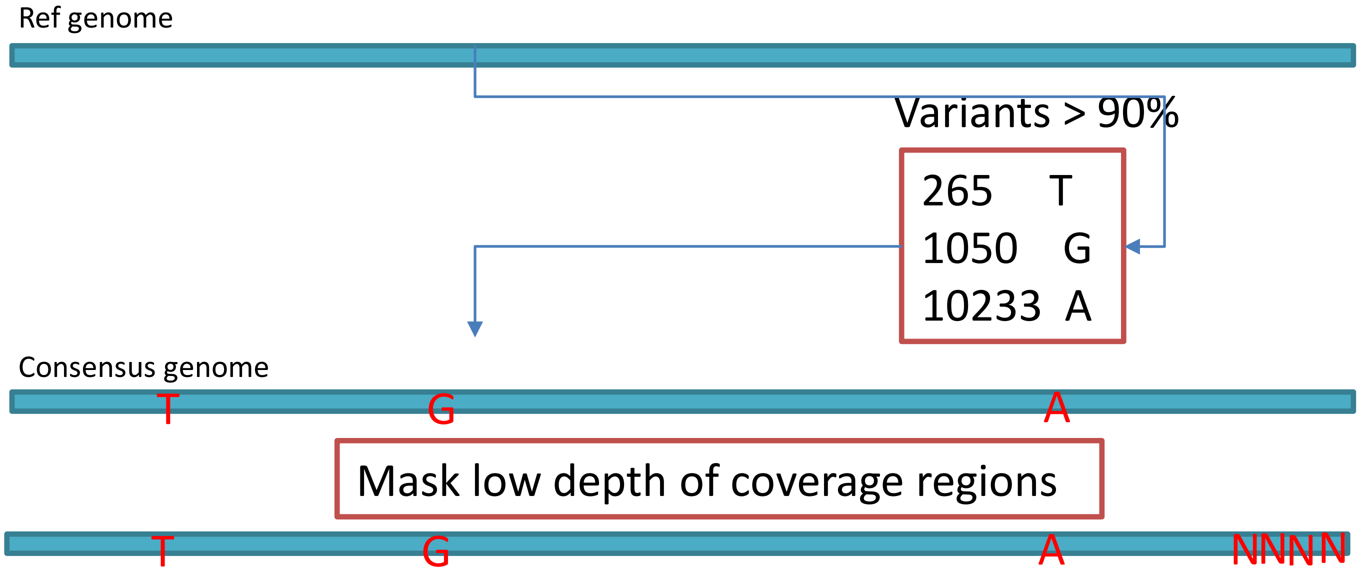
2. Percentages of all read patterns passing QC process
   - Patterns are clustered or non−redundant reads.
   - ASSEMBLED: excellent influenza read patterns.
   - UNUSABLE: poor or contaminant flu patterns.
   - CHIMERIC: flu patterns matching both strands.
   - NO MATCH: non−flu read patterns.

3. Percentages of assembled, merged−pair read counts
   - Shows the proportion of gene segments to the genome.
   - Paired−end reads have been merged into a single count unless not applicable: single−end reads have been used.

Consensus genome

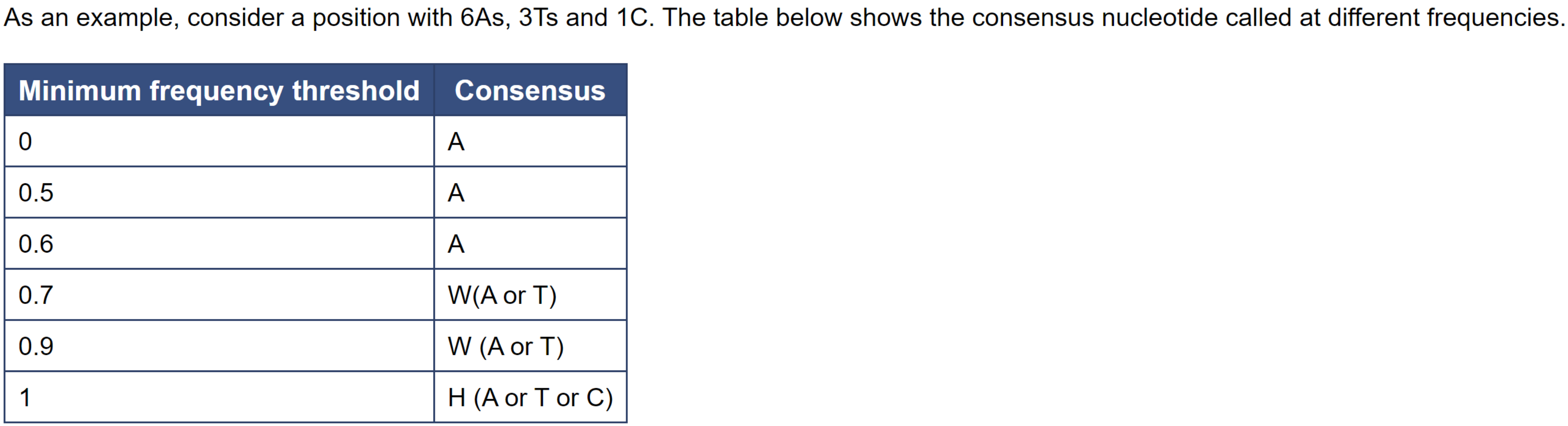
Aproximation 1

- Select variants: > 80% allele frequency
- Include variants in reference genome.
- Mask low frequency positions: <10x.



Consensus genome

- **Approximation 2 (ivar)**
  - Minimum frequency threshold is the minimum frequency that a base must match to be called as the consensus base at a position. If one base is not enough to match a given frequency, then an ambigious nucleotide is called at that position



As an example, consider a position with 6As, 3Ts and 1C. The table below shows the consensus nucleotide called at different frequencies.

Minimum frequency threshold	Consensus
0	A
0.5	A
0.6	A
0.7	W(A or T)
0.9	W (A or T)
1	H (A or T or C)

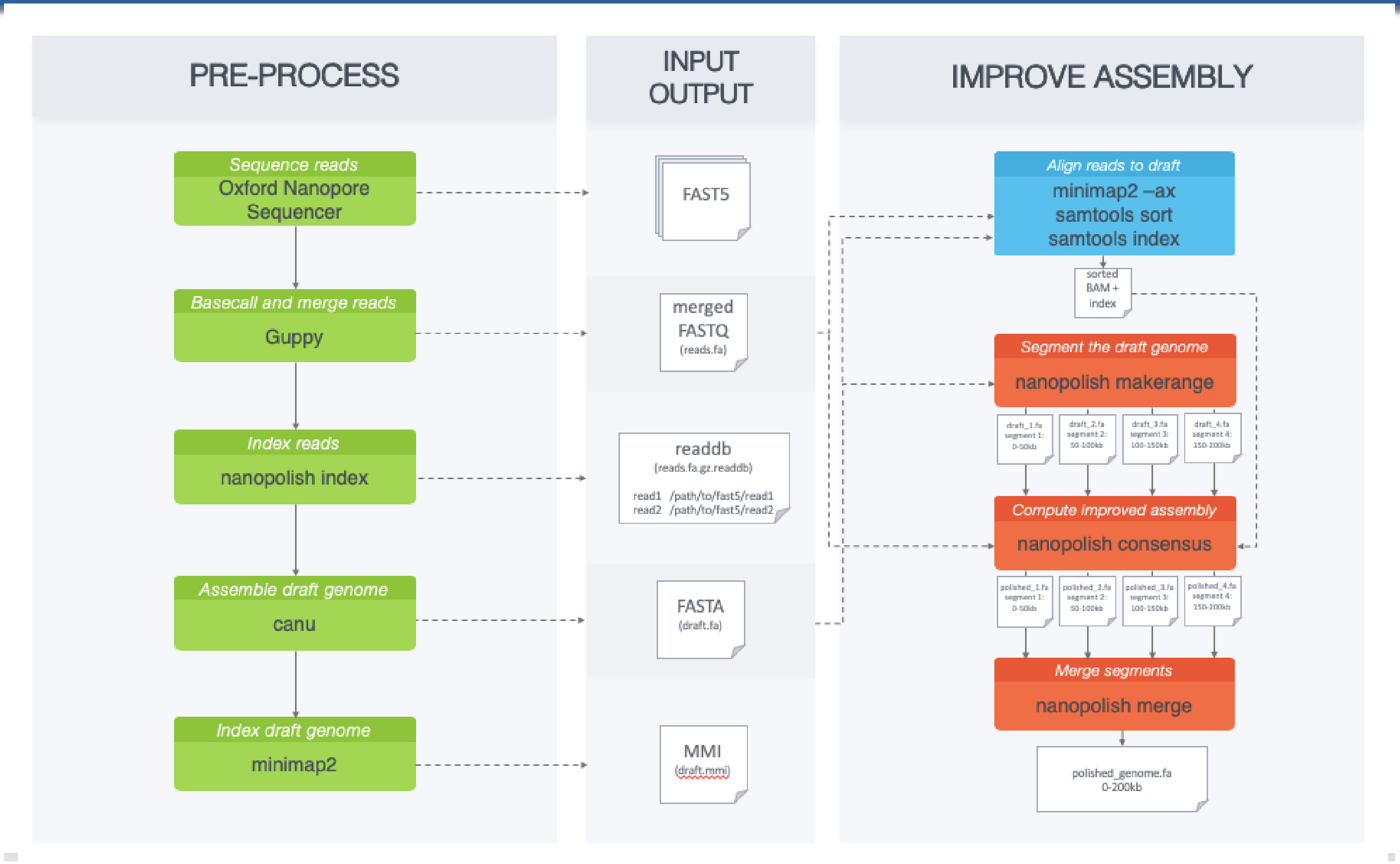
Variant Calling in Nanopore

- **Medaka** is a tool to create consensus sequences and variant calls from nanopore sequencing data. This task is performed using neural networks applied a pileup of individual sequencing reads against a draft assembly.

- medaka consensus: Creates a consensus from a draft assembly in fasta format and the nanopore corrected fastq reads.

Variant Calling in Nanopore

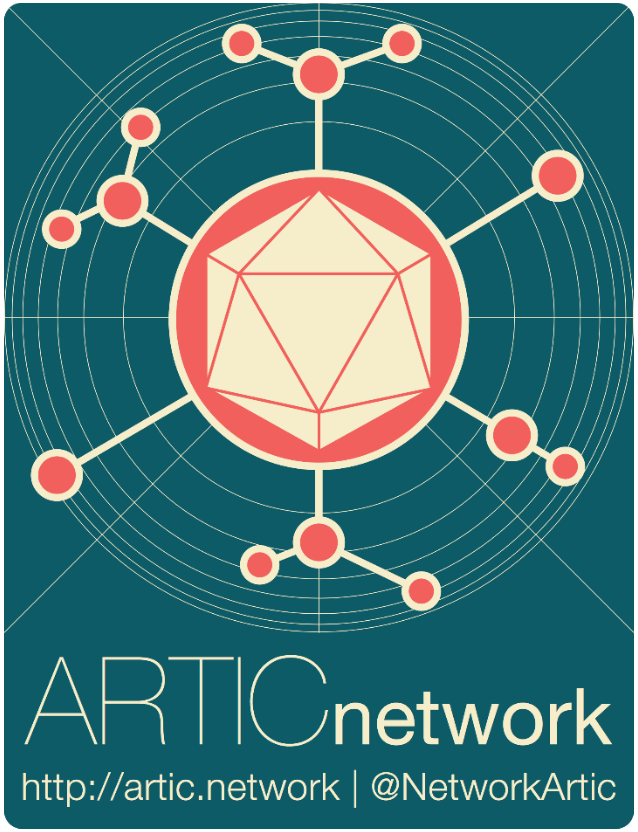
- **Nanopolish** can calculate an improved consensus sequence for a draft genome assembly, detect base modifications, call SNPs and indels with respect to a reference genome and more


PRE-PROCESS
INPUT OUTPUT
IMPROVE ASSEMBLY
Sequence reads
Oxford Nanopore Sequencer
Basecall and merge reads
Guppy
Index reads
nanopolish index
Assemble draft genome
canu
Index draft genome
minimap2
FAST5
merged FASTQ (reads.fa)
readdb (reads.fa.gz.readdb)
read1 /path/to/fast5/read1
read2 /path/to/fast5/read2
FASTA (draft.fa)
MMI (draft.mmi)
Align reads to draft
minimap2 –ax
samtools sort
samtools index
sorted BAM + index
Segment the draft genome
nanopolish makerange
draft_1.fa segment 1: 0-50kb
draft_2.fa segment 2: 50-100kb
draft_3.fa segment 3: 100-150kb
draft_4.fa segment 4: 150-200kb
Compute improved assembly
nanopolish consensus
polished_1.fa segment 1: 0-50kb
polished_2.fa segment 2: 50-100kb
polished_3.fa segment 3: 100-150kb
polished_4.fa segment 4: 150-200kb
Merge segments
nanopolish merge
polished_genome.fa 0-200kb

Variant Calling in Nanopore

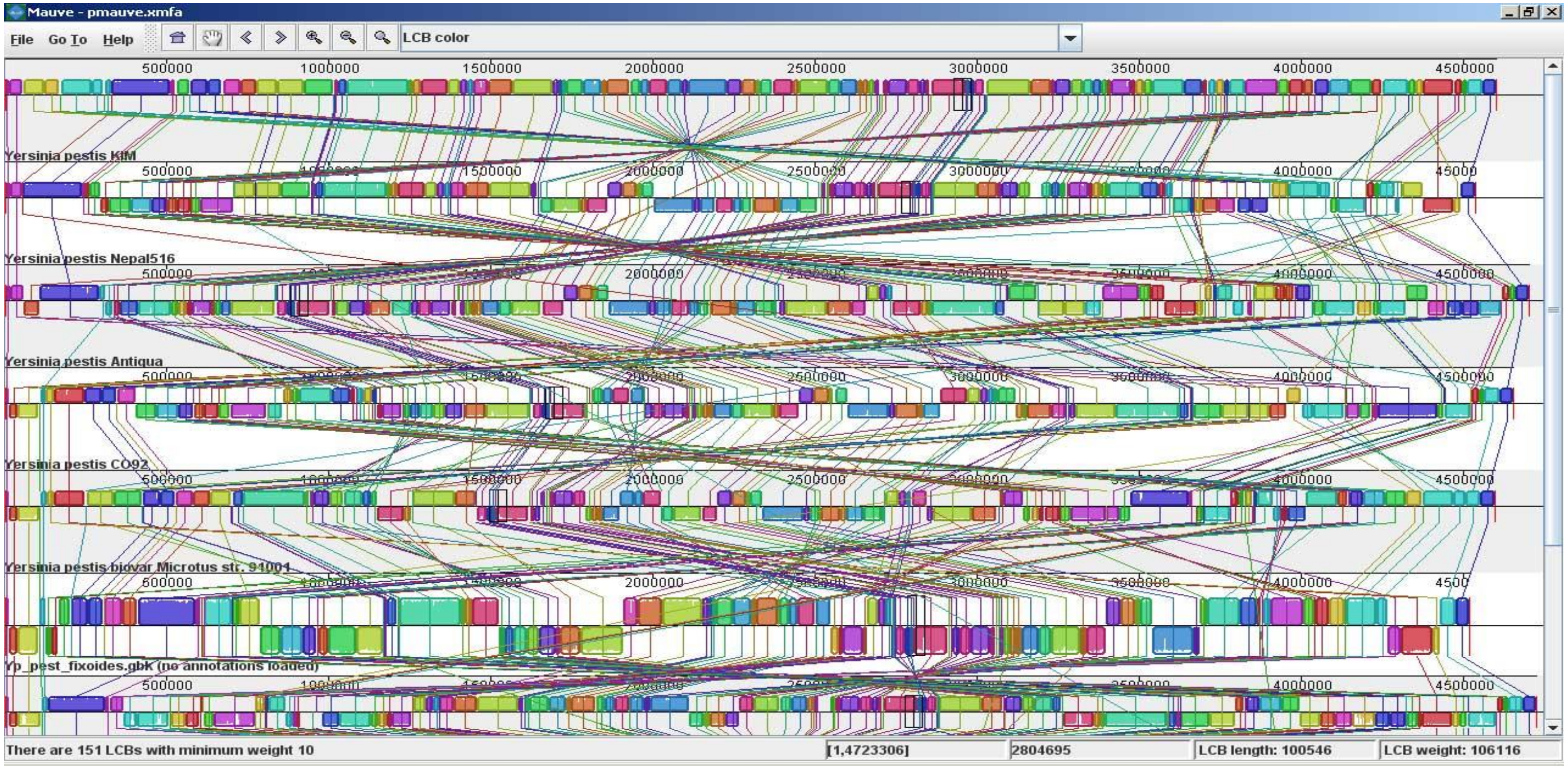
- **Artic protocol**

Full protocol from wet to drylab. Includes basecalling, preprocessing, mapping and consensus generation for Artic primers.



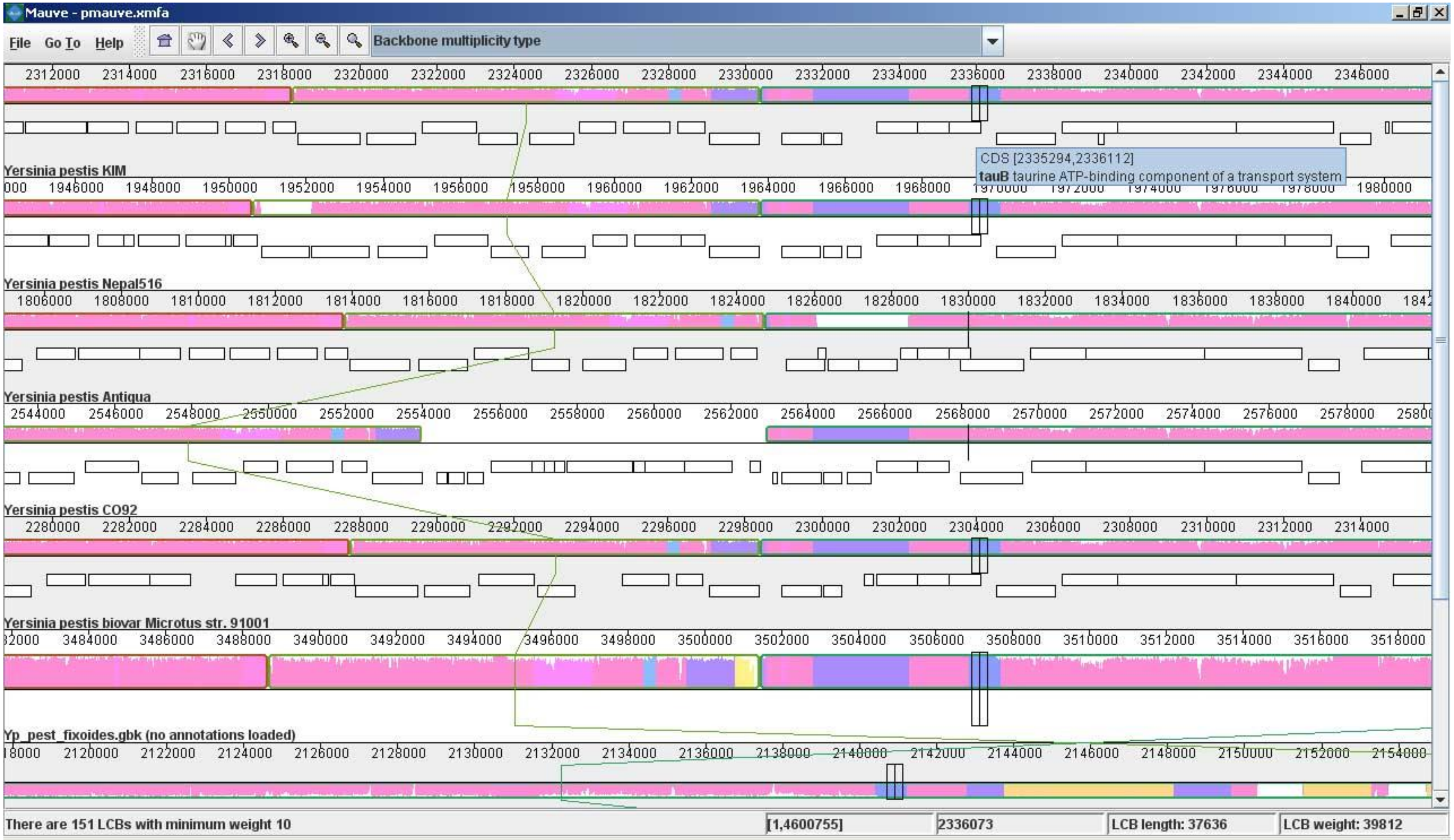
Consensus genome comparison: Mauve

Mauve is a complete genoma aligner



There are 151 LCBs with minimum weight 10 | [1,4723306] | 2804695 | LCB length: 100546 | LCB weight: 106116

Consensus genome comparison: Mauve


Mauve - pmauve.xmfa

Yersinia pestis KIM

CDS [2335294,2336112]
tauB taurine ATP-binding component of a transport system

Yersinia pestis Nepal516

Yersinia pestis Antiqua

Yersinia pestis CO92

Yersinia pestis biovar Microtus str. 91001

Yp_pest_fixoides.gbk (no annotations loaded)

There are 151 LCBs with minimum weight 10 | [1,4600755] | 2336073 | LCB length: 37636 | LCB weight: 39812

Variants annotation

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of genetic variants (such as amino acid changes).

It needs an annotation database, there are few for virus as default, commonly you need to build it using a gff file if available.

Output: annotated vcf

SnpSift converts snpeff output to a table.

Thanks for your attention!