Mapping and SNP

Christine Tranchant-Dubreuil &Francois Sabot

19th of January, 2021

IRD

TP goal

- Quality control of NGS data
- Learn to manipulate NGS data
- Having a critical look on *Mapping*
- Learn to launch a *Calling* and having a critical look

The data

Diploid Asian Rice, *Oryza sativa*

From Wikimedia

The data

Diploid Asian Rice, *Oryza sativa*

From Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

The data

Diploid Asian Rice, *Oryza sativa*

From Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones

The data

Diploid Asian Rice, *Oryza sativa*

From Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
    - SNP (1-10%),
    - indel (10b-10kb),
    - duplications...

The data

Diploid Asian Rice, *Oryza sativa*

From Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
    - SNP (1-10%),
    - indel (10b-10kb),
    - duplications...
4. Generate short & long reads for each clone...

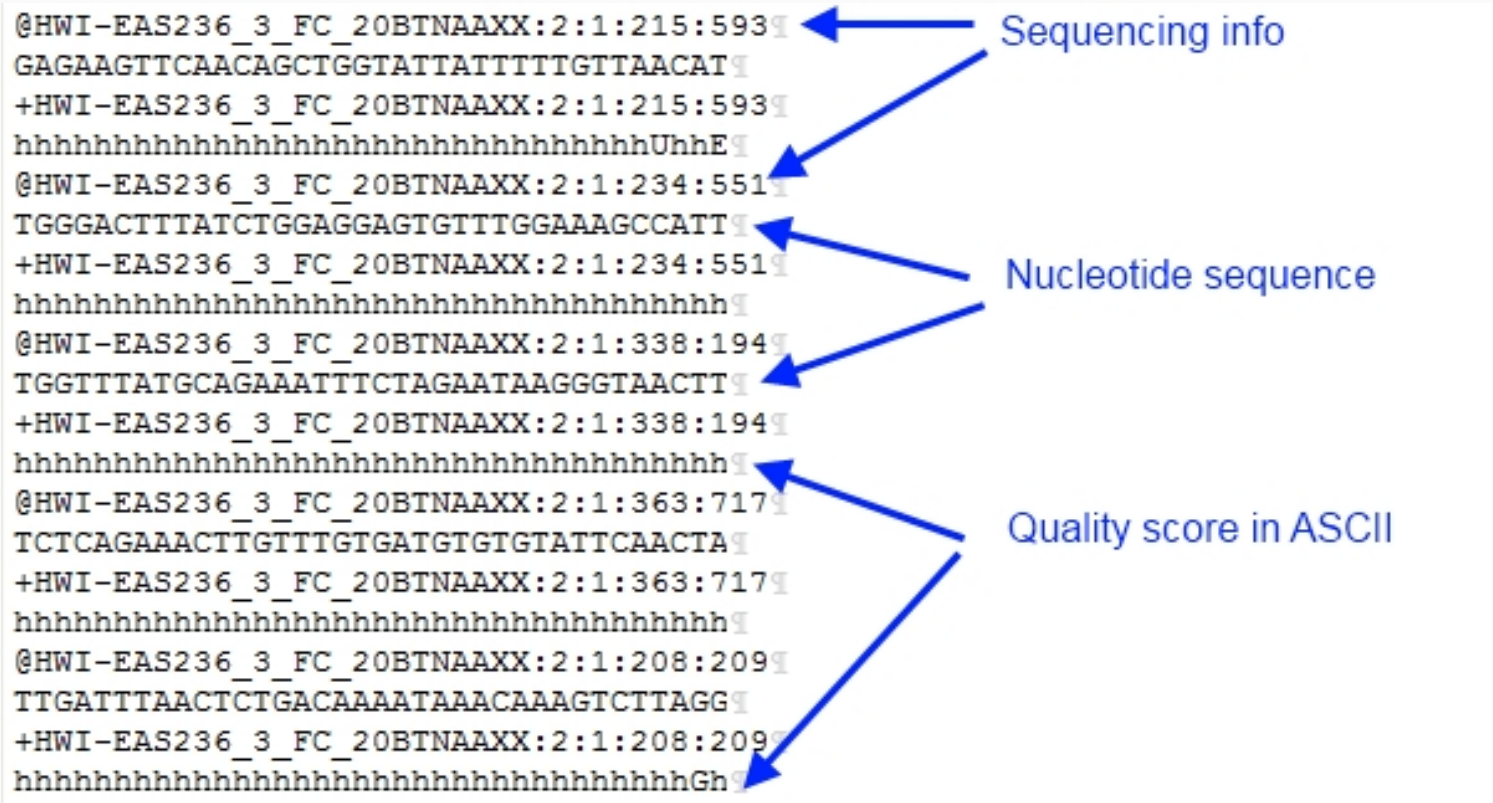
The data

Diploid Asian Rice, *Oryza sativa*

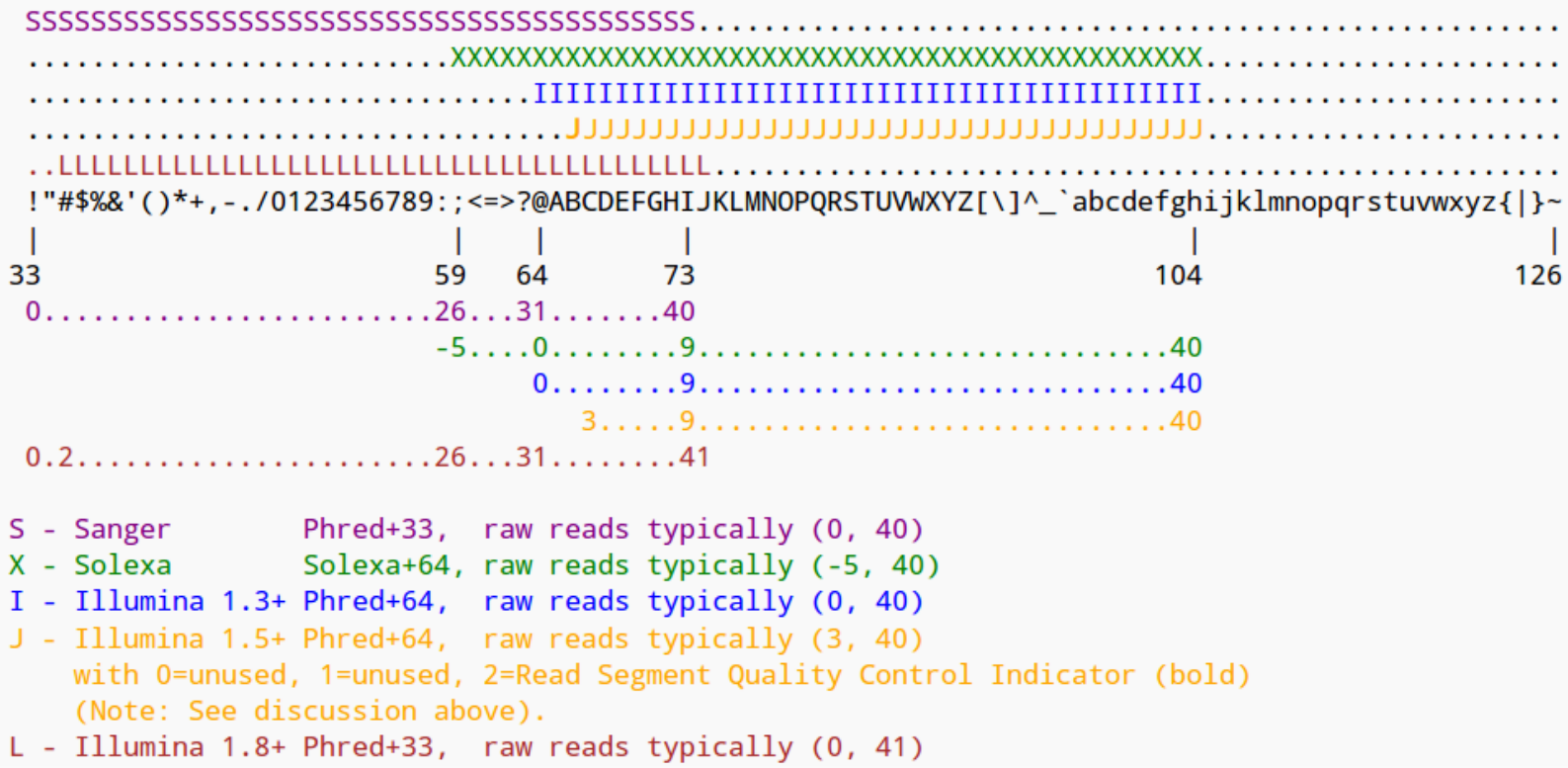
From Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
   - SNP (1-10%),
   - indel (10b-10kb),
   - duplications...
4. Generate short & long reads for each clone...
5. ~~Torturing students with these data~~

The FASTQ Format



The QPHRED Scale



List of analyses

"Classic" launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2, ...

List of analyses

"Classic" launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2, ...
2. *Cleaning mapping*: samtools, picard-tools,...

List of analyses

"Classic" launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2, ...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK, picard-tools,...
   OPTIONAL!

List of analyses

"Classic" launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2, ...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK, picard-tools,...
   OPTIONAL!
4. *SNP calling and Cleaning*: GATK,...

List of analyses

"Classic" launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2, ...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK, picard-tools,...
   OPTIONAL!
4. *SNP calling and Cleaning*: GATK,...

Between 8 and 15 different commands...

Let’s do it by hands…

We will

1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*

Let’s do it by hands...

We will

1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*
2. Look at the SAM file
3.
4.
5.
6.

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)
7	NRNM	Mate Reference NaMe (`=' if same as RNAME)
8	MPOS	1-based leftmost Mate POSition
9	ISIZE	inferred Insert SIZE
10	SEQ	query SEQuence on the reference
11	QUAL	query QUALity (ASCII-33

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

SAM format: FLAG field

numeric	binary	description
1	00000001	template has multiple fragments in sequencing
2	00000010	each fragment properly mapped according to aligner
4	00000100	fragment is unmapped
8	00001000	mate is unmapped
16	00010000	sequence is reverse complemented
32	00100000	sequence of mate is reversed
64	01000000	is first fragment in template
128	10000000	is second fragment in template

From J. Aerts, in SlideShare

Let’s do it by hands…

We will

1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it

Let’s do it by hands…

We will

1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping

Let’s do it by hands...

We will

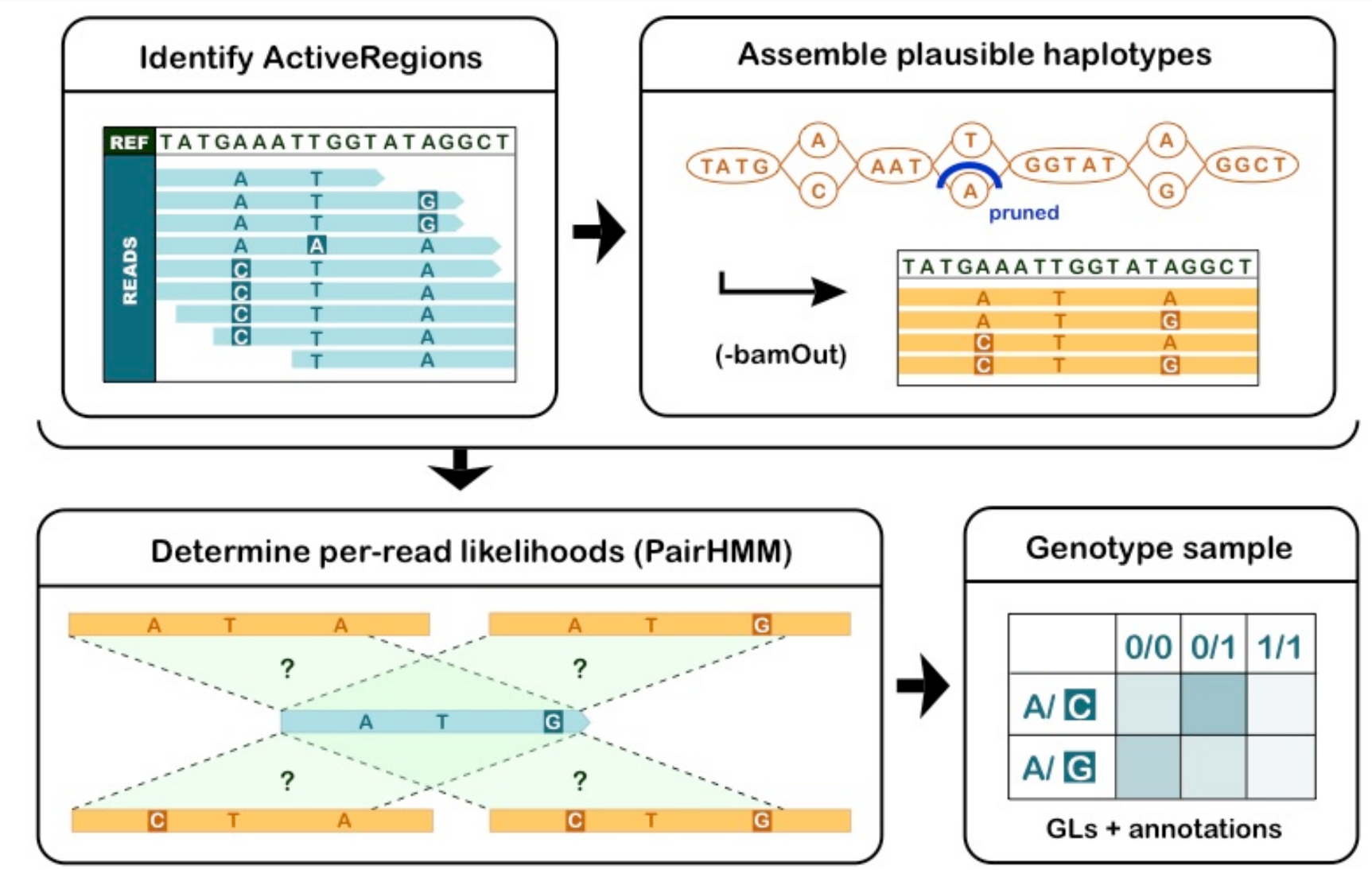
1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping
5. Mark the duplicates

Let’s do it by hands…

We will

1. Map the data of Clone 1 on *`reference.fasta`* using *`bwa`*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping
5. Mark the duplicates
6. Call SNP on this individual

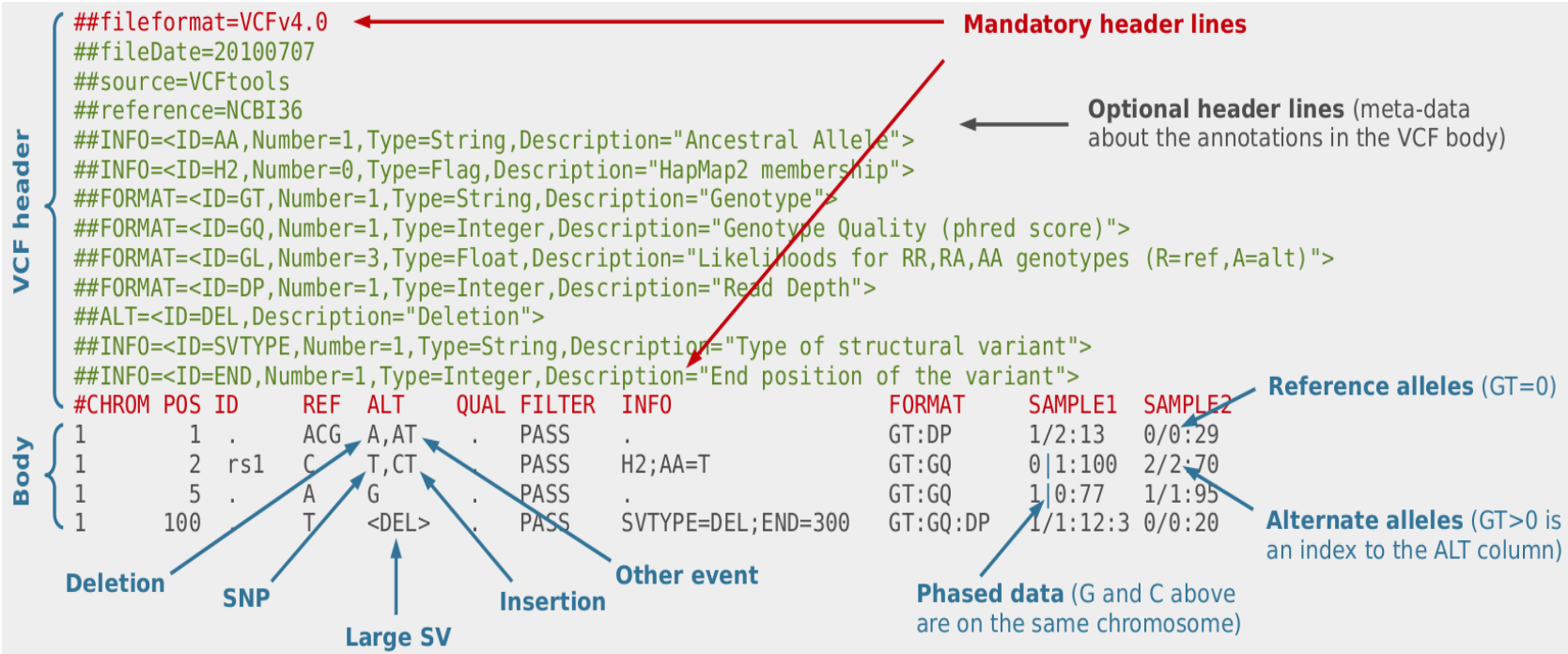
The HaplotypeCaller



From UniBe.ch Training

VCF file format for Variant Call Format

The Variant Call Format (VCF) used in bioinformatics for storing gene sequence variations



See you tomorrow…

A DEMAIN!!

Observing mapping

- Download Tablet (use Google and Tablet+NGS)

Observing mapping

- Download Tablet (use Google and Tablet+NGS)
- Transfer the BAM from **Clone 10** and the reference from the machine to your local computer (use scp or direct download from the browser)

Observing mapping

- Download Tablet (use Google and Tablet+NGS)
- Transfer the BAM from **Clone 10** and the reference from the machine to your local computer (use scp or direct download from the browser)
- Open Tablet, load an assembly

Observing mapping

- Download Tablet (use Google and Tablet+NGS)
- Transfer the BAM from **Clone 10** and the reference from the machine to your local computer (use scp or direct download from the browser)
- Open Tablet, load an assembly
- Look at the mapping and try to find SNPs

Basics filtering

- Using GATK Variant Filtration, a flag per filter

Basics filtering

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
  DP<10 or DP>20000

Basics filtering

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
  DP<10 or DP>20000
- MQ0 filter:
  MQ0 < 4 or MQ0 < 0.1 DP

Basics filtering

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
  DP<10 or DP>20000
- MQ0 filter:
  MQ0 < 4 or MQ0 < 0.1 DP
- QUAL filter:
  QUAL < 200

Basics filtering

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
  DP<10 or DP>20000
- MQ0 filter:
  MQ0 < 4 or MQ0 < 0.1 DP
- QUAL filter:
  QUAL < 200
- SNPcluster filter:
  more than 3 SNP per 10b

- *sNMF*: tool to estimate ancestry coefficients

- *sNMF*: tool to estimate ancestry coefficients
- Developed by E. Frichot and O. Francois, TIMC-IMAG

- *sNMF*: tool to estimate ancestry coefficients
- Developed by E. Frichot and O. Francois, TIMC-IMAG
- R and Command-line version

Basics Population Genomics using *sNMF*

- *sNMF*: tool to estimate ancestry coefficients
- Developed by E. Frichot and O. Francois, TIMC-IMAG
- R and Command-line version
- Much faster than ADMIXTURE or STRUCTURE, as efficient

Be “lazy” but clever

Problems with manual launches

- Long
- Fastidious
- Error prone
- Tracability and reproducibilty not ensured

Be “lazy” but clever

Problems with manual launches

- Long
- Fastidious
- Error prone
- Tracability and reproducibilty not ensured

Solution $\Longrightarrow$ Workflow Manager (or scripts...): SnakeMake, TOGGLe, NextFlow

Thanks for your attention