Genotoul Bioinfo

Aligning SGS reads
Calling SNP

Philippe Bardou
Cédric Cabau

Organisation

25 nov. (9h30-12h20)

- Sequence quality
- Read mapping

25 nov. (13h30-16h20)

- SAM format
- Visualisation

26 nov. (8h00-12h20)

- Variant calling (VCF)
- Variant filtering/annotation

26 nov. (13h30-17h30)

- TA

27 nov. (8h00-12h20)

- TA

1 dec. (8h00-10h50)

- TA

1 dec. (11h00-12h20)

- Retours TA
- Nextflow / Sarek

Organisation

All the course material (slides, input and output data, practical exercises, correction) is available online:

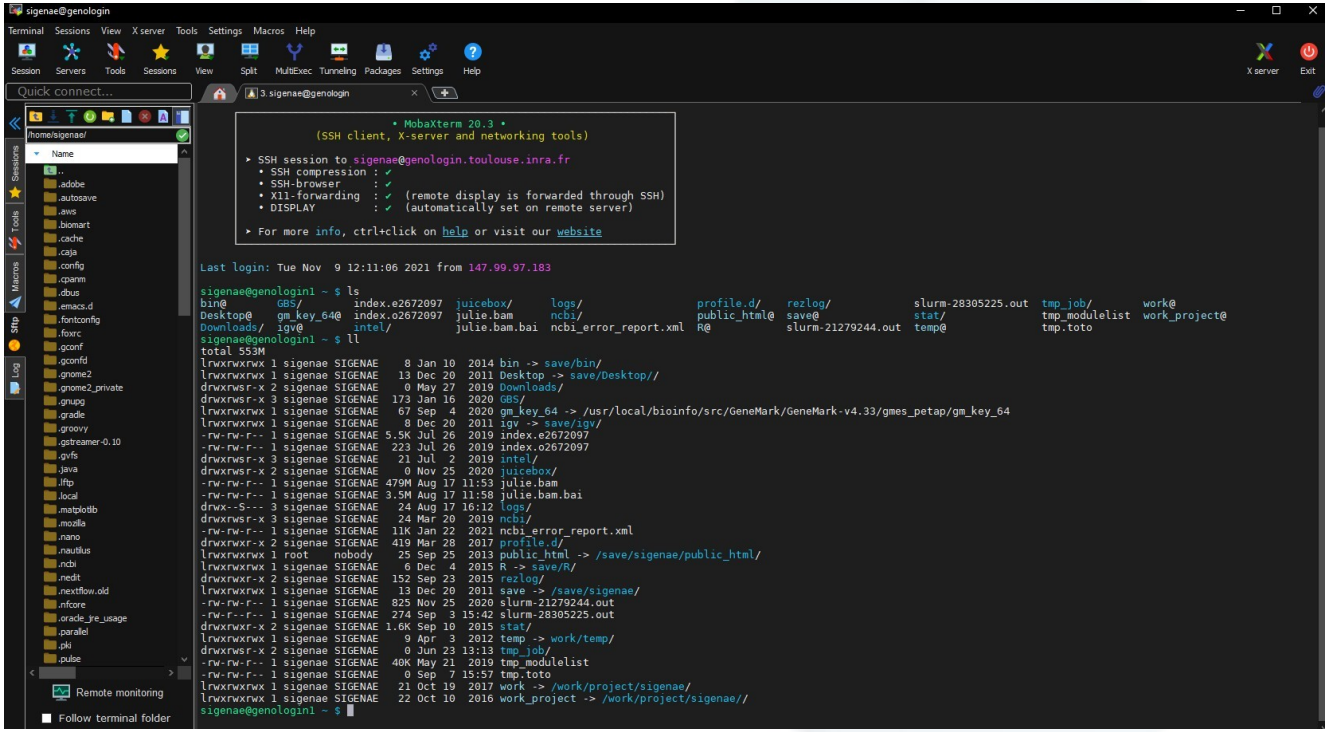
- https://web-genobioinfo.toulouse.inrae.fr/~formation/16_SGS-SNP/

What are you going to learn?

- To extract reads and reference genome from the NCBI
- To verify the read quality
- To align the reads on the reference genome
- To improve alignment and to recalibrate SGS data
- To call variants (SNP and INDEL)
- To visualise the alignments and variations
- To annotate variants
- To filter variants

What you should already know?

- How to connect to a remote unix server (mobaXterm)?
- What a unix command looks like?
- How to move around the unix environment?
- How to submit jobs?
- How to edit a file?



Genomic variants - who are you?

- Genomic variants are differences in the DNA sequence among individuals.
- Range from single nucleotide changes (SNPs) to large structural variations.
- Basis of genetic diversity and can influence traits such as physical characteristics, disease susceptibility, and response to drugs.

Types of genomic variants

- Single Nucleotide Polymorphisms (SNPs)
  - the most common type of genetic variation, involving a change in a single base pair.
- Insertions and Deletions (Indels)
  - variations involving the addition or removal of base pairs in the DNA sequence.
- Structural Variants
  - large-scale alterations in the DNA, including duplications, inversions, and translocations.

Why study genomic variants? (human context)

- Understanding disease
  - many diseases are influenced by genetic variants, including cancer, heart disease, and neurological disorders.
- Personalized medicine
  - knowledge of a patient's genomic variants can guide treatment decisions and drug dosing.
- Evolution and population genetics
  - studying genomic variants can provide insights into human evolution and population history.

Why study genomic variants? (agronomic context)

- Genomic prediction
  - identification of genomic variants causing variation in quantitative traits.
- Population genomics
  - identification of genomic regions linked to traits or evolutionary processes to study evolutionary relationships of a population, including population size, structure, demographic history and phylogenetic relationships.
- Adaptive genetic variations
  - analyze genetic variations that underlie adaptive changes.
- Crop and livestock health
  - identification of variations responsible for disease susceptibility.
  - Identify variations that modulate the response to prophylactic treatments

Methods for detecting genomic variants

- Sequencing
  - next-generation sequencing technologies can identify variants across the entire genome.
- Genotyping
  - detect specific variants of interest that can lead to major changes in phenotype.
- Bioinformatics
  - computational tools are used to analyze sequencing data and predict the effects of variants.

The 1000 Genomes Project

- International effort to establish the most detailed catalogue of human genetic variation.
- The project sequenced the genomes of over 2500 individuals from 26 populations, identifying over 88 million variants.
- The data from the 1000 Genomes Project is a valuable resource for researchers studying human genetics and disease.

The pieces of software

- Fastqc : quality control
- BWA : alignment
- Samtools & Picard-tools : manipulation of SAM/BAM files
- IGV : visualisation
- GATK : preprocess and variant calling
- SNPeff / SNPsift : annotate/filter variants

gatk

Manual Reference Pages - bwa (1)

NAME

bwa - Burrows-Wheeler Alignment Tool

CONTENTS

igv

Babraham Institute Babraham Bioinformatics FastQC

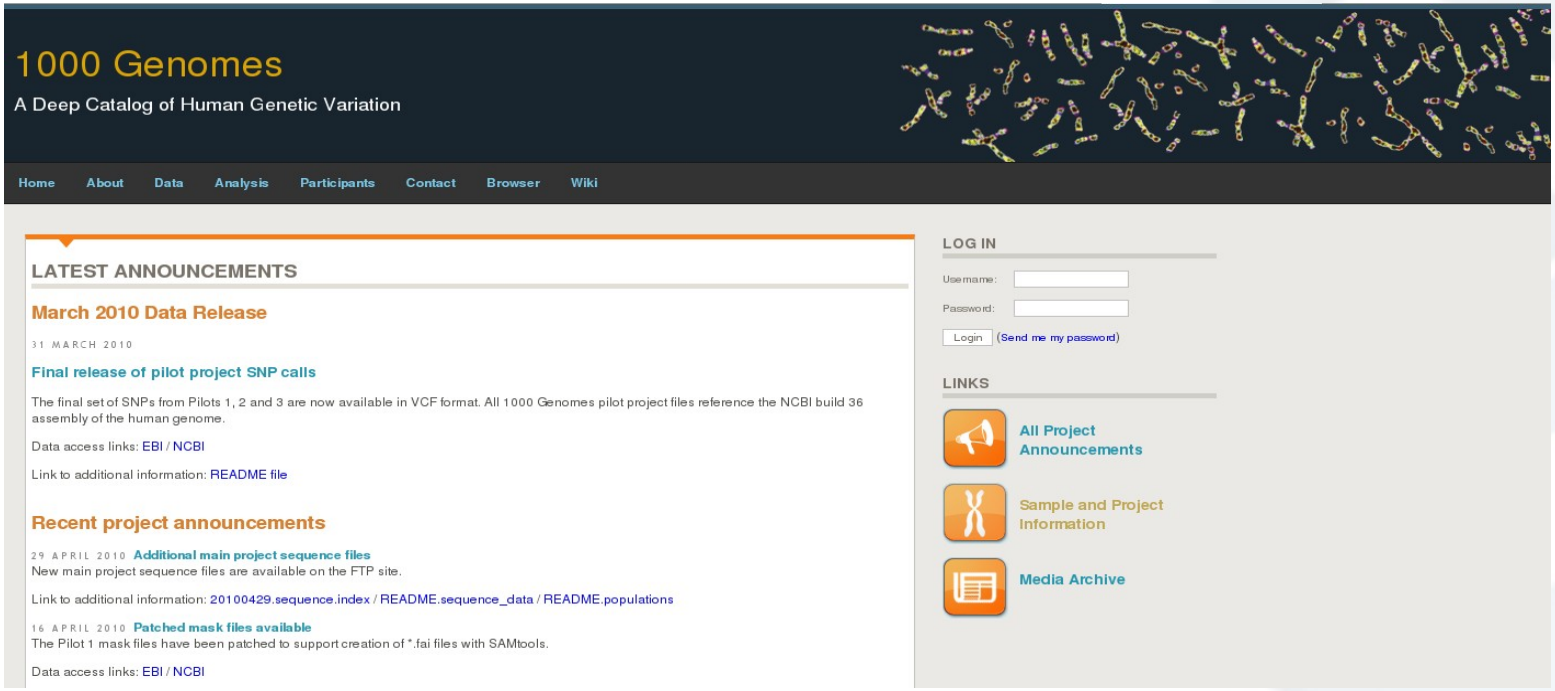
SnpEff & SnpSift

SAMtools

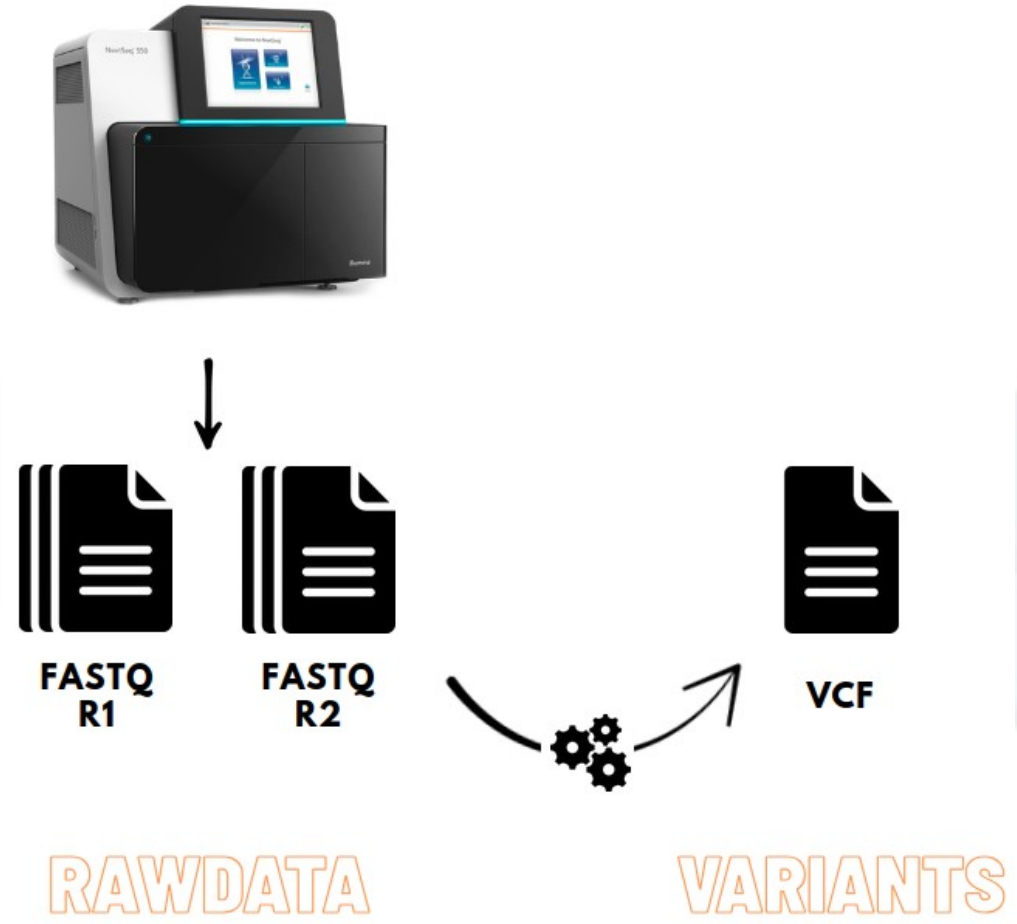
Picard

The 1000 genomes project

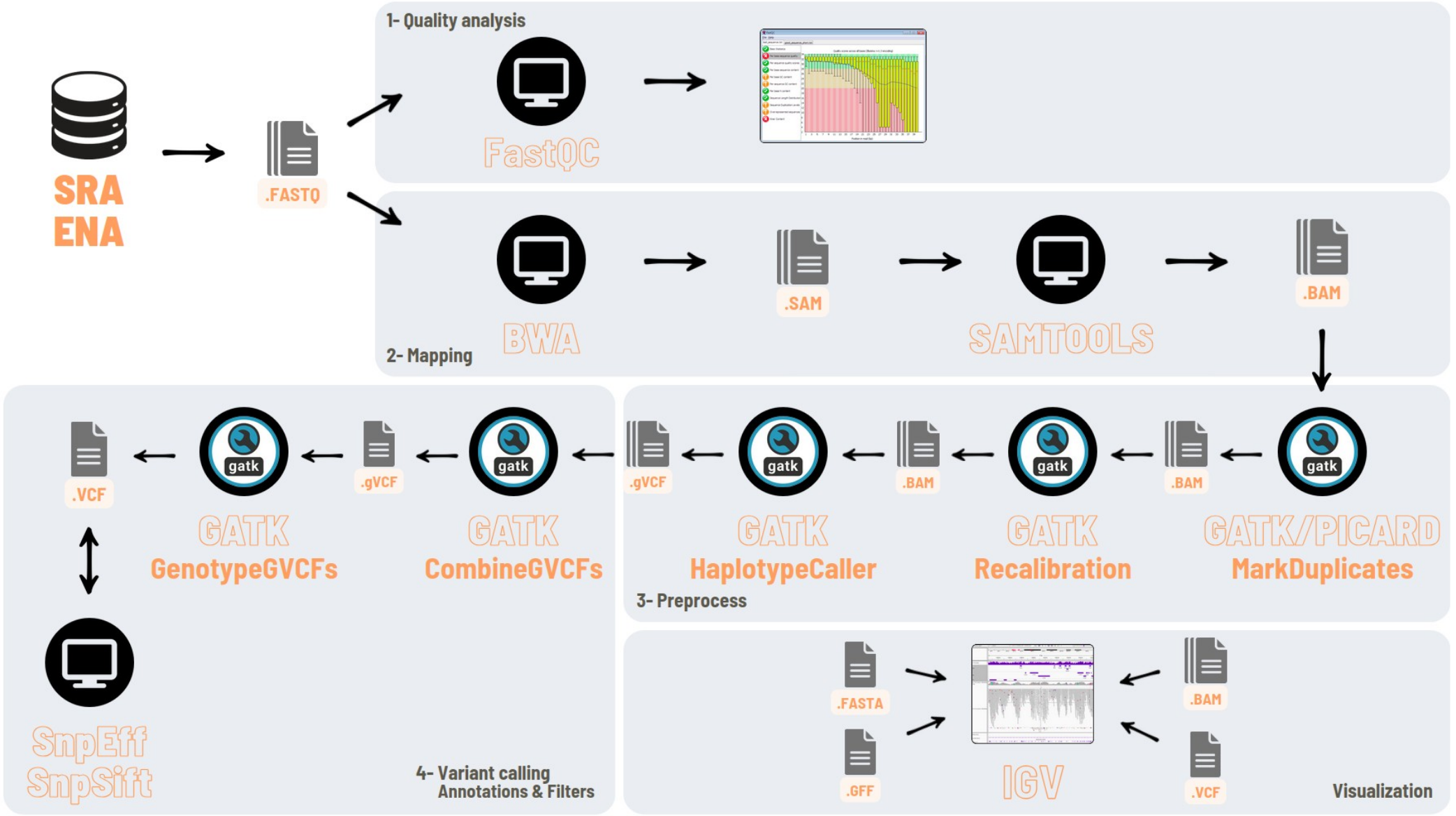
- Joint project NCBI / EBI
- Common data formats :
  - FASTQ (Raw data)
  - SAM (Sequence Alignment/Map)
  - VCF (Variant Call Format)



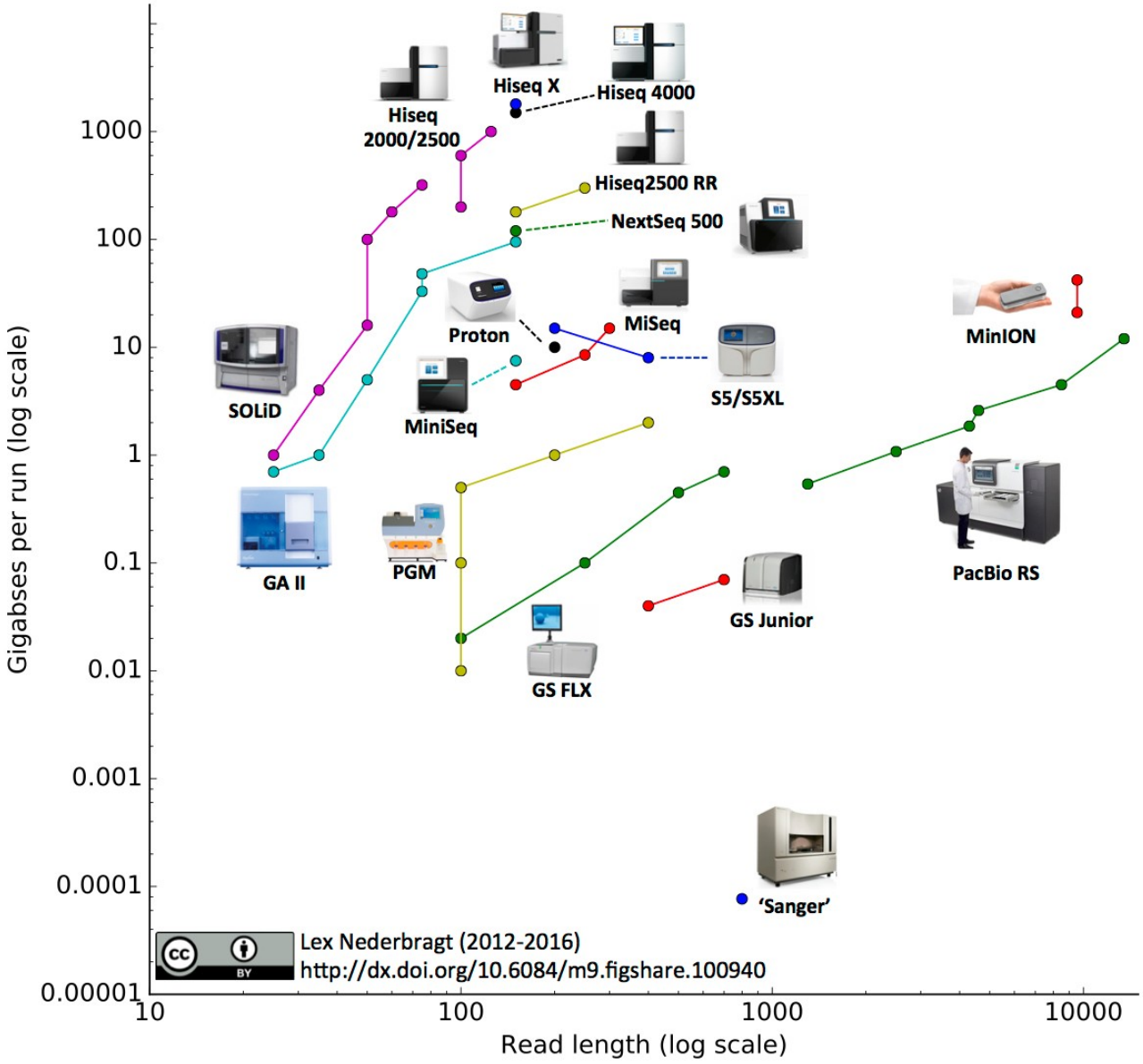
Overview



Overview

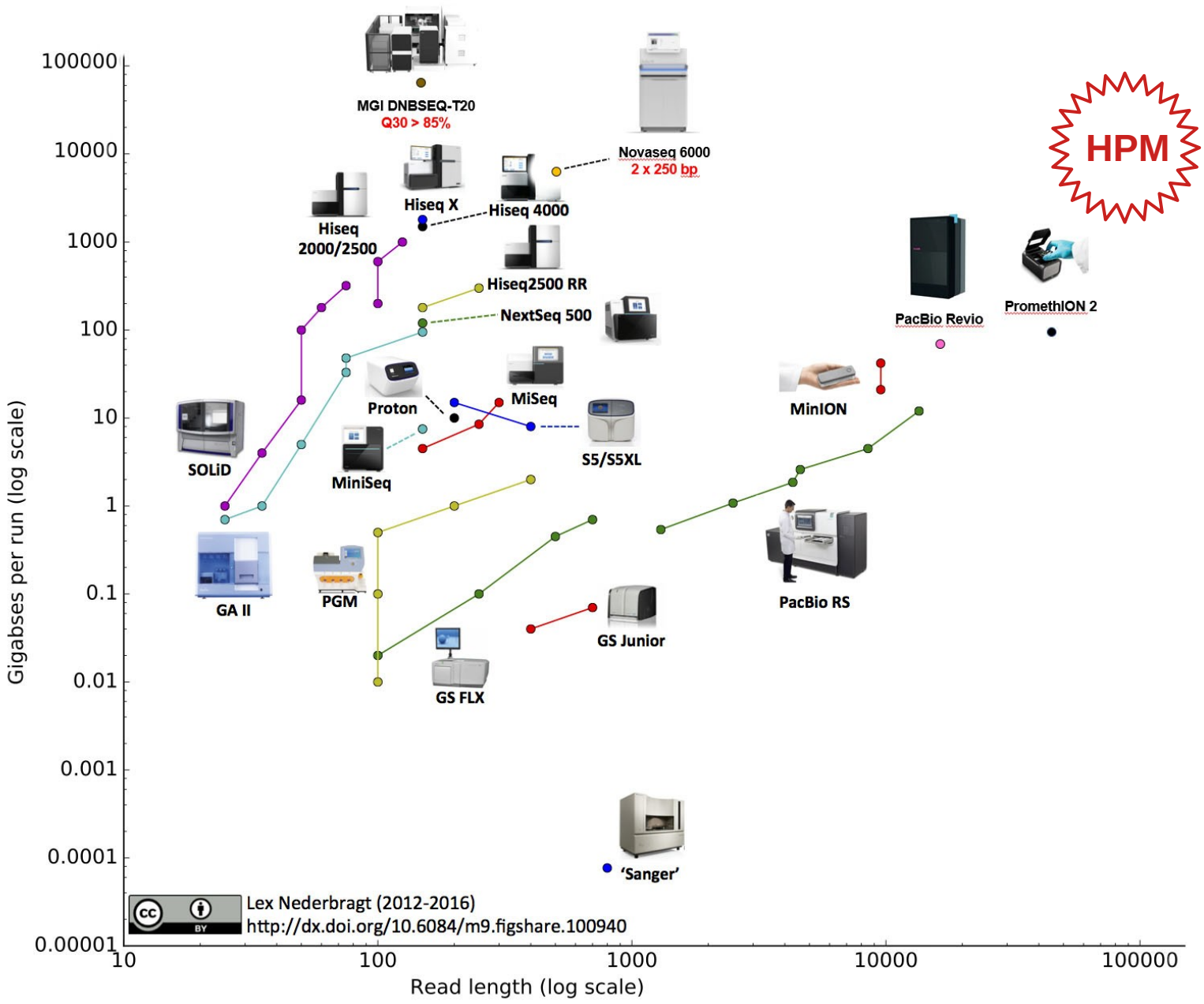

1- Quality analysis
SRA
ENA
.FASTQ
FastQC
2- Mapping
BWA
.SAM
SAMTOOLS
.BAM
gatk
GATK/PICARD
MarkDuplicates
GATK
Recalibration
GATK
HaplotypeCaller
3- Preprocess
GATK
CombineGVCFs
GATK
GenotypeGVCFs
.VCF
.gVCF
SnpEff
SnpSift
4- Variant calling
Annotations & Filters
.FASTA
.GFF
IGV
Visualization

SGS platforms



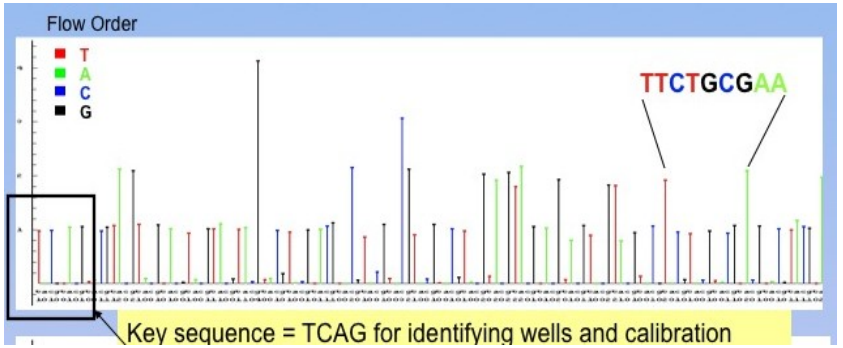
https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/#more-790

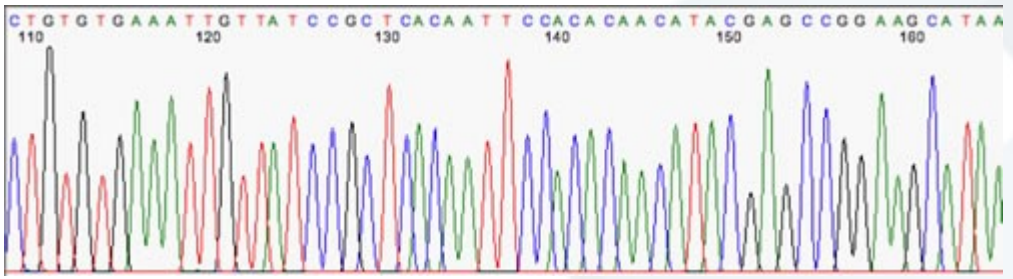
SGS platforms



Sequencing bias

Technology	Main Biases	Dominant Errors	Typical Error Rate
Sanger	Extreme GC, secondary structures, drop at end of read	Substitutions	0.001–0.01%
454 Roche	Homopolymers highly problematic	Indels ↑	~1%
Ion Torrent	Homopolymer saturation	Indels ↑	1–2%
Illumina	GC bias, PCR bias, end of read	Substitutions	0.05–0.2%
MGI / DNBseq	Moderate GC bias	Substitutions	0.05–0.15%
PacBio CLR	Few biases but depends on long DNA	Indels ↑	10–15%
PacBio HiFi	Very low bias	Substitutions	0.03–0.1%
ONT R9.4	Homopolymers, extreme GC, motifs	Indels (deletions)	4–8%
ONT R10.4.1 Q20+	Improved homopolymer handling	Substitutions ≈ Indels	~1%
ONT Duplex	Strongly reduced biases	Substitutions & indels low	0.1–0.3%





SGS platforms

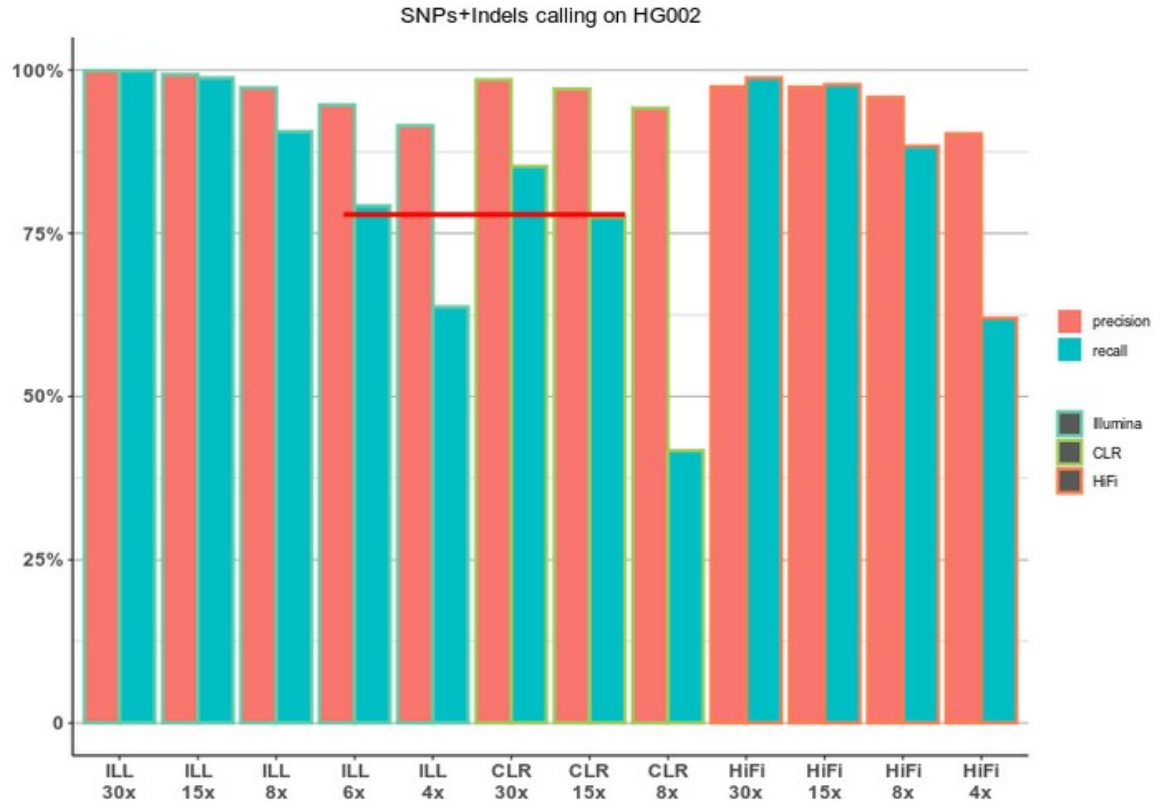
SeqOccIn

Human SNP calling

Giab as reference set

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$



- Illumina donne les meilleurs résultats
- HiFi est équivalent
- CLR gardent des hautes valeurs de précision
- CLR a des valeurs de sensibilité “correctes” sachant que Longshot ne détectent pas les Indels
- 15x en CLR a une sensibilité équivalente à 6x en Illumina

Pour mémoire, taux d’erreur

- Illumina : 0.2 %
- HiFi : 1 %
- CLR : 15%

4

Genotoul GeT Genotoul Bioinfo

INRAE

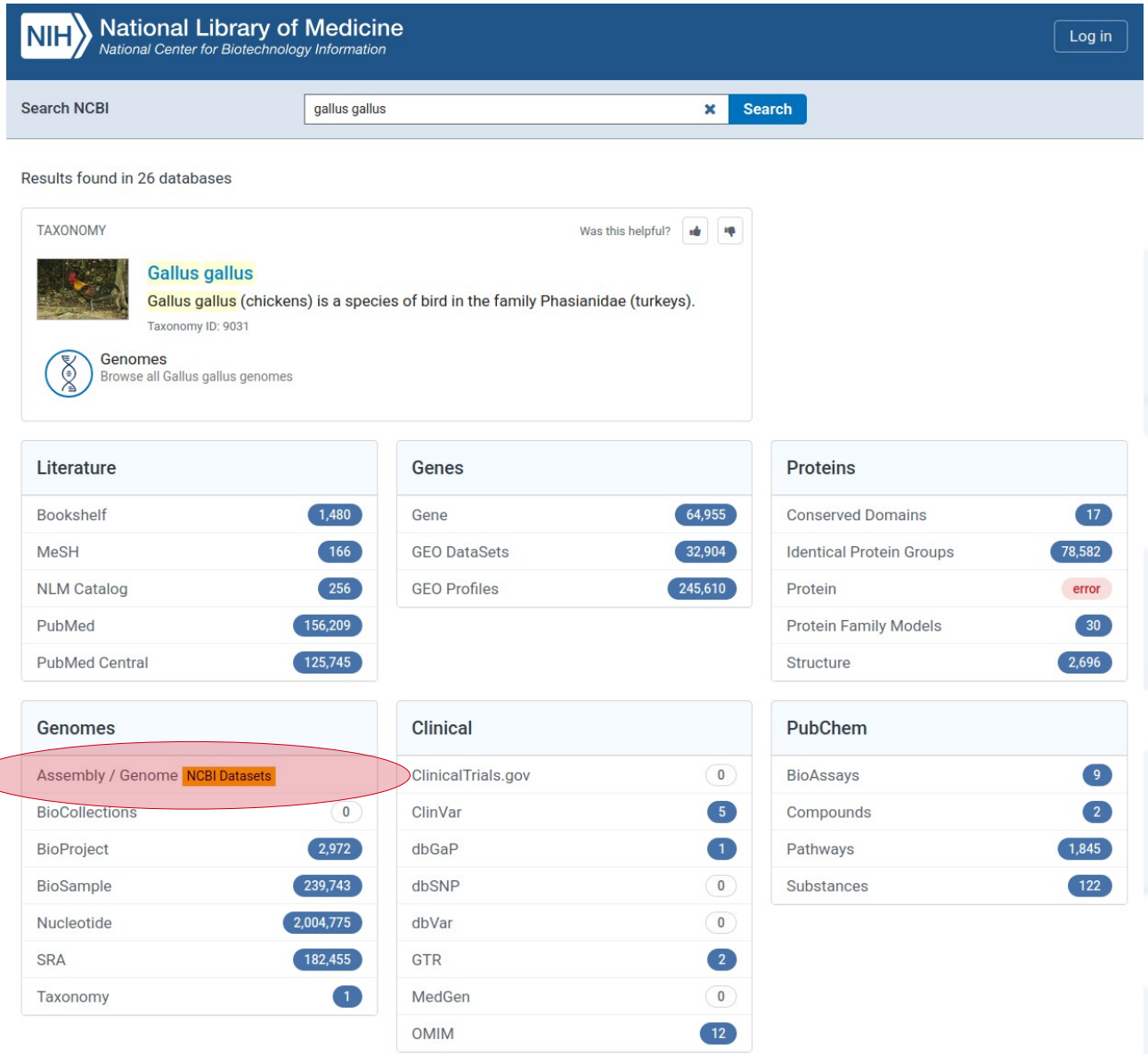
Extract from Arnaud Di-Franco - SeqOccin - 2020/12
Détection de variants Efficacité de la détection à basse couverture en longue lecture

What data will we use?

- The needed data :
  - A reference sequence :
    - Genome
    - Parts of the genome
    - Transcriptome
  - Short reads

Where to get reference genome?

Assemble your own OR use a public assembly (NCBI / EBI)

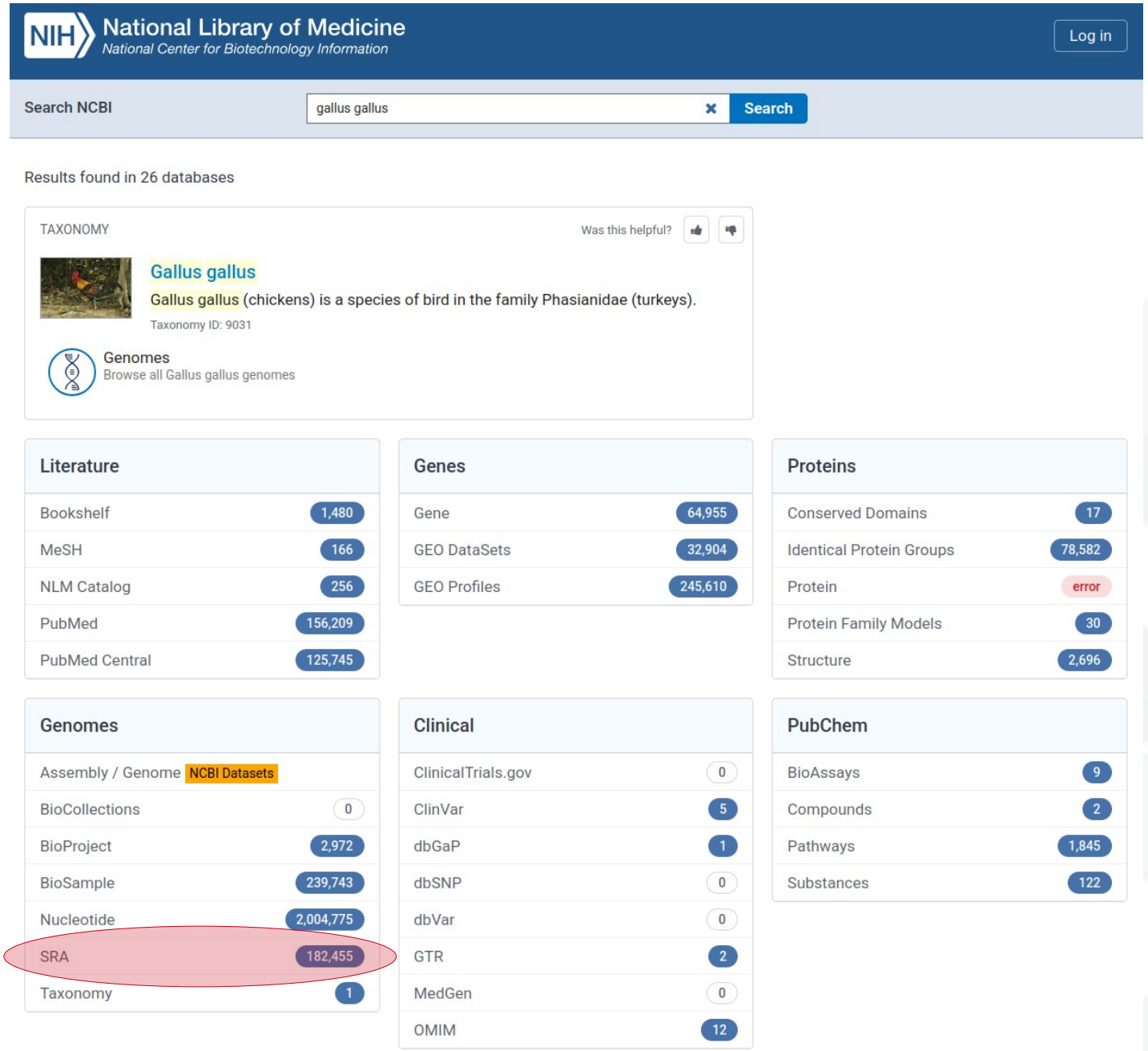


Where to get short reads?

- Produce your own sequences :
  - CNS
  - Local platform
  - Private company

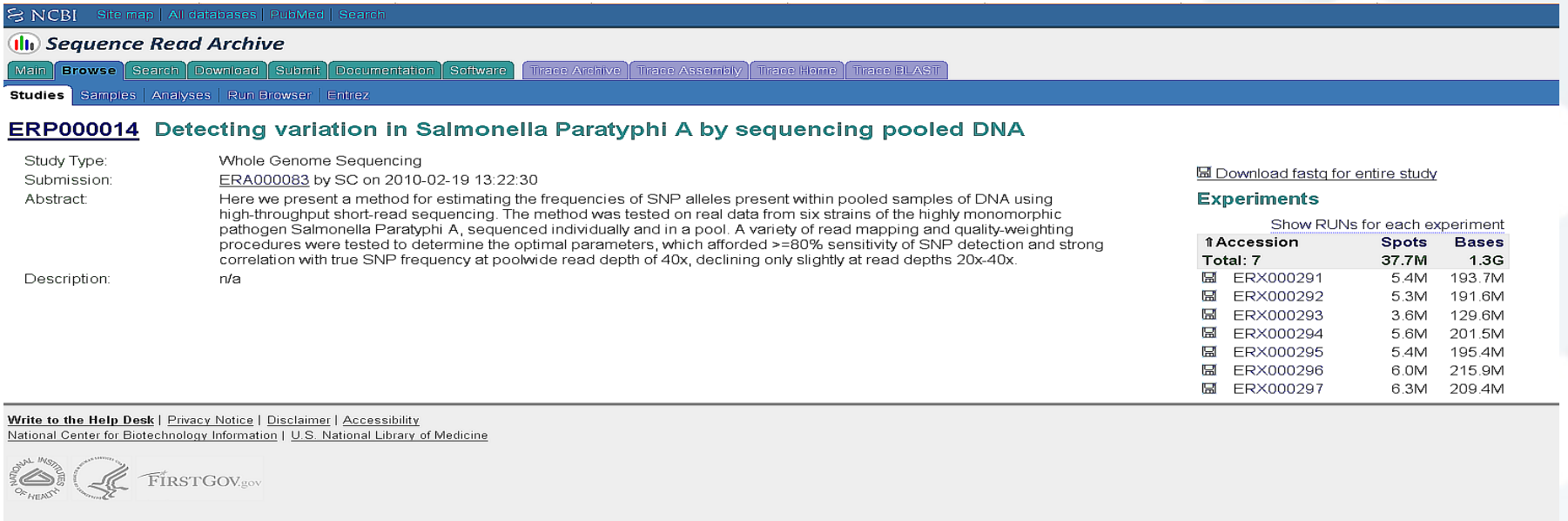
- Use public data :
  - SRA : NCBI Sequence Read Archive
  - ENA : EMBL/EBI European Nucleotide Archive

NCBI SRA?


National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
gallus gallus
Search
Results found in 26 databases
TAXONOMY
Was this helpful?
Gallus gallus
Gallus gallus (chickens) is a species of bird in the family Phasianidae (turkeys).
Taxonomy ID: 9031
Genomes
Browse all Gallus gallus genomes
Literature
Bookshelf 1,480
MeSH 166
NLM Catalog 256
PubMed 156,209
PubMed Central 125,745
Genes
Gene 64,955
GEO DataSets 32,904
GEO Profiles 245,610
Proteins
Conserved Domains 17
Identical Protein Groups 78,582
Protein error
Protein Family Models 30
Structure 2,696
Genomes
Assembly / Genome NCBI Datasets
BioCollections 0
BioProject 2,972
BioSample 239,743
Nucleotide 2,004,775
SRA 182,455
Taxonomy 1
Clinical
ClinicalTrials.gov 0
ClinVar 5
dbGaP 1
dbSNP 0
dbVar 0
GTR 2
MedGen 0
OMIM 12
PubChem
BioAssays 9
Compounds 2
Pathways 1,845
Substances 122

Meta data

- Meta data structure :
  - Experiment
  - Sample
  - Study
  - Run
  - Data file



What is a fastq file?

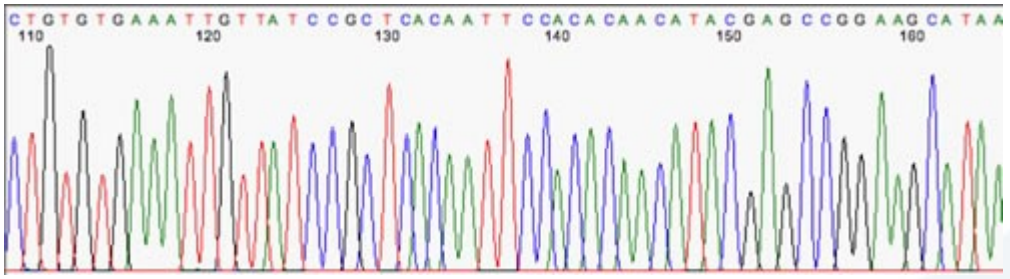
FASTQ format stores sequences and Phred qualities in a single file. It is concise and compact. FASTQ is first widely used in the Sanger Institute and therefore we usually take the Sanger specification and the standard FASTQ format, or simply FASTQ format. Although Solexa/Illumina read file looks pretty much like FASTQ, they are different in that the qualities are scaled differently. In the quality string, if you can see a character with its ASCII code higher than 90, probably your file is in the Solexa/Illumina format.

Example

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

Sequence quality

- Phred : base calling



What is Phred Quality?

Traditionally, Phred quality is defined on base calls. Each base call is an estimate of the true nucleotide. It is a random variable and can be wrong. The probability that a base call is wrong is called error probability.

Explanation about the quality values :

source http://maq.sourceforge.net/qual.shtml

Sequence quality

Phred quality scores are logarithmically linked to error probabilities

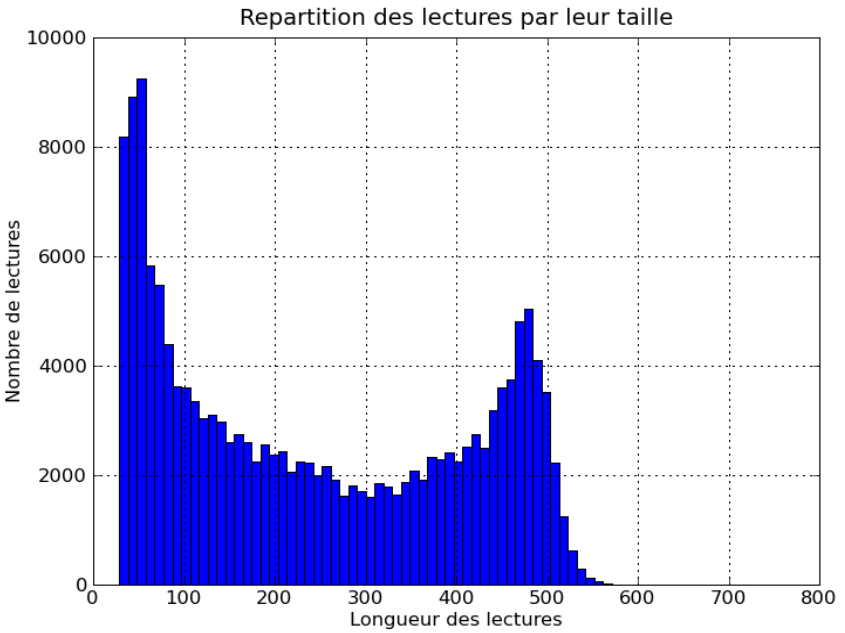
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Which reads should I keep?

- All
- Some : what criteria and threshold should I use
  - Composition (number of Ns, complexity, ...)
  - Quality
  - Alignment based criteria
- Should I trim the reads using :
  - Composition
  - Quality

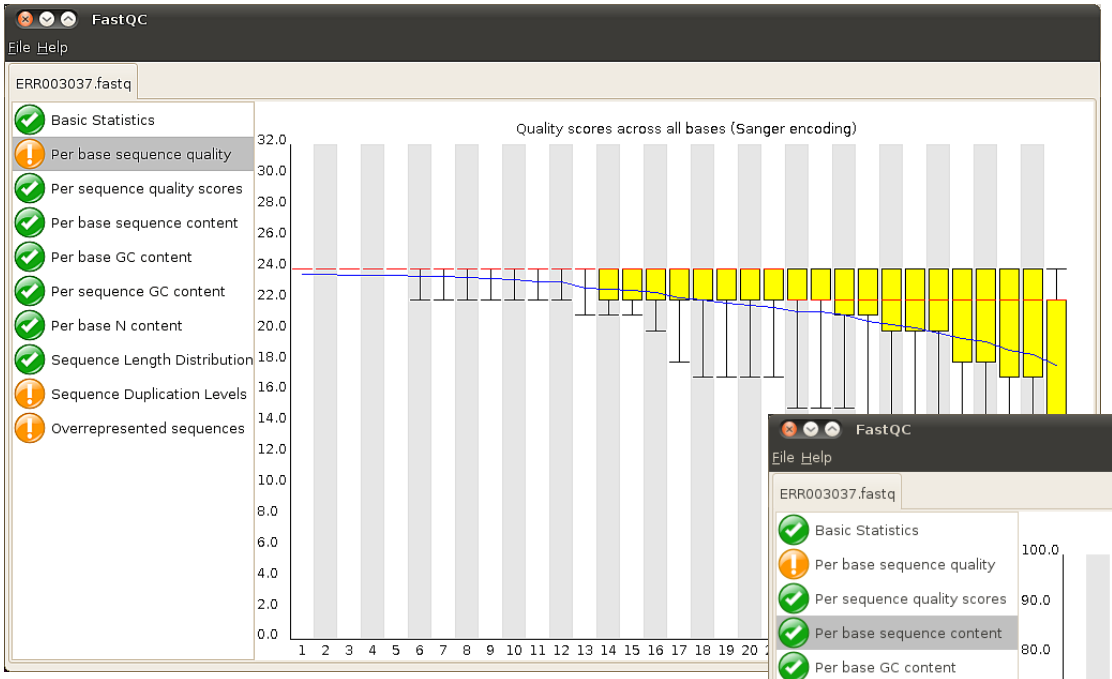
Basic reads statistics

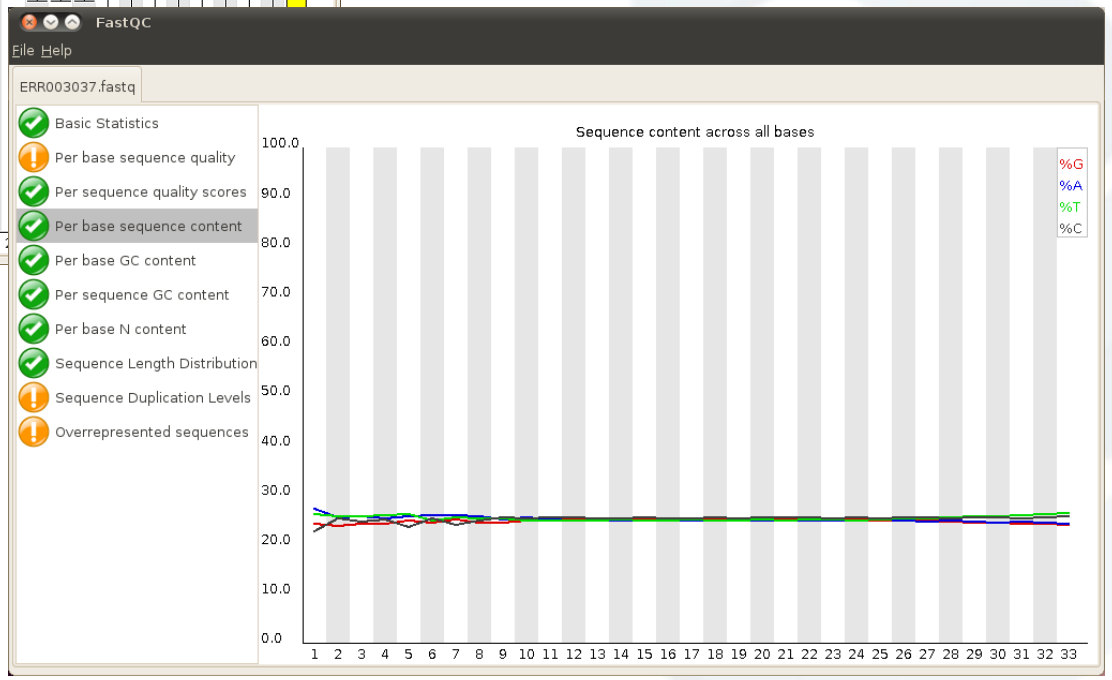
- Number of reads
- Length histogram
- Number of Ns in the reads
- Reads quality
- Reads redundancy
- Reads complexity



Sequence quality analysis

- FastQC :





Bash basics

Some basics to help you solve exercises.

Looping Constructs

The for command syntax

```
for name [ [in words ...] ; ] do commands; done
```

Example

```
for i in *.fastq.gz; do echo $i; done
```

Shell Parameter Expansion

https://www.gnu.org/software/bash/manual/html_node/Shell-Parameter-Expansion.html

The ‘$’ character introduces parameter expansion.

The basic form

```
$ string=abcd
$ echo $string
abcd
$ echo ${string}
abcd
```

`${parameter}` substitutes the value of parameter.

Shell Parameter Expansion

${parameter%pattern}

${parameter%%pattern}

If pattern matches a trailing portion of the expanded value, delete:

- the shortest matching pattern, the '%' case
- the longest matching pattern, the '%%' case

Example

```
$ file=myFastFile.fastq.gz
$ echo ${file%.*}
myFastFile.fastq
$ echo ${file%%.*}
myFastFile
```

Shell Parameter Expansion

${parameter#pattern}

${parameter##pattern}

If pattern matches the beginning of the expanded value, delete:

- the shortest matching pattern, the '#' case
- the longest matching pattern, the '##' case

Example

```
$ file=myFastFile.fastq.gz
$ echo ${file#*.}
fastq.gz
$ echo ${file##*.}
gz
```

Exercises / set 1

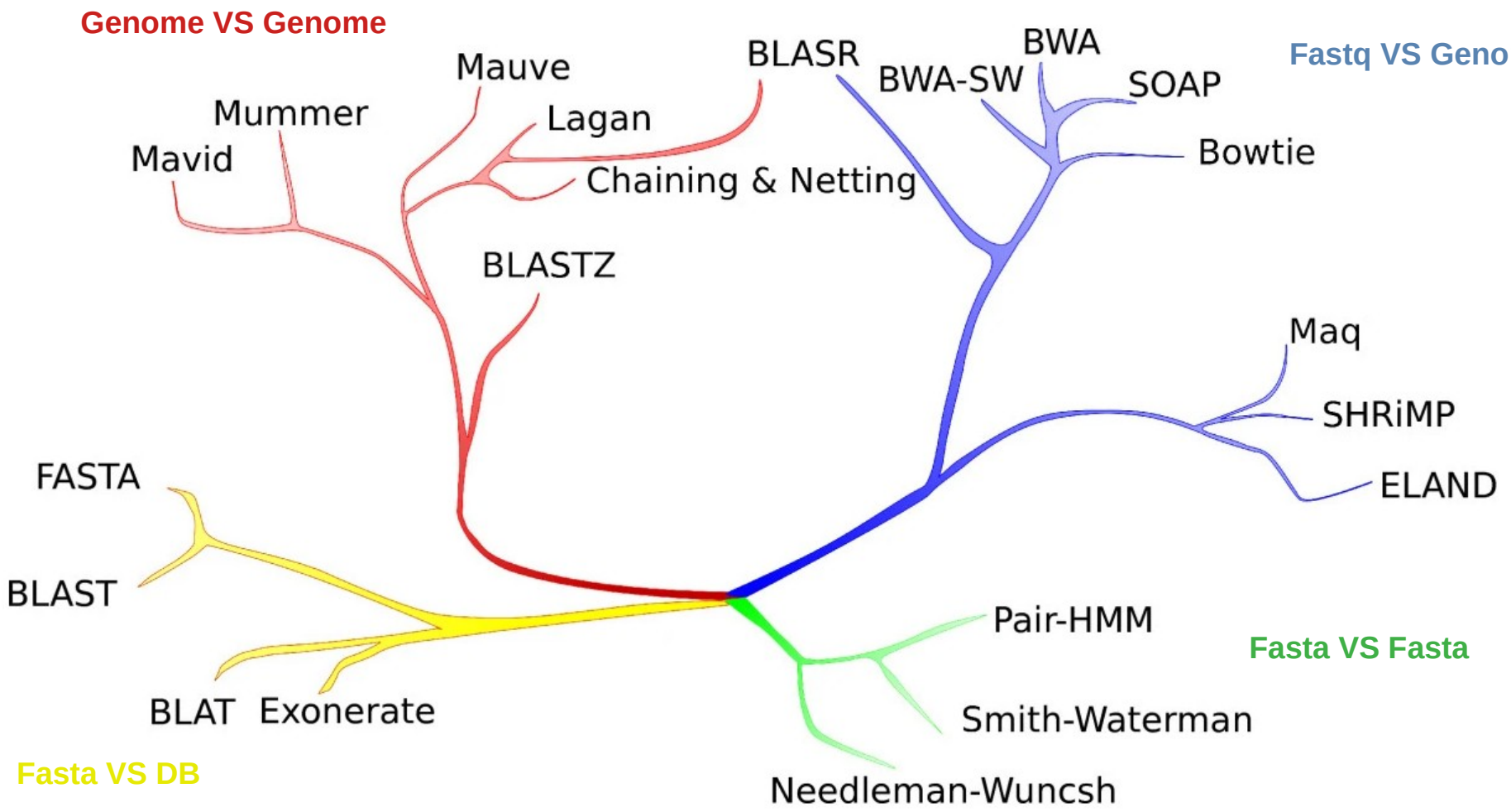
https://bios4biol.pages.mia.inra.fr/training-aln-variant-calling/

Reads alignment

The different software generations :

- Smith-Waterman / Needleman-Wunch (1970)
- BLAST (1990)
- MAQ (2008)
- BWA (2009)
- Minimap2 (2018)

Reads alignment



An illustration of relationships between alignment methods. The applications / corresponding computational restrictions shown are (green) short pairwise alignment / detailed edit model; (yellow) database search / divergent homology detection; (red) whole genome alignment / alignment of long sequences with structural rearrangements; and (blue) short read mapping / rapid alignment of massive numbers of short sequences. Although solely illustrative, methods with more similar data structures or algorithmic approaches are on closer branches. The BLASR method combines data structures from short read alignment with optimization methods from whole genome alignment.

Chaisson, M.J., Tesler, G. BMC Bioinformatics 13, 238 (2012). https://doi.org/10.1186/1471-2105-13-238

Reads alignment

Thématique	Exemples	Année (approx.)
Séquence vs séquence	Needleman–Wunsch, Smith–Waterman	1970 (NW), 1981 (SW)
Séquence vs base	FASTA, BLAST, HMMER	1985 (FASTA), 1990 (BLAST), 1998 (HMMER)
Génome vs génome	MUMmer, LASTZ, Mauve, Cactus	1999 (MUMmer), ~2004 (LASTZ), 2004 (Mauve), 2010s (Cactus)
Reads vs génome	Bowtie, BWA, STAR, HISAT2, Minimap2	2009 (Bowtie/BWA), 2013 (STAR), 2015 (HISAT2), 2018 (Minimap2)
Alignement multiple	Clustal, MAFFT, MUSCLE	1987 (Clustal), 2002 (MAFFT), 2004 (MUSCLE)
Pangénomes	VG, Giraffe	~2015 (VG), 2021 (Giraffe)

Thématique	Outils de référence	Particularités
DNA	BWA, Bowtie2, Minimap2, Novoalign	Reads courts/longs, génome de référence
RNA	STAR, HISAT2, Salmon, Kallisto	Épissage, quantification
miRNA	Bowtie, miRDeep2, ShortStack	Reads très courts, annotation
Bisulfite	Bismark, BSMAP, WALT	Méthylation, épigénétique
Pangénomes	VG, Giraffe, GraphAligner, Minigraph	Alignement sur graphes

Reads alignment

Most popular tools for mapping to a normal genomic reference (DNAseq, ChIP-Seq, sRNAseq, …) :

Scoring of aligners for various sequencing parameters based on criteria evaluated in this study; + indicates low score, ++ indicates medium score, and +++ indicates high score.

	Execution time				Memory usage				Accuracy				% Prop. paired reads			
Ins. (bp)	350		550		350		550		350		550		350		550	
RL (bp)	100	150	100	150	100	150	100	150	100	150	100	150	100	150	100	150
BWA	+	+	+	+	+	+	+	+	+++	+++	+++	+++	+++	+++	+++	+++
Bowtie2	++	++	++	++	+++	+++	++	++	+++	+++	+++	+++	++	++	+++	+++
HISAT2	+++	+++	+++	+++	+++	+++	+++	+++	++	++	++	++	+	+	+	+

https://dx.doi.org/10.3389%2Ffgene.2018.00035

Popular splice read aligners (RNAseq polyA+/total) :

Table 1: Sensitivity, precision, run times and memory usage of leading spliced aligners.

Program	Sensitivity (%)	Precision (%)	Run time (min)	Memory usage (GB)
HISAT2	97.3	94.8	26.7	4.3
TopHat2	90.6	82.6	1,170	4.3
STAR	96.3	88.3	25	28
Bowtie2	91.1	79.2	13	14

https://doi.org/10.23937/2378-3648/1410048

BWA

- Fast and moderate memory footprint (<4GB)
- SAM output by default
- **Gapped** alignment for both SE and PE reads
- Effective pairing to achieve high alignment accuracy; suboptimal hits considered in pairing.
- Non-unique read is placed randomly with a mapping quality 0
- Limited number of errors (2 for 32bp, 4 for 100 bp, ...)
- The default conguration works for most typical input.
  - Automatically adjust parameters based on read lengths and error rates.
  - Estimate the insert size distribution on the fly

http://bio-bwa.sourceforge.net/

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009
Advance Access publication May 18, 2009
Associate Editor: John Quackenbush

http://bio-bwa.sourceforge.net/bwa.shtml

Manual Reference Pages - bwa (1)

NAME

bwa - Burrows-Wheeler Alignment Tool

CONTENTS

SYNOPSIS

```
bwa index ref.fa

bwa mem ref.fa reads.fq > aln-se.sam

bwa mem ref.fa read1.fq read2.fq > aln-pe.sam

bwa aln ref.fa short_read.fq > aln_sa.sai

bwa samse ref.fa aln_sa.sai short_read.fq > aln-se.sam

bwa sampe ref.fa aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln-pe.sam

bwa bwasw ref.fa long_read.fq > aln.sam
```

BWA – cmd line

- Two steps :
  - Indexation :

```
$ bwa index GENOME.fasta
```

  - Alignment :

```
$ bwa mem \
> -R "@RG\tID:1\tSM:SRR7062654\tPL:illumina\tLB:SRR7062654\tPU:1" \
> -t4 \
> GENOME.fa \
> SRR7062654_R1.fastq.gz \
> SRR7062654_R2.fastq.gz \
> | samtools sort - > SRR7062654.bam
```

- Meaning of the read group fields required (@RG) :
  - `ID` = **Read group identifier**
  - `PU` = **Platform Unit**
  - `SM` = **Sample**
  - `PL` = **Platform/technology used to produce the read**
  - `LB` = **DNA preparation library identifier**

https://gatk.broadinstitute.org/hc/en-us/articles/360035890671-Read-groups

Exercises / set 2

Sequence Alignment/Map (SAM) format

- Data sharing was a major issue with the 1000 genomes
- Capture all of the critical information about NGS data in a single indexed and compressed file
- Generic alignment format
- Supports short and long reads (Illumina - Pacbio - ONT)
- Flexible in style, compact in size, efficient in random access

Website :

https://www.htslib.org/

Paper :

Twelve years of SAMtools and BCFtools

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li

GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

SAM format - Header section

- Header lines start with @ followed by a two-letter TYPE
- Header fields are TAG:VALUE pairs

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
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for
	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

```
@HD     VN:1.0
@SQ     SN:chr20 LN:62435964
@RG     ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG     ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

SAM format

Which informations are stored in a SAM file?

https://web-genobioinfo.toulouse.inrae.fr/~formation/16_SGS-SNP/.formats/sam.html

SAM format - Alignment section

- 11 mandatory fields
- Variable number of optional fields
- Fields are tab delimited

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. **MRNM:** Mate reference name ('=' if same as RNAME)
8. **MPOS:** 1-based leftmost mate position
9. **ISIZE:** Inferred insert size
10. **SEQQuery:** Sequence on the same strand as the reference
11. **QUAL:** Query quality (ASCII-33=Phred base quality)

SAM format - Full example

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

X?	Reserved for end users
NM	Edit distance to the reference
MD	String for mismatching positions
RG	Read group
SA	Supp. alignment
[...]	

A	Printable character
I	Signed 32-bit integer
F	Single-precision float number
Z	Printable string
H	Hex string (high nybble first)

SAM format - Flag field

FLAG: Combination of bitwise FLAGs.[12] Each bit is explained in the following table:

Bit		Description	
1	0x1	template having multiple segments in sequencing	Read paired
2	0x2	each segment properly aligned according to the aligner	Read mapped in proper pair
4	0x4	segment unmapped	Read unmapped
8	0x8	next segment in the template unmapped	Mate unmapped
16	0x10	SEQ being reverse complemented	Read reverse strand
32	0x20	SEQ of the next segment in the template being reverse comp	Mate reverse strand
64	0x40	the first segment in the template	First in pair
128	0x80	the last segment in the template	Second in pair
256	0x100	secondary alignment	Not primary alignment
512	0x200	not passing filters, such as platform/vendor quality controls	Read fails platform/vendor quality checks
1024	0x400	PCR or optical duplicate	Read is PCR or optical duplicate
2048	0x800	supplementary alignment	Supplementary alignment

https://broadinstitute.github.io/picard/explain-flags.html

SAM format - Extended CIGAR

```
Ref: GCATTCAGATGCAGTACGC
Read:  ccTCAG--GCATTAgtg
POS  CIGAR
5    2S4M2D6M3S
```

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

SAM format - Extended CIGAR

P	6	padding (silent deletion from padded reference)

```
  REF: CACGATCA**GACCGATACGTCCGA
READ1:   CGATCAGAGACCGATA
READ2:     ATCA*AGACCGATAC
READ3:    GATCA**GACCG

READ1: 6M2I8M
READ2: 4M1P1I9M
READ3: 5M2P5M
```

```
  REF: CACGATCA**GACCGATACGTCCGA
READ1:   CGATCAGAGACCGATA
READ2:     ATCAA*GACCGATAC
READ3:    GATCA**GACCG

READ1: 6M2I8M
READ2: 4M1I1P9M
READ3: 5M2P5M
```

N	3	skipped region from the reference

```
 REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
READ:           GTGTAACCC.............................TCAGAATA
```

where '...' on the read sequence indicates the intron. The CIGAR for this alignment is: `9M32N8M`.

All you need to keep in mind to work with SAM files

https://www.samformat.info/

BAM / CRAM formats

BAM (format spec.)

- Binary representation of SAM
- Compressed by BGZF library
- Greatly reduces size to about 27% of original SAM

CRAM (format spec.)

- More highly compressed alternative to BAM
- Reference based compression
- Only base calls that differ need to be stored

SAMtools

- Library and software package
- Create, sort and index BAM files from SAM files
- Remove PCR duplicates
- Merge alignments
- Visualization of alignments from BAM files
- SNP and short INDELs detection

https://www.htslib.org/doc/samtools.html

Picard

- SAMtools complementary package
- Numerous tools for manipulating SAM/BAM (n > 80)
- More format conversion than SAMtools
- More tools to collect metrics
- Natively compatible with GATK
- Provided as a single executable jar file

```
java jvm-args -jar picard.jar PicardToolName \
 OPTION1=value1 OPTION2=value2…
```

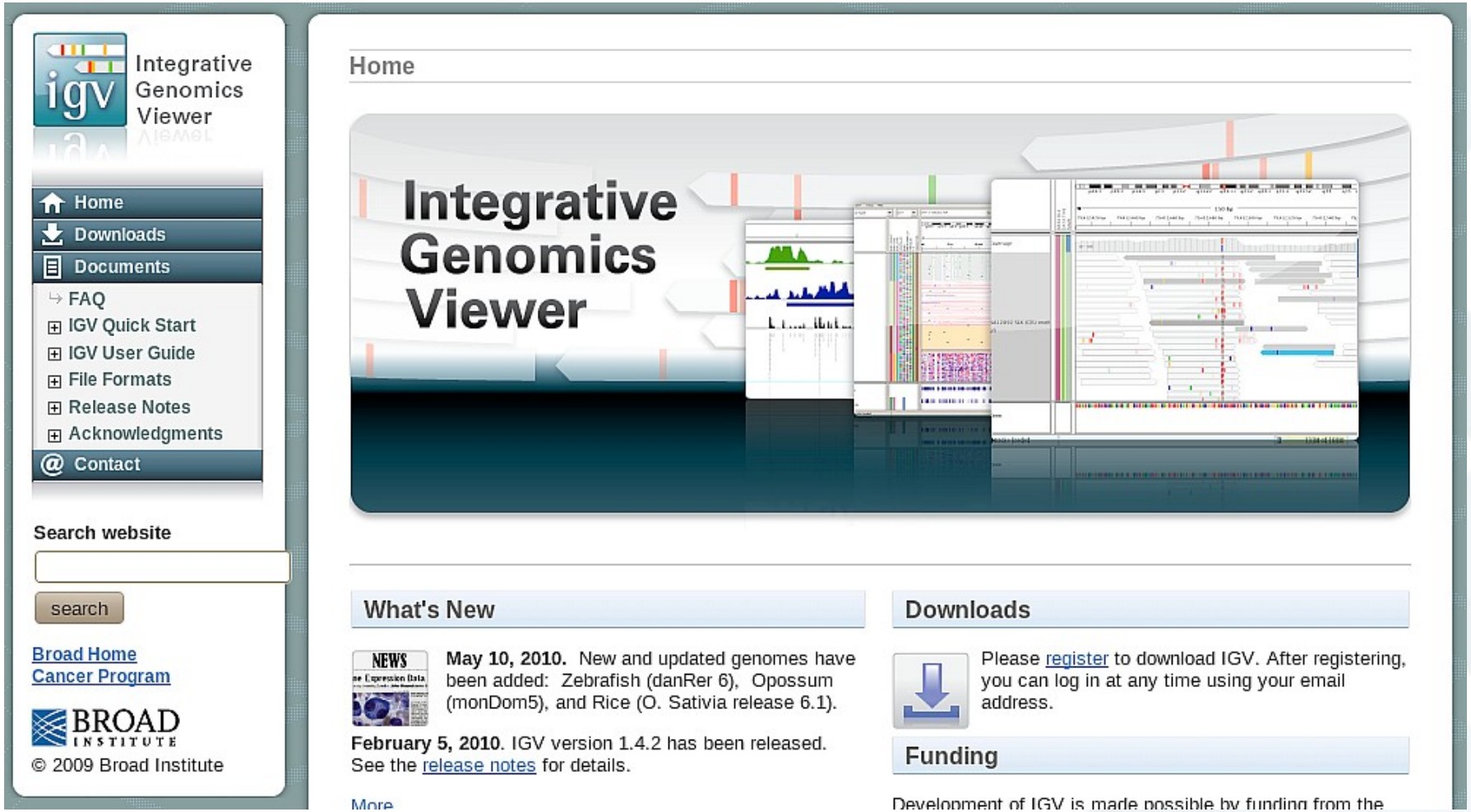
https://broadinstitute.github.io/picard/

Exercise / set 3

Visualizing the alignment - IGV

IGV Integrative Genomics Viewer

https://software.broadinstitute.org/software/igv/

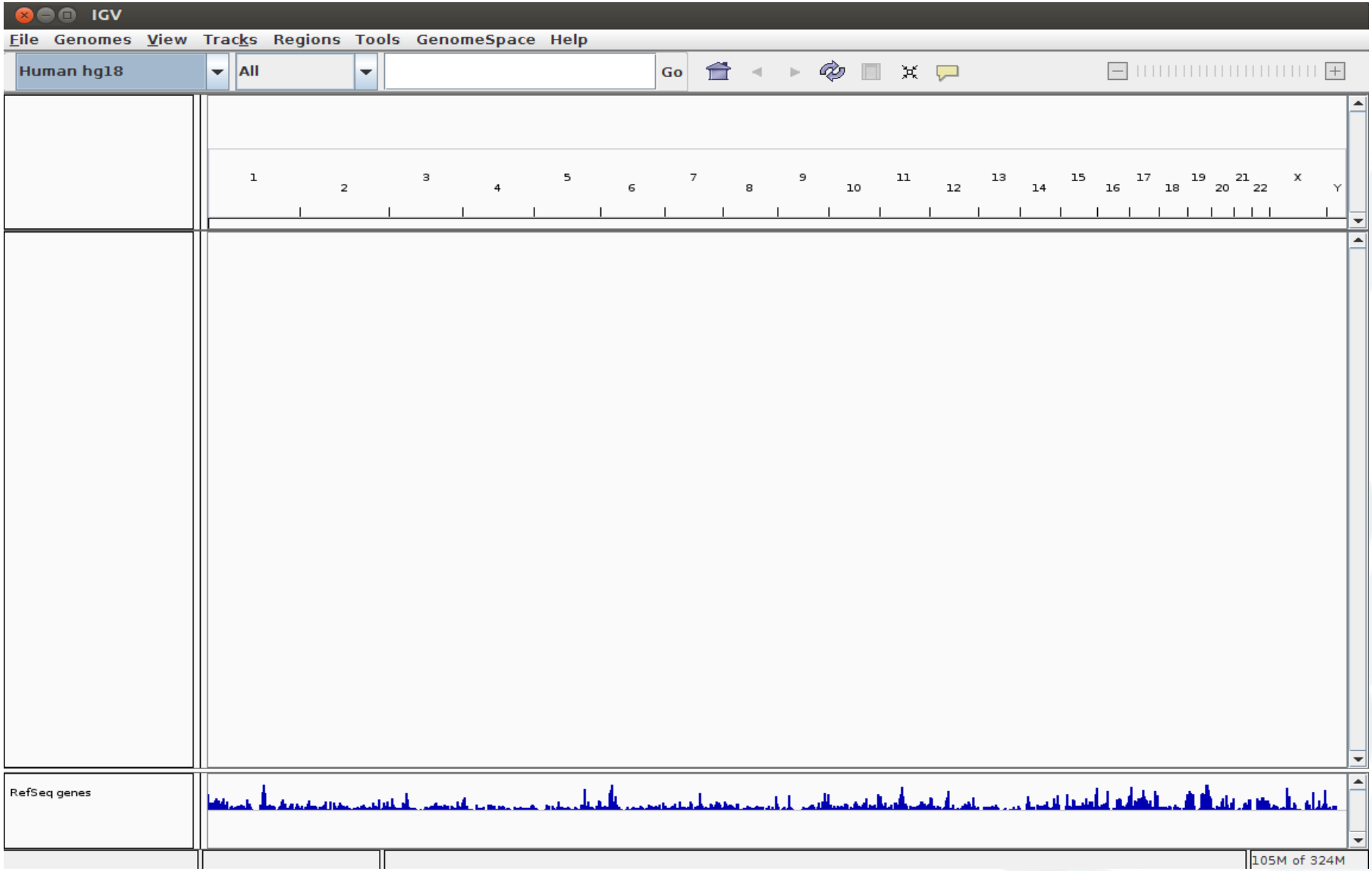


Visualizing the alignment - IGV

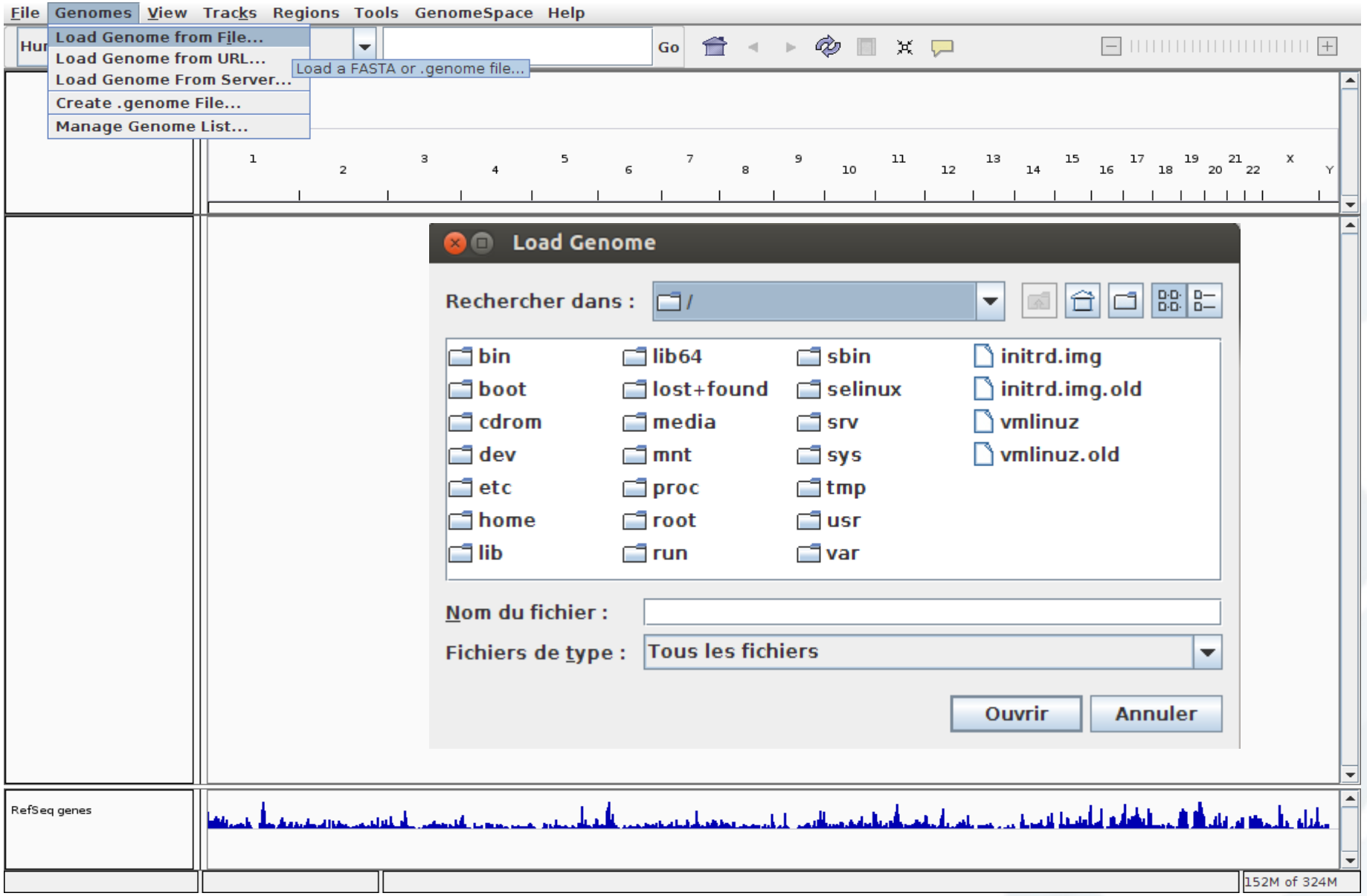
- High-performance visualization tool
- Interactive exploration of large, integrated datasets
- Supports a wide variety of data types
- Documentations
- Developed at the Broad Institute of MIT and Harvard

- BAM
- BED
- BEDPE
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- Chemical Reactivity Probing Profiles
- chrom.sizes
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- CRAM
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- RNA Secondary Structure Formats
- SAM
- Sample Info (Attributes) file
- SEG
- TDF
- Track Line
- Type Line
- VCF
- WIG

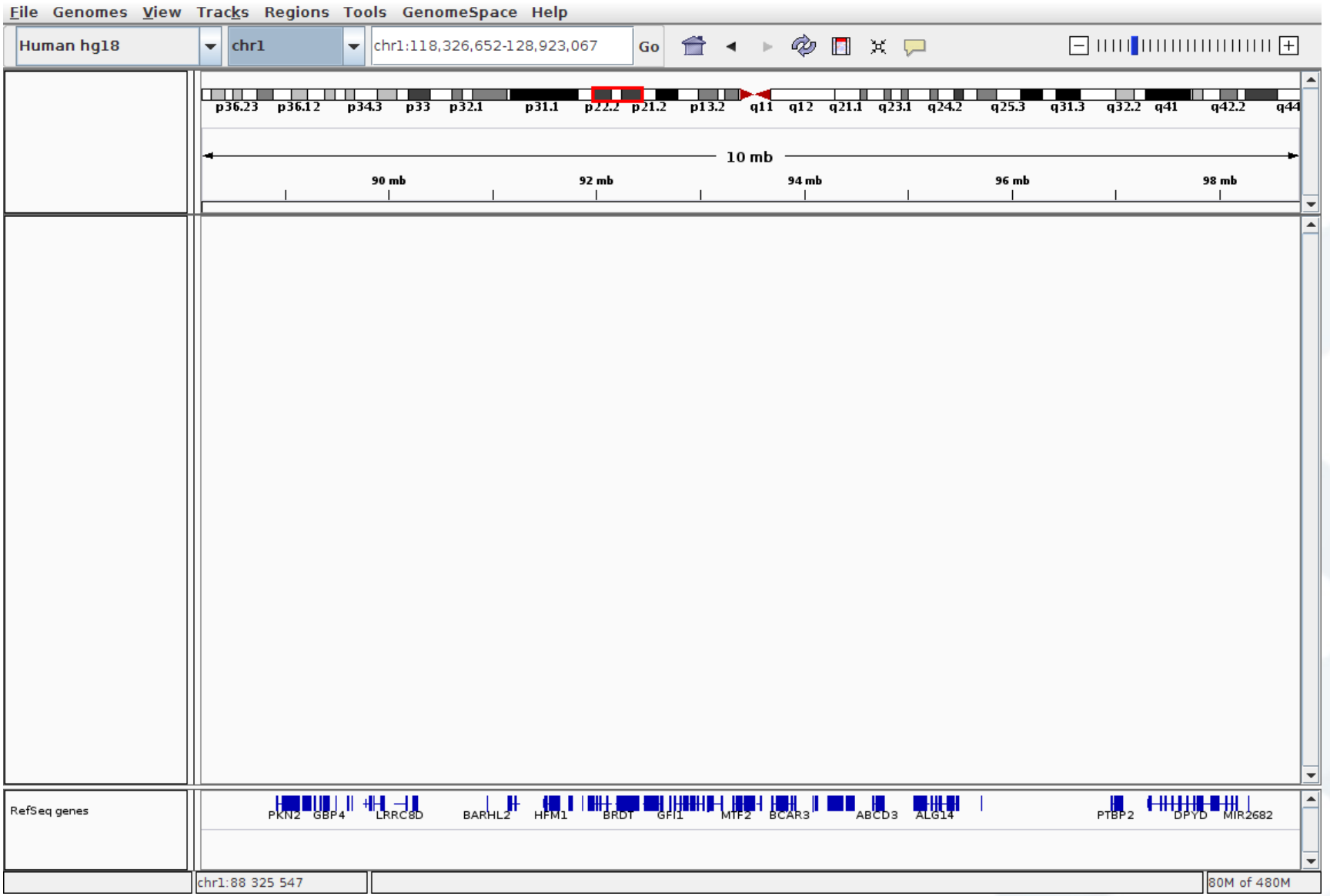
Visualizing the alignment - IGV



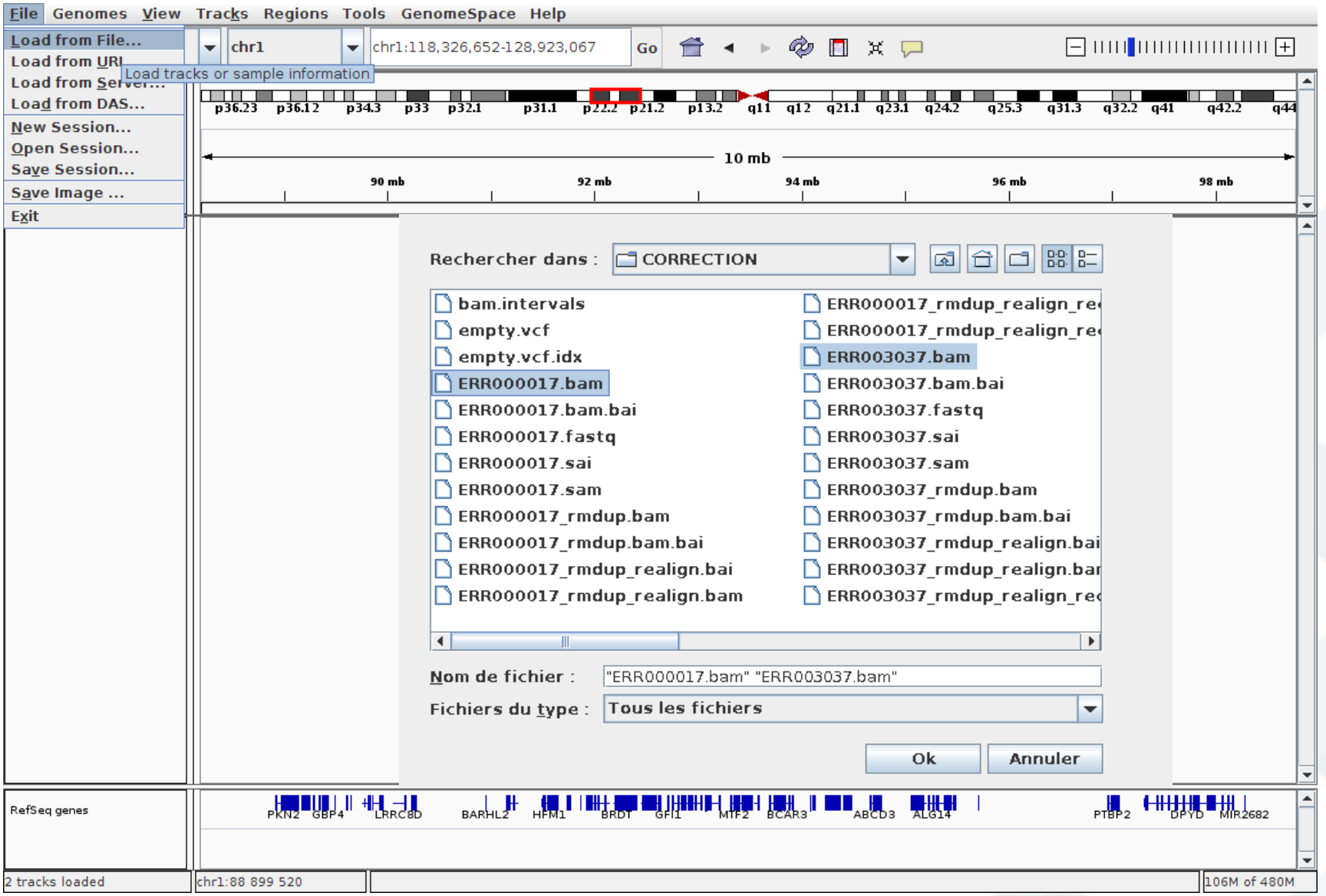
IGV - Loading the reference



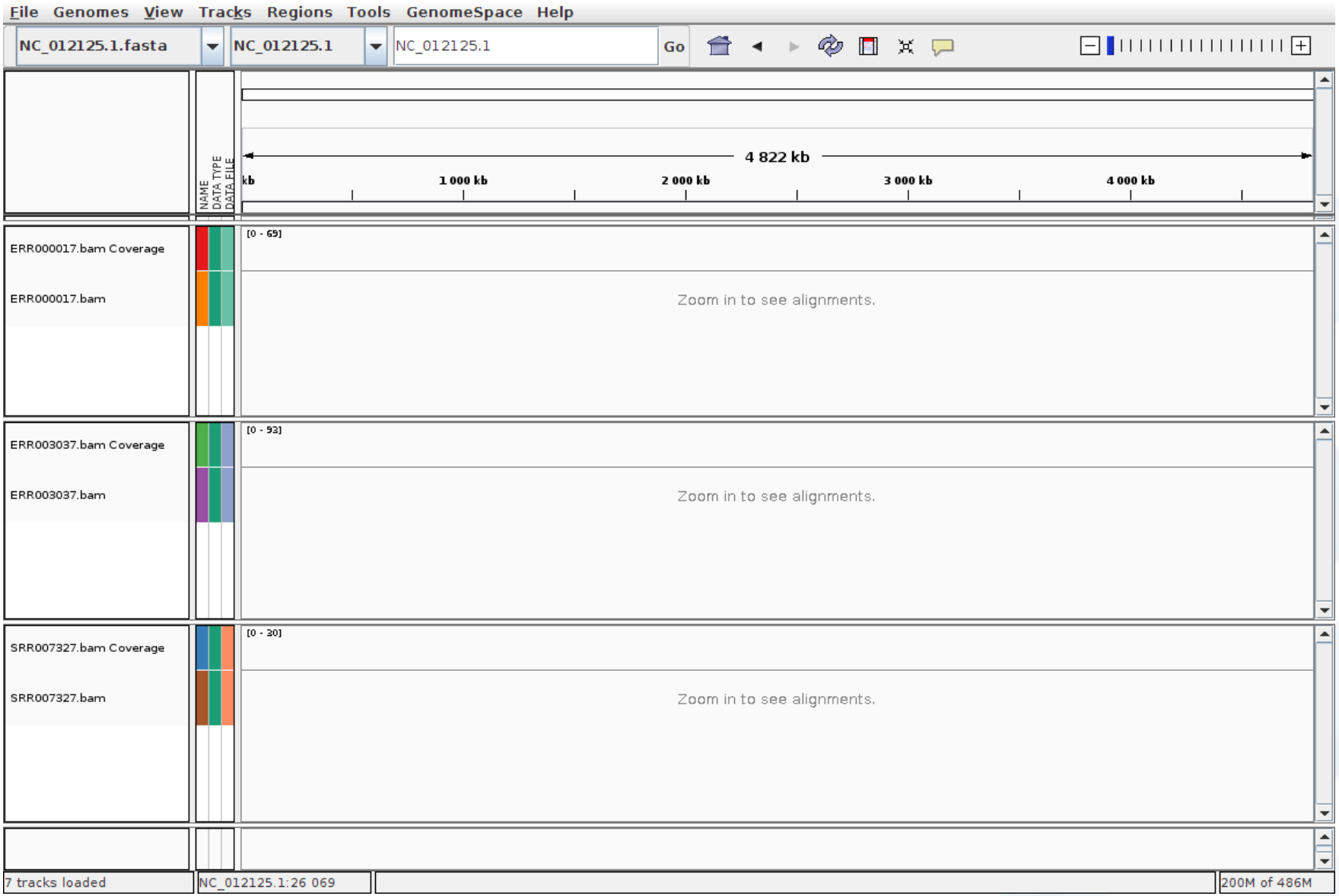
IGV - Loading the reference



IGV - Loading the bam file



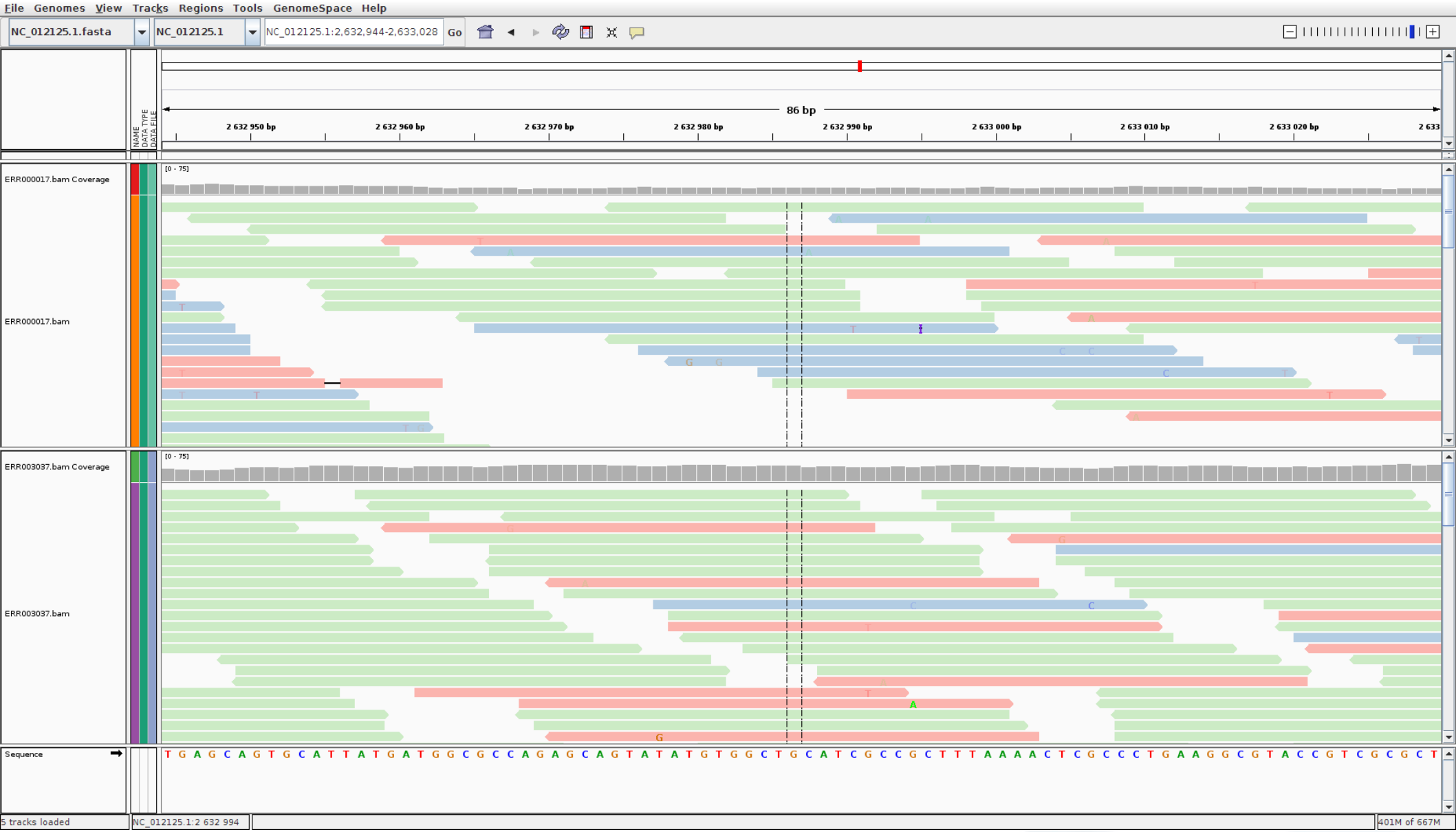
IGV - Loading the bam file



IGV - Zoom

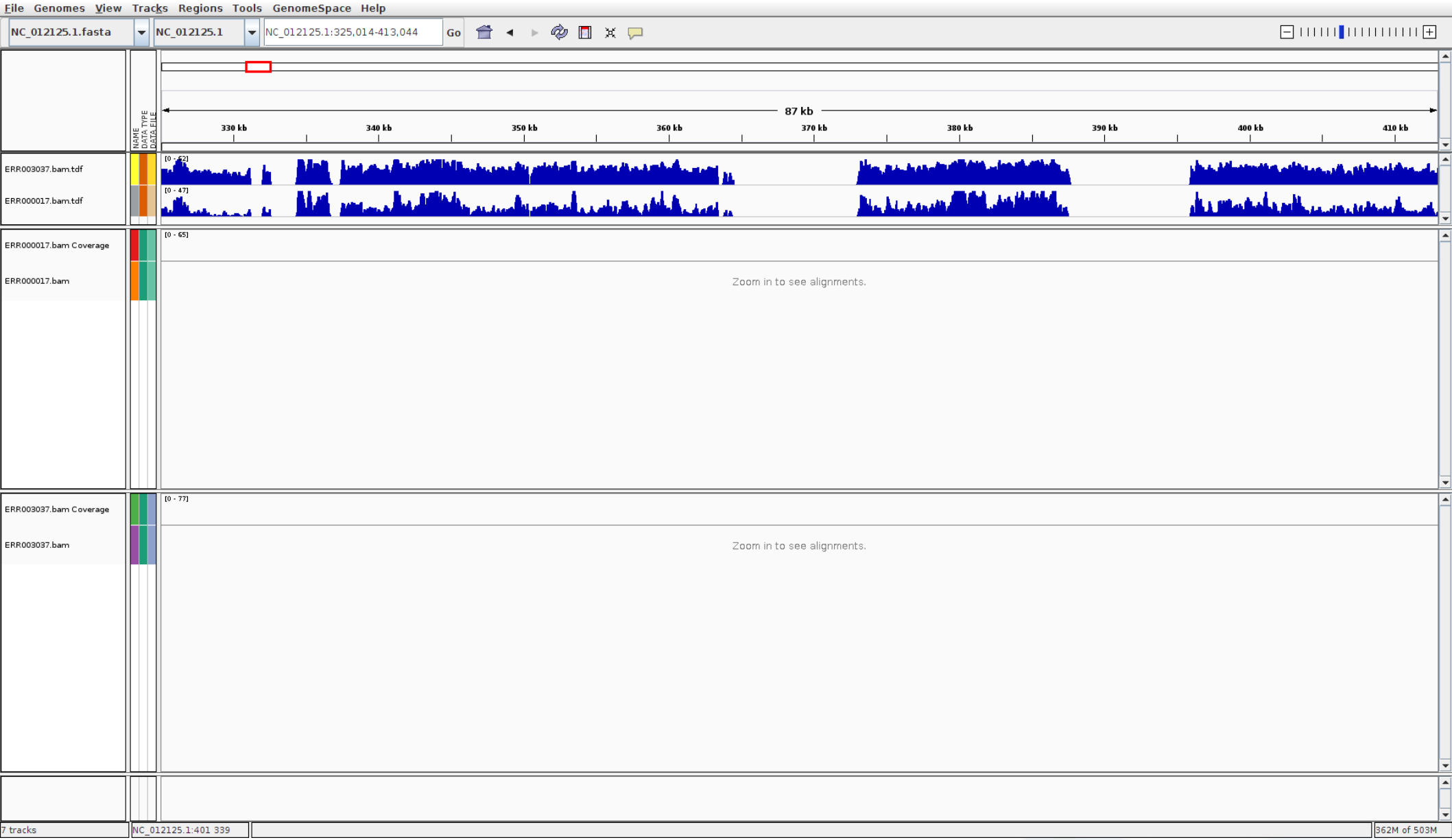

File Genomes View Tracks Regions Tools GenomeSpace Help
NC_012125.1.fasta
NC_012125.1
NC_012125.1:2,401,447-2,426,447
Go
24 kb
2 402 kb
2 404 kb
2 406 kb
2 408 kb
2 410 kb
2 412 kb
2 414 kb
2 416 kb
2 418 kb
2 420 kb
2 422 kb
2 424 kb
NAME
DATA TYPE
DATA FILE
ERR000017.bam Coverage
[0 - 69]
ERR000017.bam
ERR003037.bam Coverage
[0 - 93]
ERR003037.bam
SRR007327.bam Coverage
[0 - 30]
SRR007327.bam
Sample = 37
Read group = 37
Read name = ERR003037.1217157
Alignment start = 2414936 (-)
Cigar = 33M
Mapped = yes
Mapping quality = 37
Base = G
Base phred quality = 29
X0 = 1
X1 = 0
MD = 33
RG = 37
XG = 0
NM = 0
XM = 0
7 tracks loaded
NC_012125.1:2 414 959
217M of 486M

IGV - Zoom


File Genomes View Tracks Regions Tools GenomeSpace Help
NC_012125.1.fasta
NC_012125.1
NC_012125.1:2,632,944-2,633,028
Go
86 bp
2 632 950 bp
2 632 960 bp
2 632 970 bp
2 632 980 bp
2 632 990 bp
2 633 000 bp
2 633 010 bp
2 633 020 bp
ERR000017.bam Coverage
ERR000017.bam
ERR003037.bam Coverage
ERR003037.bam
Sequence
5 tracks loaded
NC_012125.1:2 632 994
401M of 667M

IGV - Loading an annotation file

IGV - Coverage



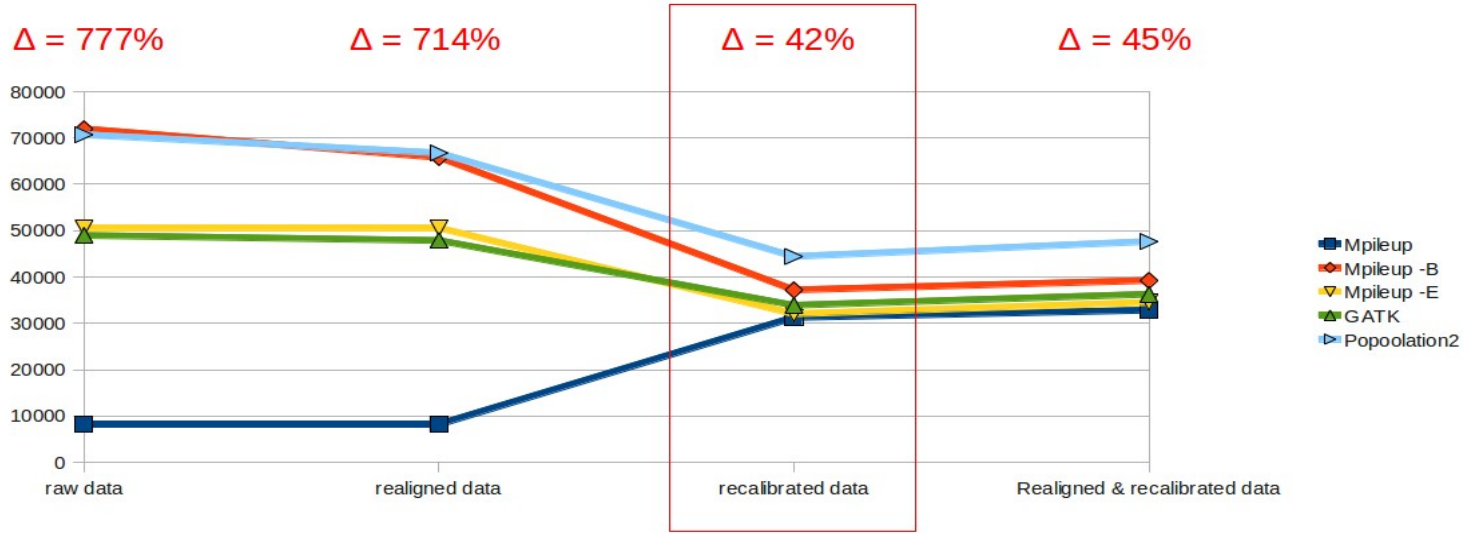
Exercise / set 4

Variant Calling methodology

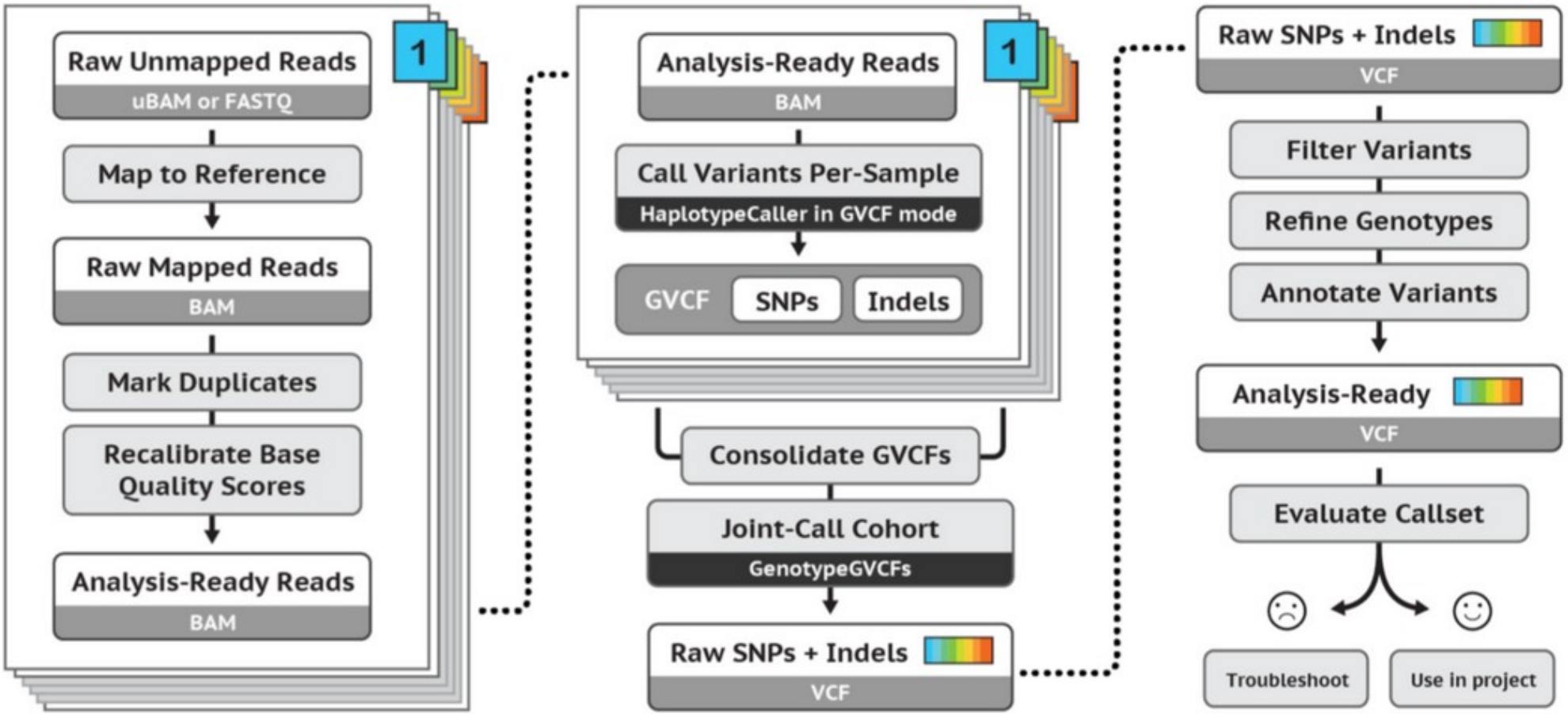
- Several Variant callers :
  - Samtools mpileup
  - GATK
  - FreeBayes
  - DeepVariant
  - ...

Why GATK ?

- 1000 genomes project
- Very used and well documentated
- RNAseq – DNAseq
- Several technologies supported
- Tested on our data :

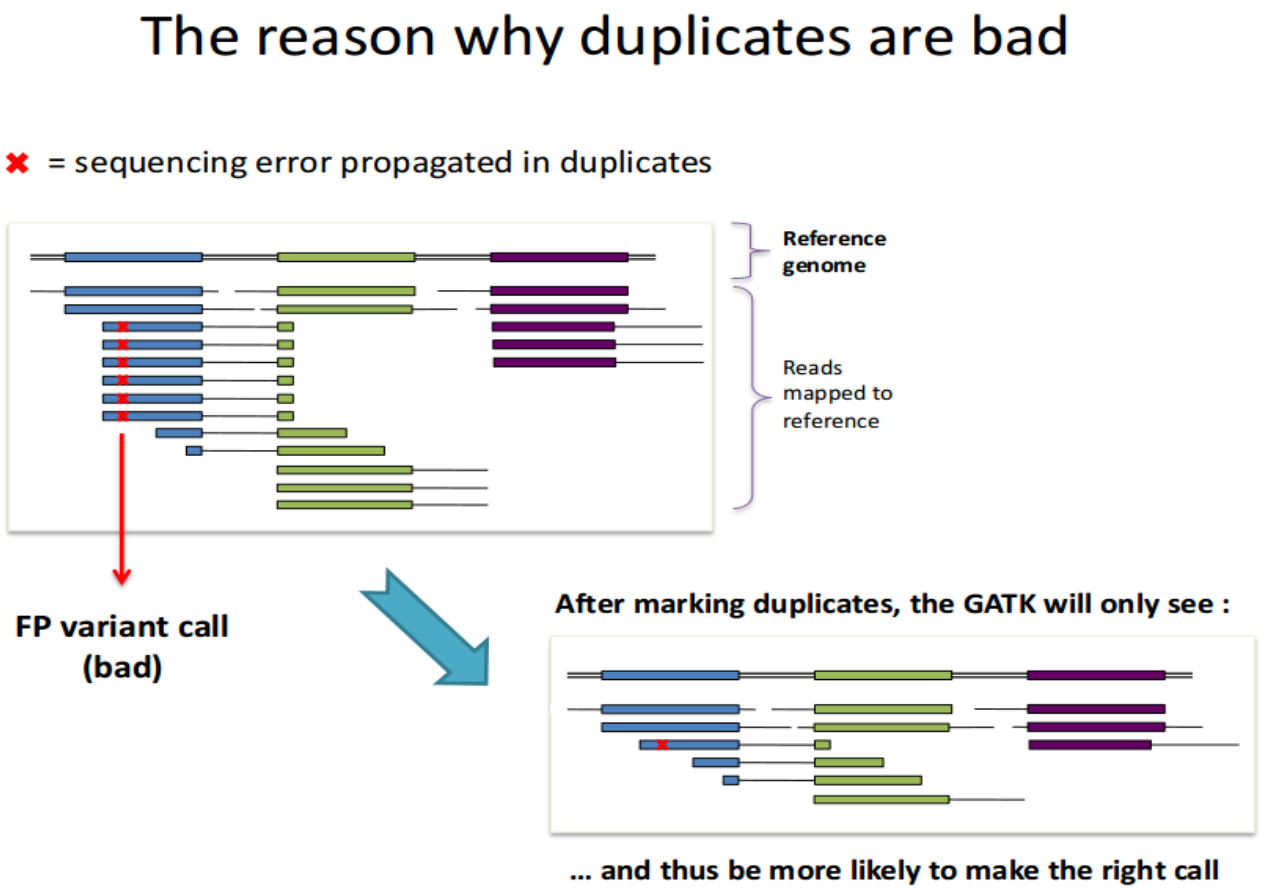


THE GATK best practices



Pre-processing – MarkDuplicates

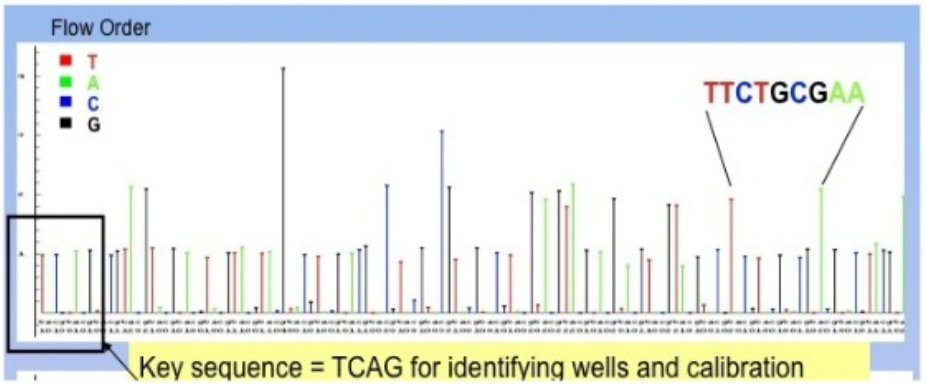
- Remove/mark duplicates (PCR and/or Optical)

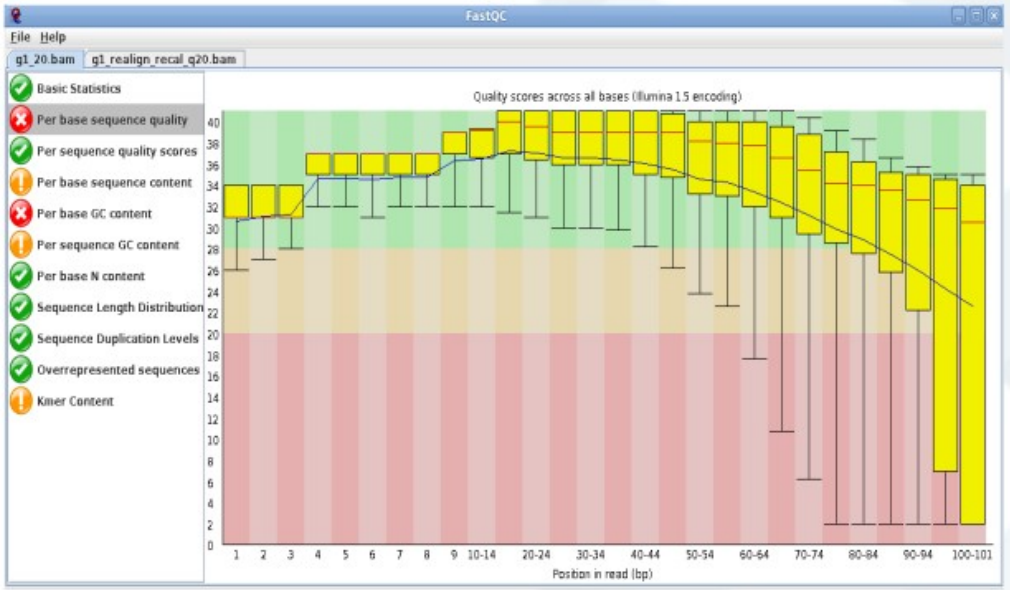



```
$ gatk --java-options "-Xms4000m -Xmx7g" MarkDuplicates \
> --INPUT SRR7062654.bam --METRICS_FILE SRR7062654.bam.metrics \
> --TMP_DIR . \
> --ASSUME_SORT_ORDER coordinate \
> --CREATE_INDEX true \
> --OUTPUT SRR7062654.md.bam
```

Pre-processing – BQSR

- Short variant calling algorithms rely heavily on the quality score
- Scores produced by the machines are subject to various sources of systematic (non-random) technical error
- The base quality provided by the sequencers is too inaccurate to be kept



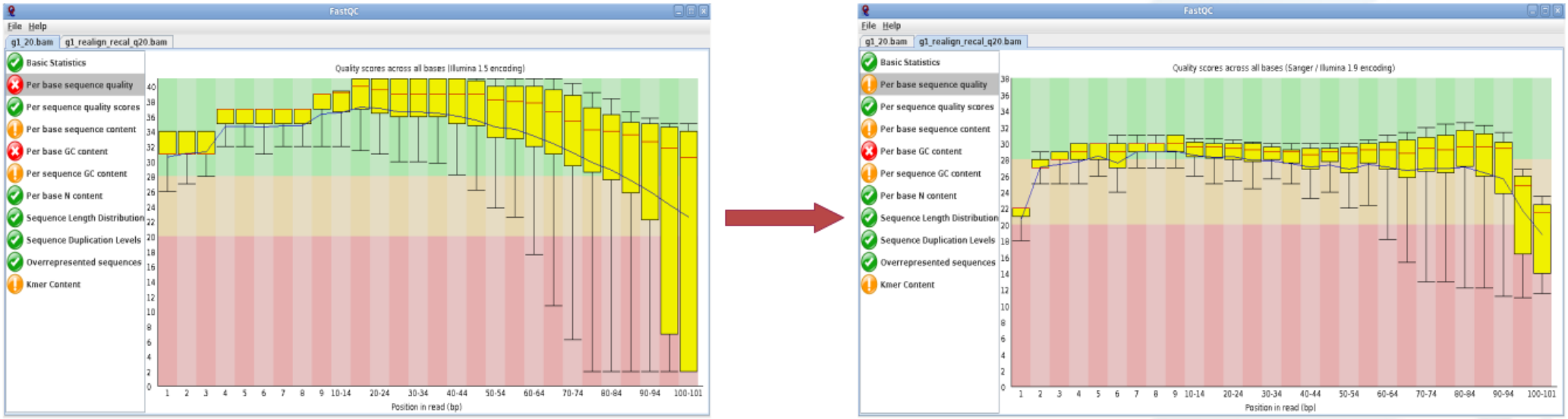


Pre-processing – BQSR step1

- **Needs knowledge of real SNP** => mask out bases at sites of real (expected) variation
- BaseRecalibrator builds the model:
  - Read group the read belongs to
  - Quality score reported by the machine
  - Machine cycle producing this base (Nth cycle = Nth base from the start of the read)
  - Current base + previous base (dinucleotide)

Pre-processing – BQSR step2

- ApplyBQSR adjusts the scores



https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-

GATK – BQSR

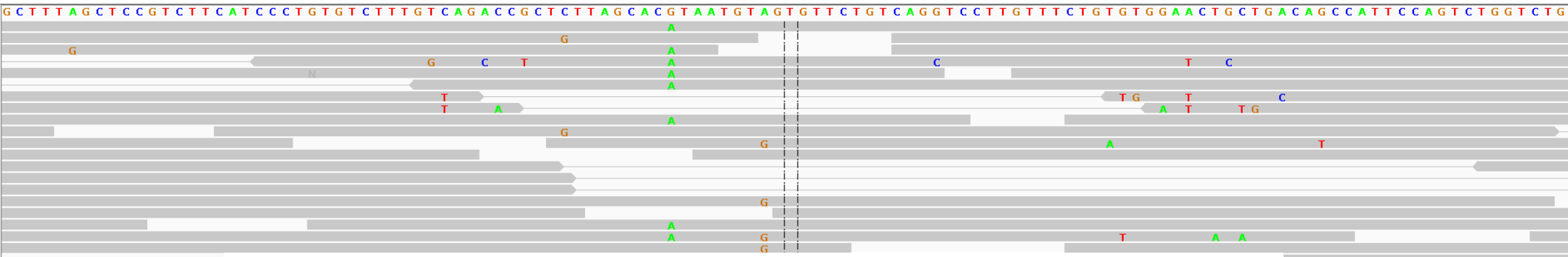
- Two steps :
  - BaseRecalibrator builds the model

```
$ gatk --java-options -Xmx7g BaseRecalibrator \
> -I SRR7062654.md.bam \
> -O SRR7062654.md.recal.table \
> --tmp-dir . \
> -R Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa \
> --known-sites Gallus_gallus_incl_consequences_chr25-26.vcf \
> --verbosity INFO
```

  - ApplyBQSR adjusts the scores :

```
$ gatk --java-options -Xmx14g ApplyBQSR \
> -R Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa \
> --input SRR7062654.md.bam \
> --output SRR7062654.md.recal.bam \
> --bqsr-recal-file SRR7062654.md.recal.table
```

Variant discovery



Base scale

Read scale

Position scale

Genotype scale

Phred-Quality Base

Mapping Quality

Forward/Reverse

ALT allele count

REF allele count

ALT / REF

Read Depth

Overall genotype association

SNP quality

$10 \Rightarrow P_{error} = 1 / 10$

$30 \Rightarrow P_{error} = 1 / 1000$

GATK – HaplotypeCaller

- Call SNPs and indels via local re-assembly of haplotypes
- How HaplotypeCaller work:
  1. Define active regions
  2. Determine haplotypes by assembly of the active regions
     1. Reassemble region and identifies haplotypes (De Bruijn-like graph)
     2. realigns each haplotype against the reference haplotype (Smith-Waterman) in order to identify potentially variant sites
  3. Determine likelihoods of the haplotypes given the read data
  4. Assign sample genotypes

```
$ gatk --java-options "-Xmx10g -Xms1000m" HaplotypeCaller \
> -R Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa \
> -I SRR7062654.md.recal.bam \
> -O SRR7062654.md.recal.g.vcf \
> -ERC GVCF
```

https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller

GATK – CombineGVCFs

- Merges one or more HaplotypeCaller GVCF files into a single GVCF
- Combine per-sample gVCF files produced by HaplotypeCaller into a multi-sample gVCF file

```
$ gatk --java-options -Xmx10g CombineGVCFs \
> -R Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa \
> -O SRR.g.vcf \
> --variant SRR7062654.md.recal.g.vcf \
> --variant SRR7062655.md.recal.g.vcf
```

https://gatk.broadinstitute.org/hc/en-us/articles/360037053272-CombineGVCFs

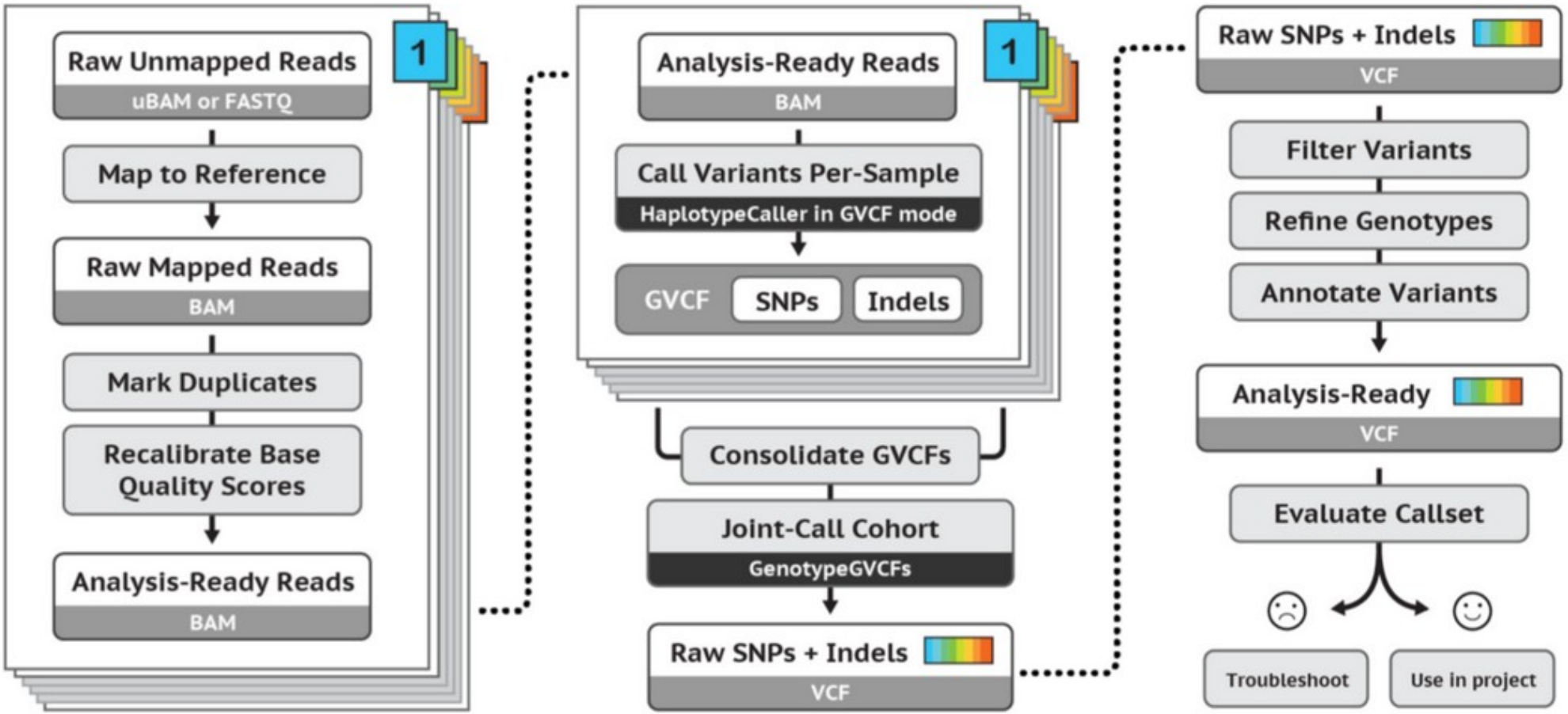
GATK – GenotypeGVCFs

- Perform joint genotyping on one or more samples pre-called with HaplotypeCaller

```
$ gatk --java-options "-Xmx10g" GenotypeGVCFs \
> -R Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa \
> -V SRR.g.vcf \
> -O SRR.vcf.gz
```

https://gatk.broadinstitute.org/hc/en-us/articles/360037057852-GenotypeGVCFs

THE GATK best practices



Exercise / set 5

VCF format

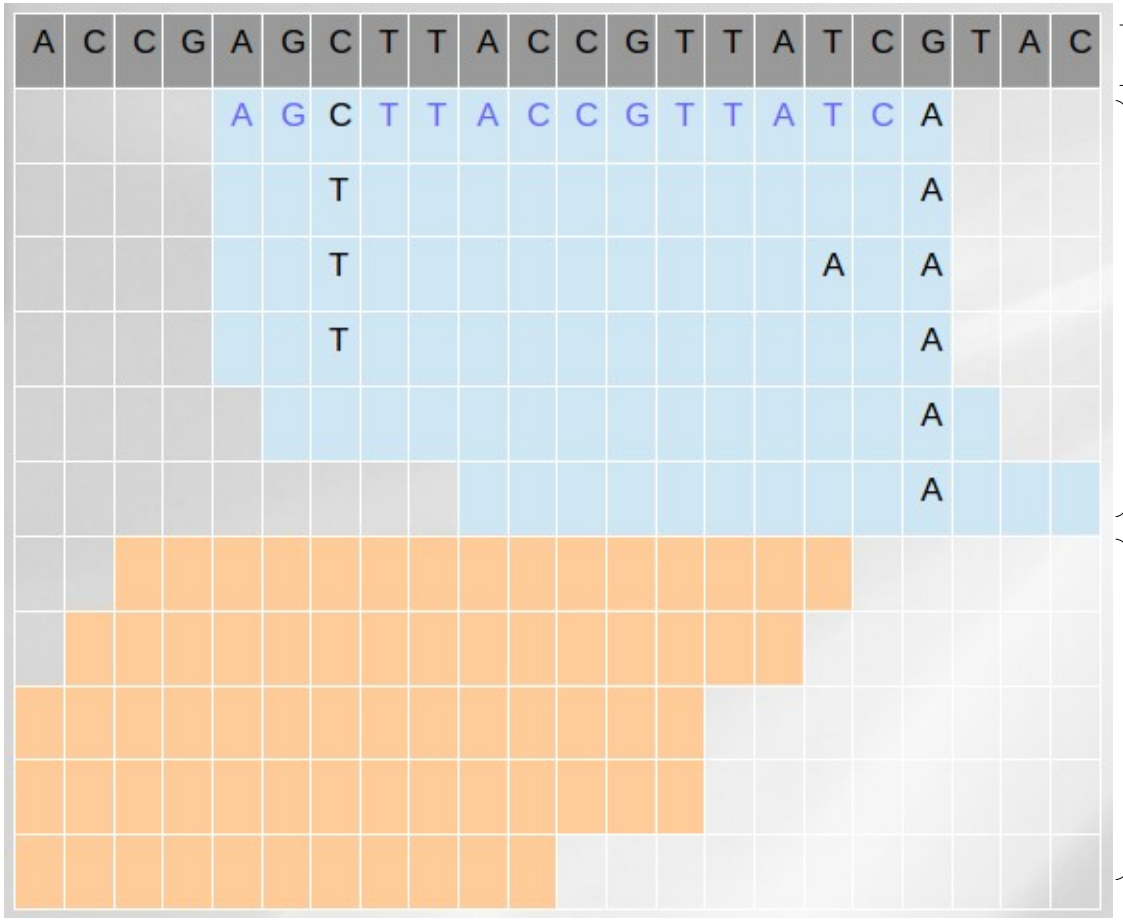
- Which informations have been stored in VCF ?

https://web-genobioinfo.toulouse.inrae.fr/~formation/16_SGS-SNP/.formats/vcf.html

Variant calling Format (VCF)

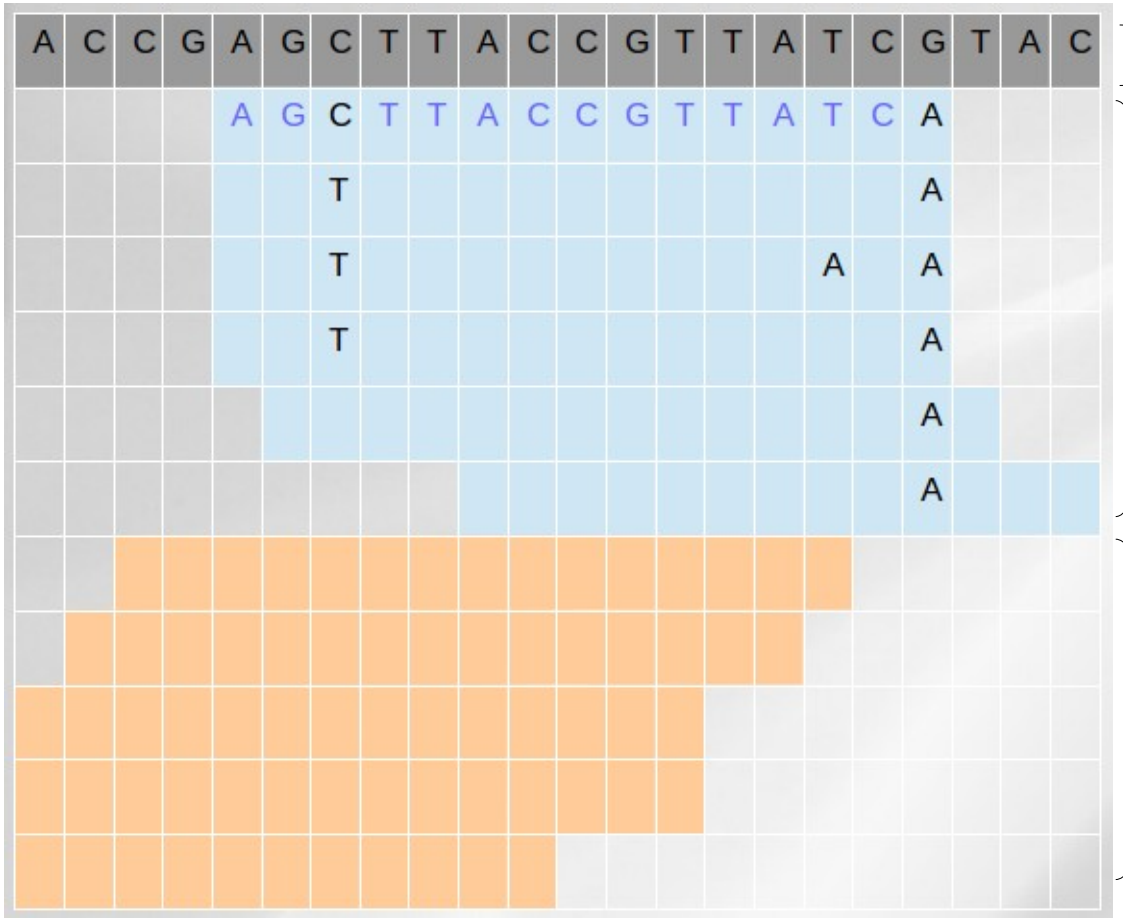
- http://vcftools.sourceforge.net/specs.html
- Tab-delimited file
- Basic documentation inside
- Header lines start with ## or #
- Give informations of data/tools/parameters used
- Variant lines represent a position on the genome

The VCF format : example



#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

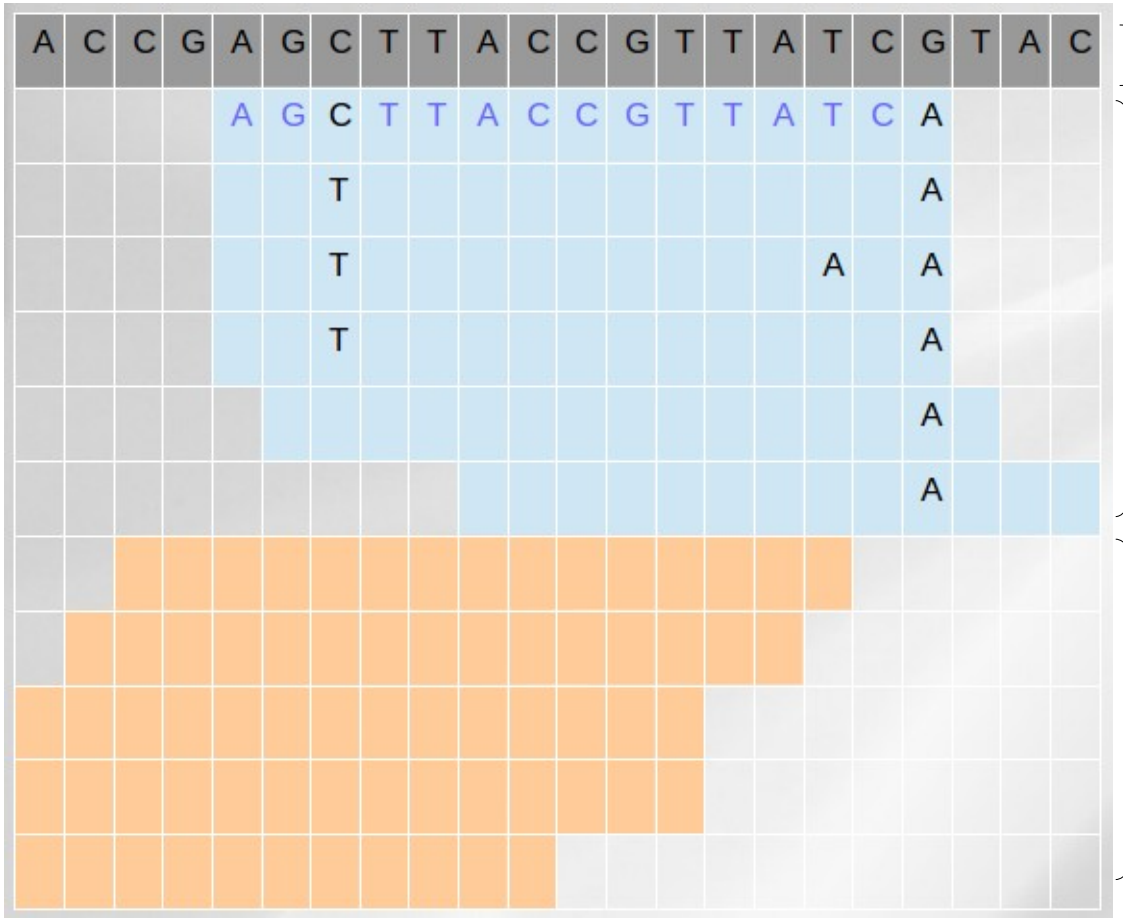
The VCF format : example



#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

The Phred scaled probability that a REF/ALT polymorphism exists at this site given sequencing data

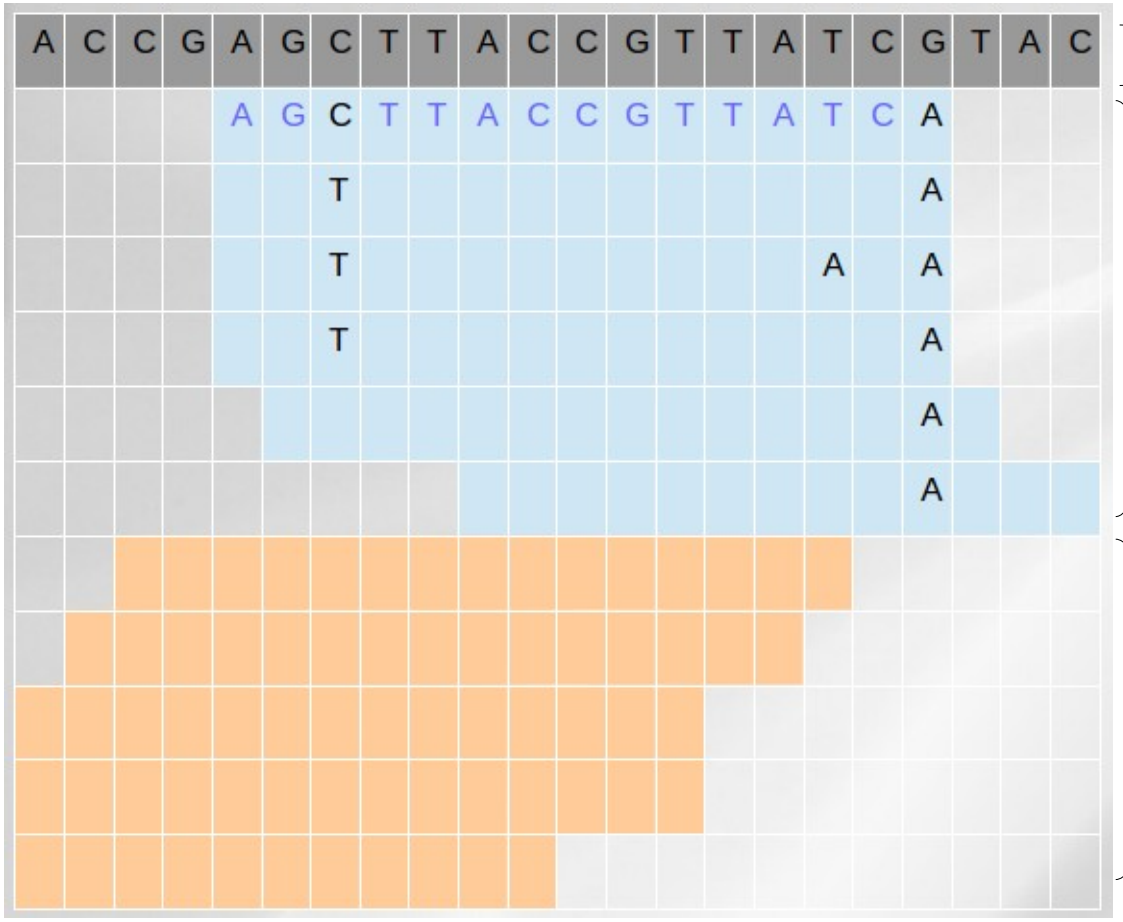
The VCF format : example



#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

[TAG=VALUE]
DP=45

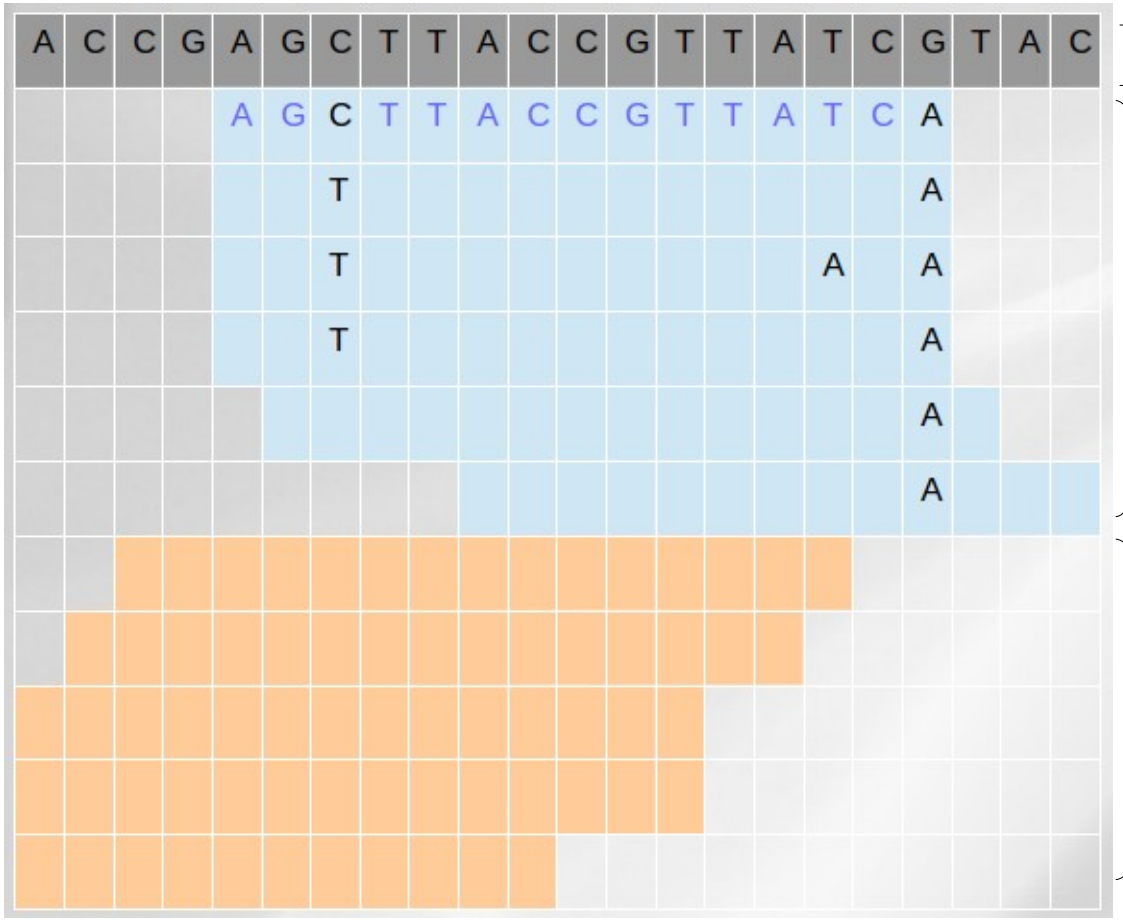
The VCF format : example



#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

0/1 : homozygous reference
0/1 : heterozygous
1/1 : homozygous alternative

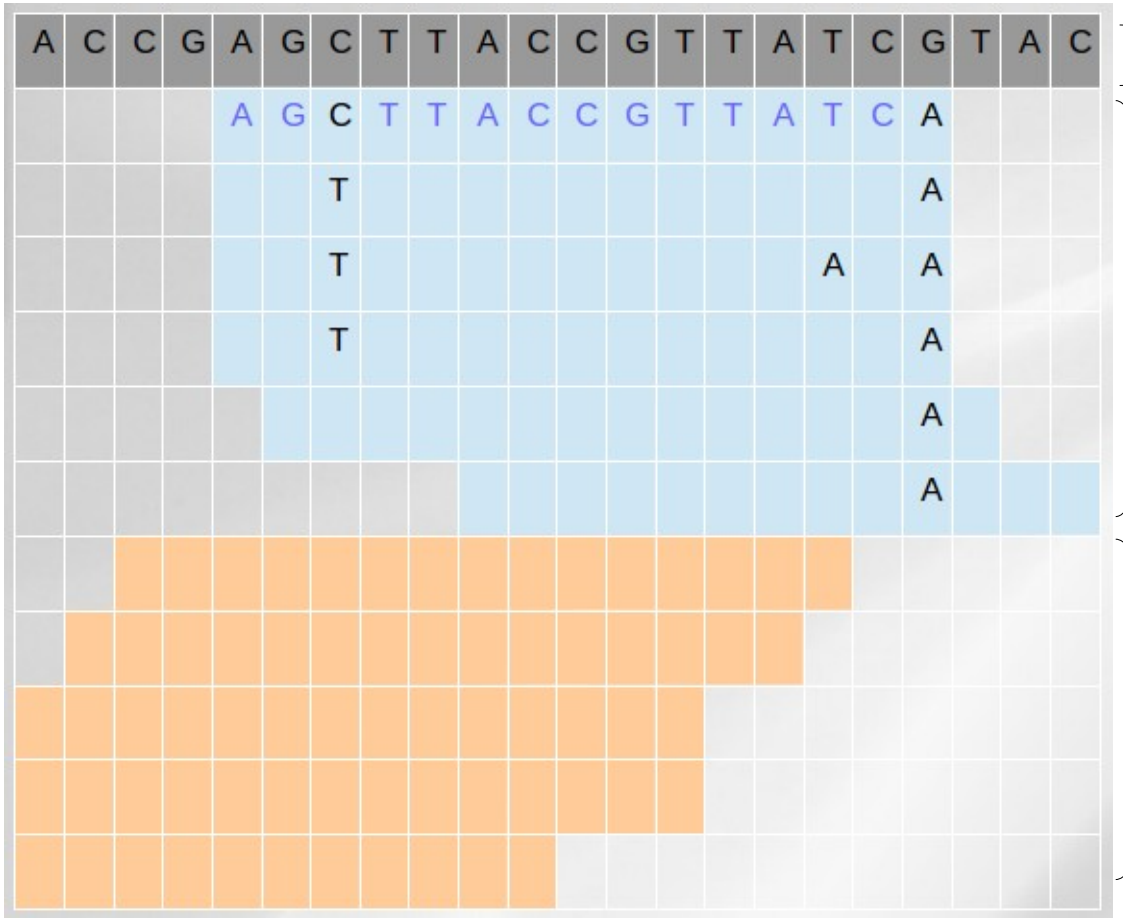
The VCF format : example



#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

count REF,count ALT *[,count ALT2...]*

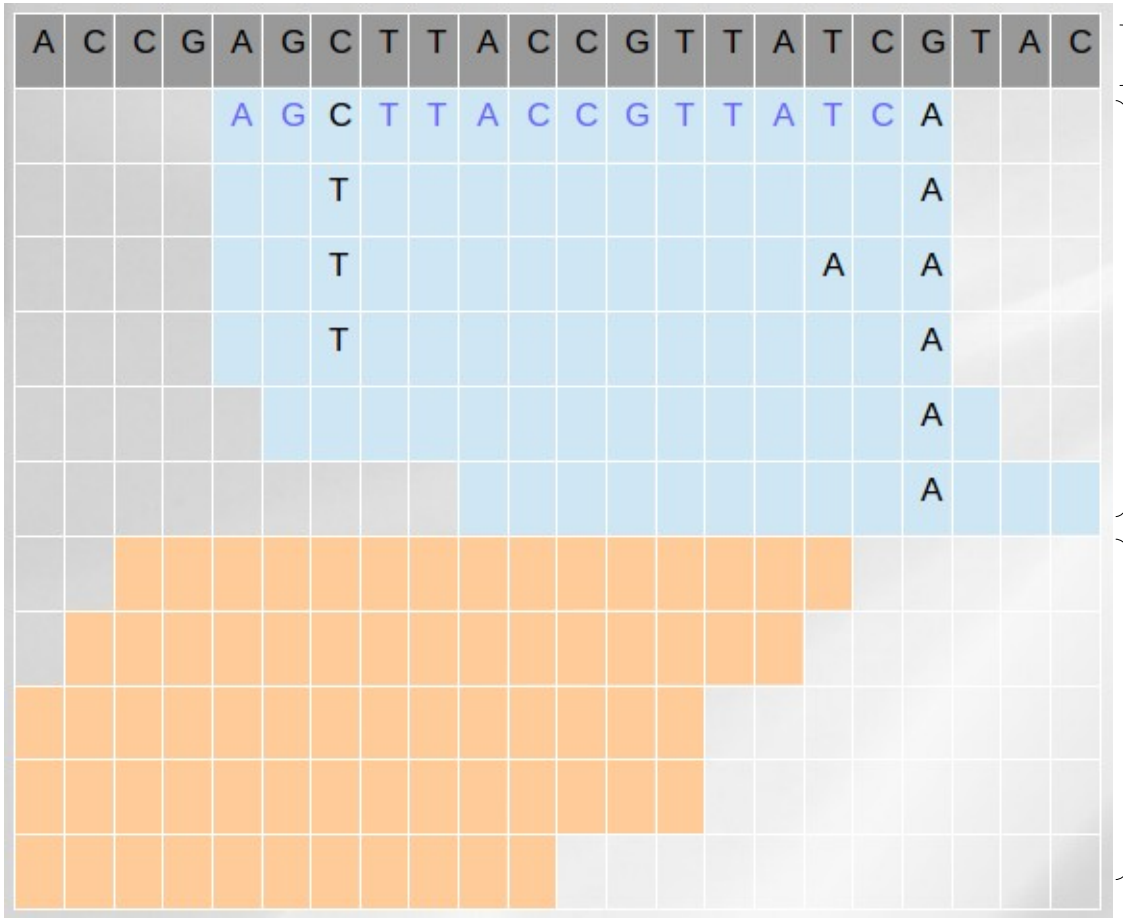
The VCF format : example




```
#CHR  POS  ID  REF  ALT  QUAL  FILTER  INFOS     FORMAT          SAMPLE_1              SAMPLE_2
chr1  7    .   C    T    247.82    .   [INFOS]   GT:AD:DP:GQ:PL  0/1:3,3:6:9.2:20,0,15  0/0:5,0:5:20.2:0,42,94
chr1  19   .   G    A    124.34    .   [INFOS]   GT:AD:DP:GQ:PL  1/1:0,6:6:9.01:10,1,6  ./.
```

Depth position

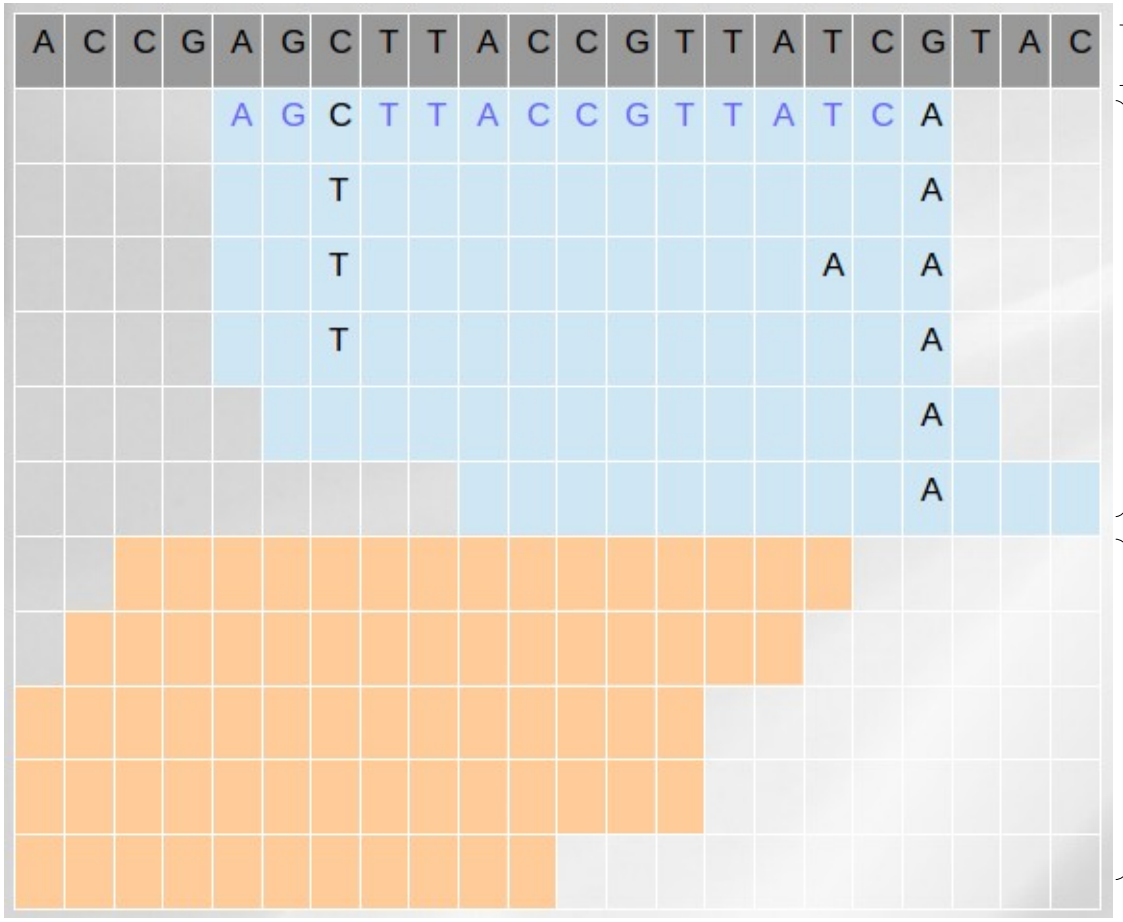
The VCF format : example




```
#CHR  POS  ID   REF  ALT  QUAL   FILTER  INFOS     FORMAT          SAMPLE_1                SAMPLE_2
chr1  7    .    C    T    247.82    .    [INFOS]   GT:AD:DP:GQ:PL  0/1:3,3:6:9.2:20,0,15    0/0:5,0:5:20.2:0,42,94
chr1  19   .    G    A    124.34    .    [INFOS]   GT:AD:DP:GQ:PL  1/1:0,6:6:9.01:10,1,6    ./.
```

The Genotype Quality, or Phred-scaled confidence that the true genotype is the one provided in GT.

The VCF format : example

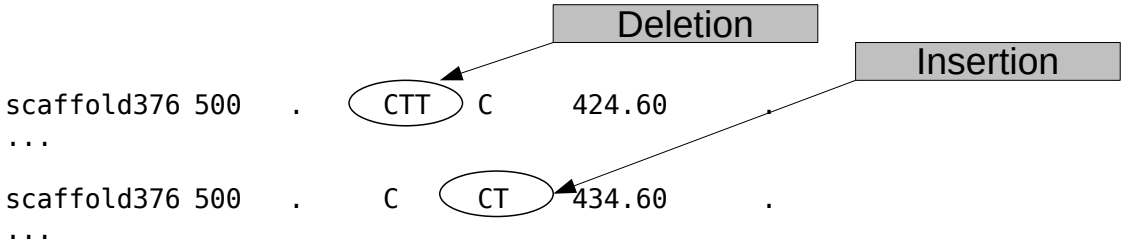


#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFOS	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:3,3:6:9.2:20,0,15	0/0:5,0:5:20.2:0,42,94
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	1/1:0,6:6:9.01:10,1,6	./.

These are normalized, Phred-scaled likelihoods for each of the 0/0, 0/1, and 1/1, without priors.
Ex : PL(0/0) = 26, which corresponds to 10^(-2.6), or 0.0025

The VCF format

- Small INDELs



- Multi-allelic variants

```
scaffold376 577   .    C    A,T  2303.19    .
...
GT:AD:DP:GQ:PL    1/2:0,10,6:16:99:394,145,118,249,0,234    1/2:0,20,6:26:99:658,160,106,498,0,480
```

The VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

The VCF header

VCF version

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

The VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

Fields description

The VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

Tools & options used

The VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

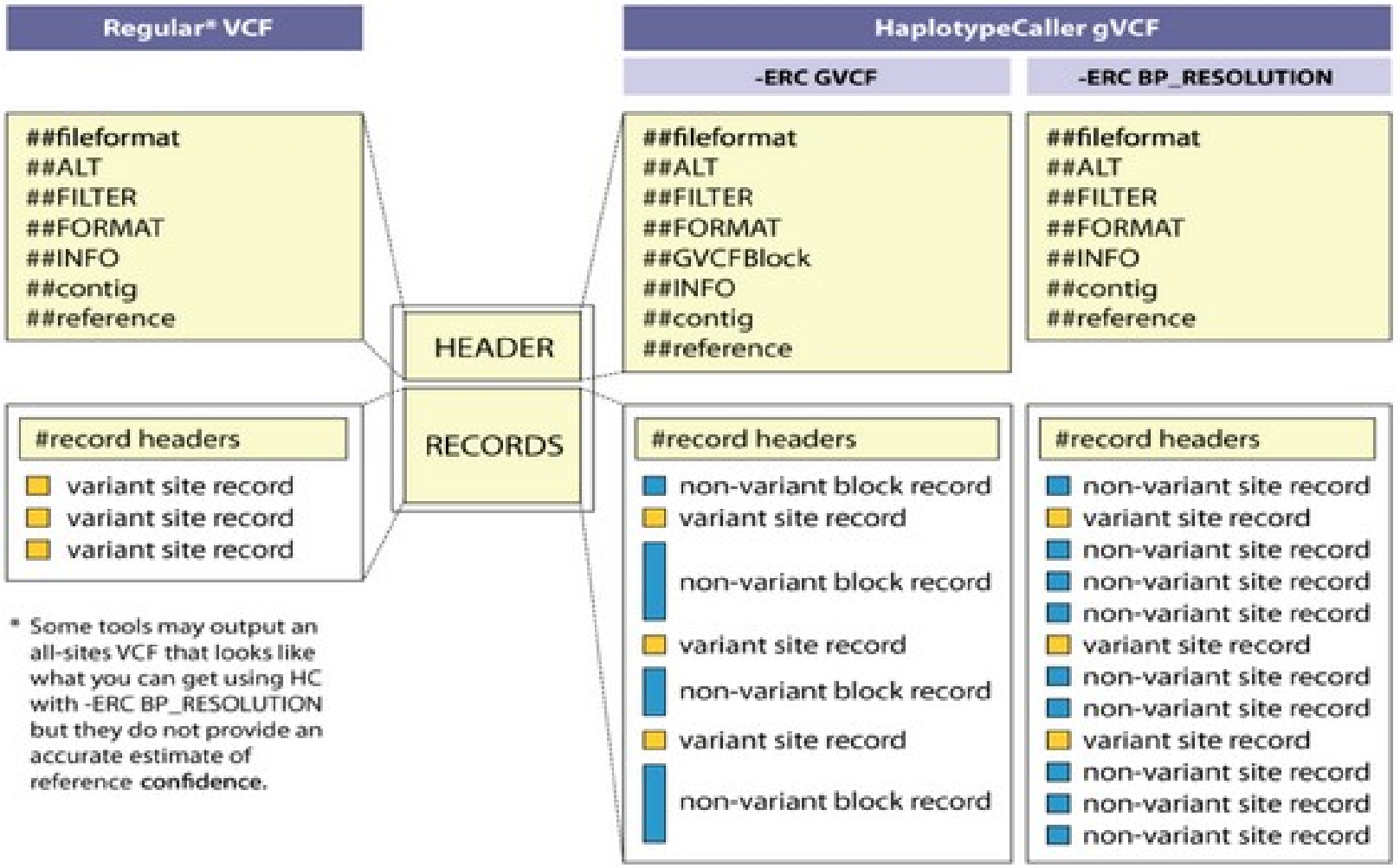
Genome informations

The VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when
compared against the Hardy-Weinberg expectation">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[L1_RG_s_realign_recal_q30.bam, L2_RG_s_realign_recal_q30.bam]
read_buffer_size=null phone_home=STANDARD gatk_key=null read_filter=[] [...]
##contig=<ID=scaffold376,length=1000>
##reference=file:///work/banks/genome.fasta
#CHROM       POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT l1     l2
```

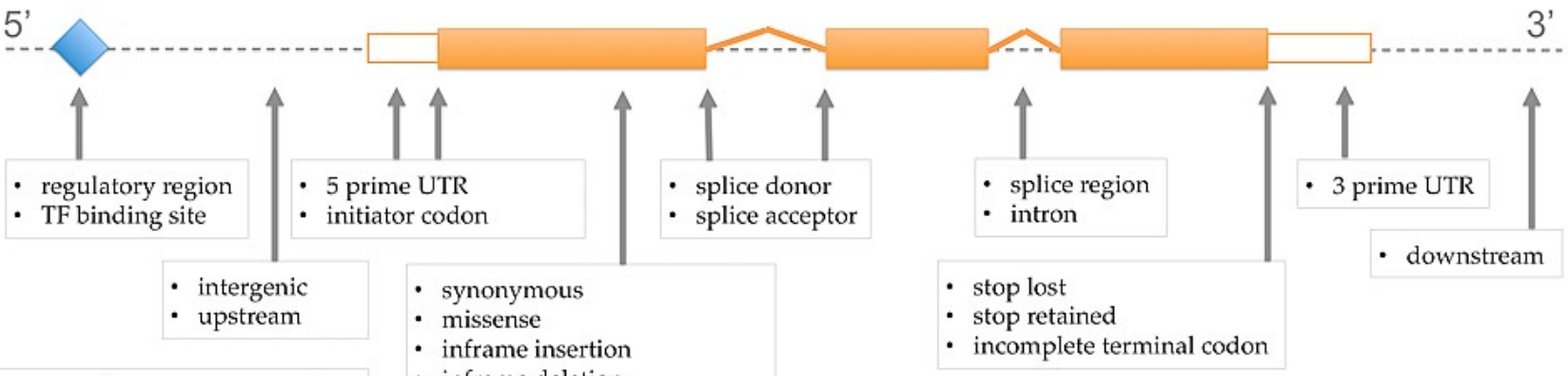
Header line

The VCF vs gVCF



Variant annotation

- Where are my SNPs ?
- Known or unknown ?
- Which effects ?



Type

Type (alphabetical order)	Count	Percent
CODON_DELETION	1	0.001%
CODON_INSERTION	2	0.002%
DOWNSTREAM	2,036	1.965%
EXON	30	0.029%
FRAME_SHIFT	6	0.006%
INTERGENIC	87,171	84.121%
INTRAGENIC	134	0.129%
INTRON	11,609	11.203%
NON_SYNONYMOUS_CODING	77	0.074%
SPLICE_SITE_DONOR	1	0.001%
SPLICE_SITE_REGION	22	0.021%
STOP_GAINED	2	0.002%
SYNONYMOUS_CODING	112	0.108%
UPSTREAM	2,324	2.243%
UTR_3_PRIME	91	0.088%
UTR_5_PRIME	8	0.008%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,036	1.965%
EXON	230	0.222%
INTERGENIC	87,171	84.121%
INTRON	11,609	11.203%
NONE	134	0.129%
SPLICE_SITE_DONOR	1	0.001%
SPLICE_SITE_REGION	22	0.021%
UPSTREAM	2,324	2.243%
UTR_3_PRIME	91	0.088%
UTR_5_PRIME	8	0.008%

Variant annotation - SnpEff

RESEARCH PAPER

Fly 6:2, 1-13; April/May/June 2012; © 2012 Landes Bioscience

A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3

Pablo Cingolani,[1-3] Adrian Platts,[4] Le Lily Wang,[1] Melissa Coon,[2] Tung Nguyen,[5] Luan Wang,[1,2] Susan J. Land,[2] Douglas M. Ruden[1,2,*] and Xiangyi Lu[1]

[1]Institute of Environmental Health Sciences; Wayne State University; Detroit, MI USA; [2]Department of Obstetrics and Gynecology; Wayne State University School of Medicine; C.S. Mott Center; Detroit, MI USA; [3]School of Computer Science & Genome Quebec Innovation Centre; McGill University; Quebec, Canada; [4]Department of Bioinformatics; McGill University; Quebec, Canada; [5]Department of Computer Sciences; Wayne State University; Detroit, MI USA

Keywords: personal genomes, *Drosophila melanogaster*, whole-genome SNP analysis, next generation DNA sequencing

We describe a new computer program, SnpEff, for rapidly categorizing the effects of single nucleotide polymorphisms (SNPs) and other variants such as multiple nucleotide polymorphism (MNPs) and insertion-deletions (InDels), in whole

ave been assigned, the citation will change accordingly.

http://snpeff.sourceforge.net/

Variant annotation - SnpEff

SnpEff input(s) :

- vcf file
- (annotation => over 2500 genomes pre-built databases / build one yourself)

SnpEff outputs :

- html report
- vcf file (information added to the INFO fields)

Table 4. Information provided by SnpEff in variant call format (VCF)

Sub-field	Notes
Effect	Effect of this variant. See details below
Codon_Change	Codon change: old_codon/new_codon
Amino_Acid_change	Amino acid change: old_AA/new_AA
Warnings	Any warnings or errors
Gene_name	Gene name
Gene_BioType	BioType, as reported by ENSEMBL
Coding	[CODING | NON_CODING]. If information reported by ENSEMBL (e.g., has 'protein_id' information in GTF file)
Trancript	Transcript ID (usually ENSEMBL)
Exon	Exon ID (usually ENSEMBL)
Warnings	Any warnings or errors (not shown if empty)

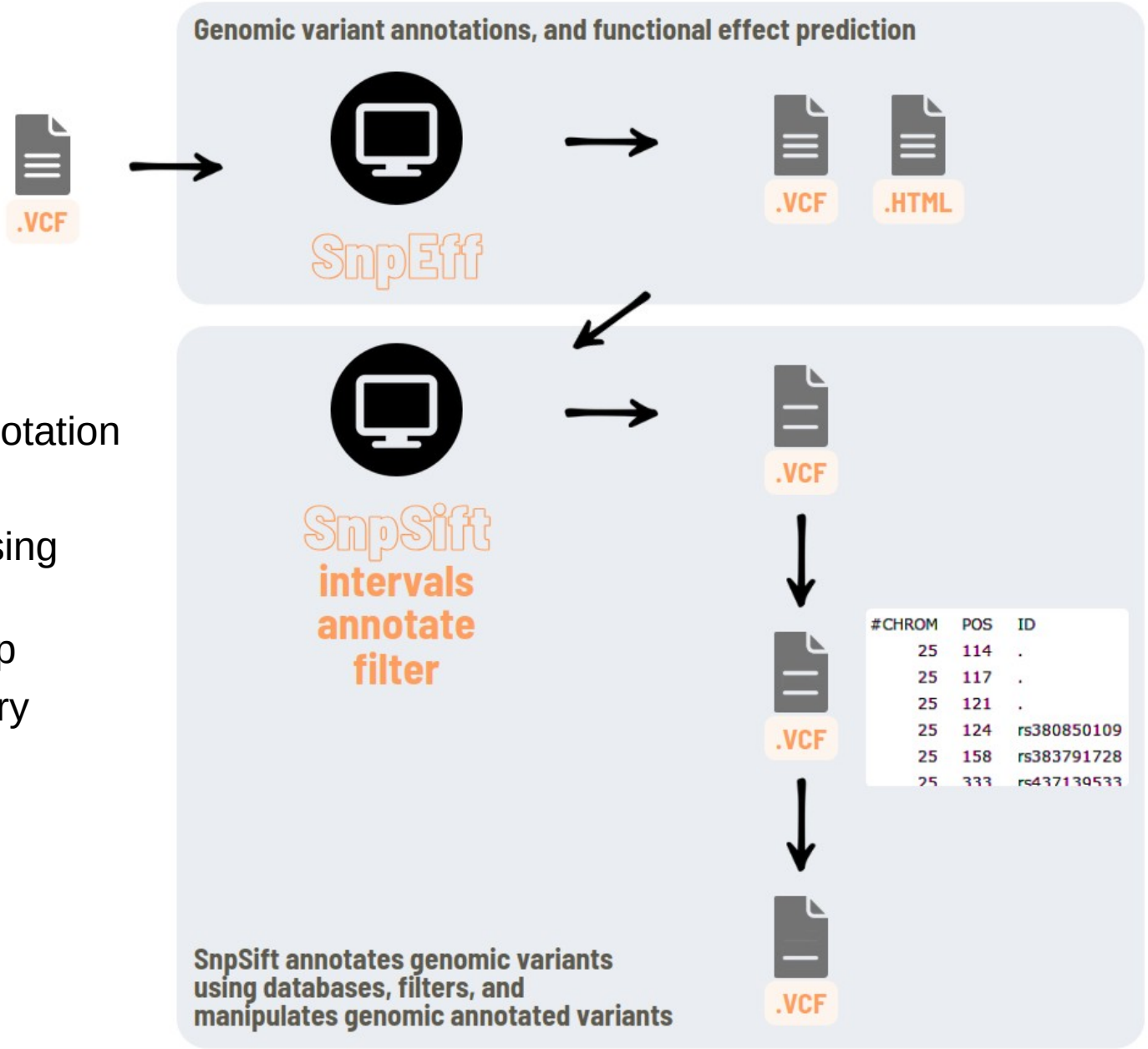
The information is added to the INFO fields using an tag 'EFF'. The format for each effect is "Effect (Effect_Impact \| Codon_Change \| Amino_Acid_change \| Gene_Name \| Gene_BioType \| Coding \| Transcript \| Exon [\| ERRORS \| WARNINGS])".

Variant annotation – pipeline

SnpEff : Variant effect and annotation

SnpSift

- Intervals - Filter variants using intervals
- Annotate SNPs from dbSnp
- Filter variants using arbitrary expressions



SnpSift - Filter

Filter variants using arbitrary expressions

http://snpeff.sourceforge.net/SnpSift.html#filter

Some examples:

I want to filter out samples with quality less than 30:
(QUAL > 30)
...but we also want InDels that have quality 20 or more:
((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
...or any homozygous variant present in more than 3 samples:
(countHom() > 3) | ((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
...or any heterozygous sample with coverage 25 or more:
((countHet() > 0) & (DP >= 25)) | (countHom() > 3) | ((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
I want to keep samples where the genotype for the first sample is homozygous variant and the genotype for the second sample is reference:
isHom(GEN[0]) & isVariant(GEN[0]) & isRef(GEN[1])

```
#CHR  POS   ID    REF   ALT   QUAL  FILTER     [INFOS]    FORMAT          SAMPLE_1                 SAMPLE_2
chr1  7     .     C     T     247.82     .     [INFOS]    GT:AD:DP:GQ:PL  0/1:2,3:5:9.2:20,0,15     0/1:2,3:5:9.01:10,1,6

chr1  19    .     G     A     124.34     .     [INFOS]    GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

SnpSift Filter – Variables

- **Fields** names: "CHROM, POS, ID, REF, ALT, QUAL or FILTER" Examples:
  - Any variant in chromosome 1:

```
"( CHROM = 'chr1' )"
```

  - Variants between two positions:

```
"( POS > 123456 ) & ( POS < 654321 )"
```

  - Has an ID and it matches the regulat expression 'rs':

```
"(exists ID) & ( ID =~ 'rs' )"
```

  - The reference is 'A':

```
"( REF = 'A' )"
```

  - The alternative is 'T':

```
"( ALT = 'T' )"
```

  - Quality over 30:

```
"( QUAL > 30 )"
```

  - Filter value is either 'PASS' or it is missing:

```
"( na FILTER ) | (FILTER = 'PASS')"
```

- **INFO field** names in the INFO field. E.g. if the info field has "DP=48;AF1=0;..." you can use something like

```
( DP > 10 ) & ( AF1 = 0 )
```

SnpSift - Filter on genotype fileds

Genotype fields

Vcf genotype fields can be accessed individually using array notation.

- **Genotype fields** are accessed using an index (sample number) followed by a variable name. E.g. If the genotypes are "GT:PL:GQ 1/1:255,66,0:63 0/1:245,0,255:99" You can write something like

```
"( GEN[0].GQ > 60 ) & ( GEN[1].GQ > 90 )"
```

  You may use an asterisk to represent 'ANY' field

```
"( GEN[*].GQ > 60 )"
```

- **Genotype multiple fields** are accessed using an index (sample number) followed by a variable name and then another index. E.g. If the genotypes are "GT:PL:GQ 1/1:255,66,0:63 0/1:245,0,255:99" You can write something like

```
"( GEN[0].PL[2] = 0 )"
```

  You may use an asterisk to represent 'ANY' field

```
"( GEN[0].PL[*] = 0 )"
```

  ...or even

```
"( GEN[*].PL[*] = 0 )"
```

Rq : You can create an expression using sample names instead of genotype numbers !

SnpSift - Filter on EFF fileds

SnpEff 'EFF' fields

SnpEff annotations are parsed, so you can access individual sub-fields:
Effect fields (from SnpEff) are accessed using an index (effect number) followed by a sub-field name.
Available `EFF` sub-fields are:

- EFFECT: Effect (e.g. SYNONYMOUS_CODING, NON_SYNONYMOUS_CODING, FRAME_SHIFT, etc.)
- IMPACT: { HIGH, MODERATE, LOW, MODIFIER }
- FUNCLASS: { NONE, SILENT, MISSENSE, NONSENSE }
- CODON: Codon change (e.g. 'ggT/ggG')
- AA: Amino acid change (e.g. 'G156')
- GENE: Gene name (e.g. 'PSD3')
- BIOTYPE: Gene biotype, as described by the annotations (e.g. 'protein_coding')
- CODING: Gene is { CODING, NON_CODING }
- TRID: Transcript ID
- RANK: Exon or Intron rank (i.e. exon number in a transcript)

For example, you may want only the lines where the first effect is a NON_SYNONYMOUS variants:

```
"( EFF[0].EFFECT = 'NON_SYNONYMOUS_CODING' )"
```

...but this probably doesn't make much sense. What you may really want are lines where ANY effect is NON_SYNONYMOUS:

```
"( EFF[*].EFFECT = 'NON_SYNONYMOUS_CODING' )"
```

May be you want only the ones that affect gene 'TCF7L2'

```
"( EFF[*].EFFECT = 'NON_SYNONYMOUS_CODING' ) &  ( EFF[*].GENE = 'TCF7L2' )"
```

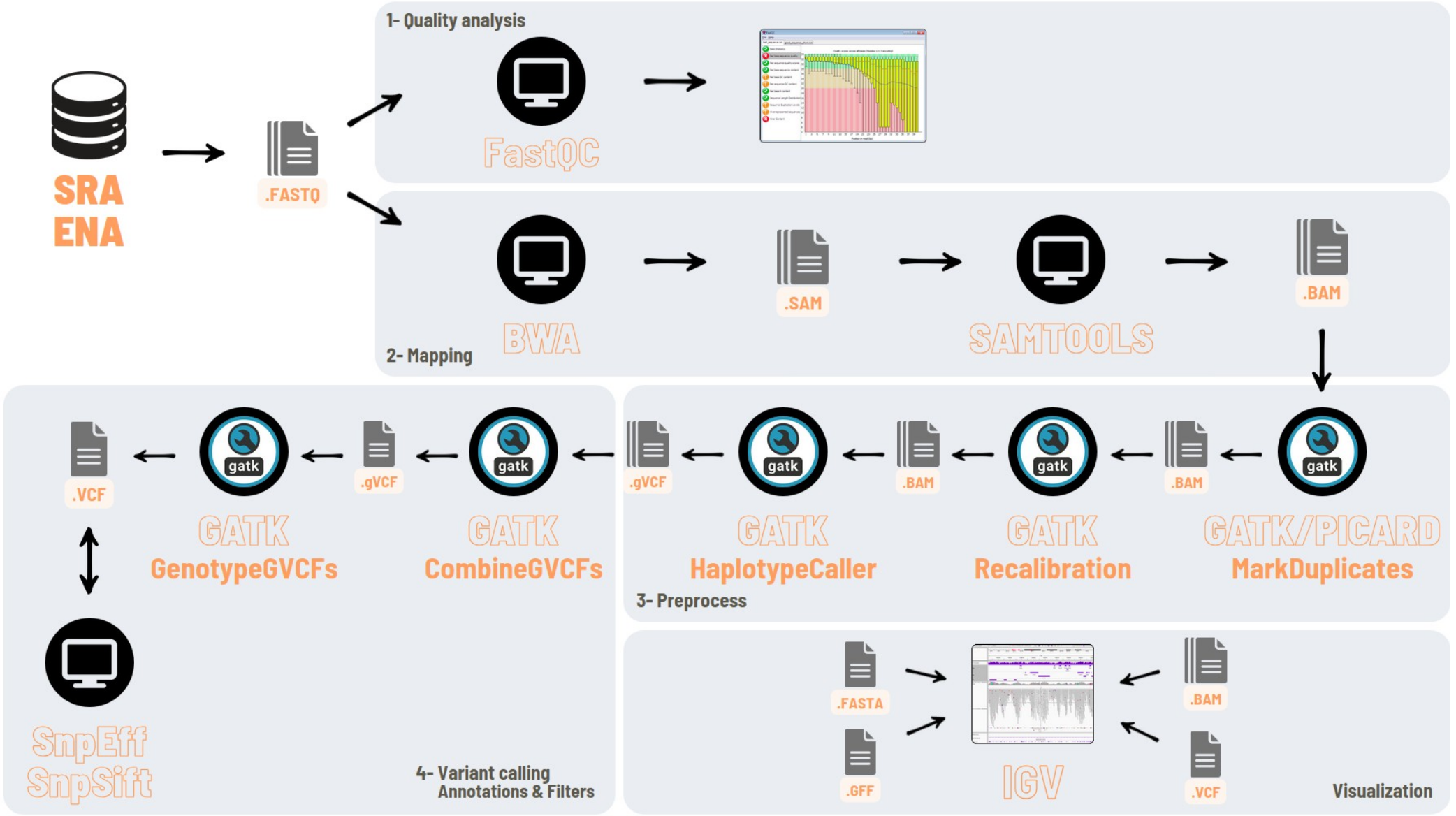
SnpSift Filter - Available operands and functions

Operand	Description	Data type	Example
=	Equality test	FLOAT, INT or STRING	(REF **=** 'A')
>	Greater than	FLOAT or INT	(DP **>** 20)
≥	Greater or equal than	FLOAT or INT	(DP **≥** 20)
<	Less than	FLOAT or INT	(DP **<** 20)
≤	Less or equal than	FLOAT or INT	(DP **≤** 20)
=~	Match regular expression	STRING	(REL **=~** 'AC')
!~	Does not match regular expression	STRING	(REL **!~** 'AC')
&	AND operator	Boolean	(DP > 20) **&** (REF = 'A')
|	OR operator	Boolean	(DP > 20) | (REF = 'A')
!	NOT operator	Boolean	**!** (DP > 20)
exists	The variable exists (not missing)	Any	(**exists** INDEL)
has	The right hand side expression is equalt to any of the items in a list consisting of separating the left hand side expression using delimiters: '&', '+', ':', ',', ':', '(', ')' ,'[', ']' Example: If the expression is: ANN[*].EFFECT **has** 'missense_variant' If left hand side (ANN[*].EFFECT) has value 'missense_variant&splice_region_variant', then it is transformed to a list: ['missense_variant', 'splice_region_variant'] Since the right hand side ('missense_variant') is in the list, the expression evaulates to 'true'	Any	(ANN[*].EFFECT **has** 'missense_variant')

SnpSift Filter - Available operands and functions

Function	Description	Data type	Example
countHom()	Count number of homozygous genotypes	No arguments	(**countHom()** > 0)
countHet()	Count number of heterozygous genotypes	No arguments	(**countHet()** > 2)
countVariant()	Count number of genotypes that are variants (i.e. not reference 0/0)	No arguments	(**countVariant()** > 5)
countRef()	Count number of genotypes that are NOT variants (i.e. reference 0/0)	No arguments	(**countRef()** < 1)
Genotype Function	**Description**	**Data type**	**Example**
isHom	Is homozygous genotype?	Genotype	**isHom(GEN[0])**
isHet	Is heterozygous genotype?	Genotype	**isHet(GEN[0])**
isVariant	Is genotype a variant? (i.e. not reference 0/0)	Genotype	**isVariant(GEN[0])**
isRef	Is genotype a reference? (i.e. 0/0)	Genotype	**isRef(GEN[0])**

Synthesis


1- Quality analysis
SRA
ENA
.FASTQ
FastQC
2- Mapping
BWA
.SAM
SAMTOOLS
.BAM
GATK/PICARD
MarkDuplicates
GATK
Recalibration
.BAM
GATK
HaplotypeCaller
.gVCF
3- Preprocess
GATK
CombineGVCFs
.gVCF
GATK
GenotypeGVCFs
.VCF
SnpEff
SnpSift
4- Variant calling
Annotations & Filters
.FASTA
.GFF
IGV
.BAM
.VCF
Visualization

Exercise / set 6

Variant recalibration

Resources for SNPs

- *True sites training resource: HapMap*
  This resource is a SNP call set that has been validated to a very high degree of confidence. The program will consider that the variants in this resource are representative of true sites (truth=true), and will use them to train the recalibration model (training=true). We will also use these sites later on to choose a threshold for filtering variants based on sensitivity to truth sites. The prior likelihood we assign to these variants is Q15 (96.84%).
- *True sites training resource: Omni*
  This resource is a set of polymorphic SNP sites produced by the Omni geno- typing array. The program will consider that the variants in this resource are representative of true sites (truth=true), and will use them to train the recalibration model (training=true). The prior likelihood we assign to these variants is Q12 (93.69%).
- *Non-true sites training resource: 1000G*
  This resource is a set of high-confidence SNP sites produced by the 1000 Genomes Project. The program will consider that the variants in this re- source may contain true variants as well as false positives (truth=false), and will use them to train the recalibration model (training=true). The prior likelihood we assign to these variants is Q10 (90%).
- *Known sites resource, not used in training: dbSNP*
  This resource is a call set that has not been validated to a high degree of confidence (truth=false). The program will not use the variants in this resource to train the recalibration model (training=false). However, the program will use these to stratify output metrics such as Ti/Tv ratio by whether variants are present in dbsnp or not (known=true). The prior likelihood we assign to these variants is Q2 (36.90%).

http://gatkforums.broadinstitute.org/gatk/discussion/1259/which-training-sets-arguments-should-i-use-for-running-vqsr