Plateforme Bioinformatique Midi-Pyrénées

- GALAXY -
Aligning SGS reads and SNP calling

- Philippe Bardou - Sarah Maman - Sabrina Rodriguez -

Organisation

First day :

Morning (9h00 -12h00)

- Initiation Galaxy

Afternoon (13h30-17h)

- Sequence quality
- Read mapping

Second day :

Morning (9h00 -12h00)

- SAM format
- Visualisation

Afternoon (13h30-17h)

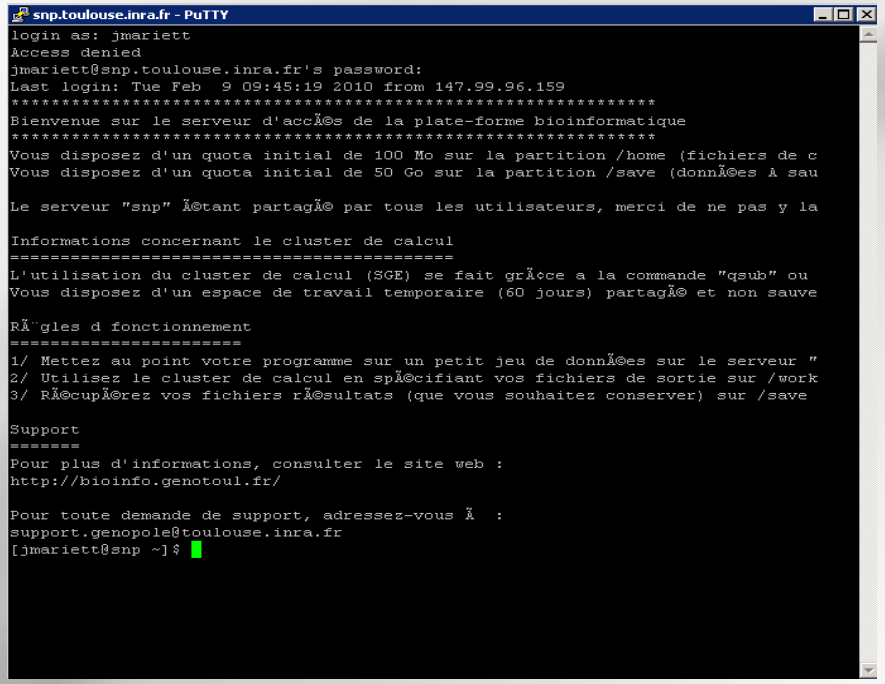
- Variant calling
- Variant annotation

What are you going to learn?

- To upload input files (extract reads and reference genome) into Galaxy
- 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 filter and annotate variants

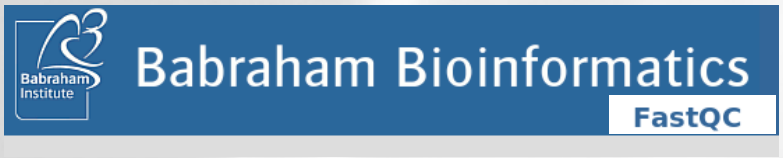
What you do not need to know?

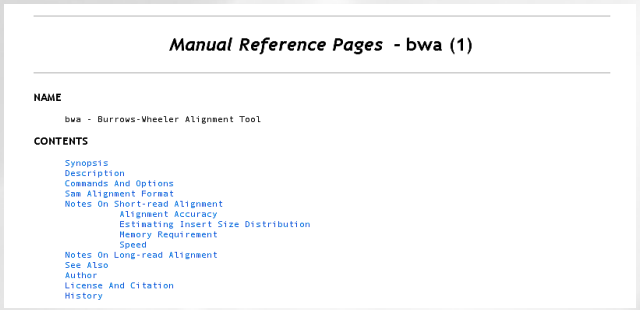
- How to connect to a remote unix server (putty)?
- What a unix command looks like?
- How to move around the unix environment?
- How to edit a file?



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 : filter, annotate, ... variants





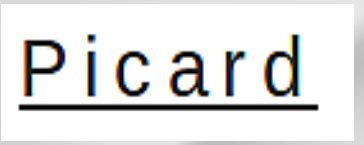
Manual Reference Pages - bwa (1)

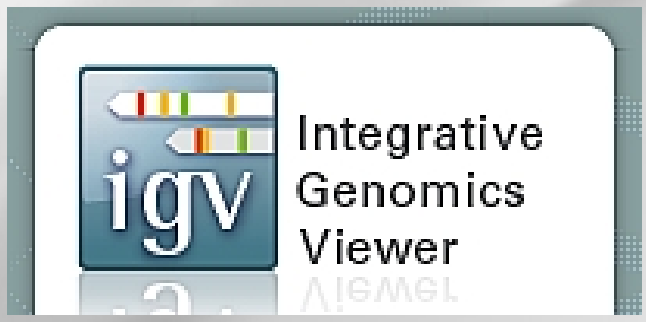
NAME

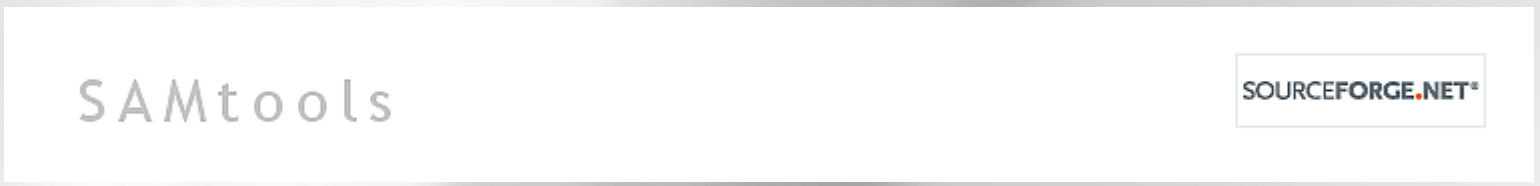
bwa - Burrows-Wheeler Alignment Tool

CONTENTS

Synopsis
Description
Commands And Options
Sam Alignment Format
Notes On Short-read Alignment
Alignment Accuracy
Estimating Insert Size Distribution
Memory Requirement
Speed
Notes On Long-read Alignment
See Also
Author
License And Citation
History

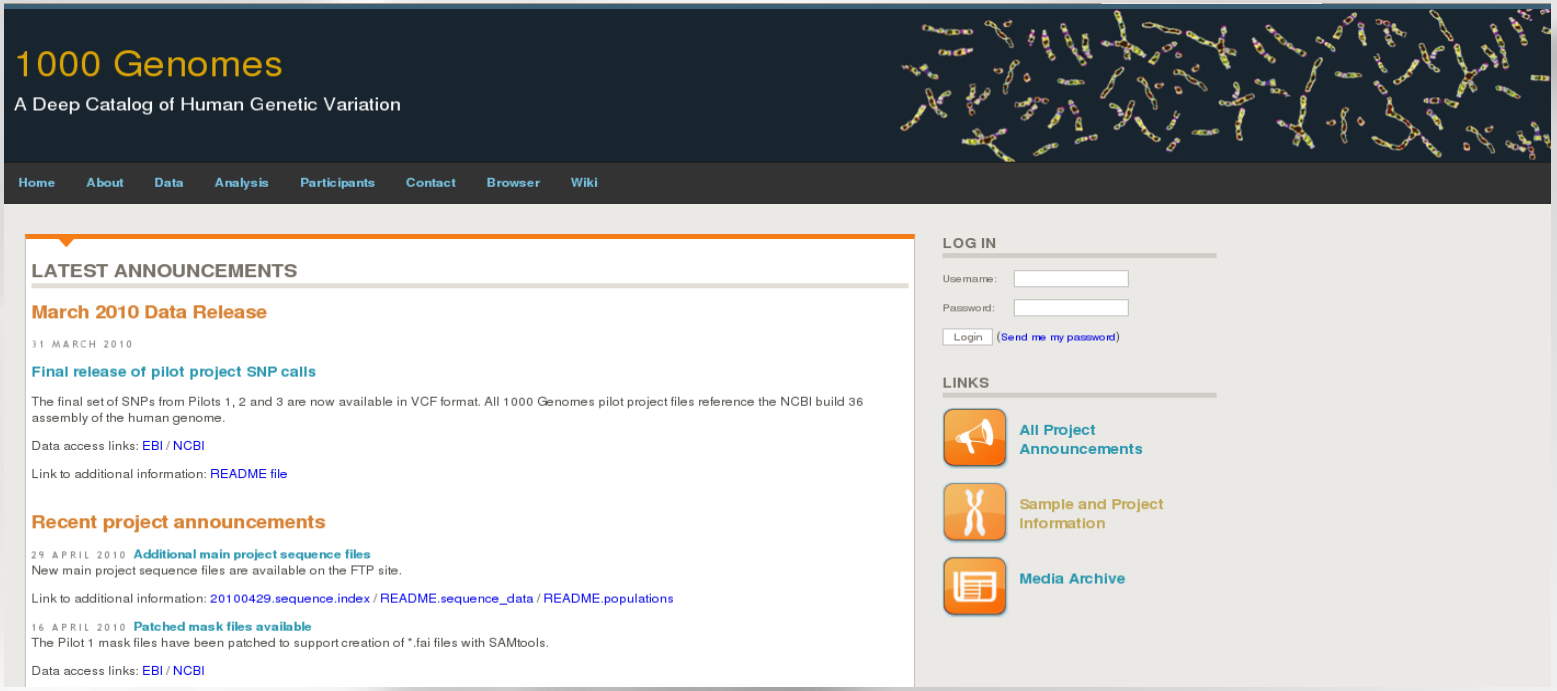






The 1000 genomes project (2007)

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



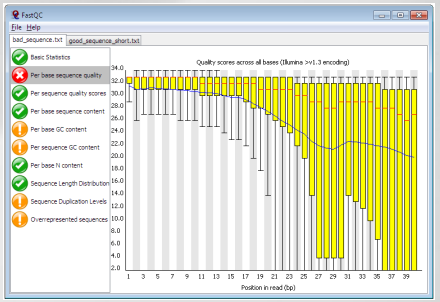
Overview

SRA
ENA
[...]

FASTQ

Quality analysis

FastQC



BWA
(aln / bwsw / mem)

Mapping

SAM

SAM manipulation

samtools (view/merge/ sort/...)

BAM

Pre-processing BAM files

Picard Rmdup/addRG

GATK realignment

BAM

GATK recalibration

BAM

GATK

BAM

Variant calling

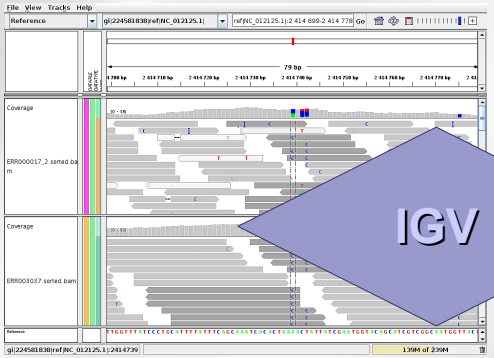
VCF

Variants annotations/filters

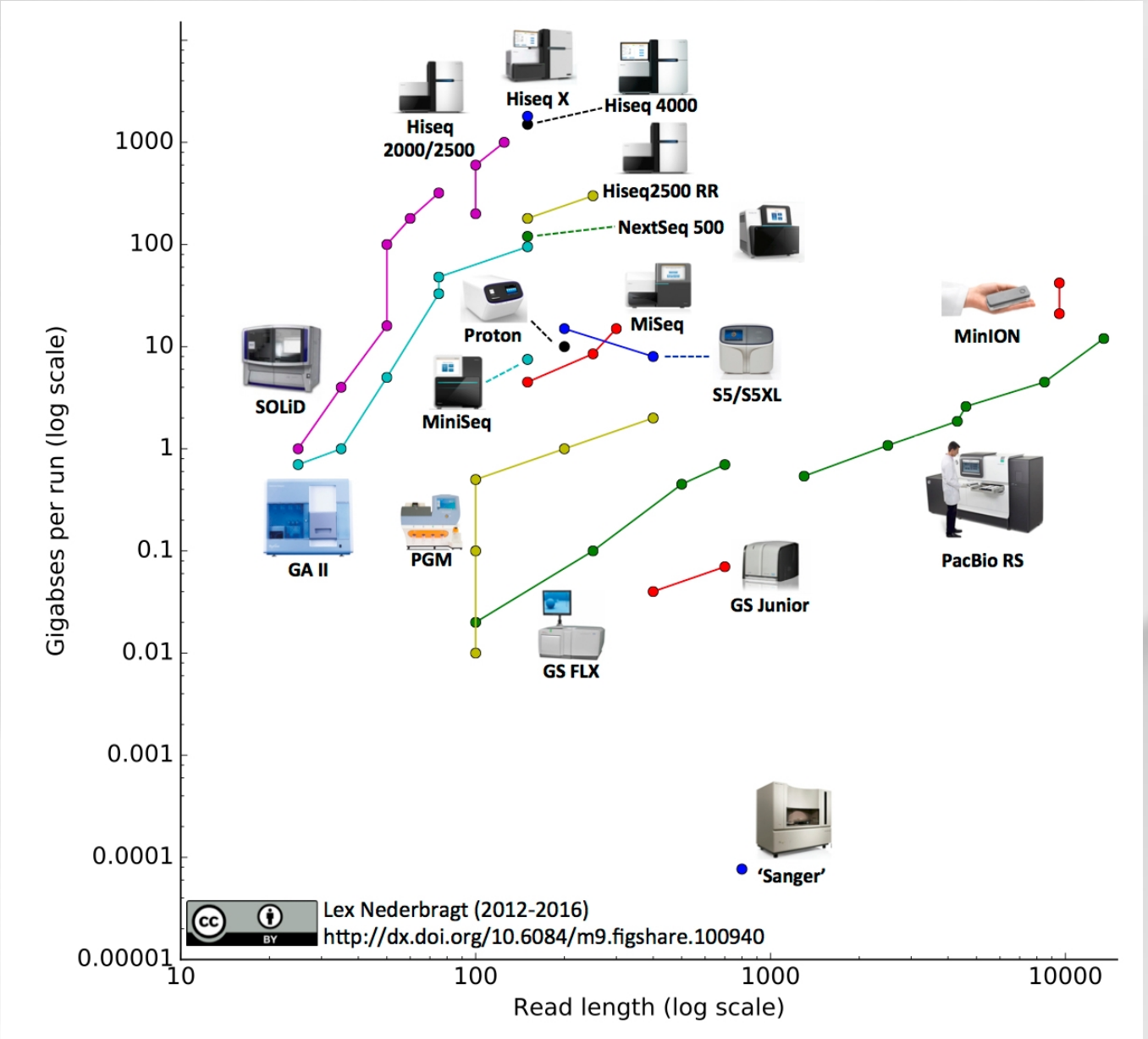
SNPeff / SNPsift

Visualization

IGV



SGS platforms



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

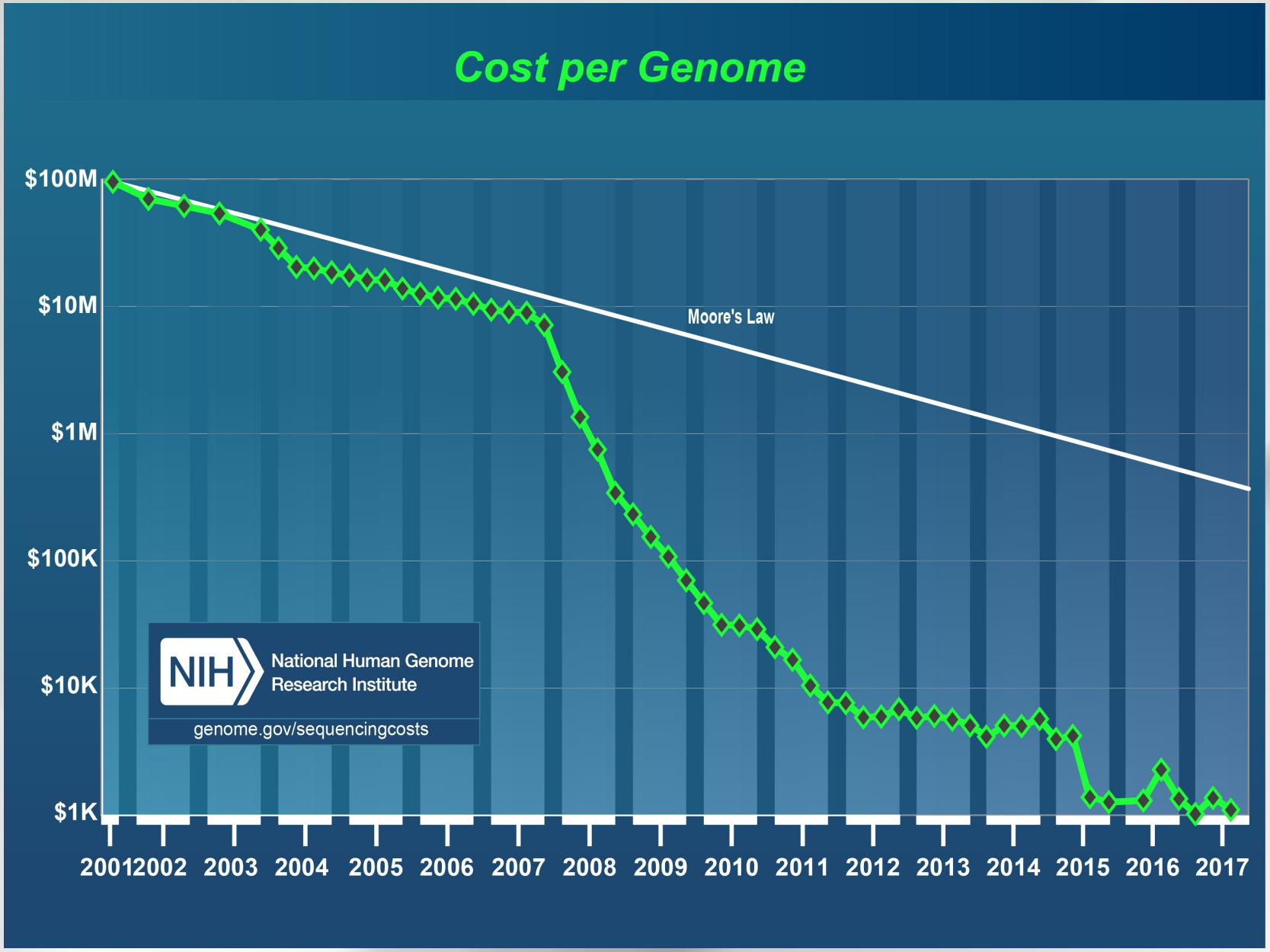
SGS platforms

>10X !




Popular Applications & Methods	NextSeq Series	HiSeq Series	NovaSeq Series	HiSeq X Series†
	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	
Exome Sequencing	●	●	●	
Targeted Gene Sequencing (amplicon, gene panel)	●	●	●	
Whole-Transcriptome Sequencing	●	●	●	
Gene Expression Profiling with mRNA-Seq	●	●	●	
miRNA & Small RNA Analysis	●	●	●	
DNA-Protein Interaction Analysis	●	●	●	
Methylation Sequencing	●	●	●	
Shotgun Metagenomics	●	●	●	

SGS platforms

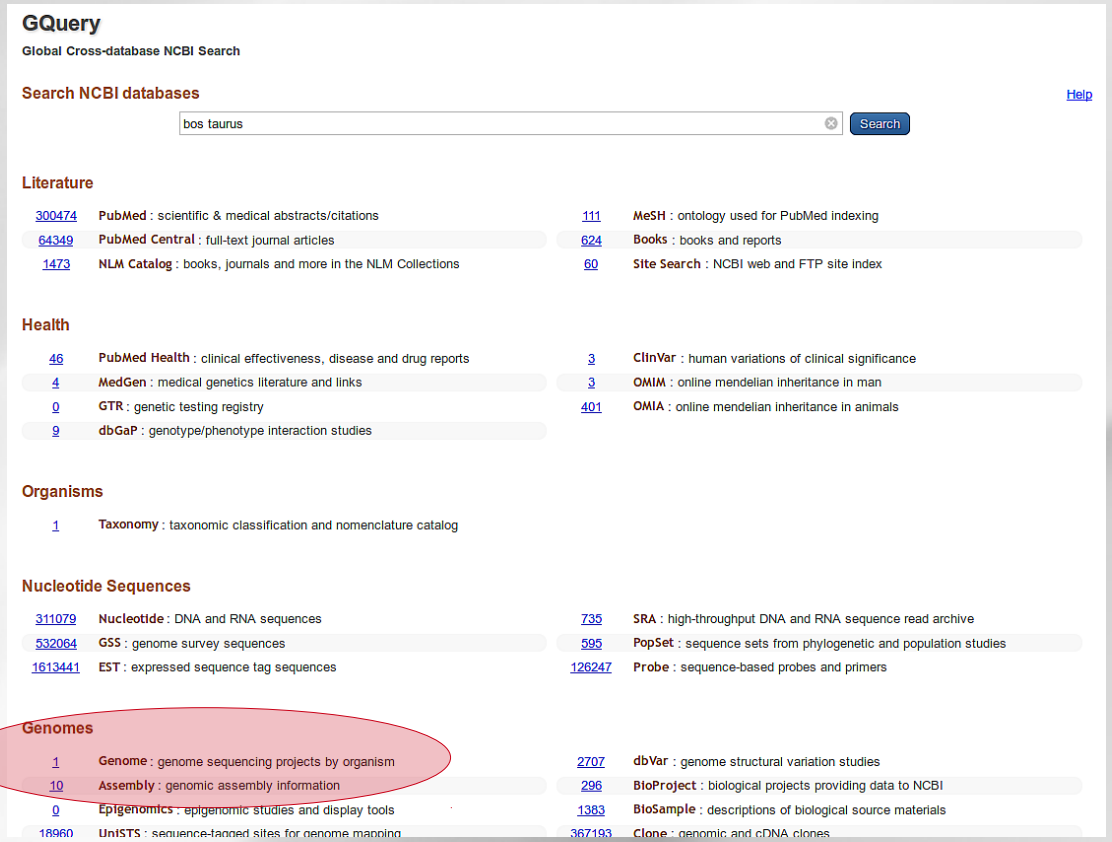


What data will we use?

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

Where to get reference genome?

- Assemble your own
- 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?

GQuery

Global Cross-database NCBI Search

Search NCBI databases

Help

ERR003037 Search

Literature

0	**PubMed** : scientific & medical abstracts/citations	0	**MeSH** : ontology used for PubMed indexing
0	**PubMed Central** : full-text journal articles	0	**Books** : books and reports
0	**NLM Catalog** : books, journals and more in the NLM Collections	0	**Site Search** : NCBI web and FTP site index

Health

0	**PubMed Health** : clinical effectiveness, disease and drug reports	0	**ClinVar** : human variations of clinical significance
0	**MedGen** : medical genetics literature and links	0	**OMIM** : online mendelian inheritance in man
0	**GTR** : genetic testing registry	0	**OMIA** : online mendelian inheritance in animals
0	**dbGaP** : genotype/phenotype interaction studies		

Organisms

0	**Taxonomy** : taxonomic classification and nomenclature catalog		

Nucleotide Sequences

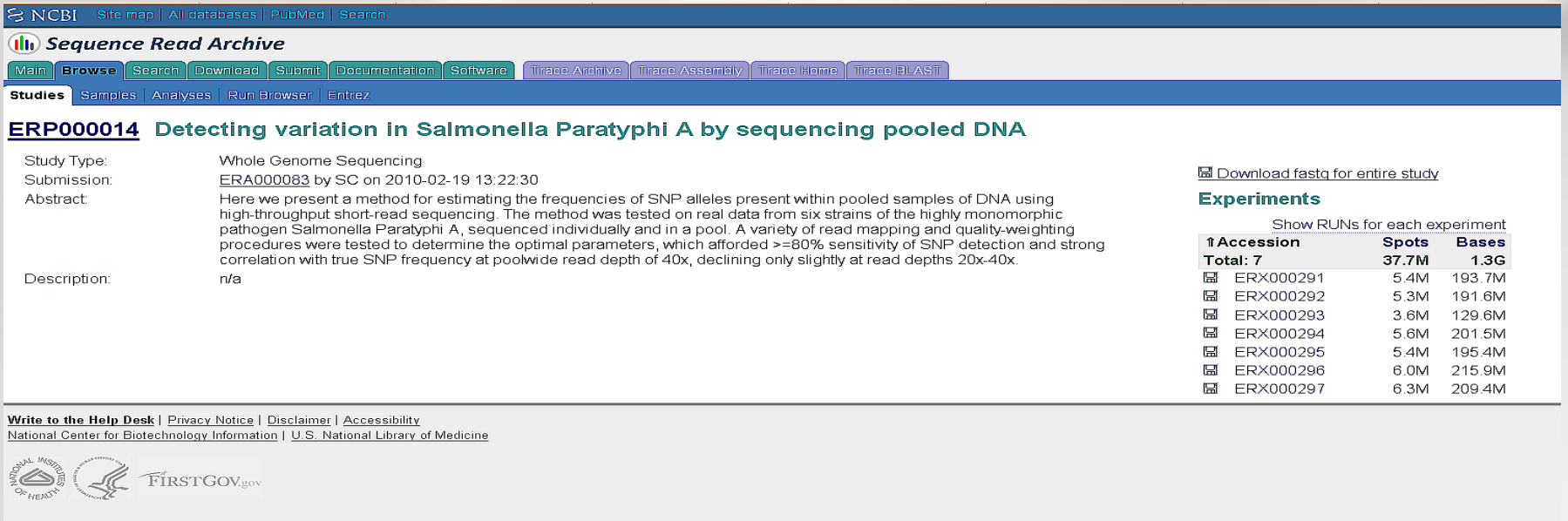
0	**Nucleotide** : DNA and RNA sequences	1	**SRA** : high-throughput DNA and RNA sequence read archive
0	**GSS** : genome survey sequences	0	**PopSet** : sequence sets from phylogenetic and population studies
0	**EST** : expressed sequence tag sequences	0	**Probe** : sequence-based probes and primers

Genomes

0	**Genome** : genome sequencing projects by organism	0	**dbVar** : genome structural variation studies

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
```

fastq file formats

Published online 16 December 2009

Nucleic Acids Research, 2010, Vol. 38, No. 6 ***1767–1771***
doi:10.1093/nar/gkp1137

SURVEY AND SUMMARY

The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants

Peter J. A. Cock[1,*], Christopher J. Fields[2], Naohisa Goto[3], Michael L. Heuer[4] and Peter M. Rice[5]

Table 1. The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

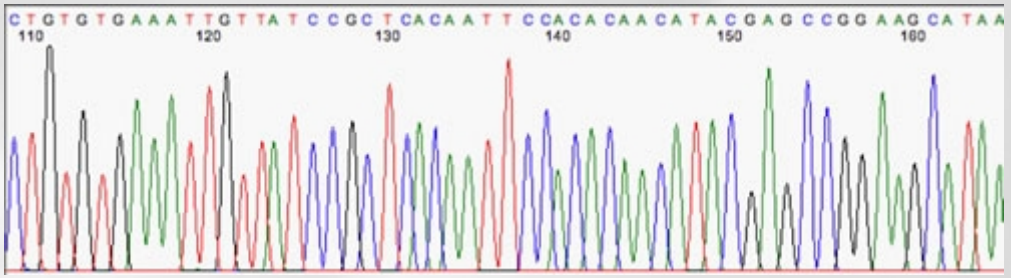
Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

$$Q_{\text{Solexa}} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

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

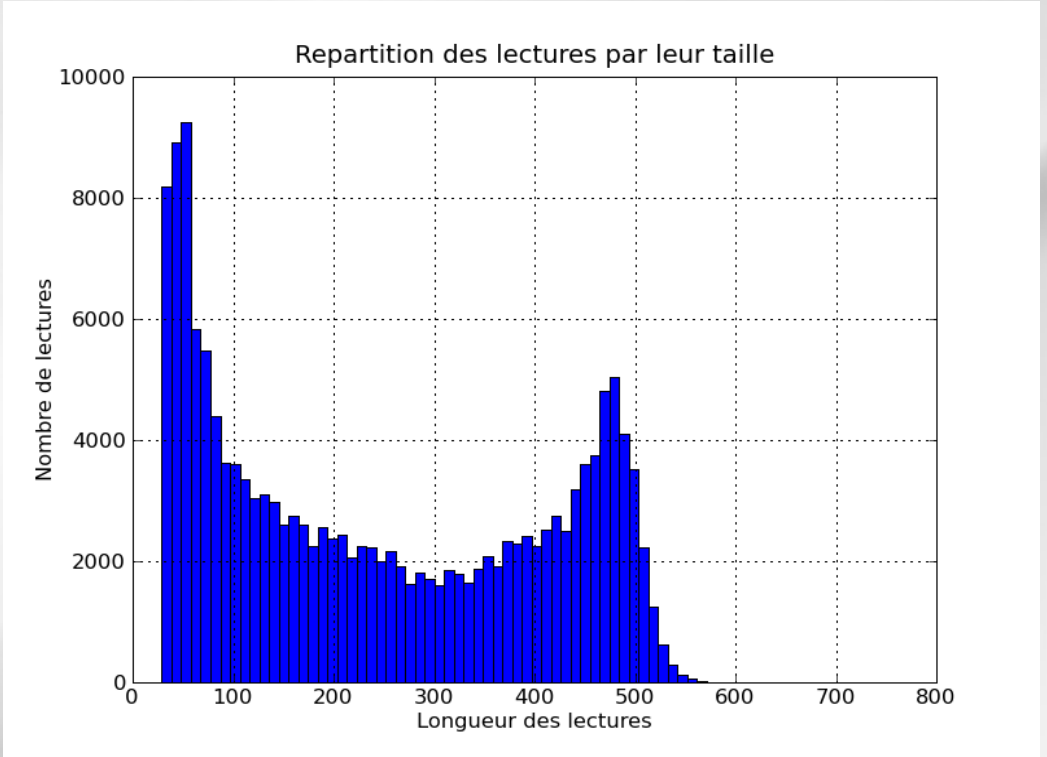
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

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

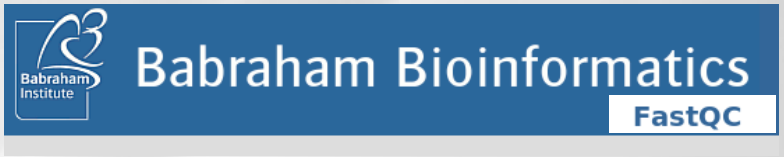
Basic reads statistics

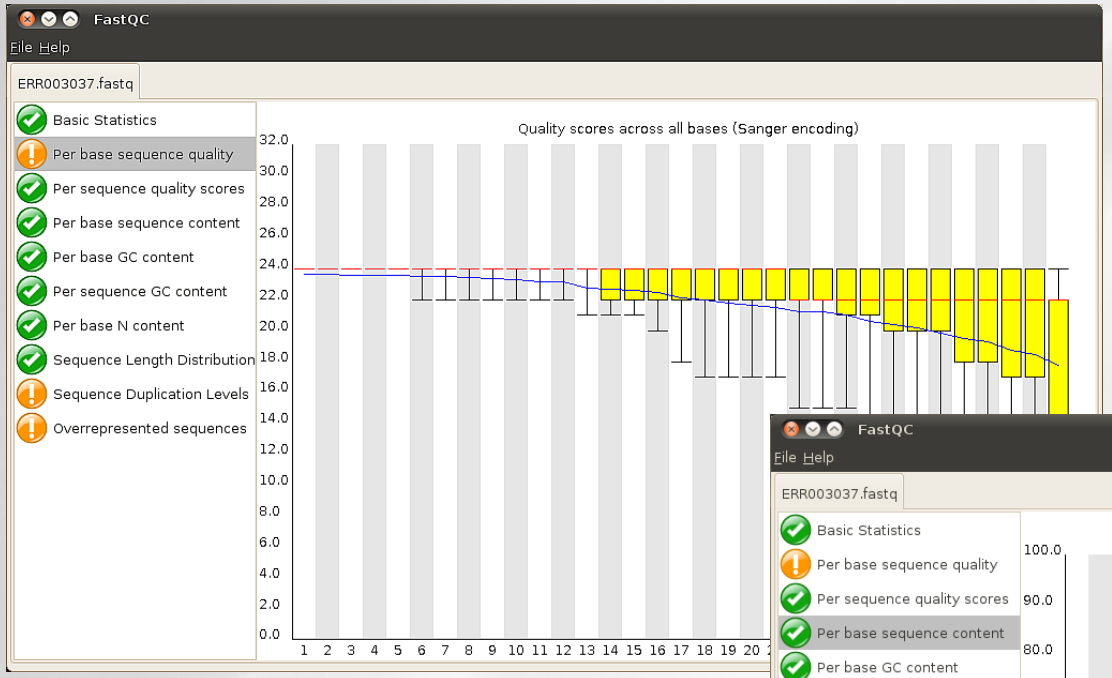
- Number of reads
- Length histogram
- Number of Ns in the reads
- Reads quality
- Reads redundancy
- Reads complexity

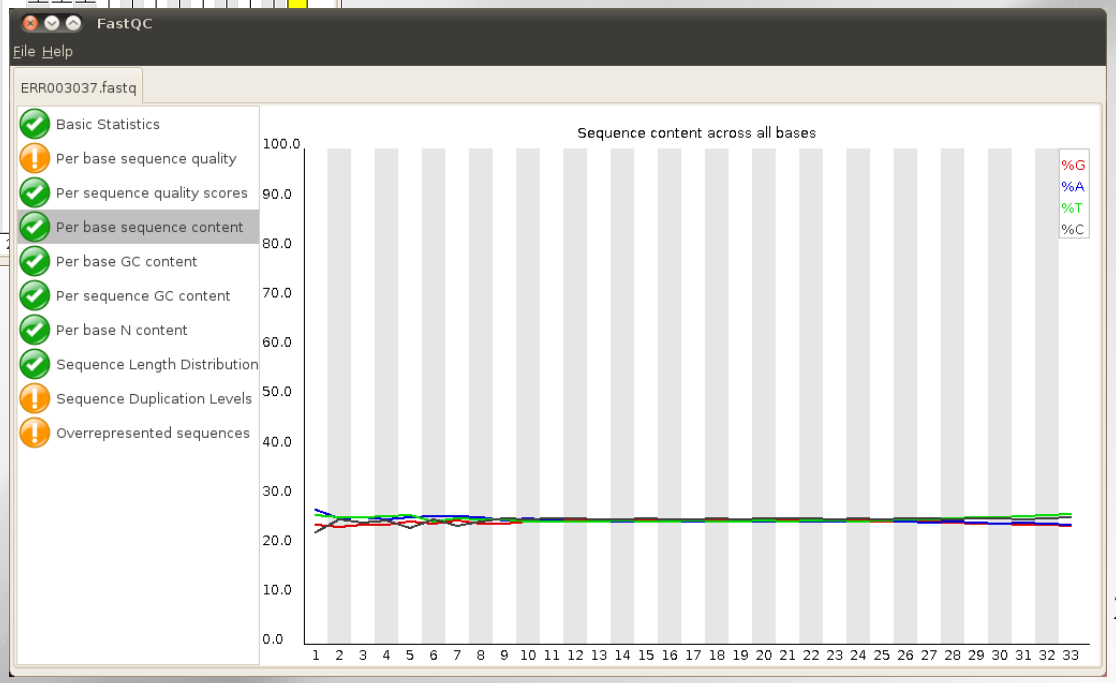


Sequence quality analysis

- FastQC :



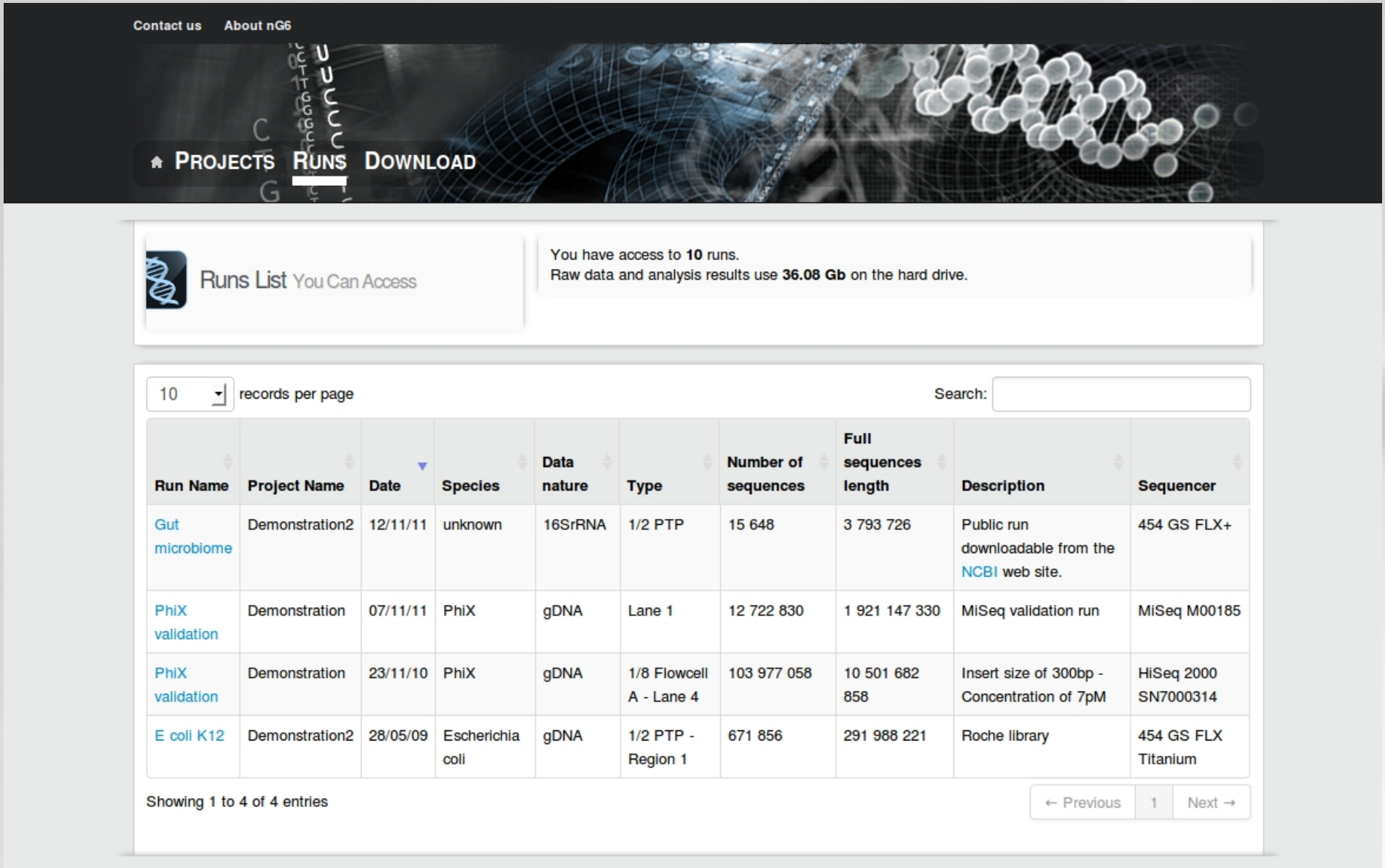




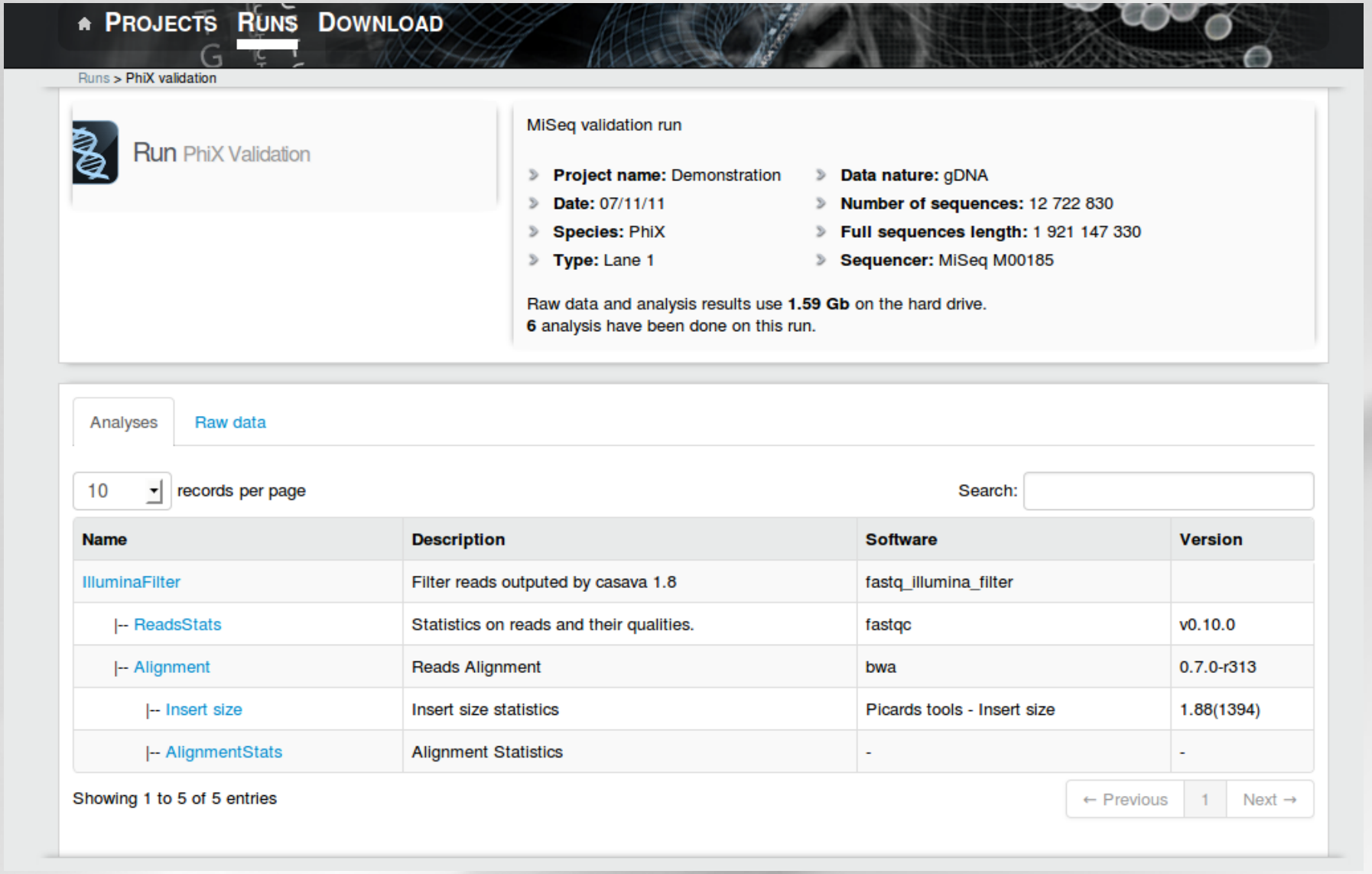
NG6 - Home


Contact us
About nG6
PROJECTS
RUNS
DOWNLOAD
User login
Enter your username and password here in order to log in on the website:
Login
Username:
User name
Password:
Password
Login
Keep up with news
June 17 2013
NG6 v2.0 is now available. This new version is based upon the jflow workflow engine instead of ergatis. This version is comming with several new features for runs administrators.
November 27 2012
NG6 is now only available in english.
September 14 2012
New user management system. 3 rights levels: administrator (in charge to run workflows), manager (in charge to manage project access) and member (browsing projects/runs/analyses).
Septembrer 9 2012
Publication of NG6 in BMC Genomics.
August 7 2012
NG6 is available in french and in english.
June 4 2012
Fix a bug in the generation of cigarline graphs produced by the AlignmentStats analysis.
November 23 2011
Since the 3rd of october 2011 the HiSeq quality encoding is in Sanger format, no longer in illumina format.
August 2nd 2011
New functionalities are available to browse your projects/runs /analyses. A hierarchical display is now available for analysis. For project administrators new functionalities are also available.
Links
GenoToul Bioinfo platform
GenoToul Genomic platform
Sigenae platform
About NG6
Next generation sequencing platforms are now well implanted in sequencing centres and some laboratories. Smaller scale machines such as the 454 junior from Roche or the MiSeq from Illumina will increase the number of laboratories hosting a sequencer. In such a context, it is important to provide these teams with an easily manageable environment to store and process the produced reads.
NG6 is a complete information system designed to answer the needs of a sequencing platform. It provides a user-friendly interface to process, store and download high-throughput sequencing data. [more]
Copyright © 2013, INRA | Designed by GenoToul Bioinfo, GenoToul Genomic and Sigenae teams | Optimized for

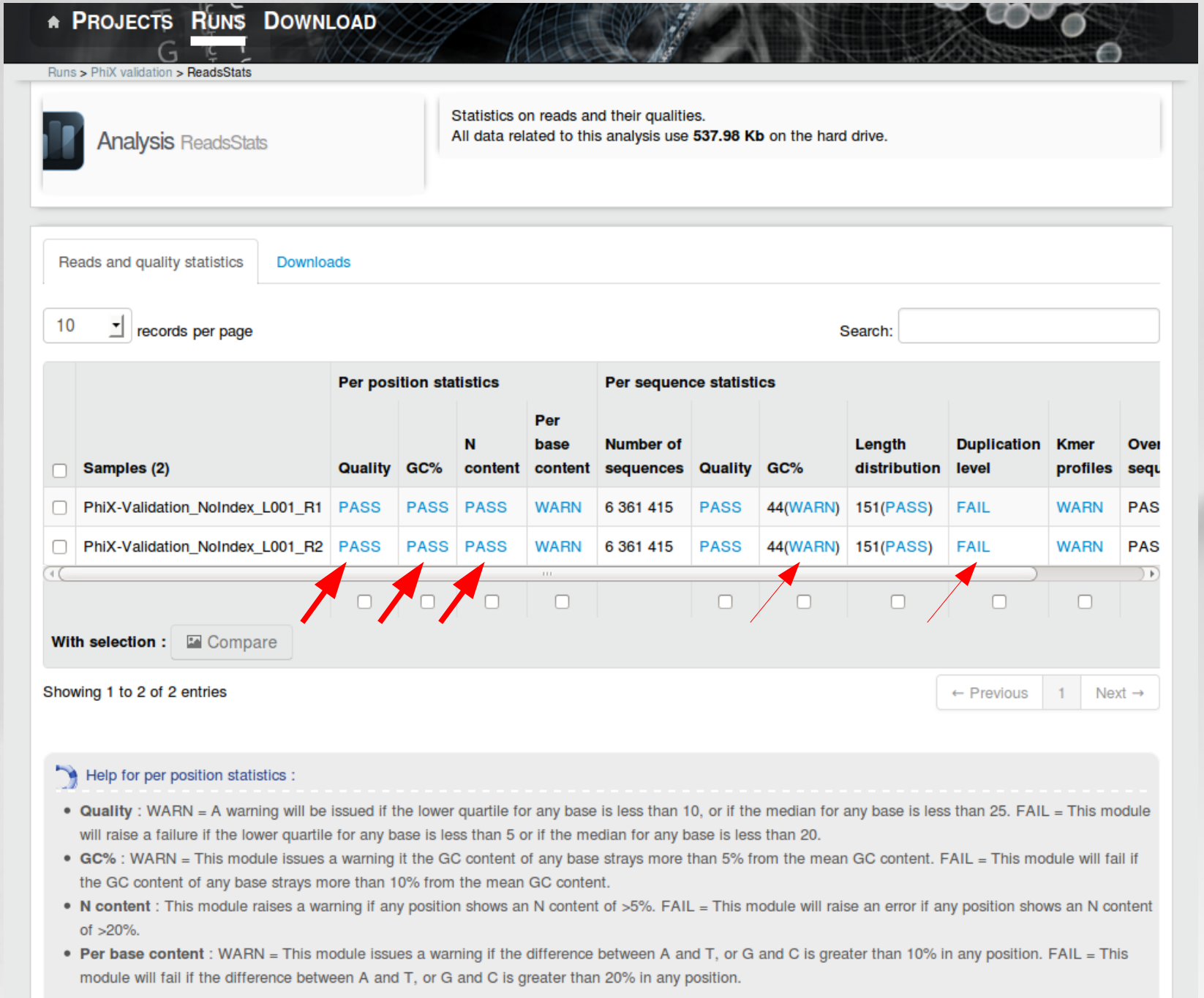
NG6 - Runs


Contact us
About nG6
PROJECTS
RUNS
DOWNLOAD
Runs List You Can Access
You have access to 10 runs.
Raw data and analysis results use 36.08 Gb on the hard drive.
10 records per page
Search:
Run Name
Project Name
Date
Species
Data nature
Type
Number of sequences
Full sequences length
Description
Sequencer
Gut microbiome
Demonstration2
12/11/11
unknown
16SrRNA
1/2 PTP
15 648
3 793 726
Public run downloadable from the NCBI web site.
454 GS FLX+
PhiX validation
Demonstration
07/11/11
PhiX
gDNA
Lane 1
12 722 830
1 921 147 330
MiSeq validation run
MiSeq M00185
PhiX validation
Demonstration
23/11/10
PhiX
gDNA
1/8 Flowcell A - Lane 4
103 977 058
10 501 682 858
Insert size of 300bp - Concentration of 7pM
HiSeq 2000 SN7000314
E coli K12
Demonstration2
28/05/09
Escherichia coli
gDNA
1/2 PTP - Region 1
671 856
291 988 221
Roche library
454 GS FLX Titanium
Showing 1 to 4 of 4 entries
← Previous
1
Next →

NG6 - Run


PROJECTS
RUNS
DOWNLOAD
Runs > PhiX validation
Run PhiX Validation
MiSeq validation run
Project name: Demonstration
Date: 07/11/11
Species: PhiX
Type: Lane 1
Data nature: gDNA
Number of sequences: 12 722 830
Full sequences length: 1 921 147 330
Sequencer: MiSeq M00185
Raw data and analysis results use 1.59 Gb on the hard drive.
6 analysis have been done on this run.
Analyses
Raw data
10 records per page
Search:
Name
Description
Software
Version
IlluminaFilter
Filter reads outputed by casava 1.8
fastq_illumina_filter
|-- ReadsStats
Statistics on reads and their qualities.
fastqc
v0.10.0
|-- Alignment
Reads Alignment
bwa
0.7.0-r313
|-- Insert size
Insert size statistics
Picards tools - Insert size
1.88(1394)
|-- AlignmentStats
Alignment Statistics
-
-
Showing 1 to 5 of 5 entries
← Previous
1
Next →

NG6 - Stats


PROJECTS
RUNS
DOWNLOAD
Runs > PhiX validation > ReadsStats
Analysis ReadsStats
Statistics on reads and their qualities.
All data related to this analysis use 537.98 Kb on the hard drive.
Reads and quality statistics
Downloads
10 records per page
Search:
Per position statistics
Per sequence statistics
Samples (2)
Quality
GC%
N content
Per base content
Number of sequences
Quality
GC%
Length distribution
Duplication level
Kmer profiles
PhiX-Validation_NoIndex_L001_R1
PASS PASS PASS WARN 6 361 415 PASS 44(WARN) 151(PASS) FAIL WARN PAS
PhiX-Validation_NoIndex_L001_R2
PASS PASS PASS WARN 6 361 415 PASS 44(WARN) 151(PASS) FAIL WARN PAS
With selection :
Compare
Showing 1 to 2 of 2 entries
← Previous
1
Next →
Help for per position statistics :
Quality : WARN = A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. FAIL = This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.
GC% : WARN = This module issues a warning it the GC content of any base strays more than 5% from the mean GC content. FAIL = This module will fail if the GC content of any base strays more than 10% from the mean GC content.
N content : This module raises a warning if any position shows an N content of >5%. FAIL = This module will raise an error if any position shows an N content of >20%.
Per base content : WARN = This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position. FAIL = This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

Exercises / set 1

- *Connection to the Galaxy website*
- *Get required data*
- *Read Quality Control*

Reads alignment

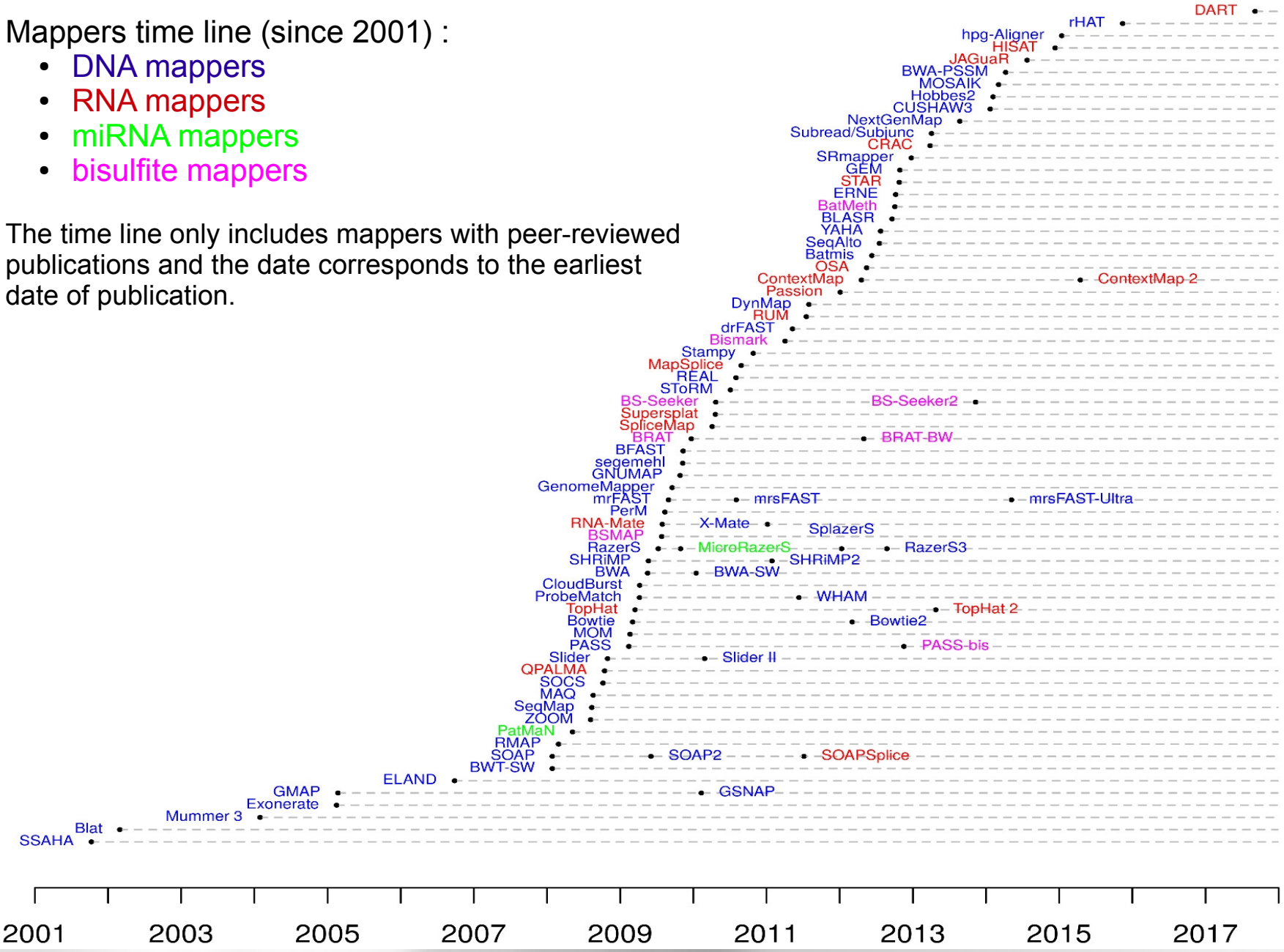
- The different software generations :
  - Smith-Waterman / Needleman-Wunch (1970)
  - BLAST (1990)
  - MAQ (2008)
  - BWA (2009)

Reads alignment

Mappers time line (since 2001) :

- DNA mappers
- RNA mappers
- miRNA mappers
- bisulfite mappers

The time line only includes mappers with peer-reviewed publications and the date corresponds to the earliest date of publication.



http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Reads alignment

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

- bowtie : fast, works well
- bowtie2 : fast, can perform local alignments too
- BWA - Fast, allows indels, commonly used for variant calling
- Subread - Very fast, (also does splice alignment)
- STAR - Extremely fast (also does splice alignment, requires at least 30 Gb memory)

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

- Tophat (most popular)
- Subread - Very fast, (also does splice alignment)
- STAR - Extremely fast (also does splice alignment, requires at least 30 Gb memory)

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

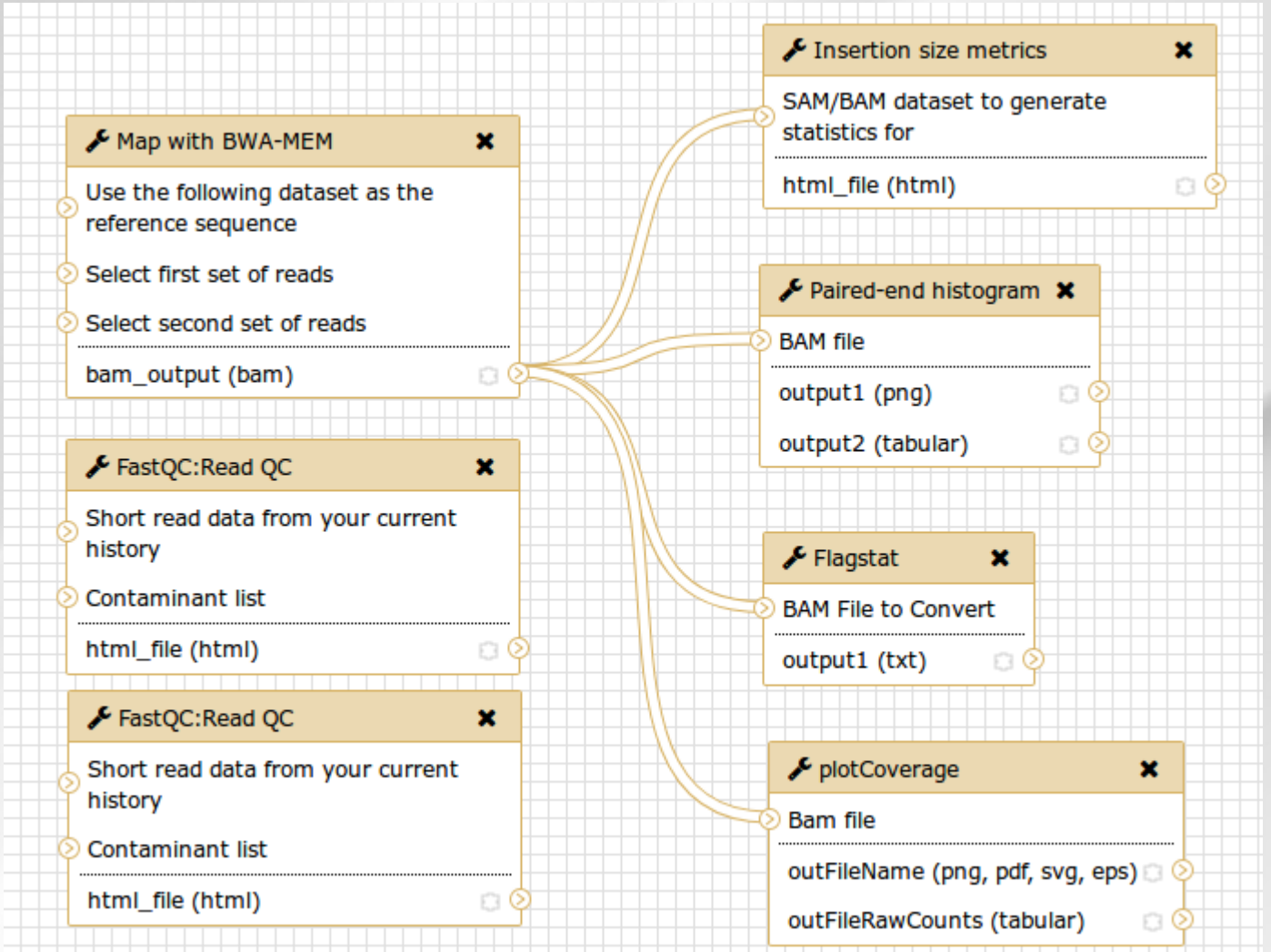
Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009
Advance Access publication May 18, 2009
Associate Editor: John Quackenbush

BWA – 3 algorithms

	mem	aln	bwasw
Read length	**>**70bp => few Mb	Short reads (~ <100bp)	Long reads (>100bp)
Genome length	More than 4Gb	< 4Gb	< 4Gb
Time processing	+++	++	-
Paired	Yes (+++)	Yes (++)	Yes (-)
Algorithm	Auto: end-to-end/SW	End-to-end	SW
Commands			
index	`bwa index ref.fasta`		
alignment	`bwa mem ref.fa r1.fq [r2.fq] > o.sam`	`bwa aln ref.fa r1.fq > o1.sai` `bwa aln ref.fa r2.fq > o2.sai`	`bwa bwasw ref.fa r1.fq [r2.fq] > o.sam`
postprocess	-	`bwa samse ref.fa o.sai r1.fq > o.sam` `bwa sampe ref.fa o1.sai o2.sai r1.fq r2.fq > o.sam`	-

Exercises / set 2

- **Aligning reads on the reference genome**
- **Get mapping statistics**



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 (454 – Solexa – Solid)
- Flexible in style, compact in size, efficient in random access

Website :

http://samtools.sourceforge.net

Paper :

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Sequence Alignment/Map (SAM) format

Aligners natively generating SAM

- BFAST, `Blat-like Fast Accurate Search Tool' for Illumina and SOLiD reads.
- Bowtie. Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in samtools-C.
- BWA, Burrows-Wheeler Aligner for short and long reads.
- GEM library. Short read aligner. Convertor provided by the developers.
- Karma, the K-tuple Alignment with Rapid Matching Algorithm.
- Mosaik. The latest version support SAM output.
- Novoalign. An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.
- SNP-o-matic, short read aligner and SNP caller.
- SOLiD BaseQV Tool. Developed by Applied Biosystems for converting SOLiD output files.
- SSAHA2 (since v2.4). Classical aligner for both short and long reads.
- Stampy, by Gerton Lunter. An accurate read aligner capable of gapped alignment for Illumina short reads. Used for indel discovery on the 1000 genomes data. Not released.
- TopHat for mapping short RNA-seq reads bridging exon junctions.

SAM format - Header section

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

```
@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 have been stored in SAM ?

http://genoweb.toulouse.inra.fr/~formation/2_Galaxy_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

Header

ERR000017_2.sam

```
@SQ     SN:ref  LN:4833080
1       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
2       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  <>>>>>>>>>>>>>>>>  XT:A:R  NM:i:2  MD:Z:5A5A6
3       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
4       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
5       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
6       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
7       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
8       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
9       16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>>>>>>>>>>>>  XT:A:R  NM:i:2  MD:Z:5A5A6
10      0   ref  4702037  25  18M     *  0  0  CTATGCAGCTATATGTTT  >>>>;>>7>>:>7>7)><  XT:A:U  NM:i:2  MD:Z:3C11G2
11      16  ref  2919865  37  18M     *  0  0  NGGTGGTTATGTCTATTC  :>>>>>>>>>>><>>>><  XT:A:U  NM:i:0  MD:Z:18
12      0   ref  2995664  37  18M     *  0  0  GTTTTGTTATGTGAATAT  ;'>>70>>7>3977+%(7  XT:A:U  NM:i:0  MD:Z:18
13      16  ref  510805   37  18M     *  0  0  ATTCTCTATGGAGTGGGT  <//>+)799+>>>>>>>>  XT:A:U  NM:i:0  MD:Z:18
14      16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>>>>>>>>>>>>  XT:A:R  NM:i:2  MD:Z:5A5A6
15      4   *    0        0   *       *  0  0  GTGCCACTCCTGGTCTTG  >8;>1;>>9-9.7/(+5$
16      16  ref  740202   0   18M     *  0  0  TTTTTTTTTTTTTTTTT  >>>>>>???????????  XT:A:R  NM:i:2  MD:Z:5A5A6
17      0   ref  1847349  37  9M1I8M  *  0  0  CATCACAATATAATCATT  >49;;>9,77>:+>6+'/  XT:A:U  NM:i:2  MD:Z:5T11
```

Alignement

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

[<TAG>:<VTYPE>:<VALUE> [...]]

```
X? : Reserved for end users
NM : Number of nuc. Difference
MD : String for mismatching positions
RG : Read group
[...]
```

```
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	Description
0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment) [1]
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped [1]
0x0010	strand of the query (0 for forward; 1 for reverse strand)
0x0020	strand of the mate [1]
0x0040	the read is the first read in a pair [1,2]
0x0080	the read is the second read in a pair [1,2]
0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
0x0200	the read fails platform/vendor quality checks
0x0400	the read is either a PCR duplicate or an optical duplicate

http://picard.sourceforge.net/explain-flags.html

SAM format - Extended CIGAR

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

op	Description
M	Alignment match (can be a sequence match or mismatch)
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

SAM format - Extended CIGAR

P	Padding (silent deletion from the padded reference sequence)

```
  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	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`.

BAM format

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

SAMtools

- Library and software package
- Creating sorted and indexed BAM files from SAM files
- Removing PCR duplicates
- Merging alignments
- Visualization of alignments from BAM files
- SNP and short INDELs detection

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

Picard

- A SAMtools complementary package
- More format conversion than SAMtools
- Visualization of alignments not available
- SNP calling & short indel detection not available

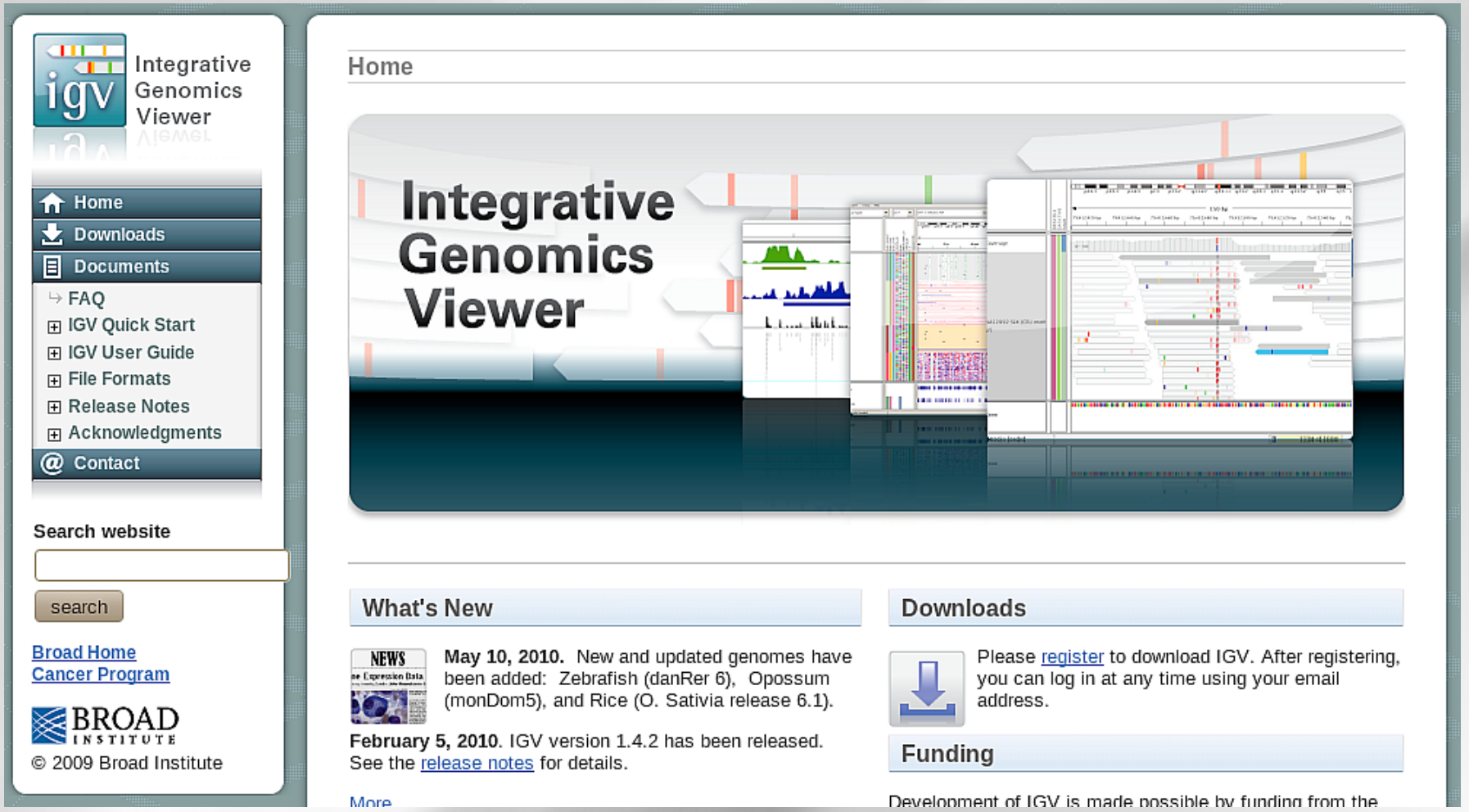
http://picard.sourceforge.net/

Exercise / set 3

- ***Manipulation of SAM files***
- ***Get informations about alignments***
- ***Run the GATK workflow***

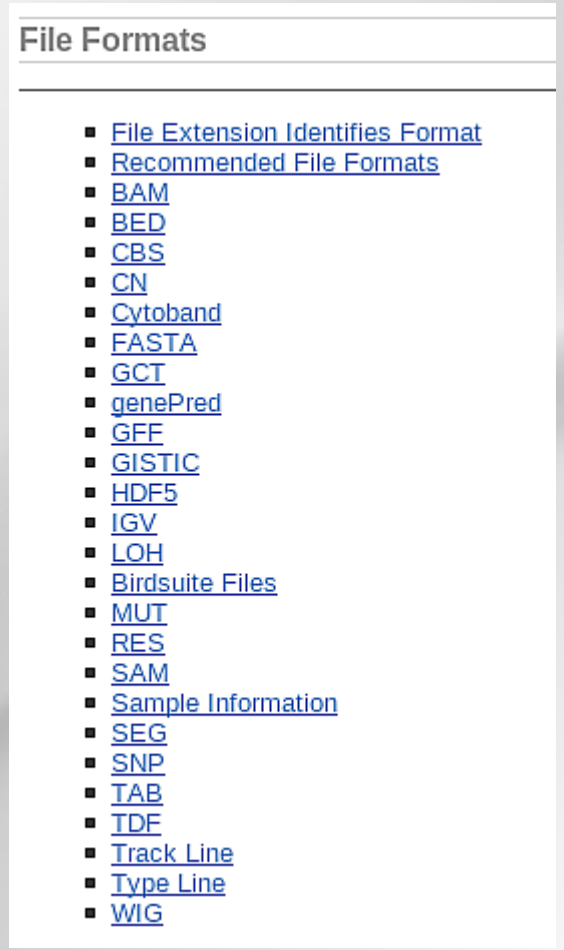
Visualizing the alignment - IGV

- IGV : Integrative Genomics Viewer
- Website : http://www.broadinstitute.org/igv



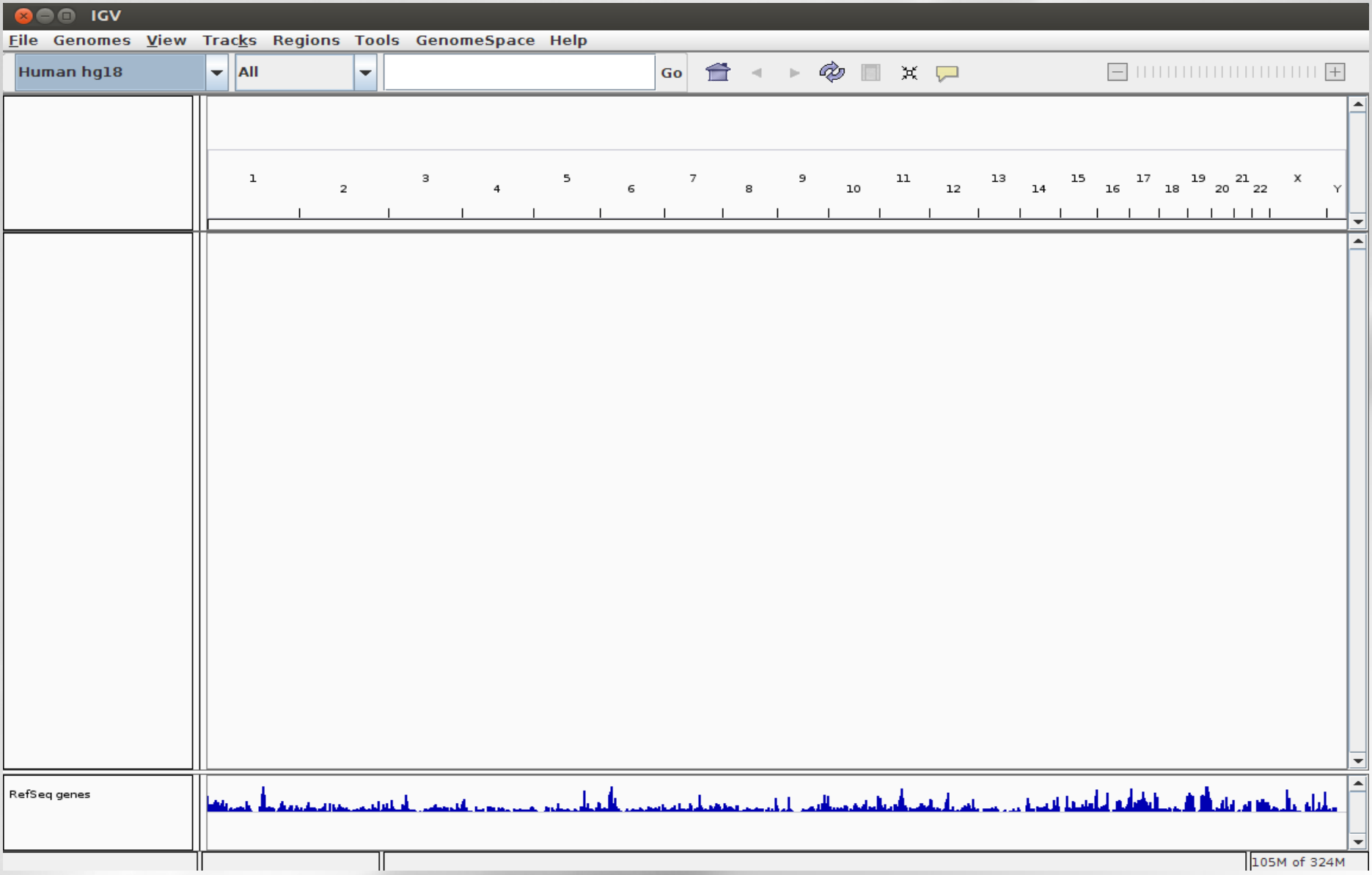
Visualizing the alignment - IGV

- 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

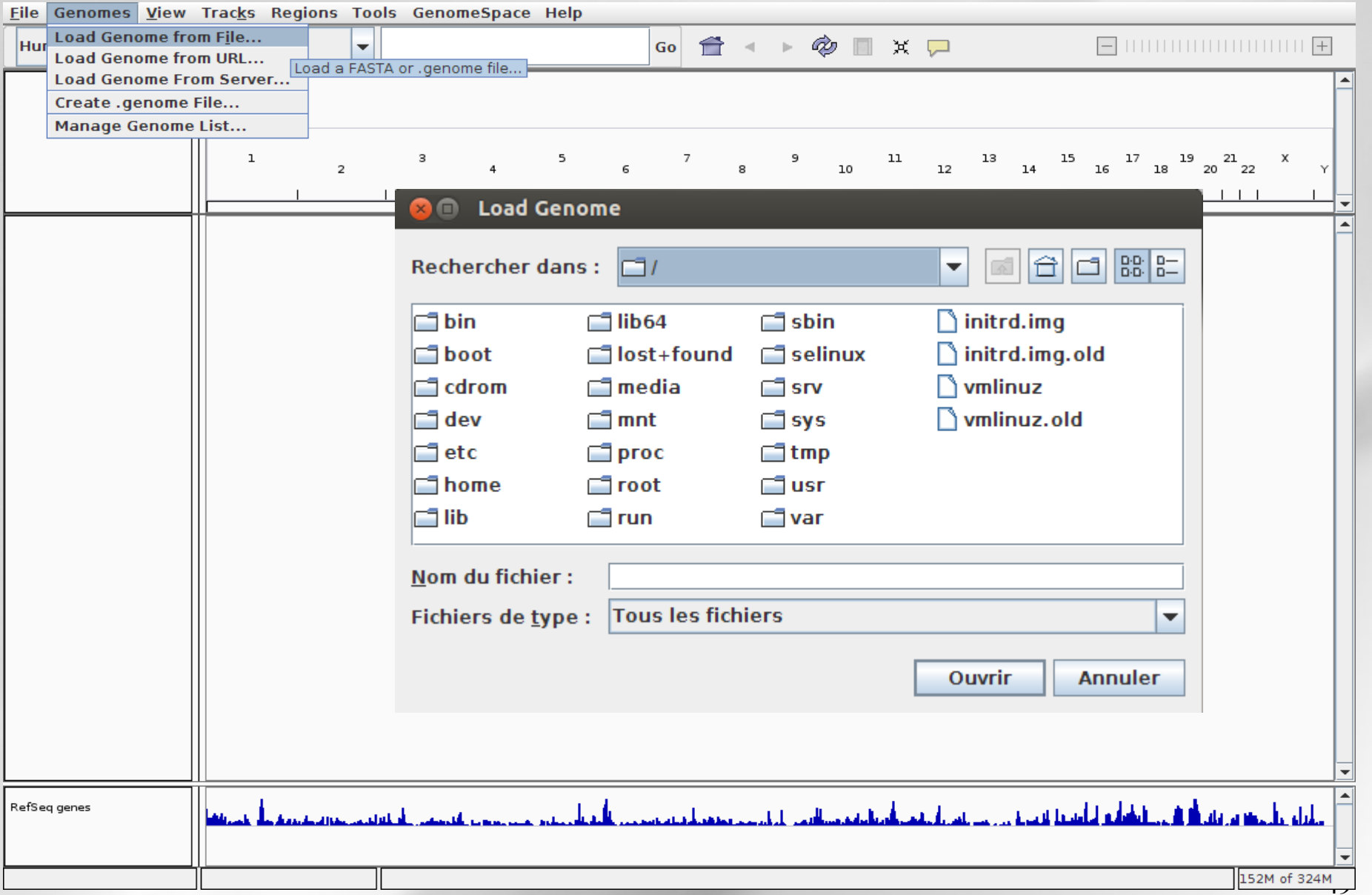

File Formats

- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- CBS
- CN
- Cytoband
- FASTA
- GCT
- genePred
- GFF
- GISTIC
- HDF5
- IGV
- LOH
- Birdsuite Files
- MUT
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- WIG

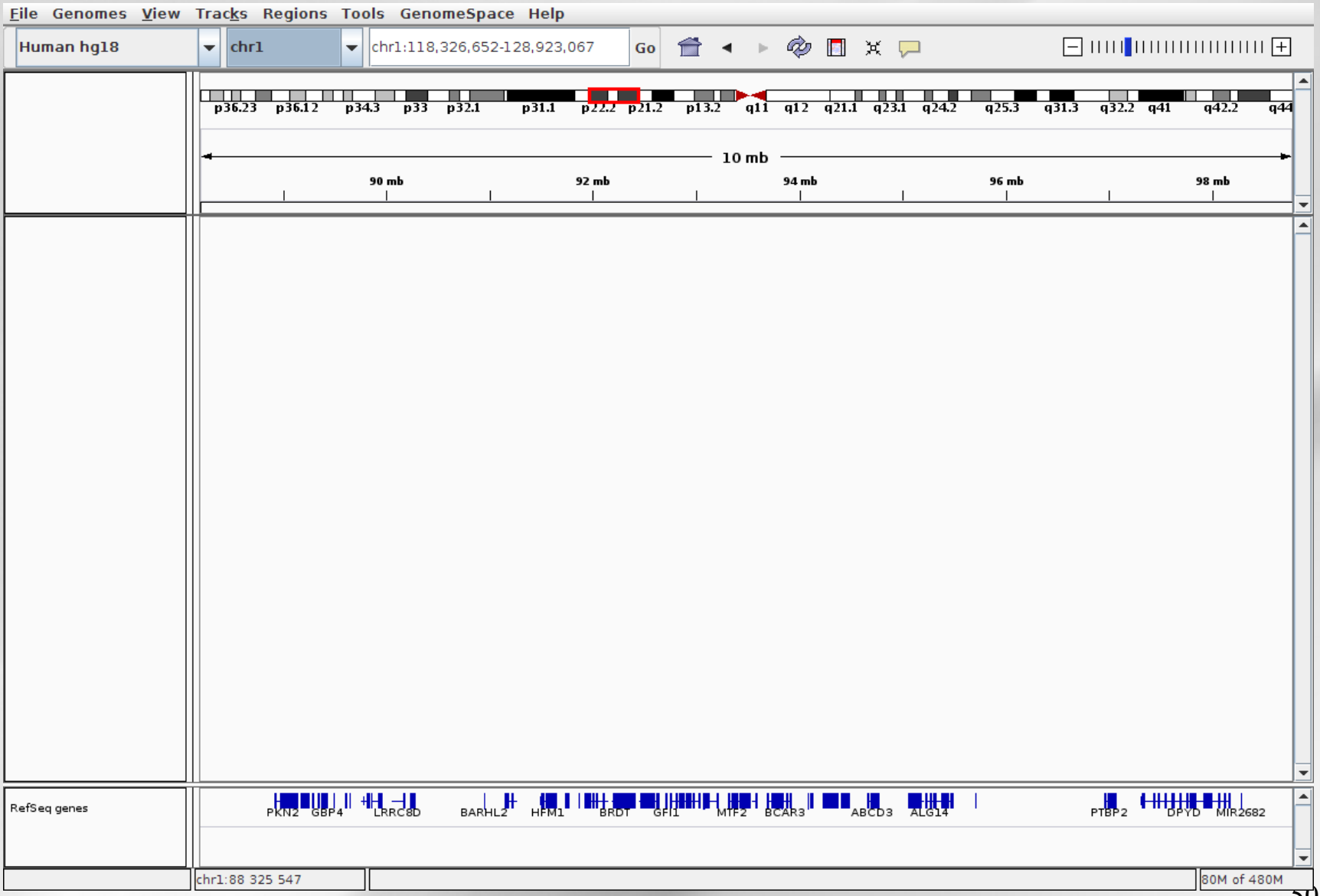
Visualizing the alignment - IGV



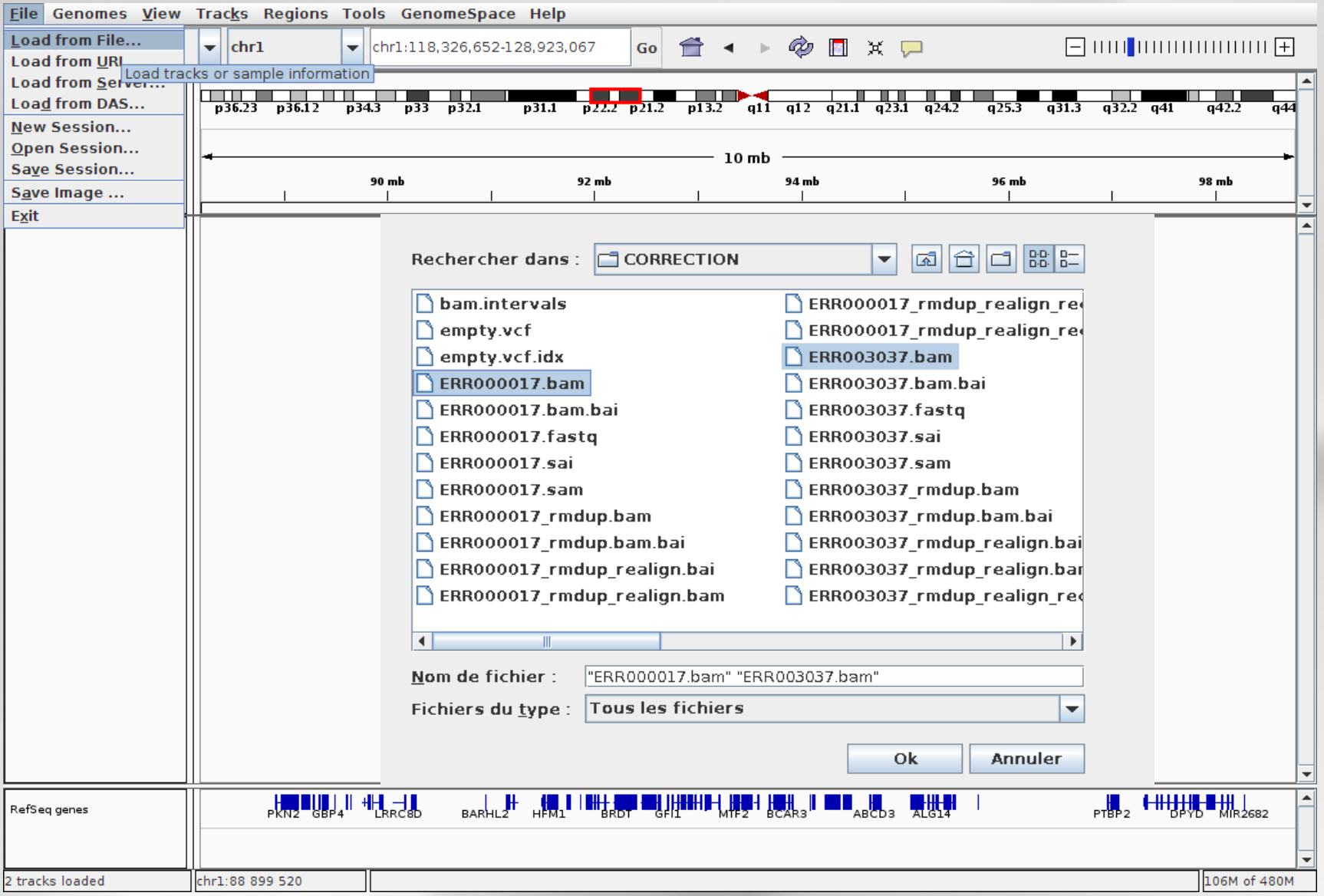
IGV - Loading the reference



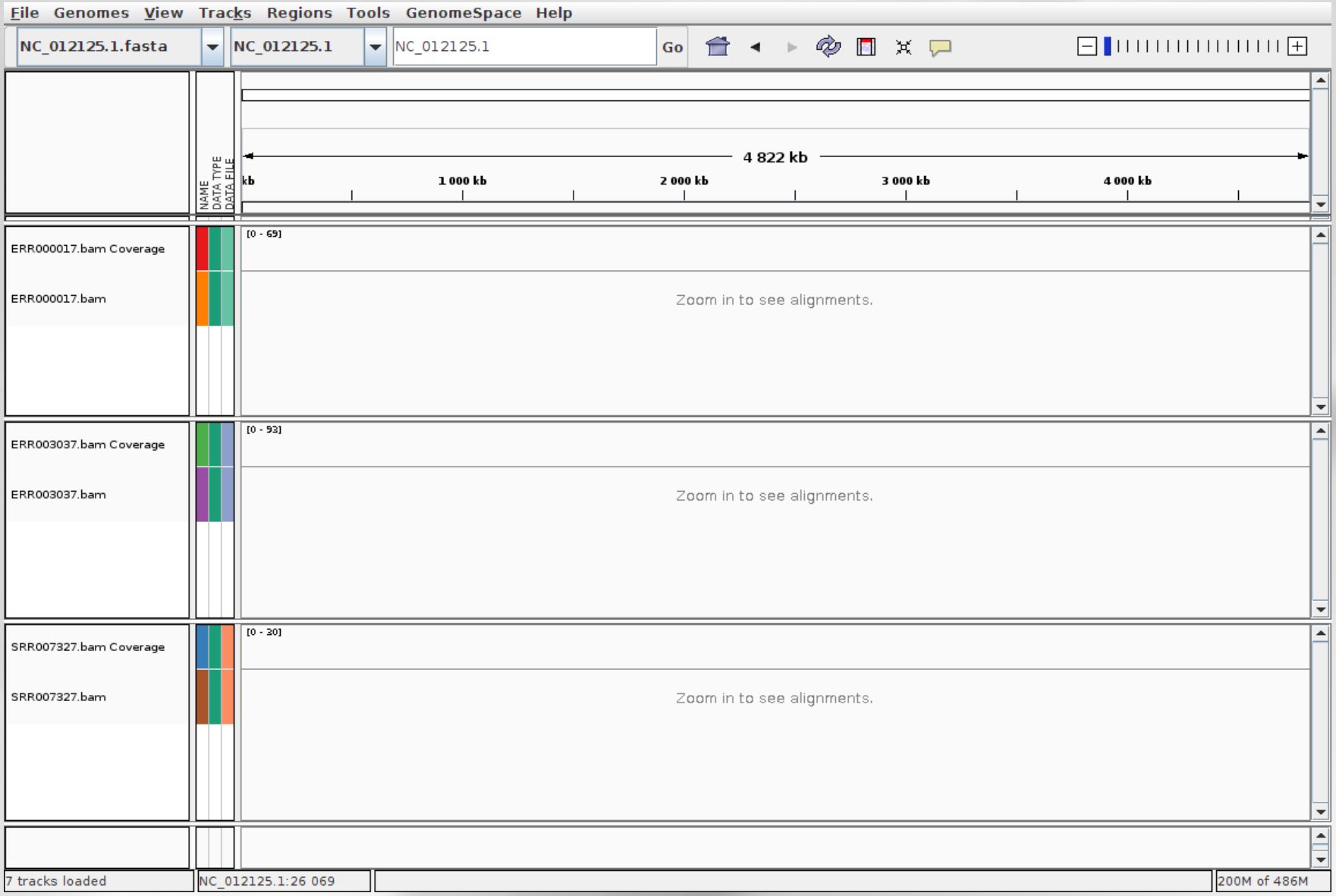
IGV - Loading the reference



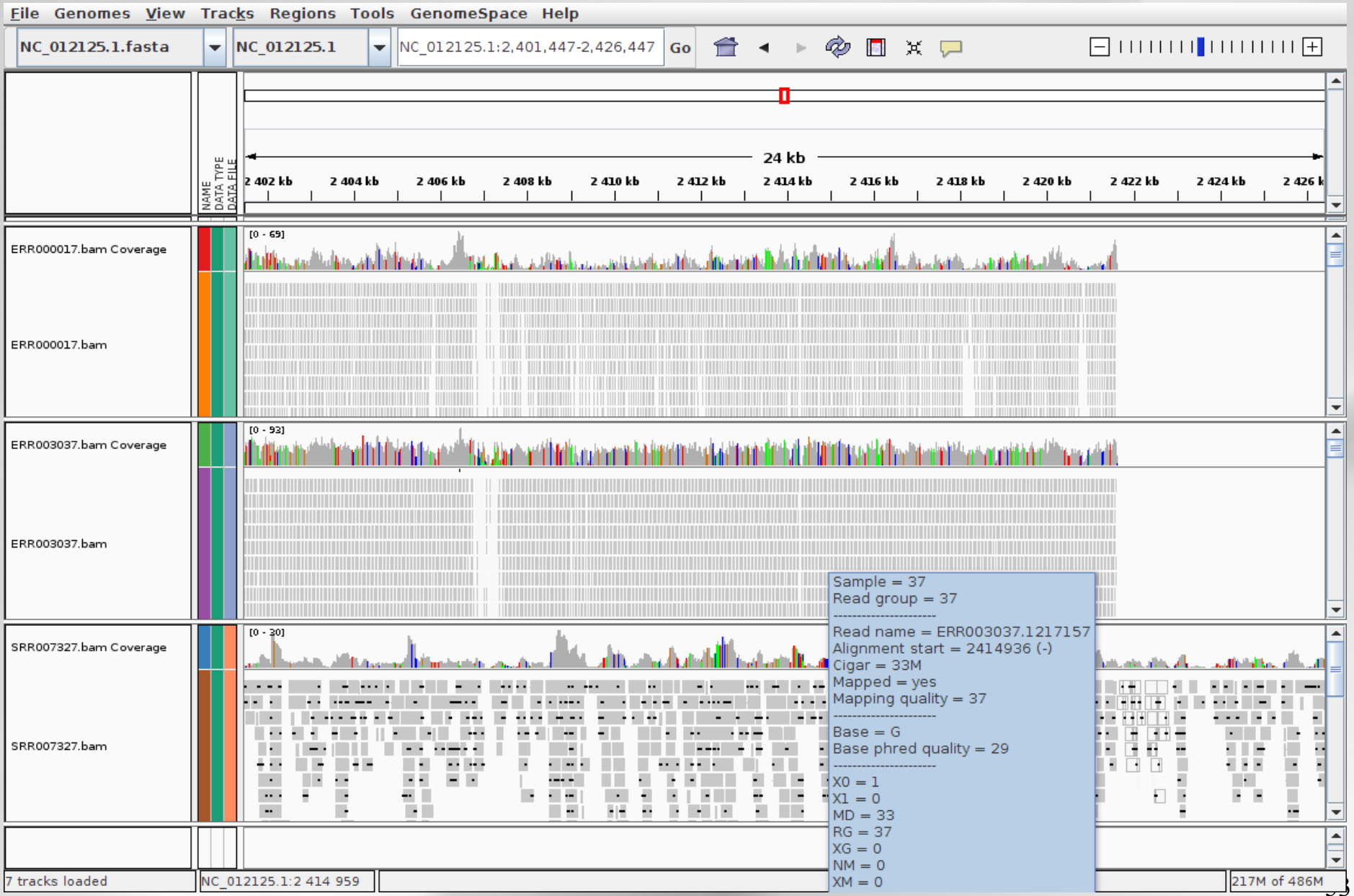
IGV - Loading the bam file



IGV - Loading the bam file

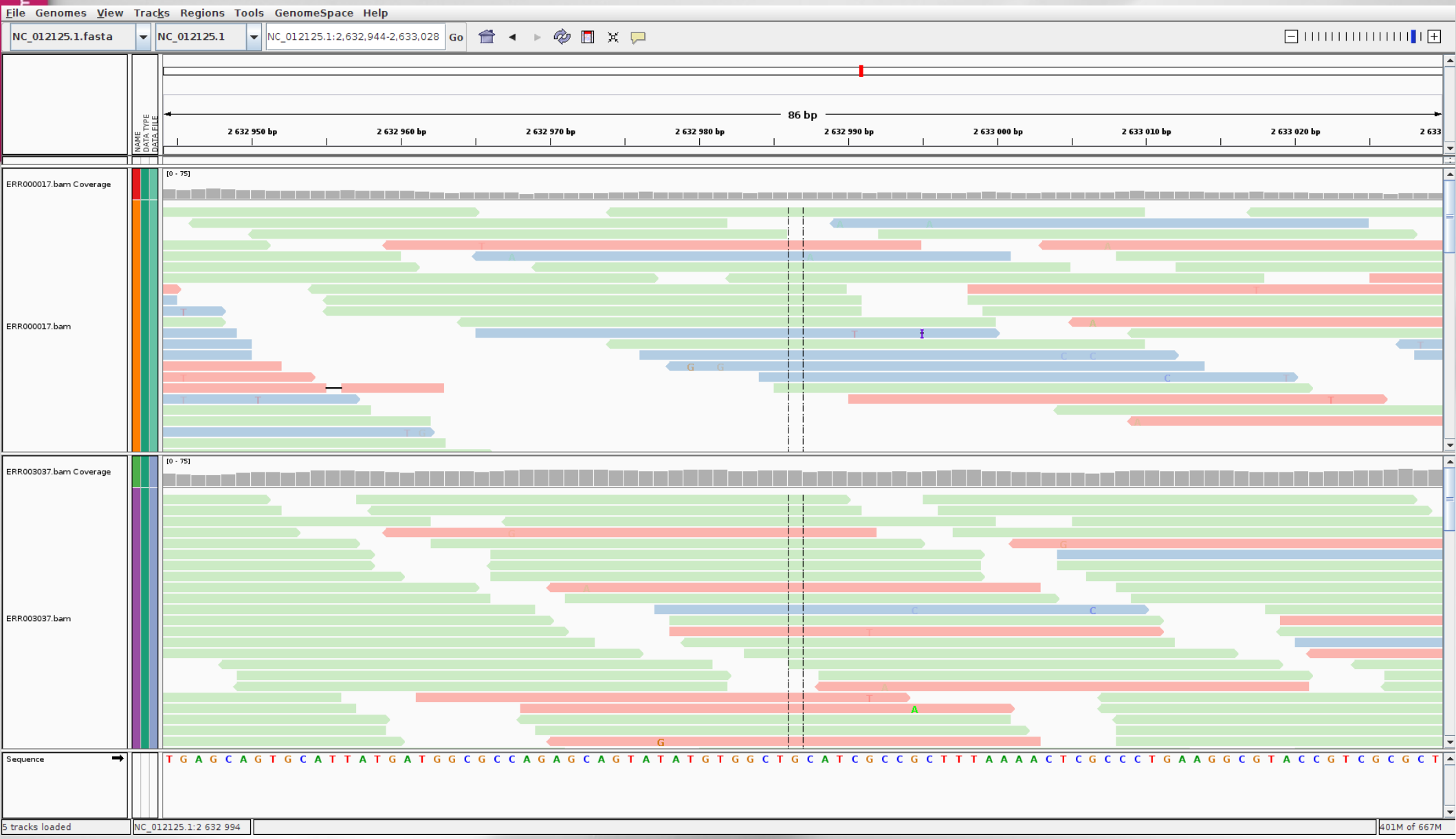


IGV - Zoom

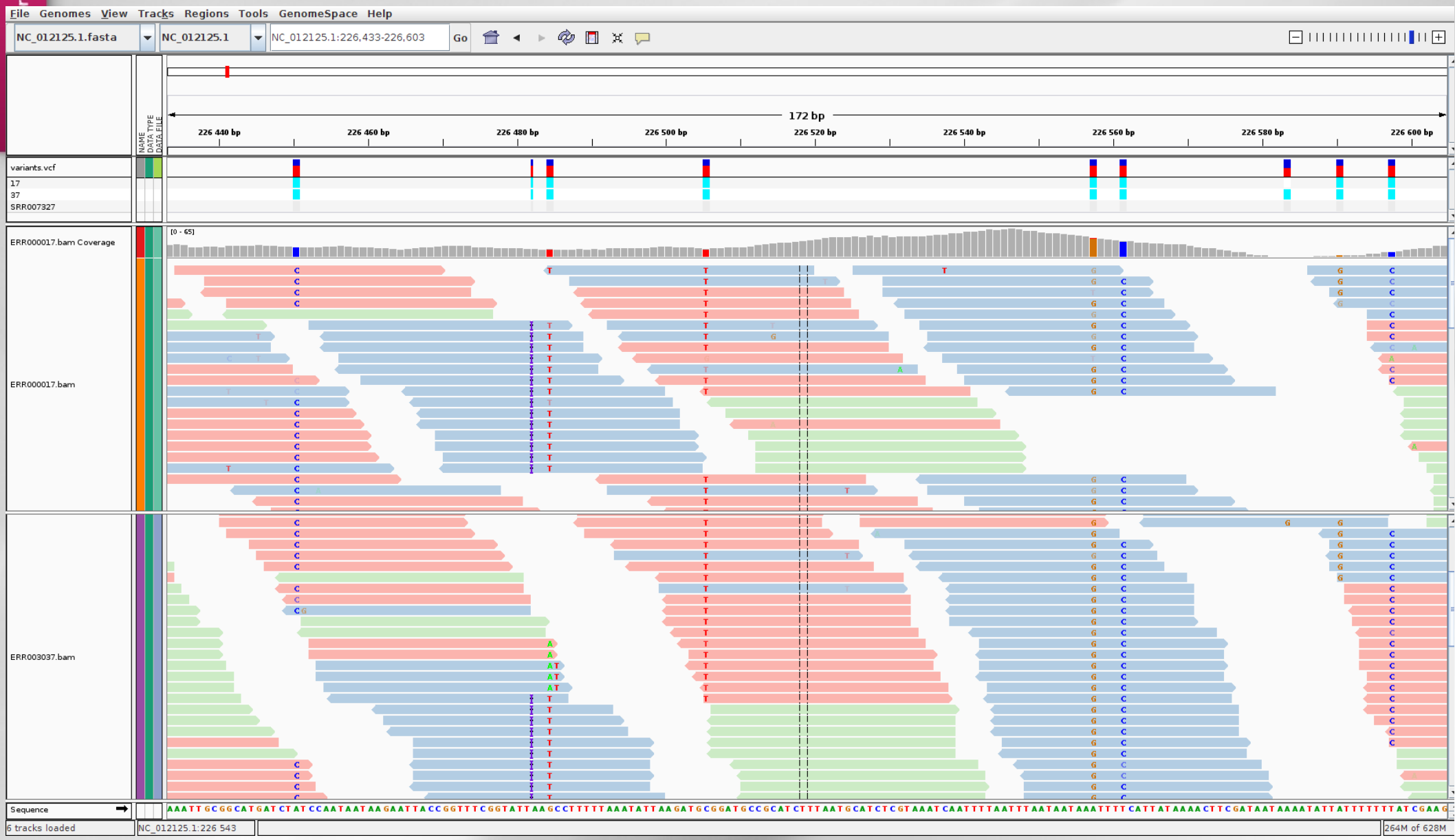

File Genomes View Tracks Regions Tools GenomeSpace Help
NC_012125.1.fasta
NC_012125.1
NC_012125.1:2,401,447-2,426,447
Go
24 kb
2 402 kb
2 404 kb
2 406 kb
2 408 kb
2 410 kb
2 412 kb
2 414 kb
2 416 kb
2 418 kb
2 420 kb
2 422 kb
2 424 kb
NAME
DATA TYPE
DATA FILE
ERR000017.bam Coverage
[0 - 69]
ERR000017.bam
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

geno toul bioinfo

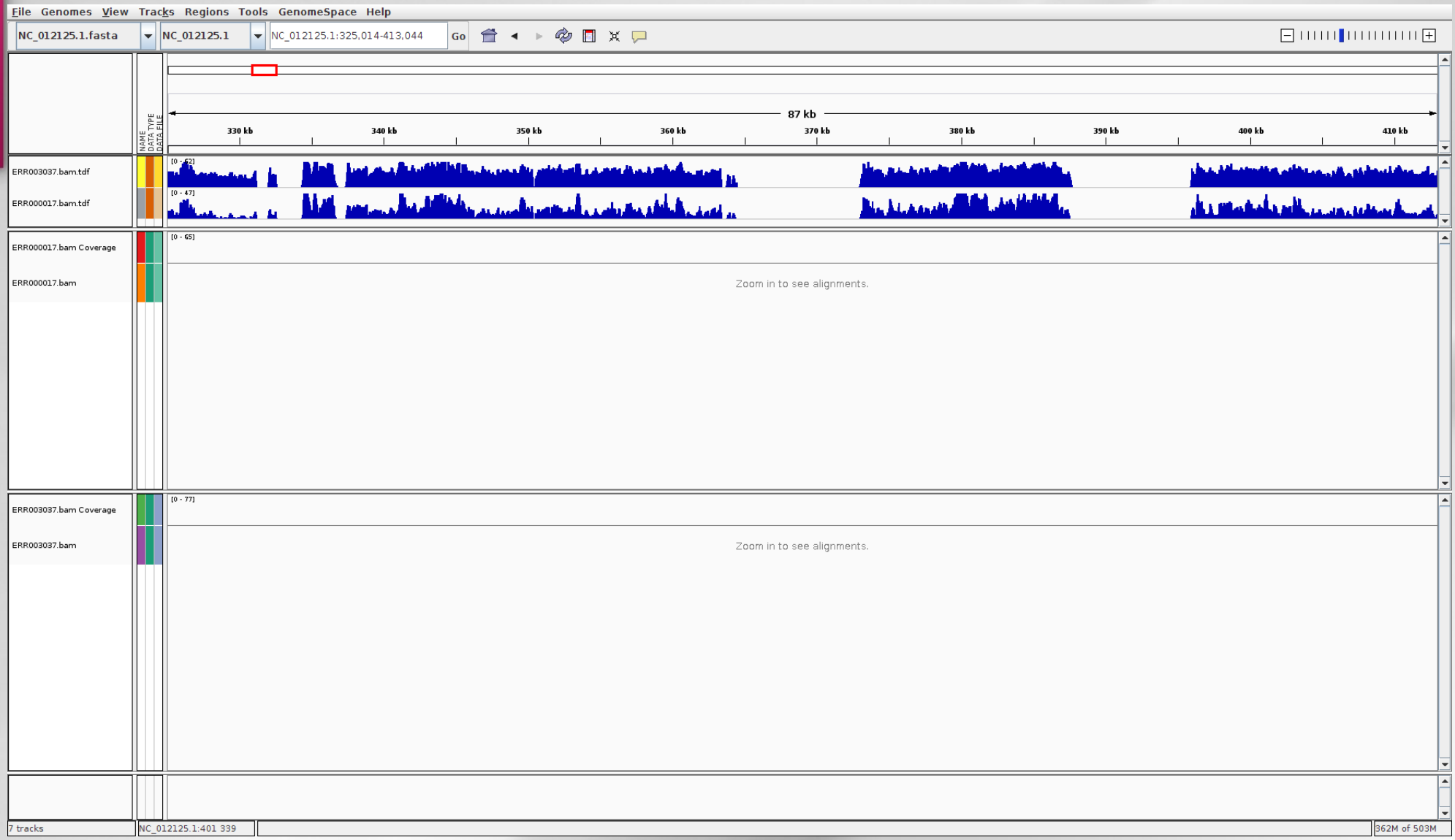
IGV - Zoom



IGV - Loading an annotation file



IGV - Coverage



Exercise / set 4

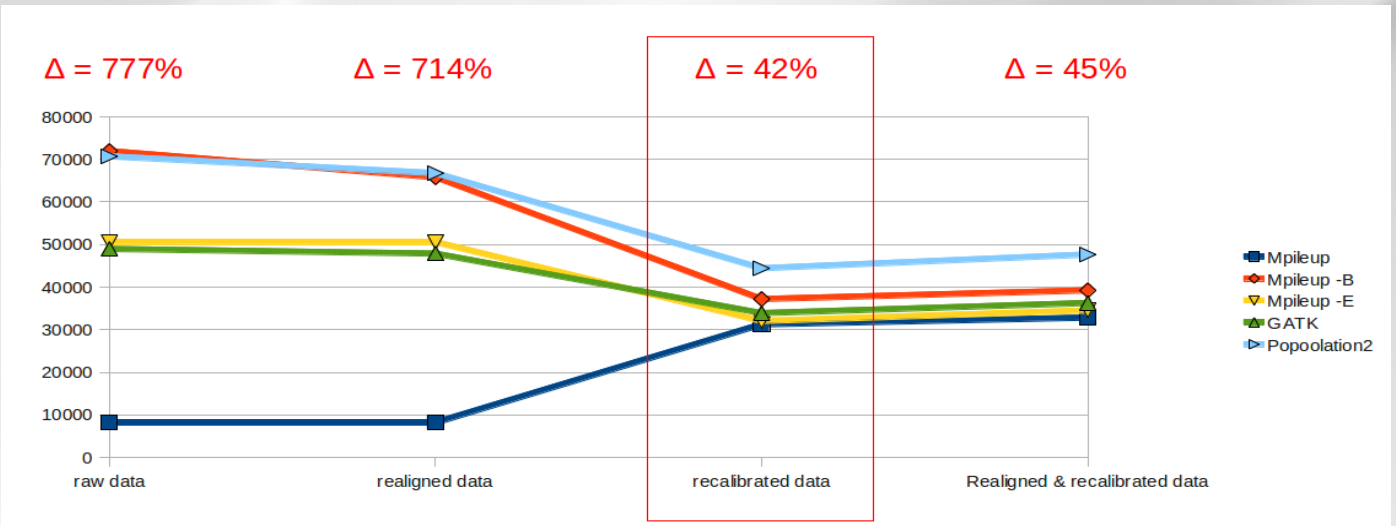
- ***Visualize FASTA and BAM files through IGV***

Variant Calling methodology

- Several Variant callers :
  - Samtools
  - GATK
  - FreeBayes
  - VarScan...

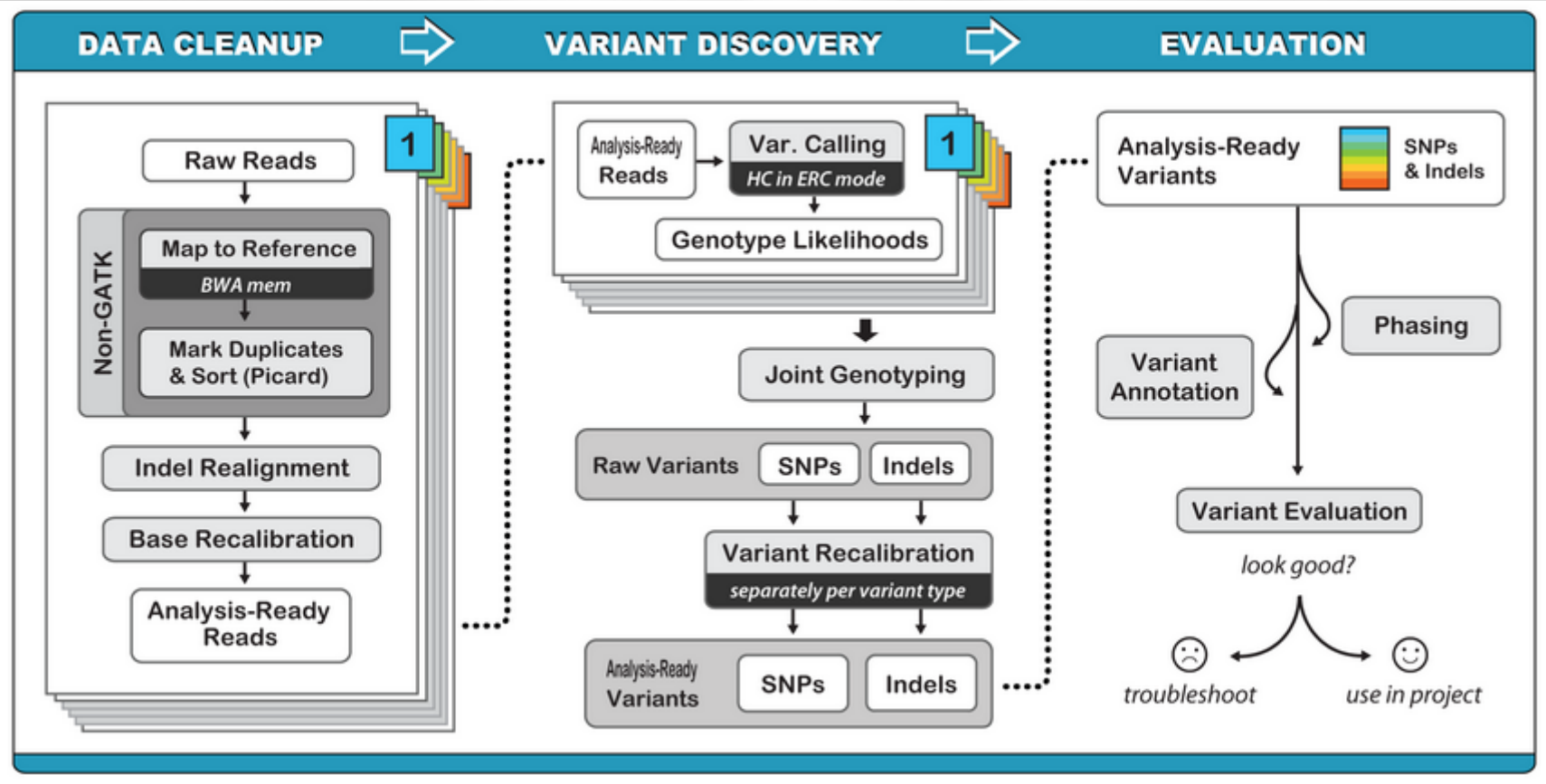
Why GATK ?

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



THE GATK best practices

- Recommended steps
- Warning : Frequently upgraded...

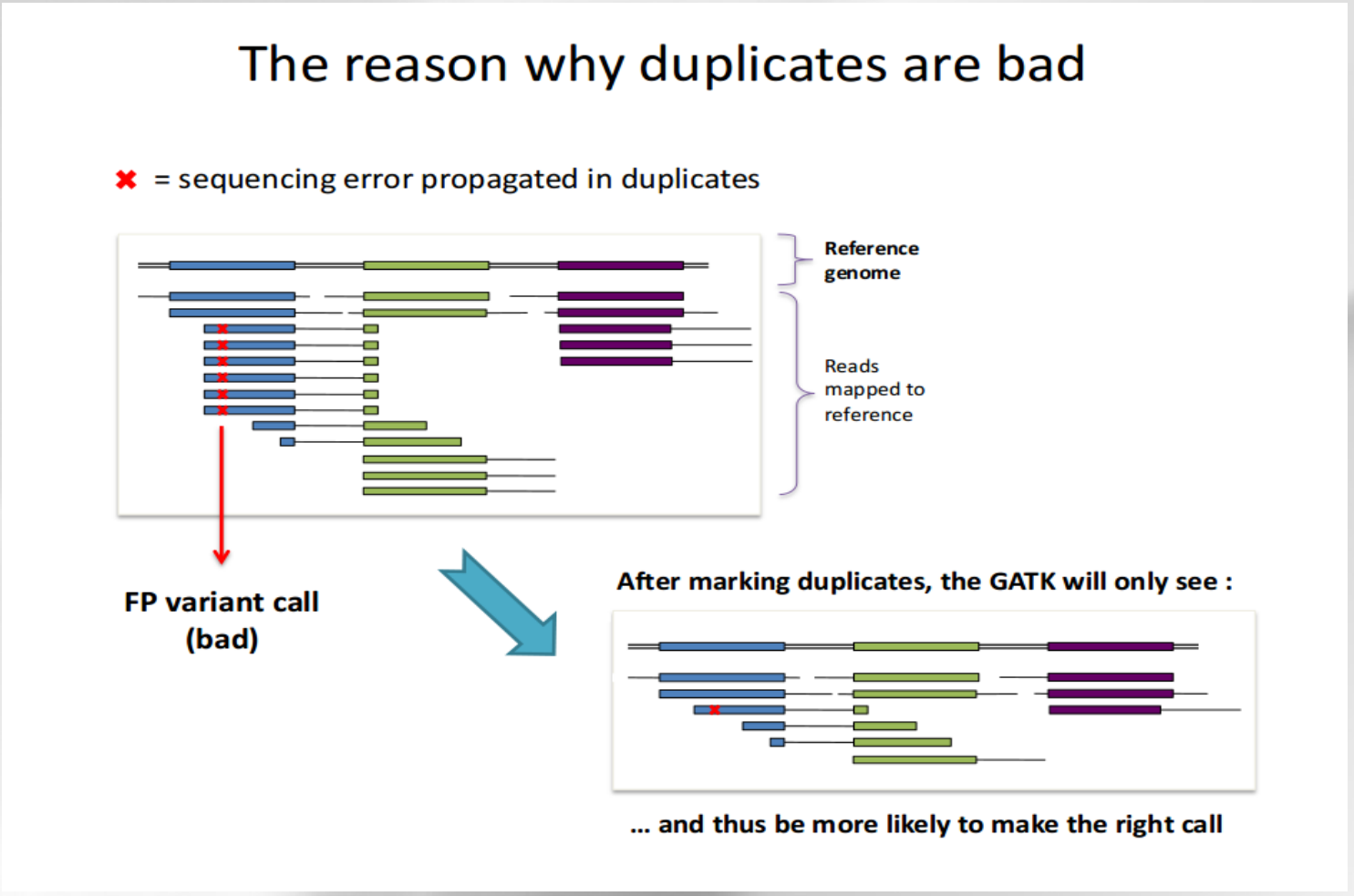


Pre-processing

- Add ReadGroups (Where samples come from ?)
- Remove duplicates

Pre-processing

- Add ReadGroups (Where samples come from ?)
- Remove duplicates



Pre-processing

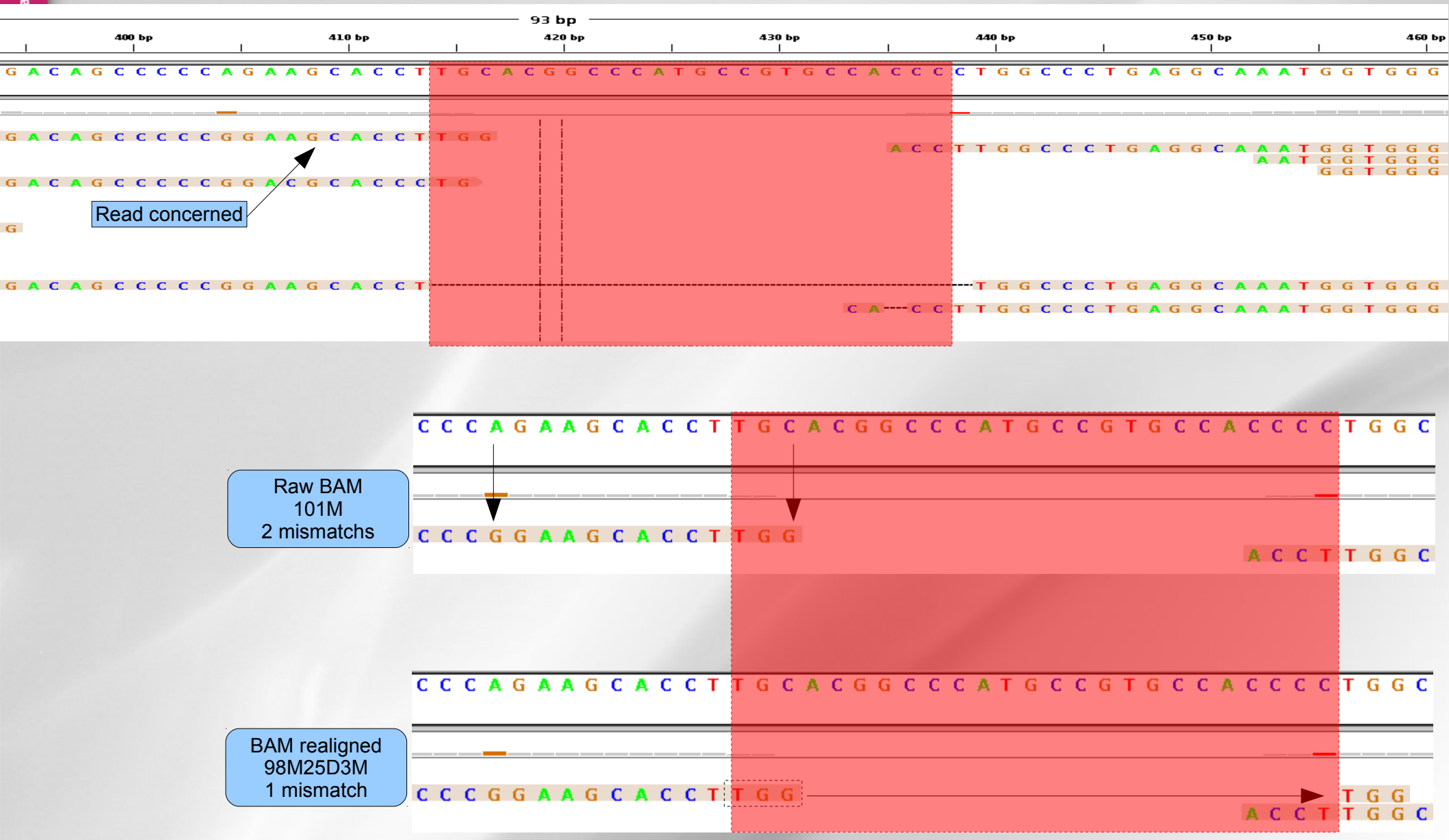
- Add ReadGroups (Where which samples come from ?)
- Remove duplicates
- Local realignment

Midi-Pyrénées

geno toul bioinfo

Local Realignment

Local Realignment

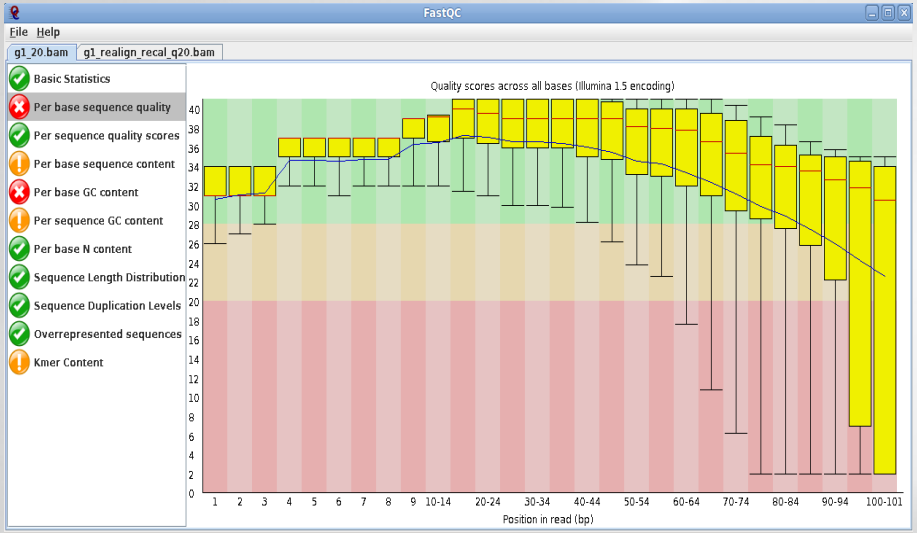


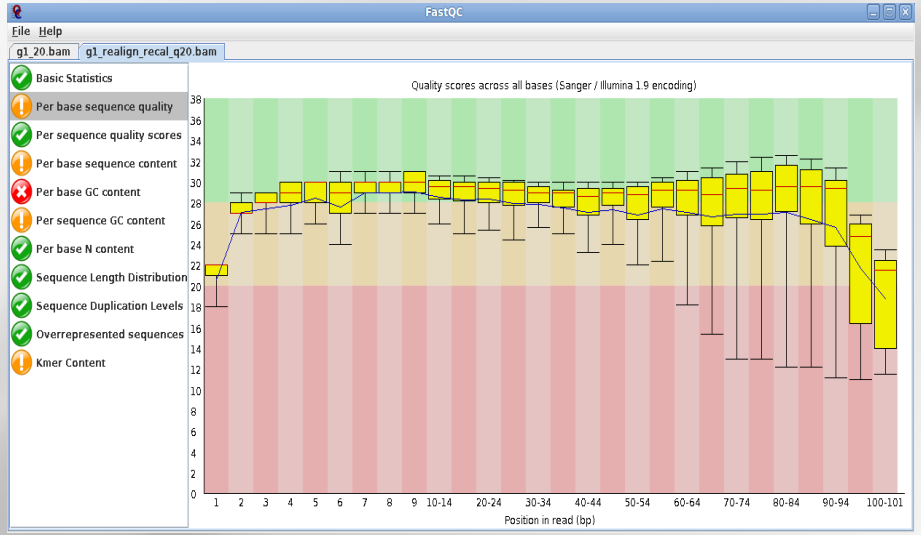
Pre-processing

- Add ReadGroups (Where samples come from ?)
- Remove duplicates
- Local realignment
- Base recalibration

Base recalibration

- The base quality provided by the sequencers is too inaccurate to be kept. They are re-computed.
- Needs knowledge of real SNP for recalibrating





Variant discovery

Base scale	Read scale	Position scale	Genotype scale
Phred-Quality Base	Mapping Quality	ALT allele count	Overall genotype association
	Forward/Reverse	REF allele count	
		ALT / REF	
		Read Depth	

SNP quality

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

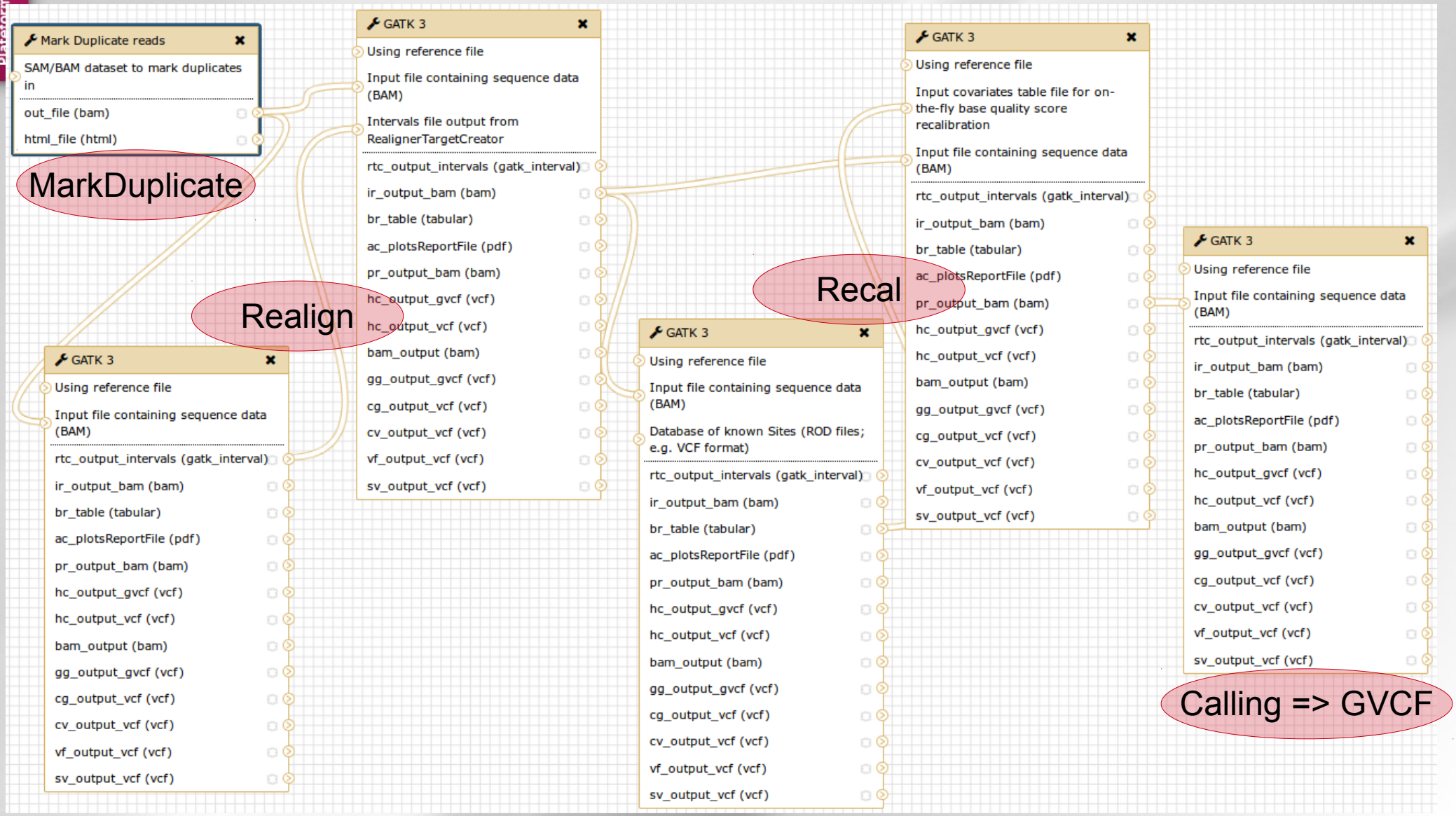
$30 \Rightarrow P_{error} = 1 / 1000$

Variant discovery

- Call variants individually for each sample
- Joint genotyping analysis
- (Variant recalibration)

GATK pipeline

- Best practices : https://www.broadinstitute.org/gatk/guide/bp_step.php



VCF format

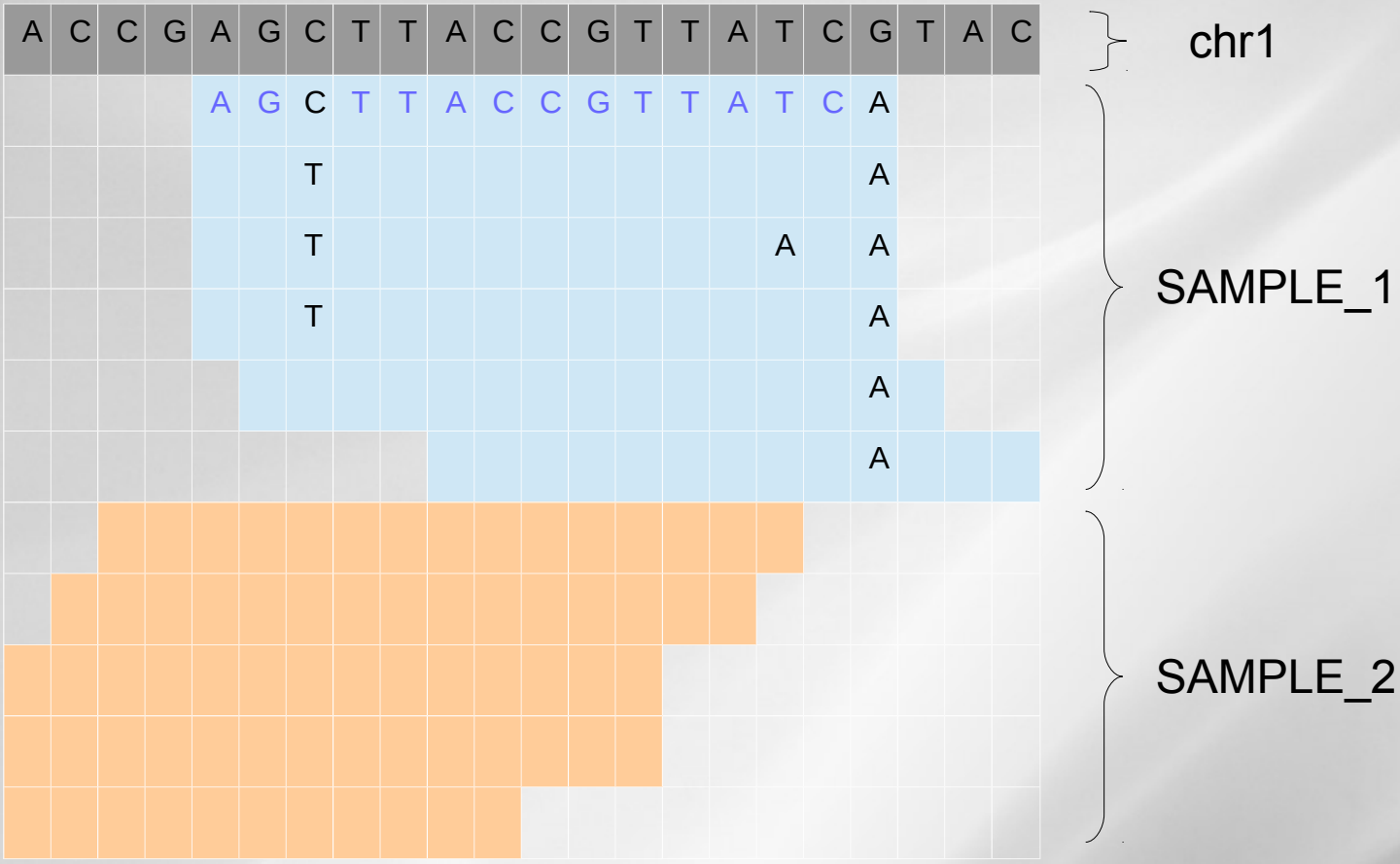
- Which informations have been stored in VCF ?

http://genoweb.toulouse.inra.fr/~formation/2_Galaxy_SGS-SNP/.formats/vcf.html

Variant calling Format (VCF)

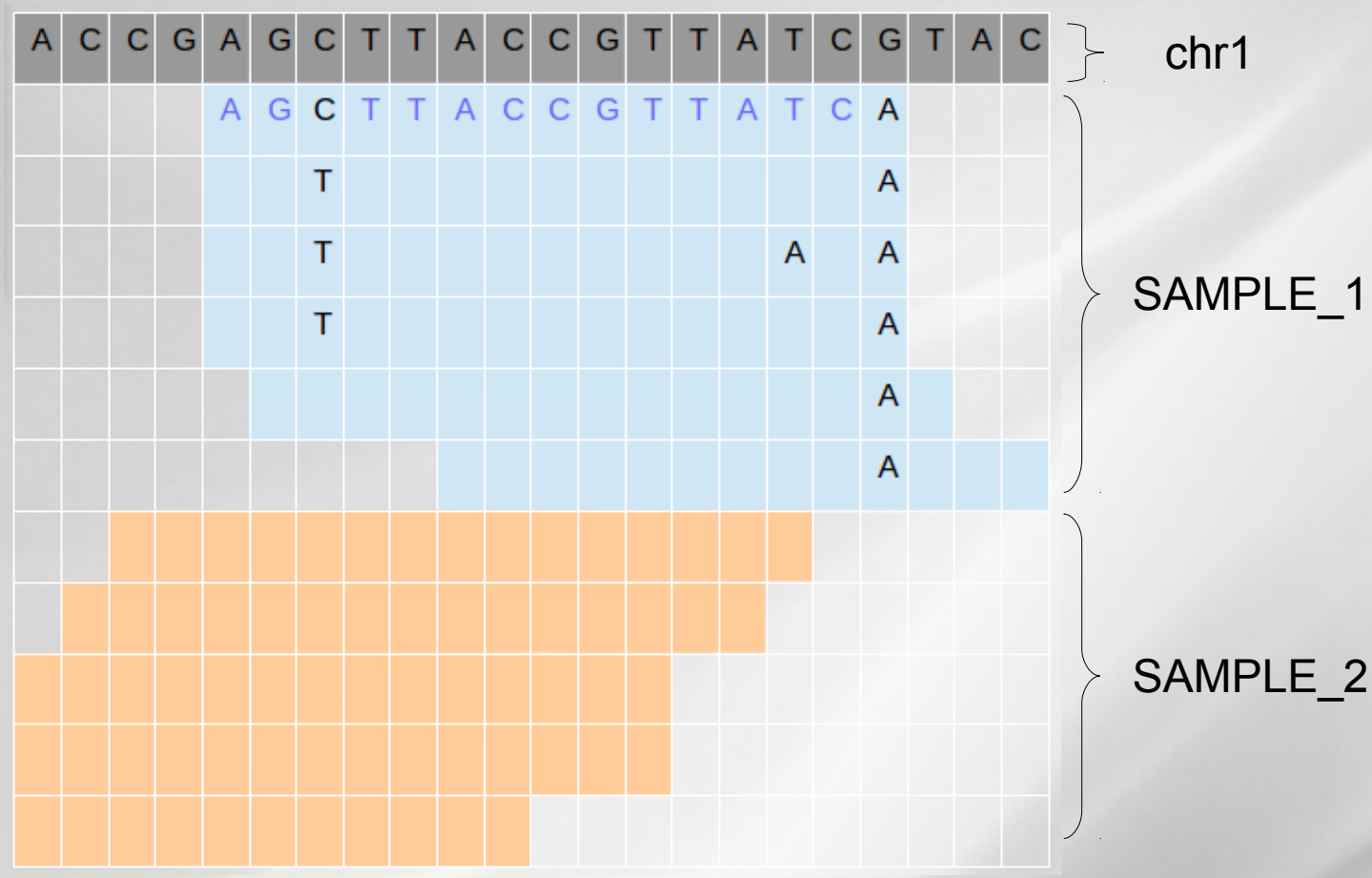
- http://vcftools.sourceforge.net/specs.html
- Tab-delimited file
- Basic documentation inside
- Header lines start with ## or #
- Give informations of data/tools/parameters used
- Variant lines represent a position on the genome

The VCF format : example



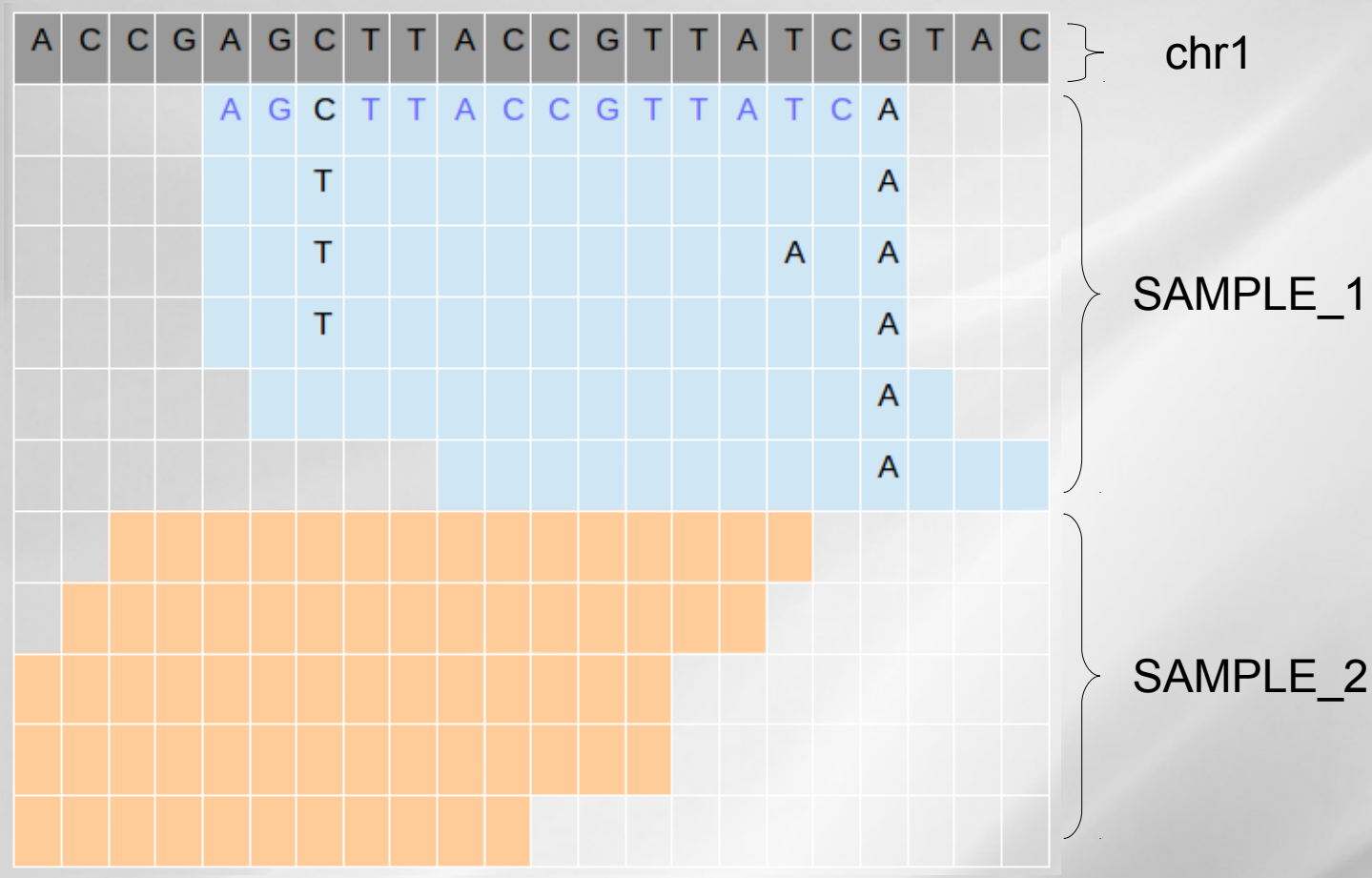

```
#CHR  POS   ID    REF   ALT   QUAL  FILTER     [INFOS]     FORMAT          SAMPLE_1                SAMPLE_2
chr1  7     .     C     T     247.82     .     [INFOS]     GT:AD:DP:GQ:PL  0/1:2,3:5:9.2:20,0,15    0/1:2,3:5:9.01:10,1,6
chr1  19    .     G     A     124.34     .     [INFOS]     GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94   ./.
```

The VCF format : example



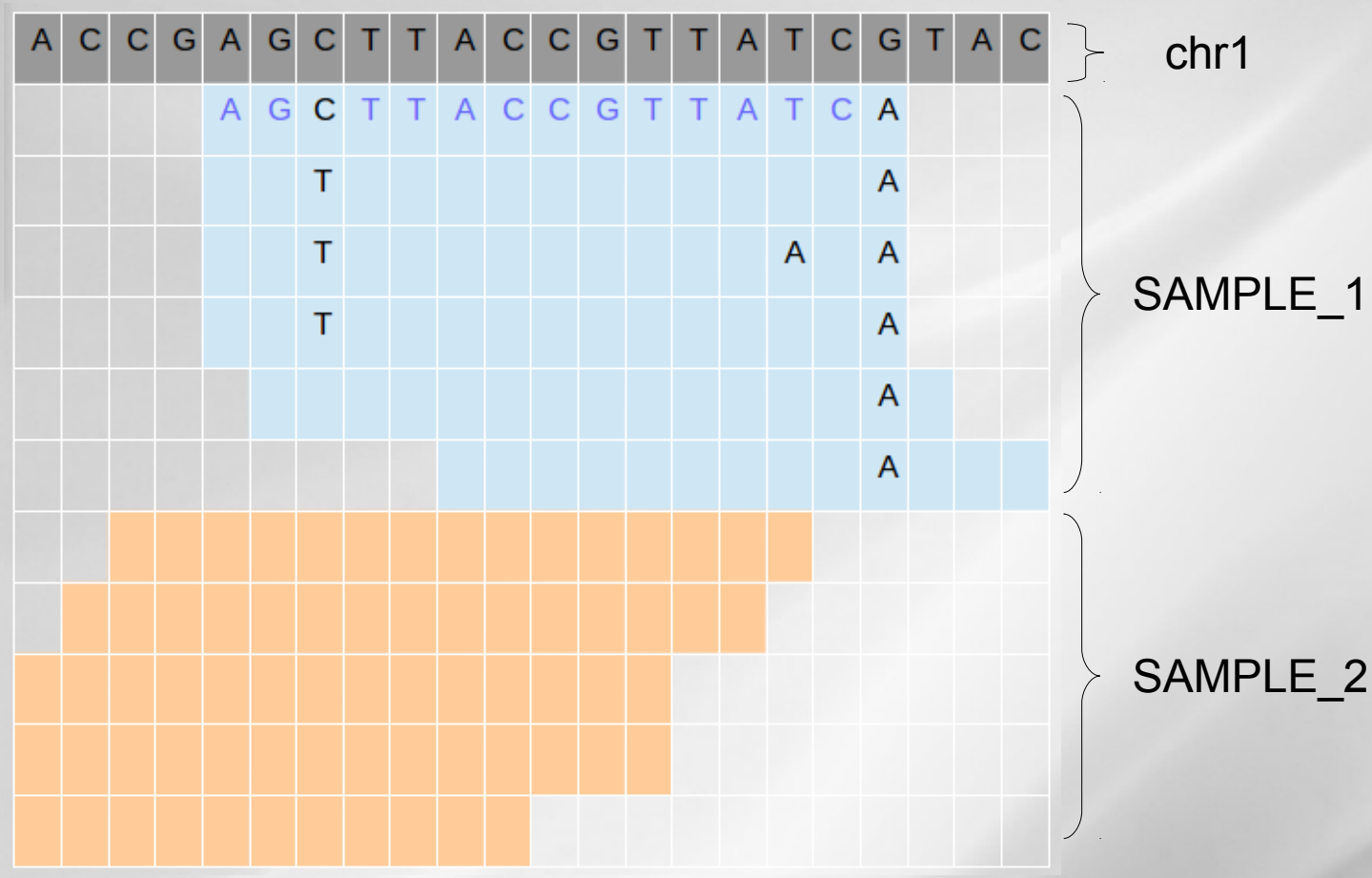

```
#CHR  POS  ID  REF  ALT  QUAL  FILTER  [INFOS]  FORMAT          SAMPLE_1                SAMPLE_2
chr1  7    .   C    T    247.82  .       [INFOS]  GT:AD:DP:GQ:PL  0/1:2,3:5:9.2:20,0,15    0/1:2,3:5:9.01:10,1,6
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

The VCF format : example



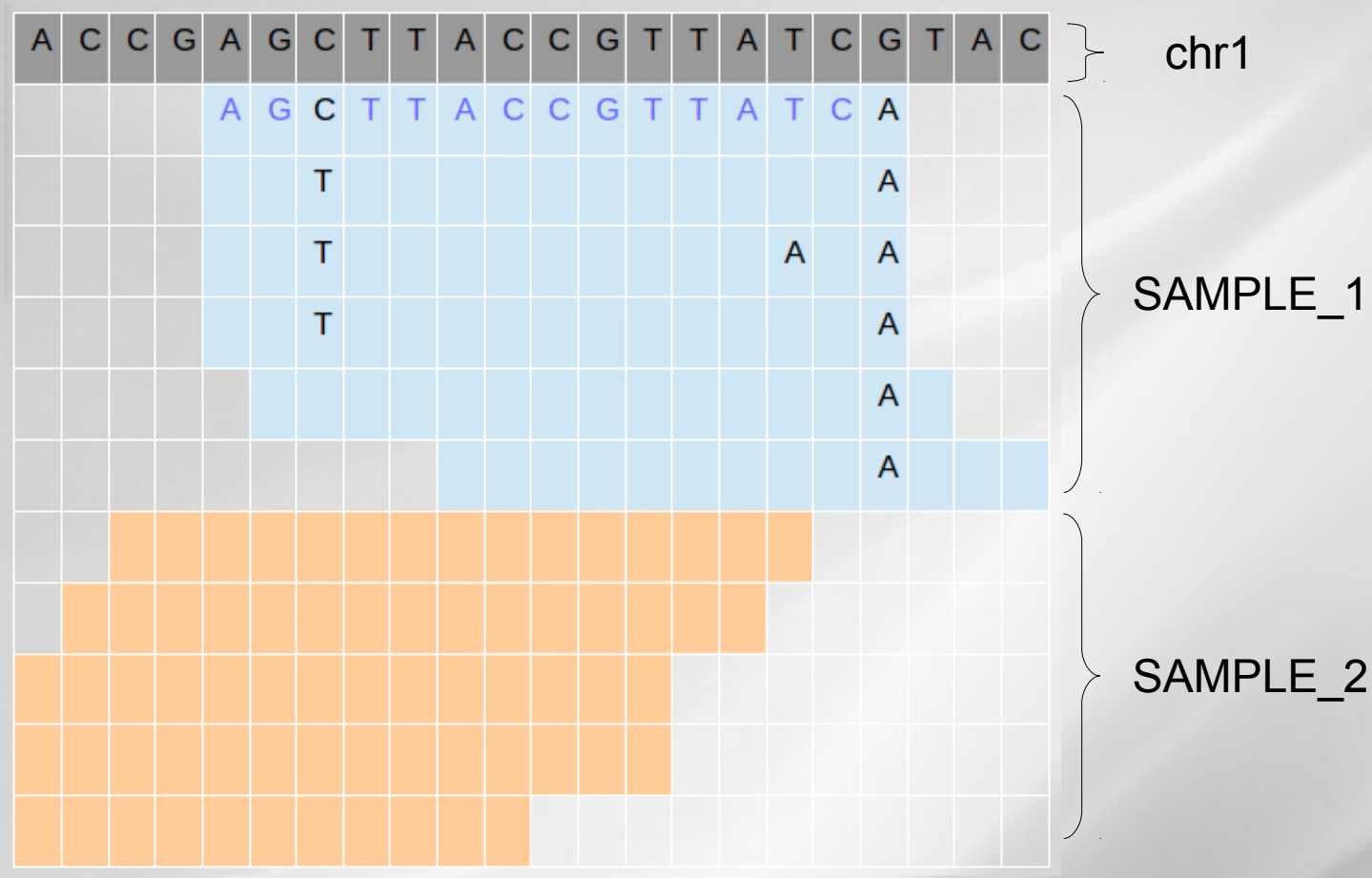

```
#CHR  POS  ID  REF  ALT  QUAL  FILTER  [INFOS]  FORMAT          SAMPLE_1                SAMPLE_2
chr1  7    .   C    T    247.82  .       [INFOS]  GT:AD:DP:GQ:PL  0/1:2,3:5:9.2:20,0,15    0/1:2,3:5:9.01:10,1,6
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

The VCF format : example



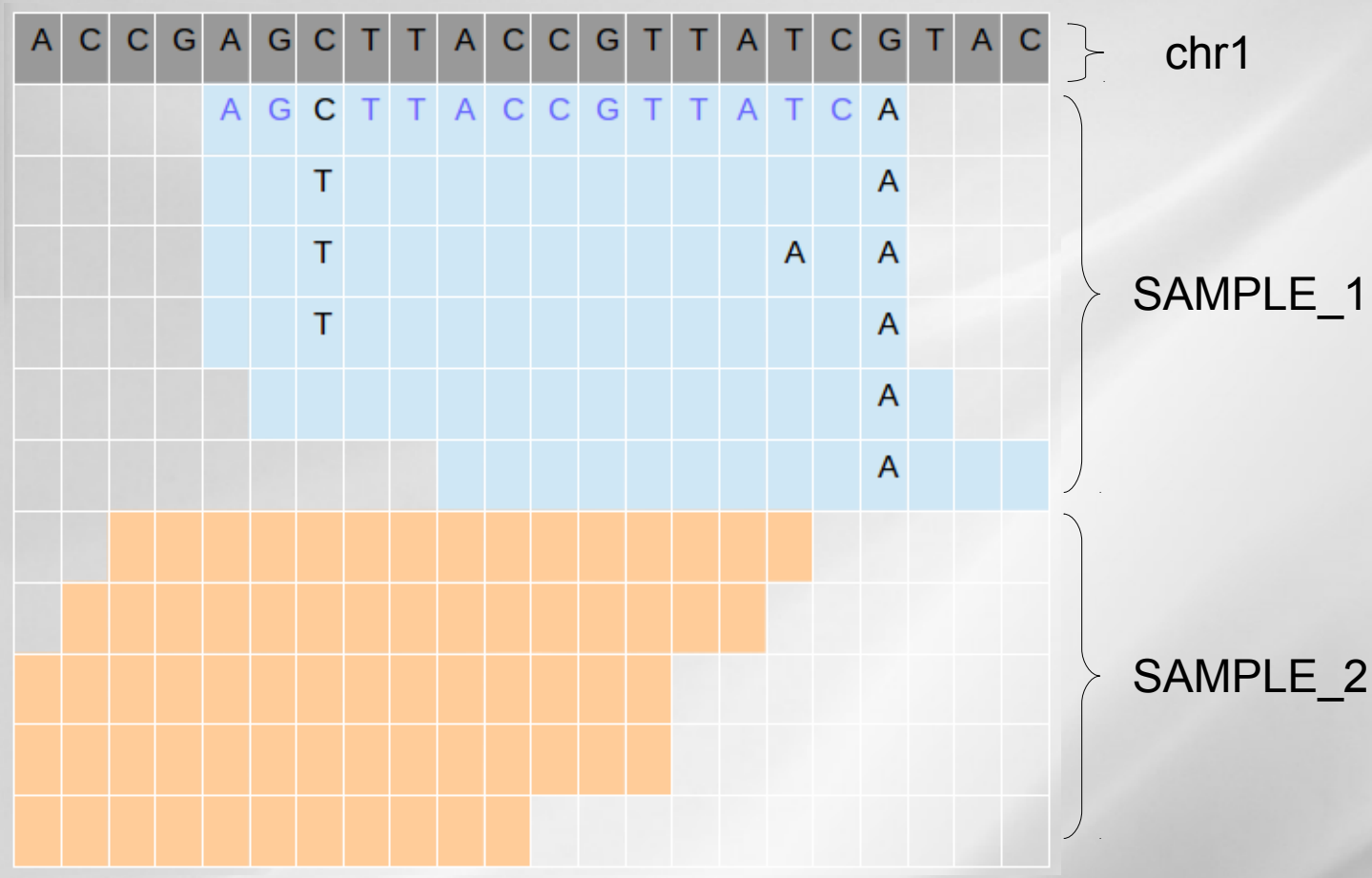
#CHR	POS	ID	REF	ALT	QUAL	FILTER	[INFOS]	FORMAT	SAMPLE_1	SAMPLE_2
chr1	7	.	C	T	247.82	.	[INFOS]	GT:AD:DP:GQ:PL	0/1:2,3:5:9.2:20,0,15	0/1:2,3:5:9.01:10,1,6
chr1	19	.	G	A	124.34	.	[INFOS]	GT:AD:DP:GQ:PL	0/0:5,0:5:20.2:0,42,94	./.

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  0/1:2,3:5:9.2:20,0,15    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    ./.
```

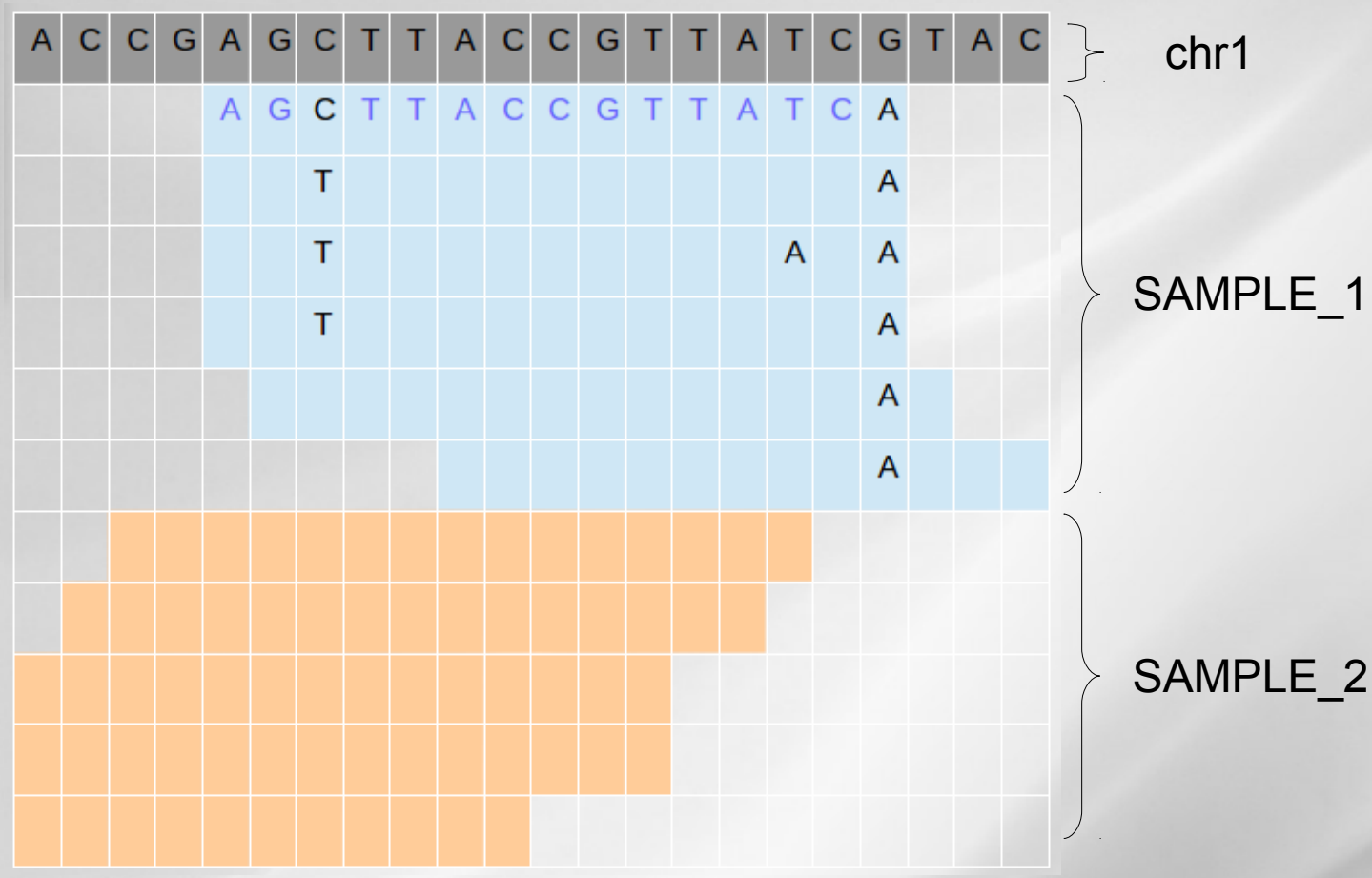
The VCF format : example



#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	0/1:2,3:5:9.2:20,0,15	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	./.

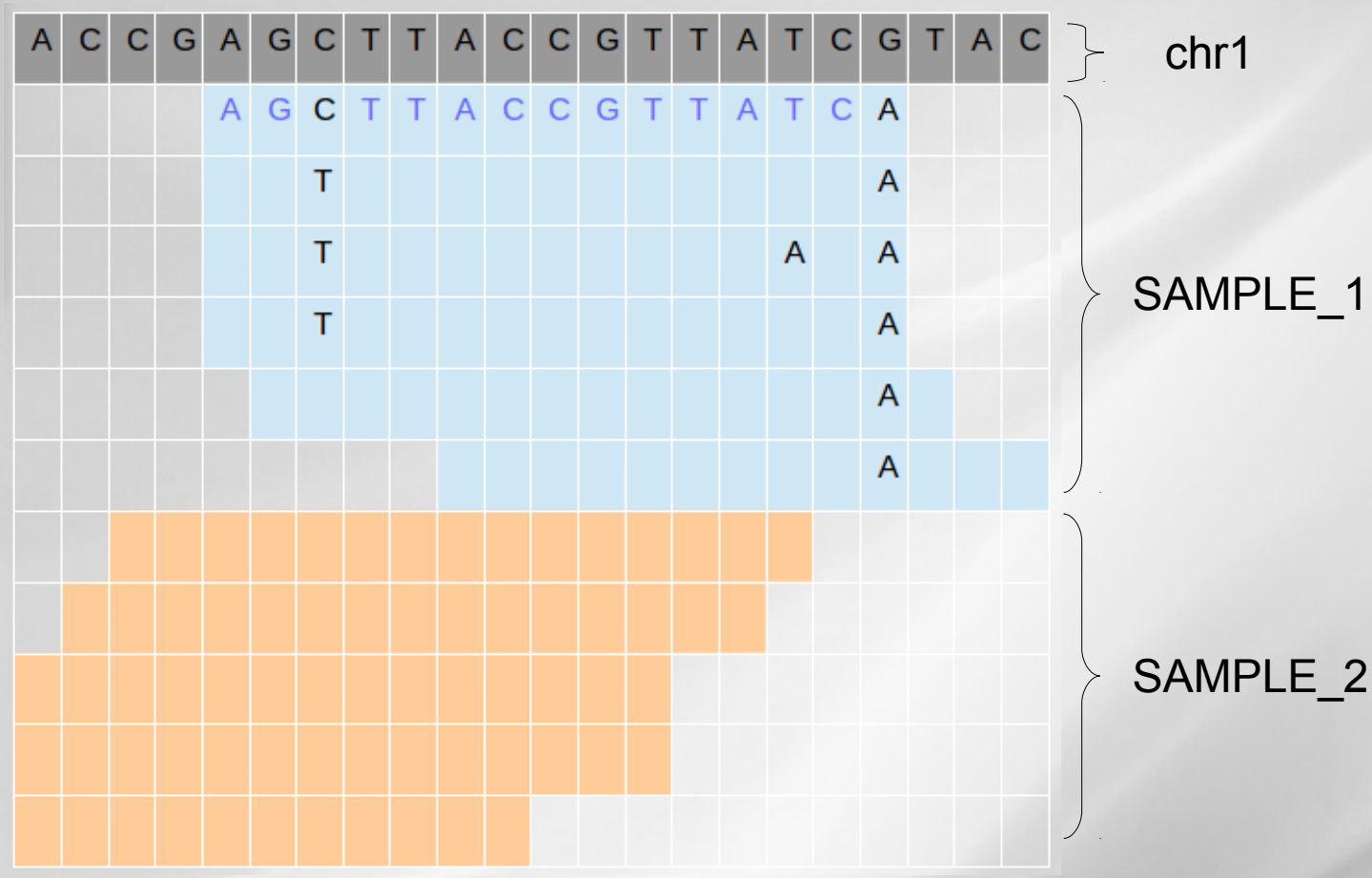
The Phred scaled probability that a REF/ALT polymorphism exists at this site given sequencing data

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	0/1:2,3:5:9.2:20,0,15	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	./.

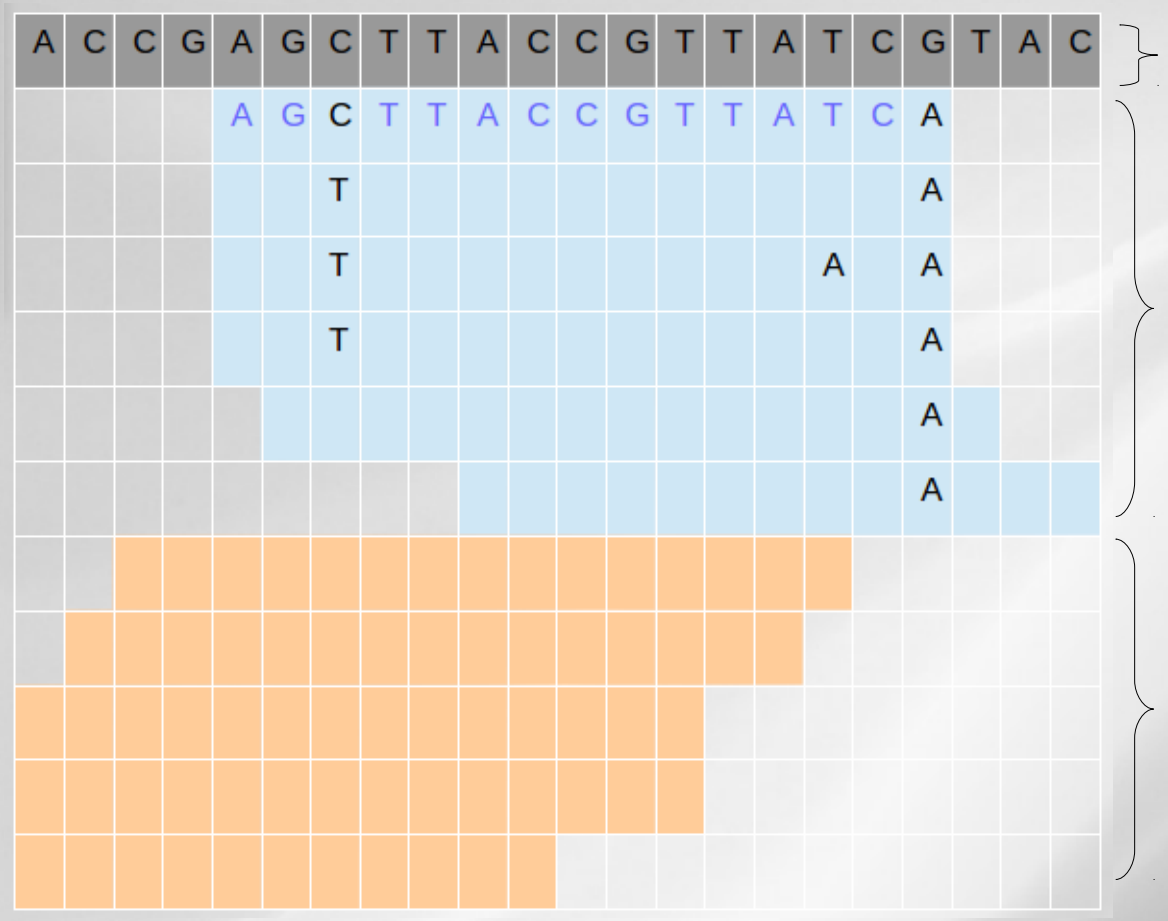
The VCF format : example




```
#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  0/1:2,3:5:9.2:20,0,15    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    ./.
```

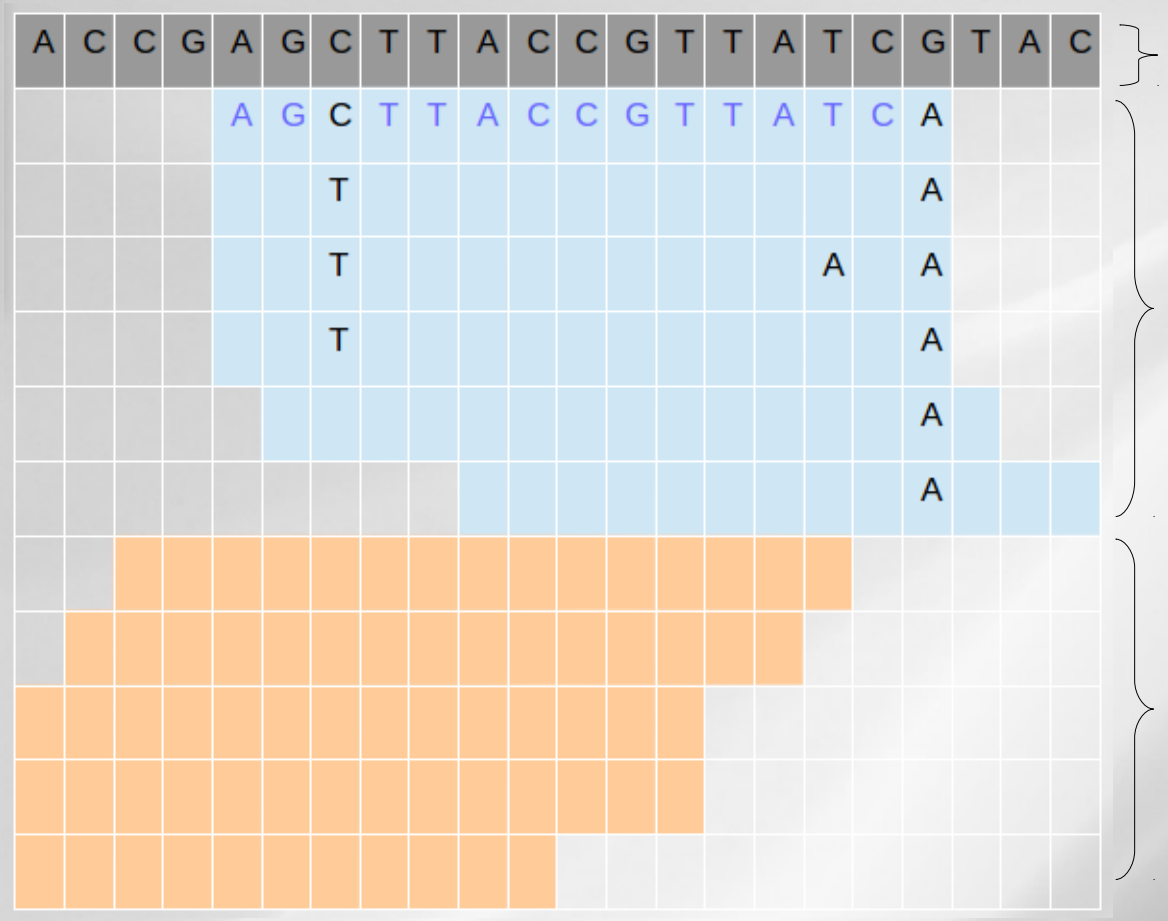
[TAG=VALUE]
DP=45

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  0/1:2,3:5:9.2:20,0,15    1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

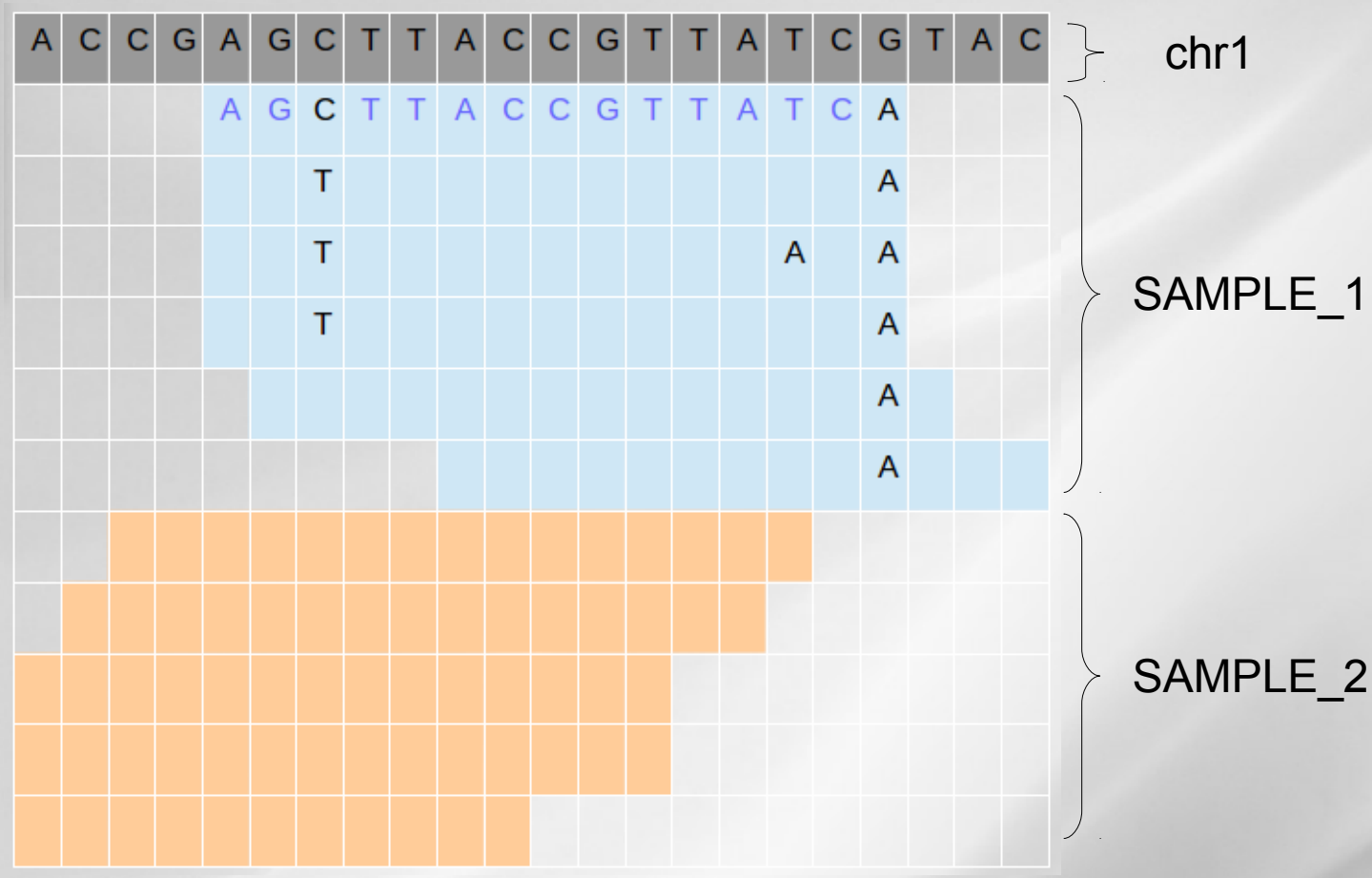
The VCF format : example




```
#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  0/1:2,3:5:9.2:20,0,15  1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94  ./.
```

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

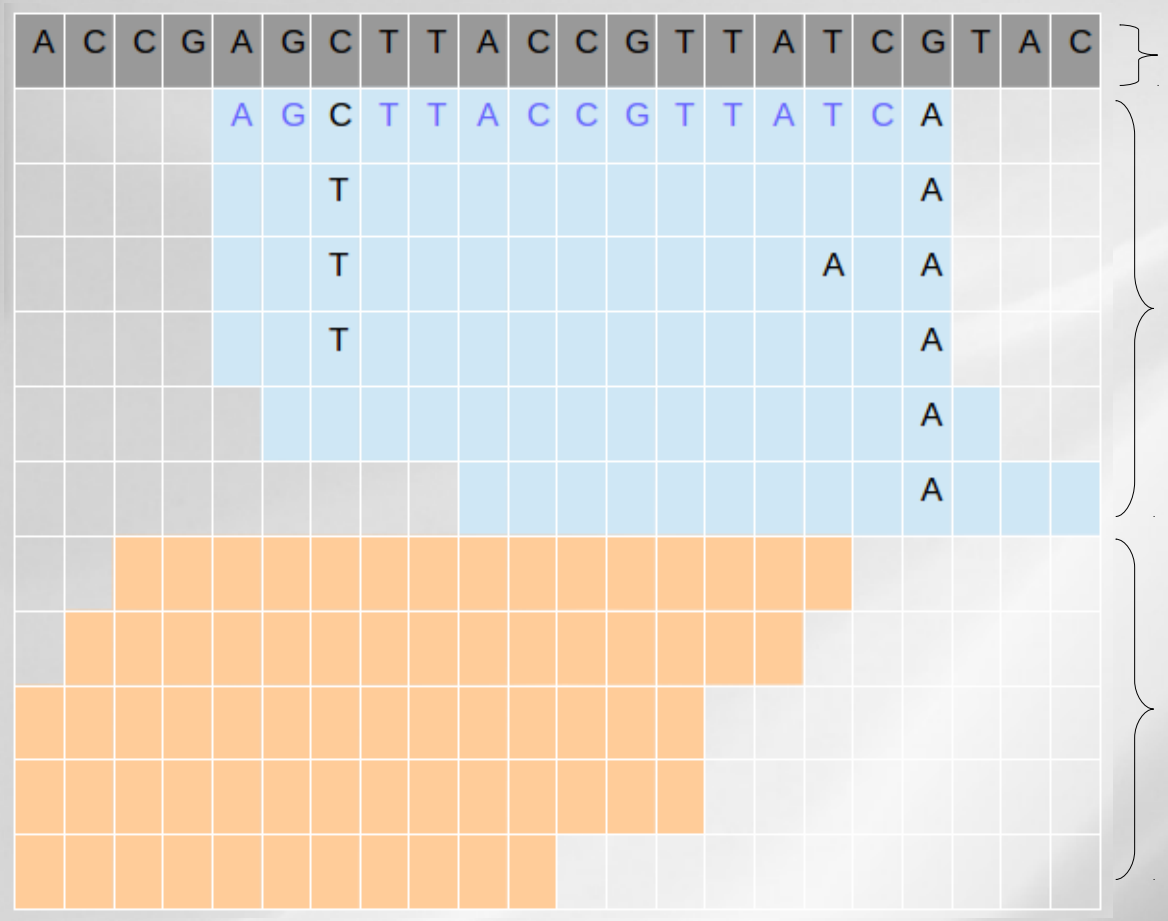
The VCF format : example




```
#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  0/1:2,3:5:9.2:20,0,15  1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94  ./.
```

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

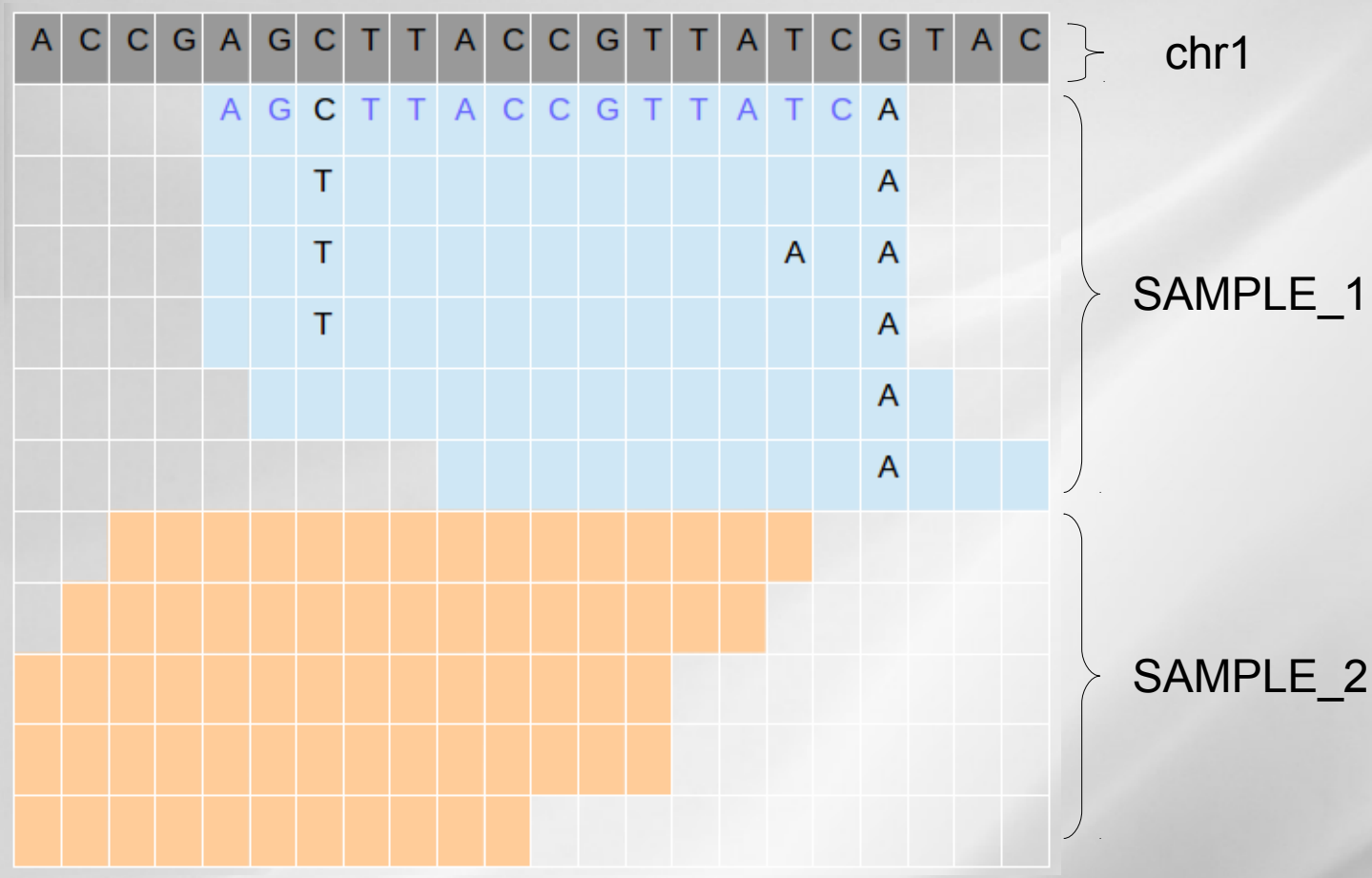
The VCF format : example




```
#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  0/1:2,3:5:9.2:20,0,15    1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

Depth position

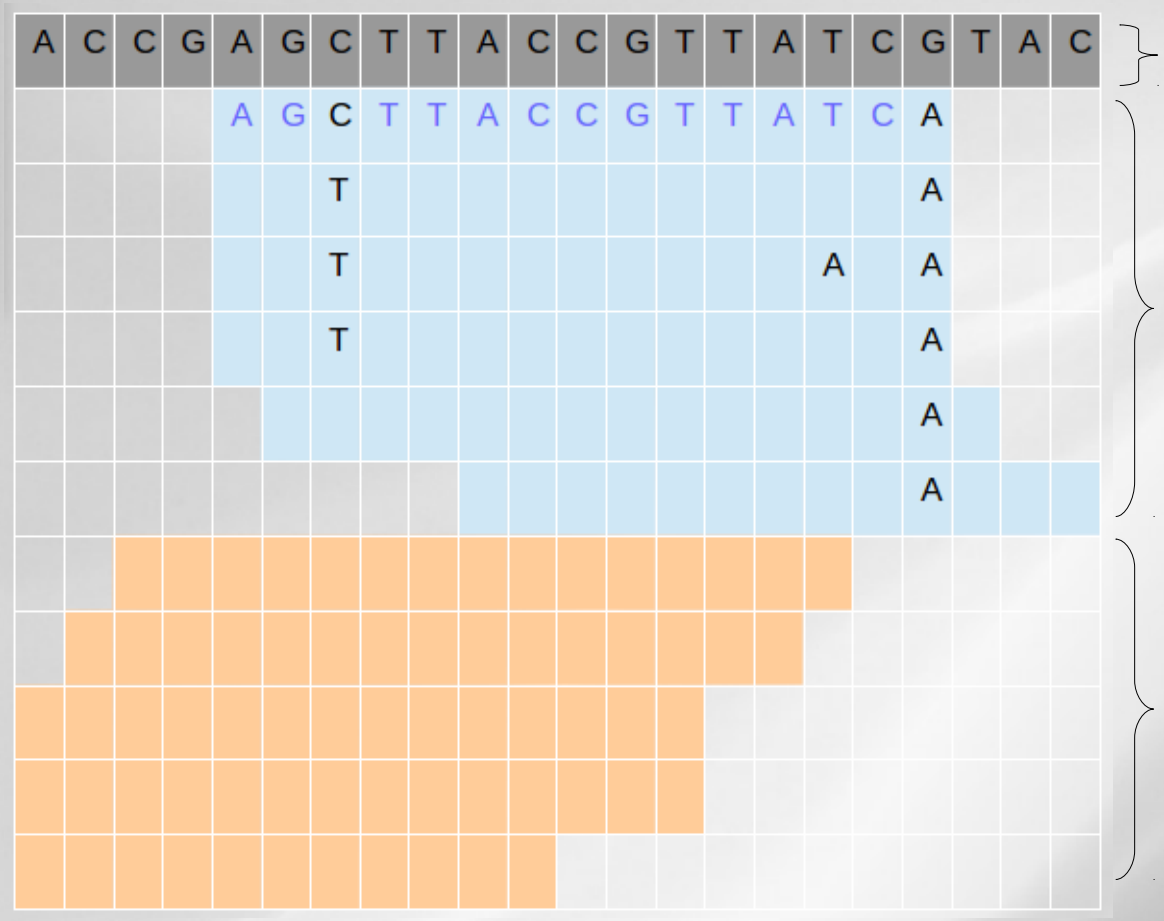
The VCF format : example




```
#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  0/1:2,3:5:9.2:20,0,15    1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

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  0/1:2,3:5:9.2:48,0,26     1/1:0,6:6:9.4:75,19,0
chr1  19   .   G    A    124.34  .       [INFOS]  GT:AD:DP:GQ:PL  0/0:5,0:5:20.2:0,42,94    ./.
```

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


```
scaffold376 500    .     CTT    C      424.60      .
...

scaffold376 500    .     C      CT     434.60      .
...
```

- 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

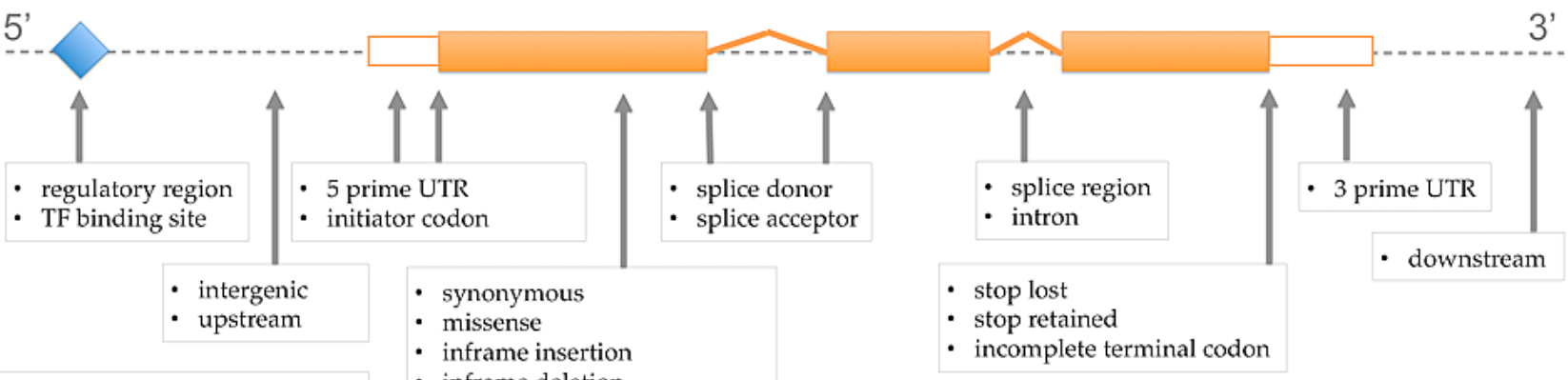
The VCF header

```
##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 11     12
```

Header line

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

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)

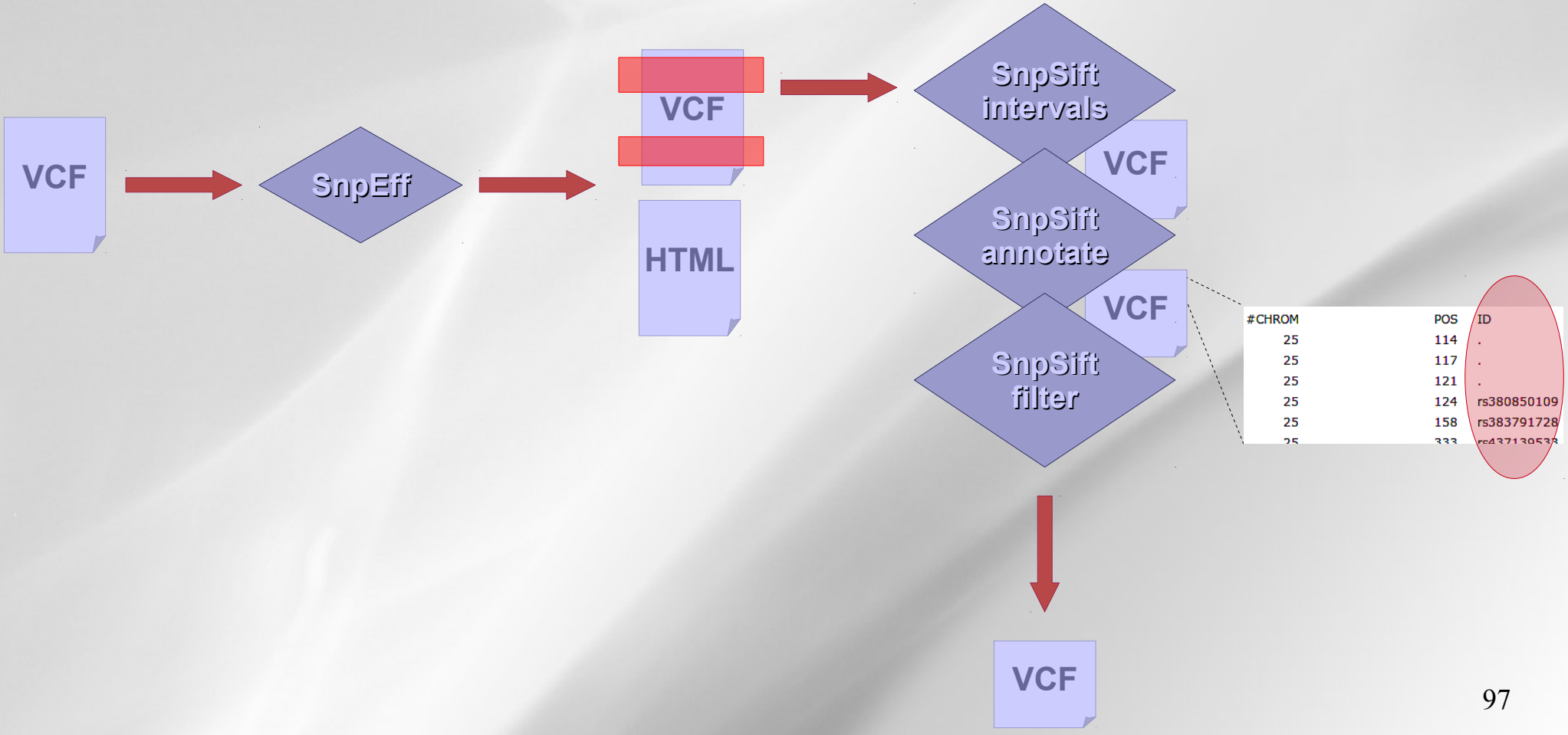
Table 4. Information provided by SnpEff in variant call format (VCF)

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
- SnpSift Annotate SNPs from dbSnp
- SnpSift Filter : 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])**

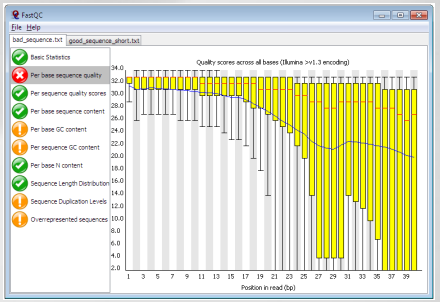
Synthesis

SRA
ENA
[...]

FASTQ

Quality analysis

FastQC



BWA
(aln / bwsw /
mem)

Mapping

SAM

SAM
manipulation

samtools
(view/merge/
sort/...)

BAM

Pre-processing BAM files

Picard
Rmdup/addRG

GATK
realignment

BAM

GATK
recalibration

BAM

GATK

Variant calling

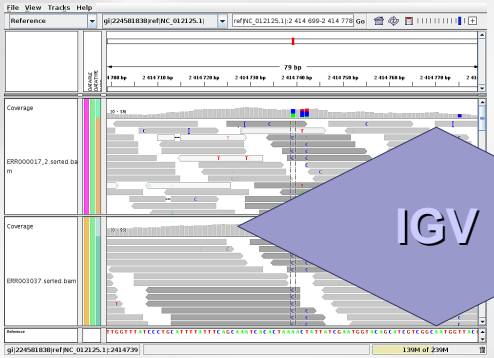
VCF

Variants
annotations/filters

SNPeff / SNPsift

Visualization

IGV



Exercise / set 5

- ***Filter, annotate variants***