
Mapping Exome-seq data

Elodie Girard

U900 INSERM - Mines ParisTech - Institut Curie

ITMO Roscoff – 11/19/2013

Dataset

- Public data: exome sequenced by the International HapMap Project
- Single-end reads of 100bp, Illumina Genome Analyzer Ix
- RNA-seq data of this exome available (Pickrell *et al.*, Nature, 2010)

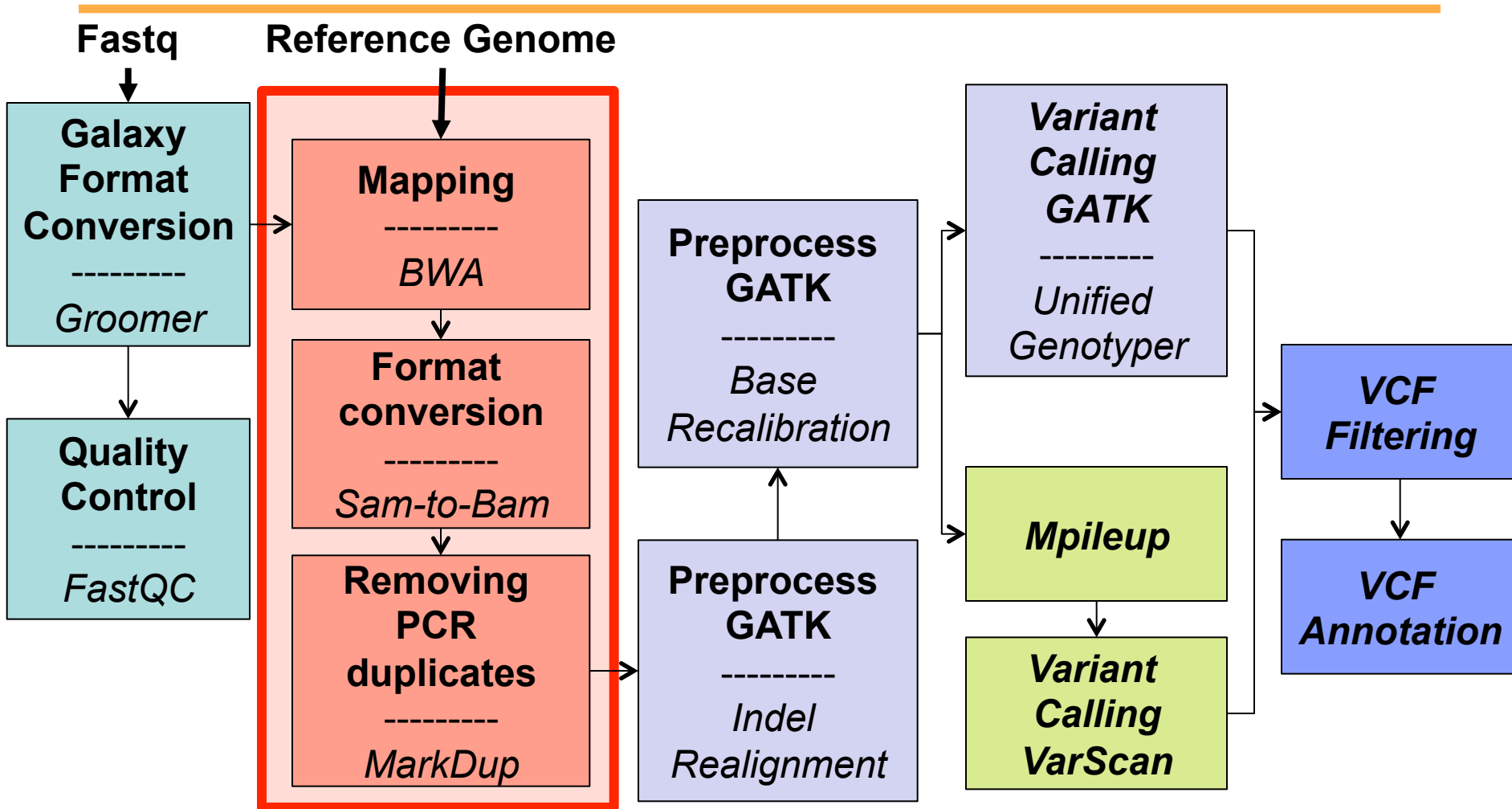
Objectives of the workshop:

- Variant calling, filtering and annotation in exome-seq data
- Observing the potential impact of these variants by looking at the corresponding RNA-seq data

Objectif of this session :

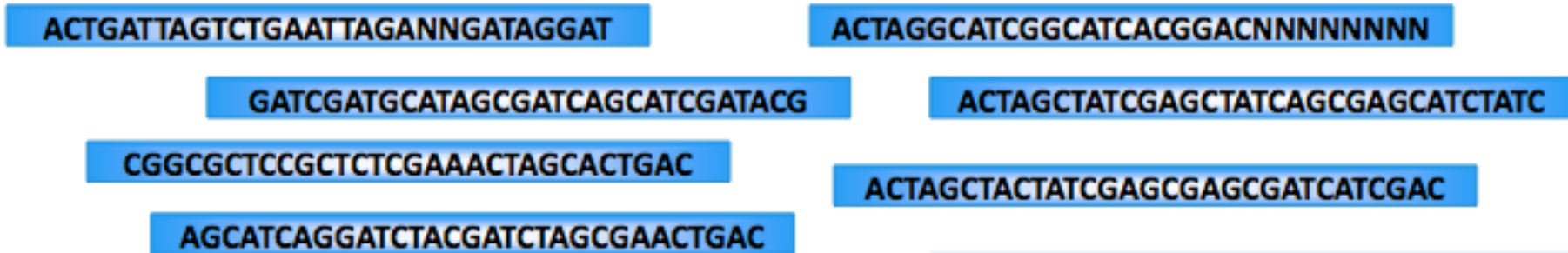
- Mapping the exome-seq data to the reference genome

Workflow



Raw data

- **Raw data:** millions of short reads of the same size (Illumina) or different sizes (Ion PGM), single-end or paired-end

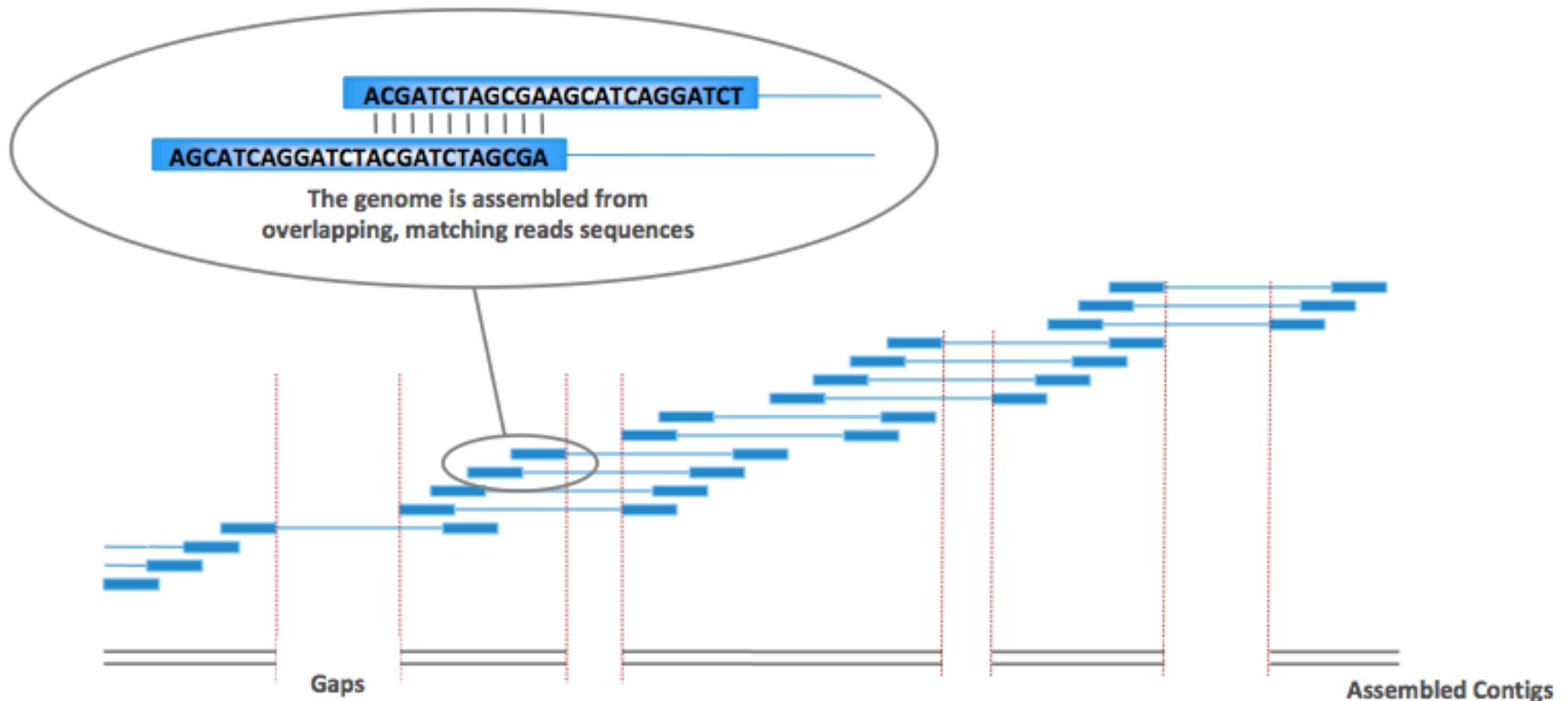


Example: fastq format

```
@SRR081222.573928
CATTCCCTAGAAGGCTGGCCTGCTCTTGCCTCTCTGTCCATTCCCTAGAAGGCTGGCTGGCCC
+
5@@-BFFGDFHHIJIIIIHHJGIIJHJHJJGHHIHHIHHIJJGAIJJFJI%%%%%%%%%
```

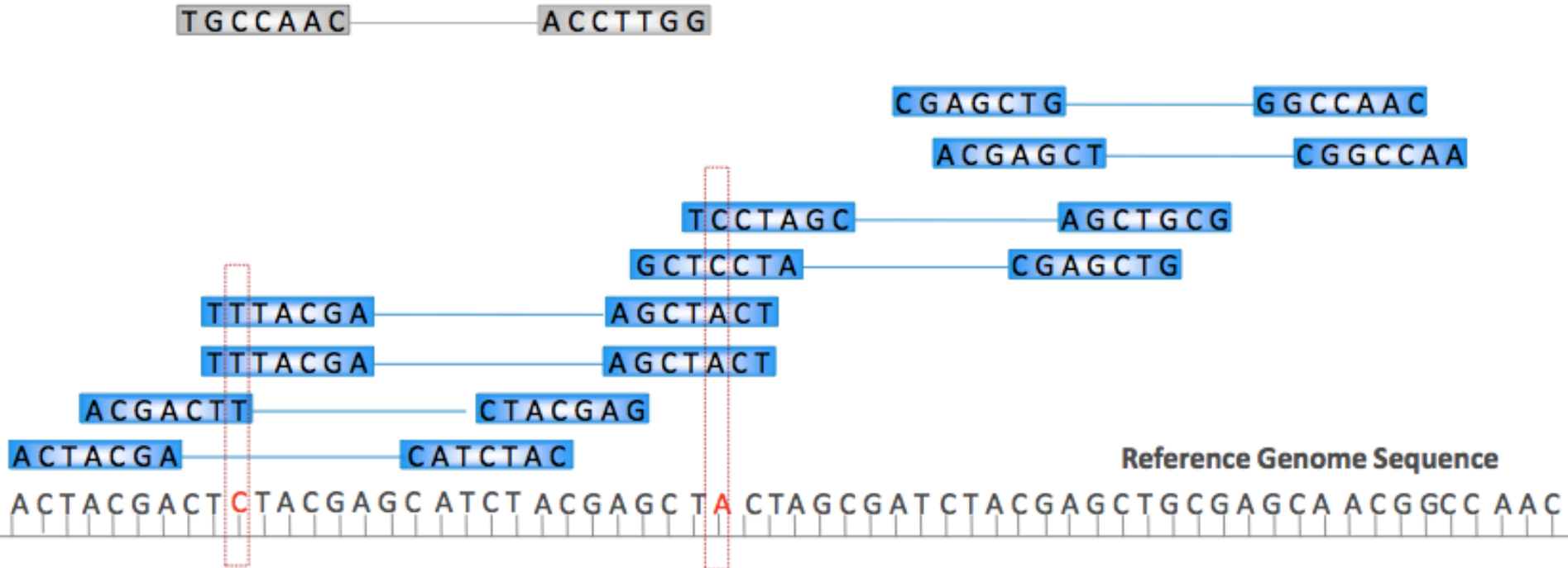
De Novo Reads Assembly

- **De Novo Reads Assembly:** used when there's no reference genome ; aims at reconstructing long scaffolds from single reads



Mapping on a reference Genome

- **Reference genome:** known sequence supposed to be **as close as possible** to the input genome and which is used as an **anchor** to organize the reads information



Reads Alignment - Vocabulary

- **Mismatch:** Incoherence between two nucleotides
- **Indels:** Insertion/Deletion into the reference genome
- **Gap:** Bridge within the read alignment (i.e. small indels)
- **Mappability:** Uniqueness of a region
 - repeated region = low mappability
 - unique region = good mappability

Global Alignment

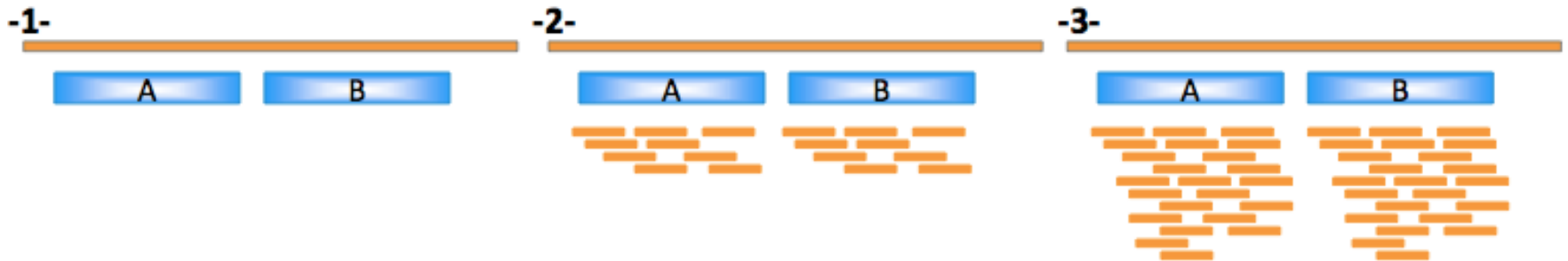
```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
|  || |  || |  |  |||  |  |  |  |  |||  |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

Local Alignment

```
                tccCAGTTATGTCAGgggacacgagcatgcagagac
                ||| ||| ||| ||| ||| ||| ||| |||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Multiple Alignments

- A read can align **multiple times** on the genome (repeated elements...)
- How to deal with these multiple alignments reads ?
- Three strategies:
 - 1- Report only unique alignment
 - 2- Report best alignments & randomly assign reads across equally good loci
 - 3- Report all (best) alignments

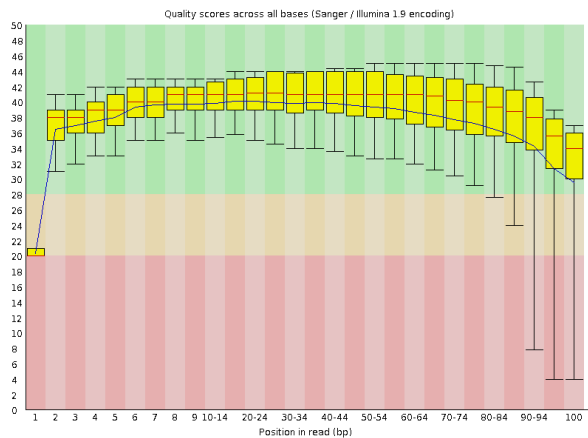


- **Mapping Quality:** quantify the probability that a read is misplaced. Low if a read has multiple alignments

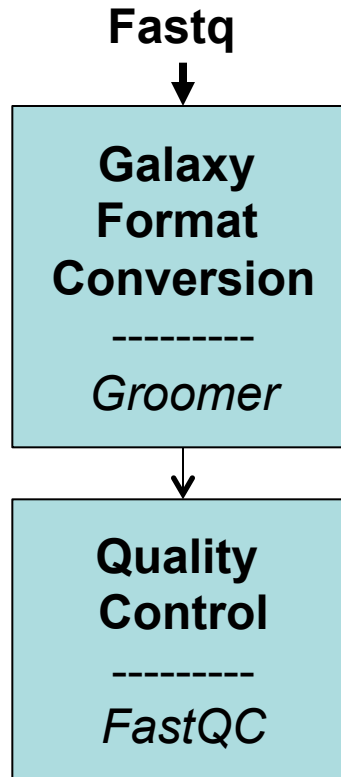
Alignment tools




- Multitude of alignment tools: BWA, Bowtie, Bowtie2, Bfast...
- How to choose the best tool ?
 - Is my sequencing technology supported ?
 - Do I have short or long reads ? Reads of different sizes ?
 - Do I want to allow gapped alignment ? Multiple alignments ?
 - Does it support single/paired-end reads ?
 - On which alignment algorithm is it based ?
 - Computational issues ? Is it used by the community ?
- A classical and performant tool for Illumina sequencing: BWA (Burrows-Wheeler Aligner)




Galaxy: summary of the previous steps






Number of reads: 5,664,374



3: FastQC FASTQ   
Groomer on data 1.html

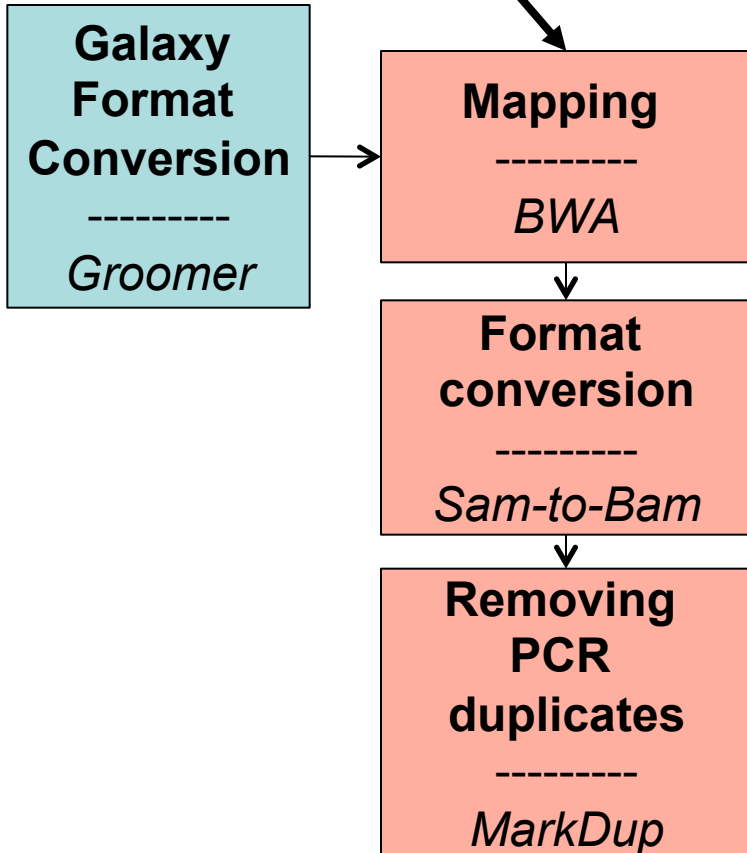
2: FASTQ Groomer on   
data 1

1:   
pickrell exon chr12.fastq

Input file of this session: FASTQ Groomer on data 1

Tools

Reference Genome



All the tools are in the left panel:
Mardi 19 Alignement des données exome-seq

ECOLE GGB GROUPE 2

Mardi 19 Pre-traitement des
donnees exome-seq

Mardi 19 Alignement des donnees
exome-seq

Map with BWA for Illumina

SAM-to-BAM converts SAM
format to BAM format

flagstat provides simple stats on
BAM files

Mark Duplicate reads

Loading the Reference Genome

1 : Go to « Shared Data » and « Data Libraries »

The screenshot shows the Galaxy/GGB interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', and 'Visualization'. The 'Shared Data' menu is open, showing options like 'Data Libraries', 'Published Histories', 'Published Workflows', 'Published Visualizations', and 'Published Pages'. The 'Data Library' section is titled 'tp-mardi-variants_exome'. A list of datasets is shown, with 'chr12.fa' selected. At the bottom, the 'Import to current history' button and the 'Go' button are highlighted.

2 : Choose this library

3 : Select « chr12.fa »

4 : Click on « Go »

Mapping with BWA

Map with BWA for Illumina (version 1.2.3)

Use a built-in index

✓ Use one from the history

genome from your history or use a built-in index?:

← 1 : « Use one from the history »

Select a reference from history:

1: chr12.fa

← 2 : Reference Genome = « chr12.fa »

Is this library mate-paired?:

Single-end

FASTQ file:

3: FASTQ Groomer on data 2

← 3 « FASTQ Groomer »

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:

Full Parameter List

← 4 « Full Parameter list »

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

Mapping with BWA

Adding a **Read Group** is required for the Variant Calling with GATK

- Informations on the sample, library used, platform...
- Attribute a read to a sample (useful in multi-samples study)
- Other tools exist to add Read Groups

At the bottom of the parameters list

Maximum occurrences of a read for pairing (sampe -o):

For paired-end reads only. A read with more occurrences will be treated as a single-end read. helps faster pairing

Specify the read group for this file? (samse/sampe -r):

Select « Yes »

Suppress the header in the output SAM file:

BWA produces SAM with several lines of header information

Execute

Mapping with BWA

Fill every box noted as « Required »

Specify the read group for

Yes

Read group identifier (ID).
alignment records. Must be

ID1

Required if RG specified. Rea

Library name (LB):

Library1

Required if RG specified

Platform/technology used to produce the reads (PL):

ILLUMINA

Required if RG specified. Valid values : CAPILLARY, LS454, ILL

Platform unit (PU):

PU1

Optional. Unique identifier (e.g. flowcell-barcode.lane for Illu

Sample (SM):

Pickrell

Required if RG specified. Use pool name where a pool is being

Suppress the header in the output SAM file:

BWA produces SAM with several lines of header information

Execute



Click on Execute

SAM-to-BAM

- **SAM Format:**

```
@SQ SN:chr12 LN:133851895
```

```
@RG ID:1 LB:sample PL:ILLUMINA SM:sample
```

Read Name / Flag / chr / Pos / MapQ / Cigar

```
SRR081222.573928 16 chr12 60124 23 100M * 0 0
```

```
GCCCCTGGGGATGTTTTGCACCAAGCCACTGTCTCCAGCTGG
```

```
BBC@GIIHGCFIEHEAIEIFFGEONDNJFINIONHNGJNNNNKNJN
```

```
RG:Z:1 XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:100 XA:Z
```

Read group affiliation

- **BAM Format:** Binary SAM Format (compressed = smaller size)

SAM-to-BAM

SAM-to-BAM (version 1.1.2)

Choose the source for the reference list:

History

← **1: Choose History**

Convert SAM file:

7: Map with BWA for Illumina on data 3 and data 1: mapped reads

← **2: Mapping output « MAP with BWA »**

Using reference file:

6: chr12.fa

← **3: « Chr12 .fa »**

Execute

← **4: Click on Execute**

Mapping Statistics

Flagstat: a tool showing some mapping statistics (from the SAMtools software)

flagstat (version 1.0.0)

BAM File to Convert:

8: SAM-to-BAM on data 1 and data 4: converted BAM

Execute

Choose the Sam-to-BAM output and execute

```
5664374 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
5523578 + 0 mapped (97.51% of mapped reads)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (-nan%:-nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (-nan%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Information on pairs
when available

Removing Duplicates

- **Duplicates reads**: different reads having the same sequence
- PCR amplification causes **molecular duplicates** and sequencing artifacts cause **optical duplicates** (*e.g.*: same cluster read twice)
- 2 Tools to **mark** (only a flag is added) or **remove** duplicates:
 - **Rmdup** (SAMtools): efficient with single-end reads
 - **MarkDuplicates** (Picard): efficient with single/paired-end reads (takes into account the 5' coordinates, the mapping orientation and all gaps in the alignment)
- The removal of the duplicates depends on the application (not appropriate for ion PGM targeted sequencing by example)

Removing Duplicates

Mark Duplicate reads (version 1.56.0)

SAM/BAM dataset to mark duplicates in:

8: SAM-to-BAM on data 1 and data 4: converted BAM

If empty, upload or import a SAM/BAM dataset.

1: Choose the SAM-to-Bam output

Title for the output file:

Dupes Marked

Use this remind you what the job was for

Remove duplicates from output file:



2: Check this box to remove duplicates

If true do not write duplicates to the output file instead of writing them with

Assume reads are already ordered:


















If true assume input data are already sorted (most Galaxy SAM/BAM should

Execute

3: Click on Execute

Removing Duplicates

10: MarkDups Dupes Marked.html			
9: MarkDups Dupes Marked.bam			
8: SAM-to-BAM on data 1 and data 4: converted BAM			
7: Map with BWA for Illumina on data 3 and data 1: mapped reads			
6: chr12.fa			

1: Click on the eye of the « html » output

2: Click on [MarkDuplicates.metrics.txt](#)

[MarkDuplicates.metrics.txt](#)
[MarkDuplicates.log](#)

Picard on line resources

- [Click here for Picard Documentation](#)
- [Click here for Picard Metrics definitions](#)

```
## METRICS CLASS net.sf.picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS UNMAPPED_READS
UNPAIRED_READ_DUPLICATES READ_PAIRS
PERCENT_DUPLICATION ESTIMATED_LIBRARY_SIZE
Unknown_Library 0 0 140796 0 0 0 ?
sample 5523578 0 1835477 0 0 0.332299
```

#Mapped	#Unmapped	#Duplicates	% Duplicates
5523578	140796	1835477	0.332299

Visualization with IGV (Integrative Genomics Viewer)

9: MarkDups Dupes
Marked.bam
307.1 MB



1: Click on the pencil to edit the attributes of this bam

Edit Attributes

Name:

MarkDups_Dupes Marked.bam

Info:

2: Edit the Database/build by choosing « Human Feb.2009 (hg19) » genome

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when

Database/Build:

Human Feb. 2009 (GRCh37/hg19) (hg19)



Save

3: Click on Save

Visualization with IGV

9: MarkDups Dupes
Marked.bam

307.1 MB

format: bam, database: hg19



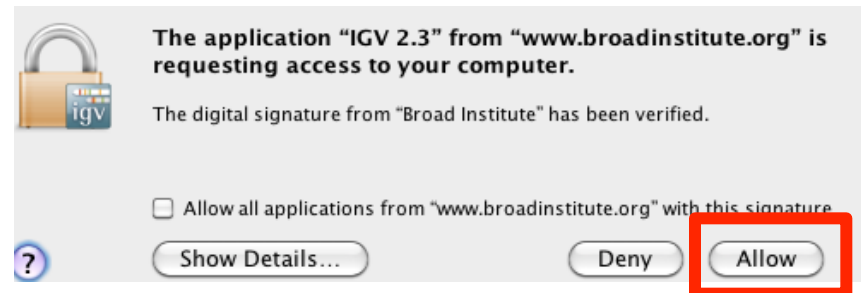
display at Ensembl Current

display with IGV web current local

display in IGB Local Web

1: Click on the name to open the options

2: Click on « web current » to download and open IGV



3: Allow IGV

Once IGV is opened, don't click on « web current » but on « local » to visualize an other BAM

Visualization with IGV

Enter « chr12:113,342,800-113,359,800 » and click on « Go »



Next Step

