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 IIx
• 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

Fastq

Reference Genome

- Galaxy Format Conversion
  - Groomer
- Quality Control
  - FastQC

Mapping
- BWA

Format conversion
- Sam-to-Bam

Removing PCR duplicates
- MarkDup

Preprocess GATK
- Base Recalibration

Variant Calling GATK
- Unified Genotyper

VCF Filtering

VCF Annotation

Preprocess GATK
- Indel Realignment

Mpileup

Variant Calling VarScan

Institut Curie

Inserm

MINES ParisTech
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
CATTCCCTAGAAGGCTGGCTGGCCC
+ 5@@-BFFGDFHHIIJIIJIIIHHJGIIJHJHJGHIIIHGIHIJGAIIJJJFJI%*%*%*%*%*%*%*%*%*%
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**

```
tccCAGTTATGTCAGgagcacgagcatgcagagac
```

```
aattgcgccgtcgttttcagCAGTTATGTCAGatc
```
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

Input file of this session: FASTQ Groomer on data 1
Tools

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 »

2: Choose this library

3: Select « chr12.fa »

4: Click on « Go »
Mapping with BWA

1: "Use one from the history"
2: Reference Genome = "chr12.fa"
3: "FASTQ Groomer"
4: "Full Parameter list"
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

Select « Yes »
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.

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

Click on Execute
SAM-to-BAM

- **SAM Format:**

  @SQ SN:chr12 LN:133851895
  @RG ID:1 LB:sample PL:ILLUMINA SM:sample

<table>
<thead>
<tr>
<th>Read Name / Flag / chr / Pos / MapQ / Cigar</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR081222.573928  16 chr12  60124  23  100M * 0 0</td>
</tr>
<tr>
<td>GCCCCTGGGGATGTTTTGCACCAAGCCACTGTCTCCAGCTGG</td>
</tr>
<tr>
<td>BBC@GIIHGCFCKCIEHEAIEIFFGEONDJFINIONHNGJNNNNKNNJN</td>
</tr>
</tbody>
</table>

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

Choose the source for the reference list:
1: Choose History

Convert SAM file:
2: Mapping output « MAP with BWA »

Using reference file:
3: « Chr12 .fa »

Execute
4: Click on Execute
Mapping Statistics

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

Choose the Sam-to-BAM output and execute

% of mapped reads

Informations 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

1: Choose the SAM-to-Bam output
2: Check this box to remove duplicates
3: Click on Execute
Removing Duplicates

1: Click on the eye of the « html » output
2: Click on MarkDuplicates.metrics.txt

Picard on line resources
- Click here for Picard Documentation
- Click here for Picard Metrics definitions

#Mapped
#Unmapped
#Duplicates

% Duplicates
Visualization with IGV (Integrative Genomics Viewer)

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

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

3: Click on Save
Visualization with IGV

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 »