PHYLOGENY – TME7

2022-2023

Marina Abakarova & Lélia Polit
marina.abakarova@etu.sorbonne-universite.fr
lelia.polit@gmail.com

12 January 2023

General rules

- Reports must be sent by e-mail, using the subject "`[PHYG] TME7`", including in the body the names of the persons who worked on it (maximum two students per group). The deadline is 19th of January.
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`).
- Your report *must be* in PDF format and named `student1_student2_TME7.pdf`. It should be simple, clear and well organized. Answers should be given in an exhaustive manner. Consider adding at the beginning a summary indicating the page of each answer.
- Source code must be well explained, commented and, most importantly, it should work without errors. Provide all needed information (*e.g.,* compiler/interpreter version) in a `README` file.
- All required materials can be found in the repository `https://github.com/abakarovaMarina/PHYG2022.git`
- A discord server is created so we can exchange our questions, answers and comments `https://discord.gg/vDmUHtDW`.

In this TME we introduce to you Next Generation Sequencing (NGS) data and its analysis. You will manipulate huge files and use specific tools, so we recommend using stationary machines to accelerate the process.

Otherwise here are the links to download the needed softwares or packages :

- Samtools: `http://www.htslib.org/download/`
- mosdepth: `https://github.com/brentp/mosdepth`
- IGV: `https://software.broadinstitute.org/software/igv/download`

At the beginning of this TME you will work with **bam files**. They are obtained after the mapping/alignment step (see figure 3). Before sequencing, during library preparation, a step called PCR can introduce GC-bias, GC-rich and AT-rich fragments are under-represented in the output of the sequencing step. Thus, one needs to correct this, as it is a very time-consuming step, it has already been performed, the bam files are "GC_corrected".

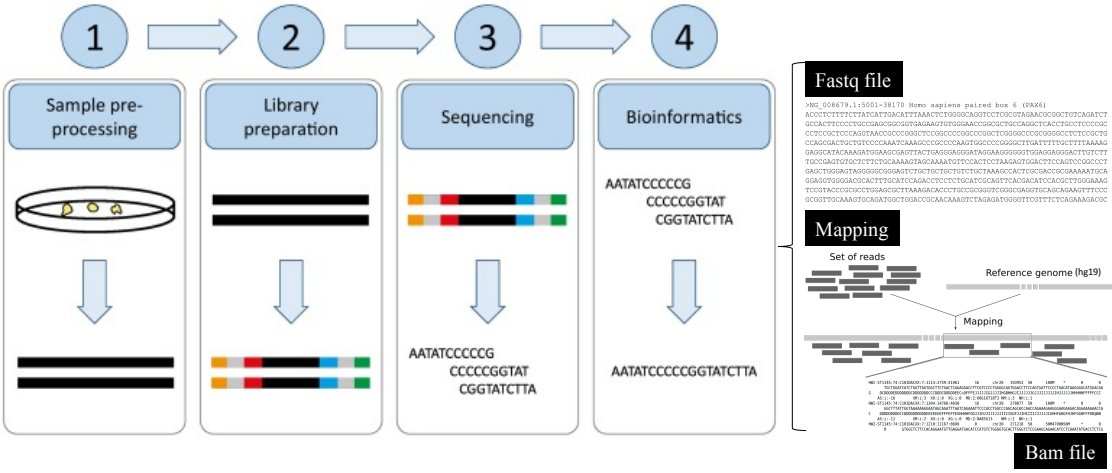

Figure 1: From DNA molecules to reads

Get the data

As mentioned previously the files are huge, thus you will only work on two (small) chromosomes: chromosome 14 and chromosome 16. You should create a directory with the following path `mkdir -p /Vrac/STUDENT_NUMBER/PHYG/TME7/` and download there `.bam` and `.bai` files for *Chagyrskaya8* (Neandertal) and *Ust'-Ishim* (Homo Sapiens) ancient humans from:

- `https://drive.google.com/drive/folders/1OddtN11e7Y76ZZnhiJBpjFhV_aGKRZ9B?usp=share_link` for chromosome 14
- `https://drive.google.com/drive/folders/1HgJR6e2sytDE7vjNgdUq_YO3Qe29TZCD?usp=share_link` for chromosome 16.

Exercise 1 – Visualisation of data on Integrative Genomics Viewer (IGV)

Load the `.bam` files for chromosome 16 for the two individuals in Integrative Genomics Viewer on the human reference genome *hg19* (you can select the assembly under the "File" menu, next to it you can also select the chromosome of interest after loading). For each individual you can see the distribution of reads (top row) and the reads themselves below.

1. Using the field of research next to the chromosome selection, go to the gene SLX1A (chr16) and have a look at the distributions' range for each individual - in particular, observe the range of coverage. Do you see any differences?
2. On the RefSeq Genes field (bottom left), right-click and select "Expanded". Look at the genes around SLX1A: BOLA2B and SULT1A3.
3. Go now to the region "chr16:29464743-29476530" (respect case!) using the white field to the left of the "Go" button. Look at the names of the genes. Do you see a similarity with the previous region? Comment it.
4. Go now to the small region "chr16:30206016-30206117". How many base pairs are contained in this region?
5. What do the letters indicated in color on the reads correspond to? Discuss the difference between the two ancient genomes.

Exercise 2 – Mosdepth

Mosdepth is a command-line tool for rapidly calculating genome-wide sequencing coverage. It measures depth from BAM or CRAM files at either each nucleotide position in a genome or for sets of genomic regions.

```
mosdepth -x --chrom [CHROM_NAME] [OUTPUT] [corrected BAM]
```

`-x` option is for fast mode

1. Run mosdepth for chr14 and chr16.
2. How many output files are there for each chromosome?
3. We are only interested in the file "chr*.per-base.bed.gz". Have a look at the file and comment it using: `gunzip -c file | head`

Exercise 3 – Bedtools, depth of coverage

Documentation: `https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html`

For the moment, we only have the **coverage values** at **each base pair**, however we want to associate these coverage values with genes.

For this you will use the `intersect` tool from the `bedtools` suite. It allows making the **intersection between genomic intervals**. We will do the intersection between the annotated genes on these chromosomes and the coverage values computed at each base pair along chromosomes 14 and 16. Two additional files are needed:

- `GRCh37.genome.txt` containing chromosomes' positions along the entire genome.
- `Homo_sapiens.hg19.genes.protein_coding_noDuplicats_sort.bed` containing genes' position. The file is composed of 4 columns `chr start stop ENSEMBL_gene_ID`
- `Homo_sapiens.hg19.87.genes.id2name.tsv` with gene_id, gene_name, gene_type

Here is the command line you will need:
`bedtools intersect -g GRCh37.genome.txt -a .bed -b .bed.gz -wa -wb | gzip -c > OUTPUT.gz`

1. What is every option for ?
2. Using `awk` find how many genes you can possibly intersect for chromosome 14 and chromosome 16? And what
3. How many genes have you effectively intersected?
4. Calculate mean depth for chromosomes 14 and 16 using the formula 1 and mean depth for each gene using the formula 2.

$$C^i = \frac{\sum_{j \in chr_i} doc_j^i}{L_i^{hg19}} \tag{1}$$

$$C_g^i = \frac{\sum_{j \in g} doc_j^i}{L_g} \tag{2}$$

where doc_j^i is depth of coverage computed by the program mosdepth at position j of a chromosome i. Pay attention to the boundaries of the coverage in the output of mosdepth.

Exercise 4 – Copy Number Variation

A gene's haploid copy number value is the ratio of the gene coverage and the chromosome coverage. Paralogs are gene copies created by a duplication event within the same genome.

1. Write a Python script to compute genes' haploid copy number value.
2. The `human_paralogs_hg19.txt` file contains the gene paralogy information for the hg19 human genome assembly. Find `SLX1A`, `SULT1A3` and `BOLA2` genes in it. Link to question 3 of exercise 1.

3. Two genes or more are considered paralogous if their percentage sequence identity is greater than 80%. What are the copy number values of AMY1, NOTCH2NL and PGA genes in hg19?

4. (Optional) When two genes are paralogous, add their copy number. Use this information to estimate to copy number value of SULT1A3/4, BOLA2/2B and SLX1A/B genes.

Annexe

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Figure 2: BAM file format explanation

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

Figure 3: CIGAR string options