**Practical Assignment MEMORANDUM**

**Module topic:** Genomics

**Contact session title:** Sequencing technologies and NGS Overview

**Trainer:** Fatma Guerfali

**NGS technologies: from theory to practice**

**Introduction**

*Next-Generation Technologies (NGS) is a catch-all term used to describe many different sequencing technologies and platforms. Although they are very different in their intrinsic specificities and their potential applications, their frequent goal for DNA-seq is to retrieve relevant information about variants. Following an analysis pipeline adapted to the biological question and the organism of interest, series of output file formats are generated at each step that allow to end up with vcf files, useful to screen for these variants. We will both cover here theoretical aspects and practical aspects in interrogating these files*

**Tools used in this session**

*http://www.illumina.com/*

**Please note**

* **Hand-in information** If you are formally enrolled in the IBT course,please upload your completed practical assignment to the Vula ‘Practical Assignments’ tab. Take note of the final hand-in date for each practical assignment, which will be indicated on Vula.

**Task 1: instructions: Basics of NGS (Next Generation Sequencing) technologies**

**Task 1: Open your browser and navigate to the Illumina webpage (http://www.illumina.com/).**

*1. Explore the website to look for information on the right device to use if you are interested in sequencing “Small Whole-Genome Sequencing”.*

*NB: Small genome sequencing (≤ 5 Mb) involves sequencing the entire genome of a bacterium, virus, or other microbe, and then comparing the sequence to a known reference (illumina).*

*1.a. Give the name of the suggested kit to perform this sequencing.*

*1.b. Give the name of the suggested devices to perform this sequencing.*

*2. Have a look at the “Nextera XT DNA Library Prep Kit” specifications.*

*2.a. Is it optimized for small genome sequencing?*

*2.b. What is particular about this kit, considering the fragmentation and adapter ligation procedure?*

*2.c. What is the minimum amount of material you could use as an input using this kit?*

*3. What are the basics of the sequencing technology used by Illumina (sequencing name and PCR amplification technique)? What is the difference between that type of sequencing and the Sanger sequencing?*

**Task 1: memo answer**

*1.a. The sequencing is recommended using the Nextera XT DNA Library Prep Kit*

*1.b. The sequencing is recommended using the MiSeq Series devices.*

*2.a. Yes*

*2.b. DNA is simultaneously fragmented and tagged with sequencing adapters in a single tube enzymatic reaction.*

*2.c. Nextera XT supports ultra-low DNA input of only 1 ng.*

*3. Sequencing-by-Synthesis and bridge PCR. Light detected along with the elongation of DNA, in real time, to determine the sequence. No need for a gel.*

*Sanger sequencing uses modified bases called ddNTP that terminate the chain elongation (lack a 3’OH). 4 different reactions, need for a gel to retrieve the sequence.*

**Task 2: instructions: Understanding the different NGS file formats**

**Task 2: We will explore the different file formats covered during the course**

*1. What kind of information is contained in a fastq file?*

*2. What is the main difference between a SAM or “.sam” file format and a BAM or “.bam” file format ? Gives examples of some of the most relevant information given in the SAM or BAM file.*

*3. At what stage of the analysis would the BAM file be generated?*

*4. At what stage of the analysis would the BAM file be required?*

*5. For what kind of variant screening would you require to generate a “.vcf” or VCF file? What is the main information included in this file under the REF and ALT fields?*

**Task 2: memo answer**

*1. A fastq file contains information about the read : unique identifier containing many kinds of informations + the read sequence + the quality associated to that read.*

*2. A BAM file format is the compressed binary counterpart of the SAM (Sequence Alignment Map) file.*

*Information stored contains read identifier, flag, CIGAR string, read sequence, quality score…*

*3. After the alignment of reads to the adequate reference genome.*

*4. For several other steps like genome coverage or the screen for large or small genomic variants.*

*5. This file is useful for variant calling such as SNPs and InDels. It contains a REF and ALT fields that give information about the nucleotide sequence in the reference in the REF field and the alternative nucleotide (mutation) in the ALT field.*

**Task 3: instructions: Manipulating a vcf file**

**Task 3: We will interrogate an annotated vcf file for specific patterns or fields of interest. For this you will need to have the “test\_Session1.vcf” file loaded on the directory you chose to work in. You will need to interrogate the file using a command-line interface. You are asked to write as an answer the exact command line you typed on your screen to address these questions.**

*1. What command line would you use to view the 20 first lines of this test.vcf file?*

*2. What command line would allow you to search for the ‘CTAGAG’ pattern?*

*3. What command line would allow you to search for the number of lines that contain the ‘CTAGAG’ pattern?*

*4. How many lines contain the ‘chr1’ pattern? (even if embedded in another pattern like chr12)*

*5. How many pseudogenes are recorded in this file?*

*6. How many mutations affecting a pseudogene are reported in this file?*

*7. What is the mutation appearing in this pseudogene?*

*8. What information is contained in the GT field information for the 1th sample?*

*NB: Remember that in this file, you have data for 4 different samples.*

*(sample1:1094PC0005 ; sample2:1094PC0009 ; sample3:1094PC0012 ; sample4:1094PC0013)*

**Task 3: memo answer**

*1. head -n 20 test\_Session1.vcf*

*2. grep CTAGAG test\_Session1.vcf*

*3. grep CTAGAG test\_Session1.vcf | wc -l*

*4. 29*

*5. 0 (none)*

*6. 1*

*7. G to C*

*8. You can do this in 2 steps in that case, and you can easily pipe this into one command line:*

*cut -f10 test\_Session1.vcf | cut -d':' -f1 > test\_Session1\_cut.vcf*