NGS2 Project only as we completed NGS1 last year

Team leader : Manar Aleslam Mattar (181027)
Group member: Manar Hashem (181013)
Methods (Sequencing analysis) :

##Main issue Manar Hashem has many problems with her laptop and changed it two days before the deadline. Also, even the new laptop has a limited memory and she couldn't complete all steps and stopped after alignment.

1- Download dataset:

Download three samples whole exome sequencing of Irish HER2+ breast cancer patients from SRA SRR7309332, SRR7309338, SRR7309325 available at (https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=475755) The the library comprised of the ERBB-family genes and included 132 regions of coding exons of the ERBB family genes, and the layout is paired.

2- Download reference genome:

- At first, we downloaded the whole genome however my lab crashed, therefore we thought about selecting the chromosomes that represent ERBB family genes (2,7,12,17) to be a reference and concatenate them in a single Fasta file. Available at : (ftp://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/dna/)

Second issue, there are multiple fasta file formats to download; We were confused between which format is the best for alignment. T**he files with "sm" in the name are "soft masked" or without masking is the best and any file with "rm" in the name should be avoided.**

The answer available in these websites:

(http://genomespot.blogspot.com/2015/06/mapping-ngs-data-which-genome-version.html)

3- Download VCF annotation file:

We downloaded the VCF files for chromosomes 2, 7, 12, 17 and concatenate them in one file. The data available in (ftp://ftp.ensembl.org/pub/release-99/variation/vcf/homo_sapiens/)

As in fasta there are multiple VCF format and we chose (homo_sapiens-chr*.vcf.gz) based on a readme file which recommends it for All germline variations from the current Ensembl release.

4- Check Quality of the data and Trimming in case of bad quality

The quality of data was checked using fastqc and resulted in that the data quality were very good so we didn't do trimming (QC files are attached in github)

5- Alignment with BWA.

We took an overview about different pipelines and aligners from this paper (Performance Assessment of Variant Calling Pipelines using Human Whole Exome Sequencing) to choose BWA aligner (https://www.biorxiv.org/content/10.1101/359109v1.full.pdf).

Alignment results :

For the first sample(SRR7309332), there were 50.35% of reads mapped and only 46.69% properly mapped. However, for SRR7309338, there were 58.49% of reads mapped and only 47.57% properly mapped. The last one, SRR7309338, there were 46.55% of reads mapped and only 42.69% properly mapped.

6- Duplicate reads were marked by Picard tools

Based on the files statistics, 265353 reads were marked as duplicates in SRR7309332, and 38802 reads were marked as duplicates for SRR7309338 sample. States for deduplicates step in SRR7309325 wasn't not done.

7- local realignment and base recalibration will be conducted with GATK

The base recalibration process involves two key steps: first one tool (BaseRecalibrator) builds a model of covariation based on the data and a set of known variants, then another tool (ApplyBQSR) adjusts the base quality scores in the data based on the model.

8- Joint variant calling which includes many steps:

A) assess genotype likelihood per-sample using HaplotypeCaller
B) combine samples using CombineGVCFs
C) Joint Genotyping using GenotypeGVCFs

It is performed only on 2 samples SRR7309338 + SRR7309332

9- Split SNPs and indels using (SelectVariants)

As we study the most common SNPs only

10- Assess the different filters in both known and novel to decide the threshold for the filtration in each filter.

The resulted graphs attached to github

10- SNP Variant filtration using VariantFiltration

11- separate the passed SNPs in a vcf file using vcftools with option (--remove-filtered-all)

##I did statistics on the passed SNPs vcf file

12- **extract known snps for further analysis in a text file:** to complete downstream analysis

Work contribution

	Manar Hashem	**Manar Aleslam**
Data analysis	**Analysis of** SRR7309325 **She has many problems related to her laptop therefore she didn't complete until the final step and analysed only one sample. She tried to solve the problems until the last hours before submission**	**Analysis of** SRR7309332, SRR7309338
Report	**Manar didn't write as she was still troubleshoot the problems of her analysis**	**100% written**