UNIVERSITÀ
DEGLI STUDI
DI MILANO

Master's Degree in Bioinformatics for Computational Genomics

Genomics - Final Project Report

Molecular Diagnosis of Rare Genetic Disorders in 5 Individuals

Professor Matteo Chiara

Alessandro Giulivo
matricola *11351A*

A.Y. 2022/2023

Contents

Figures

Tables

1 Introduction

1.1 Overview

In this final assessment project for the Genomics course at University of Milan, we applied experimental approaches studied during the course for the analysis and interpretation of human genomic data. In particular, we worked with **exome sequencing of chromosome 16** of five *TRIOs* of individuals (*mother, father, child*) where parents are known to be healthy, whilst the child is possibly affected by a **rare mendelian disease**.
By "rare mendelian disease" we mean a disorder which affects less than $\frac{1}{10^4}$ people, is caused by defects in one gene of either autosomes or sex-linked chromosomes, and is either recessive (one must have both copies of the flawed allele to be affected) or dominant (just one copy of the flawed allele can cause the disease). For our analysis, we will only consider diseases with *Autosomal Dominant (AD)* or *Autosomal Recessive (AR)* hereditary models.

The aim of the project was *to make a correct diagnosis for each child (out of the five TRIOs).*

1.2 The Data

The studied TRIOs are: **case 1642** (AR), **case 1608** (AD), **case 1765** (AR), **case 1682** (AR), **case 1705** (AR).

The majority of the workflow was performed on the *unix server* of the course, within the `BCG2023_agiulivo/finalProj` folder; a subfolder for each *case* was created with the `mkdir` command.
The data consists of:

- three `fastq` files for each case (raw DNA-sequencing reads of chr16 of the three individuals);
- a `universe.fasta` file along with its index files (our hg19 reference genome for chr16);
- an `exons16Padded_sorted.bed` file (which specifies the target regions).

The data were retrieved from the folder `BCG2023_genomics_exam`.

2 Methods

Figure 1 shows the complete pipeline carried out on each *"case"*; it will be illustrated in this section.

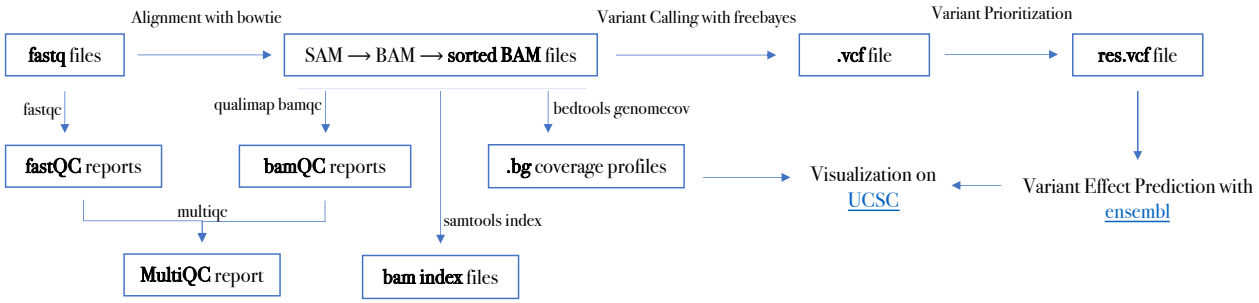

Figure 1: Workflow

2.1 Preprocessing and Variant calling

The first part of the analysis consisted in a few **pre-processing** steps:
(1) a quality control check of the reads with **FastQC**;
(2) alignment of the reads to the reference genome (`universe.fasta`) with **bowtie**. The output of this tool is in `SAM` format: we compress the `.sam` files into `BAM` format and sort the obtained `.bam` files;
(3) indexing of the latter with `samtools index`;
(4) quality control on the results with `qualimap bamqc`;
(5) computing coverage histograms of each individual's sequenced genome.

Then, for each case, all the **quality control reports** were put together in a single `html` report with **MultiQC**.

Finally, joint **variant calling** of the three individuals was done using `freebayes`.
The options used with this command were: *minimum mapping quality = 20*; *minimum alternate count = 5*; *mismatch base quality threshold = 10*; *minimum coverage = 10*; target regions to consider were specified in the `exons16Padded_sorted.bed` file; `universe.fasta` file was used as the reference genome with `bedtools genomecov`.

To perform the procedure mentioned above, the following bash script was saved in a `processCase.sh` file, and executed over each *TRIO*:

```
fastqc *.fq.gz                                                                    #(1)

for filename in *.fq.gz                                          #iterating over the three .fq files
do
    base=$(basename $filename .fq.gz)                            #filename variable
    case=$(echo ${base} | cut -f 1 -d "_")                       #case number variable
    ind=$(echo ${base} | cut -f 2 -d "_")                        #individual name variable

    echo "Aligning sample ${base}..."                                             #(2)
    bowtie2 -U ${base}.fq.gz --rg-id "${base}" --rg "SM:${ind}" -x ../uni | \
                                    samtools view -Sb | samtools sort -o ${base}.bam

    echo "Indexing sample ${base}..."                                             #(3)
    samtools index ${base}.bam

    echo "Running bamQC on sample ${base}"                                        #(4)
    qualimap bamqc --feature-file ../exons16Padded_sorted.bed  -bam ${base}.bam --outdir ${base}

    echo "Computing coverage profile on sample ${base}..."                        #(5)
    bedtools genomecov -ibam ${base}.bam -bg \
                                    -trackline -trackopts name=${ind} -max 100 > ${ind}Cov.bg
done
multiqc ./

echo "Variant Calling with freebayes..."
freebayes -f ../universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 --targets ../exons16Padded_sorted.bed \
                            ${case}_child.bam ${case}_father.bam ${case}_mother.bam  > ${case}.vcf
echo "Done"
```

2.2 Variant Prioritization Strategy

The `vcf` file obtained with the commands illustrated earlier lists all the genomic variants found in the three individuals by `freebayes`; the files for all the cases were checked to have the last three columns describing, in order, the variants of `mother`, `father` and `child`. In order to select specific variants of interest for the diagnosis of "child", knowing that parents are healthy, we need to exploit our knowledge regarding the hereditary model of the case:

- For **autosomal recessive (AR)** diseases (**case1642, case1765, case1682, case1705**), we need to search variants for which the child is homozygous (*1/1 in the `vcf` file*), whereas the parents are heterozygous (*0/1 in the `vcf` file*). This is done with the `grep` command and the results are saved in an output `case****_res.vcf` file. For example:

  ```
  grep "0/1.*0/1.*1/1" case1765.vcf > case1765_res.vcf
  ```

- For **autosomal dominant (AD)** diseases instead (**case1608**), we assume that a *de novo* mutation is the cause of the disease (*in the child, at least one allele must be different from any of the parents' alleles*), as the parents must be homozygous for the reference allele to be healthy. We look for a couple of patterns within our `vcf` file:
  - parents are homozygous for the reference allele, whilst child has at least one different allele;
  - parents have either one of two alternative alleles (both healthy), while child has another allele different from the two of the parents.

  ```
  grep "0/0.*0/0.*/1" case1608.vcf > case1608_res.vcf
  grep "[01]/[01].*[01]/[01].*/[23]" case1608.vcf >> case1608_res.vcf
  ```

2.2.1 Variant Effect Predictor

The obtained variants of interest for each case were uploaded on the Ensembl Variant Effect Predictor web tool (VEP). This tool uses gene annotations to infer the effect of the genetic variants listed in our `vcf` files.
RefSeq transcripts database was used for annotations; data about frequency of co-located variants were extracted from *1000 Genomes Global* and *gnomAD*; in order to possibly find which diseases are associated to any of our variants, we look for additional annotations which relate genes to *phenotypes*.

After running a VEP job for each of our cases, we obtained final results which were studied to make the diagnoses, shown and discussed in sections 3.2 and 3.4.

3 Results

3.1 Quality of the Data

The quality of the data was assessed using the **MultiQC reports** which were generated *for each case*. Overall, all samples had both sequencing quality (*phred score* > 28) and alignment coverage (*mean coverage* $> 10X$) high enough for the analysis to be performed.
Figure 2 shows an example. Full reports are available to download here.

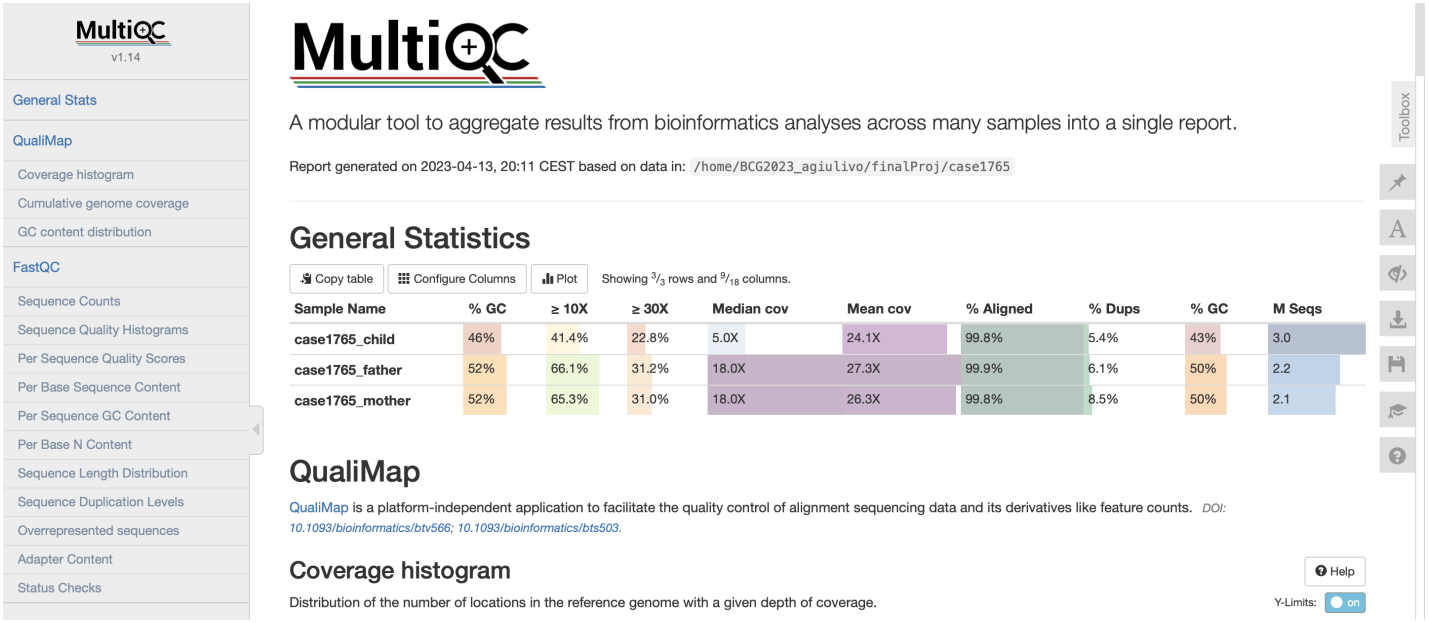

Figure 2: MultiQC Report for Case1765

3.2 Diagnoses

Variant Effect Predictor results provided us with a description of the phenotype effects caused by the variants identified in our analysis. In particular, **high impact variants** are the ones which most likely cause the disease we are looking for.
Table 1 shows the list of disease causing variants along with the diagnosed disease for each *case*.
These results are also discussed in Section 3.4

Table 1: Diagnoses.

CASE	LOCATION	REF	ALT	CONSEQUENCE	GENE	DISEASE
1642	16:89857825-89857828	ATA	A	Frameshift Variant	FANCA	**Fanconi Anemia complementation group A**
1608	-	-	-	-	-	**HEALTHY**[1]
1765	16:89882954-89882954	C	A	Stop Gained	FANCA	**Fanconi Anemia complementation group A**
1682	16:88907503-88907503	C	G	Splice Acceptor Variant	GALNS	**Mucopolysaccharidosis IV-A**
1705	16:53682877-53682877	G	T	Stop Gained, Splice Region Variant	RPGRIP1L	**Joubert Syndrome; Meckel-Gruber Syndrome**

[1]For case1608, a missense variant with moderate impact was found on the **CREBBP** gene (location: 16:3820629-3820629, REF: G, ALT: T): hence, a possible cause for Rubinstein-Taybi syndrome. However, only *PolyPhen* labelled the variant as "*possibly damaging*"; other pathogenicity predictors, .i.e., *SIFT* and *CADD*, classified it as "**tolerated**", "**likely benign**"; also, the allele frequency according to gnomAD is not very low ($>10^{-4}$). Thus, our final diagnosis for *case1608* was: **healthy**.

3.3 Visualizing the Variants on UCSC

The disease-causing variants, along with the coverage tracks, of each case were finally visualised on the UCSC Genome Browser. Figure 3 shows, as an example, the disease-causing SNV for *case1765*: the substitution of *C* with an *A* (in red) on the *FANCA* gene leads to a missense variant which disrupts gene function.

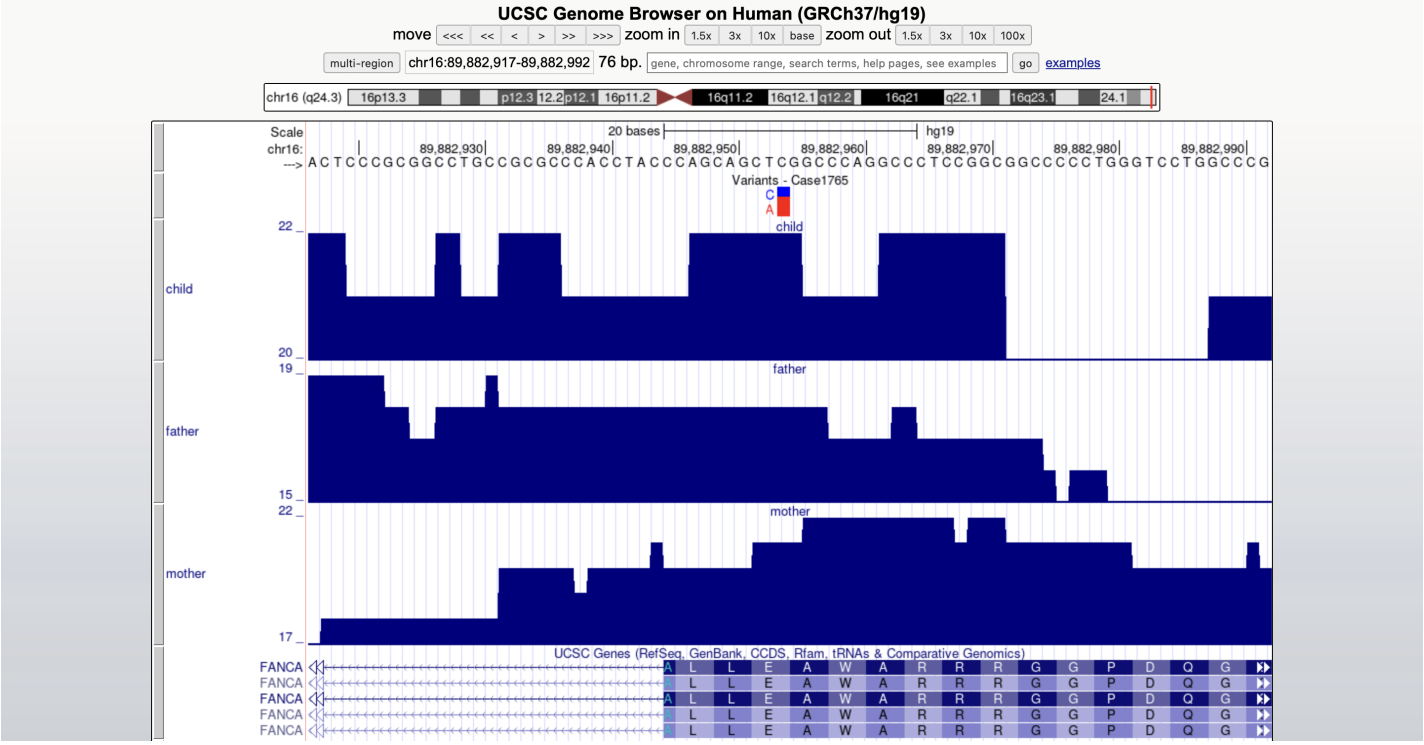

Figure 3: Case1765 disease-causing SNV on UCSC

3.4 Discussion

As the table of Section 3.2 illustrates, **four out of the five studied cases were found to have a high impact variant associated to a rare mendelian disease**.

For cases *1642* and *1765*, respectively a frameshift variant and a stop gained variant were found on the **FANCA** (Fanconi Anemia Complementation Group A) gene; mutations in this gene are the most common cause of **Fanconi Anemia**. It is a condition that affects many parts of the body, and causes bone marrow failure, physical abnormalities, organ defects, and an increased risk of certain cancers.

For case *1682*, a splice acceptor variant was found on the **GALNS** gene, which encodes galactosamine(N-acetyl)-6-sulfatase. Sequence alterations, including those that affect splicing, result in a deficiency of this enzyme, which in turn leads to Morquio A syndrome (**Mucopolysaccharidosis IV-A**). This disorder can affect an individual's appearance, organ function and physical abilities.

Then, a high impact variant for case *1705* was found on gene **RPGRIP1L**. The protein encoded by this gene is related to the Hedgehog Signaling pathway and to organelle biogenesis and maintenance. Defects in this gene are a cause of **Joubert-Meckel** syndrome, which is a lethal developmental syndrome characterized by posterior fossa abnormalities, bilateral enlarged cystic kidneys, and hepatic developmental defects.

Finally, instead, for *case1608*, no variants with a **high impact** were found to be associated with any rare disease. Therefore, we looked for variants with a moderate impact: one missense variant with moderate impact was found on the **CREBBP** gene (so, a possible cause for Rubinstein-Taybi syndrome). However, the frequency of this allele according to *gnomAD* is not very low (0.009; but should be $\leq 10^{-4}$ to be in accordance with the rareness of the diseases we are looking for). Moreover, only *PolyPhen* showed a significant score for the pathogenicity of the variant; other pathogenicity predictors such as *SIFT* and *CADD*, classified it as "tolerated", "likely benign". As a consequence, *case 1608* was diagnosed as: **healthy**.

Scripts, MultiQC reports, variant calling results and other data regarding this project are available in this GitHub repository.