UNIVERSITÀ
DEGLI STUDI
DI MILANO

MSC Bioinformatics for Computational Genomics
a.y. 2022/2023

GENOMICS PROJECT REPORT

Molecular diagnosis of rare genetic disorders in 10 individuals

Alessia Lorenzi, mat. 11613A
Lorenzo Monticelli, mat. 11271A

Index

1. Introduction

Hereditary diseases are caused by mutations in one or more genes in the DNA sequence, which lead to a gain or a loss of the full functionality of the transcript; **variants pass from parents to children** by autosomal and/or sex chromosomes and could be dominant (even *de novo*) or recessive. In order to search and identify those mutations, a useful approach is represented by the parent-child *trios* exome sequencing.
The **aim** of the study is to correctly **diagnose any possible genetic disease** that could affect a child out of a set of 10 cases.

1.1 Biological background

Mendelian diseases are conditions that result from mutation at a genomic locus and are inherited according to Mendel's laws. They have high penetrance, most of them are monogenic and individually rare, but there are almost 400 million people worldwide suffering from around 7000 different rare diseases. Even in cases where the causal disease gene is known, these diseases can display variable phenotypes, even in patients with the same mutation.
***De novo* mutations** are genetic modifications that can cause the insurgence of mendelian diseases. They are quite infrequent (~30/40 for each haploid genome), more likely to occur in the paternal germline and correlated with paternal age. Parents are usually homozygous for the same allele (e.g., c/c and c/c), but when a mutation occurs in a germ cell (e.g., c/c → c/t) it will be inherited by a heterozygous child, manifesting this mutation in the gamete of a chromosome.

2. Materials and methods

2.1 The data

Reads data about 10 different *trios* were provided by the professor in form of `.fq` files. The cases we had to analyze were: 1651, 1665, 1679, 1700, 1762, 1768, 1793, 1806, 1816 and 1841.
In addition, we were given a `universe.fasta` file, along with its index files, which is **our hg19 reference genome for chr16**, and an `exons16Padded_sorted.bed` file, which specifies the **target regions**.

2.2 Methods

Figure 1 below shows the workflow we followed for each case.

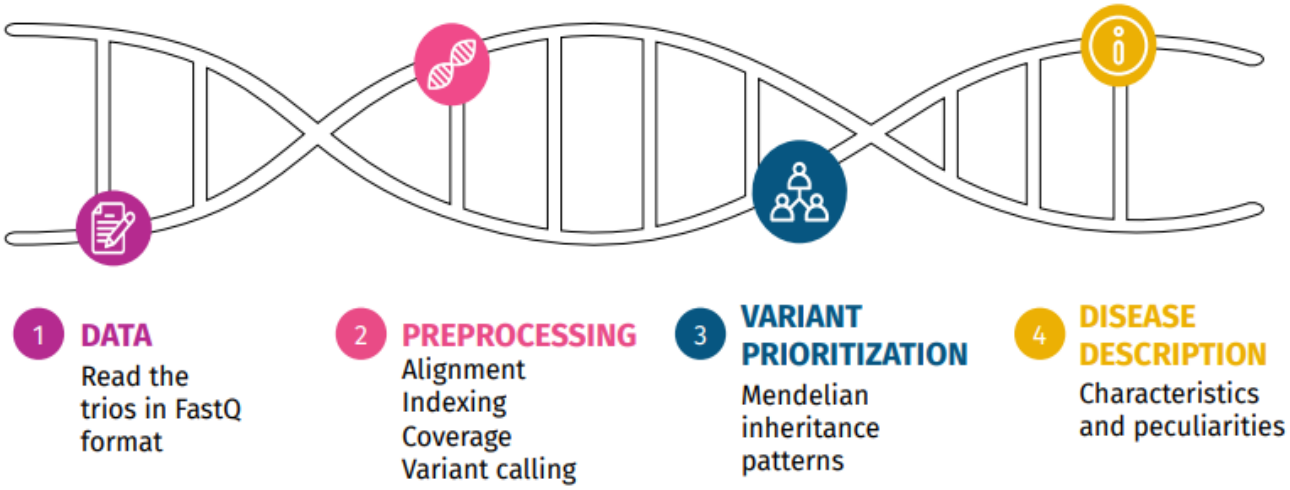

Figure 1: Workflow

2.2.1 Space organization

All our code for the analysis was **executed in the bash textual shell**, through a terminal emulator provided with the PuTTY client on Windows OS.
We created a **folder exam** in which we put all `uni*` files and the `exons16Padded_sorted.bed`. Then, **in this directory**, we organized **ten folders** in which we put all the ten different cases and called them, respectively, *case1793* exc.

2.2.2 Preprocessing and Variant calling

In the first place, we performed a few preprocessing steps:

1. We checked the **quality of the reads** running the command `fastqc *` in all the cases' folders.
2. In the ten folders we, then, wrote an sh file called `preprocess.sh`:

```
echo "Preprocessing fastq files..."

for filename in *.fq.gz
do
	base=$(basename $filename .fq.gz)

	echo "Aligning sample ${base}..."

	bowtie2 -U ${base}.fq.gz -x ../uni --rg-id "${base}" --rg "SM:${base}" |
	samtools view -Sb | samtools sort -o ${base}.bam

	echo "Indexing sample ${base}..."

	samtools index ${base}.bam

	echo "Running bamQC on sample ${base}"

	qualimap bamqc -bam ${base}.bam -gff ../exons16Padded_sorted.bed -outdir
${base}
done

echo "Done"
```

   We ran this piece of code with the command `sh preprocess.sh`. This code, in sequence:
   1. **aligned the reads** to the reference genome (**`universe.fasta`**) with **`bowtie`**;
   2. the obtained output (in `.sam` format) was **compressed** in `.bam` format and **sorted**;
   3. **indexed** the latter with **`samtools index`**;
   4. **checked the quality** of the results with **`qualimap bamqc`**.
3. We **put together all the quality control reports** in a single `html` with **`multiqc ./`**.
4. To perform the **variant calling**, we ran the following script:
   `freebayes -f ../universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 --targets ../exons16Padded_sorted.bed case1793_child.bam case1793_father.bam case1793_mother.bam > case1793.vcf`
5. Lastly, computed **coverage profiles in bedgraph** format with the following code:
   `bedtools genomecov -ibam` `case1793_father.bam -bg -trackline -trackopts 'name="father"' -max 100 > father1793Cov.bg`
   `bedtools genomecov -ibam` `case1793_mother.bam -bg -trackline -trackopts 'name="mother"' -max 100 > mother1793Cov.bg`
   `bedtools genomecov -ibam` `case1793_child.bam -bg -trackline -trackopts 'name="child"' -max 100 > child1793Cov.bg.`

In particular, the steps 3, 4 and 5 were performed in each case's folder.

2.2.3 Variant Prioritization

The `vcf` file obtained at the 4th step of the previous section contains the different genomic variants found in the individuals of the trio with freebayes, taking in order `mother`, then `father` and lastly `child`. The goal of this study is to focus on child, knowing that parents are healthy, so we need to consider different models of inheritance:

- If the disease is **autosomal recessive (AR)**, it manifests itself in the offspring if its genetic makeup is "*1/1*" in the `vcf` file, so homozygous for the variant (while both the parents are heterozygous, so "*0/1*", otherwise they would be affected as well). So, we need to select only the variants following this pattern and saving them in an output file with this command:
  `grep "0/1.*0/1.*1/1" case1793 > reccase1793.vcf`
- If the disease is **autosomal dominant (AD)** instead, we should consider that it can be caused by *de novo* mutations. We look for parents that are homozygous for the reference allele, whilst child has at least one different allele:
  `grep "0/0.*0/0.*0/1" case1806.vcf > domcase1806.vcf`

2.2.3.1 Ensembl Variant Effect Predictor

After the extraction of the variants, it's important to **evaluate their effect** and which of them is the main cause of emergence of the disease. So, we uploaded our `vcf` files on **Ensembl Variant Effect Predictor (VEP)**, that uses gene annotations to infer the effect of the genetic variants, and filtered them by different methods and settings:

- Transcripts: RefSeq
- Frequency data for co-located variants: 1000 genomes global; gnomAD (exomes) allele frequencies
- Additional annotations: phenotypes
- Pathogenicity predictions: prediction + score, SIFT + PolyPhen

Firstly, for each `vcf` file, we filtered the results by "Impact" (set as HIGH). Even if this is the most effective method, it actually doesn't take into account non-synonymous disease causing variants, for which the Impact is annotated as "moderate". So, in case of missing results, we performed a different filtration with:

- Impact: exclude LOW and MODIFIER (non protein coding)
- SIFT <= 0.2: (Sorts Intolerant From Tolerant), predicted the missense variant to be deleterious (if < 0.05) basing on the degree of protein sequence conservation
- PolyPhen2 > 0.6 (polymorphism phenotyping version 2): uses protein sequence and structure to predict the impact of a missense variant (damaging if > 0.5)

In particular, SIFT and PolyPhen use the terms "*damaging"* and "*tolerated"* to describe whether a variant is predicted to affect protein function or be functionally neutral, respectively. We emphasize that the term damaging should never be logically equated with causal for a disease phenotype, because a variant that damages a gene is not necessary damaging to an individual's health.

3. Results and analysis

3.1 Quality of the data

In order to assess the quality of the sequencing and aligning of our data, we used a MultiQC report generated for each case and evaluated the sequencing quality (*Phred score > 30*) and the alignment coverage (*Mean cov > 10X*), enough to perform the analysis (see results in Figure 2).

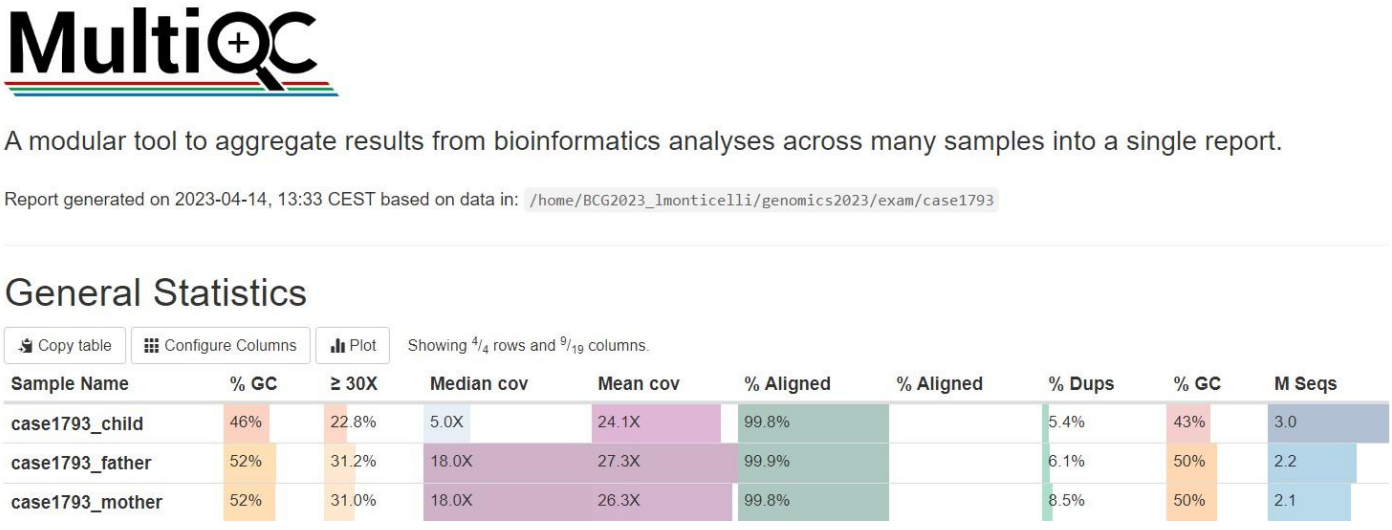

Figure 2: MultiQC report for case1793

3.2 Diagnoses

This table below collects the diagnoses found.

In particular, for *case 1679*, VEP actually showed "Rubinstein-Taybi Syndrome 1" that wasn't included in the list of pathologies. Due to the fact that the gene associated was CREBBP, we selected this type of disease.

For *case 1768* and *case 1806* instead, missense variants with moderate impact were 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 like *SIFT* classified it as "tolerated" and the allele frequency according to gnomAD is not very low (>10^{-4}). Thus, our final diagnoses for these cases were **healthy.**

Case	Location	Ref	Alt	Consequence	Gene	Disease
1651 (AR)	16:56913454-56913458	CCAT	CAT	Frameshift variant, Splice region variant	SLC12A3	**Familial hypokalemia-hypomagnesemia**
1665 (AR)	16:89811472-89811472	C	T	Stop gained	FANCA	**Fanconi anemia complementation group A**
1679 (AD)	16:3900810-3900810	G	A	Stop gained	CREBBP	**Rubinstein-Taybi syndrome due to CREBBP mutations ***
1700 (AR)	16:89806419-89806422	AGA	A	Frameshift variant	FANCA	**Fanconi anemia complementation group A**
1762 (AD)	16:2143899-2143901	GC	C	Frameshift variant	PKD1	**Autosomal dominant polycystic kidney disease**
1768 (AD)	-	-	-	-	-	**HEALTHY****
1793 (AR)	16:56553661-56553666	CCCGT	CCGT	Frameshift variant	BBS2	**Bardet-Biedl Syndrome**
1806 (AD)	-	-	-	-	-	**HEALTHY****
1816 (AD)	16:3801726-3801726	C	A	Splice donor variant	CREBBP	**Rubinstein-Taybi syndrome due to CREBBP mutations**
1841 (AD)	16:2142954-2142954	C	A	Splice donor variant	PKD1	**Autosomal dominant polycystic kidney disease**

3.3 Visualization on UCSC Genome Browser

The disease-causing variants of each case, along with the coverage tracks, were finally visualized on the **UCSC Genome Browser**. As an example, Figure 3 below shows the disease causing variant for *case 1793*.

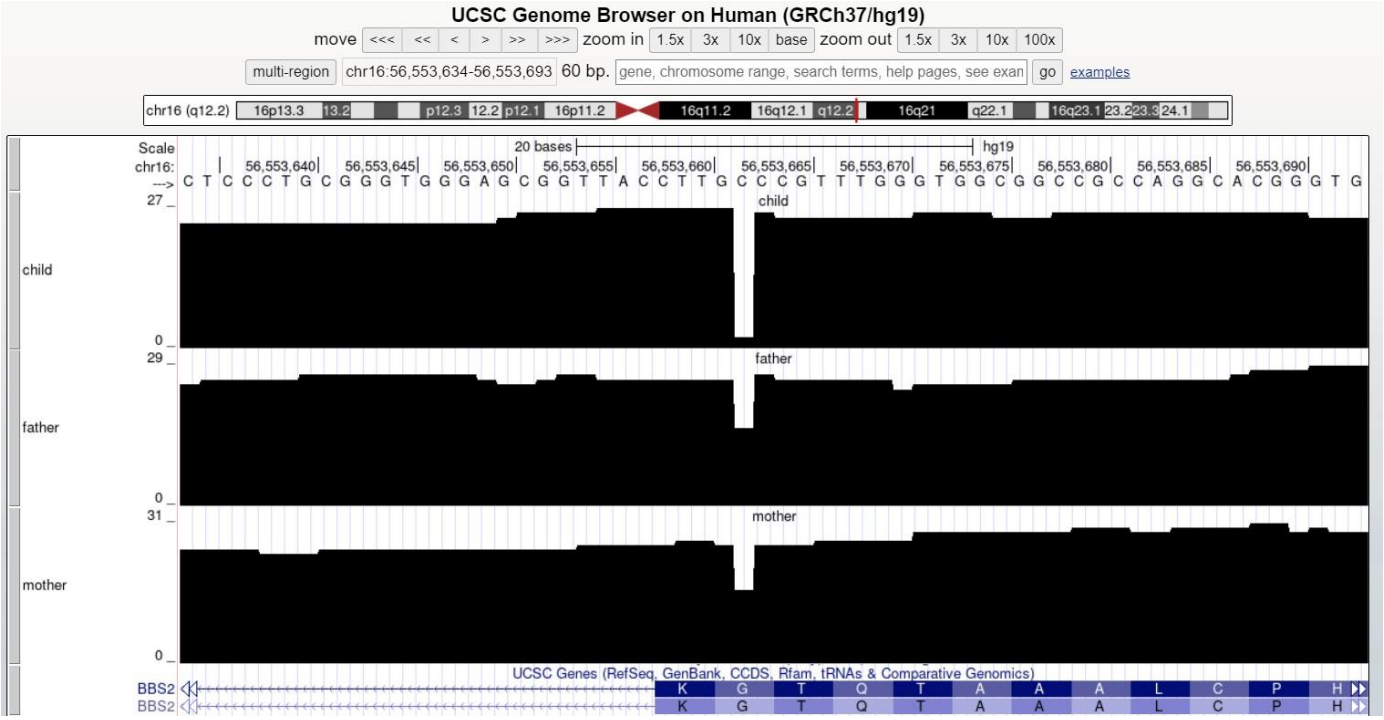

Figure 3: case1793 disease-causing frameshift mutation on UCSC Genome Browser

4. Discussion

As the Table in section 3.2 shows, only **two out of the total ten cases were diagnosed as "healthy"**, while the other eight were **diagnosed with a mutation associated to a rare mendelian disease**.

For *case 1651*, a frameshift variant/splice region variant was found on the SLC12A3 gene. The primary phenotype related to this mutation is a **familial hypokalemia-hypomagnesemia**. It is characterized by hypokalemic metabolic alkalosis with hypomagnesemia and hypocalciuria and is the most common renal tubular disorder among Caucasians. Clinical features include transient periods of muscle weakness and tetany, abdominal pains, and chondrocalcinosis.

For *cases 1679* and *1816*, respectively, a stop gained variant and a splice donor one were found on the CREBBP gene. The primary phenotype related to these mutations is **Rubinstein-Taybi syndrome**. It causes congenital anomalies, such as microcephaly, characteristic facial dimorphisms, wide thumbs and toes, and intellectual disability with behavioral problems.

For *cases 1665* and *1700*, respectively, a stop gained variant and a frameshift one were found on the FANCA gene. The primary phenotype related to these mutations is **Fanconi Anemia**. It causes an increased risk of cancer, bone marrow failure, physical defects and organ malformations.

For *cases 1762* and *1841*, respectively, a frameshift variant and a stop gained one were found on the PKD1 gene. The primary phenotype related to these mutations is **Autosomal dominant polycystic kidney disease**. It is characterized by the development of renal cysts and various extra renal manifestations, such as intracranial aneurysms and dolichoectasias, abdominal wall hernias, mitral valve prolapse, and aortic root dilatation and aneurysms.

For *case 1793*, a frameshift variant was found on the BBS2 gene. The primary phenotype related to this mutation is **Bardet-Biedl Syndrome**. It is mainly characterized by obesity, retinitis pigmentosa, polydactyly, hypogonadism and, in some cases, kidney failure.

Finally, instead, for *case 1768* and *case 1806*, no variants with a high impact were found to be associated with any rare disease. Therefore, we looked for variants with a moderate impact: missense variants with moderate impact were 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 like *SIFT* classified it as "tolerated" and the allele frequency according to gnomAD is not very low ($>10^{-4}$). Thus, our final diagnoses for these cases were **healthy.**