A. Y. 2023-2024

Project Report

MOLECULAR DIAGNOSIS OF RARE GENETIC DISORDERS

Genomics and Transcriptomics

M.Sc. Bioinformatics for Computational Genomics

Department of Biosciences, University of Milan

COMINELLI MARCO, SASSO ELENA

INDEX

1.Introduction

A **genetic disorder** is a medical condition that arises due to abnormalities, mutations, or variations within an individual's DNA or genetic material, leading to irregularities in biological processes and often resulting in observable symptoms or health complications. These genetic variations can be inherited from parents or arise spontaneously (*de novo* mutations).

Specifically **monogenic genetic diseases** are caused by mutations in a **single gene** and they can be grouped into two main categories:

- **Dominant** diseases: only one copy of the mutated gene is necessary to cause the disease.
- **Recessive** diseases: both copies of the mutated gene are necessary to cause the disease.

The current project focuses on autosomal monogenic rare genetic disorders, which are called 'autosomal' because the mutated gene is on an autosome (specifically the chromosome 16) and 'rare' due to their low frequency within the population (less than 1 in 1,000 people).
In particular, the aim of the project is to diagnose these diseases in 10 patients using both their exome sequencing data and that of their parents (**Parent-child trios exome sequencing**).

2.Materials, Methods and Analysis

To perform the diagnoses it was used the pipeline implemented in the "VarCall_pipeline.sh" script (the full script was implemented by us and can be found at this link or at /home/BCG_2024_mcominelli/VarCall_pipeline.sh on the server 159.149.160.7).
The script has been implemented to carry out all the required analyses by launching a single command: *nohup ./VarCall_pipeline.sh -ad case593 case671 case717 case731 case735 -ar case611 case659 case732 case739 case742 &.*
All the commands used are also reported in the "Commands" section at the end of the report.
The reference genome considered in this project is the hg19 genome assembly.

2.1 QC, Alignment, Indexing, Coverage

First of all, the quality of the reads (provided in the *FASTQ* format) of all three trio's members, was assessed by using the **fastqc** tool [1].
Following this, the reads were aligned to the chromosome 16 sequence using the **bowtie2** tool (which is based on the Burrows-Wheeler Transform method), providing a *SAM* file as output [2].
Subsequently, using the **samtools** tool [2], this file was converted into its binary form, the BAM file, which was then sorted [2] and indexed using the same tool [3].
At this point, the alignment quality was inspected using the **qualimap** tool [4], considering only the reads mapping on the targeted regions (the exons) of the chromosome 16 (this is an exome sequencing experiment) by specifying ***-gff ${bed_file}***.
Subsequently, all quality checks were combined into a single output using the **multiqc** tool [5].
As the final step of this initial analysis phase, the coverage for each of the three family members was computed using the **bedtools genomecov** tool [6], considering a maximum depth of 100 reads, specified by the parameter ***-max 100***.

2.2 Variant Calling

Subsequently, variant calling was conducted using the **freebayes** tool [7] (which adopts a Bayesian statistical framework to do variant calling), combining the sorted and indexed BAM files of the trio's three members to generate a unified VCF file containing all variants.
The aim of this project is to diagnose a possible rare genetic disorder in the child, so considering also the reads of the exome sequencing of the parents helps to be more robust in calling the variants of the child. The parameters used are the following:
• ***-m 20***: the aligments with a mapping quality of less than 20 were excluded
• ***-C 5***: alternate alleles were called only if at least five reads, for each individual, were observed as mutated at that specific position
• ***-Q 10***: the mismatch was called only with a sequencing quality of at least ten for that specific position
• ***--min coverage 10***: a coverage of at least ten was required for that specific position

2.2.1 Variant Prioritization

After the variant calling procedure, the freebayes tool generates a VCF file containing all the discovered variants. However, the majority of these variants may not be pertinent for the analysis, requiring a filtering step. To accomplish this, we developed a Python script named '**greppy**' (the complete script can be found at this link or at /home/BCG_2024_mcominelli/greppy on the server 159.149.160.7).
For both autosomal recessive and autosomal dominant cases, parents were considered healthy (thus 0/n in recessive cases and 0/0 in dominant cases, with 'n' being any natural number different from 0). **greppy** applies the following filtering strategies based on the type of the case:

- Autosomal recessive:
  - *Low*: This mode retains variants where the parents are 0/1 and the child is 1/1.
  - *High*: This mode retains variants where the parents are 0/n and the child is n/n (child can of course be heterozygous, e.g., mother "0/2", father "0/3" and child "2/3"), plus all the variants retained by the *low* mode.
- Autosomal dominant:
  - *Low*: This mode retains variants where both parents are 0/0 and the child is 0/n, so retaining only *de novo* mutations.
  - *High*: This mode retains variants where one of the parents has the variant (the variant can also be different from '1') and the child is heterozygous for such variant (he inherited it from the affected parent), plus all the variants retained by the *low* mode.

For this project, **greppy** was always set to *low* mode.

Following this stage, a refined VCF file, named *'filtered.vcf'*, was generated. Finally, the **bedtools intersect** tool [8] was employed to intersect this file exclusively with the target regions, resulting in the creation of *'final.vcf',* on which the subsequent analyses were conducted.

2.3 VEP

The final stage of the analyses, which led to define the diagnoses, was conducted using the Variant Effect Predictor (VEP) software. The final VCF files (one for each trio) obtained from the script were uploaded to the site, which provided detailed analyses of all the variants present in the VCF.
In particular, VEP offers crucial insights into the type of mutation caused by the variant (e.g., frameshift, stop-gained, missense, etc.), its impact (high, moderate, modifier, low), its potential pathogenicity, and the possible associated phenotypes.
To perform our analyses, we considered the hg19 genome assembly and we configured VEP to exclusively utilize the RefSeq transcript database, ensuring that the SNPs were mapped only onto RefSeq transcripts.
Then, to conduct the diagnoses, the first filter applied was '*IMPACT is HIGH*', which enabled us to identify 8 diagnoses out of 10 without further filters. In cases where no results were obtained using this filter, we analyzed the variants considering other impacts as well, applying '*Associated phenotype is defined*' as the only filter.

3.Results and Conclusions

3.1 Multiqc

The link to the complete MultiQC reports for all 10 cases is reported below.
For all of the cases, the two most problematic statistics were **per base sequence content** and **GC sequence content**. The first indicates the proportion of each base position for which each of the four normal DNA bases has been called, the second indicates the average GC content of the reads.

LINK_MULTIQC

3.2 Diagnoses

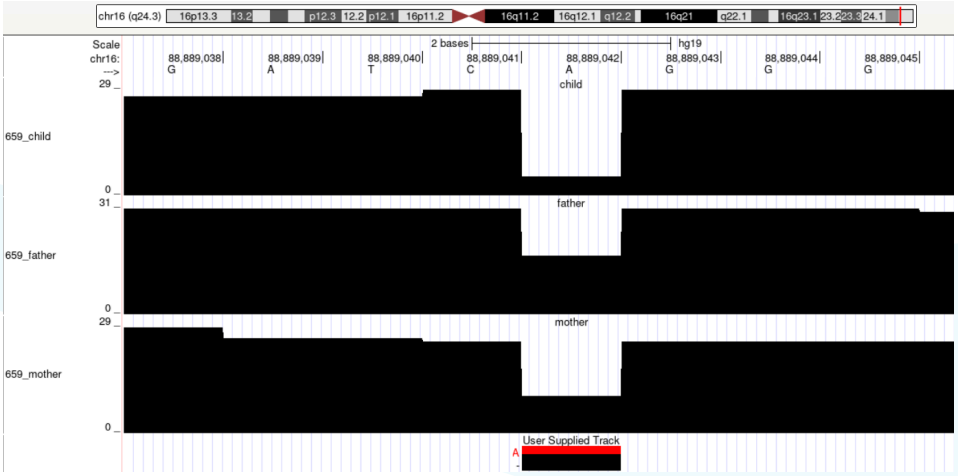
The diagnoses found for each of the 10 cases are described in the following table, with a link in the last column to the VEP webpage displaying the analyzed variant.

Case	AD/AR	Position	Consequence	Impact	AF	GnomADe	Clinical significance	Diagnosis	LINK
593	AD	16:2140777-2140777	stop_gained	HIGH	-	-	pathogenic	**Autosomal dominant polycystic kidney disease**	LINK_VEP_593
611	AR	16:89815099-89815099	stop_gained	HIGH	-	-	pathogenic	**Fanconi anemia|Fanconi anemia complementation group A**	LINK_VEP_611
659	AR	16:88889041-88889043	frameshift_variant	HIGH	-	8e-06	-	**Mucopolysaccharidosis Type IVA**	LINK_VEP_659
671	AD	NA	NA	NA	NA	NA	NA	**Not affected by a rare genetic disorder**	See the discussion below
717	AD	16:89350386-89350394	frameshift_variant	HIGH	-	-	-	**KBG syndrome**	LINK_VEP_717
731	AD	16:51174877-51174877	stop_gained	HIGH	-	-	pathogenic	**Townes-Brocks syndrome 1**	LINK_VEP_731
732	AR	16:89858416-89858416	stop_gained	HIGH	-	-	pathogenic	**Fanconi anemia|Fanconi anemia complementation group A**	LINK_VEP_732
735	AD	NA	NA	NA	NA	NA	NA	**Not affected by a rare genetic disorder**	See the discussion below
739	AR	16:89858357-89858361	frameshift_variant	HIGH	-	-	pathogenic	**Fanconi anemia|Fanconi anemia complementation group A**	LINK_VEP_739
742	AR	16:11000538-11000540	frameshift_variant	HIGH	-	-	pathogenic	**MHC class II deficiency**	LINK_VEP_742

In all cases, except for 671 and 735, a mutation responsible for causing a rare genetic disease has been identified. With the variant classified as **high-impact**, with an **allele frequency below 0.0001**, almost always with **pathogenic clinical significance** and categorized as **frameshift** or **stop-gained**, it is highly probable that this variant leads to the onset of a rare genetic disease.

However, for case 671 (and the same applies to 735), the two most suspicious variants are classified one as moderate-impact and the other as modifier-impact.
Regarding the moderate-impact variant (details can be found at these links: LINK_VEP_671_moderate, LINK_VEP_735_moderate), almost all of the gnomAD allele frequencies are greater than 0.0001, indicating that the variant is not rare. Another concerning aspect was the low quality associated with the variant in the VCF file (2.38811e-06). Then, given the fact that this variant was classified as **missense-variant**, also the prediction scores computed by SIFT, CADD and PolyPhen were evaluated and two of them (CADD and PolyPhen) considered the variant to be pathogenic, but in conclusion, considering mainly the allele frequency, the variant was not considered to be the cause of a rare genetic disease.
As for the modifier variant (details can be found at these links: LINK_VEP_671_modifier, LINK_VEP_735_modifier), all the allele frequencies were not defined; however, the quality of the variant was very low (0.020308), as expected, because the variant is located in an intron. So, also in this case as well, it was not possible to trust the variant.
In the end, it was not possible to consider the patients 671 and 735 as affected by a rare genetic disorder.

3.3 UCSC Genome Browser



As can be seen from the screenshots provided in the link below, all cases, except cases 659, 739 and 742, exhibit a good coverage in the regions where the variant of interest is located. In the three above mentioned cases (one of them, 659, is shown in the picture above), instead, the coverage in the child is notably low: this is perfectly normal as the variant is represented by a deletion causing a frameshift mutation and also recalling the fact that in those cases the children were diagnosed with an autosomal recessive disease, so they are homozygous for such deletion.

LINK_UCSC_COVERAGE

4.Commands

[1] fastqc ${file_dir}/${ad_trio}* -o ${my_wd}/${ad_trio}/QC

[2] bowtie2 -U ${file_dir}/${ad_trio}_father.fq.gz -x ${file_dir}/uni --rg-id 'SF' --rg "SM:father" | samtools view -Sb | samtools sort -o ${ad_trio}_father.bam
bowtie2 -U ${file_dir}/${ad_trio}_child.fq.gz -x ${file_dir}/uni --rg-id 'SC' --rg "SM:child" | samtools view -Sb | samtools sort -o ${ad_trio}_child.bam
bowtie2 -U ${file_dir}/${ad_trio}_mother.fq.gz -x ${file_dir}/uni --rg-id 'SM' --rg "SM:mother" | samtools view -Sb | samtools sort -o ${ad_trio}_mother.bam

[3] samtools index ${ar_trio}_father.bam
samtools index ${ar_trio}_child.bam
samtools index ${ar_trio}_mother.bam

[4] qualimap bamqc -bam ${ar_trio}_father.bam -gff ${bed_file} -outdir ./../QC/${ar_trio}_father
qualimap bamqc -bam ${ar_trio}_child.bam -gff ${bed_file} -outdir ./../QC/${ar_trio}_child
qualimap bamqc -bam ${ar_trio}_mother.bam -gff ${bed_file} -outdir ./../QC/${ar_trio}_mother

[5] multiqc ./../QC/ --outdir ./../QC/multiqc

[6] bedtools genomecov -ibam ../${ar_trio}_father.bam -bg -trackline -trackopts 'name="father"' -max 100 > ${ar_trio}_fatherCov.bg
bedtools genomecov -ibam ../${ar_trio}_child.bam -bg -trackline -trackopts 'name="child"' -max 100 > ${ar_trio}_childCov.bg
bedtools genomecov -ibam ../${ar_trio}_mother.bam -bg -trackline -trackopts 'name="mother"' -max 100 > ${ar_trio}_motherCov.bg

[7] freebayes -f ${file_dir}/universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10../bam/${ar_trio}_mother.bam../bam/${ar_trio}_child.bam../bam/${ar_trio}_father.bam> {ar_trio}.vcf

[8] bedtools intersect -a ${ar_trio}_sorted.vcf -b ${bed_file} -u > ${ar_trio}_final.vcf