From Whole Exome Sequencing to Diagnosis: A Bioinformatic Approach for Autosomal Single-Gene Disorders

Luca Lepore and Elena Lippolis

1 Introduction

Rare genetic diseases, since they affect only a small portion of the global population, represent a significant struggle for physicians tasked to differentiate among clinically similar conditions. These challenges often lead to delayed diagnoses, which can potentially worsen the disease[1].

This report focuses on autosomal single-gene disorders, which are primarily inherited through familial genetic lines. Additionally, some cases may arise from de novo mutations occurring in germ cells or during early embryonal development. In particular, autosomal dominant disorders occur in the heterozygous state, usually affecting one of the parents. Autosomal recessive disorders require both alleles at a given gene locus to be mutated instead, with neither parent affected[2].

With the introduction of Whole Exome Sequencing (WES), the analysis of Single Nucleotide Variants (SNVs) within coding exons has remarkably enhanced diagnostic accuracy for rare genetic disorders[1]. By focusing on these genetic markers, we aim to help reduce diagnostic delays, improving patient care management.

All the code and data generated for this work are available on Github.

2 Materials and Methods

2.1 List of cases

The analysis was conducted on a cohort of 10 patients, each suspected of being affected by an autosomal disease, along with their parents. Each family trio was identified by a case number and an inheritance model, as detailed in Table 1. Moreover, patients suspected of a dominant disorder were specifically evaluated for de novo mutations.

Case	584	600	622	625	630	644	696	709	717	743
Inheritance	AD	AR	AR	AD	AD	AD	AD	AR	AD	AD

Table 1: Case studies (AD: Autosomal Dominant, AR: Autosomal Recessive).

2.2 Unix-based Pipeline

To manage and analyse the WES-derived genomic data from our patients, we developed a Unix-based pipeline. This framework was designed to cover the entire analytical process, from sequence alignment and preliminary quality assessment to variant calling. These steps prepare the data for downstream analyses, performed employing the Variant Effect Predictor (VEP) for variant prioritization and the UCSC Genome Browser to validate the identified variants. The pipeline and the description of the procedural steps are provided below.

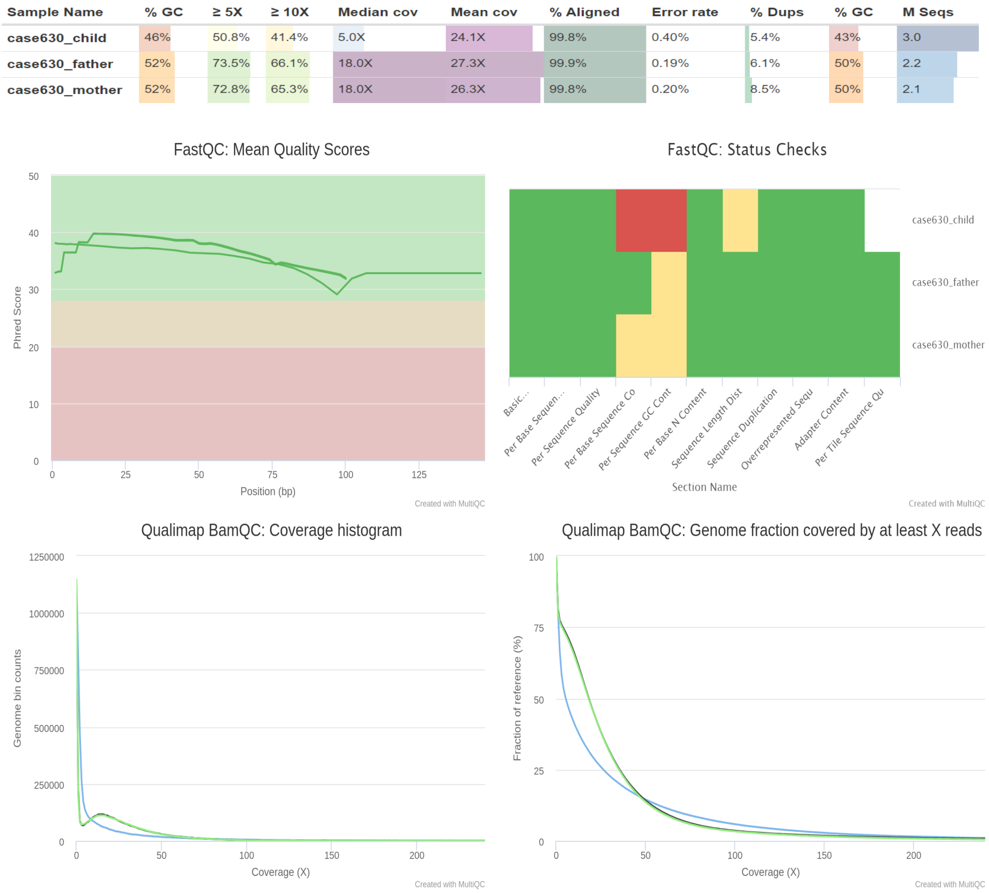
```
    cases=("case622 AR" "case717 AD" "case644 AD" "case625 AD" "case743 AD" "case584 AD" "case709 AR"
    "case600 AR" "case696 AD" "case630 AD")

#Loop for every case
for case_info in "${cases[@]}"; do
    case_name="${case_info%% *}"  # Extracts the substring before the first space (the case name)
    case_type="${case_info#* }"   # Extracts the substring after the first space (the case type: AR or AD)
    individuals=("mother" "father" "child")
    directory="/home/BCG_2024_elippolis/final_project/${case_name}"

    #if directory already exists it doesn’t create a new one
    if [ -d "$directory" ]; then
        echo "Directory and file already exist. Moving on..."
    #else create the directory and file with results
    else
        mkdir -p "$directory"
    fi

    #Sequence alignment and creation of the BAM file for each individual
    bowtie2 -U "/home/BCG2024_genomics_exam/${case_name}_mother.fq.gz" -p 8 -x /home/BCG2024_genomics_exam/uni --rg-id ’SM’
        --rg "SM:mother" | samtools view -Sb | samtools sort -o "${directory}/${case_name}_mother.bam"
    bowtie2 -U "/home/BCG2024_genomics_exam/${case_name}_father.fq.gz" -p 8 -x /home/BCG2024_genomics_exam/uni --rg-id ’SF’
        --rg "SM:father" | samtools view -Sb | samtools sort -o "${directory}/${case_name}_father.bam"
    bowtie2 -U "/home/BCG2024_genomics_exam/${case_name}_child.fq.gz" -p 8 -x /home/BCG2024_genomics_exam/uni --rg-id ’SC’
        --rg "SM:child" | samtools view -Sb | samtools sort -o "${directory}/${case_name}_child.bam"

    # Index BAM files
    for individual in "${individuals[@]}"; do
      samtools index "${directory}/${case_name}_${individual}.bam"
    done

    #Run fastQC
    mkdir "${directory}/fastqc_outputs"
    for individual in "${individuals[@]}"; do
      fastqc "/home/BCG2024_genomics_exam/${case_name}_${individual}.fq.gz" -o "${directory}/fastqc_outputs"
    done

    mkdir "${directory}/qualimap_outputs"
    # Run QC using qualimap and rename the stats file
    for individual in "${individuals[@]}"; do
          qualimap bamqc -bam "${directory}/${case_name}_${individual}.bam" -gff /home/BCG2024_genomics_exam/exons16Padded_sorted.bed
            -outdir "${directory}/qualimap_outputs/${case_name}_${individual}"
    done

    # Performing multiqc
    multiqc "${directory}" -o "${directory}/multiqc_reports" -n "multiqc_report_${case_name}.html"

    # Run FreeBayes to generate VCF
    nohup freebayes -f /home/BCG2024_genomics_exam/universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 "${directory}/${case_name}_child.bam"
    "${directory}/${case_name}_father.bam" "${directory}/${case_name}_mother.bam" > "${directory}/${case_name}.vcf" &
    wait

    # Order the patients in the VCF file to be sure they match the order of the pattern we put for the variants selection
    bcftools query -l "${directory}/${case_name}.vcf" | sort > "${directory}/${case_name}.samples.txt"
    bcftools view -S "${directory}/${case_name}.samples.txt" "${directory}/${case_name}.vcf" > "${directory}/${case_name}.sorted.vcf"

    # Select variants based on case type
    grep "^#" "${directory}/${case_name}.sorted.vcf" > "${directory}/candilist${case_name}.vcf"
    if [ "$case_type" == "AR" ]; then
        # Select variants with the recessive pattern
        grep "1/1.*0/1.*0/1" "${directory}/${case_name}.sorted.vcf" >> "${directory}/candilist${case_name}.vcf"
    elif [ "$case_type" == "AD" ]; then
        # Select variants with the dominant pattern
        grep "0/1.*0/0.*0/0" "${directory}/${case_name}.sorted.vcf" >> "${directory}/candilist${case_name}.vcf"
    else
        # In case the type (recessive or dominant) is not specified
        echo "Invalid case type: $case_type"
    fi

    # Generation of the VCF file keeping just the variants included in the target regions (-u is used to report each feature only once
    in case of overlapping exons)
    grep "^#" "${directory}/candilist${case_name}.vcf" > "${directory}/${case_name}candilistTG.vcf"
    bedtools intersect -a "${directory}/candilist${case_name}.vcf" -b /home/BCG2024_genomics_exam/exons16Padded_sorted.bed -u
        >> "${directory}/${case_name}candilistTG.vcf"

    #Generate coverage tracks with bedtools genomecov
    for individual in "${individuals[@]}"; do
       bedtools genomecov -ibam "${directory}/${case_name}_${individual}.bam" -bg -trackline -trackopts ’name="${individual}"’
        -max 100 > "${directory}/${case_name}${individual}Cov.bg"
    done
    echo "${case_name} finished."
done
```

This code performs genomic analysis on specific cases, each identified by a name and an inheritance type (AD or AR). The loop iterates through the list of cases, extracting the `case_name` and `case_type` for each and storing them as variables. For each case, a directory path is defined based on its case name.
Through the `bowtie2` command, the code aligns the sequencing data from FASTQ files to the reference genome hg19 for each family member, generating the BAM files. Once created, they are indexed through `samtools index`. The code also runs FastQC on each FASTQ file to assess raw sequencing quality and `qualimap bamqc` on each BAM file to evaluate alignment quality. Then, MultiQC aggregates the quality reports from FastQC and Qualimap into a single HTML file.
`FreeBayes` is employed to identify variants in the BAM files for each individual (mother, father, child) according to the inheritance model, and generate a VCF file as a consequence. To ensure accuracy and reliability in variant calling, the criteria set includes a minimum mapping quality of 20 (`-m 20`), at least 5 alternative reads (`-C 5`), a minimum base quality of 10 for mismatches (`-Q 10`), and a minimum total coverage of 10 reads in a site (`--min-coverage 10`). This process runs in the background and waits for completion using `wait`. The VCF file is then sorted to ensure consistent individual ordering. Depending on the case type (recessive or dominant) variants are selected and stored in a VCF file. These files are further filtered to include only the variants within target regions by utilizing `bedtools intersect`. Finally, `bedtools genomecov` generates individual coverage tracks, concluding the analysis for one case before moving on to the next.

3 Results

3.1 Quality Assessment

Quality assessment is a crucial step in establishing the integrity of sequencing data and the robustness of downstream analyses. To this end, MultiQC was employed for each family trio data to generate a comprehensive report that merges results from FastQC and Qualimap. The analysis of these reports confirmed the good quality of our data. Below, Figure 1 shows an overview of Case 630. The general statistics and the plots indicate that the data is mostly reliable for downstream analysis, although the low quality in the per base sequence content and per sequence GC content in the child's sample. The complete MultiQC reports generated are accessible at this link.



Sample Name	% GC	≥ 5X	≥ 10X	Median cov	Mean cov	% Aligned	Error rate	% Dups	% GC	M Seqs
case630_child	46%	50.8%	41.4%	5.0X	24.1X	99.8%	0.40%	5.4%	43%	3.0
case630_father	52%	73.5%	66.1%	18.0X	27.3X	99.9%	0.19%	6.1%	50%	2.2
case630_mother	52%	72.8%	65.3%	18.0X	26.3X	99.8%	0.20%	8.5%	50%	2.1

Figure 1: Summary of the quality statistics for Case 630.

3.2 Variant Prioritization

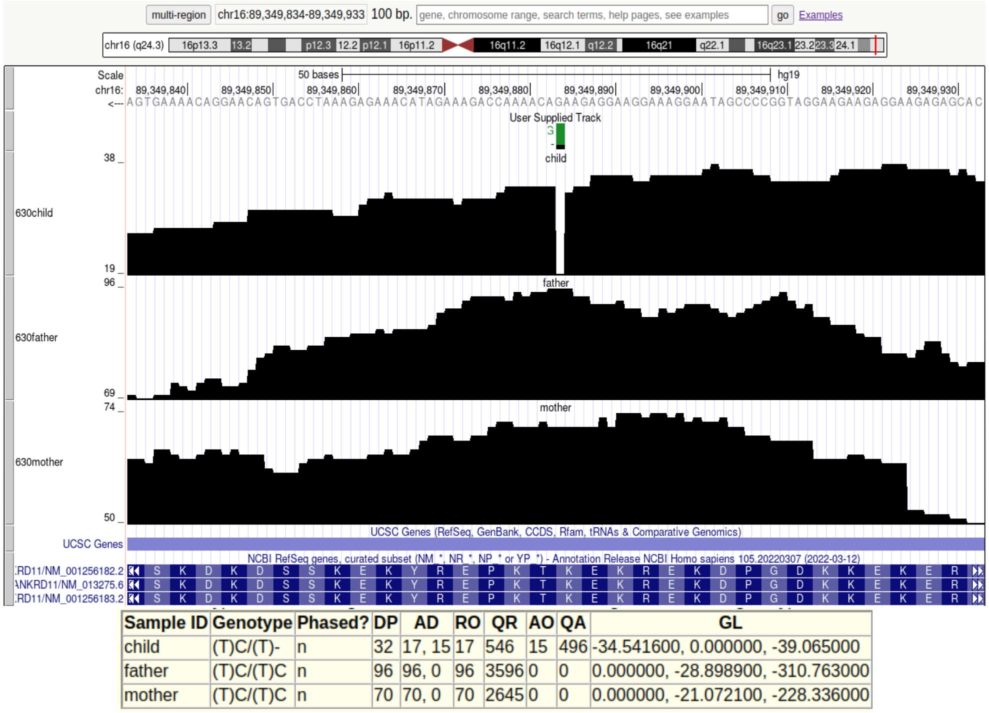
Following quality assessment, we employed the Variant Effect Predictor (VEP) tool by Ensembl to perform Variant Prioritization. For each trio analysed, the resulting VCF file was submitted to the platform to identify the most likely pathogenic variant correlating with the patient clinical presentation. Regarding the submission, the following parameters were selected: the GRCh37.p13 (hg19) assembly, the RefSeq database for transcripts annotation, the 1000 Genomes Global and gnomAD databases for variants frequency data, the phenotypes annotation, and SIFT, PolyPhen, and CADD PHRED score as pathogenicity prediction tools (with CADD consulted in case of discrepancies between the first two). The preliminary findings are reported in Table 2. The VEP outputs for all the cases are accessible at this link. For variants with similar annotations indicating the same condition, the ClinVar database was prioritized if available. This choice was based on its wide use for diagnostics applications in the clinical setting[3]. For further details regarding the variants identified and the diagnoses, refer to the Discussion section in this report.

Case	Mutated Gene	Variant Location	Consequence	REF/ALT	Impact	Potential phenotype association
584 AD	PKD1	16:2140342-2140348	Frameshift del.	(C)CATCT/(C)-	High	Polycystic kidney disease
600 AR	BBS2	16:56548534-56548538	Frameshift del.	(T)G/(T)-	High	Bardet-Biedl syndrome
622 AR	FANCA	16:89874750-89874750	Stop-gained	C/T	High	Fanconi anemia
625 AD	CREBBP	16:3820629-3820629	Missense	G/T	Moderate	Rubinstein-Taybi syndrome
630 AD	ANKRD11	16:89349883-89349885	Frameshift del.	(T)C/(T)-	High	KBG syndrome
644 AD	ANKRD11	16:89350771-89350783	Frameshift del.	(C)TT/(C)-	High	KBG Syndrome
696 AD	PKD1	16:2140149-2140151	Frameshift del.	(C)G/(C)-	High	Polycystic kidney disease adult type
709 AR	GALNS	16:88891224-88891226	Frameshift del.	(T)G/(T)-	High	Mucopolysaccharidosis MPS-IV-A
717 AD	ANKRD11	16:89350386-89350394	Frameshift del.	(T)C/(T)-	High	KBG syndrome
743 AD	ANKRD11	16:89345899-89345903	In-frame del.	(C)CTG/(C)-	Moderate	KBG syndrome

Table 2: Summary of the preliminary Variant Prioritization outcomes.

3.3 Variant validation and Coverage analysis

For each family trio, the VCF file and coverage tracks were uploaded to the UCSC Genome Browser to confirm the presence of the identified variant and to visualize the coverage depth, providing additional confidence in our results. Below, Figure 2 shows Case 630 as an example. The figure displays a pronounced gap in the coverage in the variant site, supporting the frameshift deletion mutation. The absence of this pattern in both parents confirms the de novo occurrence. Furthermore, the child's read count at this locus is consistent with heterozygosity, and the high read depth supports an optimal confidence level in variant identification. The other cases are accessible at this link


multi-region
chr16:89,349,834-89,349,933
100 bp.
gene, chromosome range, search terms, help pages, see examples
go
Examples
chr16 (q24.3)
16p13.3
13.2
p12.3
12.2
p12.1
16p11.2
16q11.2
16q12.1
q12.2
16q21
q22.1
16q23.1
23.2
23.3
24.1
Scale
50 bases
hg19
chr16:
89,349,840
89,349,850
89,349,860
89,349,870
89,349,880
89,349,890
89,349,900
89,349,910
89,349,920
89,349,930
<---
AGTGAAAACAGGAACAGTGACCTAAAGAGAAACATAGAAAGACCAAAACAGAAGAGGAAGGAAAGGAATAGCCCCGGTAGGAAGAAGAGGAAGAGAGCAC
User Supplied Track
G
-
child
38
630child
19
96
father
630father
69
74
mother
630mother
50
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
UCSC Genes
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens 105.20220307 (2022-03-12)
RD11/NM_001256182.2
ANKRD11/NM_013275.6
RD11/NM_001256183.2
S K D K D S S K E K Y R E P K T K E K R E K D P G D K K E K E R

Sample ID	Genotype	Phased?	DP	AD	RO	QR	AO	QA	GL
child	(T)C/(T)-	n	32	17, 15	17	546	15	496	-34.541600, 0.000000, -39.065000
father	(T)C/(T)C	n	96	96, 0	96	3596	0	0	0.000000, -28.898900, -310.763000
mother	(T)C/(T)C	n	70	70, 0	70	2645	0	0	0.000000, -21.072100, -228.336000

Figure 2: Coverage Profile at Locus 16:89349883-89349885 for Case 630.

4 Discussion

Identifying the genetic cause of a Mendelian disease requires prioritizing among the multitude of variants typically present in an exome. However, this is not a simple task because each variant can damage a gene differently, and not all variants predicted to be damaging impact patient health[4]. Thus, the primary rationale adopted was minimizing the risk of false negatives at the expense of a higher number of false positives (Table 2) and then proceeding to a more in-depth analysis. To identify the pathogenic variants, a series of filter options was applied. Our initial focus was on high-impact variants that could completely disrupt protein functionality. This analysis revealed frameshift variants in 7 cases and a stop-gained variant in 1 case (Table 2). Given the nature of these mutations, SIFT and PolyPhen scores, which are specific for missense variants, were not applicable. A CADD PHRED score was available just for the stop-gained variant instead, with a score of 38. Nonetheless, because of the high impact of these mutations, the respective SNVs were considered responsible for the clinical conditions observed in the patients.
In contrast, a more in-depth analysis was necessary for Case 625 and Case 743. For filtering, the following criteria were sequentially implemented: the exclusion of variants with low impact, the requirement for an associated phenotype or disease, SIFT score <= 0.2, PolyPhen score > 0.6, and CADD PHRED score > 20. Regarding the pathogenicity prediction scores, the guidelines propose stricter thresholds. However, more lenient thresholds were chosen to minimize the risk of false negatives in this phase. Concerning Case 625, a missense variant with moderate impact was identified in the CREBBP gene. The SIFT and PolyPhen scores were 0.1 and 0.966, respectively, and the CADD PHRED score was 24.2, indicating a potential pathogenic variant. However, while the global Allele Frequency was undefined, suggesting a possible causal relationship with the rare disease, further examination revealed that the subpopulation allele frequencies were above the 10^{-4} threshold (gnomADe AFR AF > 0.003). In addition, the analysis on the Genome Browser showed that this variant results in the substitution of a Threonine (T) with an Asparagine (N), amino acids that share similar chemical and physical properties. These insights indicate that this variant could not probably be the main cause of the patient's condition and the patient was classified as not affected by a rare disease as a consequence. Focusing on Case 743 instead, the examination was more complex. An in-frame deletion variant with moderate impact was identified in the ANKRD11 gene. For this type of mutation, neither SIFT nor PolyPhen scores were applicable, and no CADD PRHED score was available. Moreover, global and subpopulation allele frequencies were undefined, suggesting a possible link with a rare disease. In-frame deletions preserve the reading frame, meaning the pathogenic impact depends only on the significance of the deleted amino acid to the protein's structure and function[2]. Considering the uncertain evidence connecting this variant to the observed phenotype, a literature investigation was conducted, resulting in minimal relevant findings[5]. Therefore, even this patient was categorized as not affected by a rare disease but with a recommendation for further detailed analyses. Below, Table 3 summarizes the final results of this study.

Case	Diagnosis
584 AD	Polycystic kidney disease
600 AR	Bardet-Biedl syndrome
622 AR	Fanconi anemia
625 AD	Not affected
630 AD	KBG syndrome
644 AD	KBG syndrome
696 AD	Polycystic kidney disease adult type
709 AR	Mucopolysaccharidosis MPS-IV-A
717 AD	KBG syndrome
743 AD	Not affected

Table 3: Resulting diagnoses post variant evaluation.

5 References

1. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet. 2011 Sep 27;12(11):745-55.

2. Kumar V., Abbas A. K., Aster J. C. Robbins & Cotran *Pathologic Basis of Disease, 9th Edition.* "Elsevier" 2015

3. Landrum MJ, Lee JM, Benson M, et al. *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res. 2016 Jan 4;44(D1): D862-8

4. Eilbeck K., Quinlan A., Yandell M. *Settling the score: variant prioritization and Mendelian disease.* Nat Rev Genet 18, 599–612 (2017)

5. Parenti I, Mallozzi MB, Hüning I, et al. *ANKRD11 variants: KBG syndrome and beyond.* Clin Genet. 2021 Aug;100(2):187-200

6. Course materials: https://gt.ariel.ctu.unimi.it/v5/frm3/ThreadList.aspx?name=contents