Laura Mottarlini, mat. 30696A
Sabrina Pervin Abu, mat. 30507A

GENOMICS PROJECT
Rare genetic disease diagnosis in parents-child trios

1. INTRODUCTION

Rare genetic diseases, with a frequency within the population below 10^{-4}, are characterized by genetic mutations that can be inherited from parents or arise spontaneously. Understanding the inheritance patterns of these diseases is crucial for effective diagnosis.

These diseases can be classified into two categories based on their mode of inheritance: Autosomal Dominant Diseases and Autosomal Recessive Diseases. In Autosomal Dominant Diseases, mutations in a single allele are sufficient to cause the disease, while in Autosomal Recessive Diseases, the individual has to be homozygous for the pathogenic allele.
All of our Autosomal Dominant cases are *de novo*, which means that the parents are both healthy and the mutation occurs only in the child.

In our analysis we will refer to family trios.These are composed of the two parents and the child: the parents are healthy while the child could be affected by a rare Mendelian Autosomal Disease.
The trios we were assigned are reported in the table below.

CaseID	590	593	601	631	669	672	681	708	710	745
Inheritance Model	AD	AD	AR	AD	AD	AD	AD	AR	AD	AR

2. METHODS

Our approach involves a multi-stage analysis, incorporating different bioinformatics tools. We also developed a shell script that can be found at the following paths:

→ /home/BCG_2024_lmottarlini/script_genomics.sh
→ /home/BCG_2024_sabu/script_genomics.sh

With this script we were able to analyze all the assigned cases that we specified through an array.

First we checked the quality of each fasta file with the **FastQC** tool. We then proceeded to map each sequence with the reference sequence (/home/BCG2024_genomics_exam/uni) using **Bowtie2**, the specified parameters are:

- -p 8: is meant to specify the number of alignment threads to launch
- --rg-id "$name" --rg "SM:$name": to specify the origin of the sample (mother, father or child) in the @RG line of the .sam header, which will be later displayed in the .vcf header.

The output .sam files were converted into .bam files and sorted using **samtools**. For each .bam file we completed a quality control through **qualimap** and also created the .bg files (Bedgraph), that we later uploaded onto the genome browser, using **bedtools.**
The resulting quality reports from the sequencing and the mapping were merged into a single .html using **MultiQC.** We then created the multi-sample .vcf files for every trio using **freebayes**, the specified parameters are:

- -m 20: to set the minimum mapping quality at 20
- -C 5: a variant is considered as such if there are at least 5 reads covering it
- -Q 10: the threshold of mismatch base quality is set to 10
- -q 10: the supporting base quality of the alleles has to be greater or equal to 10
- --min-coverage 10: only sites with a coverage greater or equal to 10 are considered

Using **bcftools**, we alphabetically sorted the previously specified samples' names in order to have a consistent order in all the .vcf headers.

We concluded the analysis with variant prioritization.The cases were sorted into two arrays according to their inheritance model and using grep only the interesting variants were filtered.
The Autosomal Dominant Diseases, since being *de novo*, are characterized by the following pattern "0/1.*0/0.*0/0", while the Autosomal Recessive Diseases follow the "1/1.*0/1.*0/1" pattern.
Using **bedtools** we intersected these variants with the exome of reference found at the following path: /home/BCG2024_genomics_exam/exons16Padded_sorted.bed.

3. RESULTS AND DISCUSSION

The final .vfc files were uploaded to VEP (Ensembl Variant Effect Predictor) genome assembly GRCh37.
To determine if the variants cause a Mendelian Autosomal Rare Disease, the following parameters were taken into account:

- variant impact HIGH or, if no high impact variants are found, MODERATE
- allele frequency AF lower than 0.0001 or not defined
- SIFT score lower or equal to 0.02 (the variant is deleterious) and PolyPhen score higher or equal to 0.6 (tha variant is damaging or possibly damaging)

All the results are summarized in the table below.
The cases 590 631 and 669 report variants that are associated with autosomal dominant polycystic kidney disease. However, taking into account the allele frequency (AF), PolyPhen and SIFT values, these patients are not considered affected by any Mendelian Autosomal Rare Disease.

As to the other cases, these patients are considered to be affected with Mendelian Autosomal Rare Disease since the values of allele frequency (AF), PolyPhen and SIFT are not defined.
The variants associated with the diseases are also characterized by a high impact, as their consequences are either stop-gained or frameshift mutations.

caseID	Location	Consequence	Gene	SIFT	PolyPhen	AF	Associated phenotypes
case590	16:2140554-2140554	missense_variant	PKD1	0.11	0.006	0.0365	(see discussion)
case593	16:2140777-2140777	stop_gained	PKD1	-	-	-	Autosomal dominant polycystic kidney disease
case601	16:89858886-89858889	frameshift_variant	FANCA	-	-	-	Fanconi anemia complementation group A
case631	16:2140554-2140554	missense_variant	PKD1	0.11	0.006	0.0365	(see discussion)
case669	16:2140554-2140554	missense_variant	PKD1	0.11	0.006	0.0365	(see discussion)
case672	16:50820762-50820769	splice_acceptor_variant, coding_sequence_variant	CYLD	-	-	-	Familial cylindromatosis
case681	16:3820696-3820696	stop_gained	CREBBP	-	-	-	Rubinstein-Taybi syndrome due to CREBBP mutations
case708	16:28495408-28495408	stop_gained	CLN3	-	-	-	Neuronal Ceroid Lipofuscinosis
case710	16:3779061-3779065	frameshift_variant	CREBBP	-	-	-	Rubinstein-Taybi syndrome due to CREBBP mutations
case745	16:89838209-89838212	frameshift_variant	FANCA	-	-	-	Fanconi anemia complementation group A

For case 681, we designated the associated phenotype as *Rubinstein-Taybi syndrome due to CREBBP mutations* because we took into consideration that the specified gene affected by the mutation is CREBBP, despite also obtaining *Rubinstein-Taybi syndrome* from the VEP results.

UCSC genome browser

Focusing on case708, we uploaded the BedGraph files of the trio and the corresponding .vcf file onto the UCSC genome browser.

The results, as observed in Figure 1, confirm what was anticipated from the VEP analysis. The variant in position 16:28495408-28495408 shows a mutation of a single nucleotide from G (reference) to A (alternate).

In the considered case the coding strand is the reverse strand. The reference codon on the template strand is GTC (therefore on the coding strand is CAG).
In the child, due to the variation, the first nucleotide G is mutated into A. The resulting transcripted codon is UAG, which is a stop codon that interrupts the translation of the mRNA.

The genotypes associated with the trio (Figure 2) indicate that both parents are healthy carriers of the disease, while the child, characterized by two recessive alleles, manifests the Autosomal Recessive Disease.

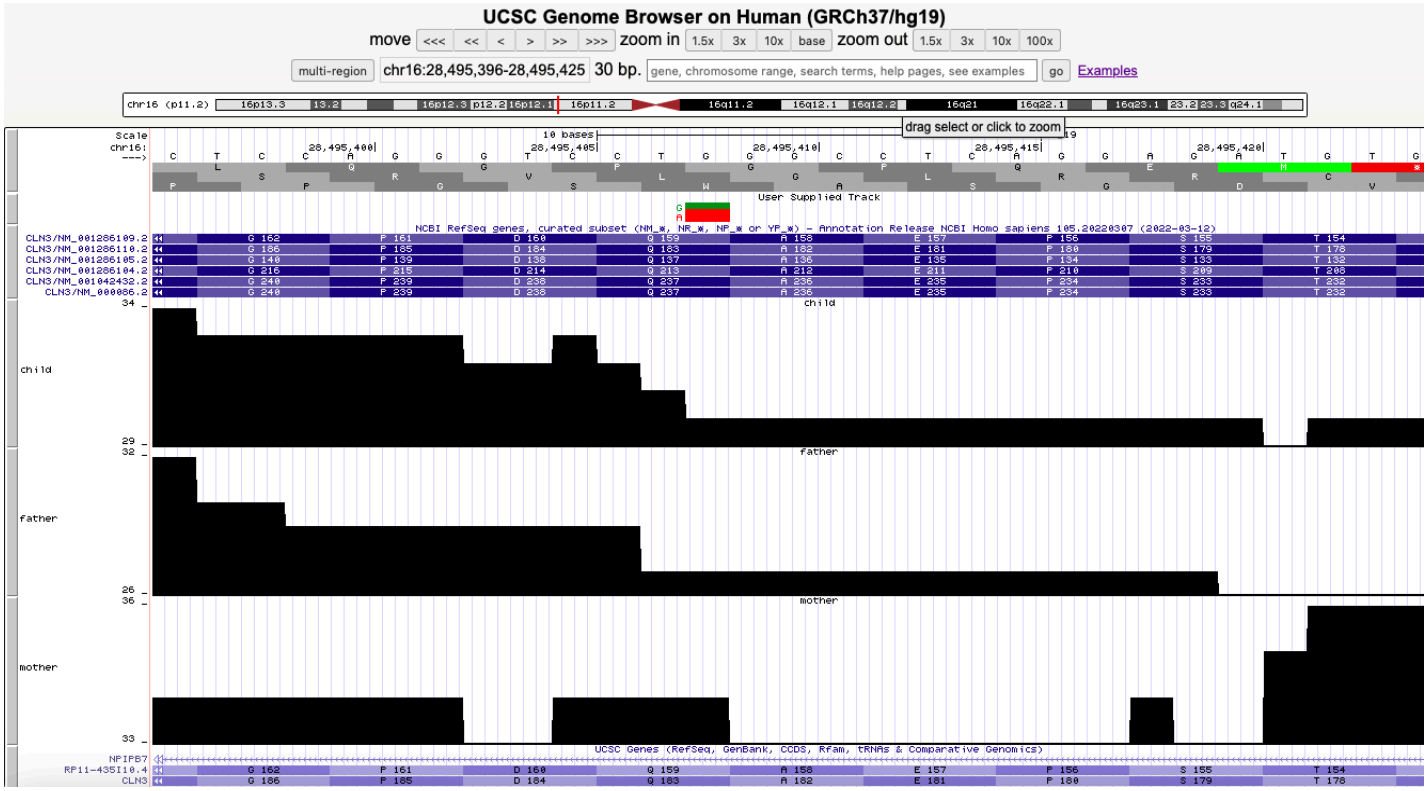


Figure1. [Genome Browser Analysis - Case708]: the stop-gained variant in the CLN3 gene is responsible for Neuronal Ceroid Lipofuscinosis.

Sample ID	Genotype	Phased?	DP	AD	RO	QR	AO	QA	GL
child	A/A	n	30	0, 30	0	0	30	991	-88.051000, -9.030900, 0.000000
father	G/A	n	27	10, 17	10	330	17	556	-41.437500, 0.000000, -21.411500
mother	G/A	n	34	14, 20	14	473	20	665	-48.936200, 0.000000, -31.895400

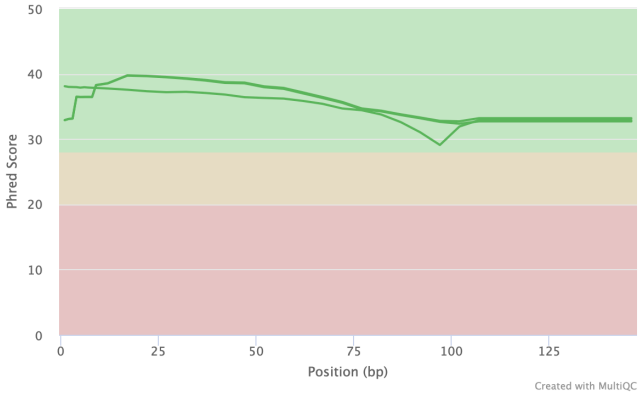
Figure 2. Genotype Table

Quality control

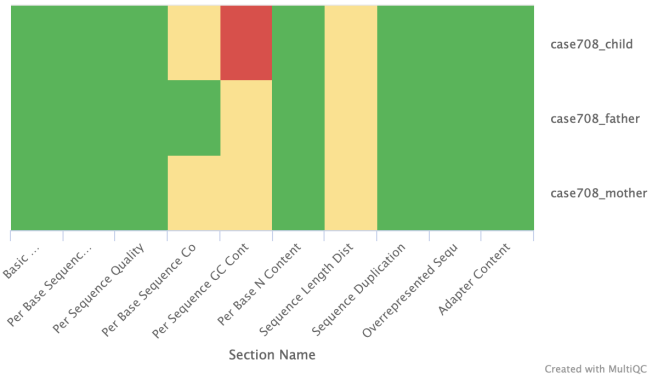
The MultiQC tool allows us to evaluate the quality of the sequencing and the mapping. After examining the results, we can state that the overall quality is good. However the GC content distribution of the child's sequencing results appear to be of lower quality. Nevertheless, our analysis can be considered reliable.

Sample Name	% GC	≥ 30X	Median cov	Mean cov	% Aligned	% Dups	% GC	M Seqs
case708_child	46%	22.8%	5.0X	24.1X	99.8%	5.4%	43%	3.0
case708_father	52%	31.2%	18.0X	27.3X	99.9%	6.1%	50%	2.2
case708_mother	52%	31.0%	18.0X	26.3X	99.8%	8.5%	50%	2.1

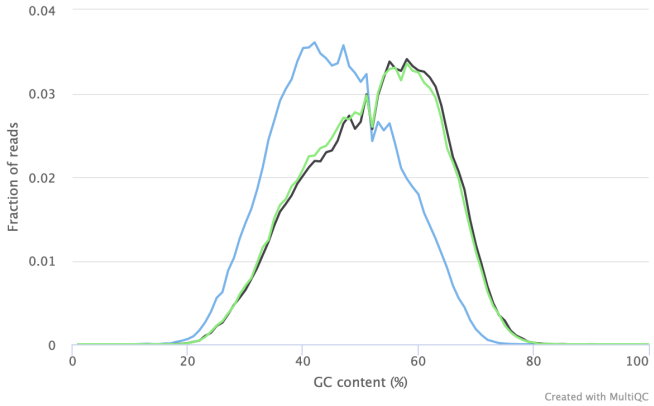
FastQC: Mean Quality Scores



FastQC: Status Checks



Qualimap BamQC: GC content distribution



```
#### make the script executable ####
### chmod +x script_genomics.sh ###
### nohup ./script_genomics.sh & ###

#!/bin/bash

if [ ! -d "project"]; then
            mkdir project
cd project

#create a directory for each assigned case
cases=(590 593 601 631 669 672 681 708 710 745)
for num in "${cases[@]}"; do
            if [ ! -d "case$num" ]; then
                        mkdir case$num
            fi
done

#check the FASTA files
names=("mother" "father" "child")
for num in "${cases[@]}"; do
            for name in "${names[@]}"; do
                        fastqc /home/BCG2024_genomics_exam/case${num}_${name}.fq.gz -o case${num}/
            done
done

#create the BAM files
for num in "${cases[@]}"; do
            for name in "${names[@]}"; do
                        bowtie2 -U /home/BCG2024_genomics_exam/case${num}_${name}.fq.gz \
                                    -p 8 -x /home/BCG2024_genomics_exam/uni --threads 4 \
                                    --rg-id "$name" --rg "SM:$name" \
                                    | samtools view -Sb -@ 4 \
                                    | samtools sort -@ 4 -o case${num}/case${num}_${name}.bam
            done
done

#create the BG files from each BAM file
for num in "${cases[@]}"; do
            for name in "${names[@]}"; do
                        bedtools genomecov -ibam case${num}/case${num}_${name}.bam \
                                    -bg -trackline -trackopts "name="$name"" -max 100 \
                                    > case${num}/${name}Cov.bg
            done
done

#do the quality check for each BAM file (with target exome)
for num in "${cases[@]}"; do
            for name in "${names[@]}"; do
                        qualimap bamqc -bam case${num}/case${num}_${name}.bam \
                                    -gff /home/BCG2024_genomics_exam/exons16Padded_sorted.bed \
                                    -outdir case${num}/case${num}_${name}
            done
done

#do multiqc
for num in "${cases[@]}"; do
            multiqc case${num}/ -o case${num}/
done

#create the multi-sample VCF files
mkdir results
mkdir final

for num in "${cases[@]}"; do
            freebayes -f /home/BCG2024_genomics_exam/universe.fasta \
                        -m 20 -C 5 -Q 10 --min-coverage 10 case${num}/case${num}_mother.bam \
                        case${num}/case${num}_father.bam case${num}/case${num}_child.bam \
                        > results/case${num}.vcf
done

#sort the sample's names in the VCF files
for num in "${cases[@]}"; do
            bcftools query -l results/case${num}.vcf \
                        | sort > results/samples_${num}.txt
done

for num in "${cases[@]}"; do
            bcftools view -S results/samples_${num}.txt results/case${num}.vcf \
                        > results/case${num}.sorted.vcf
done

rm results/*.txt

#variant prioritization
AR=(601 708 745)
AD=(590 593 631 669 672 681 710)

AD_pipeline () {
            num=$1
            grep "#" results/case${num}.sorted.vcf > results/candilist${num}.vcf
            grep "0/1.*0/0.*0/0" results/case${num}.sorted.vcf >> results/candilist${num}.vcf
            bedtools intersect -a results/candilist${num}.vcf \
                        -b /home/BCG2024_genomics_exam/exons16Padded_sorted.bed -u > final/${num}TG.vcf
}

for num in "${AD[@]}"; do
            AD_pipeline "$num" &
done


AR_pipeline () {
            num=$1
            grep "#" results/case${num}.sorted.vcf > results/candilist${num}.vcf
            grep "1/1.*0/1.*0/1" results/case${num}.sorted.vcf >> results/candilist${num}.vcf
            bedtools intersect -a results/candilist${num}.vcf \
                        -b /home/BCG2024_genomics_exam/exons16Padded_sorted.bed -u > final/${num}TG.vcf
}

for num in "${AR[@]}"; do
            AR_pipeline "$num" &
done


#END OF THE SCRIPT
```