University of Milan

Dept of Biosciences

Genomics Project Report: Molecular Diagnosis of Rare Genetic Disorders

Gabriele Ghezzi, Lorenzo Paratici
49560A, 50974A

Genomics 2025
Matteo Chiara
May 4, 2025

1. Introduction

- **Objective**: Diagnose rare Mendelian disorders in 10 individuals using exome sequencing data from chromosome 16 (hg19).
- **Context**:
    - Analysis of simulated TRIO data (father, mother, child) with children potentially affected by autosomal dominant (AD) or recessive (AR) disorders.
- **Key Tools**: Qualimap, BEDTools, BCFTools, Bowtie2, samtools, freebayes, fastqc, VEP, UCSC Genome Browser

---

2. Methods

The tools used for the analysis are all provided by the UNIX operating system and the UNIX shell.
To speed up the process we decided to build an automatic pipeline with some useful checkpoints that we used for all the cases. We decided to report, as an example, the code of only one of the three files that is present in every case (father).

2.1 Data Processing Pipeline

All steps were performed using GRCh37/hg19 as the reference genome build. Reads were aligned to hg19 using Bowtie2, and all downstream analyses were matched to this genome version to ensure consistency, especially for BEDTools intersection and VEP annotation.

- **Alignment**: bowtie2, samtools

```
bowtie2 -x $GENOME_INDEX -U $RAW_DATA_DIR/case${CASE_NUM}_father.fq.gz \
--rg-id father${CASE_NUM} --rg "SM:father${CASE_NUM}" --rg "PL:ILLUMINA" | \
samtools view -bSu - | samtools sort -@ 4 -o \
$BAM_DIR/case${CASE_NUM}_father.bam

samtools index -@ 4 $BAM_DIR/case${CASE_NUM}_father.bam
```

- **Variant Calling**: freebayes

```
freebayes -f $GENOME_FASTA -m 20 -C 5 -Q 10 --min-coverage 10 \
$BAM_DIR/case${CASE_NUM}_father.bam $BAM_DIR/case${CASE_NUM}_mother.bam
$BAM_DIR/case${CASE_NUM}_child.bam -t $EXONS_BED \
> $VCF_DIR/case${CASE_NUM}_variants.vcf
```

- **Quality Control**:
    a. FastQC for raw read quality.

```
fastqc -o $SEQ_QUALITY_DIR $RAW_DATA_DIR/case${CASE_NUM}_father.fq.gz
```

b. Qualimap for target region coverage (--feature-file exons16Padded_sorted.bed).

```
qualimap bamqc -bam $BAM_DIR/case${CASE_NUM}_father.bam -gff \
$GENOME_DIR/*.bed -outdir $MAP_QUALITY_DIR/case${CASE_NUM}_father_qualimap
```

c. MultiQC to aggregate reports.

```
multiqc ./seq_quality/case${CASE_NUM}* ./map_quality/case${CASE_NUM}* -o \
./multiqc_reports/case${CASE_NUM}
```

2.2 Variant Prioritization

- **Intersection and filtering**: Filter variants by quality and depth, keeping only those in exonic regions with high variant quality (QUAL > 30), sufficient overall coverage (INFO/DP ≥ 20), and at least one sample with read depth ≥ 15 (FMT/DP ≥ 15).

```
bcftools filter -i 'QUAL > 30 && FMT/DP>=15 && INFO/DP>=20' \
./vcfs/case${CASE_NUM}_variants.vcf | \
bedtools intersect -a stdin -b ./genome/exons16Padded_sorted.bed -header -u \
> ./vcfs/case${CASE_NUM}_targeted.vcf
```

- **Detecting samples order:**

```
samples=$(bcftools query -l ${VCF_DIR}/case${CASE_NUM}_variants.vcf)
sample_array=($samples)
mother_index=-1
father_index=-1
child_index=-1

for i in "${!sample_array[@]}"; do
      if [[ "${sample_array[$i]}" == "mother${CASE_NUM}" ]]; then
            mother_index=$i
      elif [[ "${sample_array[$i]}" == "father${CASE_NUM}" ]]; then
            father_index=$i
      elif [[ "${sample_array[$i]}" == "child${CASE_NUM}" ]]; then
            child_index=$i
      fi
done

if [[ $mother_index -eq -1 || $father_index -eq -1 || $child_index -eq -1 ]];
then
      echo "Error: Could not find the required samples (mother${CASE_NUM},
father${CASE_NUM}, child${CASE_NUM}"
      exit 1
fi
```

- **Autosomal Dominant (AD)**:
    - Filter for heterozygous variants (0/1) in children absent in parents (DNMs).

```
FILTER="GT[$mother_index]='0/0' && GT[$father_index]='0/0' && \
GT[$child_index]='0/1'"

bcftools view -i "$FILTER" ${VCF_DIR}/case${CASE_NUM}_targeted.vcf -o \
${VCF_DIR}/case${CASE_NUM}_dominant.vcf
```

- **Autosomal Recessive (AR)**:
    - Filter for homozygous variants (1/1) in children with heterozygous (0/1) parents.

```
FILTER="GT[$mother_index]='0/1' && GT[$father_index]='0/1' &&
GT[$child_index]='1/1'"

bcftools view -i "$FILTER" ${VCF_DIR}/case${CASE_NUM}_targeted.vcf -o
${VCF_DIR}/case${CASE_NUM}_recessive.vcf
```

2.3 Annotation & Visualization

- **VEP**: Annotate variants using RefSeq transcripts, including only the 'Phenotype' field from the phenotype and citation section. Filter for high-impact variants and retain only those with a gnomAD exome allele frequency (AF) $\leq 1\times10^{-4}$ (when reported).
- **UCSC Genome Browser**:
    - Upload coverage tracks (bedtools genomecov -bg).

```
bedtools genomecov -ibam ./bam/case${CASE_NUM}_father.bam -bg \
-trackline -trackopts 'name="father"' -max 100 > \
./bgs/case${CASE_NUM}_fatherCov.bg
```

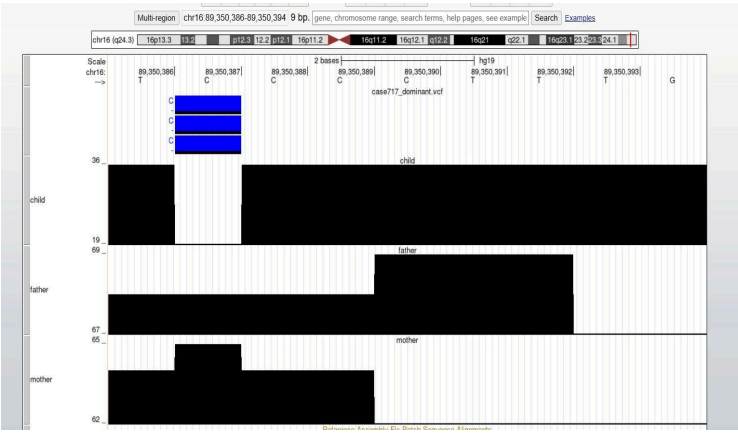
    - Visualize candidate variants and examine their locations based on VEP annotations

3. Results

3.1 Diagnostic Findings

Case	Chr	POS	Variant Type	Gene	Inheritance	Disease Diagnosis
617	16	89877443-89877450	Frameshift	FANCA	AR	Fanconi Anemia
628	16	89815164-89815174	Frameshift	FANCA	AR	Fanconi Anemia
657	16	50788249-50788254	Frameshift	CYLD	AD	Familial cylindromatosis
717	16	89350386-89350394	Frameshift	ANKRD11	AD	KBG syndrome
737	16	3779202-3779206	Frameshift	CREBBP	AD	Rubinstein-Taybi syndrome due to CREBBP mutations
596	16	89877468-89877469	Stop_gained	FANCA	AR	Fanconi Anemia
639	16	3807957-3807959	Frameshift	CREBBP	AD	Rubinstein-Taybi syndrome due to CREBBP mutations
642	16	3789616-3789616	Stop_gained	CREBBP	AD	Rubinstein-Taybi syndrome due to CREBBP mutations
645	16	3778641-3778644	Frameshift	CREBBP	AD	Rubinstein-Taybi syndrome due to CREBBP mutations
701	16	3901012-3901012	Splice_acceptor	CREBBP	AD	Rubinstein-Taybi syndrome due to CREBBP mutations

3.2 UCSC Genome Browser Example



At genomic position chr16:89350387 (band 16q24.3), a heterozygous deletion (frameshift) was detected in the child (child717) with a read depth of 36x and balanced allele support (19 reference reads, 17 alternate reads). Both parents (mother717 and father717) showed normal homozygous genotypes without evidence of the deletion, supported by 56x and 66x read depth, respectively. These findings strongly support the presence of a de novo frameshift mutation in the child.

Detailed genotypes

Genotype info key:
DP: Read Depth
AD: Number of observation for each allele
RO: Reference allele observation count
QR: Sum of quality of the reference observations
AO: Alternate allele observation count
QA: Sum of quality of the alternate observations
GL: Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy

Sample ID	Genotype	Phased?	DP	AD	RO	QR	AO	QA	GL
child717	(T)C/(T)-	n	36	19, 17	19	632	17	551	-38.325800, 0.000000, -45.410600
mother717	(T)C/(T)C	n	56	56, 0	56	2074	0	0	0.000000, -16.857700, -180.044000
father717	(T)C/(T)C	n	66	66, 0	66	2419	0	0	0.000000, -19.868000, -209.843000

Haplotype sorting order: using middle variant in viewing window as anchor.

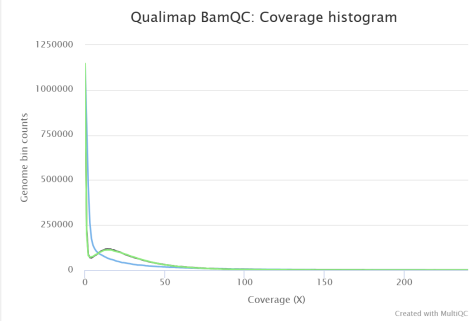
3.3 Resulting VCF files

The resulting VCF files reside on the server at `/home/BCG2025_ghezziG/genomics_project/vcfs`
The one used for VEP annotations are: case*_dominant/recessive.vcf

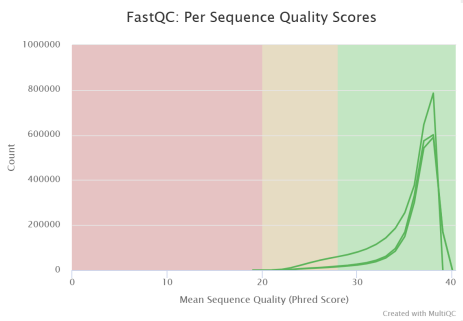
4. Quality Report

Sample Name	% GC	≥ 30X	Median cov	Mean cov	% Aligned	% Dups	% GC	M Seqs
case717_child				24.1X	99.8%	5.4%	43%	3.0
case717_child_qualimap	46%	22.8%	5.0X					
case717_father				27.3X	99.9%	6.1%	50%	2.2
case717_father_qualimap	52%	31.2%	18.0X					
case717_mother				26.3X	99.8%	8.5%	50%	2.1
case717_mother_qualimap	52%	31.0%	18.0X					

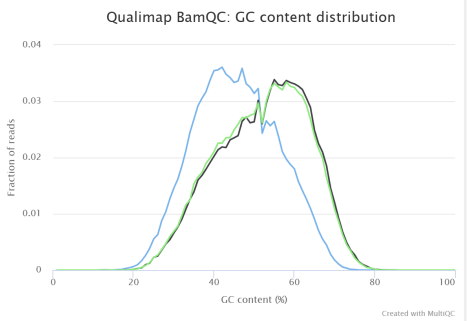
The table shows high alignment rates (>99%) and sufficient mean coverage across all three samples. GC content and duplication levels are within expected ranges. These metrics confirm good sequencing quality, enabling reliable variant analysis on exonic regions of chromosome 16.



Parents display expected exome coverage patterns with a local peak around 20×. The child shows a more uniform decline, lacking the same peak, and has reduced ≥30× coverage (22.8% vs. ~31% in parents), which may slightly reduce sensitivity for heterozygous variant detection.



Most reads in all three samples have a mean Phred score above 30, indicating high base-calling accuracy and overall reliable sequencing quality.



While all samples show a typical bell-shaped GC distribution, the child's GC peak (~43%) is notably lower than the parents' (~50–60%). This may reflect minor sequencing bias or uneven target capture, though it remains within expected ranges for exome data.

The quality metrics confirmed that the sequencing data for all three samples is sufficient for reliable variant calling and annotation. The child showed slightly reduced ≥30× coverage (22.8%) and a lower GC peak (~43%) compared to the parents, which may reflect sequencing or capture bias. However, average depth and alignment rates remained high, justifying continued analysis. No significant adapter contamination or overrepresented sequences were observed.

5. Conclusion

This pipeline enabled accurate variant detection in simulated exome trio data from chromosome 16.
Both AD and AR inheritance patterns were evaluated, enabling successful diagnoses in all examined cases.
Coverage and GC metrics showed minor deviations in the child but did not compromise overall data quality.
VEP annotation allowed functional interpretation of rare variants in known disease genes.
This approach demonstrates the power of WES in rare Mendelian disease investigation.