OXFORD GENETICS

Clinical Diagnosis of Rare Mendelian Diseases in family Trios

Caccia Riccardo[1,†] and Gautieri Giuseppe[2,†]

[1]BCG, Univesità degli Studi di Milano Statale
[2]BCG, Univesità degli Studi di Milano Statale
[†]These authors contributed equally to this work.

Abstract

In this analysis, the focus is on autosomal inheritance, which involves the 22 non-sex chromosomes in humans. Mendelian disorders inherited through autosomal chromosomes typically follow two main patterns: autosomal dominant and autosomal recessive. In autosomal dominant inheritance, the presence of a single pathogenic allele is sufficient to cause the disease. On the other hand, autosomal recessive conditions require both alleles to be altered for the phenotype to manifest. In this study, 10 family trios will be analyzed, each of them with suspected Mendelian disorders, with the goal of identifying potential disease-causing variants.

Keywords: Autosomal; Mendelian; Inheritance

The goal of this project is to obtain a clinical diagnosis for the children of ten families trios, each suspected to be affected by a rare Mendelian genetic disorder. In this specific analysis, the focus was placed on a subset of genetic diseases affecting only chromosome 16. Consequently, the final part of the pipeline was dedicated exclusively to the examination of this specific genomic region. All code, images, and results related to this project are available at this GitHub repository.

Introduction

Mendelian genetic disorders (*Kennedy, 2005*) are a specific set of diseases that are caused by a single gene, with a direct genotype variant effect on the phenotype of the individual. Following the first and second Mendel law, there are two possible patterns of diseases which in the autosomal part of the genome that do not affect the reproductive cells and can be divided into dominant and recessive. In fact, in the first Mendel law, two categories can be distinguished: dominant and recessive alleles, where one of the alleles is sufficient to display the phenotype in the first case, while both alleles are necessary in the second case.

Materials and methods

In order to perform the analysis, simulated data were provided in advance. For each case, the starting point was three zipped FASTQ files (*Cock et al., 2009*) obtained from sequencing. Since proper analysis requires mapping the reads against a reference genome, a file containing the hg19 reference was also provided. Before starting the analysis, some background information about the samples was given:

- All the parents were healthy;
- For each genetic site, only one alternative to the reference allele was considered, using a 0 (reference) / 1 (alternative) format;
- Is Known in advance the suspected type of disease for the family (Autosomal Dominant or Autosomal Recessive).

The initial step involved observing the quality of the sequencing reads using **FastQC** (*Bittencourt, 2010*), to ensure that downstream variant calls would be reliable. Once quality control was completed, the reads were mapped to the reference genome using **Bowtie2** (*Langmead et al., 2012*), with sample identifiers specified via appropriate parameters to enable variant separation in the final joint VCF file for the family.

The alignment output in SAM format was converted to BAM, sorted, and indexed using **Samtools** (*Li H et al., 2009*). At this stage, a second quality control was performed using **Qualimap** (*Garcia-Alcalde et al., 2012*).

Variant calling was then conducted using **FreeBayes** (*Garrison et al., 2012*), producing a multi-sample VCF file containing all variants across the family (the link to all VCF file to this GithHub repository). To limit the analysis to potentially disease-related variants, **Bedtools** (*Quinlan et al., 2010*) was used to intersect the variant calls with a BED file containing exon coordinates for chromosome 16. A filtering step with grep was then applied to retain only those variants matching the disease-specific pattern, excluding non-relevant ones.

Than all quality reports were aggregated using MultiQC (*Ewels et al., 2016*, here the Github repository containing the results), and coverage files were generated with Bedtools for visualization in the UCSC Genome Browser (*Karolchik et al., 2003*).

At this stage, the dataset was ready for variant interpretation and clinical assessment using tools such as Variant Effect Predictor (VEP) *McLaren W. et al., 2016*, the UCSC Genome Browser (here the Github repository containing the images), and IGV (*Thorvaldsdòttir et al., 2013*)(here the Github repository containing the

images).

The complete pipeline is described in the Data availability section and is publicly accessible on the authors' Github page.

Results

One of the family has been taken as an example in order to show the results and the analysis conducted, the table of the diagnosis containing final conclusion is reported at the end of Discussion section.
Specifically, the analysis will take as an example case 584. First of all MultiQC report were analyzed to retrieve some information about quality, the summary is visible in *Figure 1*.

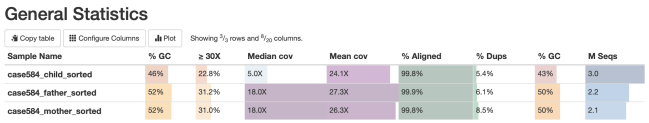

Figure 1. MultiQC statistics.

The General Statistics part in MultiQC reports provide a quantitative summary of the key metrics obtained during the quality control and bioinformatic analysis of the samples. This panel aggregates data from the tools used in the workflow, such as Fastqc and Qualimap.
It is possible to observe the good quality of both the mean coverage, higher than 24X for each member of the family, and also the fraction of aligned genome is very good since almost every read has been aligned.

In the report, it is possible to see several quality score plots and an overview of the alignment. Below, two plots have been selected: the per-sequence quality score plot in *Figure 2* and the mean quality score per position plot in *Figure 3*.

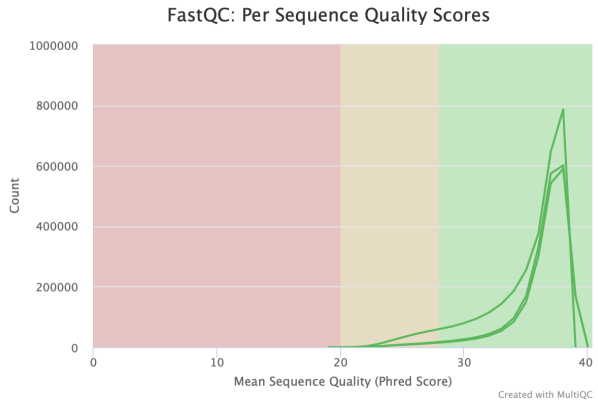

Figure 2. Sequence Quality Score.

This first quality report was extracted directly from the fastq files, taking into account only the quality of the sequence.
As the figure shows, the mean quality of the sequence is almost entirely shifted towards the very good quality, having the peak of counts with a PHRED score larger than 35.

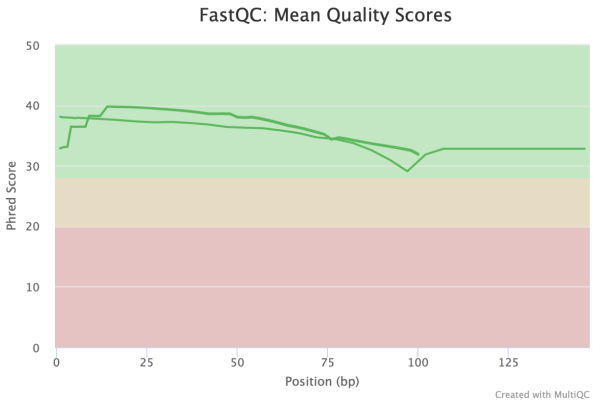

Figure3. Mean Quality Score per Position.

This second report shows the sequence quality per position, it confirms the good quality of the previous plot since the level never drops in the yellow region.

Having assessed the good quality of the reads, it is possible to look at the Genome covered plot in *Figure 4*.

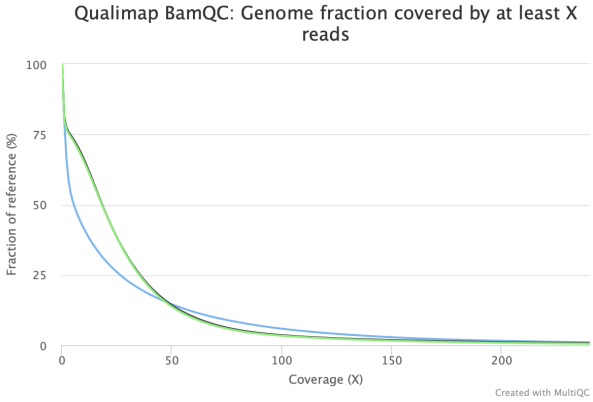

Figure 4. Genome Fraction Coverage.

This third report has been generated from the bam files after the mapping process and express the general level of coverage of the three family components.
The figure shows the coverage for genome fraction. Even if the child coverage seems to have less smooth presence with respect to the parents, also in this case the quality of both parents and child can be accepted, having an important fraction of the genome with at least 20X coverage and even a portion going higher than 50X.

After quality checks, the reads were analyzed and the results were compared using both the UCSC Genome Browser and IGV. The results of these analyses with these software are reported in Figure 5 and Figure 7, it also reported in Figure 6 a summary of the Genome Browser view.

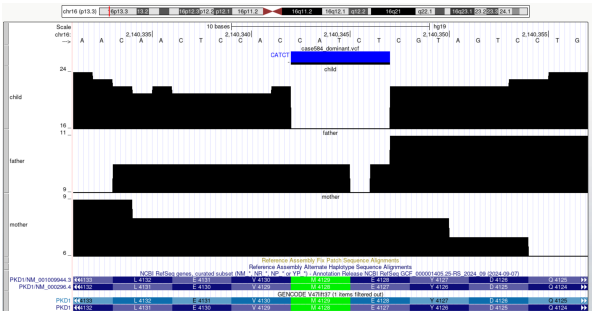

Figure 5. UCSC Genome Browser view.

	SAMPLEID	GenoType	DP	AD	RO	QR	AO	QA	GL
1	father	0/0	8	8,0	8	0	257	0	0,-2.40824,-22.9107
2	child	0/1	22	16,6	16	6	527	198	-11.0763,0,-40.3722
3	mother	0/0	8	8,0	8	0	285	0	0,-2.40824,-25.2579

Figure 6. Genotype key information.

The UCSC genome browser view shows the variant, detectable in the child with good coverage but not in the parents. Below, the affected gene (PKD1) and mutation position are indicated. The table includes read counts, quality sums for reference and alternate alleles, and genotype probability.

IGV software is then used to support the conclusion from UCSC GEnome Browser, here reported the view of the final output.

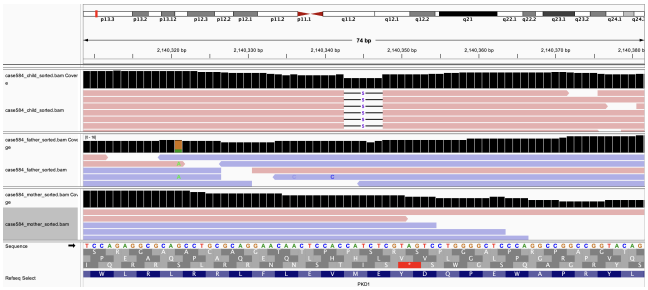

Figure 7. IGV view.

In Figure 7 the IGV representation of the same variant is showed. With this other tool is much more appreciable the presence of gap between the child genome and the parents, also here with an important coverage. Since both the representation supports the presence of the variant, the conclusion is that the child of this family should be affected by the disease.

Discussion

In order to identify a potential genetic cause of a Mendelian disease, is required to select the most relevant variants in the genome, note that not all the variants present in patient genome impact his health. In this study a pipeline was implemented to assess variants as Frameshift and stop-gain (among the most impactful). The final diagnosis was facilitated by combining quality control metrics, coverage analysis, and annotation tools like VEP and genome browsers. Notably, several cases showed variants consistent with known autosomal dominant disorders (e.g., PKD1 in polycystic kidney disease), while others matched autosomal recessive patterns (e.g., GALNS in MPS-IV-A) and others no relevant mutations at all(e.g., intron mutation, silent mutation). Some families (case 629 and case 714) were found not to carry variants, showing the pipeline's specificity in filtering out benign variants.

Data availability

Here is the entire pipeline code:

```
    # all cases
cases=("case747" "case685" "case714" "case629"
"case723" "case584" "case688" "case696" "case590"
"case730")

# recessive cases
recessive_cases="case688 case747 case685"

# reusable path
REFERENCE="/home/BCG2025_genomics_exam/universe.fasta"
BOWTIE_INDEX="/home/BCG2025_cacciaR/all_var/uni"
EXONS_BED="/home/BCG2025_genomics_exam/exons16Pad
ded_sorted.bed"
VEP_CACHE="/home/BCG2025_genomics_exam/vep_cache"

# index building if not present
if [ ! -f "${BOWTIE_INDEX}.1.bt2" ]; then
    echo "Building Bowtie2 index..."
    bowtie2-build "$REFERENCE" "$BOWTIE_INDEX"
fi

# for loop for each case
for case in "${cases[@]}"; do
    echo "Processing $case..."

    mkdir -p QC_reports_"$case"
    mkdir -p VCF_outputs_"$case"
    mkdir -p coverage_"$case"

    mother="/home/BCG2025_genomics_exam/${case}_
    mother.fq.gz"
    father="/home/BCG2025_genomics_exam/${case}_
    father.fq.gz"
    child="/home/BCG2025_genomics_exam/${case}_
    child.fq.gz"

    # FastQC report
    fastqc "$mother" "$father" "$child"
    -o QC_reports_"$case"/

    # Bowtie2 alignment
    bowtie2 -x "$BOWTIE_INDEX" -U "$mother" -S "${case}_
    mother.sam" --rg-id "mother" --rg "SM:mother"

    bowtie2 -x "$BOWTIE_INDEX" -U "$father" -S "${case}_
    father.sam" --rg-id "father" --rg "SM:father"

    bowtie2 -x "$BOWTIE_INDEX" -U "$child"  -S "${case}_
    child.sam"  --rg-id "child"  --rg "SM:child"

    # SAM → sorted BAM
    samtools view -bS "${case}_mother.sam" |
    samtools sort -o "${case}_mother_sorted.bam"

    samtools view -bS "${case}_father.sam" |
    samtools sort -o "${case}_father_sorted.bam"

    samtools view -bS "${case}_child.sam"  |
    samtools sort -o "${case}_child_sorted.bam"
```

```
# BAM indexing
samtools index "${case}_mother_sorted.bam"
samtools index "${case}_father_sorted.bam"
samtools index "${case}_child_sorted.bam"

# Qualimap
qualimap bamqc -bam "${case}_child_sorted.bam"
--gff  "$EXONS_BED"
-outdir "QC_reports_${case}/child"
qualimap bamqc -bam "${case}_father_sorted.bam"
--gff "$EXONS_BED"
-outdir "QC_reports_${case}/father"
qualimap bamqc -bam "${case}_mother_sorted.bam"
--gff "$EXONS_BED"
-outdir "QC_reports_${case}/mother"

# Variant calling
freebayes -f "$REFERENCE" -m 20 -C 5 -Q 10
--min-coverage 10 "${case}_mother_sorted.bam"
"${case}_father_sorted.bam"
"${case}_child_sorted.bam"
> "VCF_outputs_${case}/${case}_trio.vcf"

# Filtering on target region
bedtools intersect -a
"VCF_outputs_${case}/${case}_trio.vcf"
-b "$EXONS_BED" -u > "VCF_outputs_
${case}/${case}_filtered.vcf"

# Annotation with VEP
~/.vep/vep -i
"VCF_outputs_${case}/${case}_filtered.vcf"
--cache --dir "$VEP_CACHE"
--o "VCF_outputs_${case}/$
{case}_annotated.vcf" --vcf

# AR/AD classification
if [[ " $recessive_cases " =~ " $case " ]]; then
    echo "-> Recessive case detected for $case"
    grep "#"
    "VCF_outputs_${case}/${case}_annotated.vcf"
    > "VCF_outputs_${case}/${case}_recessive.vcf"
    grep "1/1.*0/1.*0/1" "VCF_outputs_${case}/$
    {case}_annotated.vcf" >>
    "VCF_outputs_${case}/${case}_recessive.vcf"
else
    echo "-> Dominant case detected for $case"
    grep "#" "VCF_outputs_${case}/${case}_trio.vcf"
    > "VCF_outputs_${case}/$
    {case}_dominant.vcf"
    grep "0/1.*0/0.*0/0" "VCF_outputs_${case}/$
    {case}_filtered.vcf" >> "VCF_outputs_${case}/$
    {case}_dominant.vcf"
fi

# MultiQC
multiqc QC_reports_"$case"/

# Coverage UCSC
bedtools genomecov -ibam "${case}_child_sorted.bam"
-bg -trackline -trackopts 'name="child"'  -max 100 >
"coverage_${case}/child${case}Cov.bg"
    bedtools genomecov -ibam "${case}_mother_sorted.bam"
    -bg -trackline -trackopts 'name="mother"' -max 100 >
    "coverage_${case}/mother${case}Cov.bg"
    bedtools genomecov -ibam "${case}_father_sorted.bam"
    -bg -trackline -trackopts 'name="father"' -max 100 >
    "coverage_${case}/father${case}Cov.bg"

    echo "Completed $case"
  done

echo "Pipeline COMPLETED for all cases!"
```

At the beginning, all case names are stored in a list to allow iteration, a second list is used to mark recessive cases in order to facilitate later analysis. For convenience key file paths were stored in variables. If not already present, the Bowtie2 index is built. Then, for each case, output directories are created and FastQC is run on the input reads. The reads from all three family members are aligned to the reference genome previously built using Bowtie2 to produce SAM files.
Once the SAM files are stored, with samtools, these files are converted to BAM format with -Sb, sorted and indexed. A Qualimap report is then created for each family members with Qualimap command, note that only chromosome 16 exons are taken in consideration with –gff, this because the locus of the disease are previously known.
A VCF file is then obtained by FreeBayes command, used for variant calling, specifying some thresholds parameters to filter the variants such as -m as mapping quality, -C as minimum count to support the variant and -q as base quality. This command used the sorted bam file of each family components and the reference genome to give a file with all the trio variants. Bedtools intersect is then used for selecting only the relevant variant in the genomes, the ones that are present in the exons. VEP, then, takes the remaining variant and annotate them with biological information.
Depending on the type of inheritance (dominant or recessive), grep variants were filtered to identify consistent genotype patterns: $0/1$ for dominant variant, $1/1$ for recessive one. A complete VCF file is then constructed over these assumptions and ready to be analyzed by Variant Effect Predictor tool from Ensembl.
Subsequently, a MultiQC report is produced using multiqc command, and then put in a dedicated directory. In the end, to computes the coverage of sequencing reads over a genome and obtaining the BED files (that needs to be uploaded to the UCSC Genome Browser to visualize the readings) bedtools genomecov was used. Bedtools is necessary to summarize how deeply and how evenly a genome has been sequenced- with -ibam a BAM file is passed as an input, with -bg the results is returned in BEDGRAPH format, -trackline and -trackpots helps with the visualization of the name in the Genome Browser, the first make the label visible to the user and the second specify the label name, and -max limit the coverage to be a certain value, all the reads that overcome this value is reported with the cap value.

Cases	Mutated gene	Variant location	REF//ALT	Consequence	Associated phenotype
case 584	PKD1	16:2143906-2143914	- // CCCACCCT / CCCCCACCCT	frameshift	Autosomal dominant polycystic kidney disease
case 590	CREBP	16:56508722-56508739	TTTTTTTTTTTCCTTAG // TTTTTTTTTTC-CTTAG / TTTTTTTTTTTTC-CTTAG	splice acceptor variant	Rubinstein-Taybi syndrome
case 629	N/A	N/A	N/A	N/A	healty
case 685	GALNS	16:88901702-88901702	G // G/A	stop gained	Mucopolysaccharidosis, MPS-IV-A
case 688	GALNS	16:88901627-88901627	C C/A	stop gained	Mucopolysaccharidosis, MPS-IV-A
case 696	PKD1	16:2140149-2140151	G // GA/A	frameshift	Autosomal dominant polycystic kidney disease
case 714	N/A	N/A	N/A	N/A	healty
case 723	CREBP	16:3830734-3830740	T // TTTATG/TTATG	frameshift	Rubinstein-Taybi syndrome
case 730	PKD1	16:2143906-2143914	- // CCCACCCT/CCCCCACCCT	frameshift	Autosomal dominant polycystic kidney disease
case 747	FANCA	16:89816207-89816212	G // GGAGC/GAGC	frameshift	Fanconi anemia

Table 1 Diagnosis table of all familiar cases

Conclusion

In conclusion, the described pipeline has been used in order to achieve the diagnosis shown in the table 1. Since the variants found in the early steps of the analysis were also spotted in clinical tools associated to various types of diseases, the quality check and thresholds used during the different steps were effective in distinguish among all the data the disease-causing ones. While most cases showed clear genetic evidence aligning with the clinical phenotype, cases 714 and 629 required deeper investigation. In the absence of relevant high-impact variants, the analysis was extended to include moderate-impact missens mutations. Although these variants were predicted as tolerated by SIFT (e.g., score = 0.12), even if an high PolyPhen scores (e.g., > 0.94, "possibly damaging") from Ensembl suggested a potential impact on protein function their extremely low allele frequency in population databases such as gnomAD (e.g., < 1e-5) reinforced their rarity.

Nevertheless, in the absence of strong functional or clinical validation, a cautious approach was adopted. Based on the available evidence, these samples were classified as likely healthy to avoid overinterpretation of uncertain findings.

Literature cited

Bittencourt S. 2010. Fastqc: a quality control tool for high-throughput sequence data. Babraham Bioinformatics. .

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2009. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. Nucleic Acids Research. 38:1767–1771.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. Multiqc: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32:3047–3048.

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. 2012. Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics. 28:2678–2679.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. .

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ *et al*. 2003. The ucsc genome browser database.

Kennedy MA. 2005. in *Mendelian Genetic Disorders*. Wiley.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nature Methods. 9:357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and samtools. Bioinformatics. 25:2078–2079.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. Genome Biology. 17.

Quinlan AR, Hall IM. 2010. Bedtools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (igv): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 14:178–192.