The ABC of variant calling: a guide for the diagnosis of genetic diseases.

Denisa Sufaj and Flavia Leotta
Università degli Studi di Milano, Milano, Italy.

Variant calling is the process by which variants are identified from sequence data: bioinformatics is important in each stage of this process and it's essential for handling genome-scale data. In our study we analyzed 10 parents-child trios' exomes to find variants in the child's 16th chromosome sequence that could lead to rare Mendelian genetic diseases, knowing that the parents were healthy. The family trios underwent WES, and we received the resulting *fastq* files: the analysis of WES data through an Unix pipeline was employed to identify and characterize genetic variations within the individuals' exomes. We aligned the sequences to the reference hg19 assembly human genome, and then the identified variants were filtered and prioritized based on their potential relevance to the patient's phenotype (observable traits) and suspected genetic condition. This process involved comparing the variants against a database of known disease-causing mutations (VEP) and assessing their predicted functional impact. In our results we show that 6 out of the 10 cases show variants that are deemed to be potentially disease-causing, which should be prioritized for further validation and confirmation through additional genetic testing.

1. Introduction

Rare genetic diseases are a significant challenge in diagnosis and treatment due to their low frequency and diverse clinical development. Globally there are more than 300 million people living with rare diseases and around 80% of these conditions have a genetic basis[1]: for this reason personalized medicine plays a crucial role in enhancing diagnostic accuracy and tailoring treatment strategies to individual patients. Recent advances in genomic sequencing technologies allow for an effective approach, and whole exome sequencing analysis is a commonly used technique.

Whole exome sequencing (WES) is a powerful technique used for genetic testing, that allows for the comprehensive sequencing of the exome (1-2% of the entire genome), which represents all the protein-coding regions of an individual's genome. The sequenced DNA is then analyzed to identify variations or mutations: they can be single nucleotide polymorphisms (SNPs), small insertions or deletions (indels), or larger region variants (frame-shifts).

Even though carrying the mutation doesn't always mean developing the disease, these genetic alterations can be inherited and passed on to offspring, creating an inherited genetic disorder. In our study we focus on two main modes of inheritance, which are autosomal dominant (AD) and autosomal recessive (AR). The latter is caused by the presence of two copies of a mutated gene, passed by each parent, while for autosomal dominant diseases, on the other hand, the presence of only one mutated allele is sufficient. The power to detect mutations involved in disease by genome sequencing is enhanced when combined with the ability to discover specific mutations that may have arisen between offspring and parents[2]. For this reason we recruited a cohort of 10 families, where the parents were healthy and the child was suspected of having either an autosomal recessive or dominant disease, to test if our bioinformatics tools are effective at detecting them. We focused on chromosome 16, one of the 23 pairs of human chromosomes, which is approximately 90 million base pairs long and contains around 800 to 900 genes. It plays a crucial role in various biological processes: a mutation on these genes could lead to various rare Mendelian autosomal diseases, like Fanconi Anemia or KBG syndrome. We received whole-exome fastq files of all the 10 parents-child trios, to each associated a case number and type of inheritance of the possible disease, as stated on *Table 1*. All the cases in our study are de novo mutations (DNMs), which rate of occurrence is given by the human intergeneration mutation rate and the haploid genome size. The mutation rate of single-nucleotide variants in humans is approximately 1×10^{-8} mutations per generation, giving rise to 45–60 de novo mutations per genome[3]. Using modern bioinformatics tools, we tried to deduce if the

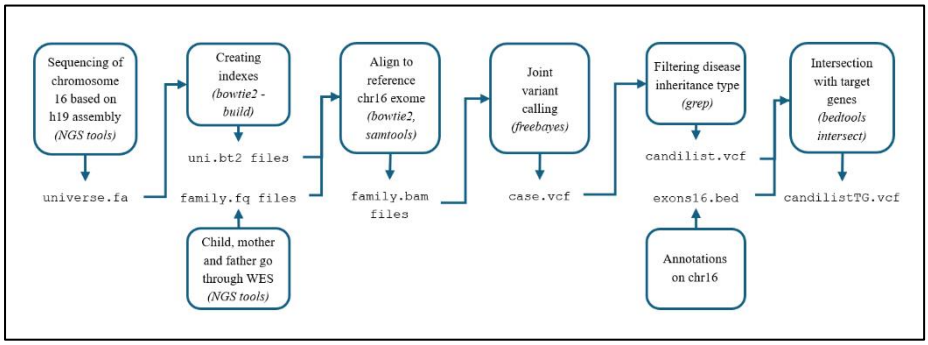


Figure 1: ***Complete Unix pipeline for variant calling.*** The *.fq, .fa, .bt2* and *.bed* files were provided and can be found on the server and path mentioned in the methods section, along with the codes used.

presence or absence of these mutations would lead to rare Mendelian genetic diseases, focusing only on monogenic variants.

1. Methods

Patient analysis was conducted in two stages, using an Unix pipeline and variant effect predictor (VEP) annotation. A schematic visualization of the pipeline can be found on *Figure 1*. All the files provided can be found on the server 159.149.160.7 at the following path: /home/BCG2024_genomics_exam/.

Alignment and quality assessment

We received the *.fq* files of all the families' trios exomes, the human chromosome 16 sequence based on the hg19 assembly in fasta format (*universe.fa* file) and already built indexes based on it (*.bt2* files). We first performed control checks on our raw sequence data (1) to get an impression on whether it had any problems of which we should've been aware before doing any further analysis.

(1) fastqc *.fq.gz

Then, using the *bowtie2* tool, the exome sequencing reads from the families' trios were aligned with the reference genome, producing *.sam files* for each member of the trio. These *.sam* files contained sequence alignments in a standardized text format. Subsequently, employing *Samtools*, the *.sam* files underwent conversion into their binary equivalent, resulting in *.bam* files that were sorted. We piped all of these commands into one, which can be found at (2), to avoid creating intermediate files for a more efficient server space usage. The bold text was changed depending on the case and the family member which *.bam* file needed to be created.

(2) bowtie2 -U **case_familymember**.fq.gz -p 8 -x uni --rg-id '**SC**' --rg "SM:**family member**" | samtools view -Sb | samtools sort -o **case_familymember**.bam

Before proceeding with the rest of variant calling we checked the quality of the alignment, using the *qualimap* tool (3). Since we're focusing on just chromosome 16 and we don't need an overview of the alignment of the whole exome, we added the *.bed* file (exons16Padded_sorted.bed) that was provided to the command, which contains the coordinates of our target regions.

(3) qualimap bamqc -bam **case_familymember**.bam -gff exons16Padded_sorted.bed -outdir **case_familymember**

Both *fastqc* and *bamqc* files can be aggregated using MultiQC (multiqc ./), which allows assessing the consistency and quality of data across all samples: no trends or patterns were found, though, so adjustment of analysis parameters wasn't needed.

Detection of potential variants and VEP analysis

To detect variants from the chromosome 16 exome sequence, we employed the freebayes tool (4) by specifying the following parameters:

- m 20: minimum number of copy observations required to call a variant.
- C 5: minimum confidence value for calling a variant, which is based on the sequencing error probability. 5 is a threshold that can provide sensitivity and specificity, both needed because the novo mutations are rare events and a higher value may not be able to detect them.
- Q 10: minimum quality score for bases considered during variant calling.
- min-coverage 10: minimum number of reads that cover each variant position. A value of 10 reduces the risk of false positives, but also can detect variants with moderate allele frequencies.

(4) freebayes -f universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 **case**_child.bam **case**_mother.bam **case**_father.bam > **case**.vcf

After this analysis, a *.vcf* file was generated that contained all the variants of the three components of the family trios. To filter it by showing only the lines that match the specified genotype pattern, either dominant or recessive, it was necessary to first check the order of the #CHR header line (5). This both ensures the integrity of the file and that the columns are in the expected order (first the child and then the parents): any interpretation of the data before checking for a deviation from the expected format could've caused errors.

(5) grep "#CHR" **case**.vcf

We then created another *.vcf* file that contained the header (6) and only the lines that matched the specified genotype pattern (7). The command (7) would change depending if the disease was autosomal dominant (7a) or recessive (7b).

(6) grep "#" **case**.vcf > candilist**case**.vcf

(7a) grep "0/1.*0/0.*0/0" **case**.vcf >> candilist**case**.vcf

(7b) grep "1/1.*0/1.*0/1" **case**.vcf >> candilist**case**.vcf

TABLE 1. ASSIGNED CASES NUMBERS AND POTENTIAL DISEASES INHERITANCE TYPE

Cases	587	591	625	626	664	668	679	711	744	749
Inheritance type	AD	AR	AD	AR	AD	AD	AD	AD	AD	AD

Assuming the child is the first in order, for recessive autosomal disease (7b) the offspring must have exhibited both recessive genes (1/1), which is hard to obtain by just the novo mutation, so we assumed that both parents were healthy carriers (0/0). In the case of dominant autosomal diseases (7a), however, since the parents were healthy, they must've been genotypically 0/0, and the child must've been heterozygote. We further filtered this file, by keeping the header (8) and intersecting it with the coordinates of only the target regions (9). The bedtools intersect tool allowed us to filter the variants based on their overlap with the exonic regions specified in the *.bed* file, and created a *.vcf* file that contained only the variants that fall within the exonic chromosome 16 regions.

(8) grep "#" **case**.vcf > **case**candilistTG.vcf
(9) bedtools intersect -a candilist**case**.vcf -b exons16Padded_sorted.bed -u >> **case**candilistTG.vcf

This final file was then uploaded on VEP (Variant Effect Predictor) that compared the coordinates found with the h19 assembly of the human exome.

Creation of coverage tracks
We created bedgraph files (*.bg*) of all the members of the family trios (10) for both a graphical representation of the aligned sequences data from the *.bam* files and to better compare them with our results from the previous analysis.

(10) bedtools genomecov -ibam **case_familymember**.bam -bg -trackline -trackopts 'name="**familymember**"' -max 100 > **familymember**Cov.bg

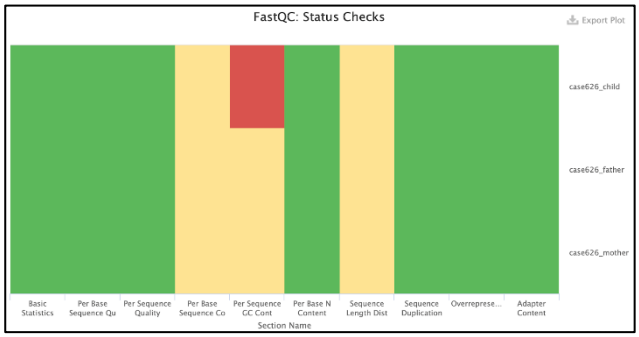
These files were then uploaded on UCSC Genome Browser, along with the *.vcf* file, to visualize genome-wide data, such as sequencing read coverage, an example is shown in *Image 2*.

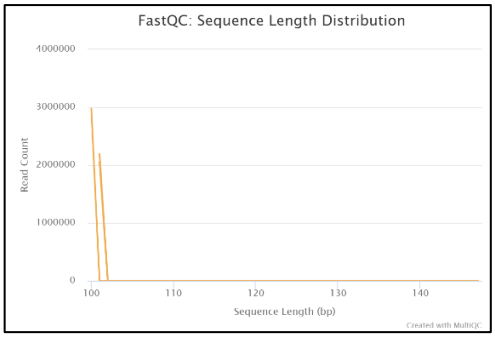
1. Results

All results are shown in *Table 2*. VEP provided information on which genes are affected by the variants and predicted potential functional impact on protein structure. We prioritized variants based on grade of impact (preferring high and moderate impact over low one) and population frequency. With a frequency lower than 10^{-4} we considered the disease a rare one, and so we diagnosed a rare Mandelian genetic disease on 6 patients out of 10.

1. Discussion

Quality assessment
Figure 2 shows the status check that we obtain from the MultiQC for the 626 case. As "Per Base Sequence Quality" and "Per Sequence Quality" suggest, we are dealing with high-quality sequencing data that can be used with confidence for further bioinformatics analyses. When the average quality of the bases (expressed as Phred scores) is high, it's not necessary to filter or modify the data. The high quality is further confirmed by the Basic Statistic measure in which there aren't sequences flagged as poor quality. "Per base N content" shows only normal (green) results too: if a sequencer is unable to make a base call with sufficient confidence, it will normally call an N rather than a conventional base but our plot indicates that the sequencer consistently made base calls with sufficient confidence.
There were no samples found with any adapter contamination > 0.1%, which are short DNA sequences used to facilitate the attachment of DNA fragments to the sequencing platform, so the sequencing process was carried out properly.
The "Per Sequence GC Content" metric indicates a deviation in the GC content distribution compared to the expected one, especially in the sample from the child. The deviation could be due to technical bias introduced during sample preparation or sequencing. All RNA-Seq libraries inherit an intrinsic bias for many reasons, one of which could be mistakes during PCR: overly steep thermoprofile does not leave sufficient time above a critical threshold temperature, causing incomplete denaturation and poor amplification of the GC-rich fraction[4]. We can theorize that the samples required for child's WES weren't prepared correctly and this resulted in not optimal GC content. The Sequence length distribution shows some irregularities, although it is entirely normal to have different read lengths. The





On the left, *Figure 2:* ***Case 626 status check***. The overall quality is acceptable for all samples, although some technical problems may have arisen during sample preparation for the child, which resulted in non optimal GC content.
On the right, *Figure 3:* ***Sequence Length Distribution for case 626***. From left, first the distribution for the child, then for the parents.

TABLE 2. VEP ANALYSIS RESULTS

Case	*Impact*	*Variant*	*Codons*	*Position*	*Phenotype*	*Case*	*Impact*	*Variant*	*Codons*	*Position*	*Phenotype*
587	High	frameshift (deletion, 1 base)	**G**GG/ -GG	**16**:2144194 -2144200	Autosomal dominant polycystic kidney disease	668	High	splice donor variant	/	**16**:89357420-89357420	KBG syndrome
591	High	frameshift (deletion, 4 bases)	**CGGC**G G/ - - - - GG	**16**:5693636 3-56936368	Familial hypokalemia hypomagnesemia	679	High	frameshift (insertion, 1 base)	CAG/ CA**G**G	**16**:89346177-89346184	KBG syndrome
625	Moderate	missense	A**C**C/ A**A**C	**16**:3820629 -3820629	Healthy	711	Moderate	missense	ACC/ A**A**C	**16**:3820629-3820629	Healthy
626	High	stop gained Splice region	**C**AG/ **T**AG	**16**:8986966 9-89869669	Fanconi Anemia	744	Moderate	missense	ACC/ A**A**C	**16**:3820629-3820629	Healthy
664	Moderate	missense	A**C**C/ A**A**C	**16**:3820629 -3820629	Healthy	749	High	frameshift (insertion, 1 base)	CCA/ CC**C**A	**16**:50788330-50788335	Familial cylindromatosis

warning was raised because sequences are not the same length as shown on *Figure 3* (the child has 100 bp long sequences, while the parents' were 101 bp long); the difference, though, is negligible.

Diagnosis discussion

In cases 587, 591, 679 and 749 a frameshift was reported. According to UCSC Genome Browser, a frameshift is a sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three. That is coherent with the prediction of a high impact variant: a frameshift not only varies the amino acid translated from the position where the variation happened, but also changes the whole sequence of the proteindownstream. If an altered protein is biologically important, it's then probably going to cause a disease.

In Case 668 we observed a splice donor variant, which affects the 5'-splice site (also called donor site) at the beginning of an intron. In most cases (98.7%), the exon/intron boundary sequences contain GT at the 5′ end of the intron[5] and a variant could disrupt the recognition of the site by the spliceosome. In this case, the allele C/C which we found on the parents' sequences, became C/T in the child: the variant is located in an intron right after the exon 5 of gene ANKRD11. This, and also the fact that we were provided exome sequences, led us to believe that the variation may have resulted in intron retention, which could cause a defective mRNA and potentially dysfunctional proteins that may cause KBG syndrome.

Case 626 proved itself a rather interesting case, showing a variant with both splice region and stop gained characteristics: this duality can be easily understood by also looking at the UCSC Genome Browser screenshot we provided at *Figure 4*. A splice region variant is a sequence variant in which a change has occurred in the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron. By looking at *Figure 4*, it's easy to notice that the variation affects the third last base pair of an exon of the gene FANCA, located on the antisense or template strand. The bases sequence, which originally was GTC, had become ATC because of the variant: once transcribed into mRNA the sequence would become UAG, a stop codon. This will probably have serious consequences: the translation of the protein would stop at the 8th exon out of 43, effectively cutting the structure much shorter than it was supposed to be. The child (genotype A/A), will in all likelihood suffer from Fanconi Anemia because both the parents were healthy carriers of the affected allele (genotype G/A).

We deemed cases 625, 664, 711 and 744 to be healthy, and our choice was supported by various reasons. They all presented the same missense variant, resulting in a different amino acid sequence but where the length was preserved. In the child, the affected codon changed from ACC to AAC which, in translation, means adding asparagine (N) instead of threonine (T) in the protein sequence. UCSC Genome Browser has a useful feature in this sense: it summarizes some of the amino acids' properties (*Table 3*) that might be useful for interpretation of the data. Since both amino acids' properties are quite similar, this variation might not have significant consequences on the protein structure, although this would need further investigation. An additional proof for this can be found on the high frequency of the existing variant in sub-populations: all of the ones featured on VEP, besides the South Asian one, are higher than our previously set threshold of 10^{-4}, with the maximum found on the African/American population (AF = 0.00322). We couldn't consider it a rare Mendelian disease inducing variant.

TABLE 3. T AND N AMINO ACIDS PROPERTIES

Property	**T**	**N**
Polarity	polar	polar
Acidity	neutral	neutral
Hydropathy	-0.7	-3.5

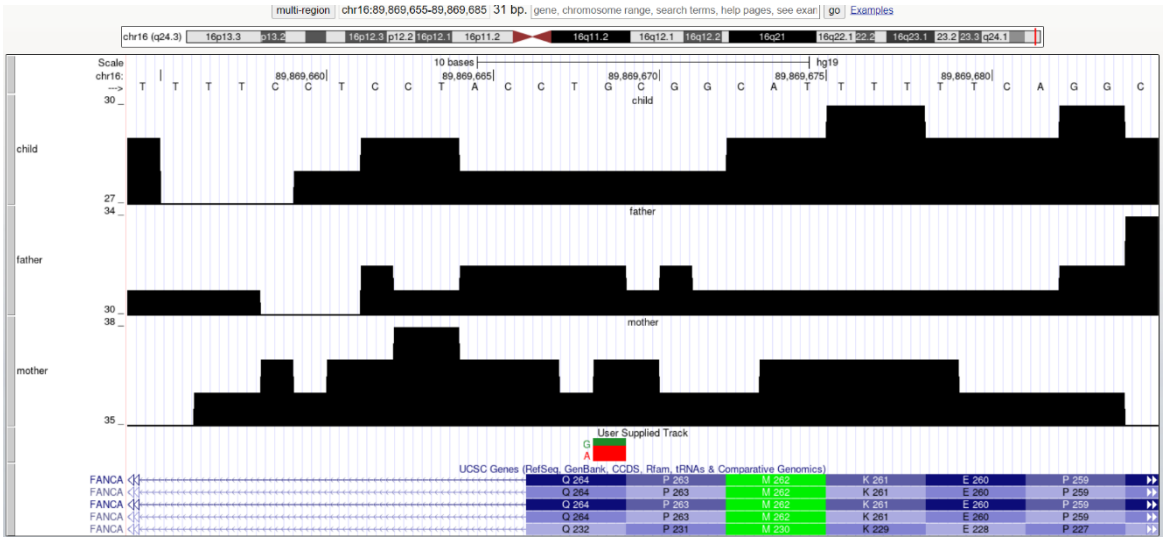


Figure 4: ***UCSC Genome Browser screenshot of case 626.*** Bedgraph files provided coverage tracks while the *.vcf* file shows the stop-gained variant.

1. Conclusion

"Tackling rare diseases through research and to enable all people living with a rare disease to receive an accurate diagnosis, care, and available therapy within 1 yr of coming to medical attention" was the vision of the International Rare Diseases Research Consortium, a global collaborative initiative launched in 2011 by the European Commission and the U.S. National Institutes of Health[6]. Today, a decade later, rare diseases are finally getting the attention they always deserved[7] but had not received for many years. Technological innovations and novel algorithms, but also databases for pathogenic or benign variants, have drastically reduced the time to diagnosis, and thanks to affordable and fast very-long-read sequencing, genetic diagnoses based on the sequence of the complete, truly personal genome have appeared on the horizon[8]. A disease is conventionally defined "rare" when the number of affected subjects is <1:2,000 (i.e., <0.05%) in the European Union and <1:200,000 (i.e., <0.0005%) in the US[9]. Translating this figure into the real world, though, rare diseases, cumulatively, are not as rare as universally perceived. While important steps have been taken to raise awareness of rare diseases and encourage the development of international and national frameworks and policies for people living with rare diseases, many challenges persist. Bioinformatics has proven to have a leading role in patient care: it is a tool that accompanies wet lab techniques starting with sequencing analysis to variant calling and annotation (*Figure 5*)[10]. These new technologies are efficient and created ad hoc for the purpose of diagnosis. Our reports, though, represent only the first step to take for patient assistance and our analysis will prove useful insight for future clinical trials.

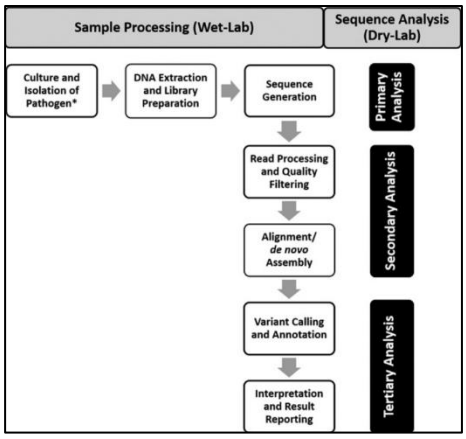


Figure 5: ***General NGS workflow.*** (Amy S. Gargis, Lisa Kalman, and Ira M. Lubin, 2016).

References

Additional materials, including each case's qualimap report and UCSC Genome Browser screenshots, can be found at the following link; https://github.com/moonmiun/Genomics_2024_project.git

References

1. "Rare diseases, common challenges". *Nat Genet* **54**, 215 (2022).
2. Jared C. Roach et al., "Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing". *Science* **328**, 636-639 (2010).
3. Mohiuddin Mohiuddin, R. Frank Kooy, Christopher E. Pearson. "De novo mutations, genetic mosaicism and human disease". *Front. Genet.* **13**:983668 (2022).
4. Aird, D., Ross, M.G., Chen, WS. et al. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." *Genome Biol* **12**, R18. (2011)
5. Abramowicz Anna and Gos Monika. "Splicing mutations in human genetic disorders: examples, detection, and confirmation". *J Appl Genet.* **59**(3): 253–268 (2018).
6. "Who we are". IRDIRC, International Rare Diseases Research Consortium. https://irdirc.org/who-we-are/ (2022).
7. Antonarakis SE, Beckmann JS. "Mendelian disorders deserve more attention." *Nat Rev Genet* **7**: 277–282 (2006).
8. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A. "The complete sequence of a human genome" Science. **376**(6588): 44-53 (2022).
9. Elisa Danese and Giuseppe Lippi. "Rare diseases: the paradox of an emerging challenge". *Ann Transl Med*. **6**(17): 329 (2018)
10. Amy S. Gargis, Lisa Kalman, and Ira M. Lubin. "Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories". J Clin Microbiol. **54**(12): 2857–2865 (2016).