Introduction

The variant calling format (VCF) is a type of file used in bioinformatics research, that specifies in a text format the different gene sequences variations. It describes in the different columns the main details needed to identify a single mutation like the chromosome on which the mutation is called, the position on said gene and the reference base at a given position of a reference. It is also important to keep in mind that variant calling is used in comparison with a reference genome, that it is not extracted from a single person, and that it is not the main human genome, but just a representation.

The variance calling method also consist in the use of public, or private tools needed to analyse the VCF file. Those tools compare the genome that we want to analyse and the reference genome, giving as a result the differences between those. Tools such as Ensembl Variant Effect Predictor permit the user to choose between different filter, thanks to which it is possible to identify the most impactful mutations from a disease point of view. Those filters include algorithms like SIFT and POLYPHEN, needed to point out nonsynonymous and missense mutations, a nonsynonymous variant being a mutation in the DNA that alters the aminoacidic sequence of a protein with two different outcomes: In the case that the single point mutation stops the production of the protein completely, it is considered a nonsense mutation, in this case the mutation changes the original amino acid to a stop codon, ending the protein synthesis sooner than the non-mutated one, while a mutation that just alters the aminoacidic sequence of the protein is considered a missense mutation.

Rare diseases usually occur thanks to both missense and nonsense mutations and can be autosomal recessive or autosomal dominant diseases. Autosomal recessive diseases can be transmitted to a progeny while both parents being healthy, in this case both parents probably consist in a heterozygous couple for this mutation, while autosomal dominant diseases can be described as a de novo mutation, in which the progeny is affected by a disease thanks to a variant, or a mutation, in the germ cell, that can both be in the sperm cell or in the egg cell. The presence of a mutation can causes an interference in different parts of the protein that is being synthesised, it can for example occur in a regulatory region, interfering with the expression mechanisms of that said protein, or it can interfere with the protein structure itself, modifying one of the structures of the protein and possibly override the primary function of the protein, making it biologically less or completely ineffective.

The study of our genome data analysis was possible thanks to a Linux interface, accessible through a terminal. The Linux interface is useful for multiple reason, it can store files easily on a server making them more accessible, and it can guarantee the use of a various number of commands and software, needed to correctly identify a variant.

Scope of analysis

Our work consisted in the analysis of genome data from a family, where both the parents were not affected by a rare disease, while the child was in fact affected by it. The different possible diseases could be transmitted both with an autosomal recessive pattern, or with an autosomal dominant pattern, to find and correctly identify the disease we initially used different methods to approach autosomal recessive and autosomal dominant diseases, after which we used different algorithms needed to identify the most probable and significant mutation to directly call the presence of a variant

Materials and methods

Linux server the main method used to store bioinformatics data; it is also needed to use the programs cited in this section, even though some of them could have a GUI that let the user interact with them

Bowtie2 It is a memory efficient genomic aligner, needed to align relatively long genomes

Samtools Needed to post-process DNA sequences into SAM, BAM and CRAM formats needed to visualize the DNA data

Bedtools Allows to intersect, merge, count and shuffle genomic data from different genomic file format, like BAM, BED and VCF

Freebayes Bayesian genetic variant detector, used to identify SNP, indels and MNP. It uses short-read alignments (BAM) for any number of individuals from a population and a reference genome, it then calculates the most likely combination of genotypes returning a VCF file

VEP Tool used to identify the impact of a variant in the genome, it uses a refence genome to compare the file in the VCF format, and thanks to a wide range of filter it makes possible to identify the disease-causing variant

UCSC Genome Browser Online genome browser, useful to visualise the position of genes, mutations, introns and many other details needed in variant calling

Procedure

To discover the kind of disease the patient had we adopted to different approaches for autosomal recessive mutations and autosomal dominant mutations, even though they are similar for the most part, in term of Unix code.

Aligning

Consisting of the alignment of the sequencies of mother, father and child to the reference genome
example taken from case510

```
bowtie2 -U /home/BCG2022_genomics_exam/case510_child.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SC' --rg "SM:child" | samtools view -Sb | samtools sort -o case510_child.bam
```

```
bowtie2 -U /home/BCG2022_genomics_exam/case510_father.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SF' --rg "SM:father" | samtools view -Sb | samtools sort -o case510_father.bam
```

```
bowtie2 -U /home/BCG2022_genomics_exam/case510_mother.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SM' --rg "SM:mother" | samtools view -Sb | samtools sort -o case510_mother.bam
```

We then obtained 3 bam file containing all the results needed to perform the Variant Calling process

Visualization of the mutation's position using UCSC

Using these pipelines

```
bedtools genomecov -ibam case457_father.bam -bg -trackline -trackopts 'name="father"' -max 100 > father457Cov.bg
```

```
bedtools genomecov -ibam case457_mother.bam -bg -trackline -trackopts 'name="mother"' -max 100 > mother457Cov.bg
```

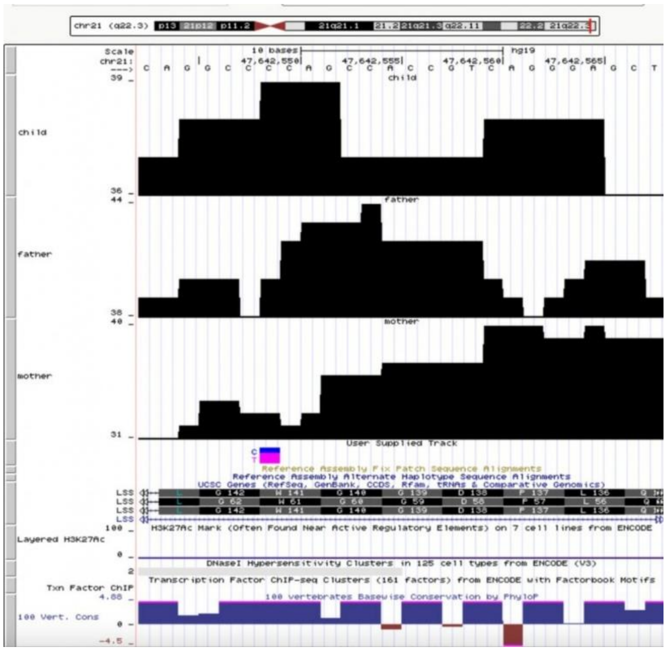
```
bedtools genomecov -ibam case457_father.bam -bg -trackline -trackopts 'name="child"' -max 100 > child457Cov.bg
```

we were able to create the three BG files loaded on UCSC with the vcf file obtained in the precedent step to track the position of the pathogenic mutation.

Results given by UCSC

In the following image we can see the position and the mutation's nature, in this case being a missense one.

The picture also underlines that the variant is placed in the 47648493 position of the chromosome 21.


chr21 (q22.3) p13 21p12 p11.2 21q21.1 21.2 21q21.3 q22.11 22.2 21q22.3
Scale
10 bases
hg19
chr21:
47,642,550
47,642,555
47,642,560
47,642,565
C A G G C C C C A G C C A C C G T C A G G G A G C T
39
child
36
44
father
38
40
mother
31
User Supplied Track
C
T
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
LSS
G 142 W 141 G 140 G 139 D 138 F 137 L 136 Q
G 62 W 61 G 60 G 59 D 58 F 57 L 56 Q
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
2
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
-4.5

Variant Calling

Performing variant calling means to ensure that the differences between reference genome and the family members’ genomic sequences are mutation and not simple errors during the sequencing procedure made il lab.

To do so we used the following pipeline

freebayes -f /home/BCG2022_genomics_exam/universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 case510_mother.bam case510_father.bam case510_child.bam > Trio510.vcf

-m Exclude alignments from analysis if they have a mapping quality less than 20 in this case.

-C Require at least this count of observations supporting an alternate allele within a single individual to evaluate the position, in this case 5.

-Q Count mismatches toward **--**read-mismatch-limit if the base quality of the mismatch is >= 10.

--min-coverage Require at least this coverage to process a site, in this case 10.

Selecting target mutations

Since AR mutations need to have an affected homozygous child and two heterozygous parents, we selected from the VCF file obtained before only the mutations corresponding to this description using this grep

example taken from case510

```
grep -v "#" Trio510.vcf | grep -e  "1/1.*0/1.*0/1" -e  "1/1.*1/2.*1/2" -e  "1/1.*1/2.*0/1" -e
"1/1.*0/1.*1/2" -e  "1/1.*1/3.*0/1" -e  "1/1.*0/1.*1/3" -e  "1/1.*1/3.*1/2" -e  "1/1.*1/2.*1/3" -e
"2/2.*0/2.*0/2" -e  "2/2.*1/2.*1/2" -e  "2/2.*1/2.*0/2" -e  "2/2.*0/2.*1/2" > case510.vcf
```

While, for AD (de novo) mutations, some changes are needed. According to the literature, this kind of mutation happens randomly in the germinal cells of the parents during replication, so parents will not have the disease, while the child will be affected by it, presenting a dominant allele different from the two of the parents

example taken from case 590

```
grep -v "#" Trio590.vcf | grep -e  "0/1.*0/0.*0/0" -e  "0/2.*0/0.*0/0" -e  "1/2.*1/1.*1/1" -e
"1/2.*2/2.*2/2" -e  "1/2.*1/3.*1/1" -e  "1/2.*1/1.*1/3" -e  "0/2.*0/1.*0/1" -e  "0/1.*0/2.*0/2" -e
"1/2.*0/1.*0/1" -e  "1/2.*0/2.*0/2"  > case590.vcf
```

In both of the pipelines we used grep -v to get rid of the lines with the symbol “#” and then we searched for the specific patterns using grep -e, that can be interpreted as a Boolean “or”. We then saved the result as a VCF file

We then completed the results keeping only the variants included in the target genome with

```
bedtools intersect -a candilist590.vcf -b /home/BCG2022_genomics_exam/targetsPad100.bed -u >
590candilistTG.vcf
```

Results and conclusions

After the upload of our vcf files on Vep, we looked for the genomic mutations occurred on coding regions, an allele frequency of 1/10000 and, in case of a missense variant, the value of PolyPhen and SIFT and the pathogenicity of the mutation.

Case 457 - ALOPECIA-MENTAL RETARDATION SYNDROME 4 (rs763705074,ClinVar)

Missense variant

Position: 21:47648493-47648493;

Case 581 - Unverricht-Lundborg syndrome (rs386833443, ClinVar)

Splice region variant synonymous variant

Position 21:45196085-45196085

Case 519 - HOLOCARBOXYLASE SYNTHETASE DEFICIENCY (rs771944310,ClinVar)

🟪 Frameshift variant

Position 21:38308962-38308964

Case 454 - ALOPECIA-MENTAL RETARDATION SYNDROME 4 (rs1569036540,ClinVar)

🟥 Stop gained

Position 21:47822302-47822302

Case 566 - TRANSCOLABAMIN II DEFICIENCY (rs1279321570,ClinVar)

🟥 Stop gained

Position 22:31018965-31018965

Case 461- AMYOTROPHIC LATERAL SCLEROSIS 1 (ENSG00000142168,MIM morbid)

🟨 Missense variant

Position 21:33036142-33036142

Case 545- MICROCEPHALIC OSTEODYSPLASTIC PRIMORDIAL DWARFISM TYPE II

(ENSG00000160299,DDG2P& MIM morbid)

🟥 Stop gained

Position 21:47822302-47822302

Case 588 - ZTTK SYNDROME (rs1555899560,ClinVar)

🟥 Stop gained

Position: 21:34927559-34927559

Case 510- HOLOCARBOXYLASE SYNTHETASE DEFICIENCY (rs119103229,ClinVar)

🟨 Missense variant 🟧 Splice region variant

Position 21:38137471-38137471

Case 590 - ZTTK SYNDROME (rs886039773,ClinVar)

🟪 Frameshift variant

Position 21:34927287-34927292