M.Sc Bioinformatics for computational genomics
A.Y. 2021-2022

Course: Genomics and transciptomics
Module: Genomics

Genomics project report

Giuseppe Barranco

Contents

1 Introduction

Trios exome sequencing for rare disease is one of the most innovative approach to identify causal mutations for inherited disease. This can be useful to recognize variants inherited from the parents causing recessive or dominant disease. Additionally it is possible to detect de-novo variants present in the offspring but not in the parents.

2 Aim of the project

The goal of the project is to identify a variant causing disease the child is affected by using the tools we discussed during classes.

3 Materials and Methods

3.1 Data

Each trio's files, BWT of reference genome and target regions were given by the professor. The trios is composed by a father and a mother that are either healty or carrier, depending on the inheritance pattern, and a child affected by a rare Mendelian autosomal disease.

3.2 Workflow

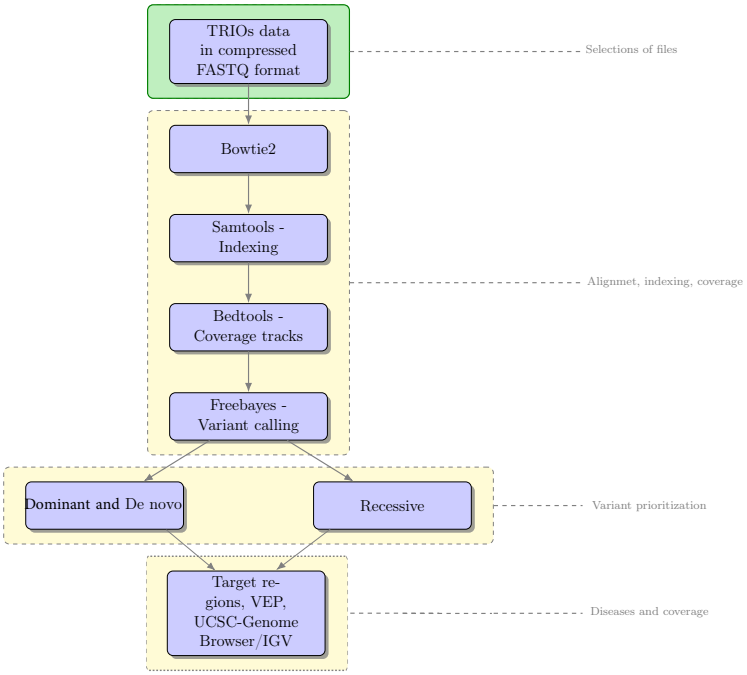


Figure 1: Workflow for the whole analysis

3.3 Prioritization strategy

Regarding the prioritization strategy, many methods have been defined in order to filter out specific variants. Since we are dealing with rare Mendelian diseases one of the straightforward approach is the reconstruction of variant segregation in the family. We have defined in our script a two different recursive selections depending on the model of inheritance of the disease (dominant/DNMs, recessive). For the DNMs, through the *-awk* command, given the fact that all the files previously generated by *freebayes* show the trio's components in the same order (Father, Mother, Child), we select the genotypes of mothers and fathers that are either homozygous (for the DNMs: *0/0*) or heterozygous (*0/1, 0/2)*. Since we want to obtain a child with heterozygous genotype only, we filter the corresponding column with (*0/1, 0/2)*.
For the recessive mutations, we apply the same criteria by setting the parameters in a different way, in order to select the genotypes of mothers and fathers that are heterozygous (*0/1, 0/2)*. Since we want to obtain a child with homozygous genotype only, we filter the corresponding column with (*1/1, 2/2)*.

We use the on-line platform VEP (Variant effect predictor) to figure out which of the variants extracted in the previous manner are more likely the most eligible according to the related filters taken into account in the analysis:

- Associated phenotypes is defined
- Impact is not LOW and not MODIFIER
- Clinal significance
- gnomAD
- SIFT and PolyPhen

Using RefSeq transcripts as database, we started setting *"Associated phenotypes"*, since we are looking for a variant that is associated to a defined disease. Then we consider the *"Impact"* that should be high or at least moderate. Till this point some variants could still be defined as "benign" so we set *"Clinal significance"* to *"pathogenic"*. Moreover, given the fact that we analyse rare Mendelian disease, we do consider the frequencies provided by The Genome Aggregation Database

[1]DNMs: De novo mutations

(gnomAD) (*freq. disease* $< 10^{-5}$*)*. Finally we set two parameters, even if sometimes there were no associated score, PolyPhen and SIFT scores. They use the same range, 0.0 to 1.0, but with opposite meanings. A variant with a PolyPhen score >0.6 is predicted to be either *"possibly damaging"* or *"damaging"* or *"deleterious"* . A variant with a SIFT score <= 0.2 is predicted to be either *"potentially damaging"* or *"possibly damaging"* or *"deleterious"*.

3.4 Executable script

We created an executable file called "project.sh" In order to make it executable we run this command in the same directory where the file is:

```
chmod u+x project.sh
```

```
./project.sh
```

This is the content:

```
#!/bin/bash

echo "START"

function align () { local var1="${1%%.*}"; var1="${var1#/home/BCG2022_genomics_exam/}";
if [[ $1 == *father* ]];
then
bowtie2 -U $1 -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SF' --rg "SM:father" | samtools view -Sb | samtools sort -o $var1.bam; echo "$1 aligned"; elif
[[ $1 == *mother* ]]; then bowtie2 -U $1 -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SM' --rg "SM:mother" | samtools view -Sb | samtools sort -o $var1.bam; echo "$1 aligned";
else
bowtie2 -U $1 -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SC' --rg "SM:child" | samtools view -Sb | samtools sort -o $var1.bam; echo "$1 aligned";
fi; }

export -f align

echo "Write the cases you want to analyse like this case###, each  separetd by a space, then press enter:"; read -a names; prova=${names[@]} ; tempo=$(date +%Y-%m-%d--%H:%M:%S ); prov="$prova $tempo"; mkdir "${prov}"
cd "${prov}";
for i in "${names[@]}"; do
find /home/BCG2022_genomics_exam -type f -name "${i}*" -exec bash -c "align \"{}\"" \;;
done

for file in *.bam;
do
if [[ $file == *father.bam* ]];
then
bedtools genomecov -ibam $file -bg -trackline -trackopts 'name="father"' -max 100 > "${file%%.*}Cov".bg; echo "Coverage of track of "${file%%.*}" generated";
elif [[ $file == *mother.bam* ]];
then
bedtools genomecov -ibam $file -bg -trackline -trackopts 'name="mother"' -max 100 > "${file%%.*}Cov".bg; echo "Coverage of track of "${file%%.*}" generated";
else
```

```
bedtools genomecov -ibam $file -bg -trackline -trackopts 'name="child"' -max 100 > "${file%%.*}Cov".bg;
    echo "Coverage of track of "${file%%.*}" generated";
fi;
done

mkdir Coverage; mv *.bg Coverage/

for file in *.bam; do samtools index $file; echo ${file} indexed; done

mkdir Indexed; mv *.bai Indexed/

printf '%s\0' *.bam | xargs -0 -n 3 sh -c 'echo "Creating vcf file for "${1%_*}; freebayes -f /home/
    BCG2022_genomics_exam/universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10 "$3" "$1" "$2" >"${1%_*}.
    vcf"; echo "Vcf file for "${1%_*} created' sh

echo "Insert cases that are dominant, if none just press enter:"; read -a names; for i in "${names[@]}";
     do find . -type f -name "${i}.vcf" -exec sh -c 'x="{}"; mv "${x##*/}" Dominant"${x##*/}"' \; ;
    done

for file in *.vcf;
do
if [[ $file == *Dominant* ]];
then
awk -F'\t' ' /^#/     { print > ("candilist"FILENAME) } ($10 ~/^0\/0/ || $10 ~/^0\/1/ || $10 ~/^0\/2/) &&
     ($11 ~/^0\/0/ || $11 ~/^0\/1/ || $11 ~/^0\/2/) && ($12 ~/^0\/1/ || $12 ~/^0\/2/){ print >>("
    candilist"FILENAME) }' "$file"; echo ${file}" filtered" ;
else
awk -F'\t' '/^#/     { print > ("candilistRecessive"FILENAME) } ($10 ~/^0\/1/ || $10 ~/^0\/2/ || $10
    ~/^1\/2/ ) && ($11 ~/^0\/1/ || $11 ~/^0\/2/ || $11 ~/^1\/2/ ) && ($12 ~/^1\/1/ || $12 ~/^2\/2/){
    print >>("candilistRecessive"FILENAME) }' "$file";     echo ${file} "filtered";
fi;
done

printf '%s\0' *candilist* | xargs -0 -n 1 sh -c 'bedtools intersect -a $1 -b /home/BCG2022_genomics_exam
    /targetsPad100.bed -u > "TG${1%.*}.vcf"' sh


mkdir FilesforVEP; mv TG* FilesforVEP/
mkdir Notfilteredcandilist; mv cand* Notfilteredcandilist/
mkdir Freebayesout; mv *.vcf Freebayesout/
mkdir Bowtie2out; mv *.bam Bowtie2out/
echo "end"
```

4 Results

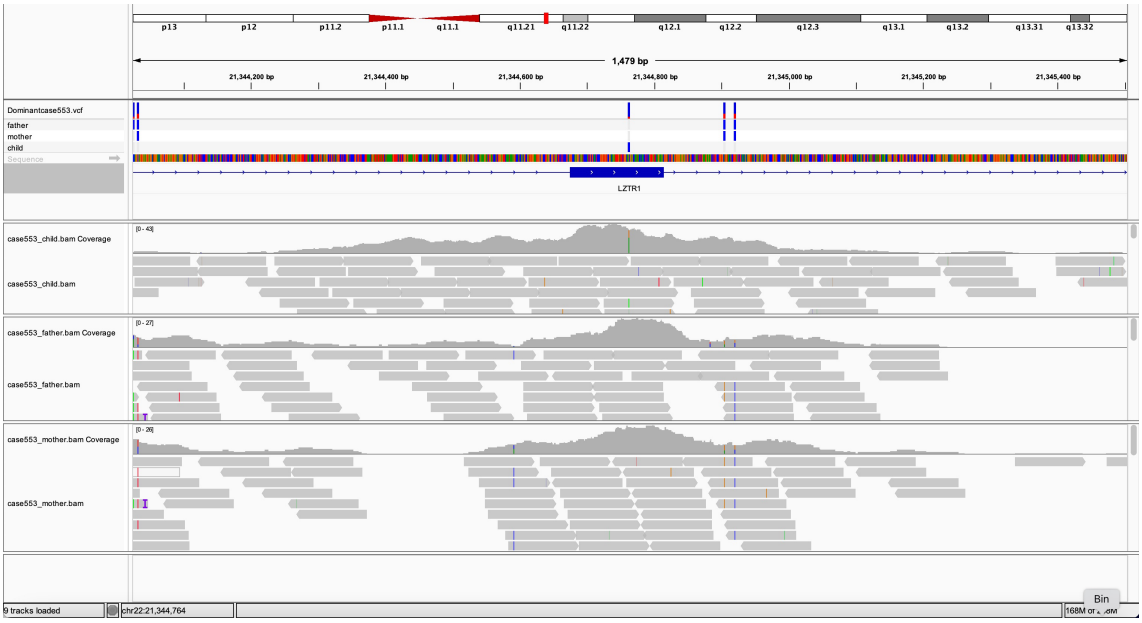
We provided a table with the all diagnosed diseases for each case, attaching the links for the variants and their consequences.

Table 1: Variants and diagnosed diseases

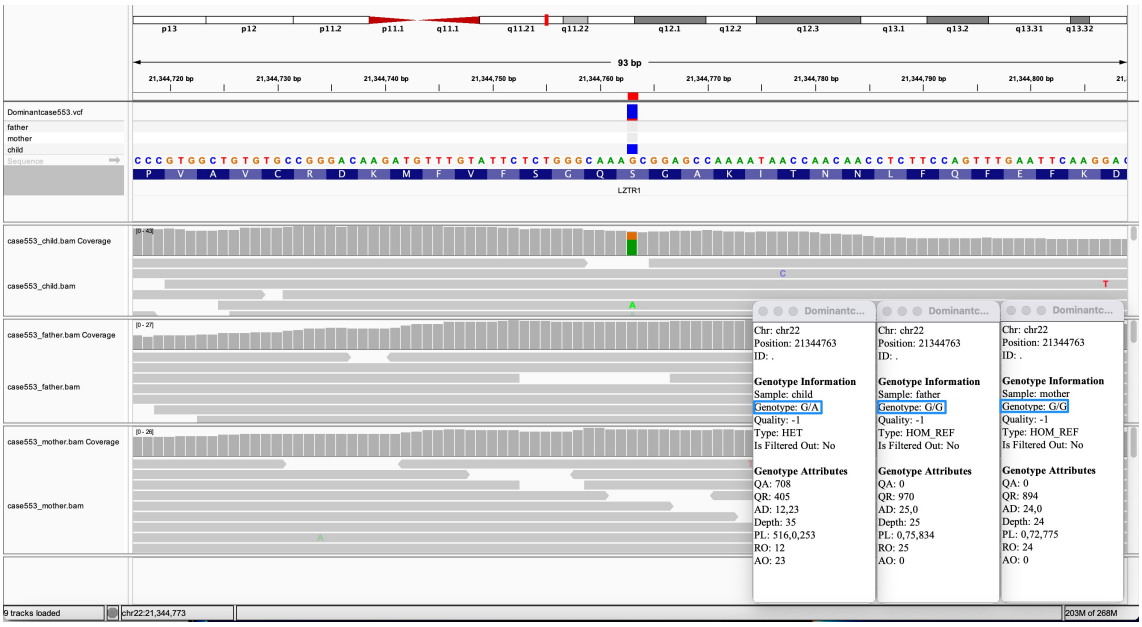
Analyzed case	Pattern of inheritance	Disease	Variant annotation	Consequence
456	AR	ALOPECIA-MENTAL RETARDATION SYNDROME 4	rs754230211	missense_variant
462	AR	AMYOTROPHIC LATERAL SCLEROSIS 1	rs121912433	missense_variant
466	AR	AMYOTROPHIC LATERAL SCLEROSIS 1	rs121912437	missense_variant
475	AR	AMYOTROPHIC LATERAL SCLEROSIS 1	rs121912448	missense_variant
476	AR	AMYOTROPHIC LATERAL SCLEROSIS 1	rs121912449	missense_variant
512	AR	HOLOCARBOXYLASE SYNTHETASE DEFICIENCY	rs148324626	stop_gained
553	AD	NOONAN SYNDROME 10	rs797045166	missense_variant
556	AD	Noonan syndrome 2	rs1555928249	frameshift_variant
566	AR	TRANSCOBALAMIN II DEFICIENCY	rs1279321570	stop_gained
572	AR	Unverricht-Lundborg disease	rs74315443	missense_variant

4.1 Disease causing variant in the UCSC genome browser

Here are reported two **screenshots** for both case n°553 and case n°512. We can appreciate for both dominant and recessive cases the coverage, that indicates the average number of reads that cover a specific target region. In this case the regions are related to the mutated genes that cause the pathology. For all the components of both trios, there is an high coverage of the target regions that should provide a reliable variant call.

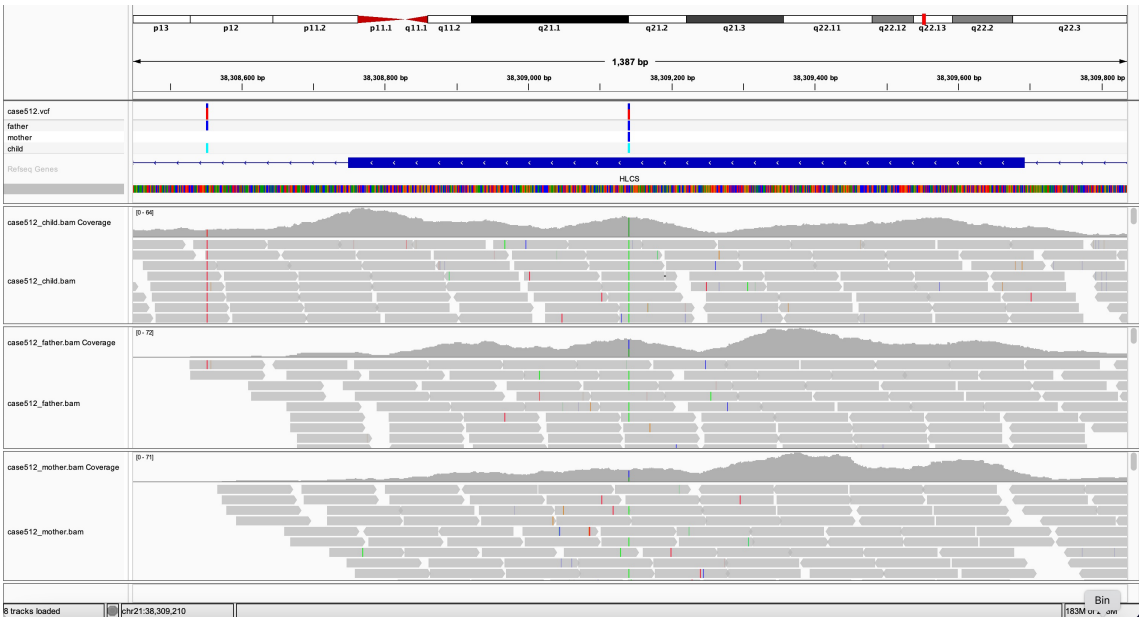


(a)



(b)

Figure 2: (a)Coverage of case n°553 - (b)Zoomed coverage of case n°553 to show the genotypes



(a)

(b)

Figure 3: (a)Coverage of case n°512 - (b)Zoomed coverage of case n°512 to show the genotypes

About the case n° 512, HCSD is a autosomal recessive disease caused by mutations in the HLCS gene (21q22.13) resulting in reduced holocarboxylase synthetase (HCS) activity. This is a stop gained mutations that leads to a premature termination codon. (C>A)

About the case n° 553, NS is a autosomal dominant disease caused by mutations in LZTR1(22q11.21), and less commonly in other genes associated with the RAS/MAPK signaling pathway. This is a missense variant. (G>A)