GENOMICS EXAM REPORT

Students: **Gabriele Marchi, Manenti Carlo**
date: 06 | 05 | 2022

Abstract:

Precision medicine is becoming gradually more accessible due to the emergence of new tools and technologies for variant calling. Thanks to Next Generation Sequencing for exome sequencing and up to date software tools for processing genomic data, variant calling became more reliable than ever. In this report we have performed variant calling on 10 trios (each composed of father, mother and an ill child) for which we had the exome of chr21 and 22. Specific variant prioritization approaches were applied depending on the case. In the end we propose a variant associated to a defined pathology for 8 of the 10 cases. While for the last two cases we strongly suggest that, for our data, the children are perfectly healthy.

Methods:

case 452, describing a typical pipeline

The raw sequences of an exome of chr21 and 22 for a father, a mother and the child are given in FASTQ format. Those sequences are aligned to the reference human genome (GRCh37) using the aligner Bowtie (in particular bowtie2).

Bowtie is an efficient software that aligns short DNA sequences to a human reference genome. Using the command bowtie2 we specify -U to pass the FASTQ file, -p 8 to speed up the process using 8 cores, also we pass the name of the file with the indexes of the reference genome with -x. Furthermore —rg and —rg-id are used to set the SM (Sample) field in the bam files to the corresponding individual (father, mother or child) for later use.

```
bowtie2 -U /home/BCG2022_genomics_exam/case452_father.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SF' --rg "SM:father" | samtools view -Sb | samtools sort -o case452_father.bam
bowtie2 -U /home/BCG2022_genomics_exam/case452_child.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SC' --rg "SM:child" | samtools view -Sb | samtools sort -o case452_child.bam
bowtie2 -U /home/BCG2022_genomics_exam/case452_mother.fq.gz -p 8 -x /home/BCG2022_genomics_exam/uni --rg-id 'SM' --rg "SM:mother" | samtools view -Sb | samtools sort -o case452_mother.bam
```

Then we use Samtools package first to convert the SAM file into a BAM file with -Sb and later to sort alignment by left most coordinate. Last we save the results in a BAM file using -o. BAM files are the standard for using and sharing NGS data given their compact file size and the indexed-access capabilities.

To access any coordinate positions of the BAM file, we must first compute the indexes using the index command of samtools.

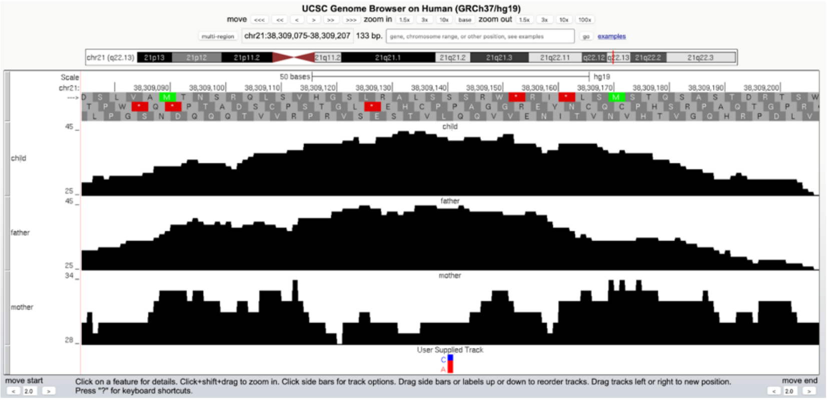
```
samtools index case452_child.bam
samtools index case452_mother.bam
samtools index case452_father.bam
```

UCSC

At this point we generate coverage tracks to use later in the University of California, Santa Cruz (UCSC) Genome Browser. We use the genomecov command to compute the histogram of coverage for the BAM file, also we had to specify -ibam because the input is a BAM file, while -bg gives a .bg file format as output. -trackline and -trackpots, are used to add a UCSC Genome Browser track line definition as the first line of the file and the name of the track, respectively. Then we combined all the positions with a depth equal or greater than 100 into a single bin(column) of the histogram using -max 100.

```
bedtools genomecov -ibam case452_father.bam -bg -trackline -trackopts 'name="father"' -max 100 > fatherCov.bg
bedtools genomecov -ibam case452_mother.bam -bg -trackline -trackopts 'name="mother"' -max 100 > motherCov.bg
bedtools genomecov -ibam case452_child.bam -bg -trackline -trackopts 'name="child"' -max 100 > childCov.bg
```

UCSC case512 (AR)



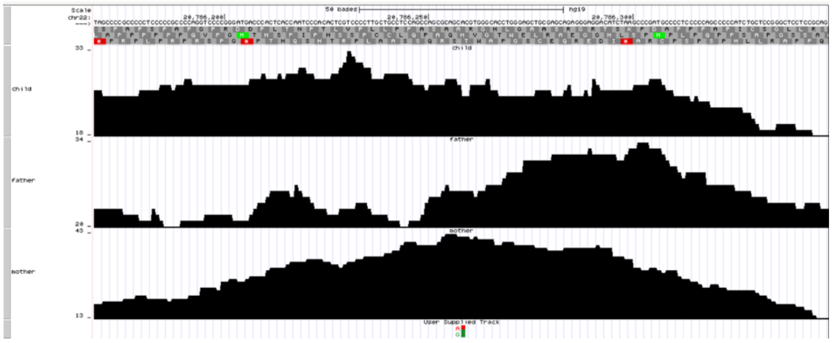
UCSC Genome Browser on Human (GRCh37/hg19)

chr21:38,309,075-38,309,207 133 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
Press "?" for keyboard shortcuts.

Using the UCSC Genome Browser we can visualize the coverage tracks for each individual and the associated disease-causing variant.

UCSC case582 (AR)



Variant calling

Plenty of variant calling tools for NGS have been developed in the past 10 years. Even though it is possible to perform variant calling on individuals and it can be desirable for laboratories processing a large number of samples, in our case we prefer joint variant calling. With joint variant calling we consider simultaneously the sample of the father, the mother, and the child. This approach offers several advantages for our analysis and probably the most important one is the reduction of variant representation differences which might be problematic.

To perform variant calling we use Freebayes, which is a Bayesian haplotype-based variant calling tool. Freebayes uses the Bayesian posterior probability to compute the conditional probability of having a SNP or insertion/deletion (indel), given the sequences provided. While haplotype-based means that the variants are called based on the literal sequences of read aligned to a specific target.

```
freebayes -f /home/BCG2022_genomics_exam/universe.fasta -m 20 -C 5 -Q 10 --min-coverage 10
case452_mother.bam case452_child.bam case452_father.bam > case452.vcf
```

We set the minimum mapping quality at 20 using -m 20 and the minimum alternative allele count at 5, with -C 5. Additionally we set the minimum coverage quality at 10 with --min-coverage 10 and the minimum base mismatch at 10 with -Q 10.

Those parameters may vary depending on the type of analysis that we want to achieve. The output is stored in Variant Calling File (VCF). Because the order of the files given by Freebayes is unpredictable we must check the header of the VCF file and annotate the order of the files.

```
grep "^#CHR" case452.vcf
```

Variant selection:

At this point we want to select only the variants with an appropriate pattern of inheritance using the command grep. Because we will need to use the command intersect of bedtools we must keep the header of the VCF file and store it in a new VCF file called candilist (candidate list).

```
grep "#" case452.vcf > candilist452.vcf
```

Thereafter we select variants with an appropriate inheritance pattern and append them to the candilist. First of all we counted how many homozygous and heterozygous variant calls are present in each individual of the family trio.

```
grep -v '#' case452.vcf | cut -f10 | cut -d ':' -f1 | sort | uniq -c (father)
grep -v '#' case452.vcf | cut -f11 | cut -d ':' -f1 | sort | uniq -c (mother)
grep -v '#' case452.vcf | cut -f12 | cut -d ':' -f1 | sort | uniq -c (child)
```

After assessing that the combination for a homozygous child are only 1/1 and 2/2, than we listed all possible combinations given the variants present in each parents. (note: in all the cases the mother presented only one variant 1/3, that we still considered in our analysis). For the purpose of the exam we did not consider the possibility that a parent may be affected by the pathology of the child.

```
ex: grep "0/1.*0/1.*1/1" case452.vcf >> candilist452.vcf
```

note: all the grep for the AR and AD cases are in a .sh file attached to the report.

Last but not least we used the command intersect of bedtools to keep only the variants included in the exons regions and discard all the possible variants found off target.

```
bedtools intersect -a candilist452.vcf -b /home/BCG2022_genomics_exam/targetsPad100.bed -u > 452candilistTG.vcf
```

-a and -b refer to the files used to perform the intersection, while -u allow to write the original entry found in the first file only once if any overlaps are found in the second file.

The output is stored in the candilistTG.vcf. After downloading the candilistTG.vcf using FileZilla, we performed variant prioritisation using Variant Effect Predictor (VEP).

Variant Prioritisation:

VEP enables us to infer the effect of the genetic variants previously selected. For all the analysis with VEP we used the GRCh37genome assembly, the RefSeq transcript database because being manually curated is more reliable than the ENSEMBL one. Also we selected the gnomAD and the 1000 Genome global minor allele frequency as frequency data for the called variants. Furthermore we used phenotypes as additional annotations and SIFT, PolyPhen and LoFtool as predictive scores to infer if a variant may be damaging or possibly damaging.

Standard filters refers to: Impact is **NOT** low, Impact is **NOT** modifier and Phenotype is defined. We did not set a threshold for the impact scores (SIFT, PolyPhen and LoFtool) as well as the AF and gnomAD AF because of a side effect of VEP. If we selected a threshold and those values were missing, VEP would automatically discard the variant. Of course this is unwanted behavior. For all the variants we also checked for a gnomAD AF lower than 10^-5 or missing (which is even lower) and a Clinical Significance that is pathogenic or not provided.

Our analysis:

To perform our analysis we relayed on nano, a user-friendly command line text editor. We used two different scripts, one till variants selection and a dedicated script for the grep. Luckily the order of the files for all cases were father, mother and last child. This facilitated the analysis because it enabled us to write only a universal script for all the AR cases.

Results:

TRIO	Variant 1	Disease	Standard VP	Impact scores
490 AR	rs76731700	AMYOTROPHIC_LATERAL_SCLEROSIS_1	YES	LoFtool confirmation (probably damaging) CAD PHERED of 12.86 SIFT and PolyPhen missing
517 AR	rs753887925	HOLOCARBOXYLASE_SYNTHETASE_DEFICIENCY	LOW IMPACT	LoFtool confirmation (probably damaging) SIFT and PolyPhen missing
526 AR	rs786200924	Infantile_cerebellar-retinal_degeneration	YES	PolyPhen and LoFtool possibly damaging. SIFT deleterious low confidence.
582 AR	rs587777657	Van_den_Ende-Gupta_syndrome	YES	SIFT and PolyPhen confirmation (deleterious and probably damaging) LoFtool missing
584 AR	rs387907086	Van_den_Ende-Gupta_syndrome	YES	SIFT and PolyPhen confirmation (deleterious and probably damaging) LoFtool missing

487 AR	rs1482760341	AMYOTROPHIC_LATERAL_SCLEROSIS_1	YES	SIFT and PolyPhen confirmation (deleterious and probably damaging)
493 AR		Empty	Special Case	
495 AR	rs119469015	GLUTAMATE_FORMIMINOTRANSFERASE_DEFICIENCY	YES	LoFtool and PolyPhen confirmation (deleterious and probably damaging)
512 AR	rs148324626	HOLOCARBOXYLASE_SYNTHETASE_DEFICIENCY	YES	LoFtool (probably damaging) and CAD PHERED (>>20) confirmation. SIFT and PolyPhen missing (so we relayed also on the CAD PHERED score)
593 AD		Empty	Special Case	

Case 493(AR) and 593(AD):

Those two cases were quite peculiar because for both cases we propose that the child is perfectly healthy, breaking the rule of the exam for which all children suffer from a rare pathological disease.

First of all we performed our analysis as for all the previous cases. The only difference being that case 593 is Autosomal Dominant, so we select only the inheritance patterns compatible with a De Novo mutation.

Using standard filters while using VEP lead us to no results, so we removed them, because in some cases the difficulties in variant prioritisation were due to the use of filters. Still without filters we were left with no results. So we started to examine the VCF without any type of variants selection (so with no grep performed). Also in this extreme case all the variants that we found were not compatible with the inheritance pattern of the disease or presented conflictual impact scores and very high frequencies in both gnomAD and AF.

This led us to believe that the child is perfectly healthy in both of this cases.

Conclusions:

The development of NGS and efficient tools for variant calling paved the way to precision medicine. In our report we analyzed ten trios exome of chr21 and 22 using up to date tools for variant calling. While in 8 of the 10 cases we found a variant strictly linked to a pathology, for the last two we were not able to identify a pathology with a plausible pattern of inheritance and which is considered of pathological interest. Hereby we strongly suggest that these two cases present healthy individuals, for what we can assess.