de Sury Riccardo Oliveto Gabriele

Genomics_Report_2024

Introduction

The sound understanding of a bioinformatics pipeline, especially when dealing with the diagnostic context, plays an important role in newer techniques for analysis of genomics data. Being able to produce all the data needed to perform such tasks in a reproducible way by using proper commands and tools, has been the main objective of this Genomics course.

We applied such acquired skills to excecute by ourselves an analisys of genomics data, in order to perform a variant calling and prioritization procedure and the following diagnoses on 10 families, each composed of mother, father and child, with the objective of identifying possible mendelian diseases in the latter one. Since the context is Mendelian Diseases, only Autosomic Dominant and Autosomic Recessive models have been considered.

Each case has been analyzed using commands and approaches that have been taught in class, by means of a Bash script written by us to completely automate the whole analysis and produce complete, consistent and reproducible results, throughout the analysis of all the families.

Methods

The starting point for our Genomic Data analysis for Variant Calling are FASTQ files, coming from an exome sequencing experiment in Human. One FASTQ file per member of the family has been provided. Only Chromosome 16 will be considered for this project: its FASTA sequence and indexing have both been provided too. All of the pipeline and the corresponding results are related to the hg19 genome assembly.

Alignment & Indexing

The first step of our pipeline consisted in the alignment of the sequencing reads to the reference genome. The command we used is the following one:

```
bowtie2 --quiet -U $fq_element -p 8 -x $index --rg-id "$family_role" --rg "SM:$family_role" | samtools view -Sb | samtools sort -o ./$sample_name.bam
```

The bowtie2 algorithm has been used in order to perform the alignment task. The command takes as input:

- The FASTQ files containing all the reads from the sequencing experiment (i.e. `-U $fq_element`)
- The genome index (only chromosome 16 in this case) (i.e. `-x $index`)

In order to set a customized SM field in the output file (SAM or BAM), we used `--rg-id "$family_role" --rg "SM:$family_role`. This will become very useful when performing the variant calling step using the freebayes tool.

The command has been concatenated in order to directly retrieve a sorted BAM file as output, which was then indexed, generating the .bam.bai files. In order to do that, the following command has been used:

```
samtools index ./$sample_name.bam
```

Quality control

A quality control report for each of the families has been created, both for assessing the quality of the reads and the quality of the alignment using respectively FASTQC and BAMQC tools. In order to obtain a single report that summarizes both of them, MULTIQC tool has been employed. The commands we used are the following ones:

```
fastqc -q -o $target_dir/$family_number/QualityCheck$family_number $fq_element
```

- Taking a fastq file as input (`$fq_element`)

```
qualimap bamqc -bam $bam_element -gff $bed_file -outdir
$target_dir/$family_number/QualityCheck$family_number/$actual_name &>
$actual_name"_output"
```

- Taking a sorted bam file as input (`-bam $bam_element`)
- By giving as input also the target regions BED file (`-gff $bed_file`), we are focusing our quality check only on those regions of interest (the exons)

```
multiqc ./
```

Variant Calling & Variant Prioritization

The process of variant calling has been performed using the freebayes command:

```
freebayes -f $reference_genome -m 20 -C 5 -Q 10 --min-coverage 10
./"case"$family_number"_child.bam" ./"case"$family_number"_father.bam"
./"case"$family_number"_mother.bam" >
$target_dir/$family_number/VariantCall$family_number/"case"$family_number".vcf"
```

The command takes as input several parameters and options:

- The BAM files that were created in the alignment step
- The reference genome (`-f $reference_genome`)
- The minimum accepted mapping quality (`-m 20`)
- The minimum number of observations supporting an alternate allele within a single individual in order to evaluate the position (`-C 5`)
- (`-Q 10`) counts mismatches if the base quality of the mismatch is >= Q.
- The minimum accepted coverage to process a site (`--min-coverage 10`)

After having obtained the raw VCF file, several steps have been added in order to refine it before the analysis on the Variant Effect Predictor (VEP).

First of all the raw VCF file has been filtered in order to retain only those variants displaying a compatible genotype, according to the right disease model. Two different approaches have been applied in order to perform such task:

```
case $disease_type in
AR) grep -E "0/1.*0/1.*1/1" "case"$family_number.vcf >> ref$family_number.vcf;;
AD) grep -E "0/0.*0/0.*0/1" "case"$family_number.vcf >> ref$family_number.vcf;;
esac
```

The general idea is to consider those genotypes in which parents do not present the disease associated with a particular variant, but the child does. Even though many other compatible genotypes could have been considered, only the most probable ones have been taken into account. This is also helpful because, by doing so, we consider a lower number of variants, making the prioritization step easier.

For **Autosomic Recessive** (`AR`) diseases, we are only considering those cases in which parents are heterozygous for the variant and the child is homozygous, presenting the variant on both alleles, making him a potential candidate for our analysis.

For **Autosomic Dominant** (**AD**) diseases, we are only considering *de novo mutations*, in which parents are homozygous for the reference alleles, and the child independently develops only one variant, which is the most probable outcome for *de novo mutations*.

Please note that, even though multiple alternative alleles could have been taken into consideration, we decided to avoid considering that scenario for simplicity.

The filtered VCF file has then been intersected with the regions of interest (the BED target file containing the exons) in order to only consider those variants that have been called in the right regions. To perform such task we used:

```
bedtools intersect -a ref$family_number.vcf -b $bed_file -u >>
ref_int$family_number.vcf
```

This command takes as input:

- A VCF file, in our case the filtered one (`-a ref$family_number.vcf`)
- The target BED file (`-b $bed_file`)
- Using the `-u` option we are reporting the original A entry once, if any overlap is found in B

Lastly, even though it was not needed, since a Bash script was used, we decided to alphabetically sort the columns of the family members. To do that, these commands have been used:

```
bcftools query -l ref_int$family_number.vcf | sort > samples$family_number.txt
bcftools view -S samples$family_number.txt ref_int$family_number.vcf >
ref_int$family_number.sorted.vcf
```

The final output of the Variant Calling module of the Bash script is a clean VCF file, ready to be loaded on VEP. As stated before, the hg19 version of the human genome has been used for this project.

On the VEP interface each case has been excecuted, by always choosing the same options:

- **RefSeq transcripts** database
- **1000 genomes** AF and **gnomAD exomes** AF
- **Phenotypes** data
- **SIFT**, **PolyPhen** and **CADD** as variant deleteriousness predictions

The VEP output has then been imported into Excel, for a better tabular structure, and the following filters have been applied:

- **Associated Phenotypes must be DEFINED**
- All the **AFs**, even the gnomAD subpopulations, **must be as rare as or even rarer than the disease** we are looking for (in our case we consider 10 ^ -4)
- The majority of **deleteriousness predictions must be concordant** (**SIFT <= 0.2**), (**PolyPhen > 0.6**), (**CADD > 20**)
- The **impact** of the variant on the gene structure **must not be MODIFIER or LOW**

Other aspects taken into consideration are that the **clinical significance** field **must not be BENIGN**, that the **alleles** taken into consideration **must not be repeated regions** (errors are more likely to occurr there) and the **consequence of the variant should not be** a **synonimous mutation**.

Variant Visualization

Bedgraph files showing the coverage track for each family member have been created using the following command:

```
bedtools genomecov -ibam $bam_element -bg -trackline -trackopts "name=$name"
-max 100 > "$name$family_number"Cov.bg
```

- Taking a bam file as input (`-ibam $bam_element`)
- By setting the option `-max 100` we are combining all positions with depth >= the specified value as a single bin in the histogram

The resulting bedgraph file (`-bg`), as well as the VCF file have then been uploaded to the Custom Track section of the UCSC Genome Browser, in order to properly visualize and more deeply characterize the variants that were spotted after the Variant Prioritization step.

Results

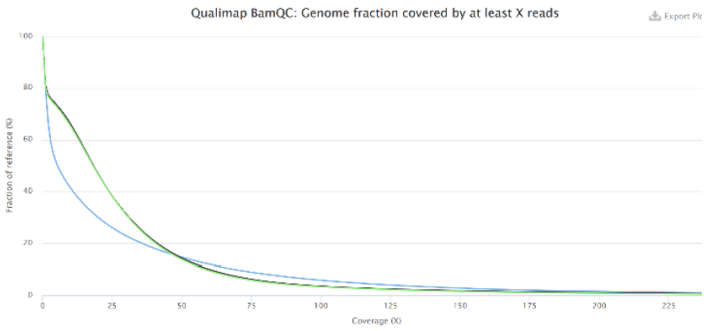
Quality control

The following is a MultiQC quality report for case 628, providing graphs both from BAMQC and FASTQC. The complete reports for each of the cases can be found in the Additional Materials section.

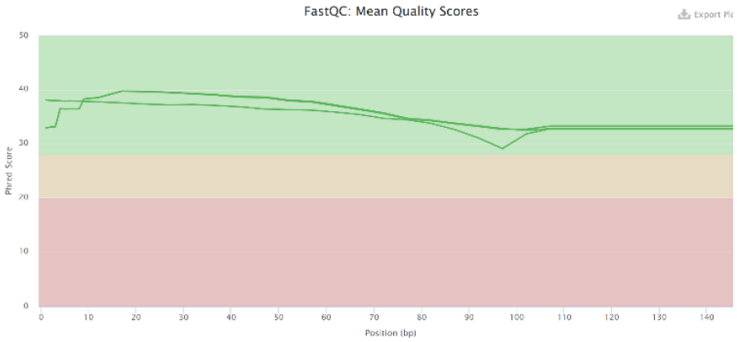
A

Sample Name	% GC	≥ 30X	Median cov	Mean cov	% Aligned	% Dups	% GC	M Seqs
case628_child	46%	22.8%	5.0X	24.1X	99.8%	5.4%	43%	3.0
case628_fathe	52%	31.2%	18.0X	27.3X	99.9%	6.1%	50%	2.2
case628_moth	52%	31.0%	18.0X	26.3X	99.8%	8.5%	50%	2.1

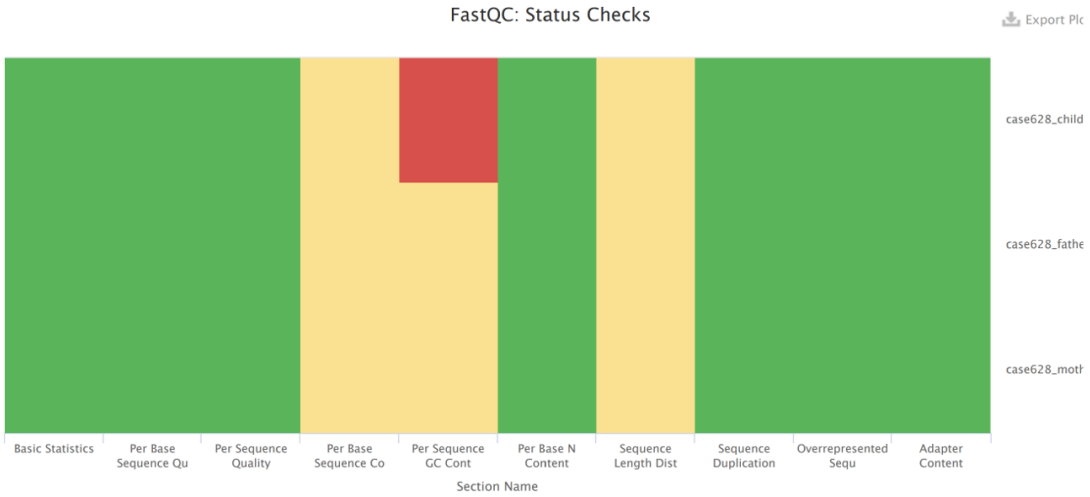
B



C



D



- In **figure A** we see a general statistics table showing that all of our samples have a very high mean coverage (over 20X for each sample), a very low percentage of duplicated reads and a very high percentage of mapped reads
- **Figure B** and **C** show respectively the coverage breadth and the mean quality scores for each position of the reads. Both plots show good results
- **Figure D** shows a simple visual recap of the report. Even though a few sections display warnings (such as Per Base Sequence Content, Per Sequence GC Content and Sequence Length Distribution), the majority of the report displays a good overall quality

Variant Calling & Variant Prioritization

A table reporting all the main information about the variant prioritization step and the subsequent diagnoses is provided. The complete Excel file and the file containing all the confidence scores can be found in the Additional Materials section.

Case Number	697	738	641	628	689
Inheritance Model	AR	AD	AR	AR	AD
Location	16:10996514-10996519	16:2142182-2142182	16:89858892-89858892	16:89815164-89815174	16:3789608-3789608
Consequence	frameshift_variant, splice_region_variant	stop_gained	stop_gained	frameshift_variant	stop_gained
Gene	CIITA	PKD1	FANCA	FANCA	CREBBP
Impact	HIGH	HIGH	HIGH	HIGH	HIGH
Disease	MHC class II deficiency	Autosomal dominant polycystic kidney disease	Fanconi anemia, Fanconi anemia complementation group A	Fanconi anemia, Fanconi anemia complementation group A	Rubinstein-Taybi syndrome due to CREBBP mutations

Case Number	610	683	657	586	681
Inherritance Model	AR	AD	AD	AD	AD
Location	-	-	16:50788249-50788254	-	16:3820696-3820696
Consequence	-	-	frameshift_variant	-	stop_gained
Gene	-	-	CYLD	-	CREBBP
Allele	-	-	HIGH	-	HIGH
Disease	None	None	Familial cylindromatosis	None	Rubinstein-Taybi syndrome due to CREBBP mutations

As we can see from the table, we were able to spot what we believe are pathogenic variants in some of our cases. However, in 3 of them we were not able to do so. In particular, the reasons for which we did not call a pathogenic variant for that particular disease model, reside in some details we want to further explain:

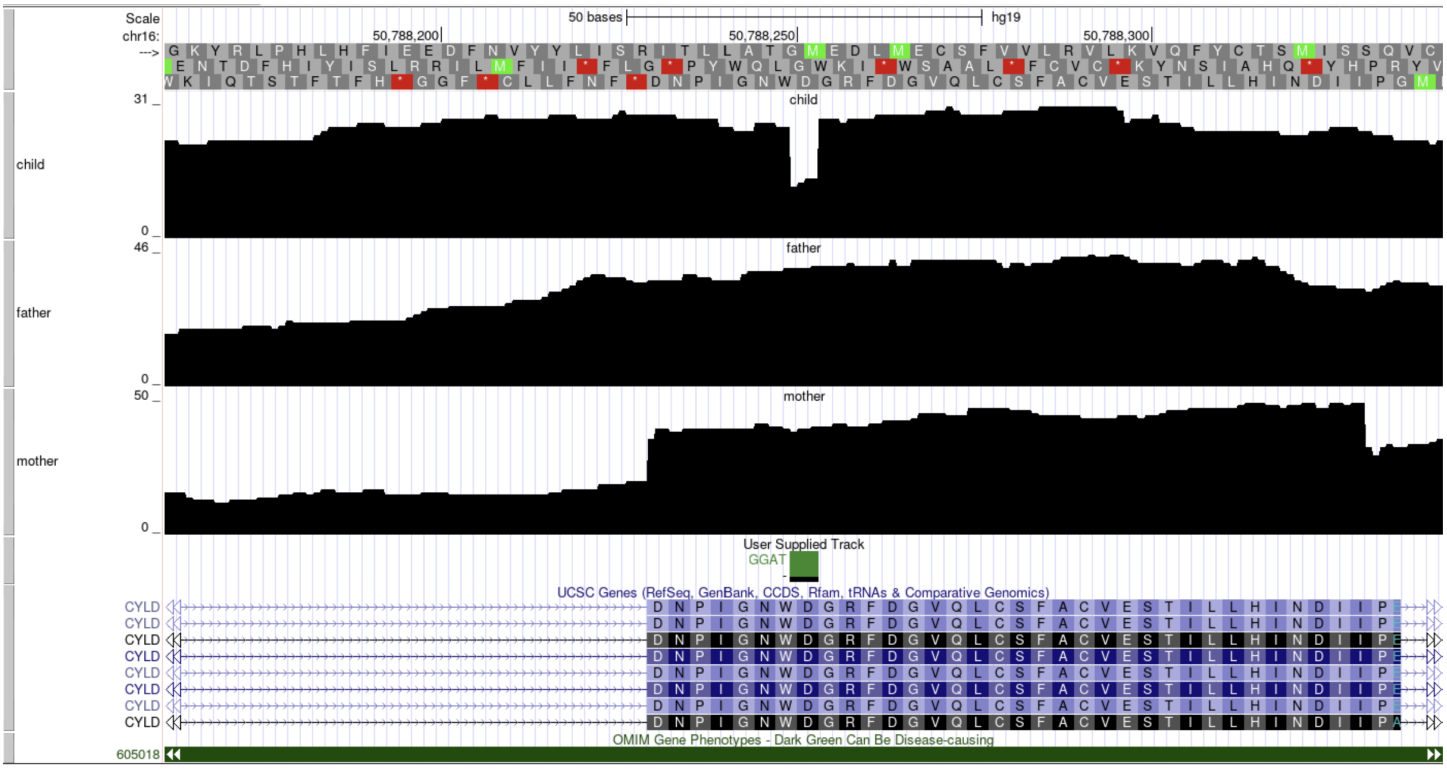
- **Case 610:** The only variant we were able to spot was located on the RBFOX1 gene. The variant, classified as an intron variant, had a LOW impact on the gene function and falls in a repeated region, which is where replication errors are more likely to occurr. Moreover, by visualizing the OMIM track on the UCSC Genome Browser, we can see the gene is color coded as light gray, meaning that no associated OMIM phenotype information is available.
- **Cases 683 & 586:** Even though we managed to spot possible variants having a MODERATE impact on the gene function, we decided not to call a pathogenic variant for 2 main reasons in both cases. First of all the confidence scores of the variants in the VCF file are very low (2.38811e-06 for both of them), and secondly some of the frequencies in the gnomAD subpopulations are indeed greater than the frequency of the diseases we are looking for.

In the other cases, we succeded to find pathogenic variants, all characterized by a very high confidence score. All the variants have an HIGH impact on the gene function (since they are either stop gain or frameshift variants). Moreover all the allelic frequencies, even for the gnomAD subpopulations, are compatible with the disease frequency we are looking for.

Variant Visualization

The disease causing variants we spotted using the VEP interface have been visualized in the UCSC genome browser, using the Custom Track section, providing both the VCF file and

the coverage tracks (the BedGraph files). All the screenshots from the UCSC Genome Browser can be found in the Additional Materials section



Genotype count: 3 (0 phased)
Alleles: (G)GGAT: 5 (83.333%); (G)-: 1 (16.667%)
Genotypes: (G)GGAT/(G)GGAT: 2 (66.667%); (G)GGAT/(G)-: 1 (33.333%)

Detailed genotypes

Genotype info key:
DP: Read Depth
AD: Number of observation for each allele
RO: Reference allele observation count
QR: Sum of quality of the reference observations
AO: Alternate allele observation count
QA: Sum of quality of the alternate observations
GL: Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy

Sample ID	Genotype	Phased?	DP	AD	RO	QR	AO	QA	GL
child	(G)GGAT/(G)-	n	26	11, 14	11	365	14	455	-32.778900, 0.000000, -24.776100
father	(G)GGAT/(G)GGAT	n	40	40, 0	40	1517	0	0	0.000000, -12.041200, -131.065000
mother	(G)GGAT/(G)GGAT	n	39	39, 0	39	1514	0	0	0.000000, -11.740200, -130.292000

Here are reported the UCSC screenshots for case 657. From the image above we can see the visualization on the reference genome of the coverage and the variant that has been spotted. The image below shows details about the genotypes of the members of the families. One interesting feature we would like to report is that, since in case 657 the child presents a deletion as a pathogenic variant, we clearly see a lower coverage in that locus (since there is one copy less for the reads to be aligned with).

These images are indeed coherent with the table in the Diagnoses sections, reporting the same affected gene and the same coordinates, as well as the same reported alleles.

Additional Materials

MultiQC reports for each case can be found here:

- MultiQC Folder

The complete Excel file, with all the VEP outputs can be found here:

- Tentative Complete.xlsx

Please note that even for those cases in which we failed to identify a pathogenic variant compatible with the criteria we are looking for, our “best guess” is provided, and the

reasons for which we did not consider such variants will be discussed in the Results section.

The confidence scores of the VCF files can be found here:

- Confidence_Scores

As stated in the previous section, even for the variants' confidence scores we decided to keep the same approach, providing our "best guess", even for those cases in which we failed to identify a pathogenic variant compatible with the criteria we are looking for. The reasons for which we did not consider such variants will be discussed in the Results section.

The variant visualization in the UCSC genome browser for each case can be found here:

- UCSC_Folder

The same rule stated before is applied in this context. Inside the folder are also the screenshot of those cases in which we failed to identify a pathogenic variant.

The complete script can be found here:

- Complete_Script_MK7