Whole exome sequencing and Variant calling to establish the genetic variants in a case of osteopetrosis in a family trio

Introduction

Osteopetrosis is a rare genetic disease that causes abnormally increased bone mass and density due to impaired bone-resorption and turnover (Coudert et al., 2015). This disease is both clinically and genetically heterogenous and it can be inherited through autosomal dominant/recessive or x-linked inheritance patterns (Povoroznyuk et al., 2021). It is important to note that the parents are unaffected and are known to be consanguineous meaning that they share a common ancestor, only the child is diagnosed with this disorder. Diagnosis is based on both radiography and genetic testing methods (R K et al., 2024).

In this study, the platform GALAXY will be used to make use of tools such as FreeBayes, DeepVarient and NaïveVarient calling to determine the differences in the exome sequencing data amongst the family trio. Through genetic testing and by using these tools, it should be possible to identify the variants that could potentially be the genetic cause of this child's disease.

Methods

Data acquisition and preprocessing

The exome sequence data was generated using the Illumina NGS platform acquired using data from a family trio where the child was diagnosed with osteopetrosis and the parents were unaffected. This data was retrieved from Zenodo/the Galaxy server (as premapped BAM files) and used for analysis. Hg19 (specifically Human Feb 2009- GRCh37/hg19)) is the reference genome used to map the data.

Quality control

Although the data had already been premapped, quality control on the BAM files was still done to assess the mapping quality, determine the base quality distribution, check the G+C content and to check for cross contamination. The bedtools convert tool was used to covert the BAM files into FASTQ files so the FASTQC tool could be used to generate a report. This process was repeated individually for all members of the family. The FASTQE tool was also used to get an easy and readable report on the overall quality of the data. Then finally, the MULTIQC tool was used to combine all these analyses together into a single report.

Read mapping and postprocessing

Since the original files were already premapped, the read mapping step was skipped and postprocessing was done. The SamTools View tool was used to filter out any unmapped reads/mates and the RmDup tool was used to remove any duplicates and to make sure that the resulting BAM file is paired ended. The RmDup tool generated 3 BAM files (one for each member of the family), and these files were used at the input for the multiple variant calling tools used below.

Variant calling

Three variant calling tools were used in this report namely FreeBayes, DeepVarient and Naive Variant calling. These tools would identify and call out all single nucleotide polymorphisms (SNP's) or insertions and deletions (Indels) present in the data by comparing the sequences to the reference genome Hg19

FreeBayes (FB):

> The data inputs were the RmDup BAM files, and the multiple file option was used to do variant calling for all family members simultaneously. The VCF file that was generated from this was run into the tool bcftools norm to split the multiallelic variants into their own rows to ensure compatibility with other downstream analysis tools (like GEMINI). The final output of this data was another uncompressed VCF file.

DeepVariant (DV):

> For DeepVariant the same input files were used but unlike FreeBayes, each variant calling was done separately for each family member. The outputs were a VCF file and an HTML report for each member. The bcftools merge tool was used to combine all the VCF outputs into one VCF file.

Naïve Variant (NVC):

> Same as above, RmDup files were used as input in the Naïve variant calling tool to generate VCF files for each member using hg19 as the reference genome. The outputs were merged into one file using the bcftools merge tool.

Variant annotation and reporting

The tools used in this section were used for the purpose of discovering any clinical or biological relevance in the data. Variants were prioritised according to this, and the results of the annotations were presented in human readable HTML reports.

The main tool used for annotation was the SNPeff eff tool at the end of each variant calling method with the VCF files as input. This tool generates 2 files, a VCF file and an HTML file which consists of summary stats for each variant caller.

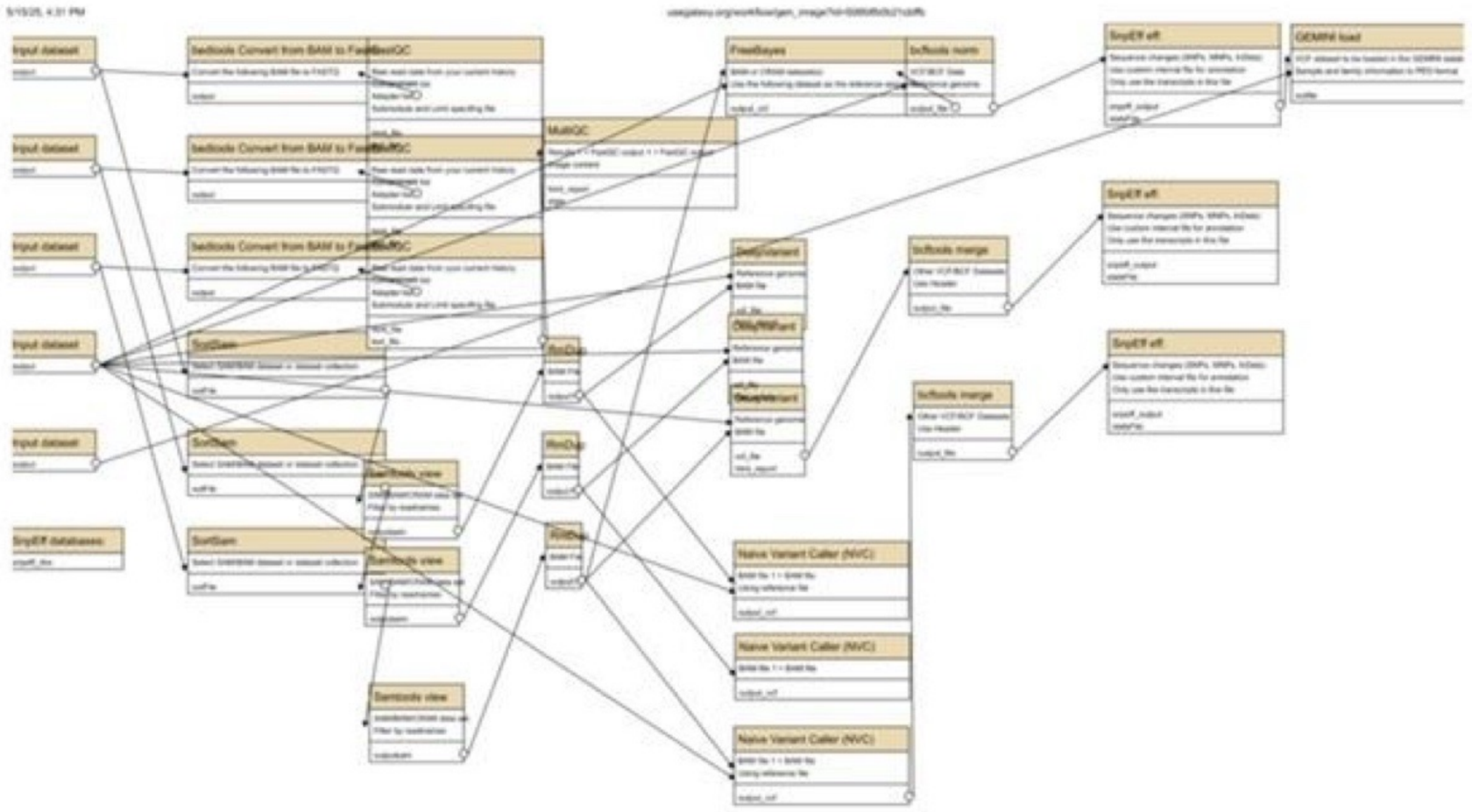

Figure 1: Full GALAXY workflow showing the steps and methods used. Link to a clearer and interactive version of the workflow in GALAXY is included under references

Results

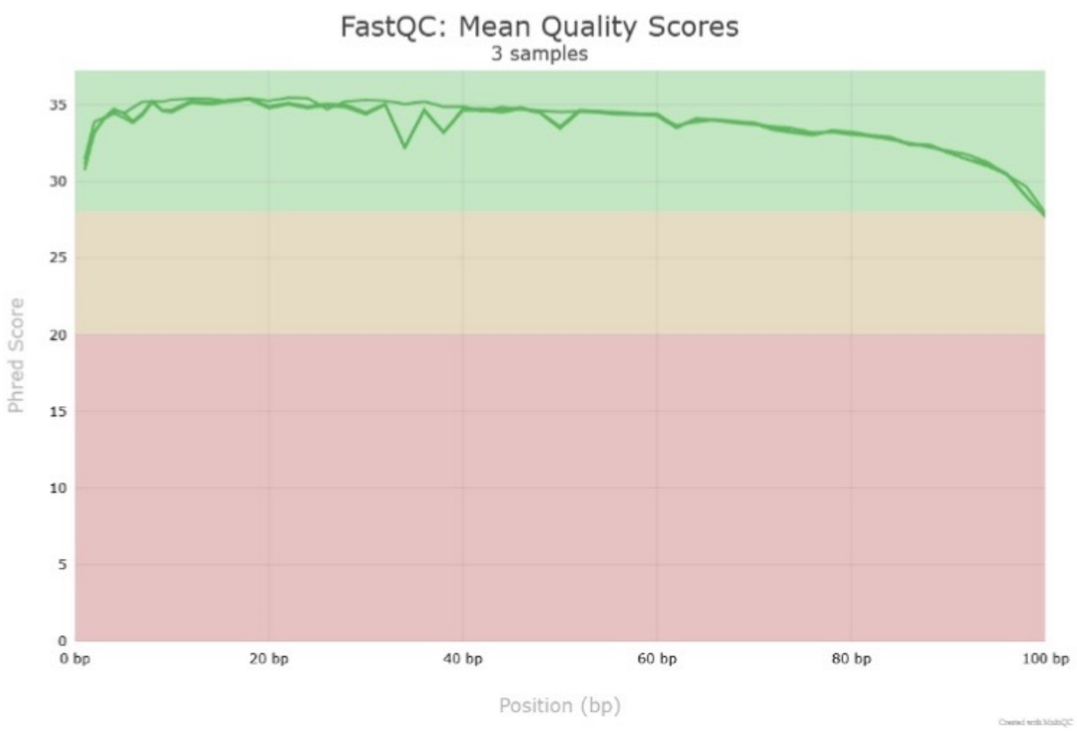

Figure 2A: The average quality scores of each premapped read from each family member

Figure 2B: The quality scores per sequence of each premapped read from each family member

	No. of Variants	Variant frequency
FreeBayes	38 647	1 in 3 755
Deep Variant	10 120	1 in 14 360
Naive Variant calling	1 825 102	1 in 80

Figure 3: Table showing the number of variants detected by each tool as well as the frequency of the variant per base pair

	FreeBayes	Deep Variant	Naïve Variant calling
SNP	35 761	7 198	1 805 185
MNP	1 101	0	0
Ins	590	850	6 327
DELs	1 042	2 072	13 590
Mixed	153	2	0

Figure 4: Table showing the different types of variants and the number of each variant in each.

	FreeBayes	Deep Variant	Naïve Variant calling
Missense	1 481 (79%)	1 169 (46%)	1 347 488 (75%)
Nonsense	22 (1%)	8 (0.3%)	62 009 (3%)
Silent	368 (19.7%)	1 367 (54%)	380 170 (21%)

Figure 5: The number of effects (and %) by functional class

	FreeBayes	Deep Variant	Naïve Variant calling
HIGH	556 (0.34%)	222 (0.57%)	128 606 (1.88%)
MODERATE	1 617 (1.04%)	1 195(3.07%)	1 342 012 (18.64%)
LOW	1 700 (1.09%)	2 305 (5.9%)	524 756 (7.68%)
MODIFIER	152 354 (97.5%)	35 138 (90.4%)	4 837 044 (70.8%)

Figure 6: Number of effects by impact (and %)

Ref	Alt	Impact	Gene	ClinVar Significance	ClinVar Disease	Family genotype (M, D, C)
C	T	Missense	MCPH1	Benign	Microcephaly	C/T, C/T, T/T
C	G	Nonsense	LPL	Likely Benign	Hyperlipidaemia, Hyperlipoproteinemia Coronary heart disease	C/G, C/G, G/G
T	C	Missense	CHRNA2	Benign	Nocturnal frontal lobe epilepsy	C/T, C/T, C/C
G	A	Nonsense	CA2	None	Osteopetrosis with renal tubule acidosis	G/A, G/A, A/A
T	C	Missense	SLC39A4	Benign	Hereditary acrodermatitis	C/T, C/T, CC

Figure 7: A table showing the results of a GEMINI analysis done on the SNPEff eff FreeBayes file

Discussion

Although the very first input files were premapped it is still essential to do quality control as a preprocessing measure. In figure 2a the quality scores of all 3 BAM files are good meaning these sequence reads are good enough to undergo proper variant analyses. Figure 2b further supports this because it's clear to see that the quality per base sequence read is high. This cleared any necessity to do any more preprocessing steps such as filtering or trimming any sequences.

Three variant caller tools were used namely DeepVarient, Freebayes and Naïve Varient Caller, to compare the outputs of each one. All 3 gave different annotations. Figure 3 shows that NVC had the highest number of variants with the highest frequency compared to the other two tools. This was initially thought to be a good thing as it represents high sensitivity, but this also indicates

the possibility that this tool generates a lot of false positives. FB comes in second with a much lower number and frequency of variants and lastly, DV with the lowest number of variants and lowest frequency rate. Same as with Naïve variant, the results from DV could also indicate a high number of false negatives but it could also mean that the DV tool is highly selective. Existing literature acknowledges that each tool has its own advantages and disadvantages e.g. FB is better at identifying SNP's whereas DV is better at identifying indels (Stegemiller et al., 2023, Supernat et al., 2018). This is also supported by the results in figure 4. There is not much existing literature for NVC. The assumed reason is that this tool is a very basic and primal form of FB and DV and it is used for less complex analyses. This also explains why the results for NVC are very different from the other tools.

All the tools indicated a higher number of SNPs with DV and NCV detecting higher Indels compared to FB. FB detected that 79% of the variants resulted in missense point mutations whereas DV and NCV variants showed a balance of both missense and silent point mutations. This is relevant because missense mutations result in changes in amino acids potentially affecting protein structure and function which could play a role in the development of osteopetrosis in the child.

Only the output from the FB data annotation could be analysed further using GEMINI. Figure 7 tabulates the top hits for the variants that could have contributed to the development of osteopetrosis. Since the reference was to chromosome 8, this suggests that this could be an autosomal disease but since both parents are unaffected the assumption is that this was autosomal recessive inheritance. A nonsense mutation in the Carbonic anhydrase II (CA2) gene was linked to the disease which causes a deficiency in this enzyme. The lack of this enzyme impairs the osteoclasts' ability to generate ions that facilitate bone reabsorption which will result in abnormal bone density and mass. Early detection and genetic testing paired with gene therapy is crucial for the treatment of this disease

Conclusion

All 3 variant caller tools had their unique advantages and disadvantages. Depending on the purpose, each tool could be used to successfully gather information on exome sequences. FreeBayes is better suited for this experiment because it allows for further analysis in GEMINI to identify and biological or clinical significance to the data. This enabled us to conclude that an inherited autosomal recessive nonsense mutation in the CA2 gene is what lead to the development of osteopetrosis in the child.

References

Coudert, A.E. *et al.* (2015) 'Osteopetrosis and its relevance for the discovery of new functions associated with the skeleton', *International Journal of Endocrinology*, 2015, pp. 1–8. doi:10.1155/2015/372156.

Povoroznyuk, V.V. *et al.* (2021) 'Остеопетроз: класифікація, патоморфологія, генетичні порушення, клінічні прояви (огляд літератури та власне клінічне спостереження)', *PAIN, JOINTS, SPINE*, 9(2), pp. 135–142. doi:10.22141/2224-1507.9.2.2019.172125. (Translated version from elicit was used)

R K, S. *et al.* (2024) 'Alberg's Schoenberg Disease – a review and Case report', *International Journal of Contemporary Dental Research*, 2(1), pp. 1–5. doi:10.62175/apdch2401.

Stegemiller, M.R. *et al.* (2023) 'Using whole genome sequence to compare variant callers and breed differences of US Sheep', *Frontiers in Genetics*, 13. doi:10.3389/fgene.2022.1060882.

Supernat, A. *et al.* (2018) 'Comparison of three variant callers for human whole genome sequencing', *Scientific Reports*, 8(1). doi:10.1038/s41598-018-36177-7.

Link to workflow: https://usegalaxy.org/u/thatooxx/w/workflow-constructed-from-history-exome-sequencing-2/json