VCF Analysis

Team Silver:
Cat Turnbull, 1370807
Daniel Shepherd, 1514996
Stefenie Pickston, 1506427

28-10-2022

Abstract

A Variant Call File or VCF holds the genetic variations for one or more individuals. Many illnesses are linked to genetic causes that could be inherited, so from analysis of this file, one may be able to understand which genes are responsible for a particular illness. Our team was tasked to find the genetic variations that were responsible for two illnesses - Sickle Cell Anaemia and Retinitis Pigmentosa - in a family of seven where some family members would exhibit symptoms from either one or both illesses. We were given the inheritance diagram for both illnesses in the family, as well as 2 VCF files: one from a population sample of 2,000 people and one from the family of seven. Our approach consisted of narrowing down the genetic variations through filtering by chromosone, inheritance and exonic regions before making comparisons with the population and consulting current biological research and resources on the illnesses.

1 Sickle Cell Anaemia (SCA)

SCA is a rare genetic blood disease caused by mutations in the *HBB* (Haemoglobin Subunit Beta) gene[1]. The *HBB* gene is located on Chromosome 11p15.4 on the negative strand, with three exons and 3,932 base pairs[2][3]. The *HBB* gene is at base position 5,225,464 to 5,227,071[3][4]. The most common mutation in the *HBB* gene is a point mutation of an A to a T; this substitution also changes the amino acids from Glutamine to Valine. This variant causes a change in the confirmation of the Hemoglobin from normal to abnormal Hemoglobin (also known as sickle Hemoglobin or HbS)[6]. There are also multiple types of Hemoglobin variations that, when seen in one allele, the variant causes no issue. When seen in two alleles, the variants cause abnormal Hemoglobin and SCA[6][7][8]. Hemoglobin is a protein responsible for carrying oxygen. SCA is a disease that affects the shape of the red blood cells, causing the blood to carry oxygen around the body ineffectively. Sickle cells are often stiff and sticky, and individuals with the disease commonly have blocked blood flow in the blood vessels of organs and limbs. The lack of blood flow can cause pain and organ damage but also increases infection risks[5].

SCA affects 1-5/10,000 individuals; onset is seen at any age

1.1 Methods

Before we began coding, we needed to know what genes affected these diseases that fit the inheritance pattern. We knew SCA's inheritance pattern was autosomal recessive, so we searched for the gene that causes this disease with the pattern of inheritance. What we found was that SCA was related to one gene called the *HBB* gene[11]. We could see what chromosome location of that gene (Chr11p15.4).

By finding this information, we narrowed our variant search to a particular region in the genome. Using Python as our programming language, we took the family VCF files containing all the variants from the selection of families with seven members (mother, father, three daughters, and two sons). We filtered the family files to create a file that only had the variants on chromosome 11; this file contained 487,930 variants. We only wanted to look at variants inside the exon regions to narrow the variants further. We refined the variants down to 18,320.

Finally, we wanted only the variants that fit the inheritance pattern of SCA and the inheritance diagrams given to us. This inheritance pattern is "father and mother were carriers of the diseased mutation but did not have the disease, daughter one and son one were also carriers, daughter two and son two had two alleles with the variant and therefore had the disease and finally daughter three had no variant alleles". We found three possible disease mutations. We also compared these mutations to the population file to see if the same or other mutations were at the same positions we saw in our family variants.

1.2 Potential Genetic Causes

Variants Position	*File*	*Mutation - Base Change*	*Mutation - Amino Acid Change*
77376796	Population	T to A	N/A
	Family	TTCCAACACAC-ACACACACGCA-CACGCACACA to TTCCAACACA-CACACACACG-CACACGCACA-CATCTCGCCGT, GTCCAACACA-CACACACACG-CACACGCA-CACA	N/A
88240118	Population	G to A	Glutamic Acid to Lysine (First base change)
	Family	G to A	Glutamic Acid to Lysine (First base change)
94801365	Population	A to G	Glutamic Acid to Glycine
	Famliy	A to T	Glutamic Acid to Glycine

Table 1: After filtering the family VCF file, three mutations were identified to fit the inheritance pattern and were in the exonic region. The population VCF file also found three variants in the exact locations. These tables show these variants of the population and family files at the particular location and the possible amino acid change that these variants could cause based on published literature that we will discuss in this section.

After filtering the files, we found three potential disease-causing mutations. We compared these mutations to the population variant files and published literature about the known disease-causing variants to determine if the change in the sequence was a pathogenic variant. Searching for published known disease-causing variants on the *HBB* gene[6][7][8] and from further research, we found an *HBB* mutation database with all the known variants in this gene; the database is called the "Leiden Open Variation Database (LOVD) – The globin gene server"[12].

The first mutation was seen at position 77376796. In the population, it was a T to an A mutation. The family data was a long sequence of nucleotides

$$TTCCAACACACACACACACGCACACGCACACA$$

to

$$TTCCAACACACACACACACGCACACGCACACATCTCGCCGT,$$
$$GTCCAACACACACACACACGCACACGCACACA$$

When looking at the mutation database, we found 49 variants that see a T to an A substitution, with 46 of these variants causing an amino acid change (On this database it has a link to the HbVar database (A database of Human Hemoglobin Variants and Thalassemia). Unfortunately, due to time constraints and the fact that we were not given the whole sequence to blast to the gene sequence to pinpoint the exact location that this variant will be was not possible, so to say which one of the 49 variants our variant is difficult. However, the database suggests that it may be possible for this variant to be a pathogenic mutation. When we look at the family variant, we see an extended sequence with what looks to be a long repeating sequence that is being replaced by two long repeating sequences, while there may be a valid mutation in the population data, the family data suggests that this variant could be due to a misalignment of the sequences when compared to the reference sequences.

The next mutation in the population and family files, at position 88240118, is a G to A substitution. This substitution could be the well-known substitution that causes a change to the beta-globin chain in the Hemoglobin protein from glutamate to lysine at the sixth position of the chain[7], creating the Hemoglobin C. variant. On the LOVD database we also saw 95 known G to A substitutions, with 87 of those variants causing an amino acid change. This shows us that there is a high possibility that the variant that we see in our family data could be a pathogenic allele, with further research and information, it could be the explanation for the disease that is presenting in the family. Like with the previous mutation, we would need the whole gene/genome sequence of the family to BLAST it against the reference gene/genome to be able to pinpoint where exactly we see that variant to investigate its pathogenicity further.

The final mutations were seen at position 94801365. In the population files, the variant we see is a change an A to G. In the LOVD database, there are 72 variants, with 62 causing amino acid changes. In the family data, there was a change in the same position from A to T. After doing some research, we found this is a common mutation for SCA is a point mutation on codon 6 of the beta haemoglobin chain, which changes the amino acid from glutamic acid to valine. The LOVD database has 66 variants, and 61 have amino acid changes. For both mutations, there is a high possibility that the variants we have seen are pathogenic alleles, and we could determine that with further research and information. Like with the previous mutations, we would need the whole gene/genome sequence of the family to BLAST it against the reference gene/genome to be able to pinpoint where exactly we see that variant to inves-

tigate its pathogenicity further.

1.3 Conclusion

Overall, in our filtering and investigations into the family VCF files, we found two mutations that could be potential disease-causing mutations that could contribute to the SCA seen within the affected family. The LOVD database and published work have given us the confidence that we need to show that when we progress our research further, there is a high likelihood that we will find a positive link between these mutations and the disease we see in these families. The next steps would be to take the whole sequence and run BLAST and multiple alignments against the HBB gene to analyse and compare where these mutations occur within the sequence and whether they fall in known variant locations seen within the published work and the LOVD database.

References

[1] Wonkam, A., Chimusa, E.R., Mnika, K., Pule, G.D., Ngo Bitoungui, V.J., Mulder, N., Shriner, D., Rotimi, C.N., Adeyemo, A. (2020). *Genetic modifiers of long-term survival in sickle cell anemia.* Clinical and Translational Medicine 10.

[2] National Institute of Health (n.d.). *HBB, hemoglobin subunit beta [Homo sapiens (human)].*NCBI. `https://www.ncbi.nlm.nih.gov/gene/3043`

[3] Gene Cards (2022, August 30). *HBB gene – Hemoglobin Subunit Beta.* `https://www.genecards.org/cgi-bin/carddisp.pl?gene=HBB`

[4] Gene Loc – Genome Locator (n.d.). *Exon Structure for HBB.* Gene Cards. `https://genecards.weizmann.ac.il/geneloc-bin/exon_struct.pl?disp_name=HBB&chr_nr=11`

[5] MalaCards – Human Disease Database (n.d.). *Sickle Cell Anemia.* Gene Cards. `https://www.malacards.org/card/sickle_cell_anemia`

[6] Melamed, D., Nov, Y., Malik, A., Yakass, M. B., Bolotin, E., Shemer, R., ... & Livnat, A. (2022). De novo mutation rates at the single-mutation resolution in a human HBB gene region associated with adaptation and genetic disease. Genome research, 32(3), 488-498.

[7] Ashorobi, D., Ramsey, A., Yarrarapu, S. N. S., & Bhatt, R. (2019). *Sickle cell trait.*

[8] Karna, B., Jha, S. K., & Al Zaabi, E. (2022). *Hemoglobin C Disease.* In StatPearls [Internet]. StatPearls Publishing.

[9] Galanello, R., & Origa, R. (2010). *Beta-thalassemia.* Orphanet journal of rare diseases, 5(1), 1-15.

[10] Teaching Health Tech (n.d.). *Codon List*. DTU. https://teaching.healthtech.dtu.dk/22110/index.php/Codon_list

[11] OMIM (n.d.). *Sickle Cell Anemia #603903*. https://www.omim.org/entry/603903#molecularGenetics

[12] Giardine, B., & Borg, J. (n.d.). *Beta Globin (HBB)*. Leiden Open Variation Database (LOVD) – The Globin Gene Server. https://lovd.bx.psu.edu/variants.php?select_db=HBB&action=view&view=0002252%2C0001926%2C0

2 Retinitis Pigmentosa (RP)

Retinitis Pigmentosa (RP) is an inherited retinal disorder that affects 1 in 4000 people worldwide[1]. There have been over 80 genes that have been linked to causing syndromic and non-syndromic RP[1]. The three most common affected genes are RHO (autosomal dominant pattern of inheritance), USH2A (autosomal recessive pattern of inheritance) and RPGR (X-linked recessive)[1]. It is a progressive disease characterised by the impairment of rod photoreceptors initially, progressing as nyctalopia and peripheral field loss, and finally, central vision loss and subsequent progressive cone involement[1]. Other symptoms are also related to RP[1]. These symptoms can begin as early as 20 years of age[3]. To follow our desired autosomal dominant inheritance pattern, we decided to only focus on the RHO gene.

The RHO gene is found on Chromosome 3q22.1 at the genomic coordinates of 129,528,639 bp to 129,535,344 with five exons[3][9]. The RHO gene codes for visual pigments called Rhodopsin necessary for vision as they are light-absorbing molecules[3]. There have been over 150 mutations identified in the RHO gene in those individuals with RP[4].

RHO gene mutations account for 20-30% of all autosomal dominant (ad) RP (adRP)[4]. Gene mutations in the RHO gene fall into two classes; Class A mutations cause severe early-onset loss of rod function. The class B mutations are a milder phenotype with retained rod function in specific retinal regions[6].

2.1 Methods

Before we began coding we needed to know what genes affect this disease that fit the inheritance pattern. We knew the inheritance pattern we wanted to look for was autosomal dominant, so we searched for the gene that causes this disease with the pattern of inheritance. We found that the RHO gene was the gene whose mutations contributed to adRP[2][3][4]. We found what chromosome that gene was located on (Chr3q22.1)[3]. This information helped narrow our variant search to a particular region in the genome. Using Python as our programming

language, we took the family VCF files containing all the variants from a selection of families with seven members (mother, father, three daughters, and two sons). Filtered that file to create a file that only had the variants on chromosome 3; this file contained 726,511 variants. To further narrow variants, we only wanted to look at variants inside the exon regions to further narrow variants. Next, we only wanted the variants that fit the inheritance pattern of adRP and the inheritance diagrams that were given to us (the father had the diseased mutation but did not have the disease, mother, daughter one, daughter three and son one had no variant alleles. Daughter two and son two had one allele with the variant and therefore had the disease). We, unfortunately, found no potential disease-causing mutations. Because of this, we opened the search to look 1000bp on either side of the exon region; this found 14 variants. We also compared these mutations to the population file we were given to see if the same or other mutations were seen at the positions that we saw in our family variants. However, we did not find any variants at the same positions in the population files.

2.2 Potential Genetic Causes

Table 2: After filtering the family VCF file there were 11 mutations that were identified to fit the pattern of inheritance; we had to search 1000bp outside for this gene, as we did not get any hits within the exonic region. We also did not find any population variants in the exact location within the population VCF files. These tables show these variants; at; the particular location and the possible amino acid change these variants could cause based on published literature that we will discuss in this section. There were also three "variants" that presented more like misalignments than variants, so we will not discuss them in this report.

Variants Position	*File*	*Mutation - Base Change*	*Mutation - Amino Acid Change*		
37845175	Family	G to A	Glycine to Arginine (Class B)	Glutamic Acid to Lysine (Class A)	Glycine to Aspartic Acid (Class B)
49248662	Family	T to C	Leucine to Proline (Class A)		
111318004	Family	T to A	Methionine to Lysine (Class B)		
112006709	Family	T to TA (indel)	-		
122259853	Family	T to A	Methionine to Lysine (Class B)		
124895774	Family	A to G	Asparagine to Glycine (Class B)		
171570115	Family	CAA to C (frameshift)	Glutamine to Proline (Class A)	Glutamine to Stop (Class B)	Glutamine to Stop (Class B)
179043889	Family	T to C	Leucine to Proline (Class A)		
179043894	Family	G to A	Glycine to Arginine (Class B)	Glutamic Acid to Lysine (Class A)	Glycine to Aspartic Acid (Class B)
182975930	Family	G to A	Glycine to Arginine (Class B)	Glutamic Acid to Lysine (Class A)	Glycine to Aspartic Acid (Class B)
184061911	Family	T to C	Leucine to Proline (Class A)		

After filtering, we found 14 variants; however, only 11 of them we deemed potential disease-causing variants, and the other three we deemed misalignments. In the 11 variants we found, we saw three G to A, three T to C, two T to A, a T to TA, an A to G and finally, a CAA to C.

After researching all these mutations using the "Rentino Genetics" mutation database, which is a database that contains known mutations in retinal genes[7][8] as well as published work[5][6] we found that all except the T to a TA have been described by literature and seen in the genome before. We will discuss each of them in more depth below. It is important to note that some of these mutations have more than one potential change, some don't, and before we can be sure that they are a disease-causing variant we would need to do further research, which would include using the whole genome sequence to BLAST the gene against a reference genome to see where precisely these changes are occurring to be more sure of the pathogenic state of these variants.

The first variant we will look at is the G to A at positions 37845175, 179043894, and 182975930. Looking at published literature[5][6] we identified three potential amino acid changes in known mutations; the location of the mutation would depend on what amino acid is altered; if the mutation occurs at codon 188 or 106, the change would be a Glycine to an Arginine[5]. If it is at codon 181, the change would be Glutamic Acid to a Lysine, and finally, if it is seen at codon 89, the change will be Glycine to an Aspartic Acid. All these variants except the E181K are Class B, so they are responsible for a milder phenotype of adRP. The fact that there are four potential known variations from the change in the sequence gives us confidence that further investigation could be a potential pathogenic variant in the family.

The second variant, T to C, is seen three times at positions 49248662, 179043889, and 184061911. This variant is responsible for one Class A mutation at two different sites[5]. This mutation at codon 88 or 131 sees a change from a Leucine to a Proline. As this is a class A mutation, a change like this is detrimental to gene function; further investigation could show a high potential of this gene becoming a pathogenic variant.

The third variant seen is a T to an A at positions 111318004 and 122259853; this variant causes a change in the amino acid from a Methionine to a Lysine at codon 207[6]. This change is a Class B change. This has some grounding to be a contributor to adRP that may be seen in that family, but without further analysis, we won't be able to be 100% sure. As there is only one mutation with that change on the Rentino Genetics database[8] we feel it is a candidate for a potential disease-causing variant but has a high likelihood that it could be nothing.

The fourth variant we saw was an indel mutation at position 112006709, a T to TA. When looking at the database and within the literature, there was no acknowledgement of this variant being seen or noted. While it is possible that this could be a disease-causing mutation, the lack of any published work makes us sceptical. We would need further information to either confirm that this is indeed a variant or whether it is just a misalignment.

The fifth variant was an A to G at position 124895774, which changes the

amino acid Asparagine to Glycine at codon 15[6]. This is a Class B mutation. We also observe one A to G change on the Retino Genetics database[8]. This variant could also possibly be a variant.

The final variant seems to be an indel of CAA (glutamine) to a C at position 171570115, which could cause a frameshift to the codons. In the published literature, there have been four mutations that involve glutamine. The first is glutamine into a proline at codon 344 (class A), and three are changed to stop codons at codon 64, 312, and 344 (class B)[5]. This variant would need further investigation before we could determine whether it was an alignment error or is an indel. We would need to look at the whole gene sequence and see where this variant lands to see the consequence of this change.

2.3 Conclusion

Overall, in our filtering and investigations into the family VCF files, we found 11 individual variants that could be potential disease-causing mutations that could contribute to the adRP seen within the affected family. The Retino Genetics mutation database and published works have given us the confidence that we need to show that when we progress our research further, there is a high likelihood that we will find a positive link between these mutations and the disease we see in these families. The following steps of our research would be to take the whole sequence and run BLAST and multiple alignments against the RHO gene to analyse and compare where these mutations occur within the sequence, to see what codons these mutations fall into to help reduce the ambiguity that we see with some of the mutations and determine whether the adRP variants seen are class A or B mutations to understand the severity of the disease symptoms.

References

[1] Gene Vision. (n.d.). *Retinitis Pigmentosa: for professionals.* `https://gene.vision/knowledge-base/retinitis-pigmentosa-for-doctors/`

[2] OMIM (n.d.). *Retinitis Pigmentosa; RP #268000.* `https://www.omim.org/entry/268000`

[3] OMIM (n.d.). *Rhodopsin; RHO #180380.* `https://omim.org/entry/180380`

[4] National Institute of Health (n.d.). *RHO gene – Rhodopsin.* MedlinePlus. `https://medlineplus.gov/genetics/gene/rho/`

[5] Iannaccone, A., Man, D., Waseem, N., Jennings, B. J., Ganapathiraju, M., Gallaher, K., ... & Klein-Seetharaman, J. (2006). *Retinitis pigmentosa associated with rhodopsin mutations: Correlation between pheno-*

typic variability and molecular effects. Vision research, 46(27), 4556-4567.

[6] Athanasiou, D., Aguila, M., Bellingham, J., Li, W., McCulley, C., Reeves, P. J., & Cheetham, M. E. (2018). *The molecular and cellular basis of rhodopsin retinitis pigmentosa reveals potential strategies for therapy*. Progress in retinal and eye research, 62, 1-23.

[7] Ran, X., Cai, W. J., Huang, X. F., Liu, Q., Lu, F., Qu, J., ... & Jin, Z. B. (2014). *'RetinoGenetics': a comprehensive mutation database for genes related to inherited retinal degeneration*. Database, 2014

[8] Rentino Genetics (n.d.). *RetinoGenetics; mutation database*. `http://www.retinogenetics.org/`

[9] National Institute of Health (n.d.). *RHO - rhodopsin [Homo sapiens (human)]*.NCBI. `https://www.ncbi.nlm.nih.gov/gene/6010`