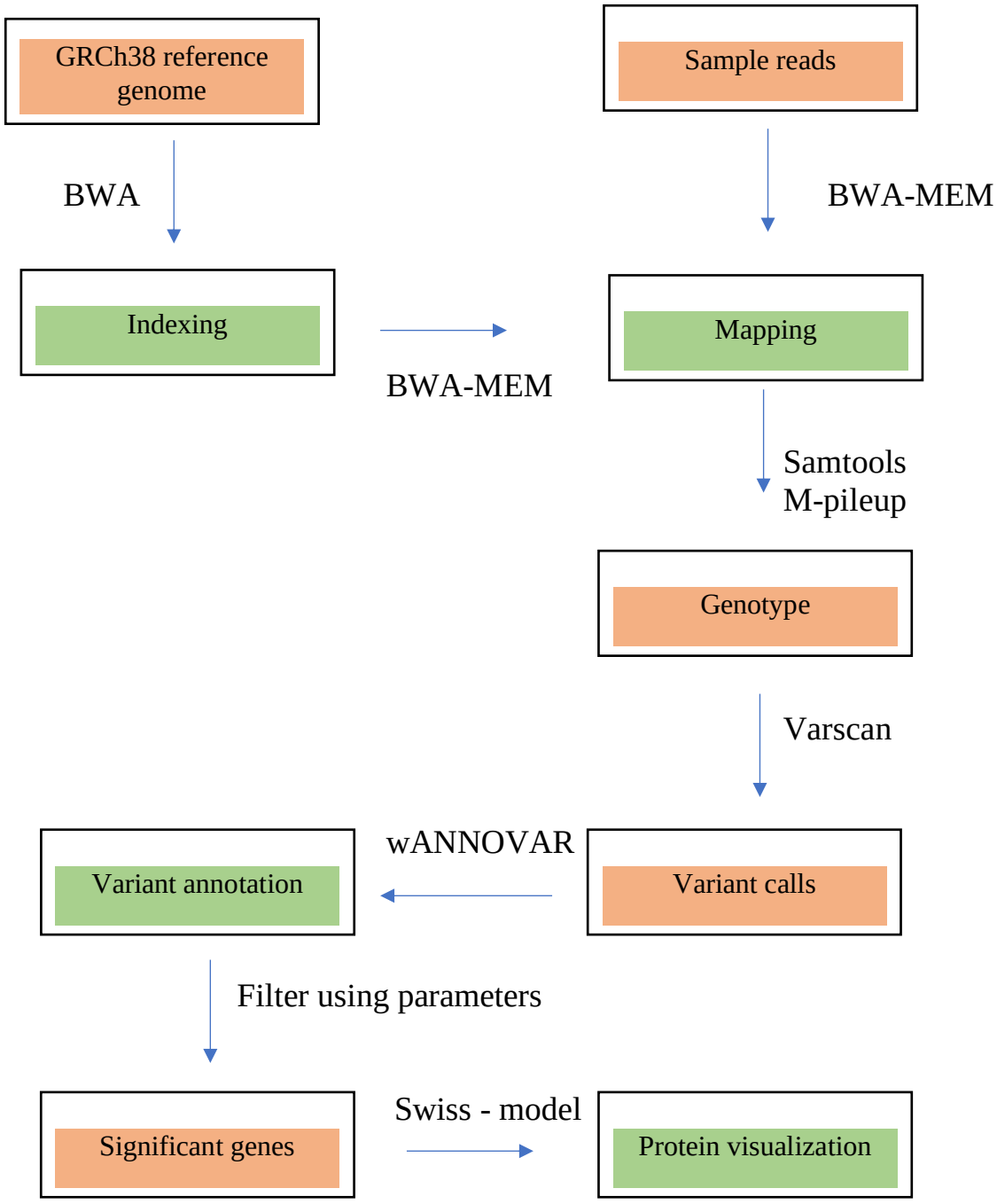
Exome Analysis

The exome analysis project was conducted by selecting a sample through the 1000 genomes phase 3 project based on the technology exome to obtain sequence data, the sample which I chose was a British Male in England and Scotland. I uploaded the reference genome (hg38) as well as the forward and reverse reads of the selected sample onto galaxy and followed the method



Detailed information for each of these steps have been tabulated and summarized below:

Cell Line source/ Biosample ID/ SRR ID	**HG00152 at coriell/SAME124593/ SRR769545**
Gender	Male
Population	British in England and Scotland
Super population	European

Table 1: General information about the sequence considered:

The sample described above was subjected to various steps mentioned in the methodology to obtain a variant calling file(vcf).

Type of Variant	**Count**
Total variants	34605
Total synonymous variants	14831
Total non-synonymous variants	18963
Protein truncating variants	521

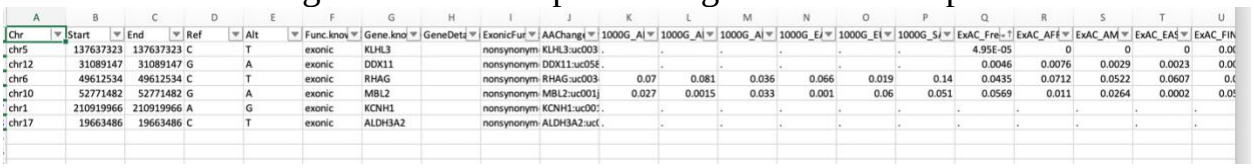
Table 2: General information on the variants obtained

Since the total number of non-synonymous variants are 18,963, I have applied some filters to narrow down my search and study those variants which are most significant. The filters used to narrow down the search included:

- The first filter that I used was on the exonic function column to narrow down only the non-synonymous variants
- The second filter that I used was on the Clinvar_SIG column to filter the pathogenic or most likely pathogenic variants

- The third filter I applied was on the CADD_phred score column to filter all values >= 25

A screenshot of the genes obtained upon adding these filters is pasted below:



Chr	Start	End	Ref	Alt	Func.kno	Gene.kno	GeneDeta	ExonicFu	AAChang	1000G_A	1000G_A	1000G_A	1000G_E	1000G_E	1000G_S	ExAC_Fre	ExAC_AF	ExAC_AM	ExAC_EA	ExAC_FIN
chr5	137637323	137637323	C	T	exonic	KLHL3		nonsynonym	KLHL3:uc003	.	.	.	.	.	.	4.95E-05	0	0	0	0.0
chr12	31089147	31089147	G	A	exonic	DDX11		nonsynonym	DDX11:uc05	.	.	.	.	.	.	0.0046	0.0076	0.0029	0.0023	0.0
chr6	49612534	49612534	C	T	exonic	RHAG		nonsynonym	RHAG:uc003	0.07	0.081	0.036	0.066	0.019	0.14	0.0435	0.0712	0.0522	0.0607	0.0
chr10	52771482	52771482	G	A	exonic	MBL2		nonsynonym	MBL2:uc001	0.027	0.0015	0.033	0.001	0.06	0.051	0.0569	0.011	0.0264	0.0002	0.05
chr1	210919966	210919966	A	G	exonic	KCNH1		nonsynonym	KCNH1:uc00	.	.	.	.	.	.	.	.	.	.	.
chr17	19663486	19663486	C	T	exonic	ALDH3A2		nonsynonym	ALDH3A2:uc	.	.	.	.	.	.	.	.	.	.	.

- When I added in the fourth filter of ExAC_Freq <= 0.01, I obtained one significant gene – namely KLHL3 which I would be focusing my further analysis on.

Information on the gene obtained after filtering steps:

Chromosome number	**Chromosome start**	**Chromosome stop**	**Gene Name**	**CADD_phred score**	**Allele frequency in population of individual**
Chr 5	137637323	137637323	KLHL3	35	6.01E-05

Information on gnome AD

Population	**Gnome AD**
African	6.53E-05
American	2.98E-05
Finnish	0.0002
Non-Finnish Europeans	3.58E-05

Information on the damaging variant:

Item	**Description**
Gene Name	KLHL3
Protein Name	Kelch-like protein 3

Variant ID	rs199469643
Variant DNA change	Reference – C and Alternate – T
Variant protein change	Arginine to Glutamine
Variant frequency in overall human population	4.95E-05
Highest frequency and the population that contains it	0.0003 is the highest allele frequency observed in Finnish population
Other populations in which the allele is observed	6.01E-05 is the allele frequency observed in non-Finnish population

The distribution of the damaging variant across populations globally from the aggregated study is illustrated below and is obtained from NCBI reference SNP report:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	20924	C=1.00000	T=0.00000
European	Sub	15836	C=1.00000	T=0.00000
African	Sub	3324	C=1.0000	T=0.0000
African Others	Sub	114	C=1.000	T=0.000
African American	Sub	3210	C=1.0000	T=0.0000
Asian	Sub	112	C=1.000	T=0.000
East Asian	Sub	86	C=1.00	T=0.00
Other Asian	Sub	26	C=1.00	T=0.00
Latin American 1	Sub	146	C=1.000	T=0.000
Latin American 2	Sub	610	C=1.000	T=0.000

The distribution of the damaging variant across populations globally from the gnomAD – Exomes study is illustrated below and is obtained from NCBI reference SNP report:

Study	Population	Group	Samp.Size	Ref	Alt

gnomAD - Exomes	Global	Study-wide	251424	C=0.999960	T=0.000040
gnomAD - Exomes	European	Sub	135356	C=0.999941	T=0.000059
gnomAD - Exomes	Asian	Sub	49006	C=1.00000	T=0.00000
gnomAD - Exomes	American	Sub	34590	C=0.99997	T=0.00003
gnomAD - Exomes	African	Sub	16254	C=0.99994	T=0.00006
gnomAD - Exomes	Ashkenazi Jewish	Sub	10080	C=1.00000	T=0.00000
gnomAD - Exomes	Other	Sub	6138	C=1.0000	T=0.0000

Damaging variant analysis

Function of the normal gene: The Kelch-like protein 3(KLHL3) gene transcribes a protein which is known to be involved in the protein ubiquitination pathway. This pathway is generally employed for protein modification and breaking down of unwanted proteins. The KLHL3 is a part of the E3 ubiquitin ligase complex which is a part of the ubiquitin-proteasome system and helps tag damaged and excess proteins.
The proteins are tagged using molecules like ubiquitin which signal to proteasomes to help breakdown and degrade the tagged proteins. This system is important to regulate the proteins during cell division and cell growth.

Gene mutations in KLHL3 are found to cause pseudo hypoaldosteronism type2A(PHA2A), which causes high blood pressure and high potassium levels in the blood. [1,2]
While the mutation has been accounted for on ClinVar there is no assertion criterion or citation. The information about this variant has been submitted by Richard Lifton Laboratory, Yale University School of Medicine.
Additionally, over 10+ SNP data submissions to NCBI by different studies and databases including Illumina, AFFY, GNOMAD, and TOPMED between 2013 and 2019.

Other instances of damaging variants on the gene:

- The KLHL3 rs7444370 variant has been studied to be a possible protective factor in the pathogenesis of females' essential hypertension in the Chinese Han population. The study of haplotype frequency distribution of rs7444370 in EH and control groups also highlighted that the CT haplotype could have a protective effect in women [3]

- Familial hyperkalemic hypertension (FHHt) is a form of arterial hypertension that is linked to mutations in WNK1 and WNK4. Using combimed linkage analysis the KLHL3 has been studied as the thirs gene responsible for FHHt. The FHHt is known to be a complex signaling pathway that ensures ion homeostasis in the distal nephron and control blood pressure indirectly [4]

Protein visualization

Protein visualization is performed by obtaining the FASTA sequence of the protein from UNIPROT. The FASTA sequence is employed by the Swiss-model tool to provide a 3-D structure of the protein in consideration and is depicted below.

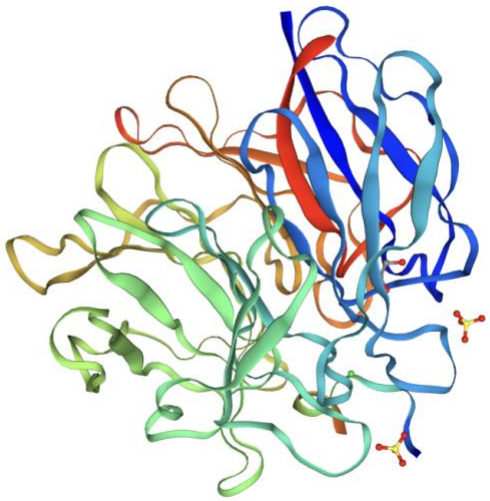


Figure 1: Structural model of the natural variant of KLHL3: The protein has 587 amino acids and three different isoforms

To obtain the variant protein, we first obtained the FASTA sequence of the variant from the pathology and biotech section of UNIPROT. This was further visualized in Swiss-model for the missense mutation variant which has Arginine substituted by glutamine highlighted by a red band pointed by a yellow arrow is represented below:

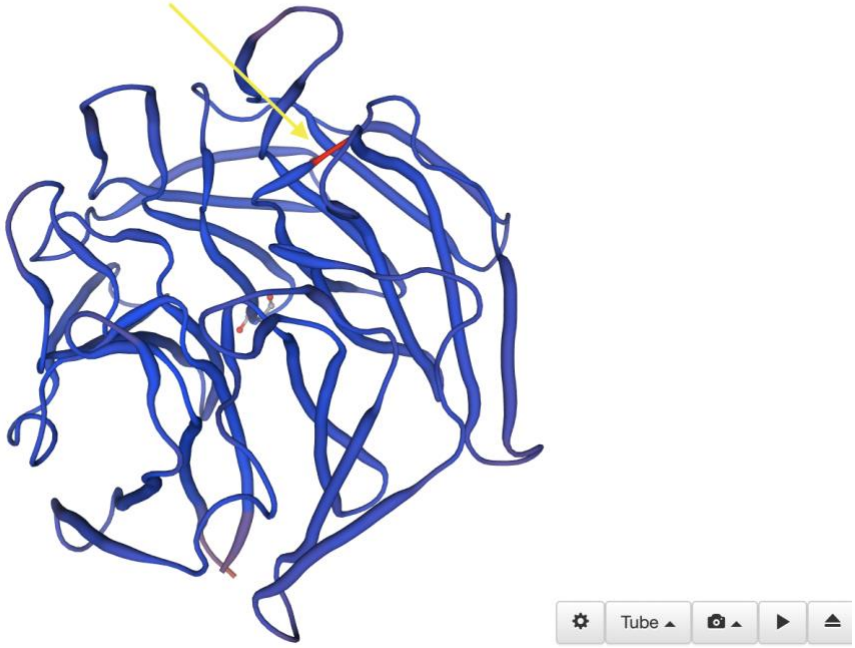


Figure 2: Mutated variant of KLHL3: The protein has 587 amino acids and is mutated at the 431st position where arginine is replaced by glutamine

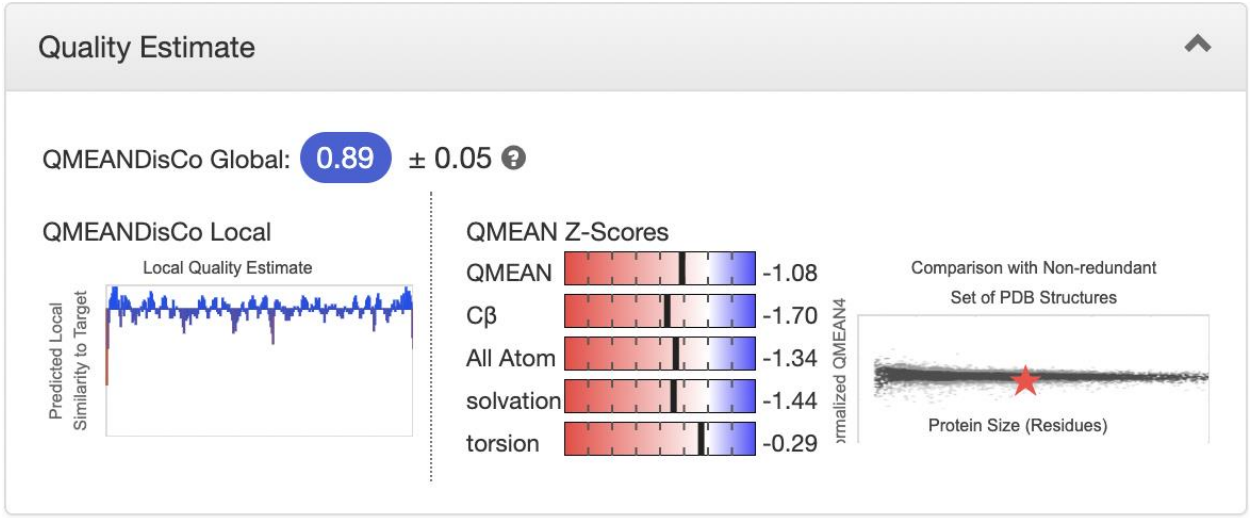


Figure 3: Table depicting quality estimate

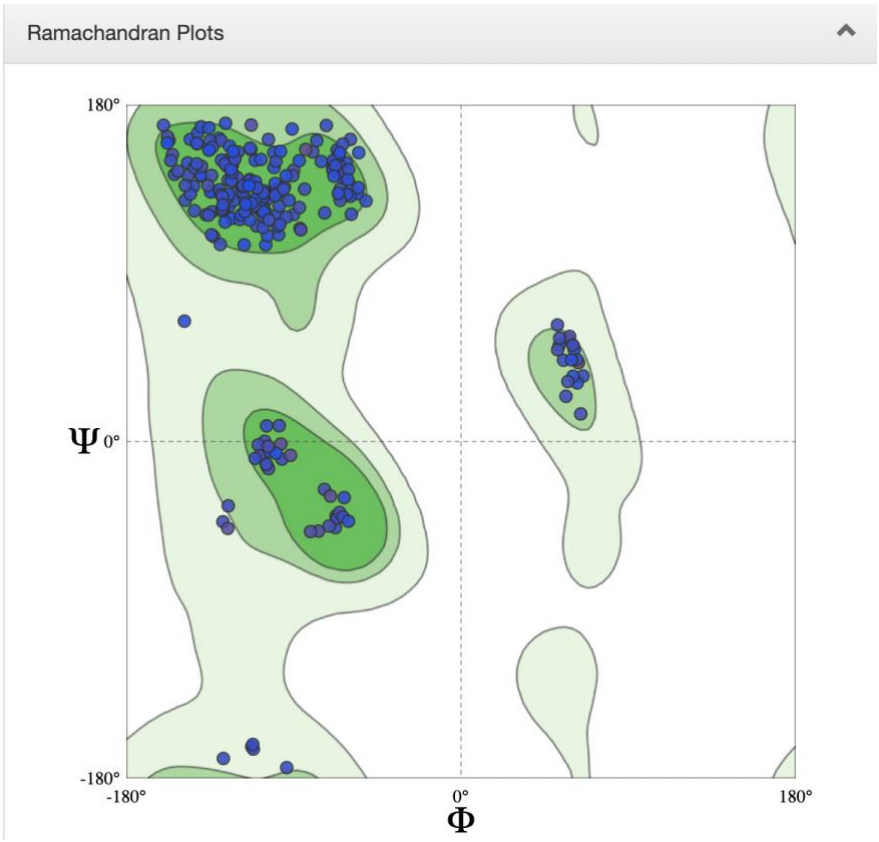


Figure 4: Table depicting Ramachandran plot of the modelled protein

The QMEANDisCo Global score is 0.89 +/- 0.05. This score can be considered as the average per-residue QMEANDisCo score which has been IDDT score. [5]
The QMEANDisCo score is measured between 0 and 1 with a higher number indicating a higher expected quality. Since the score in our case is around 0.89, we can say that this protein model has a high expected quality, and this score is calculated without coverage dependence.
The QMEAN z-score below -4 usually indicates that the model has a low quality, however for our model the z-score is -1.08 which is above the threshold declared for low quality models.
From the parameters described above along with the Ramachandran plot, it is safe to conclude that the model being depicted is reliable and the protein formed despite of the mutation at the 431st position is stable despite its deleterious nature

References:

1. Boyden LM, Choi M, Choate KA, et al. Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature*. 2012;482(7383):98-102.
2. Mori Y, Wakabayashi M, Mori T, et al. Decrease of WNK4 ubiquitination by disease-causing mutations of KLHL3 through different molecular mechanisms. *Biochem Biophys Res Commun*. 2013;439(1):30-34.
3. Li J, Hu J, Xiang D, et al. KLHL3 single-nucleotide polymorphism is associated with essential hypertension in Chinese Han population. *Medicine (Baltimore)*. 2019;98(20):e15766.
4. Louis-Dit-Picard, H., Barc, J., Trujillano, D. *et al. KLHL3* mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nat Genet* 2012; 44:456–460
5. Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, R., Haas, J., Schwede, T. QMEANDisCo - distance constraints applied on model quality estimation. Bioinformatics 2020; 36:1765-1771.