BIOL 6150 – Genomics Applied Bioinformatics

WHOLE EXOME SEQUENCE ANALYSIS OF A MALE COLOMBIAN FROM MEDELLIN, COLOMBIA

Varsha Srinivasan

INTRODUCTION

The objective of this project was to analyze the whole exome sequence (WES) of a male Colombian of American ancestry in Medellin, Colombia. The aim is to identify a rare non-synonymous pathogenic variant.

Input files

Human reference genome (hg38) obtained from the 1000 Genomes Project.

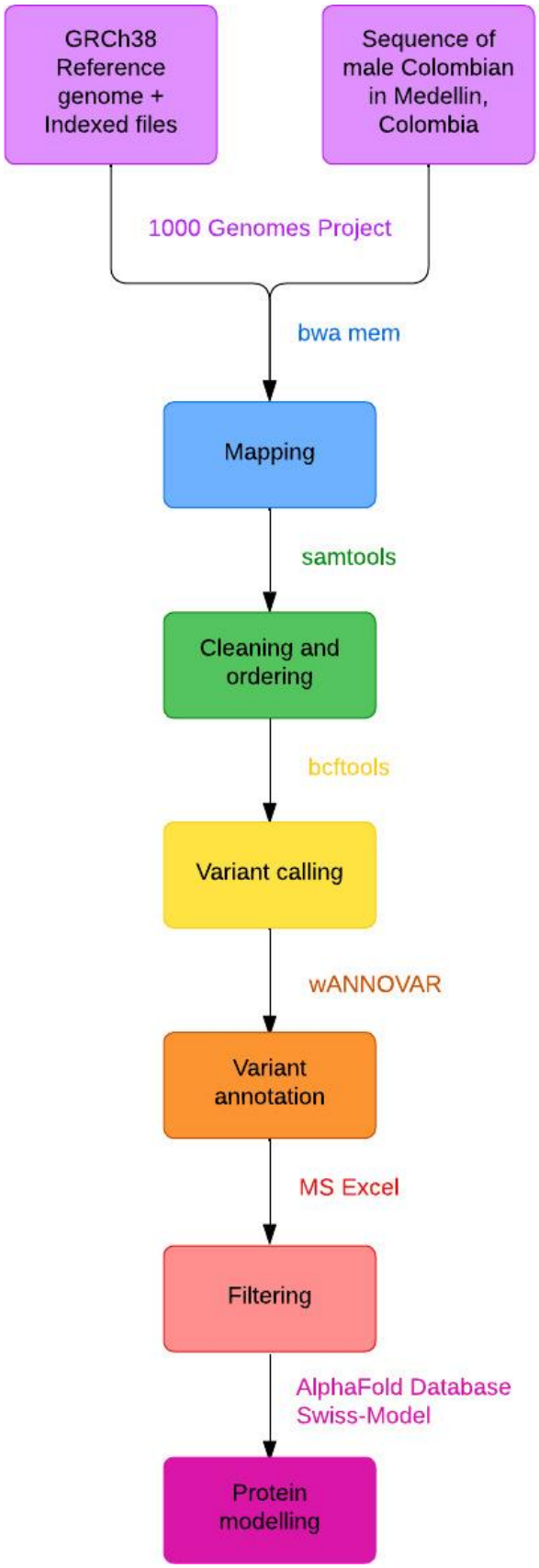
Sequence of the individual under study also obtained from the 1000 Genomes Project:

- Sample ID: HG01341
- BioSample ID: SAME124859
- Sequence ID: SRR080245
    - Forward read: SRR080245_1
    - Reverse read: SRR080245_2
- Population: Colombian in Medellin Colombia (American Ancestry)
- Sex: Male
- Cell line source: HG01341 at Coriell

WORKFLOW

1. **Downloading reference genome and sequences**
   The reference genome, corresponding index files and the sequence for the study were downloaded from the links mentioned above.
2. **Mapping using bwa mem tool in the Ubuntu terminal**
   ```
   bwa mem -M -R '@RG\tID:foo\tSM:bar\tLB:library1' GRCh38_full_analysis_set_plus_decoy_hla.fa SRR080245_1.fastq SRR080245_2.fastq > lane.sam
   ```
3. **Cleaning up read pairing information and flags using samtools on Ubuntu terminal**
   ```
   samtools fixmate -O bam lane.sam lane_fixmate.bam
   ```
4. **Sorting from name order to coordinate order using samtools on Ubuntu terminal**
   ```
   samtools sort -O bam -o lane_sorted.bam -T /tmp/lane_temp lane_fixmate.sam
   ```
5. **Variant calling using bcftools on Ubuntu terminal**
   ```
   bcftools mpileup -Ou -f GRCh38_full_analysis_set_plus_decoy_hla.fa lane_sorted.bam | bcftools call -vmO z -o SRR080245.vcf.gz
   ```
6. **Variant annotation using wANNOVAR with the following parameters:**
    a. Reference genome: hg38
    b. Gene definition: RefSeq gene
    c. Analysis: Individual
7. **Filtering the annotated variants using MS Excel with the following filters**
    a. ExonicFunction = non-synonymous SNV
    b. ExonicFunction = synonymous SNV
    c. ExonicFunction = frameshift deletion, frameshift insertion
    d. ExonicFunction = early stop
    e. ClinVar_SIG = Pathogenic, likely pathogenic
    f. ClinVar_SIG = Pathogenic, likely pathogenic, 1000G_ALL < 0.1

8. Protein modelling using the AlphaFold Database and Swiss-Model



Analytical Pipeline for WES

RESULTS AND DISCUSSION

Summary table

Genome ID: HG01341	Columbian in Medellin, Columbia (Male)
Total # of variants	8330
Total # of non-synonymous variants	4533
Total # of synonymous variants	3514
Total # of protein-truncating variants	232
Total # of frameshifts	10
Total # of clinically significant variants	5

Total # of variants = 8330

Chr	Start	End	Ref	Alt	Func.refG	Gene.refG	GeneDetai	ExonicFun	AAChange	1000G_AL	1000G_AF	1000G_AN	1000G_EA	1000G_EU	1000G_SA	ExAC_Frec	ExAC_AFR	ExAC_AM	ExAC_EAS	ExAC_FIN	ExAC_NFE	ExAC_OTH
chr1	69511	69511	A	G	exonic	OR4F5		nonsynony	OR4F5:NM_							0.9394	0.5942	0.9507	0.9994	0.9907	0.9716	0.9597
chr1	930314	930314	C	T	exonic	SAMD11		nonsynony	SAMD11:N	0.052	0.023	0.097	0.14	0.002	0.025	0.0381	0.0344	0.1602	0.1605	0.002	0.0023	0.0351
chr1	935835	935835	C	G	exonic	SAMD11		synonymo	SAMD11:N	0.053	0.02	0.094	0.14	0.004	0.029	0.0269	0.0294	0.1046	0.1138	0.0059	0.0031	0.029
chr1	961945	961945	G	C	exonic	KLHL17		nonsynony	KLHL17:NI	0.86	0.7	0.88	0.92	0.94	0.92	0.9077	0.721	0.9401	0.9079	0.9345	0.9261	0.9283
chr1	1043838	1043838	C	T	exonic	AGRN		nonsynony	AGRN:NM_													
chr1	1223251	1223251	A	G	exonic	SDF4		synonymo	SDF4:NM_	0.95	1	0.93	1	0.86	0.96	0.9166	0.9825	0.9568	0.9998	0.8569	0.8864	0.9106
chr1	1290912	1290912	G	A	exonic	SCNN1D		synonymo	SCNN1D:N	0.17	0.019	0.12	0.54	0.005	0.22	0.0959	0.0237	0.1745	0.4998	0.0081	0.0095	0.0633
chr1	1319461	1319461	C	G	exonic	INTS11		synonymo	INTS11:NM	0.71	0.3	0.85	0.84	0.92	0.83	0.8476	0.3848	0.7911	0.8258	0.8814	0.9295	0.833
chr1	1334174	1334174	T	C	exonic	TAS1R3		nonsynony	TAS1R3:NI	0.95	0.98	0.98	0.91	0.97	0.92	0.9613	0.9873	0.988	0.9128	0.9366	0.9721	0.9321
chr1	1487887	1487887	A	G	exonic	ATAD3B		synonymo	ATAD3B:N	0.49	0.85	0.34	0.63	0.099	0.36	0.2539	0.7512	0.3989	0.6383	0.2018	0.0864	0.2328
chr1	1627156	1627156	G	A	exonic	MIB2		synonymo	MIB2:NM_	0.14	0.0061	0.29	0.17	0.14	0.19	0.1822	0.0339	0.4245	0.2458	0.1998	0.1552	0.1748
chr1	1719368	1719368	T	C	exonic	CDK11A		synonymo	CDK11A:N	0.75	0.5	0.88	0.78	0.92	0.78	0.4784	0.3893	0.4877	0.4648	0.4958	0.4943	0.4844
chr1	1922670	1922670	A	G	exonic	CFAP74		synonymo	CFAP74:NI	0.6	0.56	0.62	0.64	0.54	0.66	0.5677	0.523	0.6668	0.6456	0.5492	0.5308	0.5572
chr1	1923021	1923021	C	T	exonic	CFAP74		synonymo	CFAP74:NM_001304360:exon37:c.G4647A:p.V1549V													
chr1	3021701	3021701	C	T	exonic	ACTRT2		synonymo	ACTRT2:NI	0.2	0.13	0.24	0.4	0.16	0.11	0.1941	0.151	0.3204	0.4222	0.1367	0.1753	0.1787
chr1	3385157	3385157	C	T	exonic	PRDM16		synonymo	PRDM16:N	0.34	0.039	0.33	0.62	0.27	0.54	0.3295	0.0802	0.4145	0.5966	0.3095	0.2706	0.3158
chr1	3414629	3414629	G	A	exonic	PRDM16		synonymo	PRDM16:N	0.018	0.0015	0.099	0.019	.	.	0.0136	0.0017	0.1097	0.0348	0	0.0002	0.0102
chr1	3473163	3473163	C	T	exonic	ARHGEF16		nonsynony	ARHGEF16	0.68	0.39	0.76	0.67	0.86	0.86	0.805	0.4548	0.6769	0.6607	0.9028	0.8731	0.8497
chr1	3477892	3477892	T	C	exonic	ARHGEF16		synonymo	ARHGEF16	0.97	0.9	0.99	1	1	1	0.9903	0.9036	0.9943	1	1	0.9986	0.9944
chr1	3493792	3493792	A	T	exonic	MEGF6		nonsynony	MEGF6:NM_													
chr1	3511596	3511596	T	G	exonic	MEGF6		synonymo	MEGF6:NM	0.5	0.75	0.4	0.64	0.29	0.3	0.3713	0.6994	0.468	0.6653	0.2605	0.2988	0.32
chr1	3628476	3628476	G	T	exonic	TPRG1L		nonsynony	TPRG1L:NI													
chr1	3632268	3632268	G	C	exonic	WRAP73		nonsynony	WRAP73:N	0.049	0.035	0.14	0.002	0.055	0.047	0.0673	0.0301	0.194	0.0007	0.0475	0.0642	0.0673
chr1	3707776	3707776	G	A	exonic	TP73		synonymo	TP73:NM_001126240:exon2:c.G267A:p.K89K,TP73:NM_001126241:exon2:c.G267A:p.K89K,TP73:NM_001126242:exon2:c.G267A:p.K89K,TP73:NM_00120													
chr1	3752641	3752641	C	G	exonic	CCDC27		nonsynony	CCDC27:N	0.39	0.4	0.41	0.29	0.39	0.45	0.3773	0.3998	0.4157	0.2882	0.3728	0.3702	0.3756

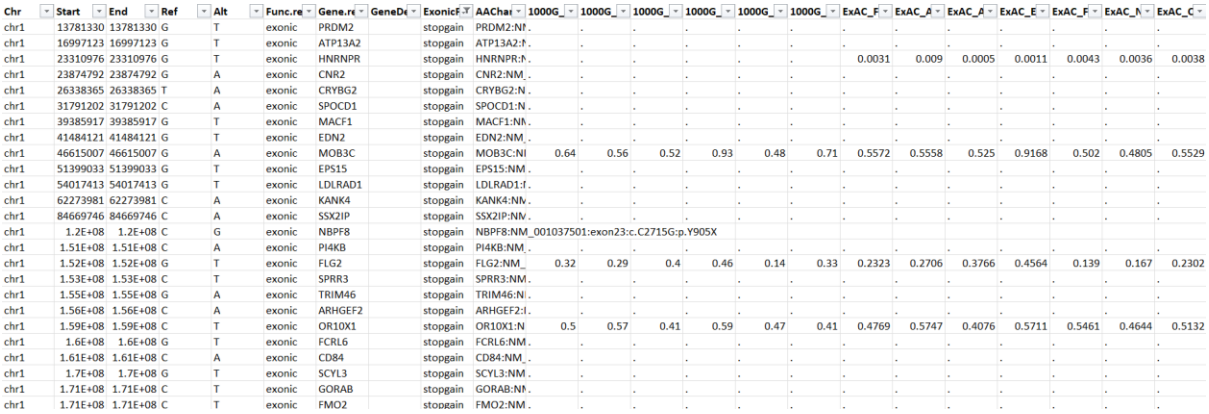
Total # of non-synonymous variants = 4533

Chr	Start	End	Ref	Alt	Func.re	Gene.re	GeneDe	ExonicF	AAChar	1000G_	1000G_	1000G_	1000G_	1000G_	1000G_	ExAC_F	ExAC_A	ExAC_A	ExAC_E	ExAC_F	ExAC_N	ExAC_C
chr1	69511	69511	A	G	exonic	OR4F5		nonsynony	OR4F5:NM_							0.9394	0.5942	0.9507	0.9994	0.9907	0.9716	0.9597
chr1	930314	930314	C	T	exonic	SAMD11		nonsynony	SAMD11:N	0.052	0.023	0.097	0.14	0.002	0.025	0.0381	0.0344	0.1602	0.1605	0.002	0.0023	0.0351
chr1	1043838	1043838	C	T	exonic	AGRN		nonsynony	AGRN:NM_													
chr1	1334174	1334174	T	C	exonic	TAS1R3		nonsynony	TAS1R3:NI	0.95	0.98	0.98	0.91	0.97	0.92	0.9613	0.9873	0.988	0.9128	0.9366	0.9721	0.9321
chr1	3473163	3473163	C	T	exonic	ARHGEF16		nonsynony	ARHGEF16	0.68	0.39	0.76	0.67	0.86	0.86	0.805	0.4548	0.6769	0.6607	0.9028	0.8731	0.8497
chr1	3493792	3493792	A	T	exonic	MEGF6		nonsynony	MEGF6:NM_													
chr1	3628476	3628476	G	T	exonic	TPRG1L		nonsynony	TPRG1L:NI													
chr1	3632268	3632268	G	C	exonic	WRAP73		nonsynony	WRAP73:N	0.049	0.035	0.14	0.002	0.055	0.047	0.0673	0.0301	0.194	0.0007	0.0475	0.0642	0.0673
chr1	3752641	3752641	C	G	exonic	CCDC27		nonsynony	CCDC27:N	0.39	0.4	0.41	0.29	0.39	0.45	0.3773	0.3998	0.4157	0.2882	0.3728	0.3702	0.3756
chr1	3836572	3836572	A	T	exonic	CEP104		nonsynony	CEP104:NI	0.3	0.06	0.49	0.34	0.43	0.33	0.3698	0.1096	0.4091	0.3225	0.3356	0.4291	0.3491
chr1	3891029	3891029	G	C	exonic	C1orf174		nonsynony	C1orf174:I	1	1	1	1	0.99	1	0.9941	0.9988	0.998	0.9999	0.9941	0.9912	0.9978
chr1	6124032	6124032	A	G	exonic	CHD5		nonsynony	CHD5:NM_	0.73	0.61	0.6	0.99	0.6	0.83	0.657	0.6074	0.604	0.9969	0.7162	0.5919	0.6765
chr1	6124092	6124092	G	A	exonic	CHD5		nonsynony	CHD5:NM_													
chr1	6128966	6128966	C	A	exonic	CHD5		nonsynony	CHD5:NM_													
chr1	6225193	6225193	T	G	exonic	ICMT		nonsynony	ICMT:NM_													
chr1	6349866	6349866	C	A	exonic	ACOT7		nonsynony	ACOT7:NM													
chr1	6469216	6469216	T	A	exonic	PLEKHG5		nonsynony	PLEKHG5:I													
chr1	6633037	6633037	A	G	exonic	THAP3		nonsynony	THAP3:NM	0.37	0.31	0.34	0.39	0.33	0.49	0.3911	0.3605	0.3266	0.4572	0.3063	0.3782	0.3925
chr1	7737443	7737443	C	G	exonic	CAMTA1		nonsynony	CAMTA1:N	0.1	0.014	0.21	0.077	0.12	0.15	0.1167	0.0331	0.1688	0.0768	0.0945	0.1232	0.1203
chr1	7827298	7827298	G	T	exonic	PER3		nonsynony	PER3:NM_													
chr1	7827433	7827433	T	C	exonic	PER3		nonsynony	PER3:NM_	0.98	0.98	0.97	1	0.95	0.98	0.9688	0.9724	0.9892	0.999	0.9645	0.9591	0.9769
chr1	7830057	7830057	T	C	exonic	PER3		nonsynony	PER3:NM_	0.19	0.097	0.29	0.051	0.29	0.26	0.2531	0.1146	0.2469	0.0538	0.3247	0.2903	0.2852
chr1	7837073	7837073	A	G	exonic	PER3		nonsynony	PER3:NM_	0.12	0.071	0.21	0.0079	0.2	0.16	0.1659	0.0907	0.1876	0.0073	0.2483	0.1839	0.1674
chr1	9263851	9263851	G	A	exonic	H6PD		nonsynony	H6PD:NM_	0.39	0.49	0.37	0.48	0.25	0.31	0.2851	0.462	0.336	0.4645	0.229	0.233	0.2622
chr1	9731862	9731862	T	C	exonic	CLSTN1		nonsynony	CLSTN1:NI													

Total # of synonymous variants = 3514

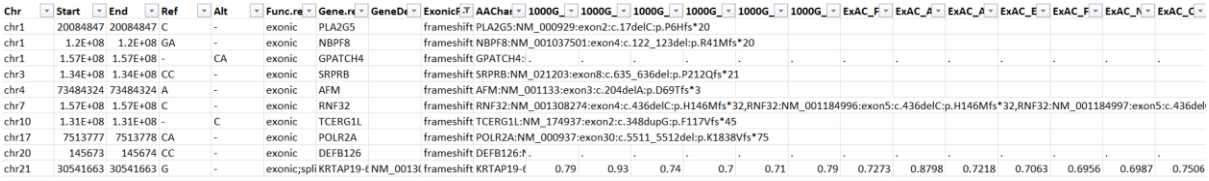
Chr	Start	End	Ref	Alt	Func.re	Gene.re	GeneDe	ExonicF	AAChar	1000G_	1000G_	1000G_	1000G_	1000G_	1000G_	ExAC_F	ExAC_A	ExAC_A	ExAC_E	ExAC_F	ExAC_N	ExAC_C
chr1	935835	935835	C	G	exonic	SAMD11		synonymo	SAMD11:N	0.053	0.02	0.094	0.14	0.004	0.029	0.0269	0.0294	0.1046	0.1138	0.0059	0.0031	0.029
chr1	961945	961945	G	C	exonic	KLHL17		synonymo	KLHL17:NI	0.86	0.7	0.88	0.92	0.94	0.92	0.9077	0.721	0.9401	0.9079	0.9345	0.9261	0.9283
chr1	1223251	1223251	A	G	exonic	SDF4		synonymo	SDF4:NM_	0.95	1	0.93	1	0.86	0.96	0.9166	0.9825	0.9568	0.9998	0.8569	0.8864	0.9106
chr1	1290912	1290912	G	A	exonic	SCNN1D		synonymo	SCNN1D:N	0.17	0.019	0.12	0.54	0.005	0.22	0.0959	0.0237	0.1745	0.4998	0.0081	0.0095	0.0633
chr1	1319461	1319461	C	G	exonic	INTS11		synonymo	INTS11:NM	0.71	0.3	0.85	0.84	0.92	0.83	0.8476	0.3848	0.7911	0.8258	0.8814	0.9295	0.833
chr1	1487887	1487887	A	G	exonic	ATAD3B		synonymo	ATAD3B:N	0.49	0.85	0.34	0.63	0.099	0.36	0.2539	0.7512	0.3989	0.6383	0.2018	0.0864	0.2328
chr1	1627156	1627156	G	A	exonic	MIB2		synonymo	MIB2:NM_	0.14	0.0061	0.29	0.17	0.14	0.19	0.1822	0.0339	0.4245	0.2458	0.1998	0.1552	0.1748
chr1	1719368	1719368	T	C	exonic	CDK11A		synonymo	CDK11A:N	0.75	0.5	0.88	0.78	0.92	0.78	0.4784	0.3893	0.4877	0.4648	0.4958	0.4943	0.4844
chr1	1922670	1922670	A	G	exonic	CFAP74		synonymo	CFAP74:NI	0.6	0.56	0.62	0.64	0.54	0.66	0.5677	0.523	0.6668	0.6456	0.5492	0.5308	0.5572
chr1	1923021	1923021	C	T	exonic	CFAP74		synonymo	CFAP74:NM_001304360:exon37:c.G4647A:p.V1549V													
chr1	3021701	3021701	C	T	exonic	ACTRT2		synonymo	ACTRT2:NI	0.2	0.13	0.24	0.4	0.16	0.11	0.1941	0.151	0.3204	0.4222	0.1367	0.1753	0.1787
chr1	3385157	3385157	C	T	exonic	PRDM16		synonymo	PRDM16:N	0.34	0.039	0.33	0.62	0.27	0.54	0.3295	0.0802	0.4145	0.5966	0.3095	0.2706	0.3158
chr1	3414629	3414629	G	A	exonic	PRDM16		synonymo	PRDM16:N	0.018	0.0015	0.099	0.019	.	.	0.0136	0.0017	0.1097	0.0348	0	0.0002	0.0102
chr1	3477892	3477892	T	C	exonic	ARHGEF16		synonymo	ARHGEF16	0.97	0.9	0.99	1	1	1	0.9903	0.9036	0.9943	1	1	0.9986	0.9944
chr1	3511596	3511596	T	G	exonic	MEGF6		synonymo	MEGF6:NM	0.5	0.75	0.4	0.64	0.29	0.3	0.3713	0.6994	0.468	0.6653	0.2605	0.2988	0.32
chr1	3707776	3707776	G	A	exonic	TP73		synonymo	TP73:NM_001126240:exon2:c.G267A:p.K89K,TP73:NM_001126241:exon2:c.G267A:p.K89K,TP73:NM_001126242:exon2:c.G267A:p.K89K,TP73:NM_0012(													
chr1	3839111	3839111	T	C	exonic	CEP104		synonymo	CEP104:NI	0.79	0.92	0.87	0.48	0.94	0.72	0.8584	0.9075	0.7321	0.4837	0.9061	0.9496	0.8733
chr1	4711993	4711993	T	C	exonic	AJAP1		synonymo	AJAP1:NM	0.53	0.52	0.5	0.66	0.37	0.58	0.4616	0.5168	0.6554	0.668	0.3899	0.3781	0.4128
chr1	6254762	6254762	T	C	exonic	GPR153		synonymo	GPR153:N	0.73	0.81	0.8	0.62	0.71	0.71	0.7092	0.7926	0.8205	0.5927	0.69	0.6961	0.7195
chr1	6644575	6644575	A	G	exonic	DNAJC11		synonymo	DNAJC11:I	0.83	0.87	0.96	0.57	0.92	0.87	0.8836	0.8781	0.9655	0.5977	0.8741	0.9093	0.8612
chr1	7745908	7745908	C	A	exonic	CAMTA1		synonymo	CAMTA1:N													
chr1	8329828	8329828	G	A	exonic	SLC45A1		synonymo	SLC45A1:N	0.24	0.54	0.29	0.17	0.33	0.2	0.2144	0.5127	0.2599	0.1689	0.2096	0.2238	0.2114

Total # of protein-truncating variants = 232



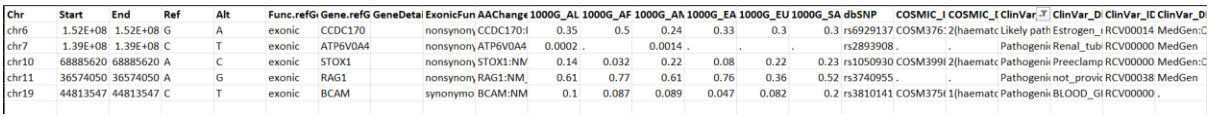
Chr	Start	End	Ref	Alt	Func.re	Gene.re	GeneDe	ExonicF	AAChar	1000G_	1000G_	1000G_	1000G_	1000G_	1000G_	ExAC_F	ExAC_A	ExAC_A	ExAC_E	ExAC_F	ExAC_N	ExAC_C
chr1	13781330	13781330	G	T	exonic	PRDM2		stopgain	PRDM2:NI	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	16997123	16997123	G	T	exonic	ATP13A2		stopgain	ATP13A2:I	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	23310976	23310976	G	T	exonic	HNRNPR		stopgain	HNRNPR:I	.	.	.	.	.	.	0.0031	0.009	0.0005	0.0011	0.0043	0.0036	0.0038
chr1	23874792	23874792	G	A	exonic	CNR2		stopgain	CNR2:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	26338365	26338365	T	A	exonic	CRYBG2		stopgain	CRYBG2:N	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	31791202	31791202	C	A	exonic	SPOCD1		stopgain	SPOCD1:N	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	39385917	39385917	G	T	exonic	MACF1		stopgain	MACF1:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	41484121	41484121	G	T	exonic	EDN2		stopgain	EDN2:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	46615007	46615007	G	A	exonic	MOB3C		stopgain	MOB3C:NI	0.64	0.56	0.52	0.93	0.48	0.71	0.5572	0.5558	0.525	0.9168	0.502	0.4805	0.5529
chr1	51399033	51399033	G	T	exonic	EPS15		stopgain	EPS15:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	54017413	54017413	G	T	exonic	LDLRAD1		stopgain	LDLRAD1:I	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	62273981	62273981	C	A	exonic	KANK4		stopgain	KANK4:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	84669746	84669746	C	A	exonic	SSX2IP		stopgain	SSX2IP:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.2E+08	1.2E+08	C	G	exonic	NBPF8		stopgain	NBPF8:NM_001037501:exon23:c.C2715G:p.Y905X													
chr1	1.51E+08	1.51E+08	C	A	exonic	PI4KB		stopgain	PI4KB:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.52E+08	1.52E+08	G	T	exonic	FLG2		stopgain	FLG2:NM_	0.32	0.29	0.4	0.46	0.14	0.33	0.2323	0.2706	0.3766	0.4564	0.139	0.167	0.2302
chr1	1.53E+08	1.53E+08	C	T	exonic	SPRR3		stopgain	SPRR3:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.55E+08	1.55E+08	G	A	exonic	TRIM46		stopgain	TRIM46:NI	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.56E+08	1.56E+08	C	A	exonic	ARHGEF2		stopgain	ARHGEF2:I	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.59E+08	1.59E+08	C	T	exonic	OR10X1		stopgain	OR10X1:N	0.5	0.57	0.41	0.59	0.47	0.41	0.4769	0.5747	0.4076	0.5711	0.5461	0.4644	0.5132
chr1	1.6E+08	1.6E+08	G	T	exonic	FCRL6		stopgain	FCRL6:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.61E+08	1.61E+08	C	A	exonic	CD84		stopgain	CD84:NM_	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.7E+08	1.7E+08	G	T	exonic	SCYL3		stopgain	SCYL3:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.71E+08	1.71E+08	C	T	exonic	GORAB		stopgain	GORAB:NM	.	.	.	.	.	.	.	.	.	.	.	.	.
chr1	1.71E+08	1.71E+08	C	T	exonic	FMO2		stopgain	FMO2:NM	.	.	.	.	.	.	.	.	.	.	.	.	.

Total # of frameshifts = 10



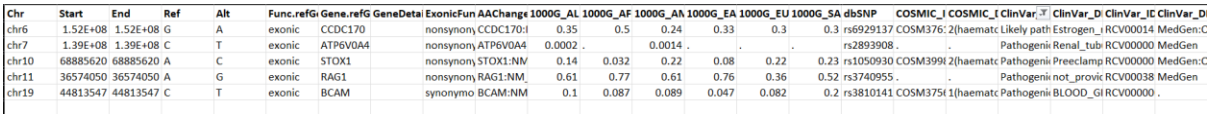
Chr	Start	End	Ref	Alt	Func.re	Gene.re	GeneDe	ExonicF	AAChar	1000G_	1000G_	1000G_	1000G_	1000G_	1000G_	ExAC_F	ExAC_A	ExAC_A	ExAC_E	ExAC_F	ExAC_N	ExAC_C
chr1	20084847	20084847	C	-	exonic	PLA2G5		frameshift	PLA2G5:NM_000929:exon2:c.17delC:p.P6Hfs*20													
chr1	1.2E+08	1.2E+08	GA	-	exonic	NBPF8		frameshift	NBPF8:NM_001037501:exon4:c.122_123delA:p.R41Mfs*20													
chr1	1.57E+08	1.57E+08	-	CA	exonic	GPATCH4		frameshift	GPATCH4:	.	.	.	.	.	.	.	.	.	.	.	.	.
chr3	1.34E+08	1.34E+08	CC	-	exonic	SRPRB		frameshift	SRPRB:NM_021203:exon8:c.635_636del:p.P212Qfs*21													
chr4	73484324	73484324	A	-	exonic	AFM		frameshift	AFM:NM_001133:exon3:c.204delA:p.D69Tfs*3													
chr7	1.57E+08	1.57E+08	C	-	exonic	RNF32		frameshift	RNF32:NM_001308274:exon4:c.436delC:p.H146Mfs*32,RNF32:NM_001184996:exon5:c.436delC:p.H146Mfs*32,RNF32:NM_001184997:exon5:c.436del													
chr10	1.31E+08	1.31E+08	-	C	exonic	TCERG1L		frameshift	TCERG1L:NM_174937:exon2:c.348dupG:p.F117Vfs*45													
chr17	7513777	7513778	CA	-	exonic	POLR2A		frameshift	POLR2A:NM_000937:exon30:c.5511_5512del:p.K1838Vfs*75													
chr20	145673	145674	CC	-	exonic	DEFB126		frameshift	DEFB126:I	.	.	.	.	.	.	.	.	.	.	.	.	.
chr21	30541663	30541663	G	-	exonic;spli	KRTAP19-(	NM_0013(	frameshift	KRTAP19-(	0.79	0.93	0.74	0.7	0.71	0.79	0.7273	0.8798	0.7218	0.7063	0.6956	0.6987	0.7506

Total # of pathogenic variants = 5



Chr	Start	End	Ref	Alt	Func.refG	Gene.refG	GeneDetai	ExonicFun	AAChange	1000G_AL	1000G_AF	1000G_AM	1000G_EA	1000G_EU	1000G_SA	dbSNP	COSMIC_I	COSMIC_I	ClinVar_	ClinVar_D	ClinVar_ID	ClinVar_DI
chr6	1.52E+08	1.52E+08	G	A	exonic	CCDC170		nonsynony	CCDC170:I	0.35	0.5	0.24	0.33	0.3	0.3	rs6929137	COSM376:	2(haemato	Likely path	Estrogen_I	RCV00014	MedGen:C
chr7	1.39E+08	1.39E+08	C	T	exonic	ATP6V0A4		nonsynony	ATP6V0A4	0.0002	.	0.0014	.	.	.	rs2893908	.	.	Pathogeni	Renal_tub	RCV00000	MedGen
chr10	68885620	68885620	A	C	exonic	STOX1		nonsynony	STOX1:NM	0.14	0.032	0.22	0.08	0.22	0.23	rs1050930	COSM3991	2(haemato	Pathogeni	Preeclamp	RCV00000	MedGen:C
chr11	36574050	36574050	A	G	exonic	RAG1		nonsynony	RAG1:NM_	0.61	0.77	0.61	0.76	0.36	0.52	rs3740955	.	.	Pathogeni	not_provic	RCV00038	MedGen
chr19	44813547	44813547	C	T	exonic	BCAM		synonymo	BCAM:NM	0.1	0.087	0.089	0.047	0.082	0.2	rs3810141	COSM3756	1(haemato	Pathogeni	BLOOD_GI	RCV00000	.

Pathogenic variant with RAF < 0.1 = 1



Chr	Start	End	Ref	Alt	Func.refG	Gene.refG	GeneDetai	ExonicFun	AAChange	1000G_AL	1000G_AF	1000G_AM	1000G_EA	1000G_EU	1000G_SA	dbSNP	COSMIC_I	COSMIC_I	ClinVar_	ClinVar_D	ClinVar_ID	ClinVar_DI
chr6	1.52E+08	1.52E+08	G	A	exonic	CCDC170		nonsynony	CCDC170:I	0.35	0.5	0.24	0.33	0.3	0.3	rs6929137	COSM376:	2(haemato	Likely path	Estrogen_I	RCV00014	MedGen:C
chr7	1.39E+08	1.39E+08	C	T	exonic	ATP6V0A4		nonsynony	ATP6V0A4	0.0002	.	0.0014	.	.	.	rs2893908	.	.	Pathogeni	Renal_tub	RCV00000	MedGen
chr10	68885620	68885620	A	C	exonic	STOX1		nonsynony	STOX1:NM	0.14	0.032	0.22	0.08	0.22	0.23	rs1050930	COSM3991	2(haemato	Pathogeni	Preeclamp	RCV00000	MedGen:C
chr11	36574050	36574050	A	G	exonic	RAG1		nonsynony	RAG1:NM_	0.61	0.77	0.61	0.76	0.36	0.52	rs3740955	.	.	Pathogeni	not_provic	RCV00038	MedGen
chr19	44813547	44813547	C	T	exonic	BCAM		synonymo	BCAM:NM	0.1	0.087	0.089	0.047	0.082	0.2	rs3810141	COSM3756	1(haemato	Pathogeni	BLOOD_GI	RCV00000	.

Pathogenic variant information

- Chromosome: chr7
- Gene: ATP6V0A4
- Variant ID: rs28939081
- ClinVar ID: RCV000005466.4
- Exonic function: Non-synonymous SNV
- DNA change: Guanine to adenine (G -> A) at position 2420
- Protein change: Arginine to glutamine (R -> Q) at position 807
- Variant frequency:
  - 1000 Genomes (All) = 0.0002
  - 1000 Genomes (American) = 0.0014
  - 1000 Genomes (African, European, East Asian, South Asian) = 0
- The highest frequency is in the American population (0.0014)
- Allele frequency:
  - Using hg19 as reference:
    - Total = 0.00003979
    - Latino/Admixed American = 0.0002540

    - Using hg38 as reference:
        - Total = 0.00001316
        - Latino/Admixed American = 0.00006576
- Popmax filtering allele frequency (95% confidence):
    - Using hg19 as reference = 0.0001348
    - Using hg38 as reference = 0
- SIFT score: 0
- Polyphen2_HDIV score: 1
- Polyphen2_HVAR score: 0.999
- CADD_raw: 7.735
- CADD_phred: 35

Pathogenic variant analysis

The gene ATP6V0A4 stands for ATPase H+ Transporting V0 Subunit A4. Vacuolar H+-ATPases are involved in a variety of cellular functions in the kidney, such as protein endocytosis from the lumen in the proximal tubule [1], recycling of transport proteins such as the water channel AQP-2 [2], insertion of vesicles into the membrane, and lysosomal degradation of proteins. In humans, it is part of the proton channel of the V-ATPase that is involved in normal vectorial acid transport into the urine by the kidney. [3]

This gene encodes a component of vacuolar ATPase (V-ATPase), a multisubunit enzyme that mediates acidification of intracellular compartments of eukaryotic cells. V-ATPase dependent acidification is necessary for such intracellular processes as protein sorting, zymogen activation, receptor-mediated endocytosis, and synaptic vesicle proton gradient generation. V-ATPase is composed of a cytosolic V1 domain and a transmembrane V0 domain. The V0 domain consists of five different subunits: a, c, c', c", and d. This gene is one of four genes in man and mouse that encode different isoforms of the a subunit. Alternatively spliced transcript variants encoding the same protein have been described. [4]

Mutations in the accessory ATP6V0A4 (a4 isoform) subunit have recently been shown to cause an inherited form of distal renal tubular acidosis (dRTA) in humans. [5] This form of dRTA may also be associated with sensorineural deafness in some cases. [6]

Previous citations

This mutation has been cited in a number of papers, a few of which have been referenced below. Stover et. al. identified mutations in genes encoding two different subunits of the renal α-intercalated cell's apical H+-ATPase that cause dRTA. It also showed that ATP6V0A4 is expressed within the human inner ear. These findings provide further evidence for genetic heterogeneity in rdRTA, extend the spectrum of disease-causing mutations in ATP6V1B1 and ATP6V0A4, and show ATP6V0A4 expression within the cochlea for the first time. [6]

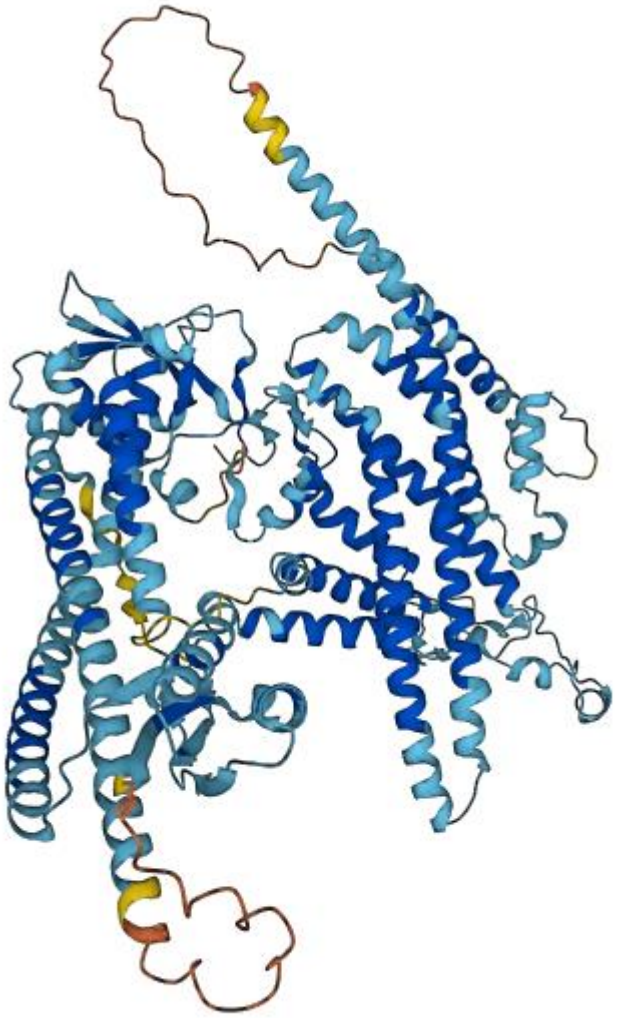
Stehberger et. al., examined the localization of the a4 subunit in mouse and human kidney and, using antigen retrieval techniques, demonstrated expression in most nephron segments. [5]

This mutation was observed in a consanguineous Tunisian family suffering from dRTA. The diagnosis of dRTA in the child described in this study was suspected by clinical features like dehydration with polyuria, non-gap severe metabolic acidosis and inability to renal

acidification, as shown by alkaline urine, hypokalemia and nephrocalcinosis. Their parents were also carrying this mutation at the heterozygous state. [7]

The hereditary forms of dRTA have received increased attention due to the advances in the understanding of the molecular mechanism, whereby mutations in the main proteins involved in acid–base transport result in impaired acid excretion. Dysfunction of intercalated cells in the collecting tubules accounts for all the known genetic causes of dRTA. [7]

Structural model of the protein



The structural model of ATPase H+ Transporting V0 Subunit A4 (ATP6V0A4) obtained using AlphaFold Database

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

The colors indicate the model confidence.

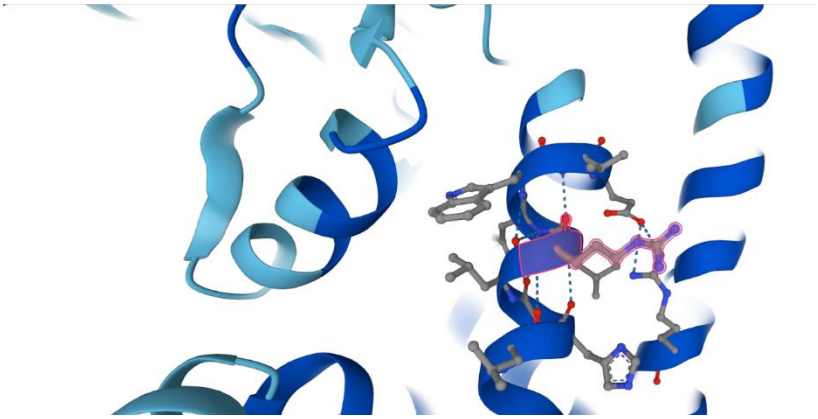
- Dark blue - Very high (pLDDT > 90)

- Light blue - Confident (70 < pLDDT < 90)
- Yellow - Low (50 < pLDDT < 70)
- Orange - Very low (pLDDT < 50)

Visualization of the altered amino acid on the protein model

Model of the protein highlighting (in pink) the region where the substitution occurs (807)

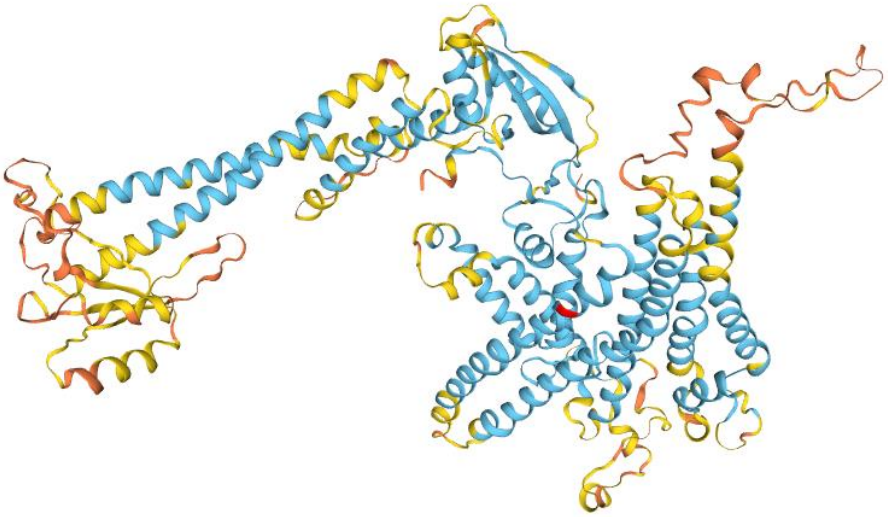
This position (807) has a pLDDT score (for 1 residue) of 94.22, which falls under the category of “very high”. This is considered to be a very good score. The regions surrounding this site also have pLDDT values over 90. This makes the structural model obtained reliable.



Zoomed in view of the original amino acid (R) at position 807

Arginine is a positively charged amino acid, whereas glutamine is a charge-neutral polar amino acid. Arginine generally prefers to be on the surface of the protein, but its amphipathic nature

can mean that part of the side chain is buried. Since the change happens at one position and is not protein truncating, the protein structure does not get altered significantly. The substitution of a positive amino acid with a neutral one is likely to affect amino acid chain reactions and affect the secondary and tertiary structure of the protein.



Model of the altered protein obtained using Swiss-Model highlighting (in red) the position where the substitution has occurred (807)

The confidence level is 0.75 at position 807 in the altered structural model. So it is considered confident. QMEANDisCo global score is the average per-residue QMEANDisCo score [8] which has been found to correlate well with the lDDT score. [9] The provided error estimate is based on QMEANDisCo global scores estimated for a large set of models and represents the root mean squared difference (i.e. standard deviation) between QMEANDisCo global score and lDDT (the ground truth). As the reliability of the prediction depends on model size, the provided error estimate is calculated based on models of similar size to the input. [10] Here, the QMEANDisCo global score is 0.66 ± 0.05, which indicates that this model is structurally sound. This is because the reliability of the model increases the closer this value gets to 1.

There is not a lot of information available in literature about this particular amino acid substitution. However, it has been reported in the fourth module of Xenopus interphotoreceptor retinoid-binding protein in *E. coli* systems. [11]

REFERENCES

[1] Marshansky V, Ausiello DA, Brown D: Physiological importance of endosomal acidification: Potential role in proximal tubulopathies. *Curr Opin Nephrol Hypertens* 11: 527–537, 2002

[2] Gustafson CE, Katsura T, McKee M, Bouley R, Casanova JE, Brown D: Recycling of AQP2 occurs through a temperature- and bafilomycin-sensitive trans-Golgi-associated compartment. *Am J Physiol Renal Physiol* 278: F317–F326, 2000

[3] https://www.phosphosite.org/proteinAction?id=15061&showAllSites=true

[4] https://www.ncbi.nlm.nih.gov/gene/50617#summary

[5] Stehberger, P. A., Schulz, N., Finberg, K. E., Karet, F. E., Giebisch, G., Lifton, R. P., Geibel, J. P., & Wagner, C. A. (2003). Localization and regulation of the ATP6V0A4 (A4) vacuolar H+-atpase subunit defective in an inherited form of distal renal tubular acidosis. *Journal of the American Society of Nephrology, 14*(12), 3027–3038. https://doi.org/10.1097/01.asn.0000099375.74789.ab

[6] Stover EH, Borthwick KJ, Bavalia C, Eady N, Fritz DM, Rungroj N, Giersch AB, Morton CC, Axon PR, Akil I, Al-Sabban EA, Baguley DM, Bianca S, Bakkaloglu A, Bircan Z, Chauveau D, Clermont MJ, Guala A, Hulton SA, Kroes H, Li Volti G, Mir S, Mocan H, Nayir A, Ozen S, Rodriguez Soriano J, Sanjad SA, Tasic V, Taylor CM, Topaloglu R, Smith AN, Karet FE: Novel ATP6V1B1 and ATP6V0A4 mutations in autosomal recessive distal renal tubular acidosis with new evidence for hearing loss. *J Med Genet 39*: 796–803, 2002 http://dx.doi.org/10.1136/jmg.39.11.796

[7] Nagara, M., Voskarides, K., Elouej, S., Zaravinos, A., Riahi, Z., Papagregoriou, G., Kefi, R., Boussetta, K., Deltas, C., Abdelhak, S., & Tinsa, F. (2014). A novel splice-site mutation in ATP6V0A4 gene in two brothers with distal renal tubular acidosis from a consanguineous Tunisian family. *Journal of genetics, 93*(3), 859–863. https://doi.org/10.1007/s12041-014-0450-4

[8] Gabriel Studer, Christine Rempfer, Andrew M Waterhouse, Rafal Gumienny, Juergen Haas, Torsten Schwede, QMEANDisCo—distance constraints applied on model quality estimation, *Bioinformatics*, Volume 36, Issue 6, 15 March 2020, Pages 1765–1771, https://doi.org/10.1093/bioinformatics/btz828

[9] Valerio Mariani, Marco Biasini, Alessandro Barbato, Torsten Schwede, lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests, *Bioinformatics*, Volume 29, Issue 21, 1 November 2013, Pages 2722–2728, https://doi.org/10.1093/bioinformatics/btt473

[10] https://swissmodel.expasy.org/docs/help#model_evaluation

[11] Baer, C. A., Van Niel, E. E., Cronk, J. W., Kinter, M. T., Sherman, N. E., Braiman, M. S., & Gonzalez-Fernandez, F. (1998). *Arginine to glutamine substitutions in the fourth module of Xenopus interphotoreceptor retinoid-binding protein. Molecular vision, 4*, 30.