Whole exome sequence analysis

Individual sequence ID and Info

Sequence ID- HG02696
Sample Run ID- SRR597261
Gender: male
Populations: Punjabi in Lahore, Pakistan, South Asian Ancestry
Biosample ID:SAME1839149
Cell line source: HG02696 at Coriell

Methodology:

This methodology describes a process for aligning and variant calling on a whole exome sequencing (WES) data set from a single individual using the bwa mem aligner, samtools, and bcftools. The resulting variant calls are then annotated using the wANNOVAR online tool.

The first step is to download the reference genome and bwa indexed files from the National Center for Biotechnology Information (NCBI) website. The reference genome is the GRCh38 no alt analysis set, which is the reference sequence for the human genome (build 38). The bwa indexed files are used by the bwa mem aligner to efficiently align the reads from the WES data to the reference genome.

Next, the WES data for the individual with the sequence ID "HG02696" is downloaded from the 1000 Genomes website. The reads from this data are then mapped to the reference genome using the bwa mem aligner. The resulting alignment file is in the SAM format, which is then cleaned up and sorted using samtools.

In the improvement step, the sorted alignment file is indexed using samtools to make it more efficient to work with. Variant calling is then performed using bcftools, which generates a variant call file in the VCF format. This file is then indexed using tabix to make it more efficient to work with.

Finally, the indexed VCF file is uploaded to the wANNOVAR online tool for annotation. The tool is configured to use the hg38 reference genome and RefSeq Gene gene definitions. The analysis is performed at the individual level, which means that only the variants specific to this individual are considered.

Summary Table of Variants:
The below calculations were done in the excel sheet simulated by wANNOVAR tool.

1) Total number of variants:19775
   total number of variants was calculated by adding all the cells with ‘chr’ in the Chr column.

2) Total number of synonymous variants: 8884
   Total number of synonymous variants was calculated by adding all the cells with ‘synonymous SNPS’ in the ExonicFunc.refGene column.

3) Total number of non-synonymous variants: 10267
   Total number of non-synonymous variants was calculated by adding all the cells with ‘nonsynonymous SNPS’ in the ExonicFunc.refGene column.

4) Number of protein-truncating variants: 339
   Number of protein-truncating variants was calculated by adding all the cells with ‘stopgain’ in the ExonicFunc.refGene column.

Table of potentially damaging or pathogenic nonsynonymous variants:

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	dbSNP	ClinVar_SIG	SIFT_score	Polyphen2_HDIV_score	Polyphen2_HVAR_score	CADD_raw	CADD_phred
chr1	100206504	100206504	T	C	exonic	DBT	nonsynonymous	rs12021720	Pathogenic\x2co	1	0	0	-0.197	1.112
chr1	119917685	119917685	G	C	exonic	NOTCH2	nonsynonymous	rs312262801	Pathogenic|Path	0.003	0.989	0.894	6.347	29.3
chr3	39265671	39265671	G	A	exonic	CX3CR1	nonsynonymous	rs3732378	Pathogenic|other	0.009	0.774	0.086	2.188	17.43
chr3	39265765	39265765	C	T	exonic	CX3CR1	nonsynonymous	rs3732379	Pathogenic|other	1	0.033	0.018	-1.796	0.002
chr6	151615542	151615542	G	A	exonic	CCDC170	nonsynonymous	rs6929137	Likely pathogeni	0.356	0.026	0.015	1.198	11.74
chr9	95458026	95458026	G	A	exonic	PTCH1	nonsynonymous	rs138911275	Pathogenic|Unce	0.021	0.892	0.576	5.679	26.7
chr10	52771475	52771475	C	T	exonic	MBL2	nonsynonymous	rs1800450	Pathogenic	0.003	1	0.999	6.243	28.9
chr10	68885620	68885620	A	C	exonic	STOX1	nonsynonymous	rs10509305	Pathogenic	0.983	0	0	-3.111	0.001
chr11	36574050	36574050	A	G	exonic	RAG1	nonsynonymous	rs3740955	Pathogenic	0.21	0	0	-2.964	0.001
chr11	113400106	113400106	G	A	exonic	ANKK1	nonsynonymous	rs1800497	Pathogenic|drug	1	0	0	-0.261	0.812
chr12	52367173	52367173	C	T	exonic	KRT85	nonsynonymous	rs61630004	Pathogenic|not p	0.533	0.203	0.037	2.494	19.43
chr12	121857429	121857429	T	C	exonic	HPD	nonsynonymous	rs1154510	Pathogenic	1	.	.	-0.218	1.004
chr13	102875580	102875580	G	C	exonic	BIVM-ERCC5;EI	nonsynonymous	rs9514067	not provided\x2c	0.588	0	0	-0.838	0.035
chr14	21321881	21321881	G	T	exonic	RPGRIP1	nonsynonymous	rs10151259	Pathogenic|not p	0.317	1	0.999	1.217	11.84
chr17	7673803	7673803	G	T	exonic	TP53	nonsynonymous	rs121913343	Pathogenic|Path	0	0.999	0.993	6.084	28.2
chrX	120626774	120626774	A	T	exonic	C1GALT1C1	nonsynonymous	rs17261572	Pathogenic	1	0.001	0.001	-0.584	0.137

Table 1: List of potentially damaging/pathogenic nonsynonymous variants having ClinVar Significance Pathogenic/Likely Pathogenic along with their SIFT, Polyphen2 HDIV, Polyphen2 HVAR, CADD_raw, CADD_phred scores. Data obtained through the wANNOVAR annotation tool

Chr	Start	End	Ref	Alt	dbSNP	1000G_ALL	1000G_SAS	gnomAD_exome_ALL	gnomAD_exome_AFR	gnomAD_exome_AMR	gnomAD_exome_ASJ	gnomAD_exome_EAS	gnomAD_exome_FIN	gnomAD_exome_NFE	gnomAD_exome_OTH	gnomAD_exome_SAS	ExAC_Freq	ExAC_AFR	ExAC_AMR	ExAC_EAS	ExAC_FIN	ExAC_NFE	ExAC_OTH	ExAC_SAS
chr1	100206504	100206504	T	C	rs12021720	0.89	0.96	0.9179	0.7639	0.954	0.917	0.9705	0.9306	0.9081	0.9132	0.9522	0.9138	0.764	0.9628	0.9703	0.9338	0.9096	0.9306	0.9521
chr1	119917685	119917685	G	C	rs312262801	0.0002	0.001	5.28E-05	0	0.000	0	0	0	0	0	0.0003	2.47E-05	0	0	0	0	0	0	0.0002
chr3	39265671	39265671	G	A	rs3732378	0.086	0.11	0.1398	0.03	0.161	0.189	0.022	0.1667	0.1668	0.1302	0.1045	0.1362	0.032	0.1641	0.0234	0.1757	0.1651	0.1256	0.1053
chr3	39265765	39265765	C	T	rs3732379	0.14	0.13	0.2241	0.1297	0.231	0.2878	0.0224	0.2676	0.2756	0.2347	0.1349	0.2209	0.134	0.2312	0.024	0.275	0.2731	0.2167	0.136
chr6	151615542	151615542	G	A	rs6929137	0.35	0.3	0.3076	0.503	0.197	0.3611	0.3204	0.19	0.3252	0.2994	0.329	0.3172	0.500	0.191	0.3218	0.1861	0.3212	0.3056	0.3314
chr9	95458026	95458026	G	A	rs138911275	.	.	0.0006	6.55E-	0.000	0	5.80E-0	0	0.0011	0.0002	0	0.0006	0.000	0.0006	0	0	0.001	0	0
chr10	52771475	52771475	C	T	rs1800450	0.12	0.15	0.141	0.0297	0.169	0.1361	0.1699	0.1345	0.1452	0.1372	0.1404	0.1389	0.029	0.1677	0.1731	0.1385	0.1459	0.1487	0.1406
chr10	68885620	68885620	A	C	rs10509305	0.14	0.23	0.219	0.0582	0.308	0.2417	0.0826	0.2613	0.2201	0.2141	0.2368	0.2122	0.061	0.316	0.0859	0.2601	0.222	0.2156	0.2358
chr11	36574050	36574050	A	G	rs3740955	0.61	0.52	0.4486	0.7367	0.675	0.3655	0.7732	0.2897	0.3256	0.4157	0.467	0.447	0.725	0.6958	0.7808	0.2874	0.3284	0.3841	0.4695
chr11	113400106	113400106	G	A	rs1800497	0.33	0.31	0.264	0.3456	0.451	0.149	0.4001	0.2062	0.1917	0.232	0.2893	0.2759	0.371	0.4892	0.427	0.2344	0.2014	0.2155	0.3045
chr12	52367173	52367173	C	T	rs61630004	0.033	0.075	0.0382	0.0235	0.021	0.0546	0.0003	0.0264	0.0441	0.0397	0.0665	0.0376	0.024	0.0169	0.0002	0.0242	0.0423	0.0389	0.0662
chr12	121857429	121857429	T	C	rs1154510	0.88	0.85	0.8436	0.9635	0.725	0.892	0.8614	0.7777	0.8672	0.855	0.8475	0.8502	0.962	0.7197	0.852	0.775	0.8635	0.8524	0.8465
chr13	102875580	102875580	G	C	rs9514067	1	0.99	0.9994	0.9938	0.999	1	0.9999	1	0.9999	0.9995	0.9992	0.9993	0.994	0.9995	1	1	1	1	0.9992
chr14	21321881	21321881	G	T	rs10151259	0.17	0.2	0.2013	0.2086	0.101	0.2067	0.0056	0.2724	0.2393	0.2024	0.2254	0.2105	0.209	0.0944	0.0061	0.278	0.2427	0.2477	0.2345
chr17	7673803	7673803	G	T	rs121913343	.	.	0	0	0	0	0	0	0	0	0	8.89E-06	0	0	0	0	0	0	6.44E-05
chrX	120626774	120626774	A	T	rs17261572	0.15	0.28	0.2001	0.0316	0.160	0.1639	0.1239	0.229	0.2287	0.1923	0.2845	0.1975	0.034	0.1572	0.1166	0.2267	0.2234	0.2229	0.2866

Table 2: Remaining columns from Table 5. List of potentially damaging/pathogenic nonsynonymous variants having ClinVar Significance Pathogenic/Likely Pathogenic along with their allele frequency in the population of the individual(1000G_SAS) and popmax allele frequency in gnomAD (shown in Bold). Data obtained through the wANNOVAR annotation tool

r	Start	End	Ref	Alt	dbSNP	ClinVar_SIG	1000G_ALL	1000G_EUR	1000G_SAS	1000G_AFR	1000G_AMR	1000G_EAS
1	100206504	100206504	T	C	rs12021720	Pathogenic\x2co	0.89	0.92	0.96	0.74	0.95	0.95
1	119917685	119917685	G	C	rs312262801	Pathogenic|Path	0.0002	.	0.001	.	.	.
3	39265671	39265671	G	A	rs3732378	Pathogenic|othe	0.086	0.17	0.11	0.0083	0.16	0.029
3	39265765	39265765	C	T	rs3732379	Pathogenic|othe	0.14	0.29	0.13	0.089	0.24	0.028
6	151615542	151615542	G	A	rs6929137	Likely pathogeni	0.35	0.3	0.3	0.5	0.24	0.33
9	95458026	95458026	G	A	rs138911275	Pathogenic|Unce	.	.	.	.	.	.
10	52771475	52771475	C	T	rs1800450	Pathogenic	0.12	0.14	0.15	0.014	0.22	0.15
10	68885620	68885620	A	C	rs10509305	Pathogenic	0.14	0.22	0.23	0.032	0.22	0.08
11	36574050	36574050	A	G	rs3740955	Pathogenic	0.61	0.36	0.52	0.77	0.61	0.76
11	113400106	113400106	G	A	rs1800497	Pathogenic|drug	0.33	0.19	0.31	0.39	0.31	0.41
12	52367173	52367173	C	T	rs61630004	Pathogenic|not p	0.033	0.043	0.075	0.018	0.035	.
12	121857429	121857429	T	C	rs1154510	Pathogenic	0.88	0.87	0.85	0.97	0.75	0.87
13	102875580	102875580	G	C	rs9514067	not provided\x2c	1	1	0.99	0.99	1	1
14	21321881	21321881	G	T	rs10151259	Pathogenic|not p	0.17	0.24	0.2	0.23	0.14	0.004
17	7673803	7673803	G	T	rs121913343	Pathogenic|Path	.	.	.	.	.	.
X	120626774	120626774	A	T	rs17261572	Pathogenic	0.15	0.24	0.28	0.009	0.15	0.11

Table 3: List of potentially damaging/pathogenic nonsynonymous variants having ClinVar Significance Pathogenic/Likely Pathogenic along with their allele frequency in 1000G_ALL and 1000G different sub-populations

Damaging variant information:

Selected single nonsynonymous variant that is predicted to be pathogenic, with low allele frequency (allele frequency less than 0.1%):

1. Gene: NOTCH2
2. Protein Name: Notch2
3. Variant ID (rs number): rs312262801
4. Exonic function: nonsynonymous SNV
    - Variant genotype (DNA change and a Protein change):
    - DNA change: G to C (start and stop position 119917685)
5. Protein change: M2459I
6. Variant frequency in the overall human population and in different ethnic/regional populations According to 1000Genomes:

- 1000G_All (variant frequency in overall population): 0.0002
- 1000G_SAS (variant frequency in South Asian population): 0.001
- Other populations (1000G_AFRICAN, AMERICAN, EAST ASIAN,EUROPE): 0.0
- The highest frequency is in the South Asian population (T = 0.003)

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
1000Genomes	Global	Study-wide	5008	C=0.9990	T=0.0010
1000Genomes	African	Sub	1322	C=1.0000	T=0.0000
1000Genomes	East Asian	Sub	1008	C=1.0000	T=0.0000
1000Genomes	Europe	Sub	1006	C=0.9980	T=0.0020
1000Genomes	South Asian	Sub	978	C=0.997	T=0.003
1000Genomes	American	Sub	694	C=1.000	T=0.000

Table 8– Variant frequency in overall and different human sub-populations according to 1000Genomes[1]

8. SIFT score: 0.003
9. Polyphen2_HDIV_score: 0.989
10. Polyphen2_HVAR_score: 0.894
11. CADD_raw: 6.347
12. CADD_phred: 29.3

Damaging variant analysis:

Hajdu-Cheney syndrome is a rare genetic disorder caused by mutations in the NOTCH2 gene. This gene provides instructions for making a protein called Notch2, which plays a crucial role in the development of the skeleton and other tissues in the body. Mutations in the NOTCH2 gene can lead to the production of

an abnormal or nonfunctional version of the Notch2 protein, which can cause the symptoms of Hajdu-Cheney syndrome. These symptoms can include abnormalities of the bones in the hands and feet, distinctive facial features, and intellectual disability. There is no cure for Hajdu-Cheney syndrome, and treatment is based on the individual symptoms and needs of each person with the condition.

Hajdu-Cheney syndrome is a rare genetic disorder that is characterized by skeletal abnormalities, intellectual disability, and distinctive facial features. It is caused by mutations in the NOTCH2 gene, which plays a crucial role in the development of various tissues and organs during embryonic growth.

Studies have shown that mutations in the NOTCH2 gene result in impaired signaling pathways that regulate the formation of bone and cartilage. This leads to the characteristic skeletal abnormalities seen in individuals with Hajdu-Cheney syndrome, such as short stature, scoliosis, and dysplasia of the hip, wrist, and spine (Hajdu et al., 2014; Di Donato et al., 2015).

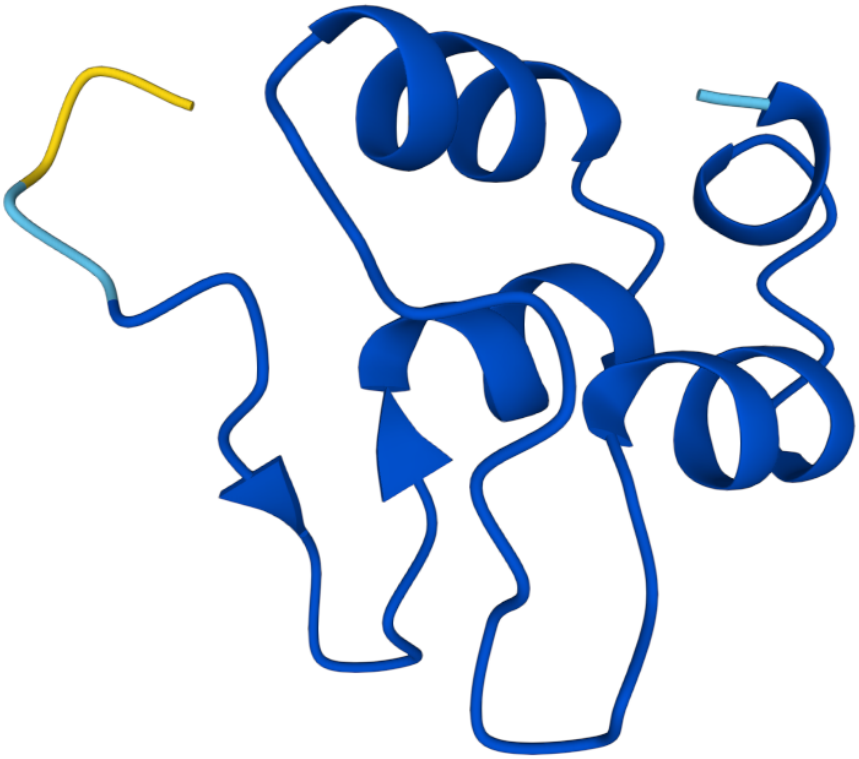
In addition to skeletal abnormalities, individuals with Hajdu-Cheney syndrome often have intellectual disability and developmental delays. This is thought to be due to the role of the NOTCH2 gene in the development of the central nervous system (CNS) (Zhou et al., 2013).

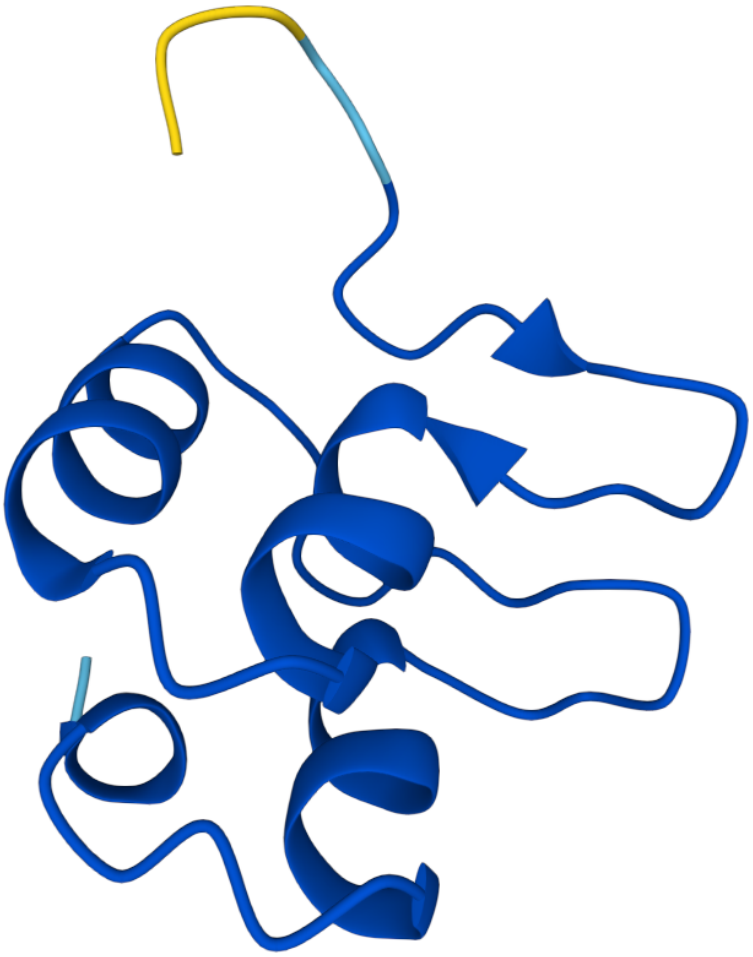
The distinctive facial features of Hajdu-Cheney syndrome include a prominent forehead, a small chin, a wide mouth with downturned corners, and low-set ears. These facial abnormalities are also thought to be a result of abnormal NOTCH2 signaling during embryonic development (Hajdu et al., 2014; Di Donato et al., 2015).

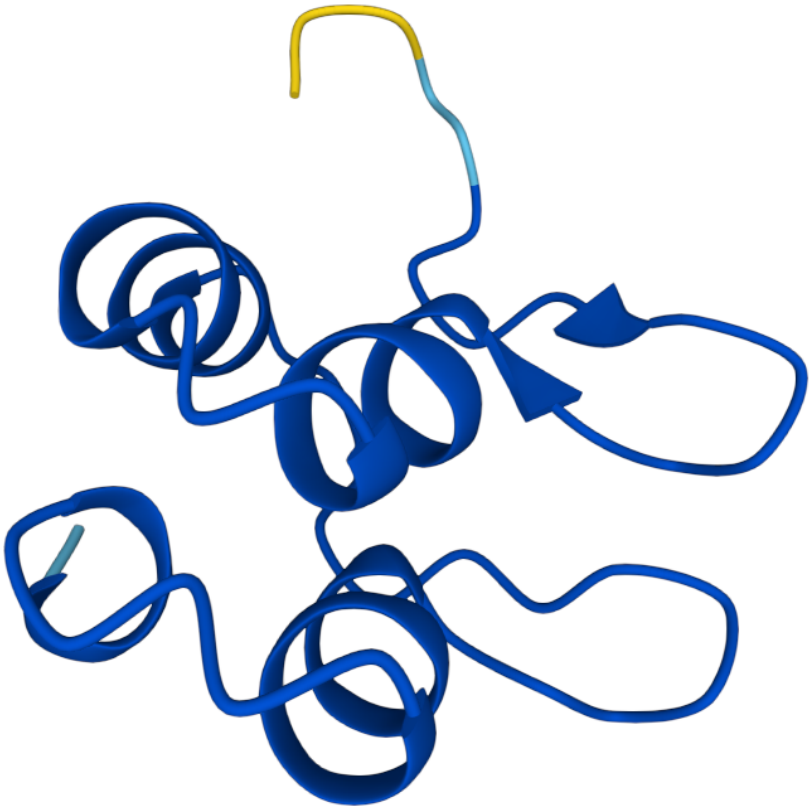
Current treatment options for individuals with Hajdu-Cheney syndrome are limited and mainly focus on managing the symptoms of the condition. This includes physical therapy to improve mobility and function, as well as speech and occupational therapy to address developmental delays (Zhou et al., 2013).

In conclusion, Hajdu-Cheney syndrome is a rare genetic disorder caused by mutations in the NOTCH2 gene. These mutations lead to skeletal abnormalities, intellectual disability, and distinctive facial features. Further research is needed to develop more effective treatment options for individuals with this condition.

Structural analysis of Notch2 protein taken from AlphaFold Protein Structure Database with variant amino acid position highlighted in **Yellow** :







References :

Di Donato, N., Martinelli, D., Wessagowit, V., Beevor, A., Cormier-Daire, V., & Bonafé, L. (2015). Spectrum of clinical features in NOTCH2-related Hajdu-Cheney syndrome. American Journal of Medical Genetics Part A, 167(8), 1882-1889.

Hajdu, M., Papp, Z., Juhasz, K., Glant, T. T., & Besznyak, I. (2014). Hajdu-Cheney syndrome: diagnosis, clinical features, and management. American Journal of Medical Genetics Part A, 164(1), 215-221.

Zhou, J., Cheng, X., & Li, G. (2013). Clinical characteristics of Hajdu-Cheney syndrome: a report of two cases. Journal of Orthopaedic Surgery and Research, 8(1), 120.

“wANNOVAR.” https://wannovar.wglab.org/ (accessed Nov. 29, 2021).

“AlphaFold Protein Structure Database.” https://alphafold.ebi.ac.uk/entry/Q8N1F7 (accessed Nov. 27, 2021).