Novel *HEXB* Single Nucleotide Polymorphism exhibits Rare Global Allele Frequency.

Valantina Persaud[1], Ashwini Vasanth[1], Sunjay Ravishankar[1], Brian McEllin[1], Saksham Gupta[1]
[1]Johns Hopkins University, Baltimore, MD

Abstract

Infantile Onset Sandhoff disease is a neurodegenerative, autosomal recessive disorder that is caused by a mutation in the Hexosaminidase Subunit Beta gene (*HEXB* gene), which codes for the β-subunit of β-hexosaminidase. Without the proper production of this enzyme, there is no breakdown of the G_{M2} Ganglioside, a lipid involved in the cell-cell recognition, cell adhesion, and signal transduction pathways within microdomains of the cells. Generally, these lipids are found in high concentrations in neurons. As this is the case, these lipids need to be broken down via a β-hexosaminidase – G_{M2} ganglioside activator complex, otherwise they buildup within the lysosomes of neuronal cells, leading to toxic vacuolation and eventually cell death. Due to this, various conditions such as Sandhoff's disease can arise. This study aims to investigate a single nucleotide polymorphism (SNP) directly associated with Sandhoff's disease identified in the *HEXB* gene, rs771103635. We used a variety of methods such as next generation sequencing (NGS), variant calling for SNPs, data analysis of various repositories, etc. to understand the general frequency of this SNP and understand its effects. Based on our research, we found that the *HEXB* SNP (rs771103635) was a rare pathogenic allele that did not occur frequently on a global scale.

Keywords: *HEXB*, Infantile Onset Sandhoff disease, Thai, NGS, GM2 gangliosidoses, neurodegeneration

Introduction

Sandhoff disease

Sandhoff disease is a rare autosomal recessive disorder characterized by abnormalities in movements and glycosphingolipid metabolism, ataxia, and blindness (NIH, 2011). The population frequency of Sandhoff disease is approximately 1 in 1,000,000 in live births. This includes three different subsets: infantile-onset, juvenile-onset, and adult-onset (Rahmani et al., 2020). Sandhoff disease is caused by mutations in the *HEXB* gene (MedlinePlus, n.d.), which encodes a subunit of beta-hexosaminidase, a glycosyl hydrolase. Mutations in this protein are implicated in diseases of neurons and neurodegenerative disorders (Dastsooz et.al., 2018). At least 88 variants of the *HEXB* gene have been identified globally in populations such as Argentina, Cyprus, Italy, the Middle East, Spain, France, Thailand, Korea, China, and Japan (Rahmani et al., 2020). The most common pathogenic variants that result in Sandhoff disease are missense or nonsense mutations, insertions, and deletions (Rahmani et al., 2020). These mutations typically impair *HEXB* function, leading to the toxic accumulation of GM2 gangliosides (Ganglioside Monosialic 2) (Dastsooz et.al., 2018). Neurons are very susceptible to these mutations as they can cause excess lipid molecules to trigger the destruction of nerve cells in the central nervous system, causing the hallmark features of Sandhoff disease (Dastsooz et.al., 2018). The accumulation of lipids within the nerve cells causes a buildup in the cells' lysosomes. As a result, these lysosomes can get disrupted and cause apoptosis and cell death (Dastsooz et al., 2018).

Infantile-onset Sandhoff disease is characterized by rapid progression of mental and motor decline. Infants, within the first six months, show symptoms such as missing developmental milestones, difficulty feeding, spasticity, cherry-red spots of the macula, and macrocephaly. They often do not live past 2-5 years of age (NORD, n.d.). The severity and onset of juvenile and adult-onset Sandhoff disease could vary. They display slower mental and motor decline and are characterized by symptoms such as muscular atrophy, dystonia, autonomic neuropathy, cognitive and psychiatric disorders, and memory loss (NORD, n.d.).

HEXB gene structure and regulation

HEXB is a protein-coding gene located at the 5q13.3 area in the human genome. It is characterized by 13 known transcripts spanning 15 exons, including four protein-coding transcripts (Figure 1). The two significant transcripts variants (201 and 207) have the largest differences in their exons and result in the alternative splicing at the 5' end. In the brain- spinal cord region, the dominant transcript is 201 (GTEx Portal, n.d.)

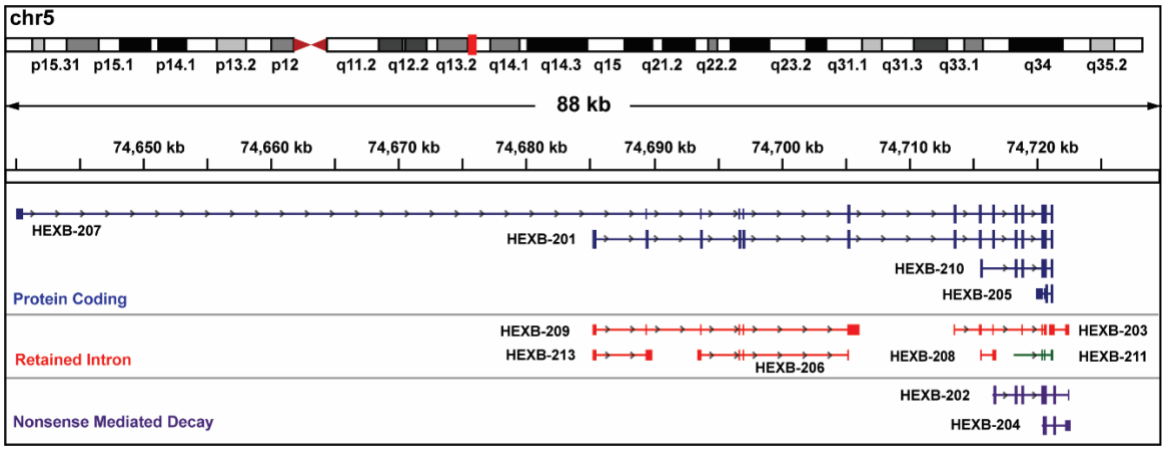


Figure 1. Gene structure of the HEXB gene. (A) *HEXB* has 13 known transcripts spanning 15 total exons (protein-coding transcripts are in blue). Exon number and total length for the protein-coding transcripts are as follows: 207 - 14 exons, 2021 base pairs (bp); 201 - 14 exons, 1812bp; 210 - 6 exons, 720 bp; 205 - 3 exons 806bp. Other transcripts include those that retained the intronic sequences (shown in red), processed transcripts (green), or transcripts that are degraded by nonsense-mediated decay (purple).

Several studies have examined the transcriptional regulation of *HEXB*. F. Norflus *et al*. (1999) performed a study to identify the elements responsible for the expression of the *HEXB* gene. Their analysis indicated the presence of a 60-base pair(bp) promoter region between 150 bp and 90 bp upstream of the start codon. Using scanning mutagenesis in this 60-bp region, they were able to identify a promoter sequence of 12-bp. This region had two potential Activating Protein-1 binding sites (AP-1) (Norflus et al., 1999). AP-1 refers to the dimeric transcription factors (TFs) that contain Jun, Fos, or Activating transcription factor (ATF) subunits. These TFs bind to the AP-1 binding sites on the DNA and regulate gene expression (Karin et al., 1997). There is evidence that mutations in the *HEXB* promoter are associated with Sandhoff disease (Norflus et al., 1999). Other studies have identified single nucleotide polymorphisms (SNPs) in regulatory regions that altered *HEXB* gene expression. One study identified two SNPs in the 3'UTR of the gene, rs75974765 and rs1048088 both of which are thought to disrupt the binding site of hsa-miR-3679-3p miRNA, generating new binding sites and is thought to result in *HEXB* gene expression (Abdelhameed et al., 2019).

The Human Regulatory Features database (chr5: 74,640,023-74,722,647) was queried to generate a comprehensive list of known regulatory sites. The query identified several regulatory elements including four predicted promoter sequences at positions chr5:74,687,800-74,688,001, chr5:74,684,400-74,687,601, chr5:74,643,600-74,644,001, and chr5:74,636,600-74,643,401. In addition, there were multiple CCCTC-binding factor (CTCF) binding sites, TF binding sites, and open chromatin regions found in this larger area (Table 1). Finally, there was one enhancer element identified at position chr5:74,670,401-74,671,000. The multiple CTCF sites were an exciting find since these sites are known to play a role in binding insulators to control gene expression (Kim et al., 2015).

Table 1. List of regulatory elements identified near the HEXB gene.

Start (bp)	End (bp)	Feature type	Feature type description	Bound start (bp)	Bound end (bp)	Regulatory stable ID
74687800	74688001	Promoter	Predicted promoter	74685402	74693799	ENSR00001096652
74680001	74680200	CTCF Binding Site	CTCF binding site	74680001	74680200	ENSR00000759062
74660136	74661400	Promoter Flanking Region	Predicted promoter flanking region	74660136	74661400	ENSR00001096647
74679802	74680600	Promoter Flanking Region	Predicted promoter flanking region	74679802	74680600	ENSR00000759061
74651672	74652307	Open chromatin	Open chromatin region	74651672	74652307	ENSR00000182615
74670401	74671000	Enhancer	Predicted enhancer region	74670401	74671000	ENSR00001096650
74690601	74690800	CTCF Binding Site	CTCF binding site	74690601	74690800	ENSR00001258834
74648401	74650000	CTCF Binding Site	CTCF binding site	74648401	74650000	ENSR00000315609
74672601	74673600	CTCF Binding Site	CTCF binding site	74672601	74673600	ENSR00000315611
74664601	74669000	Promoter Flanking Region	Predicted promoter flanking region	74664601	74669000	ENSR00000759055
74677015	74677491	TF binding site	Transcription factor binding site	74676438	74677540	ENSR00000759060
74715935	74716214	TF binding site	Transcription factor binding site	74715935	74716214	ENSR00001096653
74682201	74682400	CTCF Binding Site	CTCF binding site	74682201	74682400	ENSR00000759063
74646602	74649799	Promoter Flanking Region	Predicted promoter flanking region	74646602	74649799	ENSR00000182613
74676401	74676600	CTCF Binding Site	CTCF binding site	74676401	74676600	ENSR00001258833
74715801	74716400	CTCF Binding Site	CTCF binding site	74715801	74716400	ENSR00000759072
74684400	74687601	Promoter	Predicted promoter	74684002	74693799	ENSR00000182622
74643600	74644001	Promoter	Predicted promoter	74639769	74645999	ENSR00001096645
74681783	74682386	Open chromatin	Open chromatin region	74681783	74682386	ENSR00000182621
74650601	74650800	CTCF Binding Site	CTCF binding site	74650601	74650800	ENSR00001096646
74642201	74642600	CTCF Binding Site	CTCF binding site	74642201	74642600	ENSR00000759052
74636600	74643401	Promoter	Predicted promoter	74636002	74645999	ENSR00000182612

Pathogenic Variants in the HEXB Gene

Sandhoff disease is characterized by many different variants of the *HEXB* gene. According to the ClinVar database, there are 76 variants of the *HEXB* gene, 26 of which are SNPs (Supplementary Table 1). Similarly, a search of OMIM showed 19 allelic variants of the *HEXB* gene, including insertions, SNPs, and large deletions (Table 2). Based on the analysis of the data from ClinVar, it is evident that all the variants of the HEXB gene lead to Sandhoff disease. For example, a 16-kb deletion (RCV000004077) in one of the HEXB gene alleles encompassed the 5'UTR region, promoter region, and extended to exon 5 (Bikker et al., 1990). A heterozygous 1367A>C transversion (RCV000004081, RCV000675054) resulting in Tyr456Ser(Y456S) substitution mutation (rs10805890) was found in a female patient with juvenile-onset SD (OMIM, n.d.). The patient was heterozygous for two polymorphisms, 619A>G transition causing I207V substitution and K121R. Banerjee et.al. (1994) demonstrated that the I207V beta-chain does not self-associate when present at low concentrations. A patient with a non-functional *HEXB* allele shows a reduction in beta-chains by 50% compared to the average concentration (OMIM, n.d.a). The existing beta chains fail to associate to form the HEXB protein however they can dimerize in the presence of an abundance of normal alpha-chains resulting in partial beta-HEXA and absence of beta-HEXB (OMIM, n.d.a). The study of the allelic variants in the OMIM database leads to the inference that the variants of the *HEXB* gene could vary in nature from deletions to transversions. In addition, these variations could potentially occur across the gene. A list of the 19 samples cataloged in OMIM can be referred to in Table 2 with their corresponding mutations.

Table 2. Table view of the allelic variants of the HEXB gene linked to Sandhoff Disease. OMIM. (n.d.).

Phenotype	Mutation	SNP	gnomAD SNP	ClinVar
SANDHOFF DISEASE	HEXB, 16-KB DEL	-	-	RCV000004077
SANDHOFF DISEASE, JUVENILE TYPE	HEXB, 24-BP INS	-	-	RCV000004079
HEXOSAMINIDASE B (PARIS)	HEXB, 18-BP INS	-	-	RCV000004080
HEXB POLYMORPHISM	HEXB, ILE207VAL	rs10805890	rs10805890	RCV000403428;
				RCV000235014;
				RCV000079065;
				RCV000675619
SANDHOFF DISEASE, JUVENILE TYPE	HEXB, TYR456SER	rs121907982	rs121907982	RCV000004081;

				RCV000675054
SANDHOFF DISEASE, JUVENILE TYPE	HEXB, PRO417LEU	rs28942073	rs28942073	RCV000174009;
				RCV000004084;
				RCV000004082;
				RCV000079058
HEXB POLYMORPHISM	HEXB, LYS121ARG	rs11556045	rs11556045	RCV000004078;
				RCV000079063;
				RCV000336190;
				RCV000675618
SANDHOFF DISEASE, ADULT TYPE	HEXB, ARG505GLN	rs121907983	rs121907983	RCV000004083;
				RCV000669552
SANDHOFF DISEASE, ADULT TYPE	HEXB, PRO405LEU	rs28942073	rs28942073	RCV000174009;
				RCV000004084;
				RCV000004082;
				RCV000079058
HEXOSAMINIDASE B, HEAT-LABILE POLYMORPHISM	HEXB, ALA543THR	rs121907984	rs121907984	RCV000079061;
				RCV000987527;
				RCV000004085
SANDHOFF DISEASE, INFANTILE TYPE	HEXB, SER62LEU	rs820878	rs820878	RCV000869482;
				RCV000004086
SANDHOFF DISEASE, INFANTILE TYPE	HEXB, PARTIAL DEL			RCV000004087
SANDHOFF DISEASE, CHRONIC	HEXB, PRO504SER	rs121907985	rs121907985	RCV000004088;
				RCV001238377
SANDHOFF DISEASE, INFANTILE	HEXB, IVS8, G-C, +5	rs5030731	rs5030731	RCV000004089
SANDHOFF DISEASE, INFANTILE	HEXB, 1-BP DEL, 76A	rs1580377105		RCV000004090
SANDHOFF DISEASE, INFANTILE	HEXB, ARG284TER	rs121907986	rs121907986	RCV000004091;
				RCV000579011;
				RCV000184012
SANDHOFF DISEASE, INFANTILE	HEXB, 1-BP DEL, 965T	rs768438206	rs768438206	RCV000004092;
				RCV000673580
SANDHOFF DISEASE, ADULT	HEXB, ASP494GLY			RCV001078201

Identification of a Novel HEXB SNP rs771103635

A recent article by Tim-Aroon et al. describes an infant patient with a heterozygous genotype of c.1652G>A/c.761T>C presented with symptoms including developmental regression, axial hypotonia, exaggerated startle response to noise, macular cherry-red spot, and a cardiac defect (Tim-Aroon et al., 2021). While the c.1652G>A SNP (rs727503961) has been linked to HEXB protein dysfunction previously (Fitterer et al., 2014), this was the first instance reported of the c.761T>C (rs771103635) variation. Magnetic resonance imaging (MRI) results indicated changes that differed from the Sandhoff patients homozygous for c.1652G>A pertaining to cerebral white matter, the bilateral thalami, and corpus callosum and cerebellar peduncles (Tim-Aroon et al., 2021). rs771103635 is considered rare with an estimated allele frequency of approximately 0.00001 only in the Asian population (NCBI, 2021). A search of the NCBI Allele frequency aggregator (ALFA) confirms the low frequency of this novel SNP as depicted in Table 3 (Ensembl, 2021).

Table 3. Ensembl population genetics of rs771103635 SNP, c.761T>C variant of *HEXB*

Population	Population type	NCBI ALFA' T' allele frequency (count)	NCBI ALFA 'C' allele frequency (count)
ALFA: SAMN10492695	European	T: 1.000 (2072)	C: 0.000
ALFA: SAMN10492696	African (Others)	T: 1.000 (6)	C: 0.000
ALFA:SAMN10492697	East Asian	T: 1.000 (2)	C: 0.000
ALFA: SAMN10492698	African American	T: 1.000 (76)	C: 0.000
ALFA: SAMN10492701	Other Asian	T: 1.000 (2)	C: 0.000
ALFA: SAMN10492702	South Asian	T: 1.000 (4)	C: 0.000
ALFA:SAMN10492703	African	T: 1.000 (82)	C: 0.000
ALFA: SAMN10492704	Asian	T: 1.000 (4)	C: 0.000
ALFA: SAMN10492705	Total	T: 1.000 (2188)	C: 0.000

ALFA: SAMN11605645	Other	T: 1.000 (26)	C: 0.000

The rs771103635 SNP occurs in exon six at chr5:74,705,310 and is present in three *HEXB* transcripts – *HEXB*-207, *HEXB*-201, and *HEXB*-209 (Figure 2). The diagram indicates that rs771103635 causes an amino acid change from leucine to serine in the two protein-coding transcripts. With respect to *HEXB*-201, the change occurs at amino acid 254 (p.Leu254Ser) and for *HEXB*-207, the amino acid change occurs at position 29 (p.Leu29Ser). Molecular modeling suggests that the switch from a non-polar aliphatic amino acid (leucine) to a polar amino acid (serine) causes a change in protein structure in the active site of the enzyme leading to reduced enzyme activity (Tim-Aroon et al., 2021).

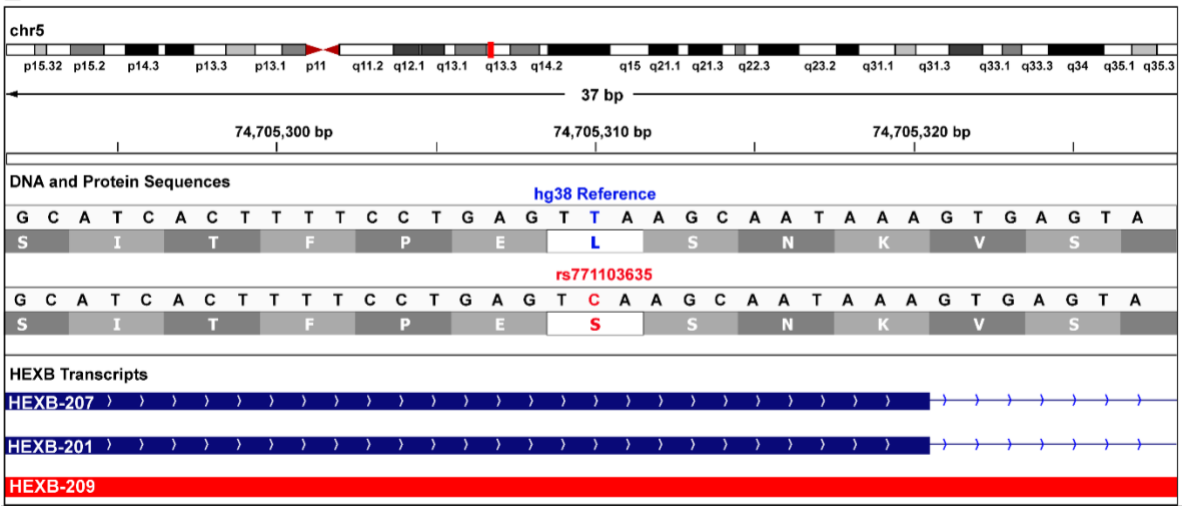


Figure 2. Location and characteristics of rs771103635. The *HEXB* SNP rs771103635 occurs in exon 6 (chr5:74,640,023-74,721,288). This SNP causes a T>C base substitution, leading to an amino acid change from Leu>Ser at position 254. This change affects the *HEXB*-201 and *HEXB*-207 protein-coding transcripts at amino acid 254 (p.Leu254Ser) and amino acid 29 (p.Leu29Ser).

Effect of rs77110365 on gene regulation

In addition to its role in altering protein function, the rs771103635 SNP may also play a role in regulating *HEXB* transcription. As previously discussed, SNPs in other regulatory elements of *HEXB* have also been linked to Sandhoff disease. PROMO is a virtual laboratory for the identification of putative transcription factor binding sites (TFBS) in DNA sequences from a species or groups of species of interest. TFBS defined in the TRANSFAC database are used to construct specific binding site weight matrices for TFBS prediction. To find the potential transcription factor binding sites (TFBS) at the site of rs771103635, the FASTA sequence of the *HEXB*-201 mRNA transcript (accession NM_000521.4) was input into the PROMO program that uses the TRANSFAC database as its source (PROMO, n.d.). Ensembl was used to retrieve the variant information of the rs771103635, which occurs at position 789 on the mRNA transcript. (Note: Ensembl does not detect the overlap of the SNP with any regulatory features (data not shown)).

The result from the PROMO program was analyzed at position 789 on the transcript. The data indicated likely transcription factors that bind at position 789: GR-alpha, FOXP3, c-Myb, HNF-1B, HNF-1C, and HNF-1A (Figure 3). Glucocorticoid receptors (GRs) are nuclear hormone receptors, with two isoforms of GR (α and β) present in humans. The GR α form is ubiquitously expressed in most tissues and its binding is known to cause changes in gene expression. In the GR signaling pathway, circulating glucocorticoids passively enter the cell cytoplasm and activate cytoplasmic GR. Once activated, GR translocates to the nucleus and their interactions with other TFs like the NF-κB, c-fos, and c-jun components of AP-1 which results in changes in expression of the target gene mRNA (Wallace & Cidlowski, 2002). The promoter sequence of the *HEXB* gene has two AP-1 binding sites to which other TFs bind and regulate gene expression (Norflus et al., 1999). The results from PROMO indicate that there are TFs associated with the SNP region that could also alter mRNA expression. Further experiments are needed to determine if rs771103635 has the potential to impact gene expression.

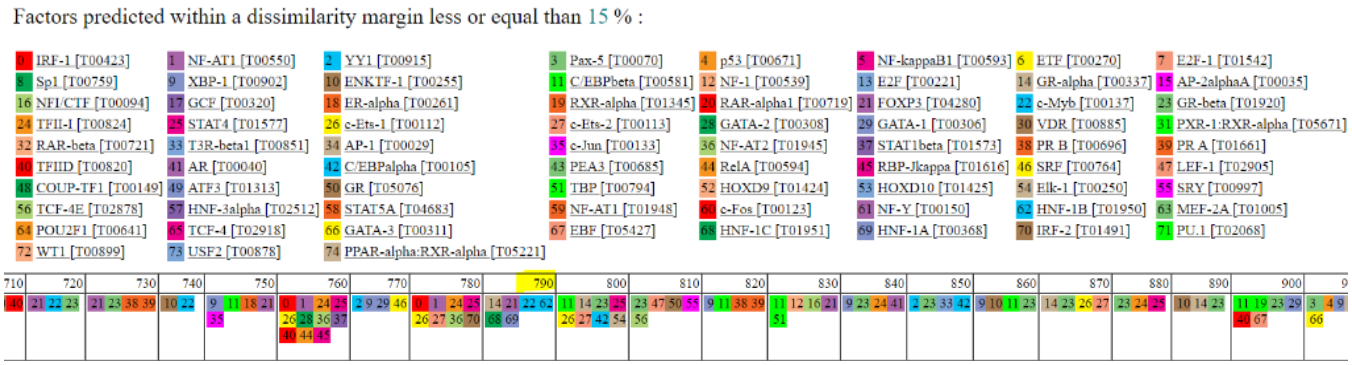


Figure 3. PROMO prediction results showing the potential transcription factors that bind to the region of the SNP. This image shows an analysis of the transcription factor binding sites in the areas closest to the analyzed SNP, rs77103635.

Hypothesis

The initial report identified rs771103635 in one patient and described how the single substitution altered protein function. While the evidence suggests that this variant is likely rare the exact frequencies are unknown. This report analyzes sequencing data collected from three individuals – two European individuals and one Thai individual. Since the *HEXB* SNP was initially identified in the Thai population, we hypothesized that the highest likelihood of a carrier was from the Thai individual. The European individuals served as our negative control. Due to the low frequency of the rs771103635 a negative result was anticipated from our analysis.

Methods

Human Regulatory Features Database Search

PROMO was used to identify the putative TFBS at the location of the rs771103635 SNP. Based on this, the potential TFs that could bind at position 789 on the HEXB-201 mRNA transcript were analyzed. Further, Ensembl was used to investigate the variant information of the rs771103635 SNP. The results from PROMO and Ensembl were then correlated.

OMIM database search for HEXB variants

OMIM database was searched using the gene name "*HEXB*". From the search results, the record specific to the *HEXB* gene with the OMIM ID: 606873 was examined. The allelic variants of the *HEXB* gene are cataloged in OMIM with the inclusion criteria being the first mutation discovered, significance, higher presence in the population, distinct phenotype, and mechanism of mutation among others (OMIM, n.d.). Most allelic variants in OMIM represent mutations that cause diseases (OMIM, n.d.). The phenotype and mutations associated with the allelic variants of HEXB gene were studied.

Sequencing Datasets

DNA sequencing of three unrelated individuals was performed in the study. All datasets were available through the International Genome Sample Resource (IGSR) repository. The samples included whole-exome sequencing data from a British male (Sample HG00234), low coverage whole genome sequencing from a Finnish female (Sample HG00171), and PCR-free high coverage DNA sequencing from a Thai female.

NGS Data Processing

The FASTQ files from the three individuals were imported into Galaxy. FASTQC was performed to assess quality. Samples were then trimmed using Trimmomatic (sliding window = 4, average quality>20) and quality was assessed with FASTQC. Reads were then aligned to the reference genome hg38 using HISAT2. Resulting BAM files were visualized with Integrative Genomics Viewer (IGV) software.

Variant Calling for HEXB SNPs

After alignment with HISAT2, FreeBayes was used to identify variants on the BAM file for all samples. This data was filtered with VCFfilter to include only variants with a read depth of >10.

Results

Next-Generation Sequencing Analysis and Interpretation

Sequencing data from three individuals were used to search for the presence of rs771103635 – a Thai female, British male and a Finnish female. Datasets were first retrieved from the ISGR and processed in Galaxy as described in the methods (quality control, trimming, and alignment – see Supplementary Figures X-Y). The resulting BAM files were then visualized in the IGV and compared against the hg38 reference genome at chr5:74,705,310. All samples displayed wild-type nucleotide (T) at the position of the indicated position (Figure 4).

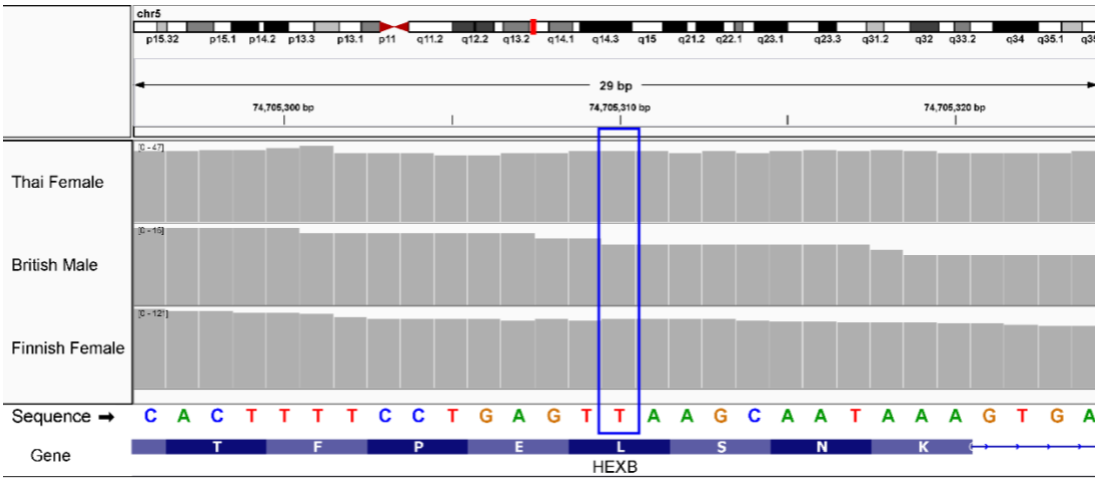


Figure 4. Sequence analysis at chr5:74,705,310. All individuals had a WT allele (T) at the site of rs771103635

The sequencing data was then scanned in IGV for evidence of SNPs in all *HEXB* exons (>10 read counts). Both the Finnish and Thai individuals were heterozygous at other locations. In exon 2, the Finnish female had a c.362A>G substitution at chr5:74,689,390 (rs11556045), leading to a p.Lys121Arg amino acid change (Figure 5A). This SNP has a 17.7% frequency in the European population (ALFA) (Phan, 2020) and is considered benign (ClinVar, n.d.a). Similarly, in exon 5 the Thai female had a c.619A>G substitution at chr5:74,697,056 (rs10805890), leading to a p.Ile207Val amino acid change (Figure 5B). As with rs11556045, this SNP is relatively common, with a 23.9% frequency in the Asian population (ALFA) (Phan, 2020). It is also considered benign (ClinVar, n.d.b).

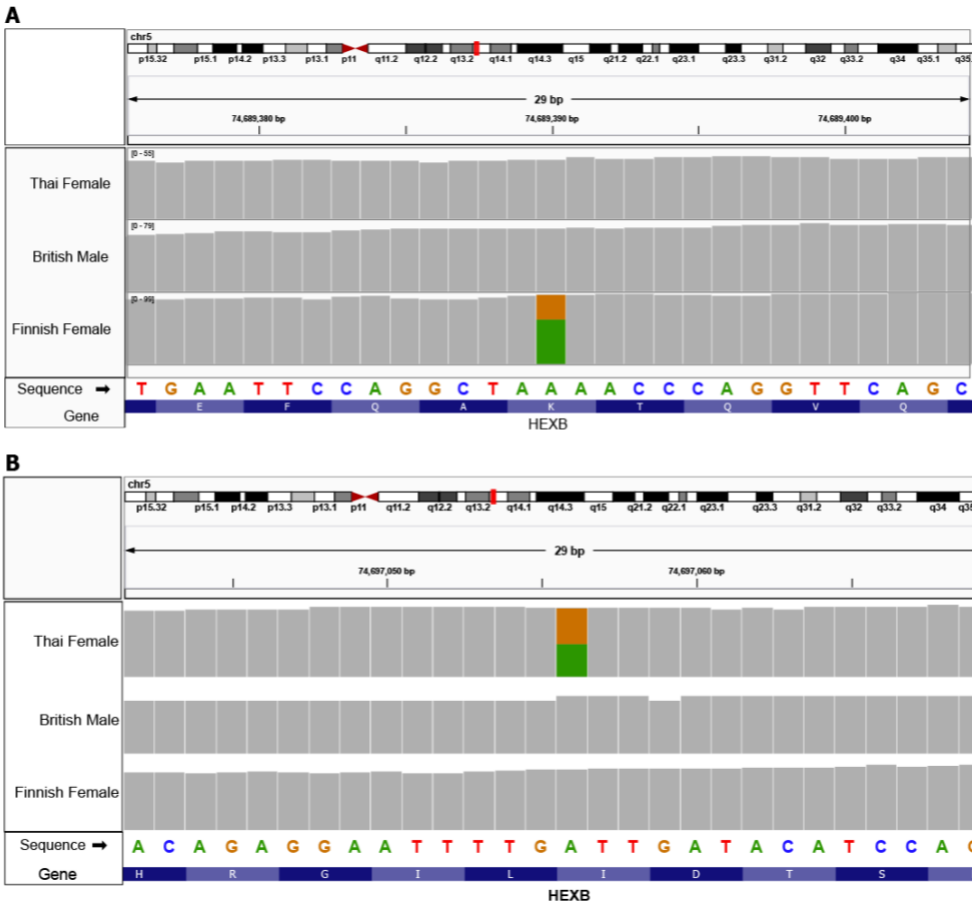


Figure 5. *HEXB* SNPs present in each sample. Other exons of *HEXB* were examined for the presence of alternate SNPs. The Finnish female was heterozygous for the rs11556045 SNP at chr5:74,689,390 in exon 2. This SNP has a c.362A>G substitution, leading to a Lys>Arg substitution at amino acid 121 (A) (dbSNP, n.d.b). In exon 5, there was evidence that the Thai female was heterozygous for the rs10805890 SNP at chr5:74,697,056. This SNP has a c.619A>G substitution, leading to a Ile>Val substitution at amino acid 207 (B) (dbSNP, n.d.a.)

To confirm this data, variant calling was also performed using the FreeBayes program for all three individuals. After searching this data, it confirmed that none of the individuals had the rs771103635 SNP. Total variants detected were as follows: Thai female – 221, British male – 4, Finnish female – 16. The drastic differences in total variant count are likely due the difference in coverage and read depth in the Thai female sample (whole genome sequencing, high depth) compared to the British male sample (whole exome sequencing) and Finnish female sample (low coverage sequencing). Most of the variants in the Thai patient were found in introns (Figure 6).

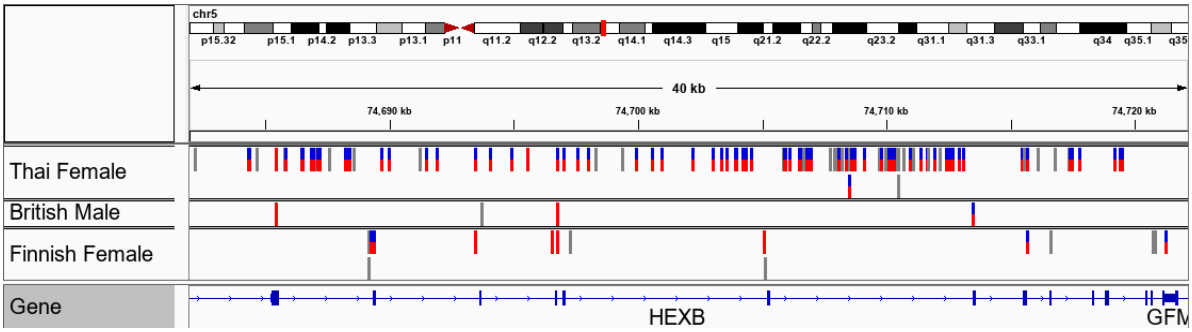


Figure 6. Variants in *HEXB* identified through FreeBayes.

Discussion

In this report, we have profiled a recently identified *HEXB* SNP, rs771103635, that was discovered in a Thai patient with infantile Sandhoff disease. The patient was heterozygous for rs77103635 and another well characterized *HEXB* mutation, rs727503961 (Tim-Aroon et al, 2021). Initial characterization of rs77103635 demonstrated that it severely impacted normal HEXB protein function, reducing activity to undetectable levels. Little is known about the prevalence of this mutation within the Thai population or globally, although evidence suggests the mutation is rare. It's also possible that rs771103635 is a novel mutation in the patient; however, we lacked access to sequencing data from immediate family members. Future collection of sequencing data from the child's parents could disprove this hypothesis.

We attempted to identify this mutation in a select group of individuals. Since Sandhoff disease is a rare autosomal recessive disorder, we reasoned that choosing a cohort of unaffected individuals might allow us to find carriers of this mutation. We chose to search DNA sequencing from a Thai female to match the population where it was first identified and two European individuals to serve as negative controls. A positive identification in the Thai female would provide support for carriers in the regional population, while an unexpected positive identification in the European individuals would indicate that this mutation is not geographically limited. However, analysis of the three DNA sequencing datasets showed that all individuals were homozygous for the hg38 reference allele at chr5:74,705,310. This result was the most likely finding, and it does not provide any additional insight into the prevalence of rs771103635. Future experiments on larger cohorts in Thailand are necessary to get a better estimation of rs771103635 frequency.

Although the most important contribution of rs771103635 to Sandhoff disease is its impact on protein function (Tim-Aroon et al., 2021), this does not rule out other regulatory phenotypes caused by the SNP. We have shown data from the PROMO program, which indicated several transcription factors are predicted to bind to the *HEXB* DNA region near the rs771103635 SNP including GR-alpha, FOXP3, and c-Myb. A ChIP-Seq approach could help determine whether the rs771103635 SNP impairs transcription factor binding at this site, potentially contributing to the neurodegenerative phenotype observed in Sandhoff disease. The predicted transcription factor GR-alpha (aka NR3C1) is highly expressed in the brain and blood (Human Protein Atlas, n.d.), and as a result it is an ideal candidate for this approach. Future experiments can be performed using Peripheral Blood Mononuclear Cells (PBMCs), which are more accessible to obtain than brain tissue. In addition, any data received is more likely to be relevant to the central nervous system (CNS) than with other TFs that show low CNS expression. DNA could be obtained from unaffected control populations and homozygous rs771103635 individuals. After crosslinking the transcription factors to DNA and immunoprecipitating GR-alpha, the DNA sequences attached to GR-alpha could be sequenced and examined for the presence of *HEXB* DNA. Any reduction in *HEXB* DNA levels with the rs771103635 SNP would indicate that the SNP impairs TF binding potentially leading to aberrant *HEXB* gene regulation.

One additional possibility is that the SNP may impact the proper splicing or stability of *HEXB* mRNA. A strategy to test this would be to analyze mRNA from individuals heterozygous for rs771103635 in the Thai population. After RNA extraction from PBMCs and cDNA library preparation, the samples could be sequenced with long-read sequencing to generate full-length *HEXB* transcripts. The use of rs771103635 carriers means that both wildtype and rs771103635 HEXB transcripts (*HEXB*-201, *HEXB*-207, and the non-coding *HEXB*-209) could be quantified. The impact of rs771103635 on *HEXB* mRNA stability could be assessed by quantifying the total *HEXB* transcript levels with rs771103635 SNP and comparing it to the wildtype transcripts. Further, any changes in the ratios of the 3 *HEXB* transcripts to each other could indicate that this novel SNP could impact the splicing patterns of *HEXB* mRNA.

In summary, NGS data was analyzed using data sets from three different individuals for the rs771103635 SNP: a Thai female, a British male, and a Finnish female. It was determined that none of the samples contained the rs771103635 c.761T>C SNP. Although some known *HEXB* SNPs were found in exon regions, none were linked to Sandhoff disease. Finally, FreeBayes was used to ensure that our SNP was not an aberrant that resulted from alignment issues. We determined that the rs771103635 SNP was not called in the VCF data output and as a result provided our IGV results with some level of confidence because our VCF data and IGV data validated each other.

References

1. "UNESCO Panel of Experts Calls for Ban on 'Editing' of Human DNA to Avoid Unethical Tampering with Hereditary Traits." *UNESCO*, 22 Mar. 2016, en.unesco.org/news/unesco-panel-experts-calls-ban-editing-human-dna-avoid-unethical-tampering-hereditary-traits.
2. Abdelhameed, T. A., Gasmelseed, M. M., Mustafa, M. I., Abdelrahman, D. N., Abdelrhman, F. A., & Hassan, M. A. (2019). Comprehensive Analysis of HEXB Protein Reveal Forty Two Novel nsSNPs That May Lead to Sandhoff disease (SD) Using Bioinformatics. https://doi.org/10.1101/853077
3. Banerjee, P., Boyers, M. J., Berry-Kravis, E., & Dawson, G. (1994). Preferential beta-hexosaminidase (Hex) A (alpha beta) formation in the absence of beta-Hex B (beta beta) due to heterozygous point mutations present in beta-Hex beta-chain alleles of a motor neuron disease patient. The Journal of biological chemistry, 269(7), 4819–4826.
4. Bikker, H., van den Berg, F. M., Wolterman, R. A., Kleijer, W. J., de Vijlder, J. J. M., Bolhuis, P. A. Distribution and characterization of a Sandhoff disease-associated 50-kb deletion in the gene encoding the human beta-hexosaminidase beta-chain. Hum. Genet. 85: 327-329, 1990. [PubMed: 1975561]
5. ClinVar. (n.d.a). NM_000521.4(HEXB):c.362A>G (p.Lys121Arg). Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/variation/3874/
6. ClinVar. (n.d.b). NM_000521.4(HEXB):c.619A>G (p.Ile207Val). Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/variation/93202/
7. dbSNP. (n.d.a). rs10805890. Retrieved from https://www.ncbi.nlm.nih.gov/snp/rs10805890?vertical_tab=true#frequency_tab
8. dbSNP. (n.d.b). rs11556045. Retrieved from https://www.ncbi.nlm.nih.gov/snp/rs11556045?vertical_tab=true#frequency_tab
9. Dastsooz, H., Alipour, M., Mohammadi, S. et al. Identification of mutations in HEXA and HEXB in Sandhoff and Tay-Sachs diseases: a new large deletion caused by Alu elements in HEXA. Hum Genome Var 5, 18003 (2018). https://doi.org/10.1038/hgv.2018.3
10. Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. Genes & development, 25(10), 1010–1022. https://doi.org/10.1101/gad.2037511
11. Ensembl. (2021). Variant: rs771103635. Retrieved from http://useast.ensembl.org/Homo_sapiens/Variation/Population?r=5:74704810-74705810;v=rs771103635;vdb=variation;vf=325665366
12. Fitterer B, Hall P, Antonishyn N, Desikan R, Gelb M, Lehotay D. Incidence and carrier frequency of Sandhoff disease in Saskatchewan determined using a novel substrate with detection by tandem mass spectrometry and molecular genetic analysis. Mol Genet Metab. 2014;111(3):382–9.
13. Furihata, K., Drousiotou, A., Hara, Y., Christopoulos, G., Stylianidou, G., Anastasiadou, V., Ueno, I., Ioannou, P. Novel splice site mutation at IVS8 nt 5 of HEXB responsible for a Greek-Cypriot case of Sandhoff disease. Hum. Mutat. 13: 38-43, 1999.[PubMed: 9888387]
14. Galaxy. (n.d.). Manipulating NGS data with Galaxy. Retrieved from https://galaxyproject.org/tutorials/ngs/
15. GTEx Portal. (n.d.). Gene Page. Retrieved from https://gtexportal.org/home/gene/HEXB
16. Human Protein Atlas. (n.d.). Retrieved from https://www.proteinatlas.org/ENSG00000113580-NR3C1/tissue
17. IGSR (n.d.a). Sample CHI-034. Retrieved from https://www.internationalgenome.org/data-portal/sample/CHI-034
18. IGSR. (n.d.). The International Genome Sample Resource. Retrieved from https://www.internationalgenome.org/
19. IGSR. (n.d.b). Sample HG00234. Retrieved from https://www.internationalgenome.org/data-portal/sample/HG00234

20. IGSR. (n.d.c). Sample HG00171. Retrieved from https://www.internationalgenome.org/data-portal/sample/HG00171
21. Karin, M., Liu, Z. g., & Zandi, E. (1997). AP-1 function and regulation. Current opinion in cell biology, 9(2), 240–246. https://doi.org/10.1016/s0955-0674(97)80068-3
22. Kim, S., Yu, NK. & Kaang, BK. CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Med 47, e166 (2015). https://doi.org/10.1038/emm.2015.33
23. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Retrieved from https://arxiv.org/abs/1303.3997
24. MedlinePlus. (n.d.). Sandhoff disease. Retrieved from https://medlineplus.gov/genetics/condition/sandhoff-disease/
25. NCBI. (n.d.). HEXB hexosaminidase subunit beta [Homo sapiens (human)]. Retrieved from ncbi.nlm.nih.gov/gene/3074
26. NCBI: ClinVar. (2021). NM_000521.4(HEXB): c.761T>C(p.Leu254Ser). Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/variation/800653/
27. NCBI: ClinVar. (2021). NM_000521.4(HEXB):c.833C>T(p.Ala278Val). Retrieved from https://www.ncbi.nlm.nih.gov
28. Niemir, N., Rouvière, L., Besse, A., Vanier, M. T., Dmytrus, J., Marais, T., Astord, S., Puech, J. P., Panasyuk, G., Cooper, J. D., Barkats, M., & Caillaud, C. (2018). Intravenous administration of scAAV9-Hexb normalizes lifespan and prevents pathology in Sandhoff disease mice. *Human molecular genetics*, *27*(6), 954–968. https://doi.org/10.1093/hmg/ddy012
29. NIH. (2011). Sandhoff disease. Retrieved from https://rarediseases.info.nih.gov/diseases/7604/sandhoff-disease
30. NIH. (n.d.). Lysosome. Retrieved from https://www.genome.gov/genetics-glossary/Lysosome
31. NINDS Sandhoff Disease Information Page. *National Institute of Neurological Disorders and Stroke (NINDS)*. https://www.ninds.nih.gov/Disorders/All-Disorders/Sandhoff-Disease-Information-Page.
32. NORD. (n.d.). Sandhoff Disease. Retrieved from https://rarediseases.org/rare-diseases/sandhoff-disease/#:~:text=Sandhoff%20disease%20is%20a%20lipid%20storage%20disorder%20characterized,the%20beta%20subunit%20of%20the%20hexosaminidase%20B%20enzyme.
33. Norflus, F., Yamanaka, S., & Proia, R. L. (1996). Promoters for the human beta-hexosaminidase genes, HEXA and HEXB. DNA and cell biology, 15(2), 89–97. https://doi.org/10.1089/dna.1996.15.89
34. OMIM. (n.d.). OMIM Frequently Asked Questions (FAQs). Retrieved from https://omim.org/help/faq#1_4
35. OMIM. (n.d.a). 606873 HEXOSAMINIDASE B; HEXB. Retrieved from https://omim.org/entry/606873?search=hexb%20gene%20variants&highlight=gene%20hexb%20variant
36. L. Phan, Y. Jin, H. Zhang, W. Qiang, E. Shekhtman, D. Shao, D. Revoe, R. Villamarin, E. Ivanchenko, M. Kimura, Z. Y. Wang, L. Hao, N. Sharopova, M. Bihan, A. Sturcke, M. Lee, N. Popova, W. Wu, C. Bastiani, M. Ward, J. B. Holmes, V. Lyoshin, K. Kaur, E. Moyer, M. Feolo, and B. L. Kattman. "ALFA: Allele Frequency Aggregator." National Center for Biotechnology Information, U.S. National Library of Medicine, 10 Mar. 2020, www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/.
37. PROMO. (n.d.). Retrieved from http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
38. Rahmani, Z. et al. (2020). Novel homozygous HEXB mutation identified in a consanguineous Iranian pedigree with Sandhoff disease. Retrieved from https://www.researchgate.net/publication/340191536_Novel_homozygous_HEXB_mutation_identified_in_a_consanguineous_Iranian_pedigree_with_Sandhoff_disease
39. Rare Disease Database. (2021). Sandhoff disease. Retrieved from https://rarediseases.org/rare-diseases/sandhoff-disease/
40. Sandhoff disease. *Genetics Home Reference (GHR)*. 2008; http://ghr.nlm.nih.gov/condition/sandhoff-disease. Accessed 10/19/2011.

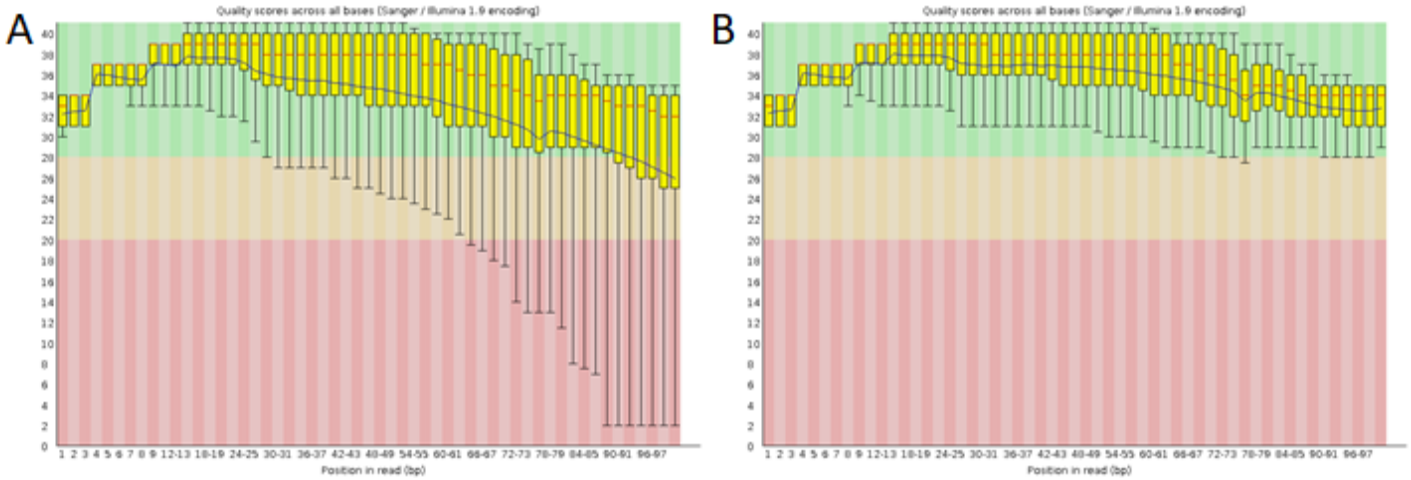
41. Tim-Aroon, T., Wichajarn, K., Katanyuwong, K., Tanpaiboon, P., Vatanavicharn, N., Sakpichaisakul, K., et.al. (2021). Infantile onset Sandhoff disease: clinical manifestation and a novel common mutation in Thai patients. BMC Pediatrics, 21(1). https://doi.org/10.1186/s12887-020-02481-3
42. UCSC Genome Browser. (n.d.). Retrieved from https://www.genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr5%3A74684134%2D74720189&hgsid=1132740911_HDmpupICyui5GFxJ7ae1ADf3eLKi
43. UniProt. (n.d.). UniProtKB - P07686 (HEXB_HUMAN). Retrieved from https://www.uniprot.org/uniprot/P07686#function
44. Wadapurkar, R.M., & Vyas. R. (2018). Computational analysis of next generation sequencing data and its applications in clinical oncology. Informatics in Medicine Unlocked. 11(75-82). https://doi.org/10.1016/j.imu.2018.05.003.
45. Wallace, A.D., & Cidlowski, J.A. (2002). Corticosteroids. Retrieved from https://www.sciencedirect.com/topics/medicine-and-dentistry/glucocorticoid-receptor-alpha

Supplementary Data

Name	Gene	Protein change	Condition(s)	Clinical significance	Accession	GRCh38 Location	dbSNP ID
NM_000521.4(HEXB):c.94C>T (p.Gln32Ter)	HEXB	Q32*	Sandhoff disease	Pathogenic	VCV000800874	74685354	rs1554034434
NM_000521.4(HEXB):c.146C>A (p.Ser49Ter)	HEXB	S49*	Sandhoff disease	Pathogenic/Likely pathogenic	VCV000555683	74685406	rs1554034452
NM_000521.4(HEXB):c.170G>A (p.Trp57Ter)	HEXB	W57*	Sandhoff disease	Pathogenic	VCV000242876	74685430	rs1114167287
NM_000521.4(HEXB):c.298C>T (p.Arg100Ter)	HEXB	R100*	Sandhoff disease	Pathogenic	VCV000623479	74685558	rs1007338250
NM_000521.4(HEXB):c.299G>C (p.Arg100Pro)	HEXB	R100P	Sandhoff disease	Pathogenic	VCV000397589	74685559	rs1060499701
NM_000521.4(HEXB):c.333G>A (p.Trp111Ter)	HEXB	W111*	Sandhoff disease	Pathogenic	VCV000557986	74689361	rs761117459
NM_000521.4(HEXB):c.445+1G>A	HEXB		Sandhoff disease	Pathogenic	VCV000633263	74689474	rs761197472
NM_000521.4(HEXB):c.508C>T (p.Arg170Ter)	HEXB	R170*	Sandhoff disease|not provided	Pathogenic/Likely pathogenic	VCV000281002	74693701	rs753823903
NM_000521.4(HEXB):c.552T>G (p.Tyr184Ter)	HEXB	Y184*	Sandhoff disease|not provided	Pathogenic	VCV000280067	74696733	rs573447174
NM_000521.4(HEXB):c.558+1G>C	HEXB		Sandhoff disease	Pathogenic/Likely pathogenic	VCV000853457	74696740	
NM_000521.4(HEXB):c.796T>G (p.Tyr266Asp)	HEXB	Y266D, Y41D	not provided|Sandhoff disease	Pathogenic/Likely pathogenic	VCV000381669	74713530	rs373979283
NM_000521.4(HEXB):c.839T>G (p.Leu280Ter)	HEXB	L55*, L280*	Sandhoff disease	Pathogenic	VCV000664928	74713573	rs1579950499
NM_000521.4(HEXB):c.841C>T (p.Arg281Ter)	HEXB	R281*, R56*	Sandhoff disease	Pathogenic	VCV000623308	74713575	rs138914144
NM_000521.4(HEXB):c.850C>T (p.Arg284Ter)	HEXB	R284*, R59*	not provided|Sandhoff disease, infantile type|Sandhoff disease	Pathogenic	VCV000003887	74713584	rs121907986
NM_000521.4(HEXB):c.1082+5G>A	HEXB		Sandhoff disease	Pathogenic/Likely pathogenic	VCV000354135	74715695	rs5030731
NM_000521.4(HEXB):c.1082+5G>C	HEXB		Sandhoff disease, infantile type	Pathogenic	VCV000003885	74715695	rs5030731
NM_000521.4(HEXB):c.1144A>T (p.Lys382Ter)	HEXB	K157*, K382*	Sandhoff disease	Pathogenic	VCV001069640	74716648	
NM_000521.4(HEXB):c.1243-2A>G	HEXB		not provided|Sandhoff disease	Pathogenic/Likely pathogenic	VCV000093197	74718795	rs398123446
NM_000521.4(HEXB):c.1250C>T (p.Pro417Leu)	HEXB	P417L, P192L	not provided|Sandhoff disease, juvenile type|Sandhoff disease, adult type|Sandhoff disease	Pathogenic	VCV000003878	74718804	rs28942073
NM_000521.4(HEXB):c.1389C>G (p.Tyr463Ter)	HEXB	Y463*, Y238*	Sandhoff disease	Pathogenic/Likely pathogenic	VCV000557658	74718943	rs1554036943

NM_000521.4(HEXB):c.1481A>G (p.Asp494Gly)	HEXB	D494G, D269G	Sandhoff disease	Pathogenic	VCV000869190	74720491	
NM_000521.4(HEXB):c.1509-26G>A	HEXB		Sandhoff disease	Pathogenic/Likely pathogenic	VCV000527971	74720617	rs201580118
NM_000521.4(HEXB):c.1514G>A (p.Arg505Gln)	HEXB	R505Q, R280Q	Sandhoff disease|Sandhoff disease, adult type	Pathogenic/Likely pathogenic	VCV000003879	74720648	rs121907983
NM_000521.4(HEXB):c.1538T>C (p.Leu513Pro)	HEXB	L288P, L513P	Sandhoff disease	Pathogenic	VCV000929001	74720672	
NM_000521.4(HEXB):c.1578T>G (p.Tyr526Ter)	HEXB	Y301*, Y526*	Sandhoff disease	Pathogenic	VCV001071217	74720712	
NM_000521.4(HEXB):c.1597C>T (p.Arg533Cys)	HEXB	R533C, R308C	Sandhoff disease	Pathogenic	VCV000435415	74720731	rs764552042

Supplementary Table 1. ClinVar variant table – *HEXB* SNPs



Supplementary Figure 2. Sequence Processing and Quality Control. FASTQC was run on all sequencing data files prior to processing, showing significant numbers of poor-quality reads. An example from the Thai female sample forward strand is shown (A). After trimming using the Trimmomatic program, FASTQC was re-run. The data indicates that reads have been effectively trimmed (B).