

**Based on PMID: 34172899**

## **KARS1 rare variant, c.1772T>A, may interfere with Lysine-tRNA activity**

### **Abstract**

Patients with KARS1 variants show neurological symptoms that are often associated with neurodevelopment. Our novel SNP c1772A>T (p. Asn591Ile) identified in Lin et al. 2021 located on exon 14 was found among five patients in three different families from Egyptian and European Ethnicity; all homozygous for N591I variant. Though KARS1 variants result in the manifestation of various neurodevelopmental diseases, it is unclear whether there is a correlation between genotypes and phenotypes. Our hypothesis is that our KARS1 variant is rare and pathogenic because it interferes with enzymatic activity of KARS1 and therefore, it is responsible for causing certain phenotypes of neurodevelopmental diseases. To test our hypothesis, we performed computational structural analysis and next generation sequencing analysis from various datasets which were generated from studying various neurodevelopmental diseases. We identified SNP N591I in the ExAc project and confirmed that it is indeed a rare variant which has an allele frequency of 0.00001647. Structural analysis also revealed that it may interfere with KARS1 enzymatic activity by making a covalent bond between KARS1 and tRNA which would prevent binding between ATP and tRNA.

### **Introduction**

*KARS1* encodes lysine-tRNA synthetase. This protein belongs to the family of aminoacyl-tRNA synthetases (ARSs). These enzymes are essential, ubiquitously expressed and highly conserved in three major kingdoms of life. Their main function is transfer RNA (tRNA) charging and editing through catalytic and anticodon recognition domains which allow them to attach an amino acid to its cognate tRNA covalently and/or remove the wrong amino acids from their own cognate tRNA [1].

ARSs catalyze esterification of an amino-acid (Lysine for KARS1) to the 3' end of a tRNA along with hydrolysis of an ATP molecule. This reaction yields aminoacyl-tRNA, AMP and PPi. This occurs in two-steps. In the first step, amino acid is activated and both Lysine and ATP bind to the catalytic site of KARS1. This causes a nucleophilic attack of the  $\alpha$ -carboxylate oxygen of Lysine to the  $\alpha$ -phosphate group of the ATP,

leading to condensation and forming of aminoacyl- adenylate (aa-AMP). aa-AMP remains bound to the enzyme and PPi is expelled from the active site. This two-step reaction is known to be universally conserved for the KARS1 protein.

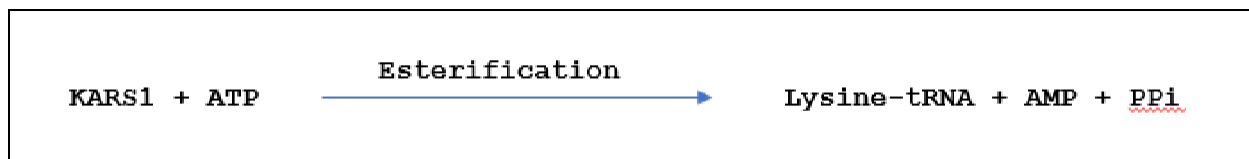


Figure 1: Esterification mechanism of KARS1 with ATP.

Unlike other ARSs, lysine-tRNA synthetase encoded by the KARS1 gene is bifunctional in cytoplasm and mitochondria [2]. The bifunctionality of the gene is achieved via alternative splicing. *KARS1* is located on the complementary strand of chromosome 16 (genomic location, hg38) and has 15 exons. There are two different isoforms (transcript variants) of *KARS1* which are known to form the 597aa cytoplasmic and 625 aa mitochondrial enzymes. While the last 576 amino acids of the two isoforms are identical, the mitochondrial enzyme differs in 49 amino acids at the N terminal with a different mitochondrial targeting sequence. Here the cytoplasmic isoform skips exon2 and splices exon 1 and exon 3, whereas the mitochondrial isoform includes exon 2 [3].

Three variants of *KARS1* in Charcot-Marie-Tooth (CMT) patients were identified in 2010: p.L133H, p.Y173SfxX7, and p.I302M. The first two variants lead to loss-of-function mutations which severely disrupt enzymatic activity [4]. Since then, numerous studies have identified pathogenic *KARS1* variants (type of mutation) and many of them are shown to alter the secondary structure of the protein, causing it to aggregate and ultimately interfering with their enzymatic activity [5]. *KARS1* variants are predominantly associated with the broad spectrum of neurological diseases such as CMT, leukodystrophies, leukoencephalopathy, epilepsy, congenital visual impairment, microcephaly, and nonsyndromic hearing impairment. There is significant phenotypic variability among patients with *KARS1* variants which seem to depend on biallelic mutations of the gene [6].

Our novel SNP c1772A>T (p. Asn591Ile) from Lin et al. 2021 located on exon 14 was found among five patients in three different families from Egyptian and European Ethnicity; all homozygous for N591I variant. The common features of these five patients include hearing loss, epilepsy, developmental delay, intellectual disability, seizures, and speech delay. Considering the neuroimaging abnormalities, three of the five patients exhibited normal MRI. Two patients from the same family were detected with reduced white matter volume, ventricular dilatation, hypoplasia of the pons and corpus callosum and cortical atrophy. Some of the patient specific abnormalities studied in these clinical

trials were leukodystrophy, behavioral abnormalities, dysmorphic facial features, uncontrolled urine and stools, polyneuropathy, hypotonia, hyporeflexia [7].

Though *KARS1* variants result in the manifestation of various neurodevelopmental diseases, it is unclear whether there is a correlation between genotypes and phenotypes. Our hypothesis is that our *KARS1* variant is pathogenic because it interferes with enzymatic activity of *KARS1* and therefore, it is responsible for causing certain phenotypes of neurodevelopmental diseases. To test our hypothesis, we performed computational structural analysis and next generation sequencing analysis from various datasets which were generated from studying various neurodevelopmental diseases. We took two different approaches to perform NGS analysis. Our first approach was to use FASTQ files from the Sequence Read Archive (SRA) database which allowed us to filter phenotypes that are associated with pathogenic *KARS1* variants. Our second approach was to obtain VCF files from the Harvard Personal Genome Project and ExAC project. We were unable to find our SNP in SRA databases and the Harvard Personal Genome Project. We were able to find our SNP in the ExAC project which identified new variants for rare diseases.

## **Materials and Methods**

### **DNA-Seq data**

Fastq raw data was taken from two datasets to perform DNA-Seq for identification and annotation of SNP N591I. We choose the genome of a boy with intellectual disabilities from the SRA database. (ERR5195739) This was a paired end human genome submitted by 'UOC Genetica Medica Poliambulatorio "Giovanni Paolo II"'. With the NextSeq100 instrument used for sequencing, this RNA-seq data was extracted from peripheral blood samples.

Next, the variant calling file (vcf) from ExAC project was selected to identify our SNP in a dataset having variants from a population of ~ 60,000 individuals. The average ExAC participant has about 54 variants previously classified as casual for a rare disorder. Such incorrectly classified data are reviewed by the ExAC project to correctly link pathogenic genetic variants with rare diseases. The ExAC project identified more than 7.4 million variants that are mostly new and by analyzing independently emerging mutations it estimated for the first time the frequency of recurrence of rare mutations. The 7.4 million variants were identified at high confidence. This finding offers understanding of rare genetic variation across populations. It also identifies 3,230 genes that show nearly no cases of loss of function.

### **DNA-Seq read mapping and Variant Detection**

DNA-Sequencing pipeline was performed for the human genome extracted from the SRA dataset. The genome was quality checked using FastQC, grooming was performed for Phred encoding of Illumina 1.8+. The sequences were trimmed to remove the misleading data from the ends of the fragments. With a sliding trimming window, 4 number of bases to average across and an average quality of 20 bases; sequences were trimmed using trimmomatic. Alignment of the trimmed sequences was performed using three programs available on Galaxy, namely; HISAT2, Bowtie and BWA (Burrows Wheelers Alignment). Built-in human (hg 19/GrCh38) genome from the Galaxy platform and all other default parameters were used for this purpose. HISAT uses hierarchical indexing for spliced alignments of transcripts. While Bowtie2 indexes the genome with an FM index, BWA concatenates multiple reference sequences into one long sequence and this multi-aligned reference is then used as a reference. These alignments were then called for SNPs using Freebayes, a haplotype-based approach which calls the SNPs based on the literal sequences of the reads aligned. Region was restricted to chromosome 16 and SNPs were scanned to identify our novel SNP. While the N591I specific SNP was not identified, we could identify previously known KARS1 variants (P127L and L233V from the human genome of a boy with intellectual disabilities).

The rare variant calling .vcf file from the ExAC project was used for annotation of our SNP. From the Galaxy web server, multiple VCF filters were applied to filter a variety of attributes. A specific filtering value of AC=2 and a KARS specific region of chr16:75,661,622-75,681,585 extracted a large number of rare variants present in KARS. VCF filter was again further applied to a much specific KARS exon 14 region to look for the desired novel SNP. VCF annotation was used on the tabix indexed file to extract SNP 1772 A<T.

## Results

### Forming a covalent bond between KARS1 and tRNA

Structural Analysis shows Asparagine at 591 position (N591) in KAR1 to be present in the binding interface for ATP and tRNA. Hence this SNP is expected to affect the binding to these substrates. Computational mutation of N591I leads to a decrease in distance between N591 and G670 of tRNA. While the distance between N591\_OD1 and G670\_OP1 was 4.985 Å, the distance between I591\_CD1 and G670\_OP1 after mutation was reduced to 3.497 Å (Figure 3 a, b) . Considering a threshold of 4 Å, this mutation may reduce the distance at the interface by ~ 1.5 Å and approach to form a covalent bond. Also, a mutation of N591 to I591 would be significant since the residue changes from being a polar, uncharged group (N591) to a Nonpolar aliphatic group (I591).

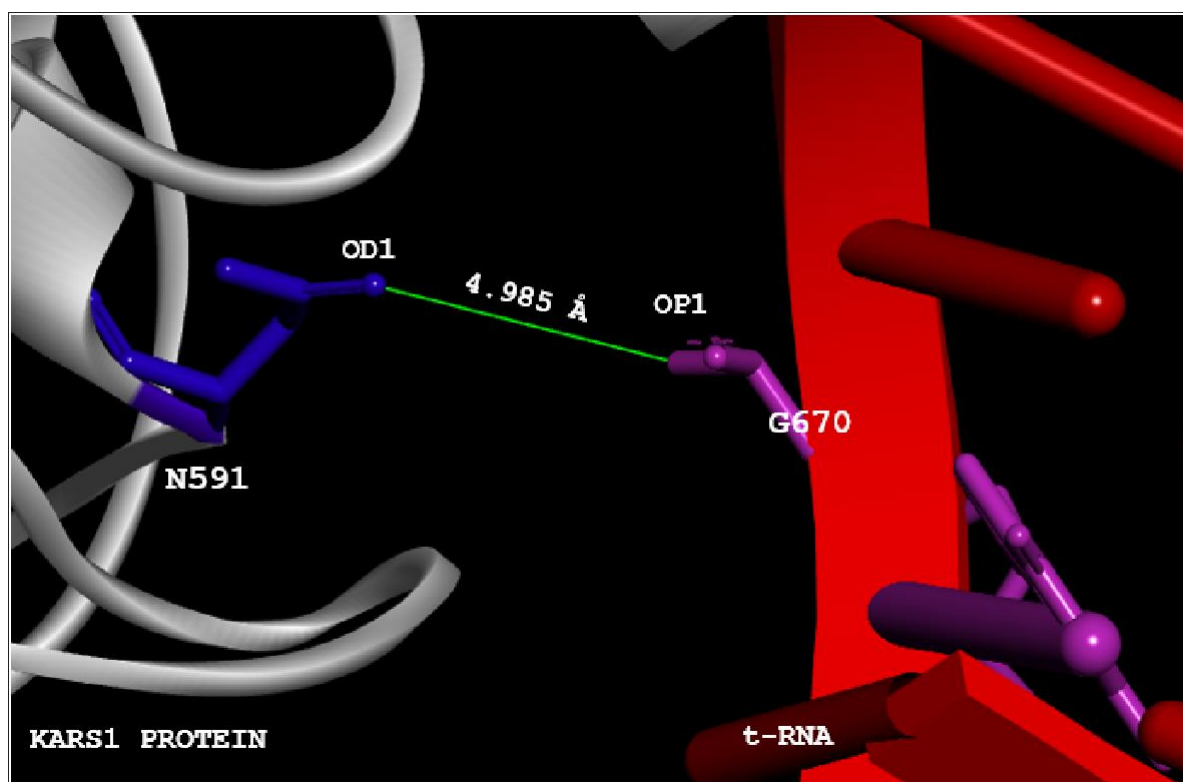


Figure 2 a) Distance between N591 and G670 before mutation. ( $4.985 \text{ \AA} > \text{threshold } 4 \text{ \AA}$ )

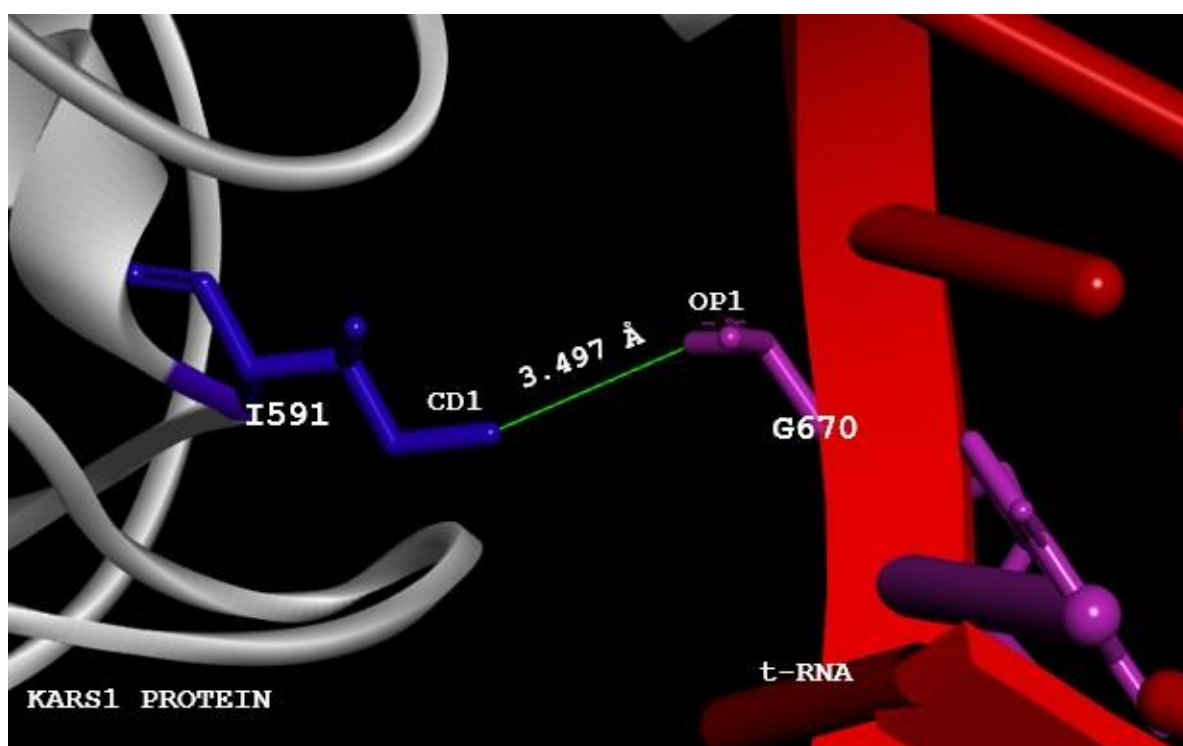
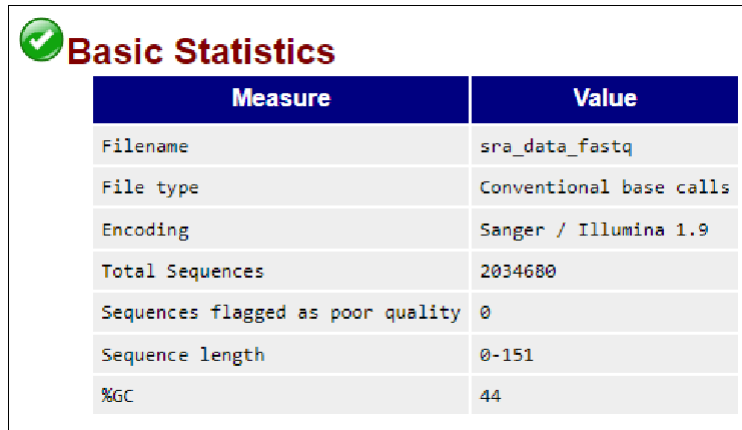


Figure 2 b) Distance between I591 and G670 after mutation ( $3.497 \text{ \AA} < \text{threshold } 4 \text{ \AA}$ )

## FASTQC and Trimming low quality ends

Finding and annotating our variant from an appropriate dataset was crucial to this study. For these purposes, Sequence Read Archive (SRA) was browsed for datasets of subjects having disorders manifested during neurodevelopment such as intellectual disability, epilepsy, hearing loss, schizophrenia, autism spectrum disorder. Next generation sequencing followed by variant calling of selected experiments from SRA revealed the absence of our variant in these datasets. One such dataset, run accession #ERR5195739 was used. Prior to running the workflows, the quality of the sequencing data obtained from the SRA dataset was analyzed. First, general statistics were collected using FastQC 0.11.4, a widely-used quality control application for data generated on Illumina platform[17]. While Fastqc showed normal results for Basic Statistics, Per base sequence quality, Per sequence quality scores, Per base N content and the Adapter content; it showed slightly abnormal values for Per Sequence GC content and highly unusual values for Per base sequence content, Sequence Length Distribution, Sequence Duplication Levels and Overrepresented Structures. Overall the sequence appeared to have stable quality.

Next the sequences were preprocessed using Trimmomatic to remove low quality ends of the transcripts [16].

A screenshot of the FastQC Basic Statistics report. It features a green checkmark icon and the title 'Basic Statistics' in red. Below the title is a table with two columns: 'Measure' and 'Value'. The table lists seven metrics: Filename (sra\_data\_fastq), File type (Conventional base calls), Encoding (Sanger / Illumina 1.9), Total Sequences (2034680), Sequences flagged as poor quality (0), Sequence length (0-151), and %GC (44).

Measure	Value
Filename	sra_data_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2034680
Sequences flagged as poor quality	0
Sequence length	0-151
%GC	44

*Figure 3: Basic Statistics of FASTQC Read Quality Report. The sequence length varies from 0-150 with a total of 2034680 sequences. Encoding method was Illumina 1.9*

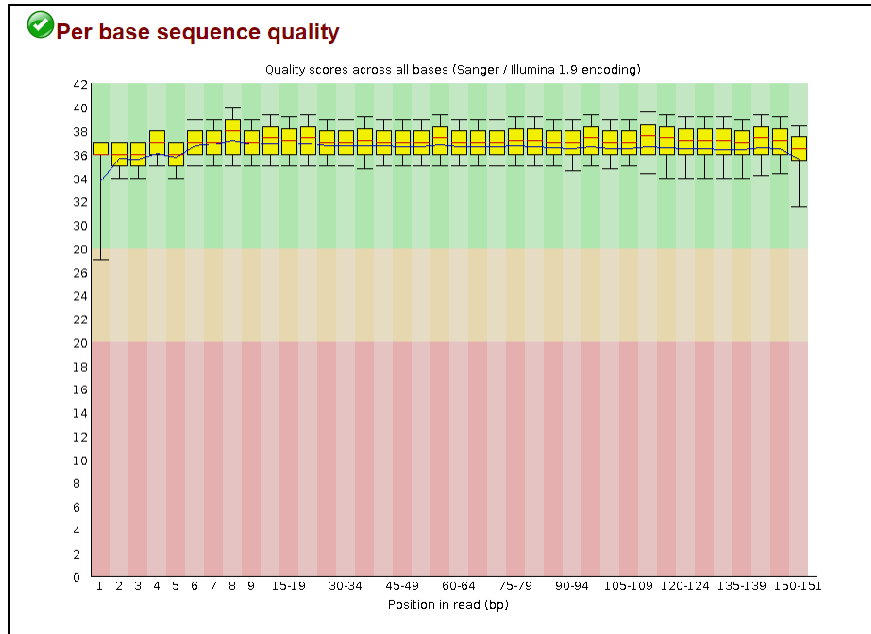


Figure 4: Per base sequence quality of ERR5195739 before sequence alignment

### Different RNA-Seq alignment algorithms call for different variants

Trimmed Sequences were aligned to the built-in human reference genome (hg19/GhCr37) using HISAT, Bowtie and BWA (Burrows Wheeler Alignment) [12,13,15]. While Freebayes identified 34 SNPs in HISAT2 based alignment, 56 SNPs for Bowtie and 67 SNPs for BWA based alignments were identified [14]. While previously known KARS1 variants (P127L and L233V) were identified using HISAT2 alignment, KARS1 SNP 1772A>T (N591I) was not identified using any alignment methods. This leads to the fact that SNP 1772 A>T is not responsible for a single phenotype but could be identified in patients having broad neurological features.

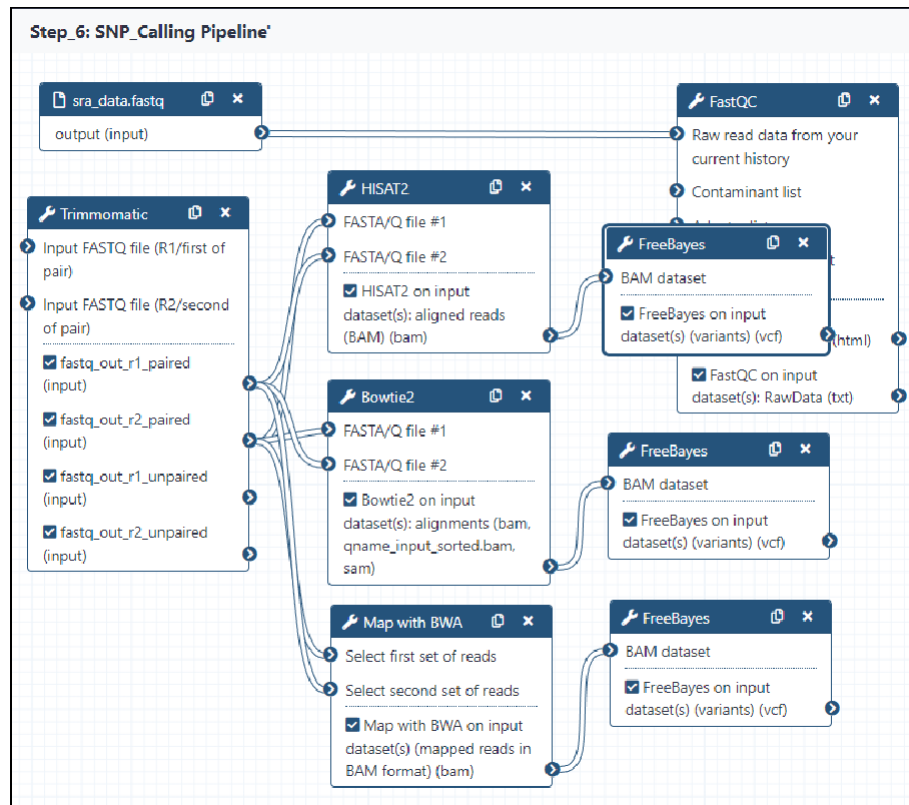


Figure 5: Workflow of NGS for SRA dataset, ERR5195739 on Galaxy

### Identification of SNP in ExAC dataset

VCF files of datasets published from 2011 through 2020 in the Personal Genome Project (PGP) were searched to identify datasets which mentioned any neurological phenotypes. On not finding our SNP in any of these files, we shifted our focus to the ExAC dataset. The ExAC dataset includes sequenced exon data from 60,706 unrelated subjects, all mapped to GRCh37/hg19 reference sequence. The variant dataset downloaded in the VCF format from GnomAD (Genome Aggregation Database) includes ~19,000,000 lines of annotated variants from the exomes of all chromosomes for these individuals. The enormous size of the file prevents the addition of a figure for this vcf file. To confirm the presence of our variant c.1772T>A (NM\_001130089.2) in this dataset, which is present at chr 16:75,662,474 (GRCh37), the results were filtered to create a manageable file. Figure 1 shows the presence of c.1772T>A.



16	75661835	rs144390077	G	C	5271.99	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75661841	.	T	C	13653.8	PASS	AC=2;AC_AFR=0;AC_AMR=2;AC_Adj=
16	75661842	.	A	T	14684.2	PASS	AC=2;AC_AFR=0;AC_AMR=2;AC_Adj=
16	75661871	.	G	A	8008.91	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75661872	.	G	A	4006.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75662474	.	T	A	1566.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75663443	.	G	C	4914.99	PASS	AC=2;AC_AFR=2;AC_AMR=0;AC_Adj=
16	75663464	.	G	A	5459.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75663465	.	G	T	6279.3	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665274	.	G	A	1853	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665288	.	C	T	1025.42	PASS	AC=2;AC_AFR=1;AC_AMR=0;AC_Adj=
16	75665353	.	C	T	6394.99	PASS	AC=2;AC_AFR=2;AC_AMR=0;AC_Adj=
16	75665403	.	G	A	3872.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665445	.	T	C	2400.85	VQSRTTrancheSNP99.60to99.80	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665482	.	C	T	7538.39	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665494	rs16941301	G	T	8.72242e+06	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665508	.	T	C	3473.85	PASS	AC=2;AC_AFR=0;AC_AMR=1;AC_Adj=
16	75665610	.	G	T	4504.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=
16	75665679	.	C	T	11703.7	PASS	AC=2;AC_AFR=1;AC_AMR=0;AC_Adj=
16	75665701	.	C	T	8027.85	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=

Figure 6: Results from VCFannotate tool showing c.1772T>A variant at highlighted position. Columns 4 and 5 in the figure show the canonical and variant alleles at the locus.

## VCF annotation for SNP

The annotation shows an alternate allele count of 2 out of the total 121412 number of alleles in the called genotype. The allele frequency of 0.00001647 suggests how rare this variant is. Despite the lower allele count and frequency of this variant, the approximate read depth of 2393284 at this locus shows the confidence of this called variant. Higher read depth also provides a better chance of finding low frequency mutations. This variant also shows adjusted allele and adjusted heterozygous counts of

2. As compared to females, the number of allele counts among males was higher. Out of the total of 54054 allele number among females and 67262 allele number among males, an allele count of 2 was found among females, while none was found in males. The ExAC VCF file only gives the total count of chromosomes observed and the count of chromosomes exhibiting the alternative allele [8]. The African/African American Chromosome Count is 10390, American Chromosome Count is 11560, South Asian Chromosome Count is 16506, East Asian Chromosome Count is 8654, Finnish Chromosome Count is 6602, Non-Finnish European Chromosome Count is the maximum among these, 66698 and the Other Chromosome Count is 906. The adjusted Chromosome Count is 121316. Non-Finnish European Allele Counts is 2. The total number of alleles in the population with maximum allele frequency for each alternate allele is 66698, and the alternate allele count for these is 2. Histogram of ages of

heterozygous allele carries 1 for the bin of 75 years. The Z-score from Wilcoxon rank sum test of Alternate Vs. Reference base qualities is 0.850. This test compares the base qualities of the data supporting the reference allele with the base qualities of data supporting the alternate allele. A Z-score close to 0 suggests little or no difference. Positive value may suggest bases supporting alternate allele have higher quality scores than the ones supporting reference allele, however, finding a difference may also point to a sequencing process that may have been biased or affected by an artifact [9].

Ensembl Variant Effect Predictor (VEP) detects the effect of variants on genes, transcripts, protein sequence, and regulatory regions. The Consequence annotations from Ensembl VEP (CSQ) are found on the ExAC VCF file. They describe that our SNP affects the canonical protein coding transcript ENST00000319410 (c.1772A>T, missense variant) at exon 14 (out of 15) of ENSG00000065427, KARS1 gene on the negative strand, leading to consequential protein change of ENSP00000325448.5 (p.Asn591Ile). The impact of the consequence is moderate. The cDNA position of this variant is 1894. The existing refSNP ID of the variant is also indicated as rs746483110. The CCDS identifier for the transcript is CCDS45532.1. The source and identifier of protein domains that overlap this region is HAMAP:MF\_00252, PROSITE\_profiles:PS50862, hmmpanther:PTHR22594:SF4, hmmpanther:PTHR22594, TIGRFAM\_domain:TIGR00499, Gene3D:3.30.930.10, Pfam\_domain:PF00152, Superfamily\_domains:SSF55681. The identifiers for UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and UniParc are Q15046, J3KRL2/H3BVA8, UPI00001405CB, respectively. The SIFT prediction for this variant affecting the protein function is deleterious(0) and PolyPhen prediction is probably\_damaging(0.998). The GENE\_PHENO score of 1 indicates that the gene is associated with a phenotype, disease, or trait. Frequency of existing variant in ExAC African/American population is A:1.647e-05, in ExAC East Asian population is A:1.649e-05, in ExAC combined other combined populations is A:2.999e-05.

Another protein coding transcript affected by this variant is ENST00000302445.3 (c.1688A>T, missense variant) at exon 13 (out of 14) of KARS1 causing a protein change of ENSP00000303043.3 (p.Asn563Ile). The CCDS identifier for the transcript is CCDS10923.1. The source and identifier of protein domains that overlap this region is PROSITE\_profiles:PS50862, HAMAP:MF\_00252, hmmpanther:PTHR22594:SF4, hmmpanther:PTHR22594, TIGRFAM\_domain:TIGR00499, Pfam\_domain:PF00152, Gene3D:3.30.930.10, Superfamily\_domains:SSF55681.

Our variant is classified as a downstream gene variant for nonsense mediated decay transcript ENST00000562875, and 3' UTR variant and non-sense mediated decay (NMD) transcript variant for another nonsense mediated decay transcript ENST00000564578. It is an intron variant for protein coding transcript

ENST00000568378 (c.147-587A>T) with no consequence on protein ENSP00000454512, and a non-coding transcript exon variant for retained intron transcript ENST00000569298 (n.434A>T). All these results are given for the negative strand.

## Discussion

Patients with KARS1 variants show neurological symptoms that are often associated with neurodevelopment. Therefore, our initial approach to find our SNP was to use the SRA database by searching numerous disorders manifested during neurodevelopment such as intellectual disability, epilepsy, hearing loss, schizophrenia, autism spectrum disorder. Search was also performed using a combination of parameters such as “Human AND brain AND RNA-seq” to filter miRNA-sequencing or any such strategies. An example of our analysis is shown in the SRA Database Approach section with EXR500401. We analyzed fastq files on Galaxy from the following SRA run accession numbers: ERR103430, ERR1378632, ERR1739481 and many more, however, our SNP did not exist in any of the datasets.

There are mainly two limitations in our datasets from this section. One is that we were only able to select studies that had manageable fastq file sizes (the largest file we had from the initial assignment submission was about 500Mb) and if we wanted to complete the analysis in time, we could not select many of them due to the file sizes being too large. With better computing power and larger storage space, we would be able to analyze more datasets. Another limitation was that we were only able to use datasets that were publicly available. There are many deposited datasets on SRA/dbGaP that require access which none of us had. Unfortunately, this limited the number of studies we could use. Although we had these limitations, we were nonetheless able to run a decent number of datasets available through the SRA database. Yet, we did not find our SNP in these datasets. After the initial assignment submission, we continued our search in SRA datasets by expanding the file size limit and phenotypes of patients. Though we ran dozens more of SRA datasets from patients with epilepsy, Aicardi-Goutières syndrome, attention-deficit/hyperactivity disorder, and posttraumatic stress disorder, we were not able to find our SNP in any of them.

Our next approach was to go through PGP. We went through reports of datasets published from 2011 through 2020 to identify datasets which mentioned any neurological phenotypes. Once we identified these, we downloaded their files (tsv format) which contain all of SNPs picked up from an individual. Though there are a decent number of datasets which contain neurological phenotypes, we could not find our SNP in any of them. It is likely that our SNP is considered a rare variant and that

symptoms are manifested during neurodevelopment in patients. We observed some SNPs found in these datasets that are linked to neurodegenerative diseases. We realized that we might not be able to find our SNP in adults who had healthy neurodevelopment.

Our final approach was to use datasets from the ExAC project which picked up many new variants that are relatively rare. Since our variant was exonic, this dataset was also chosen. We were able to find our SNP, though the dataset was relatively large, showing ~19,000,000 lines of results. Finding our SNP in the collection of rare variants datasets suggests this is probably a rare variant. We discovered our SNP in the ExAC project data corresponding to two heterozygous patients for N591I variant. The allele frequency is calculated as 0.00001647. Among the five patients identified in the study by Lin et al., three individuals were homozygous, and two other individuals were heterozygous for N591I variant. Given that the allele frequency for the heterozygous variant is 0.00001647, this may indicate that being homozygous for the N591I variant is even rarer. This rarity of this variant in general certainly contributes to why we were not able to identify this variant in the “small” patient datasets that we have run.

The search for our SNP also led us to question whether homozygosity for this variant is required for developing severe neurological symptoms. Based on the study from Lin et al., two individuals were heterozygous for the variant. This may indicate that the heterozygous mutation of the variant is haploinsufficient. Computational structural analysis indicated that having amino acids changed from N591 to I591 may increase the chance of forming a covalent bond between KARS1 and tRNA. If we were able to test this in wet lab experiments, it would be valuable to compare how both heterozygous mutations and homozygous mutations would affect its enzymatic activity.

In conclusion, we confirmed that our novel variant is indeed a rare variant which has an allele frequency of 0.00001647. Structural analysis revealed that the KARS1 variant is present in the binding interface for ATP and tRNA. The KARS1 variant is in close proximity to tRNA and that these two may form a covalent bond on mutation. This may prevent binding between ATP and tRNA, and also cause steric hindrances resulting in interference of KARS1 enzymatic activity.

## Reference

[1] Kwon, N. H., Fox, P. L., & Kim, S. (2019). Aminoacyl-tRNA synthetases as therapeutic targets. *Nature Reviews Drug Discovery*, 18(8), 629–650.  
<https://doi.org/10.1038/s41573-019-0026-3>

[2] Fuchs, S. A., Schene, I. F., Kok, G., Jansen, J. M., Nikkels, P. G. J., van Gassen, K. L. I., Terheggen-Lagro, S. W. J., van der Crabben, S. N., Hoeks, S. E., Niers, L. E. M., Wolf, N. I., de Vries, M. C., Koolen, D. A., Houwen, R. H. J., Mulder, M. F., & van Hasselt, P. M. (2019). Aminoacyl-tRNA synthetase deficiencies in search of common themes. *Genetics in Medicine*, 21(2), 319–330. <https://doi.org/10.1038/s41436-018-0048-y>

[3] Tolkunova, E., Park, H., Xia, J., King, M. P., & Davidson, E. (2000). The human lysyl-tRNA synthetase gene encodes both the cytoplasmic and mitochondrial enzymes by means of an unusual alternative splicing of the primary transcript. *The Journal of biological chemistry*, 275(45), 35063–35069. <https://doi.org/10.1074/jbc.M006265200>

[4] McLaughlin, H. M., Sakaguchi, R., Liu, C., Igarashi, T., Pehlivan, D., Chu, K., Iyer, R., Cruz, P., Cherukuri, P. F., Hansen, N. F., Mullikin, J. C., Biesecker, L. G., Wilson, T. E., Ionasescu, V., Nicholson, G., Searby, C., Talbot, K., Vance, J. M., Züchner, S., ... Antonellis, A. (2010). Compound heterozygosity for loss-of-function lysyl-trna synthetase mutations in a patient with peripheral neuropathy. *The American Journal of Human Genetics*, 87(4), 560–566. <https://doi.org/10.1016/j.ajhg.2010.09.008>

[5] Ardisson, A., Tonduti, D., Legati, A. et al. KARS-related diseases: progressive leukoencephalopathy with brainstem and spinal cord calcifications as new phenotype and a review of literature. *Orphanet J Rare Dis* 13, 45 (2018). <https://doi.org/10.1186/s13023-018-0788-4>

[6] Sun, C., Song, J., Jiang, Y., Zhao, C., Lu, J., Li, Y., Wang, Y., Gao, M., Xi, J., Luo, S., Li, M., Donaldson, K., Oprescu, S. N., Slavin, T. P., Lee, S., Magoulas, P. L., Lewis, A. M., Emrick, L., Lalani, S. R., ... Zhang, V. W. (2019). Loss-of-function mutations in lysyl-trna synthetase cause various leukoencephalopathy phenotypes. *Neurology Genetics*, 5(2). <https://doi.org/10.1212/nxg.0000000000000316>

[7] Lin, S. J., Vona, B., Barbalho, P. G., Kaiyrzhanov, R., Maroofian, R., Petree, C., Severino, M., Stanley, V., Varshney, P., Bahena, P., Alzahrani, F., Alhashem, A., Pagnamenta, A. T., Aubertin, G., Estrada-Veras, J. I., Hernández, H., Mazaheri, N., Oza, A., Thies, J., Renaud, D. L., ... Varshney, G. K. (2021). Biallelic variants in KARS1 are associated with neurodevelopmental disorders and hearing loss recapitulated by the knockout zebrafish. *Genetics in medicine : official journal of the American College of Medical Genetics*, 23(10), 1933–1943. <https://doi.org/10.1038/s41436-021-01239-1>

[8] Pedersen, B.S., Layer, R.M. & Quinlan, A.R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol* 17, 118 (2016).  
<https://doi.org/10.1186/s13059-016-0973-5>

[9] Rank Sum Test, GATK (2020).  
<https://gatk.broadinstitute.org/hc/en-us/articles/360035531952-Rank-Sum-Test>

[10] Kobayashi, Y., Yang, S., Nykamp, K. et al. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med* 9, 13 (2017). <https://doi.org/10.1186/s13073-017-0403-7>

[11] GDC VCF Format, National Cancer Institute GDC documentation, Retrieved from: [https://docs.gdc.cancer.gov/Data/File\\_Formats/VCF\\_Format/](https://docs.gdc.cancer.gov/Data/File_Formats/VCF_Format/)

[12] Langmead, B., Trapnell, C., Pop, M. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).  
<https://doi.org/10.1186/gb-2009-10-3-r25>

[13] Kim, D., Paggi, J.M., Park, C. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37, 907–915 (2019).  
<https://doi.org/10.1038/s41587-019-0201-4>

[14] Yao, Z., You, F.M., N'Diaye, A. et al. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* 21, 360 (2020).  
<https://doi.org/10.1186/s12859-020-03704-1>

[15] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

[16] Anthony M. Bolger, Marc Lohse, Bjoern Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, Volume 30, Issue 15, 1 August 2014, Pages 2114–2120, <https://doi.org/10.1093/bioinformatics/btu170>

[17] Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., & Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in genetics*, 4, 288.  
<https://doi.org/10.3389/fgene.2013.00288>