SNP rs2291418 role in Alzheimer's Disease

Jaimee Beckett[1], Heidi Lee[1], Jiamin Lian[1], Josie Jenyne Lu[1], Kennedi Todd[1]
[1]Johns Hopkins University, Baltimore, MD

Abstract

MicroRNAs (miRNAs) are small RNAs ranging from 16 - 27 nucleotides in length. They can alter gene expression by binding to promoters or the 3' or 5' UTR of mRNAs. Rs2291418 is a SNP potentially associated with Alzheimer's disease (AD) and is located in the intronic region of *MGAT4B* and the exonic region of *MIR1229*. The focus of our study is on *MIR1229* because its target is *SORL1*, which is involved in the processing and trafficking of amyloid precursor protein (APP), the precursor to Aβ. Accumulation of Aβ in brain tissue leads to build-up of plaques and potential development of AD, both as late-onset AD and familial AD. SNP rs2291418 is thought to contribute to AD progression through its regulation of *SOLR1*, as the SNP has been found to cause an increase in the production of *MIR1229*. An increase of the mature product of *MIR1229*, hsa-mir-1229-3p, leads to a greater downregulation of *SORL1* and an increase of Aβ, which increases the risk of developing AD. In our study, we examine a miRNA-Seq data set to identify the presence of the SNP among AD samples and control samples to determine if there exists a relationship between the SNP and AD patients. Three out of twenty samples had the rs2291418 SNP. One of these was a control sample and two AD samples. All three were female samples. There was no significant difference between SNP association in control vs AD groups but there was in female vs male.

1. Introduction

1.1 miRNA overview

MicroRNAs (miRNAs) are non-coding RNAs, averaging about 22 nucleotides in length. miRNA studies have shown that miRNAs are involved in gene regulation [26]. Most miRNAs bind to the 3' UTR of target mRNAs to repress gene expression. However, miRNAs bound to the 5' UTR and coding sequences have shown to have a silencing effect. miRNAs bound to promoter regions induce transcription. The miRNA strand with the lower 5' stability or 5' uracil will load into AGO and become the guide strand, then the unloaded strand is the passenger strand that unwinds from the guide strand [26]. miRNA-induced silencing complex (miRISC), which is the guide strand and the AGO protein, interacts with the target mRNA. The degree of complementarity to the target mRNA will determine the slicing of the target or the translational inhibition, which leads to target mRNA decay [26]. Gene regulation from miRNA is dependent on the shuttling of miRISC in the cell and the availability of and abundance of miRNAs.

miRNAs are transcribed from DNA sequences, which lead to primary miRNA (pri-miRNA) and then processed into pre-miRNA (precursor miRNA) and finally mature miRNAs.The transcription of pri-miRNA occurs within the nucleus by RNA polymerase II or III [21] and it is then processed into pre-miRNA. This processing occurs from the proteins

DiGeorge Syndrome Critical Region 8 (DGCR8) and a ribonuclease III, Drosha. DGCR8 recognizes specific motifs and an N6-methyladenylated GGAC within pri-miRNA while Drosha cleaves the pri-miRNA duplex resulting in a 2 nucleotide 3' overhang on the pre-miRNA hairpin structure [21]. Then, pre-miRNAs are transported to the cytoplasm by exportin 5 (XPO5)/RanGTP complex and processed by Dicer, a RNase III endonuclease. Dicer removes the terminal loop of the pre-miRNA structure and results in a mature miRNA duplex. With pre-miRNA-1229-3p, the directionality of the miRNA strand is at the 3' end [26].

1.2 MIR1229 targets and downregulates SORL1

MIR1229 is potentially associated with Alzheimer's Disease (AD). One product of *MIR1229* is processed into the mature hsa-miR-1229-3p, which downregulates sortilin-related receptor 1 (SORL1). Among all the identified variants on MIR1229, rs2291418 is the one that has the highest allele frequency [19]. Rs2291418 is located at chr5:179,798,324 (hg38) and changes from reference G to variant A. It is located in the intronic region of *MGAT4B* and exonic region of *MIR1229* (**Figure 1**). Rs2291418 may cause increased stability of the pre-miRNA-1229 hairpin structure which increases mature miRNA-1229-3p production[16]. Ghanbari, M. et al. 2016 showed HEK293 cells with the SNP increased hsa-miR-1229-3p by 70% (pval = 0.002). This increase in miRNA-1229-3p may change the equilibrium between the GQ structure and the extended hairpin structure through the GQ destabilization (**Figure 2**)[16].

In terms of population frequencies, Amish have the highest allele frequency, and East Asians have the lowest (**Table 1**). Age group 50-55 shows the highest level of variant carriers, and there is a significant level of variant carriers in people with age <30 [18].The global allele frequency, from samples from the ALFA project, shows that this SNP occurs 0.03514 [31]. The global allele frequency data can be broken down into more subsets based on region. There are a few subsets that have an allele frequency that is a bit higher than the global allele frequency. Biosample SAMN10492696, which is the geographical category "African others", has an allele frequency 0.057 for this SNP. "African others" is described to be a group with African ancestry. African and African American show an allele frequency of 0.0431 and 0.0426, respectively, which are above the global allele frequency. The groups Asian, East Asian, and Other Asian have an allele frequency of 0.00 for this SNP; these groups do not include South Asians. From dbSNP, the SNP is more common for people with ancestry from Africa and least common in Asians, not including South Asians [31].

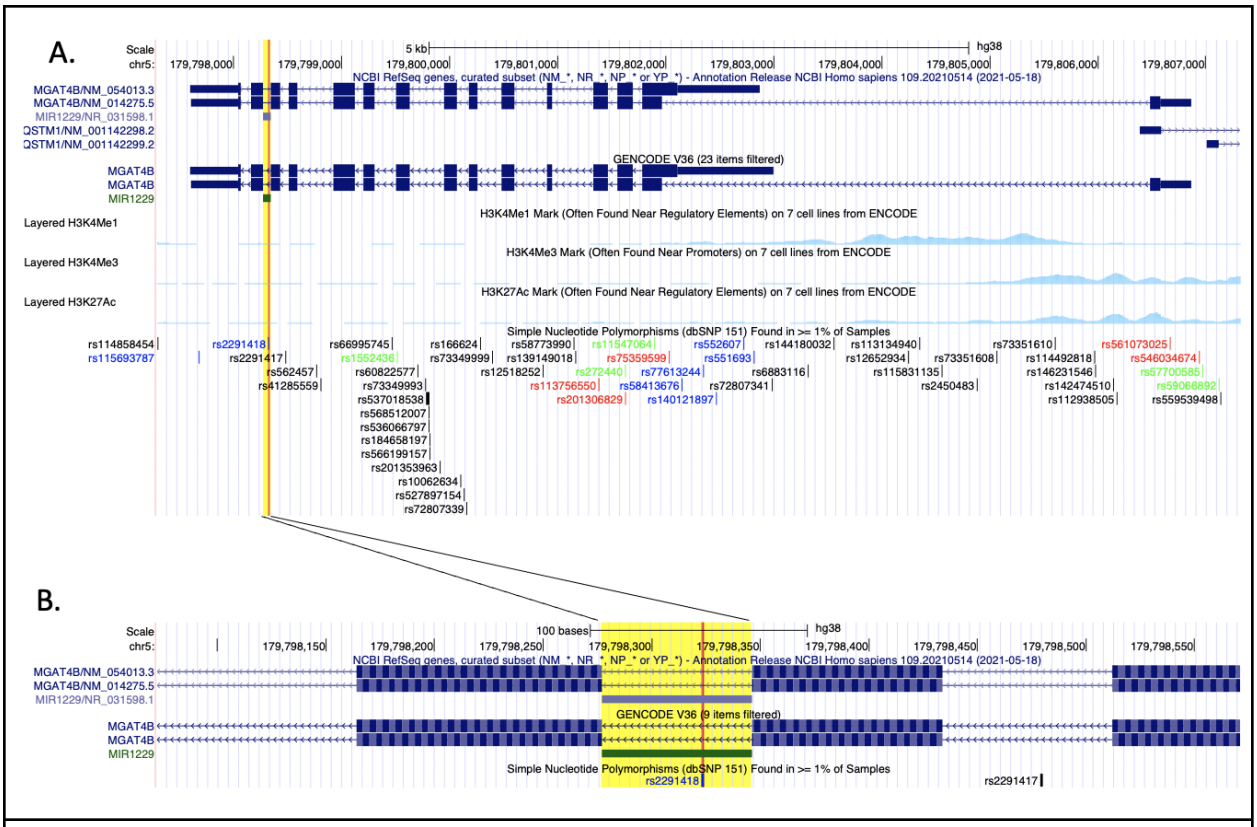

Figure 1. rs2291418 location in *MGAT4B* and *MIR1229*

The UCSC Genome Browser was used to visualize rs2291418 (highlighted in red) in *MGAT4B* and *MIR1229* (highlighted in yellow). The tracks NCBI RefSeq genes, GENCODE V36 and dbSNP151 are shown. **(A)** A zoomed out view showing the entire length of *MGAT4B* which is on the - strand and spans from hg38chr5:179,797,610-179,806,866 . Three additional histone modification tracks from ENCODE are shown [14]. **(B)** A zoomed in view of *MIR1229* which is on the - strand and spans from hg38 chr5:179,798,278-179,798,346

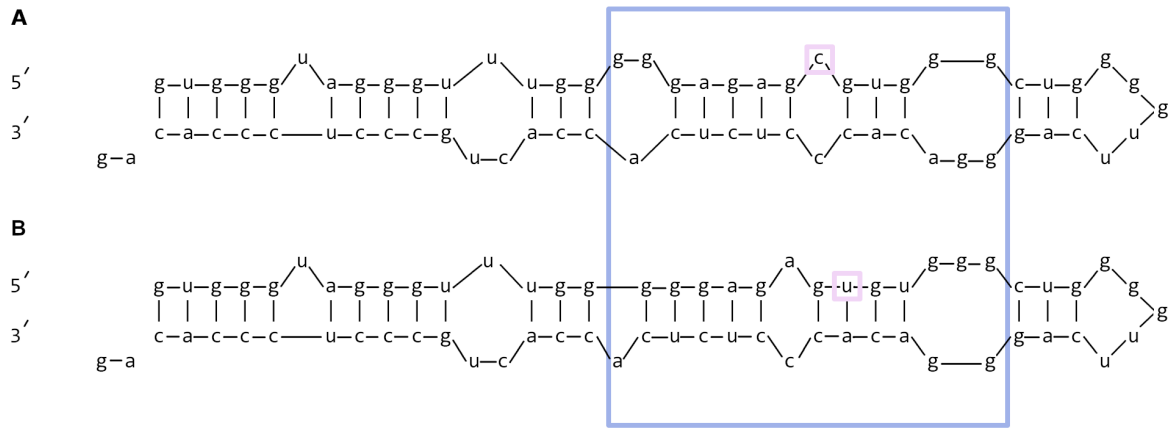

Figure 2. Stem-loop structures of the wild type *MIR1229* and *MIR1229* with rs2291418
The stem-loop structures of the *MIR1229* gene are shown with the wild type structure (**A**) and with the SNP rs2291418 present (**B**) [5, 16]. The location of the SNP is marked with a purple box and the region of the structure that is different between the two are marked with a blue box. The G → A change in the variant causes a C → U change in the miRNA.

Table 1. Population Frequencies of rs2291418 [18].

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Amish	57	912	2	0.06250
African/African-American	1935	41440	60	0.04669
European (non-Finnish)	2362	67996	52	0.03474
Ashkenazi Jewish	112	3472	2	0.03226
Other	66	2088	3	0.03161
Latino/Admixed American	358	15286	10	0.02342
Middle Eastern	7	316	0	0.02215
South Asian	83	4836	0	0.01716
European (Finnish)	70	10622	0	0.006590
East Asian	11	5178	0	0.002124
XX	2737	77816	77	0.03517
XY	2324	74330	52	0.03127
Total	**5061**	**152146**	**129**	**0.03326**

1.3 SORL1 and its role in Alzheimer's disease

SORL1 is a hybrid receptor and is involved in cargo transport, chaperone-like activity,

signaling, and intracellular sorting. The expression of *SORL1* is widespread throughout the brain but is most prominent in the hippocampus [35]. In 2004, *SORL1* was discovered to be associated with familial AD due its role in amyloid precursor protein (APP) processing and trafficking. In 2007 it was found that the overexpression of *SORL1* reduces the amount of beta amyloid (Aβ), where the lack of expression of *SORL1* will increase the amount of Aβ [8, 35].

SORL1 interacts with cytosolic adaptors for anterograde and retrograde movement of APP between the trans-Golgi network (TGN) and early endosomes, causing the delivery of the APP to endocytic compartments, which leads to its degradation. Without SORL1 binding, APP is free to undergo cleavage to form Aβ in the endosomal pathways, leading to increased Aβ levels (**Figure 3**). Increased Aβ can cause the formation of Aβ-containing plaques, which is a neuropathological hallmark of AD [23, 30]. *SORL1* also functions as a sorting factor and redirects newly made Aβ from the neurons to the lysosomes, where Aβ is degraded; when this pathway is interfered with, this may cause familial form of AD [35].

Increased levels of hsa-miR-1229-3p may explain the association of the rs2291418 variant in AD patients [16]. An increase in hsa-miR-1229-3p would decrease *SORL1* transcription leading to a decrease in trafficking of APP and Aβ to degradation pathways. This decrease could cause an increase in the plaques found in the brain which interfere with neuron-neuron communication at synapses [16]. In this study, we aim to analyze the SNP rs2291418 frequency in control vs. AD samples. We will be using a miRNA-Seq dataset from a study of post-mortem Alzheimer's Disease brain samples (PRJNA670793) submitted by Janssen Research & Development, LLC [25].

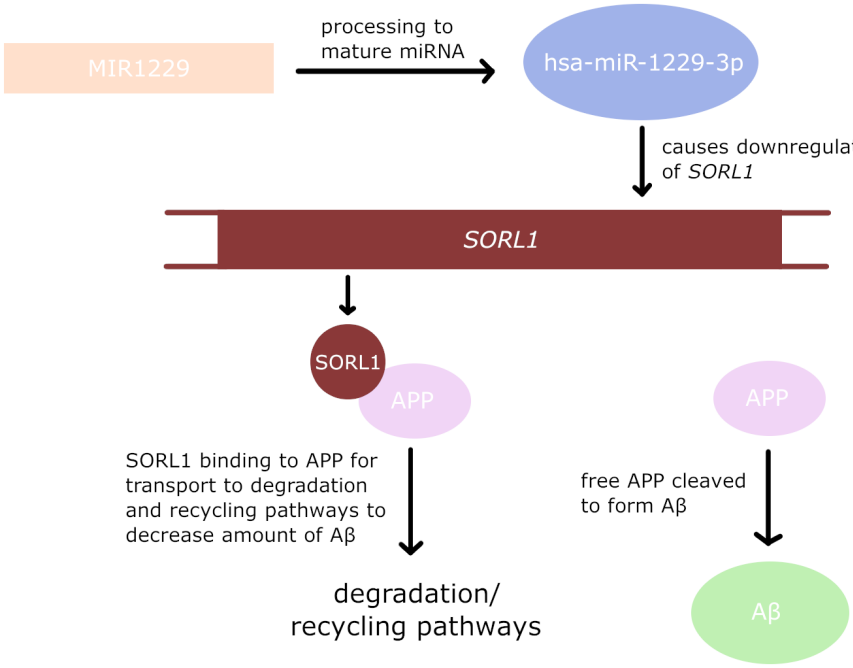

Figure 3. Regulatory pathway
MIR1229 is processed into the mature hsa-miR-1229-3p, which acts upon *SOLR1* to cause its downregulation. One function of SORL1 is binding to APP to transport it to degradation and recycling pathways to decrease the amount of Aβ in brain tissue. APP that is not bound is free to be processed into Aβ. Aβ accumulates and forms plaques, which have been found to be associated in AD [16].

2. Methods

2.1 Non-coding region Investigation

Galaxy and the UCSC Main Table Browser were used to investigate the non-coding regions of *MGAT4B*, *MIR1229*, and *SORL1*. [1, 19] To find regulatory elements in each these regions, data was extracted from the ENCODE cCandidate Cis-Regualtory Elements (cCREs) table for each gene. To find the cCREs in the non-coding regions, we selected elements that had at most 80% overlap. The phastConsElements100way table was used to look for conservation within the genes. No overlap was selected to get conservation that only occurs in the noncoding regions. [7, 10, 19]

2.2 SNP Investigation

Galaxy, BioMart, and IGV were all used to identify SNPs in *MIR1229*. Galaxy and the UCSC Main Table Browser were used to retrieve data from the snp151Common, dbSnp153, and genomadGenomesVariants3_1 tables [1, 19, 28]. BioMart was used to retrieve data from the Human Short Variants (SNPs and index excluding flagged variants) (GRCh38.p13) dataset and the Human Structural Variants (GRCh38.p13) dataset [9]. The first dataset was searched with the gene stable ID filter value ENSG00000221394, and the second dataset was searched with the filters: chromosome = 5, start = 179798346, and end = 179798278. Sequence similarity for *MGAT4B* hg19 and *MGAT4B* hg38 was performed using blastn suite-2 sequences to determine if data from both reference genomes could be used. IGV was then used to visualize the SNPs [28].

2.3 Regulation Investigation

The UCSC Genome Browser was used to visualize regulatory information in *MGAT4B* and *MIR1229*. The following tracks were used: the University of Washington ENCODE group's DNaseI Hypersensitivity track, the nucleosome occupancy track, the H3K4me1, H3K27ac and H3K4me3 tracks from the Bernstein Lab at the Broad Institute, and the transcription factor track containing data from the Myers Lab at the HudsonAlpha Institute for Biotechnology [7, 10, 19]. GeneCards was used to further investigate the transcription factors surrounding *MGAT4B* [31]. H3K27ac, H3K4me3, and H3K4me1 data from ChIP-seq tracks of control dorsolateral prefrontal cortex (DLPFC) were obtained from Gene Expression Omnibus (GEO) and displayed using IGV [28].

2.4 hsa-miR-1229-3p Target Investigation

Targets of hsa-miR-1229-3p and hsa-miR-1229-5p were identified using biomaRt [9]. The code can be found in **Figure S1.** KEGG PATHWAY Database and the GENERIF_SUMMARY in DAVID were used to investigate which of those target genes were involved in Alzheimer's Disease pathogenesis [14, 15, 17].

2.5 NGS Investigation

Galaxy was used to create an miRNA-seq workflow [1]. The miRNA-seq dataset (PRJNA670793) from Li and Cai was obtained from the Sequence Read Archive (SRA) for variant identification [25]. There were 157 samples (IFG, MTG, or STG tissue) taken postmortem from either Alzehimer's Disease patients or control patients. miRNA-seq was done using the Illumina HiSeq4000. 20 samples were used from this dataset to see if they contain variant rss2291418. The samples used are shown in **Table 2**. Fastq files were obtained using the NCBI SRA accession on Galaxy [25]. FastQC and MultiQC were used to assess the quality of the samples [3, 11]. Quality trimming was done with Trimmomatic [6]. A sliding window of 4 bp was used with an average Phred score of 30. Leading and trailing quality scores were set to 30. Reads < 16 bp were discarded. FastQC/MultiQC was run on the trimmed fastq files. Bowtie2 was used to align the trimmed reads to the Human (*Homo sapiens*)(b38): hg38 reference

genome with the default parameters. Bowtie2 is a read aligner that is not splice aware. Normally, a splice aware aligner is necessary for mRNA-seq data [22]. However, splice junctions do not matter with miRNA-seq data. BAM files were then filtered using Samtools view to focus on *MIR1229* and sorted by coordinate using SortSam [24]. Finally, FreeBayes was used to call variants [12]. VCF files were viewed in the UCSC Genome Browser hg38 at *MIR1229* [19].

Table 2. Samples used for SNP analysis.
All samples are from inferior frontal gyrus of the brain. The SRR accession and sex are listed in each group.

Control Samples	AD Samples
SRR12881019 - M	SRR12881018 - M
SRR12881020 - M	SRR12881021 - M
SRR12881027 - F	SRR12881022 - F
SRR12881029 - F	SRR12881026 - F
SRR12881030 - M	SRR12881051 - F
SRR12881031 - M	SRR12881052 - M
SRR12881049 - F	SRR12881054 - M
SRR12881050 - M	SRR12881057 - M
SRR12881055 - F	SRR12881059 - F
SRR12881064 - F	SRR12881065 - F

2.6 Statistical Analysis

A chi-square test was performed to see if there was a significant difference between AD vs. control samples and male vs female samples. The global allele frequency was used to calculate expected values because the ethnicity of the samples are unknown. A confidence level of 0.95 was used.

3. Results

3.1 SNP Investigation

The UCSC Genome Browser dbSNP151 track revealed one SNP in *MIR1229*, rs2291418 (**Figure 1**). There were two clusters of SNPs found in the intronic regions of the 5' and 3' end of the *SORL1* gene. Three SNPs found in the 5' end are associated with Alzheimer's Disease are rs668387, rs689021, and rs641120 and two SNPs found in the 3' end are rs2070045 and rs3824968 [27, 29]. Finally, the genome browser revealed the following *SORL1* SNPs: rs12285364 and rs2282649 in the intronic region, and rs2070045 in the exonic region, which are associated with an increase in risk, and rs1010159, rs641120 and rs668387 which are

reduced-risk SNPs located in the intronic region [34]. A summary of these SNPs are shown in **Table 3**.

The results from the snp51Common table revealed only one SNP in *MIR1229* (rs2291418), while the All dbSNP 153 (dbSnp153) and genomadGenomesVariants3_1 tables displayed many more. The BioMart search on the Human Short Variants dataset returned 127 results, and the search on the Human Structural Variants had 182 results.

Table 3. *SORL1* SNPs Associated with AD
SNPs identified by Reitz et al. and Rogaeva et al.

dbSNP Reference	Region/Type	AD Risk
rs668387	intron 6	decrease
rs641120	intron 6	decrease
rs2070045	exon 25 - synonymous	increase
rs2282649	intron 38	increase
rs1010159	intron 39	decrease
rs12285364	intron 9	increase

3.2 Regulation Investigation

In **Figure 1**, the H3K4me1 track shows that H4K3me1 signals upstream near the TTS of both *MGAT4B* transcripts are stronger compared to the rest of the regions in HUVECs, and the H3K4me3 and H3K27ac tracks show signals at peaks immediately upstream of *MGAT4B*, as well as one large peak in the H3K27ac track from approximately 179,243,400 - 179,243,800 [19]. No significant signal of the histone marks was observed at the location of rs2291418 [19]. Other significant transcription factors include NF-KappaB1 and SP1 in the promoter region [32].

3.3 hsa-miR-1229-3p Target Investigation

Targets of *miR1229* were identified with biomaRt. There were 299 results for hsa-miR-1229-3p, all of which had experimental validation and came from TarBase miRNA Target Prediction, and zero results for hsa-miR-1229-5p. Twenty-one target genes had multiple target regions [9].

DAVID reported 135 (50.4%) genes identified in the brain in the UP_TISSUE category. The target genes associated with Alzheimer's Disease were: CELF1, FUS, TARDBP, ADARB1,

CAST, CSNK1D, CLOCK, PURA, SETX, SIRT1, and SORL1 [14, 15]. Next, KEGG identified the following ed networks involved in Alzheimer's Disease : calcium signaling (nt06410), Ubiquitin-proteasome system (nt06420), Unfolded protein response (UPR) signaling (nt06412), Apoptosis (nt06414), AGE-RAGE signaling (nt06417), Oxidative phosphorylation (nt06418), and Microtubule-based transport (nt06419) [17]. There are 369 genes listed in the KEGG pathway entry, and the intersection of this list and the list of 299 genes found in the brain was taken. The results were RB1CC1, ATF4, IRS4, LRP5, FZD5, CASP3, ATP2A2, and CHUK. **Table 4** contains the 19 genes that are both associated with Alzheimer's Disease and targets of hsa-miR-1229-3p.

Table 4. Targets of hsa-miR-1229-3p that are associated with Alzheimer's Disease.
The genes in this table were identified as genes associated with Alzheimer's Diseases by either DAVID [14, 15] or KEGG [17]. All genes in this table except *SORL1* were identified by TarBase as targets of hsa-miR-1229-3p. Ghanbari et. al. (2016) experimentally validated *SORL1* as a target of hsa-miR-1229-3p [13].

Ensembl ID	Symbol	DAVID or KEGG
ENSG00000149187	CELF1	DAVID
ENSG00000089280	FUS	DAVID
ENSG00000120948	TARDBP	DAVID
ENSG00000197381	ADARB1	DAVID
ENSG00000153113	CAST	DAVID
ENSG00000141551	CSNK1D	DAVID
ENSG00000134852	CLOCK	DAVID
ENSG00000185129	PURA	DAVID
ENSG00000107290	SETX	DAVID
ENSG00000096717	SIRT1	DAVID
ENSG00000137642	SORL1	DAVID
ENSG00000023287	RB1CC1	KEGG
ENSG00000128272	ATF4	KEGG
ENSG00000133124	IRS4	KEGG
ENSG00000162337	LRP5	KEGG
ENSG00000163251	FZD5	KEGG
ENSG00000164305	CASP3	KEGG

ENSG00000174437	ATP2A2	KEGG
ENSG00000213341	CHUK	KEGG

3.4 NGS Investigation

Every sample analyzed with FastQC had a Phred score above 30, and all adapter content was less than 0.1% . MultiQC results are shown in **Figure 4**. The Bowtie results show the alignment rates range from 95.98% - 97.35% in all samples. The SNP, rs2291418, was found in samples SRR12881026 (Female with AD), SRR12881029 (Female control), and SRR12881065 (Female with AD) samples. The full results are found in **Table 5.** The VCF files of the three samples with rs2291418 were viewed in the UCSC Genome Browser (**Figure 5**).

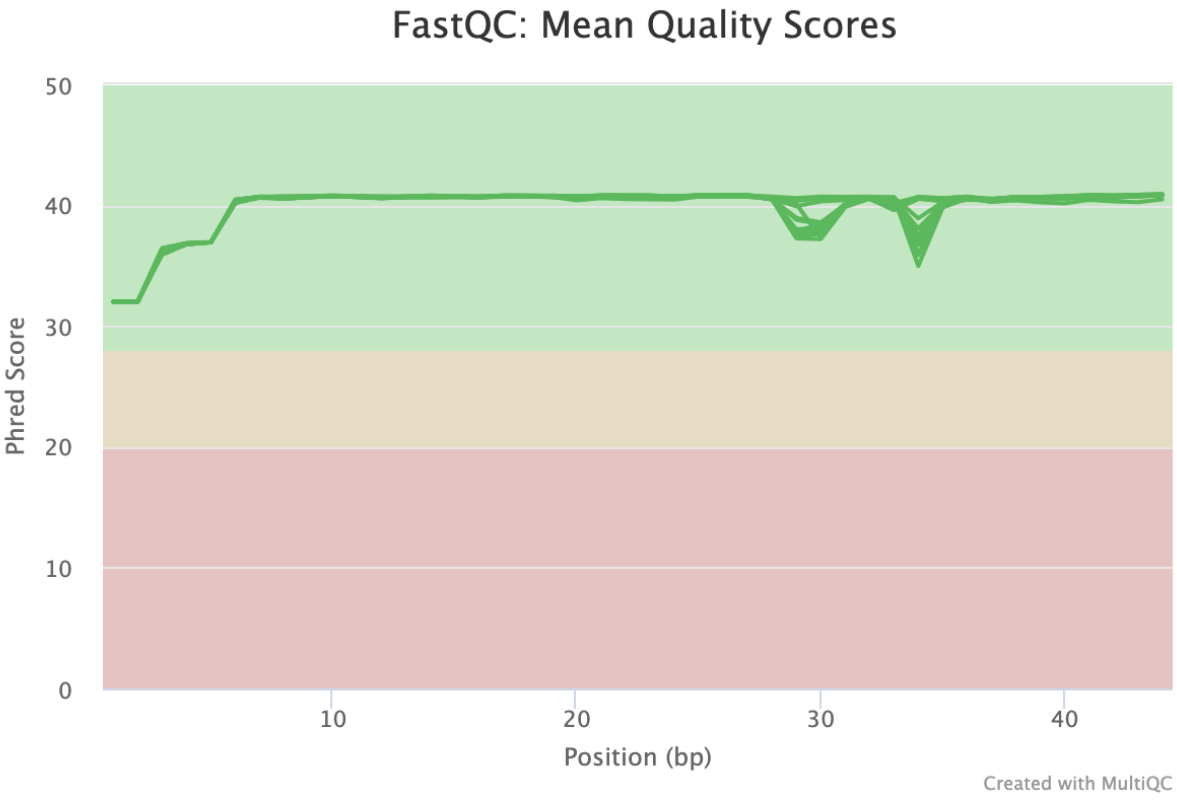

Figure 4. MultiQC Report
Mean quality scores on all 20 quality trimmed FASTQ files. All files have mean Phred scores above 30. A Phred score of 30 indicates there is 99.9% base call accuracy.

Table 5. rs2291418 SNP Analysis

Results from the FreeBayes analysis are shown in the table. If the sample has the SNP, the position, nucleotide change and quality are included. Three SNPs were found; two were control samples and one was an AD sample. All SNPs were in female samples.

Accession	Sex	Group	rs2291418?
SRR12881018	M	AD	no
SRR12881019	M	C	no
SRR12881020	M	C	no
SRR12881021	M	AD	no
SRR12881022	F	AD	no
SRR12881026	F	AD	Yes, 179798324, G>A, 27.15
SRR1281027	F	C	no
SRR12881029	F	C	Yes, 179798324, G>A, 40.35
SRR12881030	M	C	no
SRR12881031	M	C	no
SRR12881049	F	C	no
SRR12881050	M	C	no
SRR12881051	F	AD	no
SRR12881052	M	AD	no
SRR12881054	M	AD	no
SRR12881055	F	C	no
SRR12881057	M	AD	no
SRR12881059	F	AD	no
SRR12881064	F	C	no
SRR12881065	F	AD	Yes, 179798324, G>A, 2.63

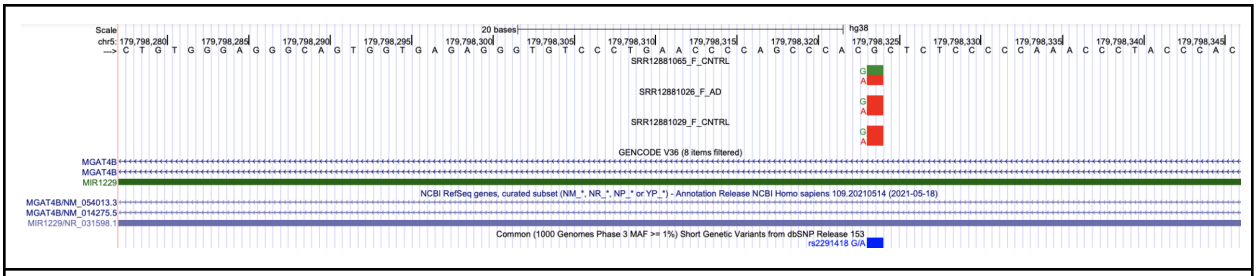

Figure 5. rs2291418 view in UCSC
VCF files of the three samples with rs2291418 in the UCSC Genome Browser (hg38). http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr5%3A179798278%2D179798346&hgsid=1139833439_11PPxm8FXEVUbkZPJX3H0LFDiaVB

3.5 Statistical significance

Chi-square analysis was performed and no statistically significant association of SNP rs2291418 was found between the AD vs control group (pval = 0.0982, $\alpha = 0.05$). However, there was a significant SNP association found between male vs females (pval = 0.0029, $\alpha = 0.05$).

4. Discussion

4.1 SNP Investigation

The phenotype of rs2291418 causes an overproduction of pre-miRNA-1229, which leads to the downregulation of *SORL1* [16]. The downregulation of *SORL1* leads to the accumulation of Aβ peptides, which is a phenotype of interest for AD. Two groups of SNPs were found on the 5' and 3' end of the *SORL1* gene in the intronic regions. These SNP groups found in *SORL1* have strong associations with Alzheimer's disease, leading to the belief that they lead to the downregulation of the *SORL1* gene. This was shown in studies done by Rogaeva, E. et al. (2007) and by Reitz, C. et al. (2011), which showed the two groups of SNPs in data sets sectioned off by race/regions [27, 29].

Wang et al. summarized 35 previous studies and reported that rs12285364, rs2070045, and rs2282649 of *SORL1* are risk factors for AD. In contrast, rs1010159, rs641120, and rs668387 of *SORL1* were associated with a decreased risk of AD. However, their subgroup analysis showed that the SNPs & AD association was affected by ethnicity. The SNPs of rs641120 (G > A), rs668387 (C > T), and rs689021 (G > A) were associated with a reduced risk of AD in Caucasians only [34]. While the AD-risk-associated *SORL1* variants (rs12285364, rs2070045, and rs2282649) alter the physiological role of SORL1 in the processing of APP holoprotein, SNP rs2291418 (*mir1229*) involves in the pathogenesis of AD through changing the equilibrium between the G-quadruplex structure formed by pre-miR-1229 and the extended hairpin structure, shifting the equilibrium towards the extended hairpin structure, thereby

increasing the levels of the mature hsa-miR-1229-3p that regulates the translation of SORL1 [16].

4.2 Regulation Investigation

H3K27ac is associated with active enhancers, H3K4me1 gene enhancers , and H3K4me3 activation of transcription. In **Figure 1 panel A**, it is clear from the histone modification tracks there are active enhancers upstream *MIR1229* / *MGAT4B* (shorter transcript) and downstream the long *MGAT4B* transcript.

4.3 hsa-miR-1229-3p Target Investigation

There were no results for hsa-miR-1229-5p since hsa-miR-1229-3p is strongly favored. Of all the target genes, 21 had multiple target regions. **Table 4** contains the 19 genes that are both targets of hsa-miR-1229-3p and associated with Alzheimer's Disease, as identified by either DAVID or KEGG. Changes in the levels of hsa-miR-1229-3p may have an effect on the regulation of these genes as they can be targeted by the miRNA. Further research investigating these effects, specifically in Alzheimer's, may lead to a better understanding of the pathology of this disease. There is currently no strong evidence of an effect of rs2291418 on these genes, aside from *SORL1*.

4.4 SNP rs2291418 Frequencies Investigation

In the study, ten control samples (5 female, 5 male) and ten Alzheimer's Disease samples (5 female, 5 male) were analyzed to see if they had the rs2291418 SNP. This was done by viewing BAM files in the UCSC Genome Browser, with results displayed in **Table 5**. The SNP was found in three samples, SRR12881026 (Female with AD), SRR12881029 (Female control), and SRR12881065 (Female with AD) . All three were female samples, and the SNP was not found in any of the male samples. According to the Alzheimer's Association, two-thirds of Americans with AD are women and women have a higher risk of developing AD than men [2]. While all three of the samples in this study with the SNP were women, the frequency of the SNP between the two groups from gnomAD are 54.1% women and 45.9% men [18]. Due to the small sample size of twenty samples as preliminary data for the study, there can't be any conclusions drawn between the possible effect of the SNP on the sex difference in AD risk. Further data should be gathered from expansion of the data set to include more samples for both AD and control samples. Another consideration for our results is the differences in the prevalence of the SNP within different ethnic groups, as the Amish population has the highest frequency of the SNP [18]. The study from Li and Cai did not list the ethnicity of the people where the samples were collected from.

While there is still the limitation of the small preliminary data set, the SNP was also present in both a control sample and in an AD sample. In Ghanbari et al, there has been an association found between the SNP in miR-1229-3p, which is the mature miRNA from the gene containing the SNP, and an increased risk of AD through the regulation of *SORL1* [13]. An

expansion of the data set could show an association of the SNP between control samples and AD samples, as well as any correlations between female and male samples, as the original study did not distinguish between the sex of the samples. Discovering differences between the development of AD between different sexes or ethnic groups can lead to better preventative methods or treatments for these specific groups.

References

1. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., & Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research, 44(W1), W3–W10. https://doi.org/10.1093/nar/gkw343
2. Alzheimer's Association. (n.d.). Women and Alzheimer's. Alzheimer's Association. https://www.alz.org/alzheimers-dementia/what-is-alzheimers/women-and-alzheimer-s
3. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Beinke, S., & Ley, S. C. (2004). Functions of NF-kappaB1 and NF-kappaB2 in immune cell biology. The Biochemical journal, 382(Pt 2), 393–409. https://doi.org/10.1042/BJ20040544
5. Berezikov, E., Chung, W. J., Willis, J., Cuppen, E., & Lai, E. C. (2007). Mammalian mirtron genes. Molecular cell, 28(2), 328–336. https://doi.org/10.1016/j.molcel.2007.09.028
6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
7. Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research, 46(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081
8. Dodson, S. E., Gearing, M., Lippa, C. F., Montine, T. J., Levey, A. I., & Lah, J. J. (2006). LR11/SorLA expression is reduced in sporadic Alzheimer disease but not in familial Alzheimer disease. Journal of neuropathology and experimental neurology, 65(9), 866–872. https://doi.org/10.1097/01.jnen.0000228205.19915.20
9. Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols, 4(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97
10. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74. https://doi.org/10.1038/nature11247
11. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England), 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354
12. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012
13. Ghanbari, M., Ikram, M. A., de Looper, H., Hofman, A., Erkeland, S. J., Franco, O. H., & Dehghan, A. (2016). Genome-wide identification of microRNA-related variants associated with risk of Alzheimer's disease. Scientific reports, 6, 28387. https://doi.org/10.1038/srep28387

14. Huang, D., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1–13. https://doi.org/10.1093/nar/gkn923
15. Huang, D., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44–57. https://doi.org/10.1038/nprot.2008.211
16. Imperatore, J. A., Then, M. L., McDougal, K. B., & Mihailescu, M. R. (2020). Characterization of a G-Quadruplex Structure in Pre-miRNA-1229 and in Its Alzheimer's Disease-Associated Variant rs2291418: Implications for miRNA-1229 Maturation. International journal of molecular sciences, 21(3), 767. https://doi.org/10.3390/ijms21030767
17. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27
18. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7
19. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome research, 12(6), 996–1006. https://doi.org/10.1101/gr.229102
20. Kumar, S., Iwata, Y., Zomorrodi, R., Blumberger, D.M., Fischer, C.E., Daskalakis, Z.J., Mulsant, B.H., Pollock, B.G., Graff-Guerrero, A. and Rajji, T.K. (2020), Dorsolateral prefrontal cortex metabolites and their relationship with plasticity in Alzheimer's disease. Alzheimer's Dement., 16: e045879. https://doi.org/10.1002/alz.045879
21. Kwak, Pieter Bas, et al. "The Microrna Pathway and Cancer." Wiley Online Library, John Wiley & Sons, Ltd, 19 July 2010, onlinelibrary.wiley.com/doi/10.1111/j.1349-7006.2010.01683.x.
22. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923
23. Lee, J. H., Barral, S., & Reitz, C. (2008). The neuronal sortilin-related receptor gene SORL1 and late-onset Alzheimer's disease. Current neurology and neuroscience reports, 8(5), 384–391. https://doi.org/10.1007/s11910-008-0060-8
24. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
25. Li, Q. S., & Cai, D. (2021). Integrated miRNA-Seq and mRNA-Seq Study to Identify miRNAs Associated With Alzheimer's Disease Using Post-mortem Brain Tissue Samples. Frontiers in neuroscience, 15, 620899. https://doi.org/10.3389/fnins.2021.620899
26. O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in endocrinology, 9, 402. https://doi.org/10.3389/fendo.2018.00402

27. Reitz, C., Cheng, R., Rogaeva, E., Lee, J. H., Tokuhiro, S., Zou, F., Bettens, K., Sleegers, K., Tan, E. K., Kimura, R., Shibata, N., Arai, H., Kamboh, M. I., Prince, J. A., Maier, W., Riemenschneider, M., Owen, M., Harold, D., Hollingworth, P., Cellini, E., … Genetic and Environmental Risk in Alzheimer Disease 1 Consortium (2011). Meta-analysis of the association between variants in SORL1 and Alzheimer disease. Archives of neurology, 68(1), 99–106. https://doi.org/10.1001/archneurol.2010.346
28. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature biotechnology, 29(1), 24–26. https://doi.org/10.1038/nbt.1754
29. Rogaeva, E., Meng, Y., Lee, J. H., Gu, Y., Kawarai, T., Zou, F., Katayama, T., Baldwin, C. T., Cheng, R., Hasegawa, H., Chen, F., Shibata, N., Lunetta, K. L., Pardossi-Piquard, R., Bohm, C., Wakutani, Y., Cupples, L. A., Cuenco, K. T., Green, R. C., Pinessi, L., … St George-Hyslop, P. (2007). The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nature genetics, 39(2), 168–177. https://doi.org/10.1038/ng1943
30. Selkoe D. J. (2001). Alzheimer's disease: genes, proteins, and therapy. Physiological reviews, 81(2), 741–766. https://doi.org/10.1152/physrev.2001.81.2.741
31. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res., 9, 677–679.
32. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current protocols in bioinformatics, 54, 1.30.1–1.30.33. https://doi.org/10.1002/cpbi.5
33. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic acids research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100
34. Wang, Z., Lei, H., Zheng, M., Li, Y., Cui, Y., & Hao, F. (2016). Meta-analysis of the association between alzheimer disease and variants in GAB2, PICALM, and SORL1. Molecular Neurobiology, 53(9), 6501-6510. doi:http://dx.doi.org/10.1007/s12035-015-9546-y
35. Yin, R. H., Yu, J. T., & Tan, L. (2015). The Role of SORL1 in Alzheimer's Disease. Molecular neurobiology, 51(3), 909–918. https://doi.org/10.1007/s12035-014-8742-5

Supplementary Materials

```
library(biomaRt)
#Find miR targets
regulation <- useMart("ENSEMBL_MART_FUNCGEN",
                      dataset="hsapiens_mirna_target_feature",
                      host='www.ensembl.org')
miR_targets <- getBM(attributes = c("chromosome_name", "chromosome_start",
                                    "chromosome_end", "chromosome_strand",
                                    "display_label","feature_type_description",
                                    "evidence", "gene_stable_id"),
                     filters="external_identifier", values=c("hsa-miR-1229-3p",
                                                             "hsa-miR-1229-5p"),
                     mart = regulation)

#extract the gene_stable_ids to use as a filter
target_ensemble_id <- miR_targets[,'gene_stable_id']

#Find the Entrez IDs and HGNC Symbols for each target
ensembl <- useMart("ENSEMBL_MART_ENSEMBL",dataset="hsapiens_gene_ensembl",host=
                   'www.ensembl.org')
targets <- getBM(attributes = c("entrezgene_id", "hgnc_symbol",
                                "ensembl_gene_id"),
      filters = "ensembl_gene_id", values = target_ensemble_id, mart=ensembl)

#merge the results together
target_names <- merge(miR_targets, targets, by.x="gene_stable_id",
                      by.y="ensembl_gene_id")

library(xlsx)
write.xlsx(target_names, file="miR1229_targets.xlsx", row.names = F)
```

Figure S1. Extracting *MIR1229* targets from biomaRt.
R Studio and the biomaRt library were used to find targets of hsa-miR-1229-3p and hsa-miR-1229-5p using the ENSEMBL_MART_FUNCGEN mart. The attributes and filters are shown above. The results did not include the target's gene name so ENSEMBL_MART_ENSEMBL was used to find them using the previous results as filter values. [9].