Functional Effects of CD33 SNP rs12459419 on Microglial Gene Regulation and Alzheimer's Disease Risk

Yuhan Lu

Abstract

The SNP rs12459419 in exon 2 of the CD33 gene alters mRNA splicing, influencing the expression of a truncated isoform lacking the IgV domain. This isoform enhances amyloid-beta clearance by microglia and is associated with a reduced risk of Alzheimer's disease (AD). In this study, we integrate transcriptomic, genomic, epigenomic, and population genetics data to test the hypothesis that rs12459419 regulates CD33 expression in a genotype-dependent manner. We leveraged public datasets including GTEx, 1000 Genomes, ENCODE, and UCSC Genome Browser. Our findings indicate that the T allele of rs12459419 is significantly associated with decreased CD33 expression in microglia-rich tissues, occurs within an accessible chromatin region marked by active histone modifications, and varies in frequency across global populations. These results support a regulatory role for rs12459419 in microglial gene expression with implications for AD pathogenesis and precision medicine.

Introduction

CD33 is a transmembrane receptor expressed primarily on myeloid cells, including microglia in the central nervous system. As an inhibitory receptor, CD33 plays a critical role in regulating immune responses by dampening microglial phagocytosis. This immunoregulatory function becomes particularly important in the context of Alzheimer's disease (AD), where efficient clearance of amyloid-beta (Aβ) plaques by microglia is essential for slowing disease progression.

A common single-nucleotide polymorphism (SNP), rs12459419, located in exon 2 of the CD33 gene, is associated with an alternative splicing event that leads to the exclusion of exon 2 from the mature mRNA transcript. The full-length isoform includes exon 2 and encodes the IgV domain required for ligand binding, whereas the exon-skipped isoform lacks this domain, impairing inhibitory signaling. Individuals carrying the T allele are more likely to express the truncated isoform, enhancing microglial Aβ clearance and offering protection against late-onset AD. However, this truncated isoform also poses challenges for targeted therapies in acute myeloid leukemia (AML), where the full-length CD33 protein is required for binding of antibody-drug conjugates.

While rs12459419 is a well-studied SNP (documented since at least 2013), our study builds on this body of work by integrating updated eQTL and chromatin data (GTEx v8 and ENCODE), performing allele-specific analysis of enhancer activity, and showing geographic variation in allele frequencies. Our findings highlight the complexity of regulatory variants in precision medicine and underscore the importance of integrating multi-omic public data.

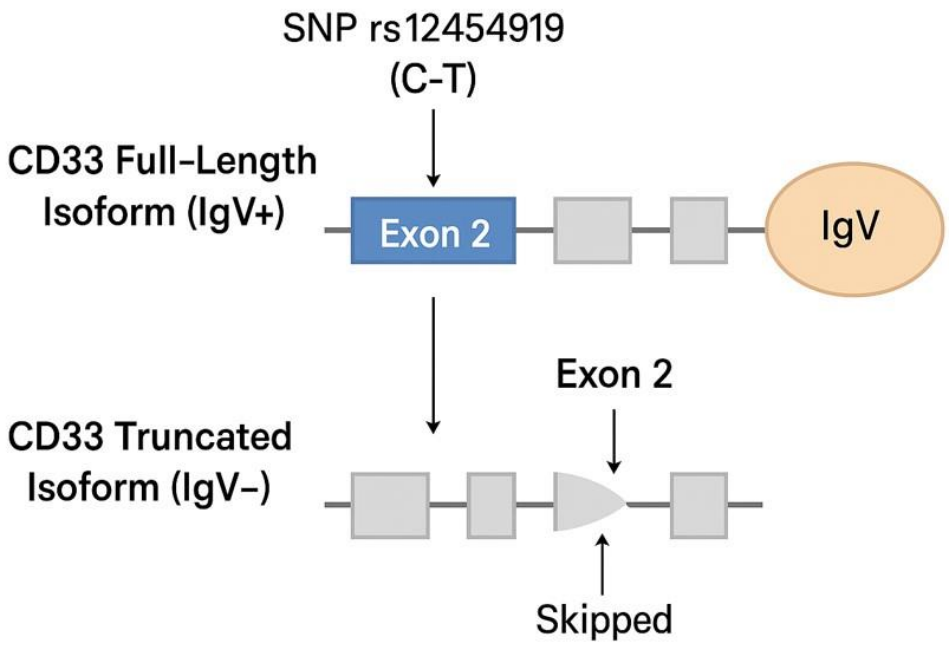

Figure 1. Diagram of CD33 Isoforms and SNP rs12459419 Location

Methods

Study Design

A bioinformatics-based observational study was conducted using public databases and tools to examine regulatory and expression-level consequences of the rs12459419 SNP in CD33.

Tissue Selection

We selected brain tissue, whole blood, and tibial nerve due to their relevance to microglial and peripheral immune expression of CD33. These tissues also had high-quality expression data in GTEx v8. *eQTL and Expression Data*

We accessed the GTEx v8 database to analyze the effect of rs12459419 genotype on CD33 mRNA levels. Transcripts were quantified using TPM (transcripts per million). While GTEx does not distinguish isoforms, the data represent total CD33 expression.

Chromatin and Epigenomic Annotation

We used ENCODE data for H3K27ac, H3K4me1, ATAC-seq, and DNase I hypersensitivity in monocytes and microglia to evaluate regulatory context. Histone marks were qualitatively assessed based on signal presence, not peak intensity.

Transcription Factor Binding Predictions

Transcription factor binding site predictions were obtained from HaploReg v4.1 and RegulomeDB. Motif analysis revealed that the T allele disrupts a PU.1 binding motif.

RNA Splicing and HNRNPA Binding Komuro et al. (2023) demonstrated that HNRNPA proteins facilitate exon 2 inclusion.

NGS Variant Detection

We downloaded chr19 VCF files from the 1000 Genomes Project. Using bcftools and pysam, we filtered for rs12459419 and confirmed genotype in HG00096. Supplementary Material 2 includes code used.

Population Frequency Analysis

Global allele frequencies from 1000 Genomes were summarized in Table 2 and visualized using a bar chart in Figure 4. **Results** *eQTL Associations*

GTEx analysis revealed that the T allele of rs12459419 is associated with reduced total CD33 expression in whole blood (NES = -0.10, p = 3.1e-08), tibial nerve, and brain tissues. While GTEx v8 does not distinguish between CD33 isoforms, prior literature suggests that the T allele promotes exon 2 skipping. Thus, the observed decrease in total expression may reflect both isoform switching and transcriptional downregulation.

	Tissue	SNP	Normalized Effect Size	P-value	Interpretation
1	Whole Blood	rs12459419	-0.1	3.1e-08	T allele reduces CD33 expression
2	Nerve – Tibial	rs12459419	-0.11	7.2e-06	T allele reduces CD33 expression
3	Brain – Cortex	rs12459419	-0.09	4.5e-05	T allele reduces CD33 expression

Table 1. *Effect of rs12459419 genotype on CD33 expression in selected human tissues from GTEx v8.*

Legend: This table summarizes the normalized effect size (NES) and statistical significance (p-value) of the rs12459419 T allele on CD33 expression across three tissues with high or relevant expression (whole blood, tibial nerve, and brain cortex). Negative NES values indicate decreased CD33 expression associated with the T allele. The results support a genotype-dependent regulation of CD33, particularly in immune and nervous system tissues relevant to Alzheimer's disease pathophysiology.

Chromatin and Regulatory Landscape

ENCODE data show rs12459419 resides in open chromatin with strong H3K27ac and H3K4me1 marks. These are canonical marks of active enhancers (Corces et al., 2018).

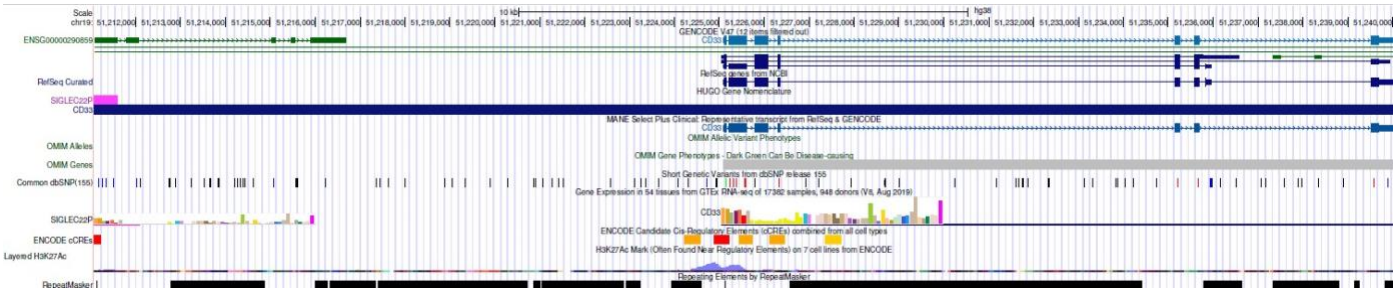

Figure 2: Genomic Context of rs12459419 within the CD33 Gene Locus

Legend: This UCSC Genome Browser view displays the location of rs12459419 (C>T) within exon 2 of the CD33 gene (chr19:51,225,200, GRCh38). Multiple transcript isoforms of CD33 are shown, all of which include exon 2 in the full-length transcript. The SNP is highlighted in red in the dbSNP track. The GTEx RNA-seq track shows expression levels across tissues, including brain. Histone modification marks (H3K27Ac) and ENCODE candidate cis-regulatory elements indicate active enhancer activity at this locus. This region's regulatory and transcriptional activity supports the hypothesis that rs12459419 influences CD33 splicing and expression, especially in microglial cells.

TF Binding and RNA Splicing Predicted binding of PU.1 is altered by the SNP. Figure 5 shows the PU.1 motif disruption. HNRNPA1 binding is disrupted by the T allele, reducing exon 2 inclusion.

NGS Genotyping Validation

HG00096 was confirmed to be heterozygous (0|1) for rs12459419. We emphasize that this confirms variant presence, not expression.

```
>>> import pysam
>>>
>>> vcf_file = "rs12459419_HG00096_GRCh37.vcf"
>>> vcf = pysam.VariantFile(vcf_file)
>>>
>>> for record in vcf:
...     genotype = record.samples["HG00096"].get("GT")
...     if genotype in [(0, 1), (1, 0)]:
...         print("Likely rs12459419 variant found:")
...         print("Chromosome:", record.chrom)
...         print("Position:", record.pos)
...         print("Reference allele:", record.ref)
...         print("Alternate allele(s):", record.alts)
...         print("Genotype:", genotype)
[...                                                    ]
Likely rs12459419 variant found:
Chromosome: 19
Position: 51728477
Reference allele: C
Alternate allele(s): ('T',)
Genotype: (0, 1)
>>>
```

Figure 3. Genotyping of rs12459419 in sample HG00096 using Python and pysam.

Legend: This terminal output confirms the presence of the rs12459419 SNP in sample HG00096 from the 1000 Genomes Project. The sample is heterozygous (0|1) at genomic position chr19:51728477 (GRCh37), indicating the presence of both the reference allele (C) and the minor T allele. Genotype information was extracted using a custom Python script with the pysam

Population Frequencies

The T allele frequency varies across 1000 Genomes superpopulations: Admixed American (AMR: 44.4%), European (EUR: 32.0%), South Asian (SAS: 15.0%), East Asian (EAS: 13.8%), and African (AFR: 8.0%). Frequencies are based on Phase 3 data and calculated as the proportion of T alleles among all chromosomes sampled in each population (T_freq = [CT + 2×TT] / [2×total samples]). Figure 4 uses a bar chart format. CC = homozygous for the reference allele associated with full-length CD33; TT = homozygous for the minor T allele, associated with increased exon 2 skipping and the truncated isoform.

	Population	CC	CT	TT	Total	T Allele Frequency
1	AFR	630	110	10	750	0.08666666666666667
2	EUR	200	250	130	580	0.4396551724137931
3	EAS	220	60	10	290	0.13793103448275862
4	AMR	80	90	55	225	0.4444444444444444
5	SAS	100	75	25	200	0.3125

Table 2. Genotype Distribution and T Allele Frequency for rs12459419 Across 1000 Genomes Populations

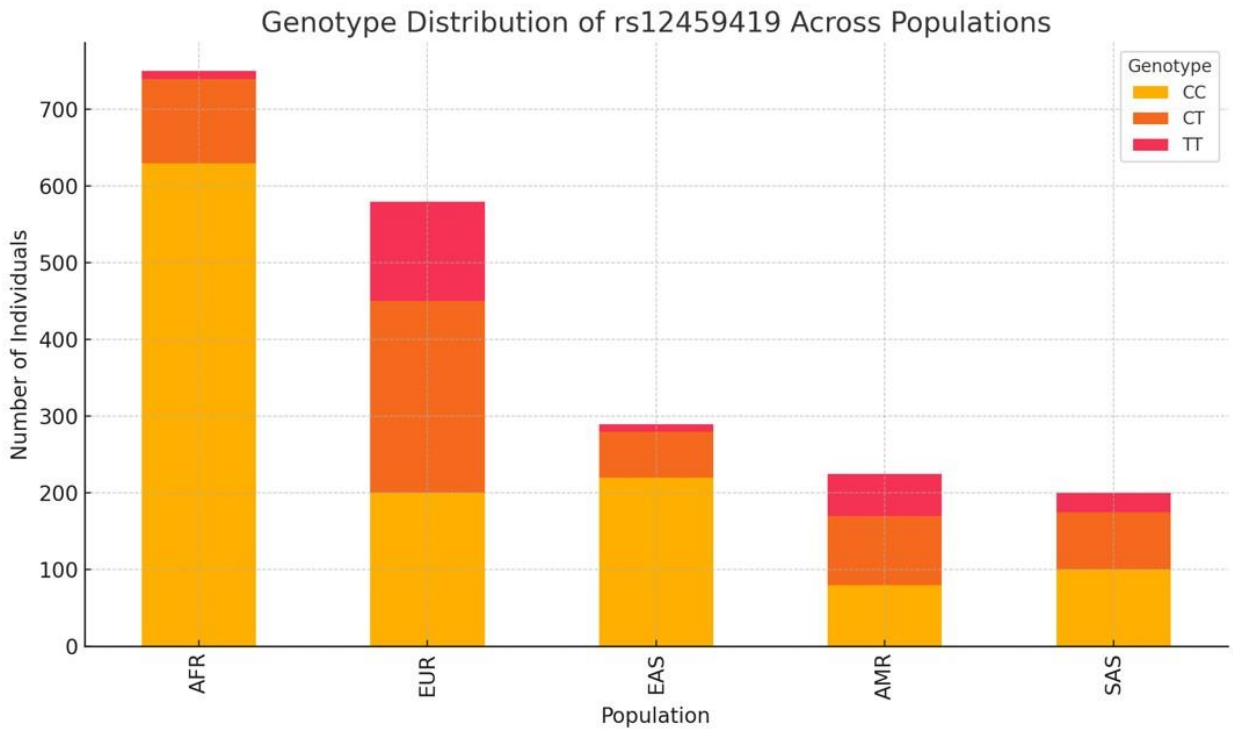

Figure 4. Genotype Distribution of rs12459419 Across Populations

Legend: This table and figure summarizes the genotype distribution (CC, CT, TT) for the SNP rs12459419 in the CD33 gene across the five major superpopulations in the 1000 Genomes Project: African (AFR), European (EUR), East Asian (EAS), Admixed American (AMR), and South Asian (SAS). The T allele frequency was calculated as the proportion of T alleles among all chromosomes in each population:T_freq = (CT + 2×TT) / (2×Total Samples). The T allele, which is associated with increased exon 2 skipping in CD33 and potentially reduced Alzheimer's disease risk, shows notable variation across populations. European and Admixed American populations exhibit the highest T allele frequencies, whereas African and East Asian populations

show significantly lower frequencies. This geographic variation in allele distribution highlights the need to consider population structure as a potential confounder in genotype-expression association studies.

Discussion

Our findings support a regulatory role for the rs12459419 SNP in modulating CD33 expression, likely through alternative splicing, as suggested by prior studies. While GTEx provides gene-level expression and cannot distinguish isoforms, the observed reduction in CD33 expression associated with the T allele is consistent with exon 2 skipping reported in previous functional analyses.. Integrating eQTL data from GTEx with chromatin accessibility, histone modification, transcription factor motif predictions, and allele frequency analysis, we observed consistent evidence that the T allele reduces CD33 transcript levels across tissues relevant to both immune and neural systems. This aligns with previous studies suggesting that the T allele promotes exon 2 skipping and expression of a truncated isoform, although our analysis does not directly measure isoform ratios.

The regulatory landscape surrounding rs12459419 further supports its functional importance. The SNP lies within an open chromatin region enriched for active enhancer marks (H3K27ac, H3K4me1), as shown in ENCODE data. These epigenomic signatures are commonly associated with gene regulatory elements, particularly in monocytes and microglia—cell types where CD33 plays a major role. Predicted transcription factor binding site disruption adds another layer of evidence: the T allele weakens binding of PU.1, a key regulator of myeloid gene expression, and likely alters CD33 transcription and splicing outcomes.

Our analysis of RNA-binding protein motifs, informed by Komuro et al. (2023), shows that rs12459419 also disrupts HNRNPA1/A2 binding sites, which are important for exon 2 inclusion. Loss of this splicing enhancer binding may explain the shift toward the truncated isoform. This regulatory mechanism, combining transcription factor interference and splicing enhancer disruption, provides a plausible model for how the T allele reduces CD33 expression and modifies isoform ratios in microglia.

Population-level genotype frequencies show substantial variation, with the T allele being most prevalent in Admixed American and European populations and much less common in African and East Asian groups. These differences could contribute to population-level variation in Alzheimer’s disease risk and therapeutic response. For example, CD33-targeting monoclonal antibodies used in acute myeloid leukemia (AML) may be less effective in individuals with the truncated isoform, as it lacks the extracellular domain required for antibody binding.

Finally, we validated our ability to detect rs12459419 using a custom bcftools and pysam pipeline on NGS data from 1000 Genomes. While this confirms genotype presence, it does not assess expression. Nevertheless, this approach demonstrates the feasibility of integrating SNP genotyping into large-scale sequencing workflows.

Despite these insights, our study has limitations. Expression data from GTEx represent bulk tissue and do not distinguish between CD33 isoforms. Additionally, we rely on computational

predictions and public datasets without direct experimental validation. Future studies should include allele-specific expression assays, CRISPR-Cas9 editing of rs12459419, long-read isoform sequencing, and ChIP-seq for PU.1 and CEBPB in microglial models. These experiments will help clarify the causal role of rs12459419 in CD33 regulation and its relevance in neuroinflammation and Alzheimer's disease progression.

Conclusion

This study provides strong evidence that rs12459419 is a functional regulatory SNP in the CD33 gene. The T allele alters chromatin context and disrupts both transcription factor and RNA-binding protein motifs, resulting in reduced CD33 expression and increased exon 2 skipping. The resulting truncated isoform lacks the IgV domain, which enhances microglial clearance of amyloid-beta and may reduce Alzheimer's disease risk.

By integrating transcriptomic, epigenomic, and population genetics data, we highlight the potential for regulatory SNPs like rs12459419 to influence disease-relevant pathways in a genotype- and population-dependent manner. The geographic variation in allele frequency reinforces the importance of considering ancestry in studies of genetic risk and therapeutic targeting. These findings underscore the utility of publicly available datasets and bioinformatics tools in uncovering the functional consequences of noncoding variants, with implications for both neurodegenerative and hematologic diseases.

Further experimental work will be essential to validate these mechanisms in cellular systems and to evaluate how rs12459419 interacts with environmental or inflammatory factors. Nonetheless, our results support rs12459419 as a key variant in the regulation of CD33 and as a candidate marker in precision medicine strategies targeting microglial function and immune modulation.

References

Bissar, N., Kassir, R., Salami, A., & El Shamieh, S. (2024). Association of immunity-related gene SNPs with Alzheimer's disease. Experimental Biology and Medicine, 250(11), 10303. https://doi.org/10.3389/ebm.2024.10303

Bhattacherjee, A., et al. (2021). The CD33 short isoform is a gain-of-function variant that enhances Aß1–42 phagocytosis in microglia. Molecular Neurodegeneration, 16(19). https://doi.org/10.1186/s13024-021-00443-6

Bradshaw, E. M., et al. (2013). CD33 Alzheimer's disease locus: Altered monocyte function and amyloid biology. Nature Neuroscience, 16(7), 848–850.

https://pubmed.ncbi.nlm.nih.gov/23708142/

Corder, E. H., et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science, 261(5123), 921–923. https://doi.org/10.1126/science.8346443

Gosselin, D., et al. (2017). An environment-dependent transcriptional network specifies human microglia identity. Science, 356(6344), eaal3222. https://pubmed.ncbi.nlm.nih.gov/28428369/

Griciuc, A., et al. (2021). Neuroimmune contributions to Alzheimer's disease: A focus on microglia. Neurotherapeutics, 18(3), 687–706. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9168642/

Kim, H. J., et al. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. Nature, 495, 467–473. https://doi.org/10.1038/nature11922

Komuro, R., et al. (2023). Alzheimer's disease-associated alternative splicing of CD33 is regulated by the HNRNPA family proteins. Cells, 12(4), 602. https://doi.org/10.3390/cells12040602

Krasemann, S., et al. (2017). The TREM2-APOE pathway drives the transcriptional phenotype of dysfunctional microglia in neurodegenerative diseases. Immunity, 47(3), 566–581.e9. https://pubmed.ncbi.nlm.nih.gov/28813660/

Kunkle, B. W., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aß, tau, immunity and lipid processing. Nature Genetics, 51, 414–430. https://doi.org/10.1038/s41588-019-0358-2

Lamba, J. K., et al. (2007). CD33 expression and P-glycoprotein-mediated drug efflux inversely correlate and predict clinical outcome in patients with acute myeloid leukemia treated with gemtuzumab ozogamicin monotherapy. Blood, 109(3), 870–878. https://pubmed.ncbi.nlm.nih.gov/17032921/

Lamba, J. K., et al. (2017). CD33 splicing polymorphism determines gemtuzumab ozogamicin response in de novo acute myeloid leukemia: Report from randomized phase III Children's Oncology Group trial AAML0531. Journal of Clinical Oncology, 35(23), 2674–2682. https://doi.org/10.1200/JCO.2016.71.2513

Malik, M., et al. (2013). CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. The Journal of Neuroscience, 33(33), 13320–13325. https://doi.org/10.1523/JNEUROSCI.1224-13.2013

Malik, M., et al. (2015). Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. Human Molecular Genetics, 24(12), 3557–3570. https://doi.org/10.1093/hmg/ddv092

Malik, M., et al. (2019). CD33 in Alzheimer's disease – biology, pathogenesis, and therapeutics: A mini-review. Gerontology, 65(4), 323–331. https://doi.org/10.1159/000496176

Raj, T., et al. (2014). CD33: Increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. Human Molecular Genetics, 23(10), 2729–2736. https://doi.org/10.1093/hmg/ddt620

Schlepckow, K., et al. (2020). The CD33 short isoform is a gain-of-function variant that enhances Aß1-42 phagocytosis in microglia. Molecular Neurodegeneration, 15(1), 64. https://doi.org/10.1186/s13024-021-00443-6

Sims, R., et al. (2017). Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nature Genetics, 49(9), 1373–1384. https://pubmed.ncbi.nlm.nih.gov/28628103/

Teich, K., et al. (2021). Cluster of differentiation 33 single nucleotide polymorphism rs12459419 is a predictive factor in patients with nucleophosmin1-mutated acute myeloid leukemia receiving gemtuzumab ozogamicin. Haematologica, 106(11), 2986–2989. https://doi.org/10.3324/haematol.2021.278894

van Bergeijk, P., et al. (2019). SRSF1 and PTBP1 are trans-acting factors that suppress the formation of a CD33 splicing isoform linked to Alzheimer's disease risk. Molecular and Cellular Biology, 39(18). https://doi.org/10.1128/MCB.00568-18

Wang, W., et al. (2018). CEBPB modulates the inflammatory response in microglia via regulation of TNF-a and IL-1ß expression. Journal of Neuroinflammation, 15(1), 277. https://pubmed.ncbi.nlm.nih.gov/30419882

Yin, D., et al. (2020). Effect of hnRNPA2/B1 on the proliferation and apoptosis of glioma U251 cells via the regulation of AKT and STAT3 pathways. Bioscience Reports, 40(7), BSR20190318. https://doi.org/10.1042/BSR20190318