Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Identification of Novel Human-Specific Long Non-Coding RNAs in Spinal Cord Astrocytoma Through RNA-Seq Analysis

**Vaibhavi Jain**
Institute of Biosciences and Technology , MGM University
**Archana Panche**
archana.harke@gmail.com

Institute of Biosciences and Technology , MGM University
**Abhishek Nagargoje**
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology

**Research Article**

**Keywords:** Human spinal cord aging, long non-coding RNA (lncRNA), differential expression, co-expression analysis, functional enrichment, evolutionary conservation, SNP, etc

**Posted Date:** November 6th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11631304/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

**Identification of Novel Human-Specific Long Non-Coding RNAs in Spinal Cord Astrocytoma Through RNA-Seq Analysis**


## Abstract

Human spinal cord aging is associated with molecular changes that may affect neuronal function and tissue homeostasis, but the role of long non-coding RNAs (lncRNAs) in this process remains poorly understood. In this study, RNA-sequencing data from young and old human spinal cord samples were analyzed using an integrated bioinformatics pipeline, including transcript assembly, novel lncRNA identification, differential expression analysis, co-expression analysis, functional enrichment, evolutionary conservation, and SNP analysis. A total of 1,669 novel lncRNAs were identified, among which three transcripts (MSTRG.24062.6, MSTRG.35781.3, and MSTRG.57931.6) were significantly differentially expressed between the young and old age groups. Functional enrichment of co-expressed genes highlighted processes related to protein targeting to membrane, synaptic signaling, neuron projection organization, calcium ion transport, and membrane potential regulation. KEGG pathway analysis identified adrenergic signaling in cardiomyocytes as the most significantly enriched pathway. Conservation analysis revealed that all identified novel lncRNAs were human-specific and unmapped to the mouse genome. Genetic variant analysis identified 11,243 variants across 496 novel lncRNAs, with SNPs accounting for more than 95%. Transitions were more frequent than transversions, and variant distribution varied across chromosomes and lncRNAs. A weak positive correlation between SNP count and lncRNA expression suggested a limited association between genetic variation and expression levels. Overall, this study provides a comprehensive characterization of novel lncRNAs associated with human spinal cord aging and establishes a foundation for future studies investigations into their potential regulatory roles in age related molecular changes in spinal cord aging.



**Keywords:** Human spinal cord aging, long non-coding RNA (lncRNA), differential expression, co-expression analysis, functional enrichment, evolutionary conservation, SNP, etc.


## 1. Introduction

Astrocytomas are among the most common primary tumors of the central nervous system (CNS), arising from astrocytes, the specialized star-shaped glial cells that support neuronal function in the brain and spinal cord. These tumors are of malignancy and are classified by the World Health Organization (WHO) into four grades. Low-grade astrocytomas (grades I and II) generally exhibit

slow-growing behavior, whereas the high-grade forms grade III anaplastic astrocytoma and grade IV glioblastoma multiforme (GBM) are known for their aggressive nature, rapid progression, and poor clinical outcomes (Louis et al., 2007). Among these, GBM stands as the most severe and challenging clinical entity, exhibiting remarkable resistance to conventional therapies such as surgery, chemotherapy, and radiotherapy (Stupp et al., 2005). In fact, despite advances in multimodal management strategies, the median survival for GBM patients remains approximately 15 months, underscoring the urgent need for novel molecular targets and biomarkers that can improve diagnostic precision, prognostic accuracy, and therapeutic intervention (Ostrom et al., 2015; Thakkar et al., 2014).

Although most studies focus on intracranial astrocytomas, spinal cord astrocytomas represent a rarer yet clinically important subgroup. These tumors can present as either low-grade or high-grade lesions, and patient outcomes are largely influenced by tumor grade and anatomical location (Konar et al., 2019). High-grade spinal cord astrocytomas often cause severe neurological deficits and are difficult to treat surgically due to their placement near vital neural structures. This leads to notable challenges in patient management and quality of life. The global incidence is about 0.1– 0.5 cases per 100,000 annually, with pediatric cases representing about 50% of all pediatric spinal tumors and adults about 30% of primary spinal cord neoplasms (Sundaram et al., 2019). This underscores the need for more molecular and clinical research.

Messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs) play important roles in regulating gene expression and cellular function. While mRNAs encode proteins required for normal cellular processes, ncRNAs regulate gene expression at transcriptional, post-transcriptional, and epigenetic levels. Dysregulation of both coding and non-coding RNAs has been associated with tumor initiation, progression, invasion, and resistance to therapy in several human cancers (Statello et al., 2021; Yao et al., 2019). Among the different classes of ncRNAs, long non-coding RNAs (lncRNAs), which are transcripts longer than 200 nucleotides with little or no protein-coding potential, have emerged as important regulators of diverse biological processes. They regulate gene expression through interactions with DNA, RNA, and proteins, thereby influencing chromatin organization, transcription, RNA processing, and signal transduction (Statello et al., 2021; Quinn & Chang., 2016).

Increasing evidence suggests that abnormal lncRNA expression contributes to multiple hallmarks of cancer, including uncontrolled cell proliferation, apoptosis, invasion, metastasis, angiogenesis, and therapeutic resistance (Schmitt & Chang, 2016; Peng et al., 2017). Several well-characterized lncRNAs, including HOTAIR, MALAT1, H19, and NEAT1, have been reported to function as oncogenic regulators in gliomas by modulating cell proliferation, migration, invasion, and epithelial-mesenchymal transition. Their altered expression has also been associated with poor patient

prognosis, indicating their potential as diagnostic and prognostic biomarkers as well as therapeutic targets (Peng et al., 2017; Statello et al., 2021). Despite these advances, the expression profiles and biological functions of novel lncRNAs in spinal cord astrocytoma remain largely unexplored.

Recent advances in high-throughput RNA sequencing (RNA-seq) and bioinformatics have enabled the identification of previously unannotated lncRNAs and facilitated the investigation of their biological functions through transcript assembly, differential expression analysis, co-expression networks, functional enrichment, evolutionary conservation, and genetic variation analyses. These approaches have greatly improved our understanding of lncRNA-mediated regulatory mechanisms in both normal physiology and disease (Yao et al., 2019; Statello et al., 2021). Furthermore, sequence variation within lncRNAs, particularly single nucleotide polymorphisms (SNPs), may alter RNA structure, molecular interactions, and regulatory activity, thereby contributing to disease susceptibility and functional diversity (Carrieri et al., 2012; Gong et al., 2015).

Despite growing evidence that long non-coding RNAs (lncRNAs) play important roles in glioma biology, their role in spinal cord astrocytoma is still not well understood. Most previous studies have focused on brain gliomas, while very few have investigated novel lncRNAs in spinal cord tumors. Therefore, this study used RNA-sequencing data and an integrated bioinformatics approach to identify and characterize novel lncRNAs associated with spinal cord astrocytoma. Differential expression, co-expression, functional enrichment, conservation, and SNP analyses were performed to better understand their potential biological roles. These findings provide a useful foundation for future studies on the molecular mechanisms of spinal cord astrocytoma and the potential use of lncRNAs as diagnostic biomarkers and therapeutic targets.

## 2. Materials and Methodology

### 2.1 Data Collection and Preprocessing

RNA-Seq data from spinal cord astrocytoma samples were downloaded from the NCBI Sequence Read Archive (SRA) using accession number SRP337025. The data contained Illumina single-end sequencing reads with particular age ranges such as 18, 20, 23, 56, 67, 69 y/o downloaded using the SRA Toolkit (fasterq-dump version 3.0.3)(NCBI SRA Toolkit). The raw FASTQ files were first checked for quality and cleaned before further analysis using FastQC (version v0.12.1)(Andrews S., 2010) and cutadapt (version 4.4)(Martin, 2011) and fastp (version 0.23.4)(Chen et al., 2018). The good-quality reads were aligned to the human reference genome (Homo_sapiens.GRCh38.dna.primary_assembly.fa)(Harrison et al., 2024) using HISAT2 (version

2.2.1)(Kim et al., 2019) and processed through Samtools (version 1.19.2)(Danecek et al., 2021), Bamtools (version 2.5.2)(Barnett et al., 2011).

### 2.2 Novel lncRNA Detection

The alignment files (BAM files) were then analyzed with StringTie (version 2.2.1) (Pertea et al., 2015) to assemble and measure transcripts using reference annotation file (gencode.v40.annotation.gtf)(Frankish et al., 2019). This helped identify both known and possible new transcripts, including mRNAs and long non-coding RNAs (lncRNAs). Possible novel lncRNA transcripts were separated and compared with existing reference annotations using gffcompare (version v0.12.10)(Pertea, G., & Pertea, M. 2020) to identify which transcripts were already known and which were unannotated. Transcripts labeled as MSTRG and marked with class code 'u' by StringTie were tested for coding potential using the CPC2 version 0.1 program(Kang et al., 2017). Only non-coding transcripts that were not found in the NONCODEv5 database for BLASTDB (BLAST identity ≥80%, coverage ≥80%) were kept using noncode reference genome file (NONCODEv6_human.fa.gz)(Zhao L., 2021). This step identified new lncRNAs. Their names and IDs were retrieved from the merged GTF file. Based on sequence similarity and query coverage, the transcripts were classified as known or novel for further analysis.

### 2.3 Expression Quantification and Filtering

Expression levels of known mRNAs and novel lncRNAs were quantified with Stringtie and featureCounts (Subread package) using reference annotation (GTF file). Read count matrices were generated raw gene-level expression from aligned RNA-seq BAM files. From obtained count matrices, genes with zero counts across all replicates were removed to reduce noise. The filtered raw count matrix was subsequently used for normalization and differential expression analysis using downstream and many more for further investigation.

### 2.4 Differential Expression Analysis

The DESeq2(1.42.1) (Love M., 2014) and dplyr(1.2.1) (Wickham H., 2023) package in Rstudio(version 4.3.3) (R Core Team, 2024) with Biocmanager(version 3.18) (Morgan M., 2023) was used to find genes from normalised data that showed different expression levels between tumor samples and age groups with Wald Statistics Test through Deseq2(Love M., 2014)**.** Transcripts with an adjusted p-value ≤0.05 and absolute log2 fold change ≥1 were marked as differentially expressed. boxplots, volcano and Heatmaps plots were constructed for further analysis to understand their correlation, distribution and cluster pattern.

### 2.5 Co-expression Correlation Analysis

Pearson correlation analysis was performed between differentially expressed novel lncRNAs and protein-coding genes(mRNAs) to identify significantly co-expressed genes using normalized data of mRNAs and lncRNAs . In between this analysis Python(3.12.7) (Python Software Foundation, 2024) scripts were used with it, where Only strong correlations ($|r| > 0.99$) with statistically significant associations (adjusted p-value < 0.05, corrected using the Benjamini–Hochberg false discovery rate method from wald statistics) were retained to predict the potential biological functions of the novel lncRNAs. The final output file included correlation coefficients, adjusted p-values, log2 fold changes, and transcript length information for both lncRNAs and mRNAs which were used for distribution of gene lengths for protein-coding mRNAs and novel lncRNAs in spinal cord astrocytoma samples through Histogram plot.

### 2.6 Gene Ontology (GO) Analysis

The significantly co-expressed mRNAs with lncRNAs that were used for GO and KEGG for Biological function to study regulatory pathways. In total 38 coexpressed mRNAs were obtained. mRNAs were tested for enrichment using a hypergeometric test, compared against a background of lncRNAs. FDR-corrected p-values were used to find significantly enriched GO and Kegg terms using biomaRT(version 2.58.2) (Durinck S., 2009) and gprofiler2(version 0.2.4) (Kolberg L., 2020) through Biocmanager.

### 2.7 Variation Region Analysis

SNP analysis was performed using the Genome Analysis Toolkit (GATK) (Mckenna A., 2010) pipeline on the aligned RNA-seq BAM files. The BAM files were first sorted and indexed, followed by duplicate marking with GATK MarkDuplicates. RNA-seq alignments spanning splice junctions were processed using GATK SplitNCigarReads before SNP calling. SNPs were identified for each sample using GATK HaplotypeCaller in GVCF mode. Individual GVCF files were combined using CombineGVCFs, and joint genotyping was performed using GenotypeGVCFs to generate a multisample VCF file. High-confidence SNPs were obtained after quality filtering using VariantFiltration and were subsequently annotated using reference genome . SNPs overlapping the identified novel lncRNA regions were extracted using BEDTools (version 02.31.1) (Quinlan A. R., 2010) intersect. The identified SNPs were further analyzed using R (ggplot2 and dplyr) and Python (pandas and matplotlib) to evaluate chromosome-wise SNP distribution, SNP density, SNP types, transition/transversion (Ti/Tv) ratio, SNP burden per lncRNA, and the correlation between lncRNA expression and SNP count.

### 2.8 Conservation and LiftOver Analysis

To assess the evolutionary conservation for all high-confidence identified novel lncRNAs to evaluate the overall conservation pattern through their genomic coordinates from the human reference genome (GRCh38) were mapped to the mouse reference genome (GRCm39/mm39) using the UCSC LiftOver tool (Kent W. J., 2002). Mouse was selected because it is a well-established model organism for studying neurological disorders, and conserved genomic regions are more likely to have retained biological functions across mammals. Successfully mapped lncRNA loci were considered mouse-conserved, whereas unmapped loci were classified as having no detectable mouse orthologous region. The conservation results were summarized and visualized as bar plots using R (ggplot2).

# 3. Results

## 3.1 Sequencing Quality Control and Alignment Assessment

Raw RNA-seq reads from Six spinal cord astrocytoma samples divided into Young and Old Groups were subjected to quality control, filtering, and alignment to the human reference genome (GRCh38). Alignment percentage ranged from 93% to 97%, indicating high-quality mapping across all samples as per shown in Table no. 1. After filtering, about 11-28 million reads were retained per sample. Duplication rates ranged from 7.3% to 55.8%, which is acceptable for RNA-seq datasets and may reflect highly expressed transcripts. Overall, the sequencing data were of sufficient quality for downstream transcriptomic analyses.

**Table no. 1 Sequencing quality control and alignment statistics for spinal cord astrocytoma RNA-seq samples.**

| Samples with age (SRA ID) | % Aligned | % Duplication | Reads After Filtering |
|---|---|---|---|
| SRR15890983 (69 y/o) | 97.09% | 32.0% | 11933960 |
| SRR15891001 (67 y/o) | 93.74% | 7.3% | 18018940 |
| SRR15891003 (56 y/o) | 96.20% | 10.6% | 28806144 |
| SRR15890993 (18 y/o) | 96.67% | 10.8% | 19686240 |
| SRR15890985 (20 y/o) | 95.44% | 55.8% | 17254529 |
| SRR15890987 (23 y/o) | 96.91% | 8.8% | 19622077 |

### 3.2 Expression Category and Age-wise Expression Changes

The sample-wise distribution of log2-normalized mRNA expression showed similar expression patterns across all six spinal cord samples, indicating successful normalization and good overall data quality (Figure no. 1-B). Although small differences in median VST-normalized expression and interquartile range were observed, no major variation was detected between the young and old groups. In the mRNAs distribution 18Y Male and 67Y Male samples showed relatively higher median log2-normalized expression, whereas the 20Y Female, 23Y Male, and 69Y Male samples showed comparatively lower median expression.

A few outlier genes were observed across all samples. Based on the mean log2-normalized expression values, mRNAs were classified into three expression groups: low (1.57–4.76), moderate (4.76–6.45), and high (6.45 upto 12).

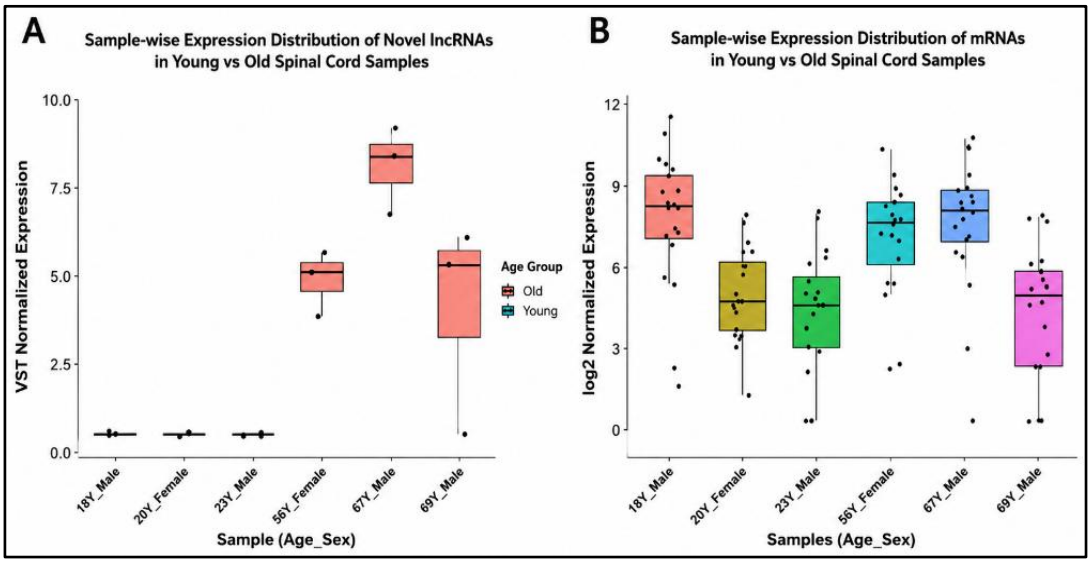


**Figure no. 1:** Sample-wise expression distribution in young and old spinal cord samples. **A)** VST-normalized expression distribution of mRNAs across samples. **B)** log2-normalized expression distribution of mRNAs across the same samples.

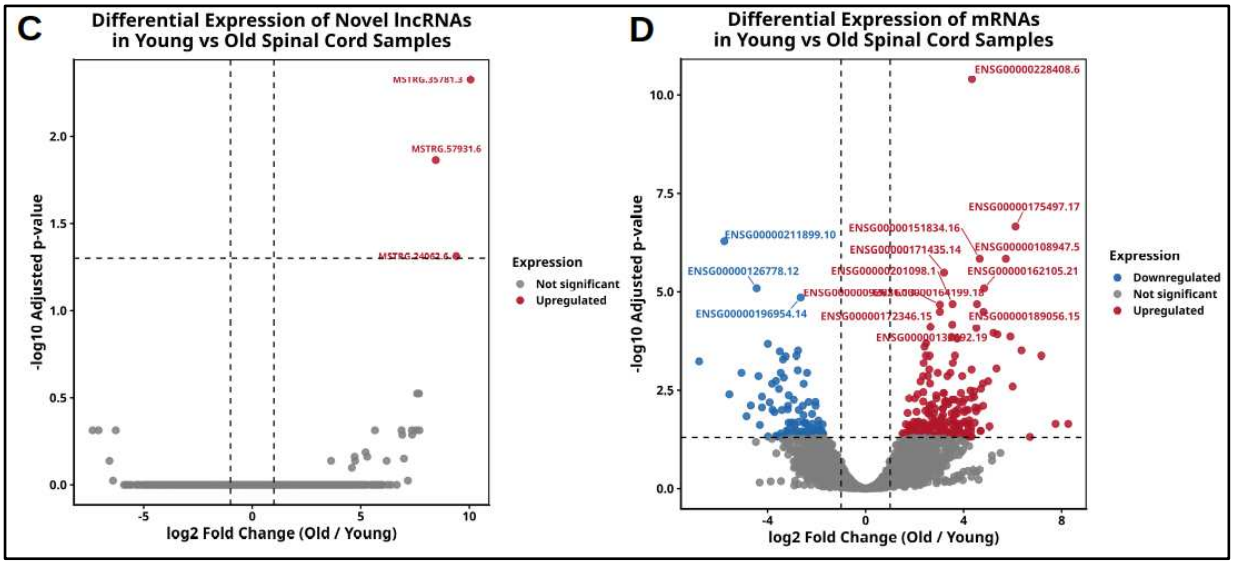
In contrast, the sample-wise distribution of VST-normalized novel lncRNA expression showed a clear difference between the young and old spinal cord samples (Figure no. 1-A). The three young samples (18Y Male, 20Y Female, and 23Y Male) showed consistently low VST-normalized expression with very little variation. However, the old samples (56Y Female, 67Y Male, and 69Y Male) showed much higher VST-normalized expression levels. Among them, the 67Y Male sample exhibited the highest median VST-normalized expression, followed by the 69Y Male and 56Y Female samples. In addition, the old samples showed greater variation in VST-normalized lncRNA expression than the young samples. Based on the mean VST-normalized expression values, novel lncRNAs were grouped into low (0.0–0.73), moderate (0.73–4.99), and high (4.99–10.0) expression categories.

Overall, these results indicate that log2-normalized mRNA expression remained relatively stable between the young and old spinal cord samples, whereas VST-normalized novel lncRNA expression showed increased expression and greater variability in the old group. These findings provide further insights into the potential involvement of novel lncRNAs in spinal cord aging

### 3.3 Differential expression analysis of mRNAs and lncRNAs

Differential expression analysis was performed using DESeq2 to identify age-associated transcriptional changes between the Young (18Y Male, 20Y Female, and 23Y Male) and Old (56Y Female, 67Y Male, and 69Y Male) spinal cord samples. For both mRNA and novel lncRNA datasets, transcripts with an adjusted $P$ value $< 0.05$ and absolute $\log_2$ fold change $> 1$ were considered significantly differentially expressed. The volcano plots demonstrated distinct transcriptional differences between the two age groups, with significantly upregulated transcripts in the Old group represented by positive $\log_2$ fold changes, whereas negatively regulated transcripts showed higher expression in the Young group.

For the mRNA expression profile there total 18692 mRNAs, the volcano plot (Figure no. 2-D) revealed a broad distribution of differentially expressed genes, indicating extensive age-associated alterations in protein-coding gene regulation. In these 301 mRNAs showed significantly where several in upregulation from the Old spinal cord group, including ENSG00000228408.6, ENSG00000175497.17, ENSG00000151834.16, ENSG00000108947.5, and ENSG00000162105.21, suggesting increased activity of specific transcriptional programs during aging. Conversely, genes such as ENSG00000211899.10, ENSG00000126778.12, and ENSG00000196954.14 showed reduced expression in the Old group, indicating relatively higher expression levels in Young spinal cord samples. These findings demonstrate that aging is




accompanied by substantial remodeling of the mRNA transcriptome, involving both activation and suppression of multiple gene expression pathways.

**Figure no. 2: Volcano plots** showing differential expression between young and old spinal cord samples. **C)** Novel lncRNAs and **D)** mRNAs showing significantly up-regulated and down-regulated and non-significant transcripts. Dashed vertical lines indicate the |log2 fold change| threshold of 1 and the dashed horizontal line indicates the adjusted p-value threshold of 0.05.

In parallel, novel lncRNA expression identified 1,669 novel lncRNAs from the spinal cord RNA-seq dataset. Among these, only three novel lncRNAs (MSTRG.57931.6, MSTRG.24062.6, and MSTRG.35781.3) showed significant differential expression between Young and Old groups. The

given plot (Figure no. 2-C) demonstrated that these lncRNAs were associated with the Old spinal cord group, exhibiting positive $log_2$fold changes and adjusted *P* values. Expression profiling using VST-normalized values further confirmed that these transcripts showed consistently higher expression in Old samples compared with Young samples. Among the identified lncRNAs, MSTRG.24062.6 showed expression restricted mainly to the Old group, whereas MSTRG.57931.6 and MSTRG.35781.3 displayed elevated expression particularly in the 67-year-old male sample with minimal or absent expression in most Young samples.

Together, the mRNA and novel lncRNA expression indicate age-related transcriptional alterations in the human spinal cord. While mRNA profiling revealed widespread changes involving numerous protein-coding genes, the identification of a small number of highly significant novel lncRNAs suggests that these transcripts may represent specific regulatory molecules associated with spinal cord aging. The distinct expression patterns of these age-associated mRNAs and lncRNAs provide potential candidates for further functional investigation through co-expression, pathway enrichment, and conservation studies.

Heatmap analysis revealed distinct expression patterns between Young and Old human spinal cord samples (Figure no. 3-E and F). Among the 1,669 novel lncRNAs identified, only three novel lncRNAs showed significant differential expression between the two age groups.

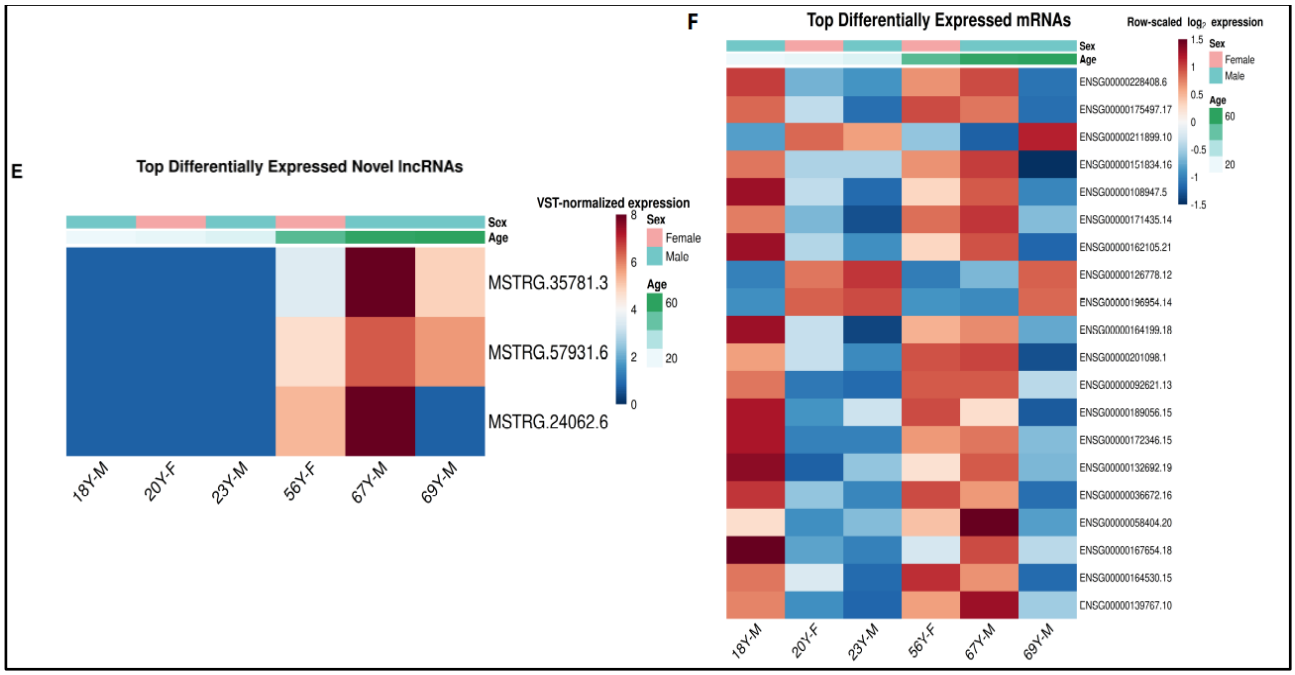



**Figure no. 3:** Clustering Heatmaps of top differentially expressed transcripts in young and old spinal cord sample: **E)** novel lncRNAs (VST-normalized expression range:- ~2-8) and **F)** mRNAs (log-transformed expression range: ~-1.5 to +1.5). Color intensity represents expression levels.

These lncRNAs were MSTRG.24062.6, MSTRG.57931.6, and MSTRG.35781.3. All three novel lncRNAs showed an overall higher expression in Old samples compared with Young samples(Figure no. 3-E). The highest expression of these transcripts was observed in the 67-year-old male sample. In comparison, their expression was relatively lower across most of the Young samples. These results indicate that the identified novel lncRNAs may be associated with age-related transcriptional changes in spinal cord tissue.

Similarly, the heatmap analysis of the top 20 differentially expressed mRNAs showed clear differences between Young and Old samples(Figure no. 3-F). The expression patterns of several mRNAs differed substantially between the two age groups. This indicates that mRNA expression is also influenced by age in human spinal cord tissue. The observed differences in both lncRNA and mRNA expression support the presence of age-associated transcriptional changes. Overall, the heatmap analysis identified distinct age-related expression patterns in both novel lncRNAs and mRNAs. These findings highlight the potential involvement of these differentially expressed transcripts in the molecular changes associated with spinal cord aging.

### 3.4 Co-expression and correlation of mRNAs and lncRNAs with Gene Length

Figure no. 4 shows the distribution of gene lengths for significantly co-expressed novel lncRNAs and protein-coding mRNAs identified from spinal cord transcriptome analysis. The histogram was used to compare the lengths of these two types of transcripts. The average length of novel lncRNAs was approximately 2302 bp, while protein-coding mRNAs showed a slightly higher average length of approximately 2941 bp. Most of the identified lncRNAs and mRNAs were present within the shorter gene-length range. However, protein-coding mRNAs showed a wider distribution, with some transcripts having larger gene lengths compared with novel lncRNAs. This indicates that the identified lncRNAs and mRNAs have different structural length patterns in the analyzed spinal cord samples.

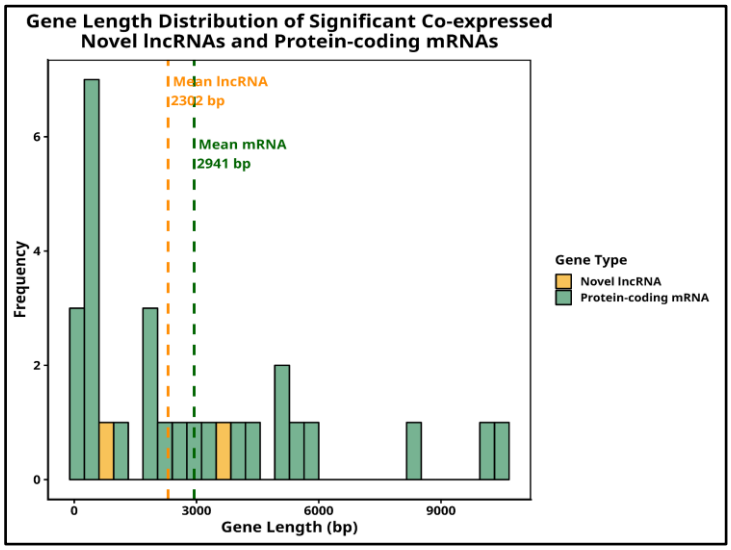



**Figure no. 4:** Gene length distribution of significant co-expressed novel lncRNAs and protein-coding mRNAs.

To check whether the difference in gene length between novel lncRNAs and protein-coding mRNAs was statistically meaningful, two statistical tests were performed. An unpaired two-sample t-test was used to compare the average lengths of the two groups, while the Mann–Whitney U test was used to compare the overall distribution of gene lengths between lncRNAs and mRNAs. In which 38 significantly co-expressed mRNAs identified with three novel lncRNAs. These tests help determine whether the observed difference occurred due to a true biological difference or random variation. Based on the obtained statistical results, the difference in gene length between novel lncRNAs and mRNAs was evaluated for significance.

Overall, this analysis provides information about novel lncRNAs and protein-coding mRNAs identified in spinal cord samples. The gene length comparison helps in understanding the basic genomic characteristics of these transcripts and supports the characterization of novel lncRNAs identified in this study.

### 3.5 Regulatory Pathways of Co-expressed lncRNA with mRNA

To investigate the potential biological functions of 38 genes co-expressed with the identified novel lncRNAs with mRNAs, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment were performed. The enrichment revealed that the co-expressed mRNA genes were mainly associated with neuronal functions, membrane organization, protein localization, and cellular transport processes.

KEGG pathway analysis identified adrenergic signaling in cardiomyocytes (KEGG:04261) as a significantly enriched pathway ($p = 0.00017$)(Figure no. 4-H). Although this pathway is primarily characterized in cardiac tissues, several genes involved in adrenergic signaling are also associated with ion transport, membrane potential regulation, and signal transduction processes, suggesting possible involvement of these co-expressed genes in cellular communication and neuronal signaling mechanisms.

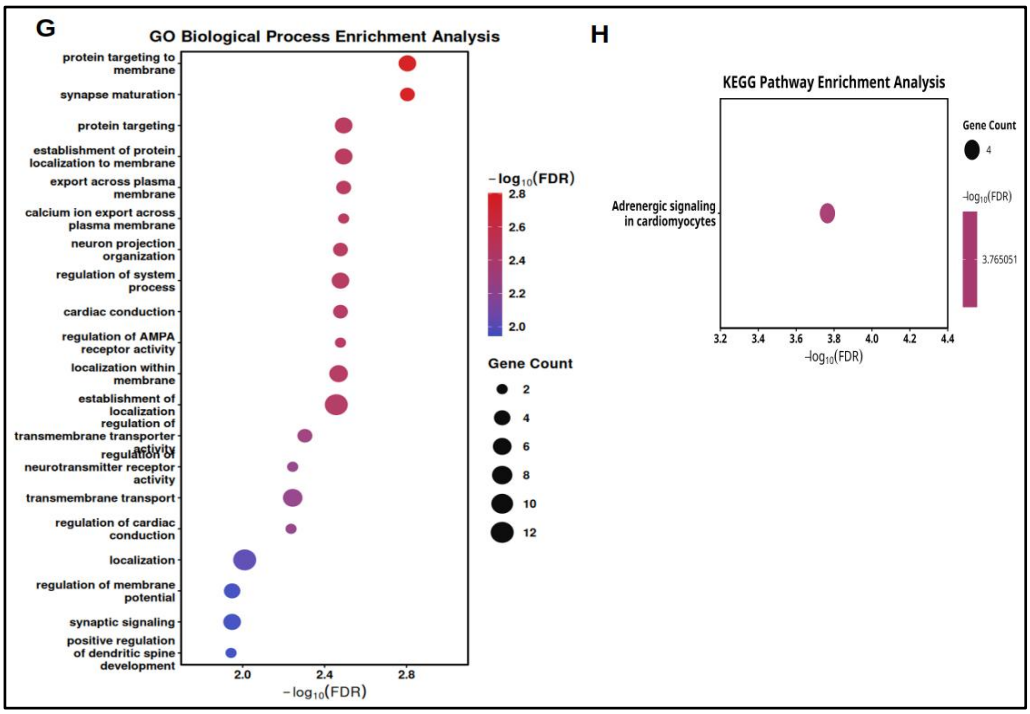



**Figure no. 5: GO Biological Process (G) and KEGG pathway enrichment (H)** analysis of co-expressed genes. Bubble size represents the number of genes enriched in each term or pathway, while color intensity indicates enrichment significance expressed as −log10(FDR); higher values correspond to greater statistical significance after false discovery rate correction.

GO biological process enrichment analysis showed significant enrichment of multiple terms related to protein targeting and membrane localization, including *protein targeting to membrane* (GO:0006612, p = 0.00157), *protein targeting* (GO:0006605, p = 0.00321), and *establishment of protein localization to membrane* (GO:0090150, p = 0.00321)as per shown in (Figure no. 5-G). These findings indicate that the lncRNA-associated genes may contribute to the Protein trafficking and membrane-associated cellular functions.

Furthermore, several enriched GO terms were related to neuronal development and synaptic functions, including *synapse maturation* (GO:0060074, p = 0.00157), *neuron projection organization* (GO:0106027, p = 0.00333), *synaptic signaling* (GO:0099536, p = 0.01125), *chemical synaptic transmission* (GO:0007268, p = 0.0287), and *regulation of neuronal synaptic plasticity* (GO:0048168, p = 0.01305). Additional enrichment was observed for processes involved in dendritic spine development, including *regulation of dendritic spine development* (GO:0060998) and *dendrite morphogenesis* (GO:0048813), suggesting a potential role of these lncRNA–mRNA interactions in neuronal structure maintenance and synaptic regulation.

The analysis also identified enrichment of genes involved in ion transport and membrane excitability, including *transmembrane transport* (GO:0055085, p = 0.00569), *calcium ion transport* (GO:0006816, p = 0.0323), *regulation of membrane potential* (GO:0042391, p = 0.01125), and *monoatomic ion transport* (GO:0006811, p = 0.0287). These processes are important for maintaining neuronal electrical activity and intracellular signaling.

The functional enrichment results suggest that the identified lncRNA-associated mRNAs are predominantly involved in neuronal signaling, synaptic organization, membrane protein localization, and ion homeostasis. These findings indicate that the novel lncRNAs participate in age-associated regulation of spinal cord function by influencing genes involved in neuronal communication and cellular maintenance.

### 3.6 SNP Variation in Novel lncRNAs

#### 3.6.1 Genome-wide Distribution of SNP Variants in Novel lncRNAs

Variant analysis of the identified novel lncRNAs revealed a total of 11,243 genetic variants across 496 novel lncRNAs. Among these, 10,707 (95.2%) were single nucleotide polymorphisms (SNPs), while 345 (3.1%) were deletions and 191 (1.7%) were insertions. No multiple nucleotide polymorphisms (MNPs) or complex variants were detected (Figure no. 6-I). These results indicate that SNPs are the predominant form of genetic variation present in the identified novel lncRNAs.

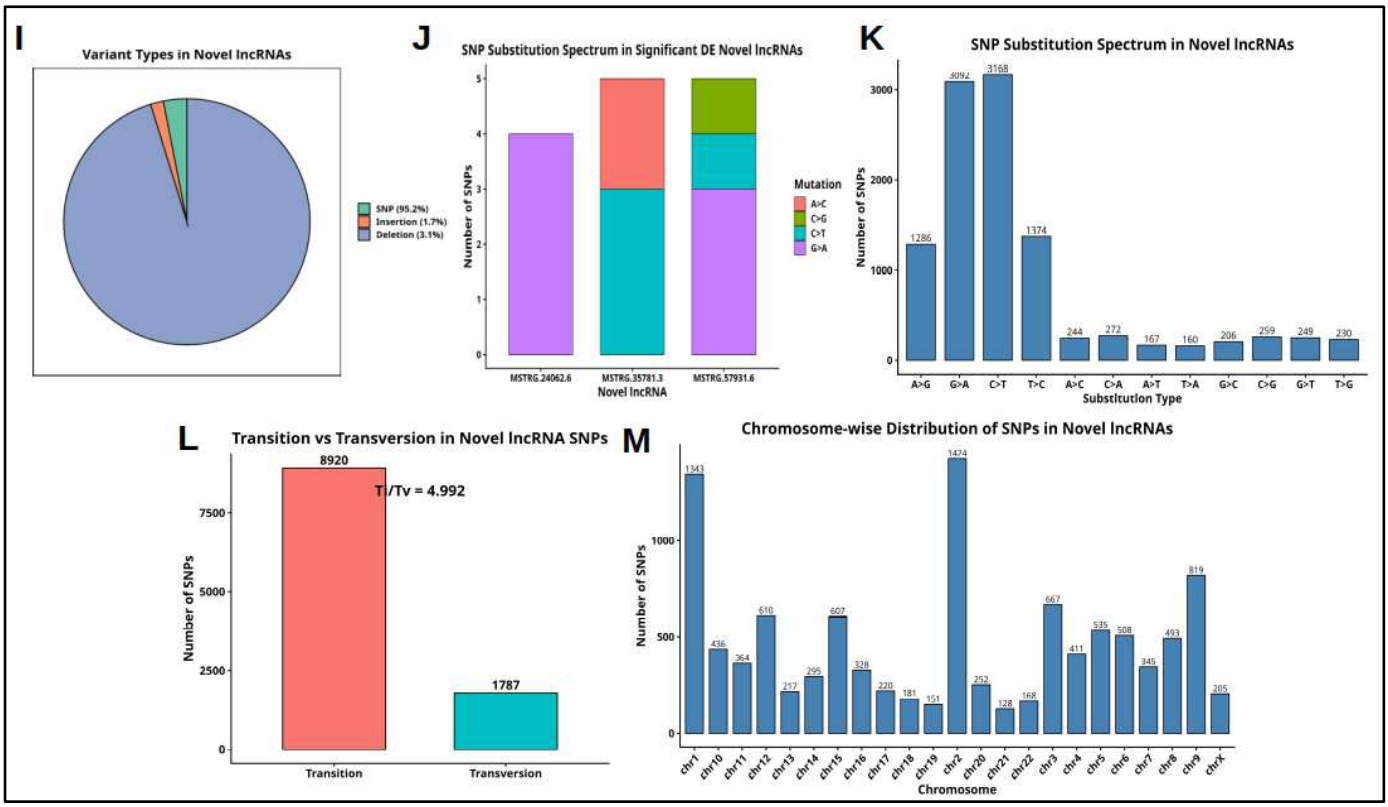



**Figure no. 6:** Variant landscape of novel lnRNAs showing variant types, SNP substitution spectra transition/transversion analysis, and chromosomal distribution of SNPs.

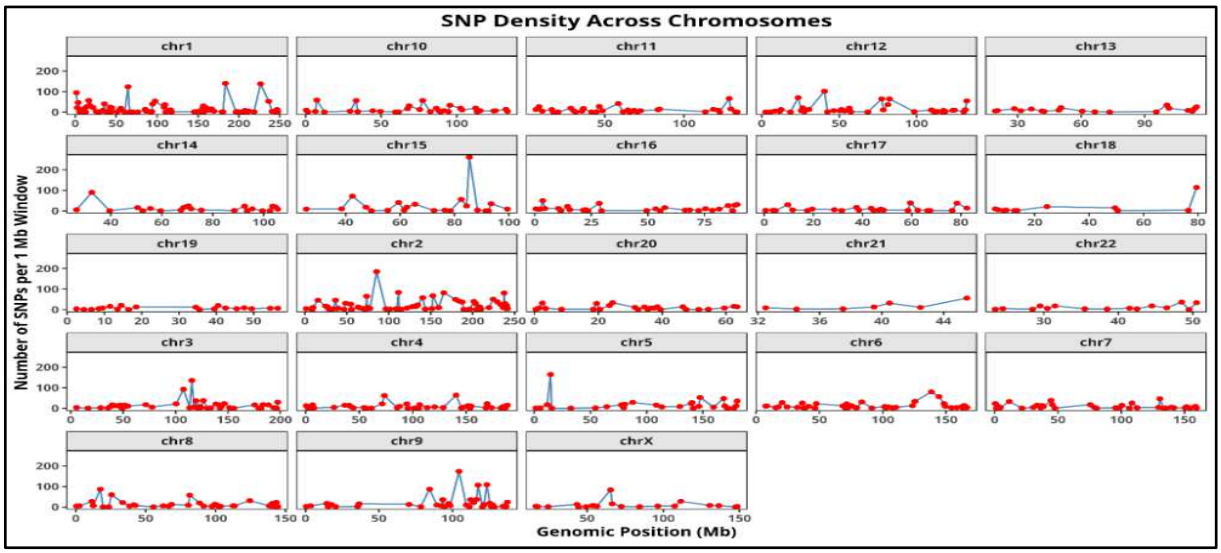
The SNP substitution analysis showed that all twelve possible nucleotide substitution types were represented (Figure no. 6-K). Among them, the transition substitutions C>T (3,168 SNPs) and G>A (3,092 SNPs) were the most frequent, whereas transversion substitutions occurred at much lower frequencies. This pattern indicates that transition mutations are the dominant type of nucleotide substitution in novel lncRNAs.

### 3.6.2 Transition Mutations are More Frequent than Transversions

To characterize the mutational pattern of SNPs, substitutions were classified into transitions and transversions. Out of the 10,707 SNPs, 8,920 (83.3%) were transition mutations, whereas 1,787 (16.7%) were transversion mutations, resulting in a transition/transversion (Ti/Tv) ratio of 4.992 (Figure no. 6-L). The predominance of transition mutations agrees with the substitution spectrum, where C>T and G>A substitutions were the most abundant mutation types.

### 3.6.3 Chromosome-wise Distribution of SNPs

The distribution of SNPs across chromosomes was not uniform (Figure no. 6-M and Figure no. 7). SNPs were detected on all 23 chromosomes, with chromosome 2 containing the highest number of




SNPs (1,424; 13.3%), followed by chromosome 1 (1,343; 12.5%), chromosome 9 (819; 7.7%), chromosome 3 (667; 6.2%), chromosome 12 (610; 5.7%), and chromosome 15 (607; 5.7%). In contrast, chromosome 21 showed the lowest number of SNPs (128; 1.2%), followed by chromosome 19 (151; 1.4%), chromosome 22 (168; 1.6%), chromosome 18 (181; 1.7%), and chromosome X (205; 1.9%).

**Figure no. 7:** Genome-wide distribution of SNP density across chromosomes in novel lncRNAs

These findings demonstrate that SNPs are distributed across the entire genome but are enriched on specific chromosomes. The chromosome-wise SNP density analysis further showed that SNPs were unevenly distributed within chromosomes (Figure no. 7). Several genomic regions exhibited higher SNP density than surrounding regions, indicating localized SNP-rich regions rather than a uniform distribution.

### 3.6.4 SNP Burden Among Novel lncRNAs

The number of variants differed considerably among individual novel lncRNAs (Figure no. 8). Among the 496 novel lncRNAs analyzed, MSTRG.21530.3 contained the highest number of variants (212), followed by MSTRG.32040.2 (181), MSTRG.59732.10 (170), MSTRG.5129.1 (144), MSTRG.46277.8 (141), MSTRG.40751.4 (133), MSTRG.1963.3 (129), MSTRG.4318.3 (129), MSTRG.59960.9 (119), and MSTRG.13257.4 (107). The observed variation in SNP burden suggests that some novel lncRNAs accumulate substantially more sequence variation than others.

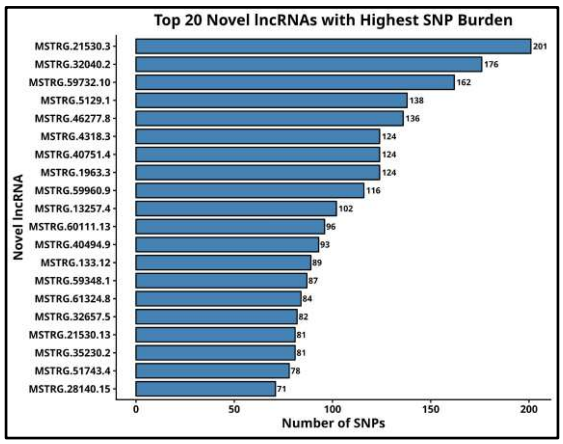



**Figure no. 8:** Distribution of SNP Counts Among the Top 20 Novel lncRNAs

### 3.6.5 Association Between SNP Count and lncRNA Expression

The relationship between SNP count and transcript expression was evaluated using average TPM values (Figure no. 9). A weak but statistically significant positive correlation was observed between SNP count and expression level (Pearson r = 0.168, $P = 1.69 \times 10^{-8}$; Spearman ρ = 0.273, $P = 1.83 \times 10^{-20}$). These results indicate that although transcripts with higher expression tend to contain slightly more SNPs, the overall association between expression level and SNP burden is relatively weak.

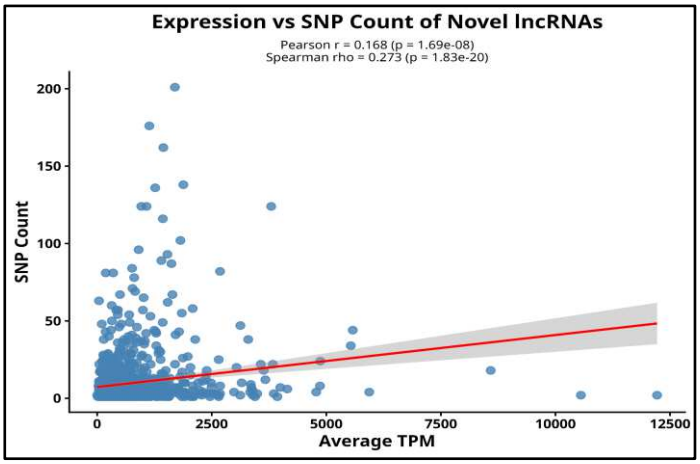



**Figure no. 9:** Association Between Novel lncRNA Expression Levels and SNP Counts

### 3.6.6 Distribution of SNPs Across Biological Samples

To examine the distribution of variants among biological samples, SNP counts were summarized across the ten sample groups (B1–B10) (Figure no. 10). The highest number of variants was observed in B10 (1,099 variants), followed by B1 (1,048) and B9 (968). The remaining groups contained 950

(B3), 935 (B2), 926 (B8), 908 (B7), 906 (B5), 868 (B6), and 820 (B4) variants. Hierarchical clustering of the top SNP-rich novel lncRNAs showed differences in SNP abundance across samples, with several lncRNAs displaying consistently higher SNP counts across multiple biological replicates.

The heatmap of the top 50 novel lncRNAs shows the distribution of SNPs across the ten spinal cord samples (B1–B10). Each row represents a novel lncRNA, while each column represents an individual sample. The color intensity indicates the number of SNPs detected in each lncRNA, with light yellow representing fewer SNPs and dark red representing a higher number of SNPs. The color scale ranges from 0 to approximately 35 SNPs, where increasing color intensity reflects an increase in SNP count. Hierarchical clustering grouped lncRNAs with similar SNP distribution patterns across the samples. Some novel lncRNAs showed consistently higher SNP counts in multiple samples, whereas others contained very few or no SNPs, indicating differences in genetic variation among the identified novel lncRNAs.

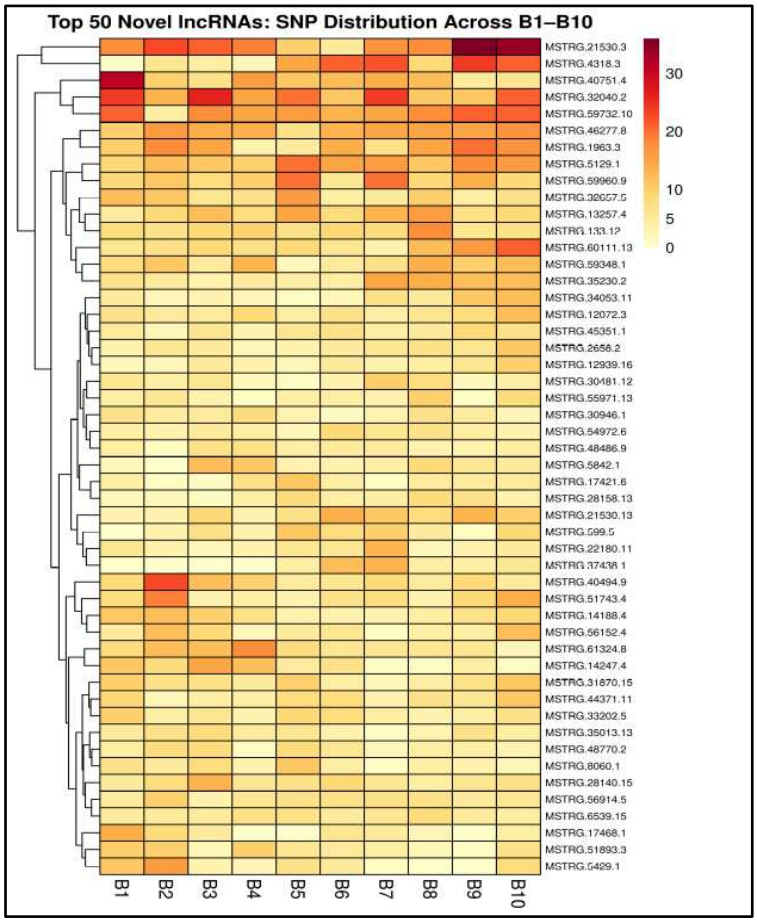



**Figure no. 10:** Hierarchical clustering heatmap of SNP burden across genomic bins in the top 50 novel lncRNAs

### 3.7 Understanding of Human-Specific lncRNA Regulators in Spinal Cord Astrocytoma

UCSC liftOver analysis revealed that all 1669 tumor-associated lncRNAs identified in spinal cord astrocytoma were human-specific, as none could be successfully mapped from the human genome (GRCh38) to the mouse genome (GRCm39). Among these, 3 highly significant novel lncRNAs were analyzed in previous differential expressions.

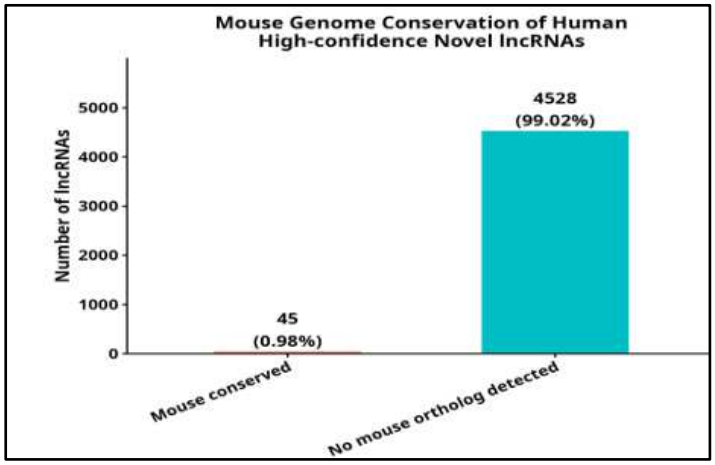



**Figure no. 11:** Cross-species conservation of human high-confidence novel lncRNAs in mice.

In contrast, analysis of reference GENCODE lncRNAs showed that only 99.02% were human-specific, while the remaining 0.98% exhibited conservation in the mouse genome (Figure no. 11). These findings indicate a striking lack of evolutionary conservation among astrocytoma-associated lncRNAs compared with the broader lncRNA population.

## 4. Discussion

The present study provides a comprehensive analysis of novel long non-coding RNAs (lncRNAs) associated with human spinal cord aging using differential expression, co-expression, functional enrichment, conservation, and SNP analyses. Among the 1,669 novel lncRNAs identified, only three transcripts (MSTRG.24062.6, MSTRG.35781.3, and MSTRG.57931.6) were significantly differentially expressed between the young and old spinal cord samples. These lncRNAs showed higher expression in the older group, suggesting that they may be involved in age-related changes in the human spinal cord. Previous studies have shown that lncRNAs play important roles in regulating gene expression, RNA processing, and maintaining cellular functions during aging and neurological disorders (Yao et al., 2019; Statello et al., 2021).

Another important finding was that novel lncRNA expression showed greater variation in the older samples, whereas mRNA expression remained relatively stable between the two age groups. This suggests that aging may affect lncRNA expression more than protein-coding genes. Similar changes in gene regulation during aging have been reported in previous studies, where regulatory RNAs become more variable with increasing age (Lopez-Otín et al., 2013; Statello et al., 2021).

Functional enrichment analysis suggested that the genes co-expressed with the significant lncRNAs are mainly involved in protein targeting to membranes, synapse maturation, neuron projection organization, synaptic signaling, calcium ion transport, membrane potential regulation, and other neuronal functions. These processes are essential for normal neuronal communication and spinal cord function. Previous studies have also reported that changes in synaptic signaling, ion transport, and membrane organization are closely associated with aging and neurodegeneration (Mattson & Arumugam, 2018; Hou et al., 2019). In addition, KEGG analysis identified adrenergic signaling in cardiomyocytes as the most significant pathway. Although this pathway is mainly known for its role in the heart, several genes involved in this pathway also participate in calcium signaling and signal transduction in neurons, suggesting that they may also contribute to spinal cord function (Lefkowitz & Shenoy, 2005).

The lncRNA conservation analysis showed that none of the 1,669 novel lncRNAs could be mapped to the mouse genome, indicating that these transcripts are human-specific. This finding agrees with previous studies showing that lncRNAs are generally less conserved than protein-coding genes and often perform species-specific regulatory functions (Peng et al., 2018; Statello et al., 2021). Therefore, some regulatory mechanisms involved in human spinal cord aging may not be completely represented in mouse models.

The SNP analysis identified 11,243 genetic variants across 496 novel lncRNAs, with more than 95% of the variants being SNPs. Transition mutations were much more common than transversions, with C>T and G>A being the most frequent substitution types. This mutation pattern is similar to that reported in previous studies of the human genome (Lynch, 2010; Alexandrov et al., 2020). SNPs were found on all chromosomes but were not evenly distributed, and chromosome 2 contained the highest number of SNPs. Some novel lncRNAs also contained many more SNPs than others, suggesting that genetic variation differs among lncRNAs. Previous studies have reported that SNPs can influence lncRNA structure and function by affecting their interactions with other molecules (Carrieri et al., 2012; Gong et al., 2015). Although a weak positive correlation was observed between SNP count and lncRNA expression, the results suggest that SNP accumulation and expression changes may occur independently during spinal cord aging.

One limitation of this study is the small number of RNA-seq samples, which limits the ability to draw strong biological conclusions. In addition, the functions of the identified lncRNAs and the effects of the detected SNPs were predicted using bioinformatics analyses and require further validation using larger sample sizes and experimental approaches such as qRT-PCR and functional studies.

Overall, this study provides new information about novel lncRNAs associated with human spinal cord aging. The identified age-related lncRNAs, enriched biological pathways, human-specific conservation, and SNP profiles provide a useful foundation for future studies to better understand the molecular mechanisms involved in spinal cord aging.

## 5. Conclusion

Novel long non-coding RNAs (lncRNAs) identified in this study provide new insights into the molecular changes associated with human spinal cord aging. Among the 1,669 novel lncRNAs identified, only three transcripts (MSTRG.24062.6, MSTRG.35781.3, and MSTRG.57931.6) were significantly differentially expressed between the young and old spinal cord samples, suggesting that they may have important regulatory roles during the aging process.

Functional enrichment analysis indicated that the co-expressed genes of these lncRNAs are involved in biological processes related to protein targeting to membrane, synaptic signaling, neuron projection organization, calcium ion transport, and regulation of membrane potential. These findings suggest that the identified lncRNAs may contribute to maintaining neuronal communication and spinal cord function during aging. Conservation analysis further showed that all identified novel lncRNAs were human-specific, highlighting their potential importance in human biology and the limitations of conventional animal models for studying these regulatory transcripts.

Genome-wide SNP analysis identified 11,243 genetic variants across 496 novel lncRNAs, with single nucleotide polymorphisms (SNPs) representing more than 95% of all detected variants. Transition mutations were more frequent than transversions, and SNPs showed uneven distribution across chromosomes and individual lncRNAs. Although only a weak positive correlation was observed between SNP count and lncRNA expression, the results indicate that sequence variation and expression changes may independently contribute to lncRNA regulation during spinal cord aging.

Although further validation using larger sample sizes and experimental studies is required, these findings provide a valuable resource for understanding the expression patterns, biological functions, evolutionary conservation, and genetic variation of novel lncRNAs in the aging human spinal cord. The identified candidate lncRNAs may serve as promising targets for future studies aimed at

understanding the molecular mechanisms of spinal cord aging and exploring their potential as biomarkers or therapeutic targets for age-related spinal cord disorders.

## 6. References


1) Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Wu, Y., ... Stratton, M. R. (2020). *The repertoire of mutational signatures in human cancer*. Nature, 578(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

2) Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* (Version 0.12.1). Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3) Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics, 27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

4) Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., ... Gustincich, S. (2012). *Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat*. Nature, 491(7424), 454–457. https://doi.org/10.1038/nature11508

5) Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

6) Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

7) 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

8) Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F.. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research, 47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

9) Gong, J., Liu, W., Zhang, J., Miao, X., Guo, A.-Y., Han, L., ... Li, C.-Y. (2015). *LNCediting: A database for functional effects of RNA editing in long non-coding RNAs*. Nucleic Acids Research, 43(Database issue), D181–D185. https://doi.org/10.1093/nar/gku1067

10) Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., *et al.* (2024). Ensembl 2024**.** *Nucleic Acids Research, 52*(D1), D891–D899. https://doi.org/10.1093/nar/gkad1049

11) Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., & Gao, G. (2017). CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research, 45*(W1), W12–W16. https://doi.org/10.1093/nar/gkx428

12) Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature *Biotechnology, 37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

13) 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

14) Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., & Peterson, H. (2020). gprofiler2: An R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research, 9*(ELIXIR), 709. https://doi.org/10.12688/f1000research.24956.2

15) Konar, S. K., Maiti, T. K., Bir, S. C., Kalakoti, P., Bollam, P., Nanda, A. (2019). *Primary spinal astrocytomas: A literature review*. *Cureus, 11*(9), e5931. https://doi.org/10.7759/cureus.5931

16) Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

17) López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). *The hallmarks of aging*. Cell, 153(6), 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039

18) Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., & Kleihues, P. (2007). *The 2007 WHO classification of tumours of the central nervous system*. *Acta Neuropathologica, 114*(2), 97–109. https://doi.org/10.1007/s00401-007-0243-4

19) Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

20) Mattson, M. P., & Arumugam, T. V. (2018). *Hallmarks of brain aging: Adaptive and pathological modification by metabolic states*. Cell Metabolism, 27(6), 1176–1199. https://doi.org/10.1016/j.cmet.2018.05.011

21) McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data*. Genome Research, 20(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

22) Morgan, M., Ramos, M., et al. (2023). *BiocManager: Access the Bioconductor Project Package Repository* (Version 3.18). Bioconductor. https://bioconductor.org/packages/BiocManager/

23) NCBI SRA Toolkit (Version 3.0.3) [Computer software]. National Center for Biotechnology Information.https://github.com/ncbi/sra-tools

24) Ostrom, Q. T., Gittleman, H., Liao, P., Vecchione-Koval, T., Wolinsky, Y., Kruchko, C., & Barnholtz-Sloan, J. S. (2015). CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. *Neuro-Oncology, 17*(Suppl 4), iv1-iv62.

25) Peng, Z., Liu, C., & Wu, M. (2018). *New insights into long noncoding RNAs and their roles in glioma*. *Molecular Cancer, 17*(1), 61. https://doi.org/10.1186/s12943-018-0812-2.

26) Peng, W.-X., Koirala, P., & Mo, Y.-Y. (2017). *LncRNA-mediated regulation of cell signaling in cancer*. Oncogene, 36(41), 5661–5667. https://doi.org/10.1038/onc.2017.184

27) Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology, 33*(3), 290–295. https://doi.org/10.1038/nbt.3122

28) Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research, 9*, 304. https://doi.org/10.12688/f1000research.23297.2

29) Python Software Foundation. (2024). *Python (Version 3.12.7)* [Computer software]. https://www.python.org/

30) Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

31) Quinn, J. J., & Chang, H. Y. (2016). *Unique features of long non-coding RNA biogenesis and function*. Nature Reviews Genetics, 17(1), 47–62. https://doi.org/10.1038/nrg.2015.10

32) R Core Team. (2024). *R: A Language and Environment for Statistical Computing* (Version 4.3.3). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

33) Schmitt, A. M., & Chang, H. Y. (2016). *Long noncoding RNAs in cancer pathways*. Cancer Cell, 29(4), 452–463. https://doi.org/10.1016/j.ccell.2016.03.010

34) Statello, L., Guo, C.-J., Chen, L.-L., & Huarte, M. (2021). *Gene regulation by long non-coding RNAs and its biological functions*. Nature Reviews Molecular Cell Biology, 22(2), 96–118. https://doi.org/10.1038/s41580-020-00315-9

35) Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J. B., Belanger, K., Brandes, A. A., Marosi, C., Bogdahn, U., Curschmann, J., Janzer, R. C., Ludwin, S. K., Gorlia, T., Allgeier, A., Lacombe, D., Cairncross, J. G., Eisenhauer, E., & Mirimanoff, R. O. (2005). *Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma*. *New England Journal of Medicine, 352*(10), 987–996. https://doi.org/10.1056/NEJMoa043330

36) Subramanian, A., Capalbo, A., Iyengar, N. R., Rizzo, R., di Campli, A., Di Martino, R., ... Luini, A. (2019). Steady-state regulation of COPII-dependent secretory cargo sorting by inositol trisphosphate receptors, calcium, and penta EF hand proteins. *eLife, 8*, e43231.

37) Sundaram, S., Johnson, P. C., & Raghavan, M. (2019). Primary spinal cord astrocytoma: A population-based analysis. *Cureus, 11*(7), e5217.

38) Thakkar, J. P., Dolecek, T. A., Horbinski, C., Ostrom, Q. T., Lightner, D. D., Barnholtz-Sloan, J. S., & Villano, J. L. (2014). *Epidemiologic and molecular prognostic review of glioblastoma*. *Cancer Epidemiology, Biomarkers & Prevention, 23*(10), 1985–1996. https://doi.org/10.1158/1055-9965.EPI-14-0275

39) Venkataramani, V., & Winkler, F. (2025). Reactive astrocytes in glioma: Emerging opportunities and challenges. *Brain Pathology, 35*(2), e13123.

40) Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation* (R package version 1.2.1). https://CRAN.R-project.org/package=dplyr

41) Yao, R.-W., Wang, Y., & Chen, L.-L. (2019). *Cellular functions of long noncoding RNAs*. Nature Cell Biology, 21(5), 542–551. https://doi.org/10.1038/s41556-019-0311-8

42) Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., ... He, S. (2021). NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research, 49*(D1), D165–D171. https://doi.org/10.1093/nar/gkaa1046