Decoding the function of Long Non-Coding RNAs in Subgroup 3 Medulloblastoma: Integrative SNP and eQTL Analysis

Laureando
Nicola Greco

Relatore
Monica Ballarino

SAPIENZA
UNIVERSITÀ DI ROMA

Decoding the function of Long Non-Coding RNAs in Subgroup 3 Medulloblastoma: Integrative SNP and eQTL Analysis

Facoltà di Farmacia e Medicina, Ingegneria dell'informazione, informatica e statistica, Medicina e Odontoiatria, Scienze Matematiche, Fisiche e Naturali
Dipartimento di Medicina Molecolare
Corso di laurea in Bioinformatics

Nicola Greco
Matricola 1929247

Relatore
Monica Ballarino

Correlatore
Alessandro Palma

A.A. 2022-2023

Index

Abstract

This project examines the role of long non-coding RNAs (lncRNAs) in medulloblastoma (MB), a malignant brain tumor primarily affecting children worldwide. The study took advantage of a bioinformatics pipeline to decode the regulatory activity of three specific lncRNAs that have been identified as dependent on the activity of the MYC oncogene in MB. Data from the Genotype-Tissue Expression (GTEx) project and dbSNP database were used to identify expression quantitative trait loci (eQTLs) within the genomic loci of these lncRNAs. Analysis revealed 164 cis-eQTLs associated with lncMB1, establishing links between seven SNPs and eight nearby genes. The variations consistently showed similar associations with multiple genes, particularly RHOT1, a gene involved in mitochondrial transport and associated with Parkinson's disease and various cancers. Furthermore, the prediction of the secondary structure of lncMB1 RNA molecule unraveled possible effects of two genetic variations on its function and stability.

These findings highlight the importance of eQTLs in getting a more comprehensive understanding of the regulatory mechanisms and functional implications of lncRNAs in this devastating disease. Results could contribute to get an improved knowledge of gene regulation in MB and provide new insights and hypotheses for further studies.

1. Introduction

1.1 LncRNAs

RNA molecules have historically been classified into messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Until recently, these were thought to comprise the bulk of the RNA content in cells. In the early 1990s, the Human Genome Project first brought attention to the idea that a substantial portion of the genome is transcribed into RNA molecules that lack protein-coding potential. These RNA molecules are referred to as non-coding RNAs (ncRNAs). Although initially, the majority of these ncRNAs were believed to constitute "junk DNA" with no biological relevance, it has now become increasingly clear that many non-coding RNA molecules play vital roles in cell biology and regulation of gene expression. [1]

Among non-coding RNAs, a significant portion is represented by long ncRNAs (lncRNAs). They are broadly defined as transcribed RNA molecules greater than 500 nucleotides in length and lacking an open reading frame of significant length (less than 100 amino acids). Even though the majority of lncRNAs has not yet been functionally characterized, in literature already exists several examples of lncRNAs that are biologically functional, and the number of examples is rapidly increasing. LncRNAs are involved in various aspects of gene regulation, including chromosome dosage compensation, imprinting, epigenetic regulation, nuclear and cytoplasmic trafficking, transcription, mRNA splicing, and translation. In such a manner, lncRNAs play important roles in the regulation of numerous biological processes, such as proliferation, cell cycle control, apoptosis, differentiation, and maintenance of pluripotency. Due to the flexible nature of RNAs, they also have the ability to interact with diverse cellular components, such as DNA, RNA, and proteins, and are involved in vast intracellular networks.

[1] Cipriano, Andrea, and Monica Ballarino: *The ever-evolving concept of the gene: the use of RNA/protein experimental techniques to understand genome functions*, Frontiers in molecular biosciences 5 (2018).

One of the primary functions of lncRNAs is to regulate gene expression. They exert their actions through a variety of mechanisms, such as acting as transcriptional regulators, recruiting chromatin-modifying enzymes, and altering the chromatin architecture, thereby shaping the epigenetic landscape of the genome. Some lncRNAs interact directly with transcription factors or transcriptional co-regulators, while others recruit chromatin remodeling complexes. For example, Xist and Tsix are lncRNAs that play crucial roles in X-chromosome inactivation in female mammals.

LncRNAs also take part in alternative splicing regulation of mRNAs. Malat1 is an example of an lncRNA that is known to regulate alternative splicing in various tissues. In addition, many lncRNAs have been implicated in miRNA sponge activity, where they compete with mRNAs for binding to miRNAs, thereby regulating mRNA stability. Thus, lncRNAs play important roles in the post-transcriptional regulation of gene expression. [2]

LncRNAs also play critical roles in different pathologies, such as cancer, cardiovascular, and neurodegenerative diseases. They can act as oncogenes or tumor suppressors depending on the context and modulate crucial cellular processes for oncogenesis, such as cell proliferation, apoptosis, and metastasis. For instance, lncRNA HOTAIR is an oncogenic lncRNA that plays a role in breast cancer metastasis.[3]

Overall, it is now clear that lncRNAs are biologically relevant molecules with essential roles in diverse biological processes. However, their functional characterization remains challenging, and many of their mechanisms of action and targets are yet to be decoded. As we continue to unravel the complexities and potential of non-coding RNA molecules, lncRNAs and other ncRNAs represent a promising therapeutic target for multiple diseases and offer potential interventions for gene regulation in the future.

[2] Zhu, JuanJuan, et al.: *Function of lncRNAs and approaches to lncRNA-protein interactions,* Science China Life Sciences 56 (2013).

[3] Gibb, Ewan A., Carolyn J. Brown, and Wan L. Lam: *The functional role of long non-coding RNA in human carcinomas*, Molecular cancer 10.1 (2011).

1.2 GTEx project and eQTLs

Understanding the relationship between genotypes and phenotypes is essential for finding genetic variations underlying complex diseases. By combining information on genetic variants with gene expression data, it has become possible to identify expression quantitative trait loci (eQTLs). These are genomic locations with polymorphic variations that are associated with gene expression levels and are usually mapped by statistical analysis that establish associations between polymorphisms and changes in gene expression levels among a large number of individuals. eQTLs enable the mapping of genomic regions, which can control the transcriptome's behavior, and can be utilized to identify candidate genes responsible for driving phenotypic variations.[4] Several studies have already exploited eQTLs of SNPs located within the locus of lncRNAs to study their potential ability to regulate expression of nearby genes[5].

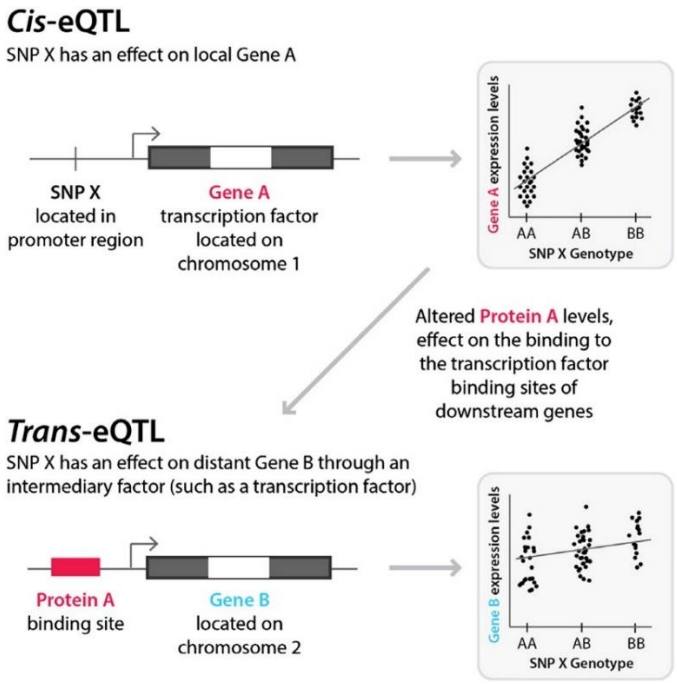


***Figure 1**. From genome to function by studying eQTLs (Harm-Jan Westra, Lude Franke; 2014)*

[4] Westra, Harm-Jan, and Lude Franke: *From genome to function by studying eQTLs,* Biochimica et Biophysica Acta (BBA)-molecular basis of Disease 1842.10 (2014): 1896-1902.

[5] Shiyang Liu, Nathan Harmston, Trudy Lee Glaser: *Wnt-regulated lncRNA discovery enhanced by in vivo identification and CRISPRi functional v alidation*, Genome Medicine, 2020

The Genotype-Tissue Expression (GTEx) project was initiated to create a comprehensive atlas of genetic diversity and gene expression differences across human tissues and organs. In the 8th version of GTEx, it was conducted a comprehensive examination of 15,201 RNA-sequencing samples from 49 different tissues of 838 postmortem donors and by analyzing RNA expression within individual tissues and treating gene expression levels as quantitative traits, they identified variations in gene expression or alternative splicing that are highly correlated with genetic variation.[6] An important distinction in eQTLs is based on the position of the gene to which the loci is associated to: cis-eQTLs are genetic variants that influence gene expression levels in a proximal genomic region, usually within 100 kilobase pairs of the regulated gene; trans-eQTLs, in contrast, are genetic variants that modulate gene expression levels at a remote distance and can be located on the same chromosome or different chromosomes (Figure 1).

1.3 Medulloblastoma

Medulloblastoma is a primary malignant brain tumor that predominantly affects the pediatric population. It is the most common pediatric brain tumor, accounting for approximately 20% of all childhood brain tumors. Categorized as an embryonal neuroepithelial tumor of the cerebellum, it is a high-grade tumor that tends to spread through the cerebrospinal fluid. Despite recent advances in treatment and diagnosis, the prognosis for patients with medulloblastoma remains poor, with a 5-year survival rate ranging from 50-70%. The rapidly proliferating nature of the tumor frequently results in increased intracranial pressure, causing various symptoms such as headaches, vomiting, and cranial nerve palsies.

Recent advances in molecular biology have led to improved genetic characterization of medulloblastoma, enabling the classification of the tumor into four molecular subgroups, each with unique features, including driver genes and clinical outcomes. Group 3 (G3) is the

[6] GTEx Consortium: *The GTEx Consortium atlas of genetic regulatory effects across human tissues*, Science 369.6509 (2020).

most aggressive subgroup and is associated with the worst survival outcome at five years and the highest rate of metastasis at diagnosis. Although a common G3 driver pathway has not yet been identified, a unique c-MYC signature has been shown to be specific to this subgroup and occurs in approximately 17% of G3 patients. MYC, plays a pivotal role in coordinating the gene expression patterns required for the proliferation, expansion, and maintenance of somatic and stem cells, thereby ensuring cellular growth and homeostasis.[7] Nevertheless, recent studies have shown that small peptides that inhibit c-MYC expression can successfully improve the outcomes for G3 patients. This features was used by the authors of a the article "Identification and Functional Characterization of Novel MYC-Regulated Long Noncoding RNAs in Group 3 Medulloblastoma" to underscore lncRNAs that are under the regulation of MYC in cell lines derived from MYC-driven primary MB patient samples. Among the identified MYC-dependent lncRNAs, three non-coding transcripts – that were renamed lncMB1, lncMB2 and lncMB3 - were found to affect G3 MB cell survival after knockdown, and thus selected for further analysis.[8] Here I will use a Bioinformatics pipeline to decode the function and mechanism of action of this three lncRNAs, using already known polymorphisms that colocalize with them and eQTLs to identify possible target genes that can mediate their oncogenic activity.

[7] Northcott, Paul A., et al.: *Medulloblastoma*, Nature reviews Disease primers 5.1 (2019).
[8] Rea, Jessica, et al.: *Identification and functional characterization of novel MYC-regulated long noncoding RNAs in group 3 medulloblastoma,* Cancers 13.15 (2021).

2. Methods

2.0 Software

Everything I have done in this thesis was performed on R programming software. All the code and scripts can be found on my GitHub repository: https://github.com/Nicogreco2001. Here all the information of my system:

```
─ Session info ─────────────────────────────────────────
 setting  value
 version  R version 4.3.0 (2023-04-21 ucrt)
 os       Windows 10 x64 (build 19045)
 system   x86_64, mingw32
 ui       RStudio
```

Now I will introduce all the R packages that I have used during the analysis. They are Bioconductor packages and were mainly used for genetic data handling and visualization:

- The rtracklayer library in Bioconductor provides a versatile framework for interacting with genome browsers, including UCSC. It allows manipulation of annotation tracks in various formats and offers export/import capabilities.

  Version used: 1.60.0
- The Gviz package facilitates the visualization of genomic data, integrating with external sources like Ensembl and UCSC. It supports multiple annotation file types and offers customization options while simplifying the creation of publication-ready figures.

  Version used: 1.44.0
- The GenomicRanges package is essential for efficiently representing and manipulating genomic annotations and alignments in high-throughput sequencing data analysis. It provides specialized data structures and algorithms for tasks such as gene model manipulation, read alignment representation, and overlap detection.

  Version used: 1.52.0
- The biomaRt package in Bioconductor seamlessly integrates BioMart data resources with data analysis software, offering comprehensive gene annotation and retrieval capabilities, including genomic sequences and single nucleotide polymorphism information.

Version used: 2.56.0

- The GenomicInteractions package in Bioconductor specializes in manipulating and exploring chromatin interaction data. It integrates with other Bioconductor packages, provides plotting functions and summary statistics, and enhances the visualization and analysis of chromatin interaction data.
  Version used: 1.34.0

2.1 dbSNPs: Datasets

Given the genetic coordinates of the three MYC-dependent lnRNAs:

- lncMB1-ENSG00000214708: chr17:32141226-32143135;
- lncMB2-ENSG00000253123: chr8:37326606-37331991;
- lncMB3-ENSG00000278484: chr10:120984966-120985596,

The analysis started by downloading from dbSNP a dataset containing information about known polymorphisms within a ± 1MB interval from the genetic coordinates of three lnRNAs. These datasets were obtained as part of the analysis to investigate the genetic variation in proximity to these lnRNAs. dbSNP, developed by the National Center for Biotechnology Information (NCBI) and the National Human Genome Research Institute (NHGRI), is a widely used public database that houses a comprehensive collection of genetic variations. It includes various types of polymorphisms, such as single nucleotide polymorphisms (SNPs), indels, microsatellite markers, and more.

Each entry in the dataset contains the following information:

- *chr*: Chromosome number where the variation is located.
- *pos*: Position of the variation on the chromosome.
- *variation*: The specific variation observed (e.g., nucleotide change).
- *variant_type*: Type of variant (e.g., single nucleotide variant - snv).
- *snp_id*: Identifier for the SNP in dbSNP.
- *clinical_significance*: Clinical significance of the variation (e.g., benign).
- *validation_status*: Validation status of the variation.

- *function_class*: Functional classification of the variant (e.g., coding sequence variant, intron variant).
- *gene*: Gene associated with the variation.
- *frequency*: Frequency of the variant in different populations from several databases and experiments.

2.2 dbSNPs: Processing and Analysis

Despite the curated nature of the dbSNP database, the acquired datasets were disordered, necessitating data processing. This step aimed to organize and improve the dataset's reliability for exploring genetic variations associated with the lnRNAs. During the dataset processing, I extracted the frequency of each polymorphism coming from the gnomAD database, which is a trusted and widely utilized resource for human genetic variations. I specifically discarded frequencies derived from other experiments. Additionally, I eliminated redundant classifications within the *function_class* variable. These steps ensured a refined dataset for further analysis (Figure 2).

chr	pos	variation	variant_type	snp_id	clinical_significance	gene	GnomAD_freq	function_class
17	32997763	T>C	snv	28956	benign	SPACA3	C:0.025946	coding_sequence_variant
17	31352228	G>A,C	snv	964288	benign	NF1	NA	intron_variant
17	31203332	A>C	snv	1013946	benign	NF1	A:0.427744	intron_variant
17	31203344	C>T	snv	1013947	benign	NF1	C:0.435074	intron_variant
17	32487830	C>T	snv	1042845	benign	CDK5R1	T:0.002631	coding_sequence_variant
17	31376984	T>C	snv	1048317	benign	NF1	T:0.341545	exon_variant

***Figure 2**. Head of the refined final dataset of SNPs used for the downstream analyses*

Utilizing the processed and refined dataset, statistical analyses were performed, focusing on three key aspects:

1. The number of variations associated with each gene was determined.
2. The clinical significance and functional class of these variations for each gene were thoroughly examined.
3. The distribution of frequencies among the variations was investigated, excluding polymorphisms lacking an associated gnomAD frequency.

These statistical analyses provided valuable insights into the genetic landscape and shed light on important aspects of variation patterns.

2.3 eQTLs: Datasets

Subsequently, my focus shifted towards exploring expression quantitative trait loci (eQTLs). To investigate these associations, I obtained the relevant datasets from the GTEx project website (https://gtexportal.org/home/). These datasets encompassed variant-gene pairs that exhibited significant correlations (padj > 0.05) in European-American donors across 49 different tissues analyzed in the 8th release of the project. These eQTLs datasets provided valuable information regarding pairs of genetic variations, predominantly SNPs, and genes that demonstrated a high level of correlation within the tissue samples of the 838 postmortem donors. This correlation indicated a statistically significant change in gene expression between individuals carrying the genetic variation and those without it. GTEx project's cis-eQTL mapping employed FastQTL, a method using linear regressions to identify optimal associations between genotypes and molecular phenotypes, using several covariates factors, such as sex, to prevent spurious associations.

The dataset containing significant associations between genetic variants and gene expression levels consists of the following columns:

- *phenotype_id*: GENCODE/Ensembl gene ID associated with the expression.
- *variant_id*: Unique identifier for the genetic variant in a specific format.
- *tss_distance*: Distance between the variant and the transcription start site (TSS) of the gene, where positive values indicate downstream positions and negative values indicate upstream positions.
- *maf*: Minor allele frequency, representing the frequency of the minor allele observed in the set of donors for a given tissue.
- *ma_samples*: Number of samples in which the minor allele is present.
- *ma_count:* Total count of minor alleles across individuals.
- *pval_nominal*: Nominal p-value indicating the significance of the association between the variant and gene expression.

- *slope*: Regression slope, representing the effect size of the variant on gene expression.
- *slope_se*: Standard error of the regression slope.
- *pval_nominal_threshold*: Nominal p-value threshold for calling a variant-gene pair significant for the gene.
- *min_pval_nominal*: Smallest nominal p-value observed for the gene among all tested variant-gene pairs.
- *pval_beta*: Beta-approximated permutation p-value for the gene.

Each row in the dataset represents a significant variant-gene pair, providing information on the genetic variant, its proximity to the gene's transcription start site, the frequency and count of the minor allele, the statistical significance of the association, and other relevant attributes (Figure 3).

	phenotype_id	variant_id	tss_distance	maf	ma_samples	pval_nominal	slope	slope_se	pval_beta
289674	ENSG00000263006.6	chr18_98761_T_TAACCCAAACCCG_b38	-10304	0.3361340	64	1.80e-06	-0.515339	0.1010470	0.0000001
166544	ENSG00000272983.1	chr10_39006780_C_A_b38	869443	0.0546219	13	0.00e+00	-1.130650	0.1867400	0.0000118
128185	ENSG00000182722.5	chr7_65103710_T_A_b38	250356	0.3781510	77	1.22e-05	-0.474795	0.1027930	0.0000001
17533	ENSG00000230325.1	chr1_236583143_C_T_b38	42697	0.3235290	66	6.00e-07	0.683174	0.1276590	0.0000954
42474	ENSG00000004534.14	chr3_49870590_G_A_b38	-69417	0.4201680	84	3.66e-05	-0.280422	0.0646914	0.0007203
230622	ENSG00000011638.10	chr16_21145692_A_G_b38	-12685	0.3907560	75	3.90e-06	0.605354	0.1233910	0.0034386

***Figure 3**. Head of the table related to the eQTL dataset*

This dataset enables the exploration of cis-eQTLs and their impact on gene expression levels, facilitating the study of lncRNAs and identification of possible genes target of their regulatory activity with the hypothesis that an eQTL SNP within or around of a lncRNA might influence its physiological regulatory activity, by modify its interaction with transcription factors or epigenetic modifiers, thereby altering the expression of nearby protein coding genes.

2.4 eQTLs: Search

In order to identify variation-gene pairs within the lncMBs loci, I conducted a search within the downloaded GTEx dataset. This involved examining the eQTLs dataset and identifying

rows that corresponded to polymorphisms found in the dbSNPs dataset. By matching these variations, I was able to pinpoint the relevant variation-gene pairs associated with the lncMBs loci. Upon analyzing the GTEx dataset within the lncMBs loci, the following results were obtained: for lncMB1, a total of 164 cis-eQTLs were identified, associating 7 SNPs with 8 genes; in the case of lncMB2, 60 cis-eQTLs were discovered, linking 9 SNPs to 1 gene; lastly, lncMB3 exhibited 15 cis-eQTLs, connecting 10 SNPs with 3 genes.

2.5 eQTLs: Analysis

In order to gain a better understanding of the results, an analysis was conducted on the identified eQTLs within the lncMB RNAs loci. The analysis focused on key aspects, including the genes involved in the variant-gene associations, the number of associations for each SNP, the genomic position of the polymorphisms, and the magnitude of the change in gene expression associated with each eQTL. This analysis provided valuable insights into the regulatory mechanisms and functional consequences within the lncMB RNAs loci. Here for the genomic visualization of the data I used the R package *biomaRt* to recover from the ENSEMBL server information on the eQTLs genes - such as coordinates, transcript lengths and strand sign– while *Gviz*, *GenomicRanges* and *GenomicInteractions* for genomic data handling and visualization.

2.6 lncMB1 structure predictions

During the final phase of this study, a specific focus was placed on lncMB1, as its eQTLs exhibited significant interest. The examination delved into the genetic variants involved in the eQTLs gene-variation pairs and their potential influence on the lncRNA. Initially, relevant information about the lnMB1 transcript, including precise details such as the transcription start site (TSS), exon and intron coordinates, and the nucleotide sequence, was obtained from Ensembl. Subsequently, SNPs located on the lncRNA exons were selected. To explore their effects, the RNAfold structure prediction algorithm was utilized. RNAfold is one of the tools provided by the Vienna RNA Websuite[9], which offers a range of tools for folding,

[9] Lorenz, Ronny, et al.: *ViennaRNA Package 2.0*, Algorithms for molecular biology 6 (2011).

designing, and analyzing RNA sequences. This algorithm employs a partition function to calculate base-pair probabilities in addition to the minimum free energy (MFE) structure and enables the folding of RNAsThe predictions were performed for both the wild-type transcripts (without any mutations) and the transcripts with identified point mutations. This facilitated a comparison to evaluate the potential impact of the SNPs on the molecular structure of the lncRNA.

3. Results

3.1 dbSNP Analysis

In this section, I present the results of the processing and analysis of the dbSNPs dataset:

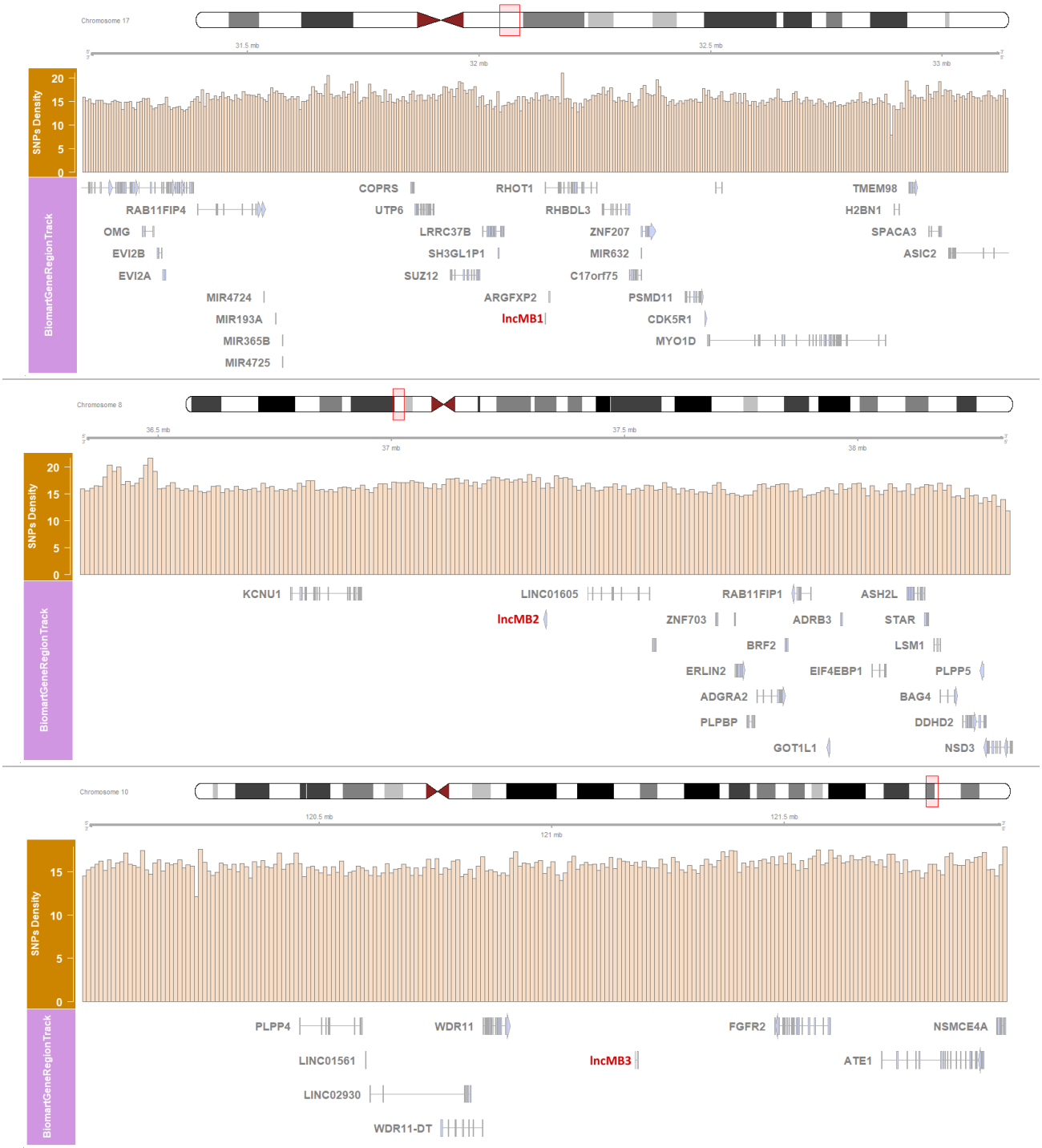


Figure 4.1 *SNPs distribution across the genomic loci of lncMB1 (upper panel), lncMB2 (mid panel) a nd lncMB3 (lower panel).*

Charts in figure 4.1 depict the distribution of polymorphisms across the genetic loci of the lncMBs. The upper track of the graph presents yellow bars representing SNP density, which is calculated as the percentage of known polymorphisms within a window size of approximately 8000 base pairs. On average, the SNP density is 15, with peaks observed in regions that are relatively lacking in protein coding genes.

Figure 4.2 *SNPs frequency distribution across the genomic loci of lncMB1 (upper panel), lncMB2 (mid panel) and lncMB3 (lower panel).*

Similarly, to the previous charts, the Figure 4.2 illustrate a metric that calculates the density of SNPs, taking into account the frequency of each polymorphism. This approach assigns higher weight to SNPs that are more prevalent in the population, resulting in taller bars indicating a higher density of SNPs with higher frequencies. Once again, it is worth noting that the taller bars predominantly appear in regions with limited coding sequence content.

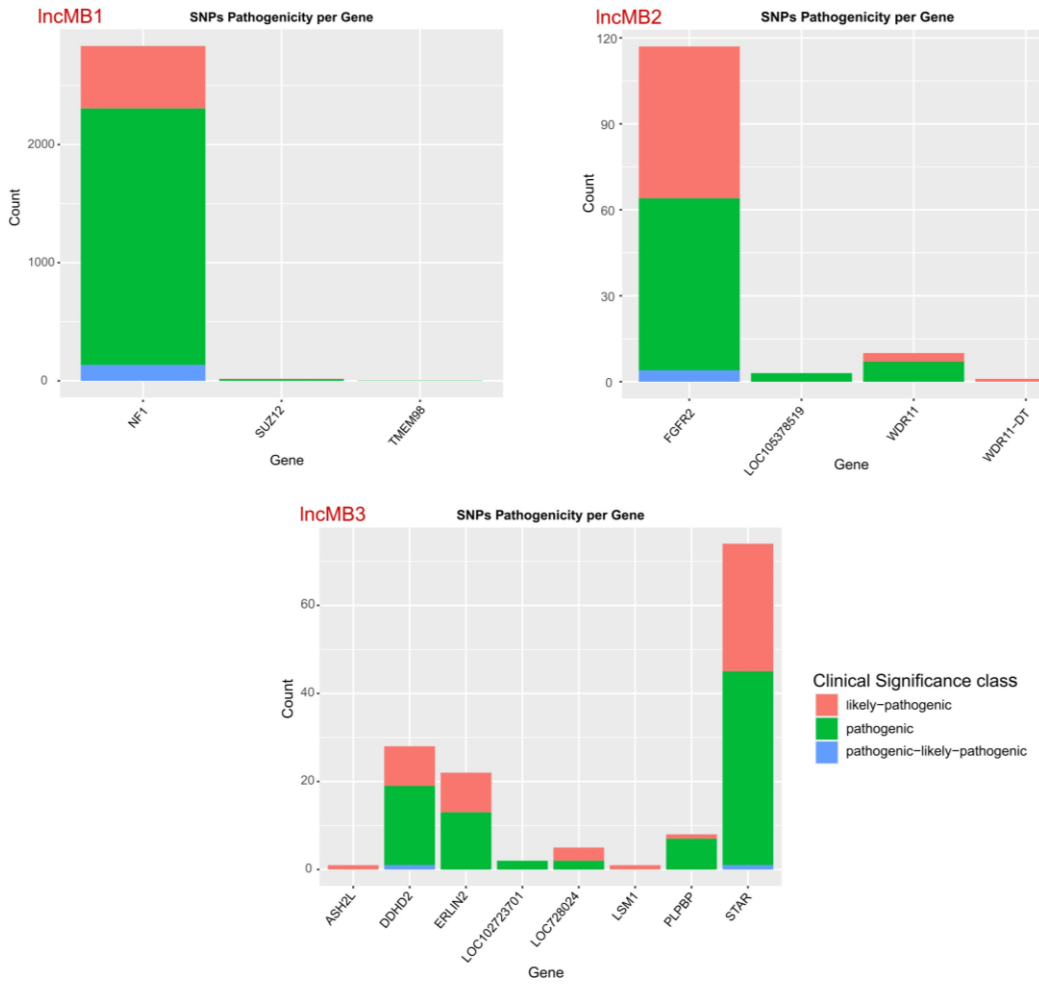


***Figure 5**. SNPs density across the genomic loci of lncMB1, lncMB2 and lncMB3.*

The barplots represent the findings of the clinical significance analysis conducted on polymorphisms classified as "pathogenic" within the dbSNPs datasets for the three MB long non-coding RNAs. The bars in green indicate the presence of known pathogenic polymorphisms, while the ones in orange and blue represent likely pathogenic variants. Notably, it is worth mentioning that all the SNPs associated with known health issues are primarily found in protein coding genes, with none of them being identified in the three MYC-dependent lncRNAs examined.

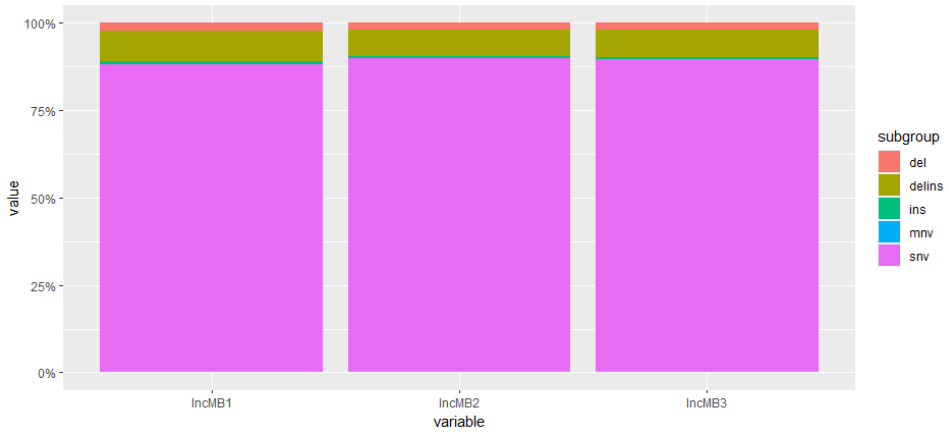
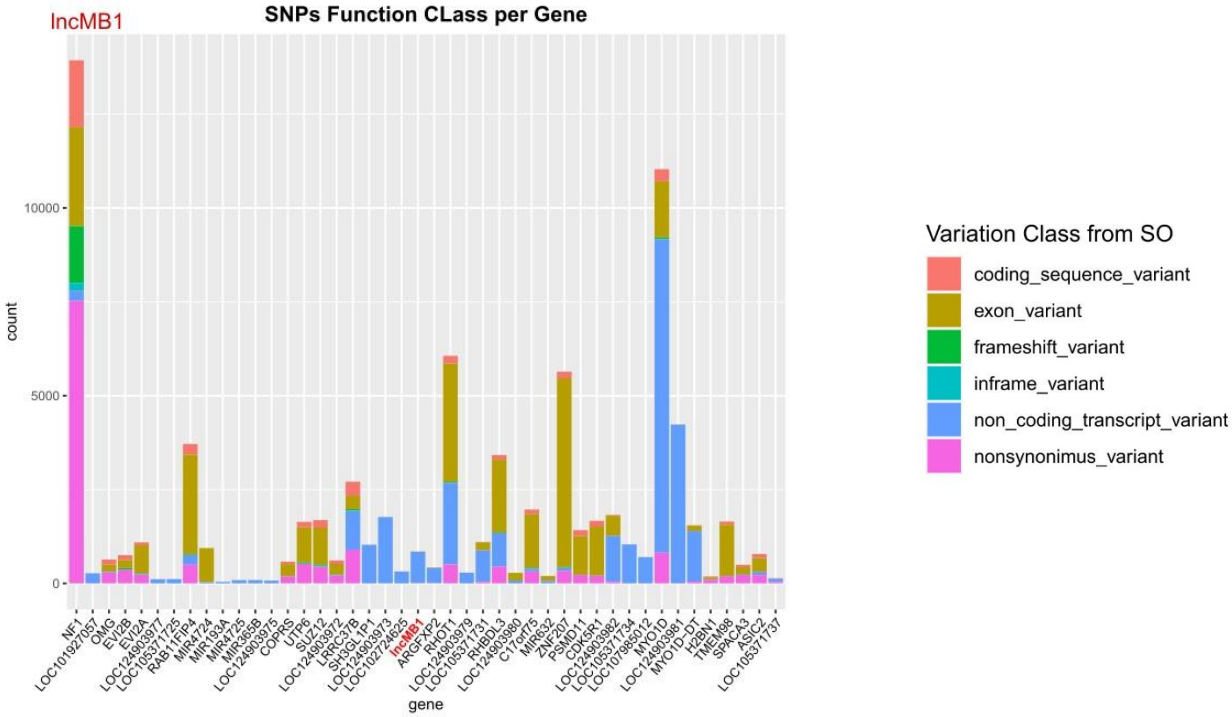


***Figure 6**. Distribution of genetic variations in the three lncRNAs.*

Displayed in the bar charts of Figure 6 are the distributions of various types of genetic mutations observed across the three datasets. Remarkably, the predominant category in all three loci is single nucleotide variation (SNV), accounting for approximately 90% of the mutations. The remaining 10% comprises a combination of insertions, deletions, and indels mutations. Notably, multiple nucleotide variations (MNV) are rare, representing a negligible portion of the overall mutations across the datasets.



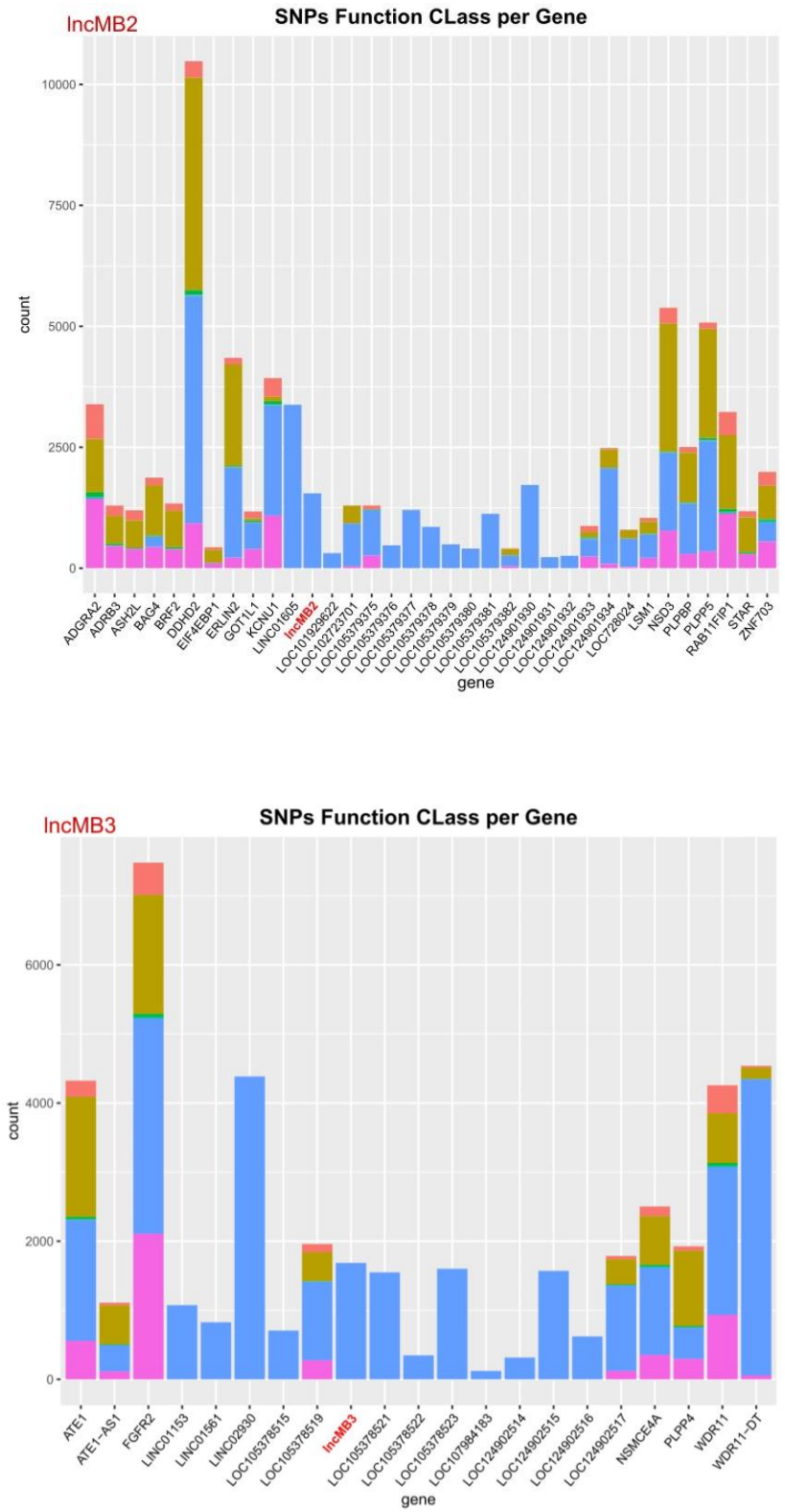


Figure 7. Barplots showing SNPs counts for each gene according to the variant class (shown in the legend).

Plots in figure 7, each corresponding to a lncMB, illustrate the count of SNPs categorized based on Sequence Ontology. The classes include coding sequence variant, exon variant,

frameshift variant, inframe variant, non-coding transcript, and non-synonymous variant. As expected, all lncMBs, and generally all lncRNAs, exclusively exhibit mutations annotated as non-coding transcript variants.

3.2 eQTLs Analysis

For what concerns the eQTLs analysis, here are exposed the results.

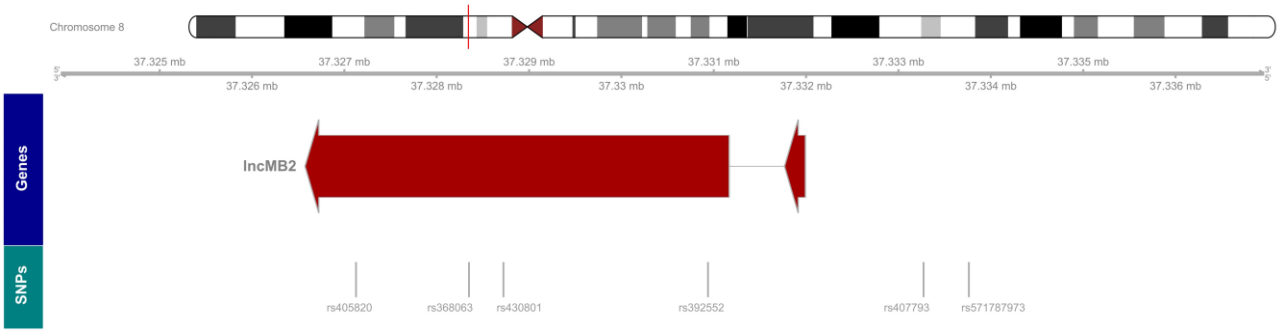


***Figure 8**. Chromosomal alignment of lncMB2 showing the position of analyzed SNPs.*

Multiple variant-gene associations were discovered within the genomic locus of lncMB2 (figure 8). Among these associations, six single nucleotide polymorphisms (SNPs) were identified, with four of them located within the second exon of the lncRNA transcript, while the remaining two were positioned upstream of the gene. Subsequent analysis revealed that all six SNPs, which were involved in a total of 60 cis-eQTLs, exhibited a significant correlation only with changes in the expression levels of the lncRNA itself (Figure 9).

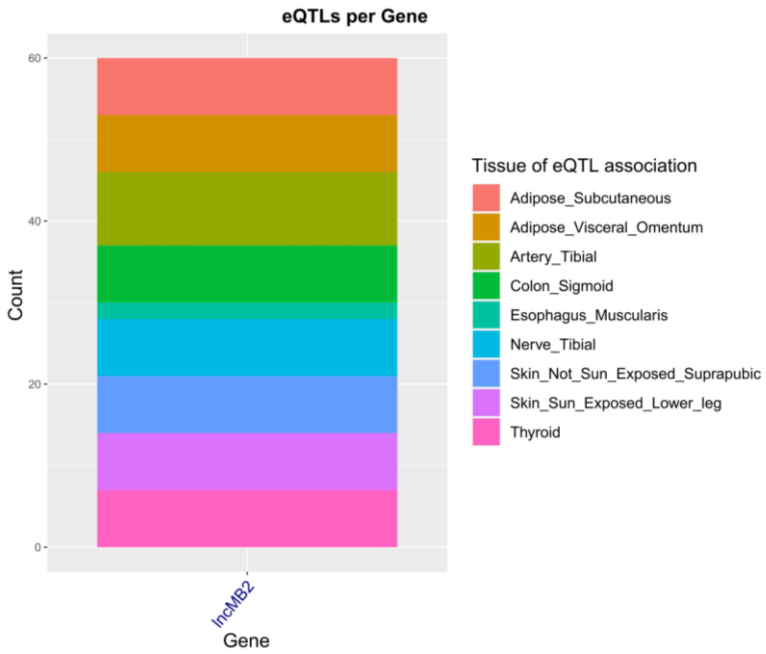


***Figure 9**. Chart showing the number of eQTL associations per gene for lncMB2, colored by tissue.*

The graph presented in figure 9 illustrates the variation-gene associations, organized according to genes, specifically focusing on the lncMB2 gene. Each association is color-coded based on the tissue in which it was identified. Notably, no mutations within the genomic locus of the lncRNA were found to be associated with changes in the expression levels of nearby protein coding genes. This lack of evidence suggests no apparent indications of a regulatory role of the lncRNA towards other genes.

In the case of lncMB3 (Figure 10), a total of 15 expression quantitative trait loci (eQTLs) were identified within the results of the GTEx project. These eQTLs established associations between 10 polymorphisms located within the genetic locus of the lncRNA and three distinct genes.

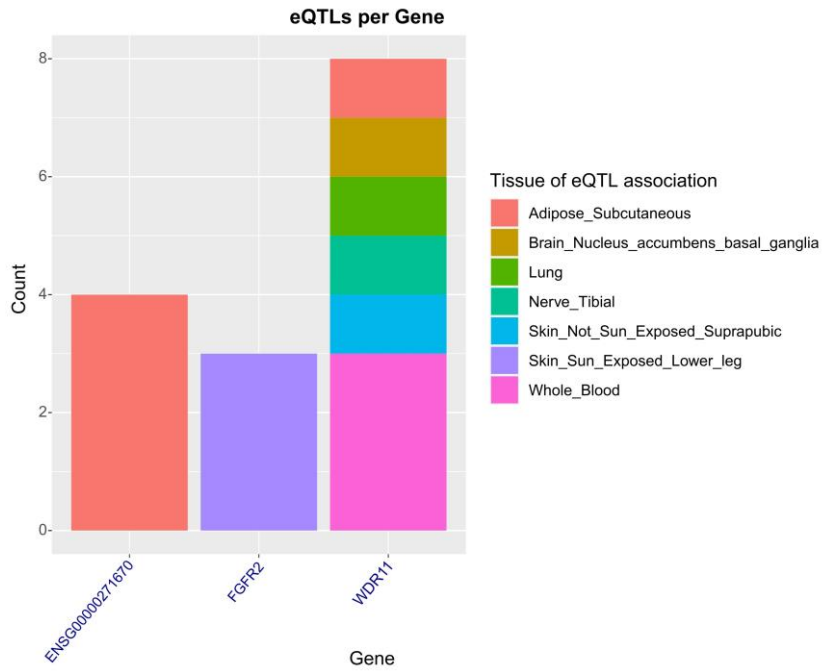


***Figure 10**. Chart showing the number of eQTL associations per gene for lncMB3, colored by tissue.*

The three genes involved in these associations are FGFR2, WDR11, and ENSG00000271670, the latter being a long non-coding transcript situated in proximity to lncMB3. In the chart provided in figure 10, the y-axis represents the number of eQTL associations between SNPs and each gene, with the associations color-coded based on the tissue in which they were observed. On the x-axis, the genes associated with the eQTLs are listed. Notably, among these three genes, WDR11 exhibited changes in expression levels in response to variations in six different tissues.

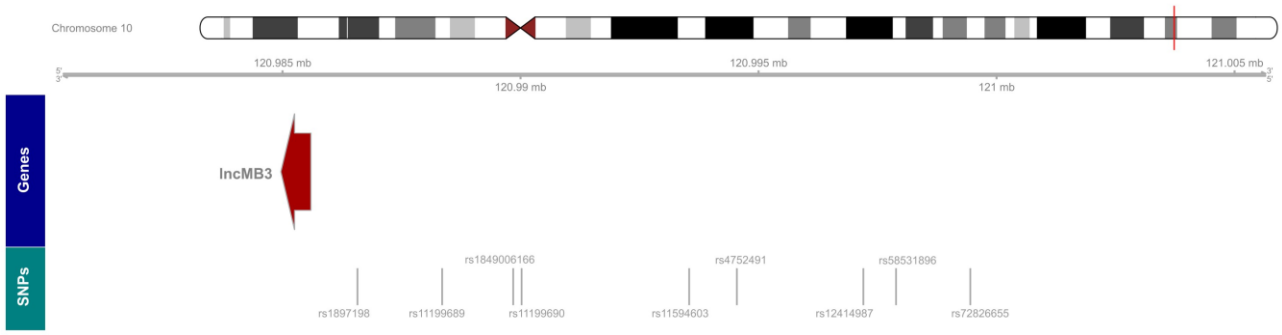


***Figure 11**. Chromosomal alignment of lncMB3 showing the position of analyzed SNPs.*

Upon examining the genomic location of the SNPs, it becomes apparent that all of the SNPs involved in the GTEx project's eQTL associations, which were annotated for lncMB3 in the

dbSNP database (designated as ENSG00000278484 in its Ensemble), were found to be situated outside the transcripts (figure 11).

For lncMB1, a total of 164 cis-eQTLs were identified, linking 7 SNPs to 8 genes (Figure 12).

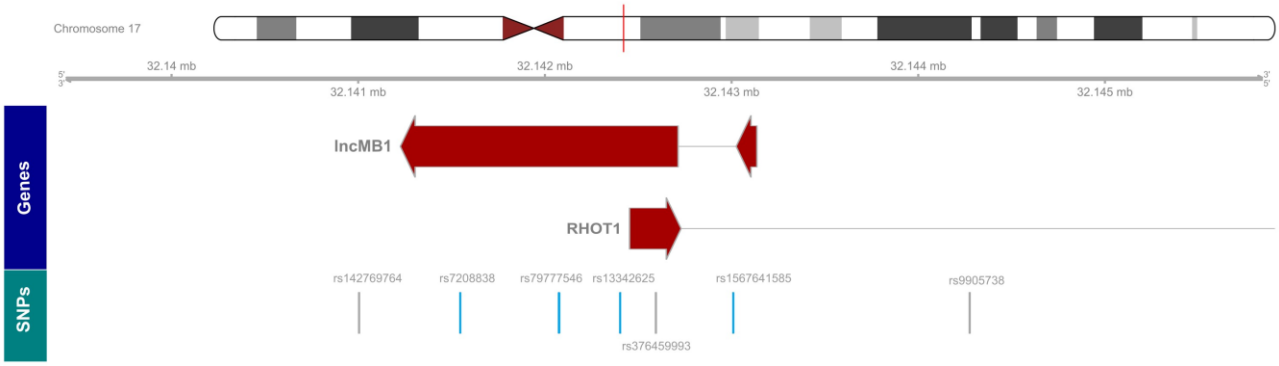


***Figure 12**. Chromosomal alignment of lncMB1 showing the position of analyzed SNPs.*

As visible from figure 12, lncMB1 is as an antisense transcript to the RHOT1 gene, which plays a crucial role in mitochondrial transport and is known to be involved in Parkinson's disease and various cancers. Among the eQTLs-linked genetic variations, four are situated within the larger second exon of the lncMB1 transcript, one is located within the intron near the splicing site of the first exon, approximately 10 bps away, and two are positioned outside the body of the lncRNA. Consequently, the subsequent analysis focused solely on the polymorphisms found within the spliced transcript and excluded those associated with the RHOT gene, highlighted in light-blue in Figure 12.

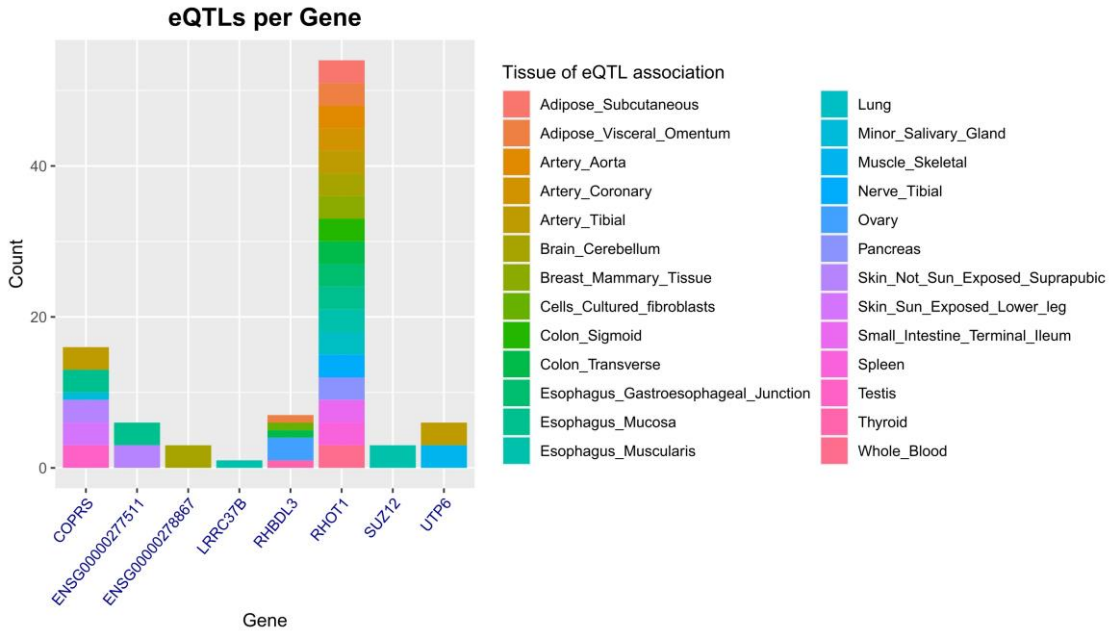


***Figure 13**. Chart showing the number of eQTL associations per gene for lncMB1, colored by tissue.*

In the provided chart in figure 13, the number of eQTL associations per gene is displayed, with each gene-variant association colored according to the tissue in which the GTEx project observed the association. Among the seven genes that exhibited statistically significant changes in expression in presence of genetic variations within the lncMB1 locus, RHOT1 displayed the highest number of associations between its expression change and the presence of one of the four variations in the lncMB1 transcript.

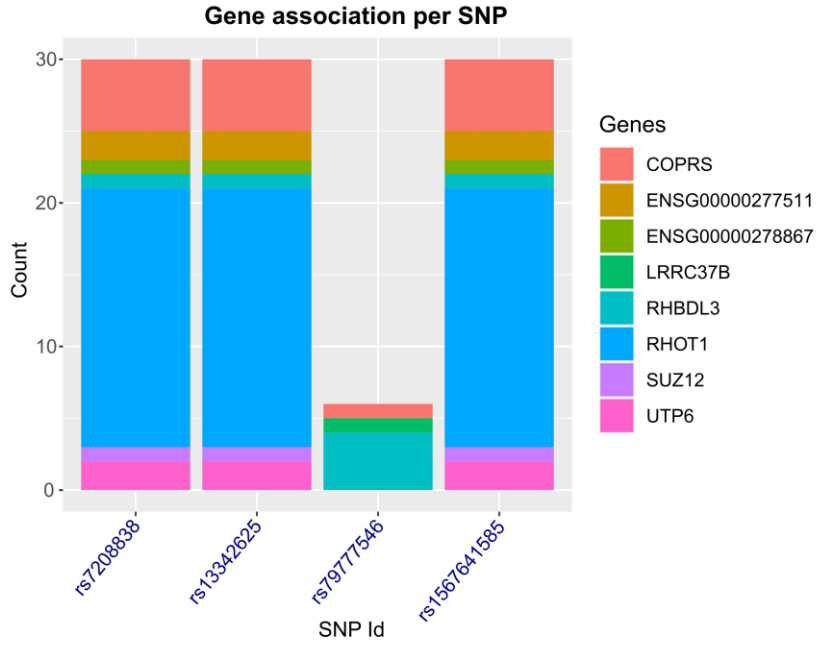


***Figure 14**. Barplot showing the number of eQTLs association per SNP, colored by gene.*

By examining the eQTLs from a variation-centric perspective, it becomes even more apparent that RHOT1 is the gene most closely associated with variations in lncMB1, with 17 associations observed in different tissues for each of the three out of the four lncMB1 transcript variations. Interestingly, these three SNPs (rs7208838, rs13342625, and rs1567641585) share an identical pattern of eQTL associations (Figure 14). Each of these SNPs is associated with a change in expression of RHOT1 in 17 different tissues analyzed by the GTEx project. The same holds true for COPRS, where all three SNPs are associated with a statistically significant change in its expression in five tissues. Similar patterns are also observed for the other genes, with all three variations associated with the same number of tissues.

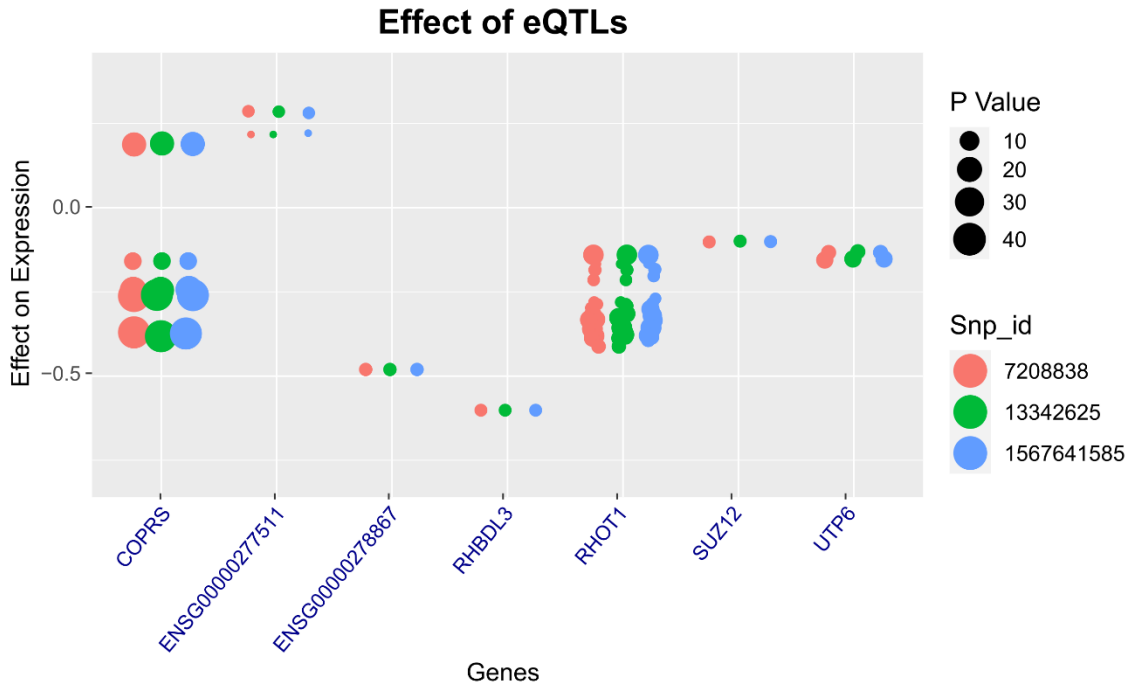


Figure 15. Dotplot showing the effect of the SNPs on the expression levels of the different genes on the x axis.

The plot in figure 15 presented here illustrates the effect of the presence of SNPs (specifically, the three SNPs that share a similar pattern of gene-variation associations) on gene expression. On the y-axis there is the slope of the fitted linear regression model, which quantify the observed change in gene expression, while the color and size of the points correspond to the associated SNP and its statistical significance, with larger points indicating

more statistically significant results (size is proportional to $-log(pval\_beta)$, where *pval_beta* is the adjusted p-value of the linear regression model). Notably, the three variants not only exhibit associations with changes in expression of the same genes in a consistent pattern but also show similar magnitudes of expression change. For example, the change in expression of UTP6, which is registered in two tissues for each variation, demonstrates a magnitude of approximately -0.15 for each association.

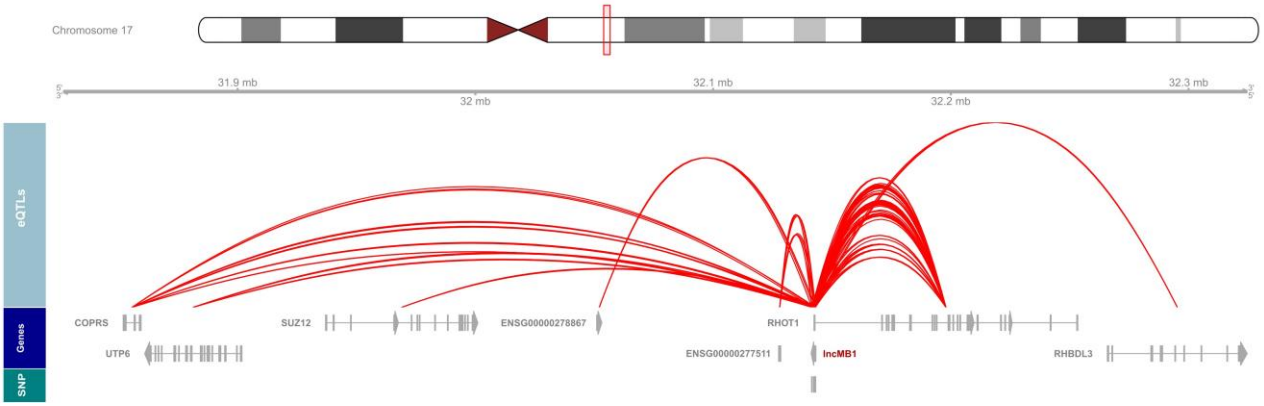


***Figure 16**. Chromosomal alignment showing: lncMB1 and Genes involved in eQTLs in the mid track; rs7208838, rs13342625, and rs1567641585 polymorphisms in the lower track; eQTLs association in the upper track plotted as red arcs, connecting the SNP with the target gene. Higher the arc, stronger the absolute value of the expression change.*

This means that individuals carrying any of the three variations tend to exhibit a nearly identical decrease in UTP6 expression across all three SNPs. The same pattern applies to all seven genes, where the presence of the three variations is associated with similar magnitudes of expression change in multiple tissues.

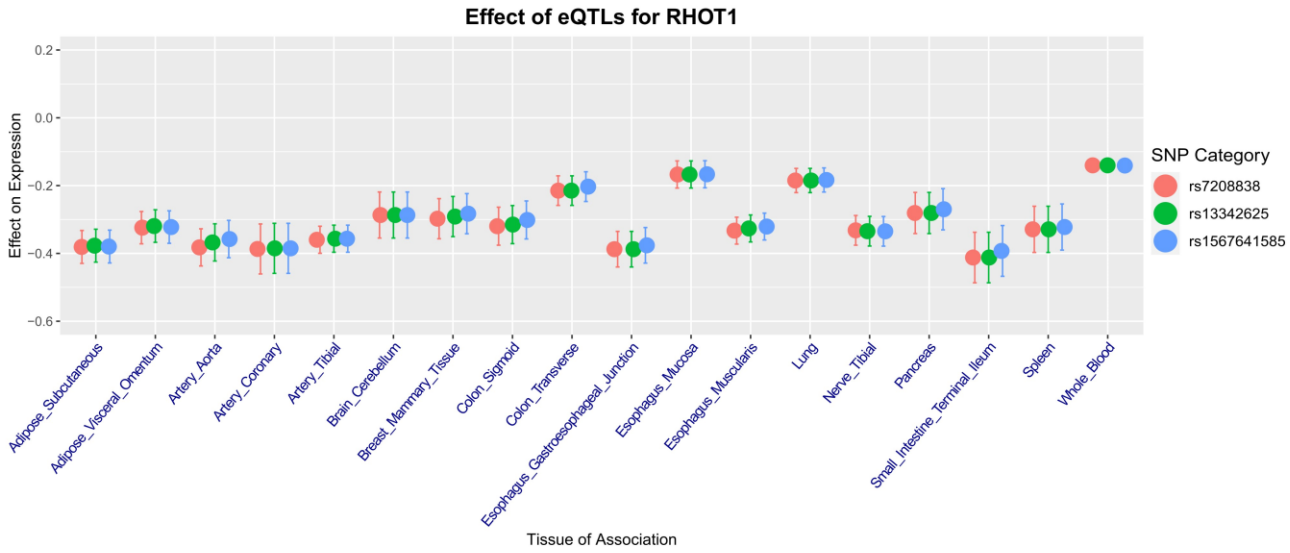


***Figure 17**. Dotplot showing the effect of eQTLs of RHOT1 gene in different tissues.*

Chart in figure 17 depicts the magnitude of expression changes specifically for the eQTLs involving RHOT1, while the tissue types associated with each eQTL are shown on the x-axis. The chart provides clear evidence that not only are the variations consistently linked to the same genes, with equal numbers of associations and similar magnitudes, but the associations also occur in the same tissues for all three variations. Furthermore, it is noteworthy that the observed changes in expression for the variations within the same tissue are nearly identical.

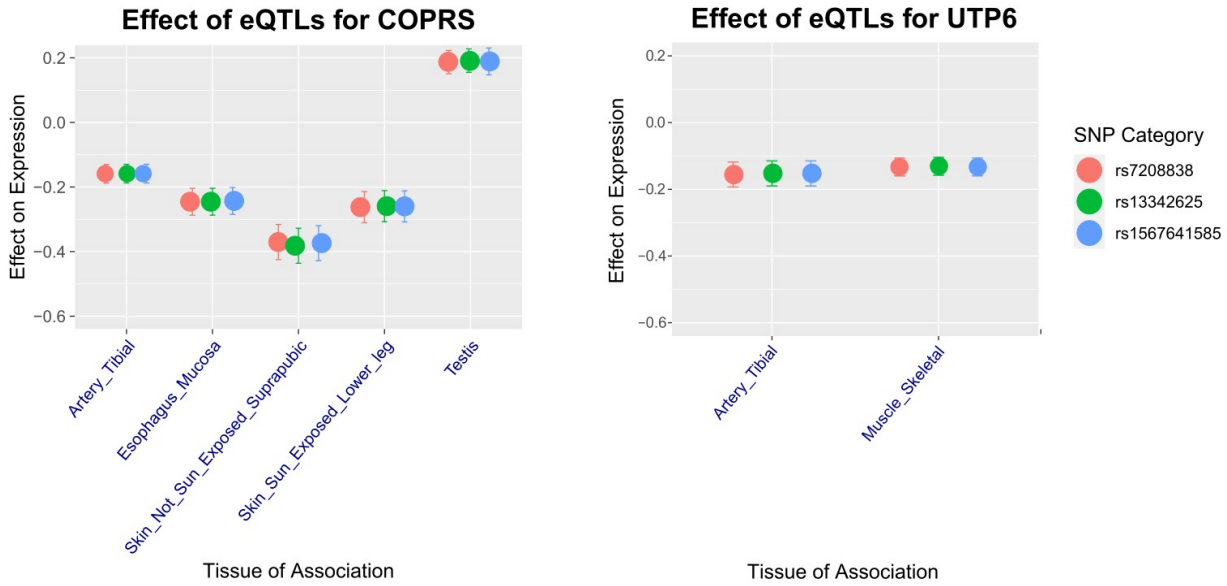


***Figure 18**. Dotplot showing the effect of eQTLs of COPRS 8left) and UTP6 (right) genes in the reported tissues.*

The same pattern holds true for the COPRS and UTP6 genes, as demonstrated in the graphs, as well as for the remaining four. Hence, individuals harboring any of these three polymorphisms within the lncMB1 transcript, as indicated by the GTEx project findings, exhibit comparable alterations in the expression of multiple genes. Specifically, these changes are consistent across multiple aspects, including the genes affected by the expression changes, the specific tissues where these changes are observed, and the magnitude of the expression alterations.

3.3 Structural Analysis

The following images in figure 19 depict the results of the RNAFold software's structure prediction for the lncMB1 RNA molecule. The folded RNA exhibits several secondary structures, including hairpins and stem-loops. Nucleotides within these structures are color-coded based on their base-pair probability, with darker colors indicating more probable structures. Zoomed-in rectangles highlight the RNA domains containing the nucleotides affected by the investigated variations, namely rs7208838 and rs13342625. However, rs1567641585 is not included in the analysis as it is located within an intron.

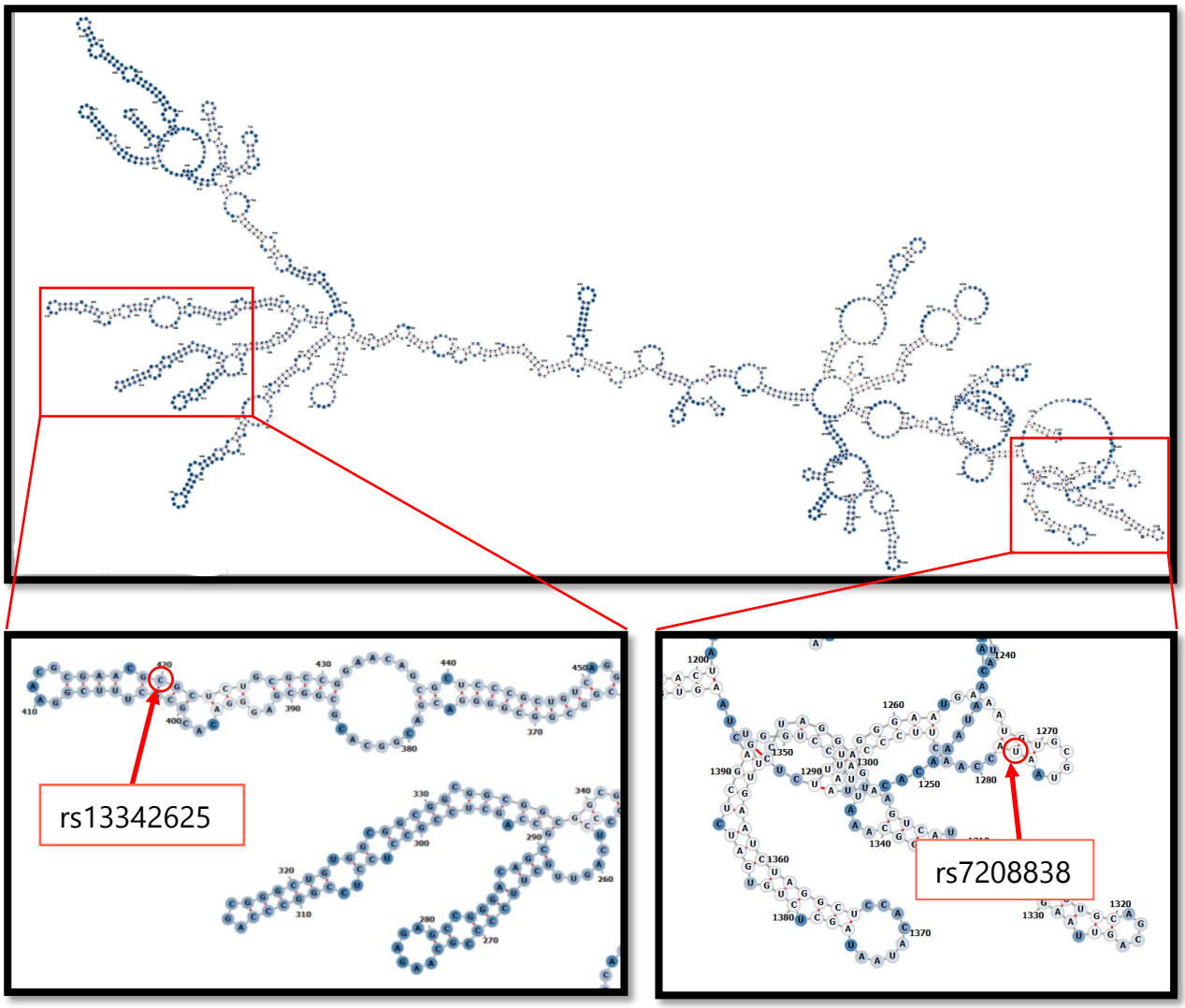


Figure 19. Output of RNAfold secondary structure prediction.

Subsequently, I conducted secondary structure predictions considering the rs7208838 and rs13342625 mutations in the transcript. The objective was to investigate potential effects of these mutations on the RNA structure, which could potentially account for the observed expression changes in nearby genes as measured by eQTLs.

The rs13342625 mutation involves a transversion from cytosine (C) to adenine (A) at position chr17:32142404, corresponding to nucleotide 420 in the RNA transcript. Since the lncRNA is on the negative strand, this mutation results in the substitution of a cytosine at position 420 with an adenine (Figure 20).

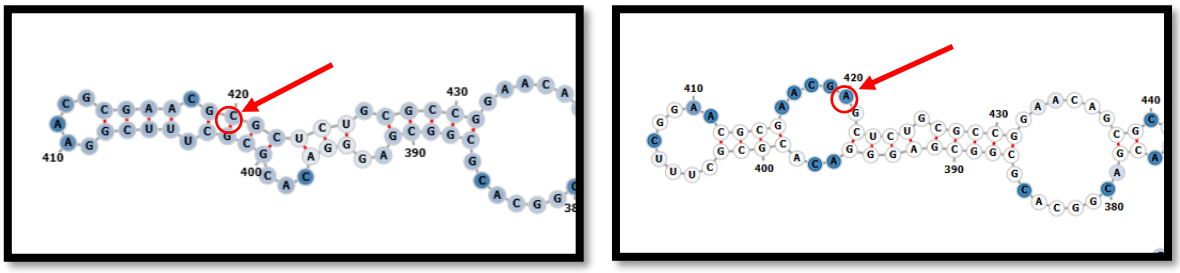


Figure 20. Effect of rs13342625 mutation on RNAfold secondary structure prediction.

According to the RNAFold results, this mutation impacts the local secondary structure by disrupting a highly energetically favorable hairpin, leading to the formation of a loop.

Similarly, the rs7208838 mutation is a transversion from thymine (T) to cytosine (C) at the genomic position chr17:32141547. Consequently, this mutation alters the spliced transcript by replacing a uracil at position 1277 with a cytosine (Figure 21).

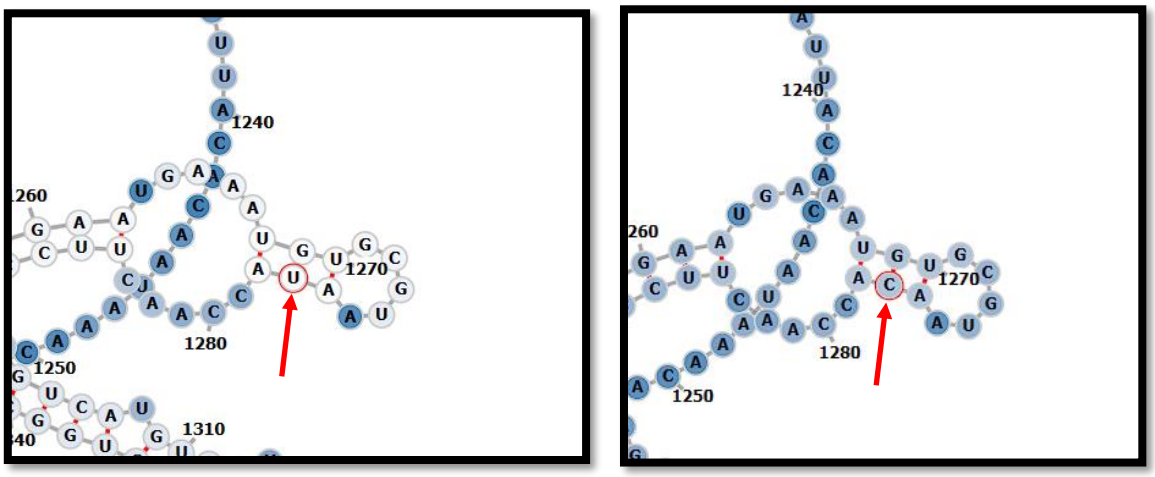


Figure 21. Effect of rs7208838 mutation on RNAfold secondary structure prediction.

As indicated by the prediction, this mutation appears to affect the structure of lncMB1, causing the original wobble base pair, guanine-uracil (G-U), to be replaced by a canonical Watson-Crick cytosine-guanine (C-G) base pair. This structural modification significantly influences the energetic stability of the structure, as evidenced by the darker coloration of the hairpin, thereby increasing the overall stability of the structure.

These findings shed light on the potential impact of the investigated mutations on the secondary structure of lncMB1 and provide insights into their potential role in mediating expression changes in nearby genes.

4. Discussion and conclusions

The aim of this research was to decode the functional roles and underlying mechanisms of three specific long non-coding RNAs (lncRNAs) that have been identified as dependent on the activity of the MYC oncogene. These lncRNAs were previously identified in a study by P. Laneve et al. (2021), where MYC inhibition in cell lines derived from primary medulloblastoma (MB) patient samples with MYC-driven tumors demonstrated their involvement in carcinogenesis. By utilizing a bioinformatics pipeline and analyzing data from the Genotype-Tissue Expression (GTEx) project and dbSNP database, insights were gained into the genetic variations and gene expression changes associated with these MYC-dependent lncRNAs.

In this study, it was firstly explored the distribution of polymorphisms across the genetic loci of MYC-dependent lncRNAs, finding a higher SNP density in regions lacking protein coding genes. Furthermore, it was conducted a clinical significance analysis of pathogenic polymorphisms and discovered that known pathogenic polymorphisms were primarily found in protein coding genes, with none identified in the MYC-dependent lncRNAs examined. This is probably due to the lack of studies on the role of lncRNAs in disease pathogenesis, hence further genetic and molecular studies are needed to decode the link between these RNA molecules and disease mechanisms.

The analysis was subsequently delved into expression quantitative trait loci (eQTLs) within the genomic loci of MYC-dependent lncRNAs, with a specific focus on lncMB1. It revealed a total of 164 cis-eQTLs associated with lncMB1, establishing links between seven SNPs and eight genes. The variation-gene associations in lncMB1 exhibited a remarkable pattern, with three specific variations consistently showing similar associations with multiple genes. Additionally, the changes in expression of these variations occurred across different tissues and are similar in value. This consistent change in expression across different tissues and may indicate the involvement of common regulatory mechanisms that lead to similar outcomes when disrupted by various genetic variations. Among the genes, RHOT1 displayed the highest number of associations. The RHOT1 (Ras Homolog Family Member T1) gene is known for its involvement in mitochondrial transport and its link to Parkinson's disease and

various cancers. This association indicate that lncMB1 may play a crucial role in modulating RHOT1 expression, potentially influencing mitochondrial function and contributing to disease progression. To gain insights into the molecular structure of lncMB1, the RNAfold software was used to predict its secondary structures and examine the effects of two mutations (rs7208838 and rs13342625) on the RNA structure. The rs13342625 mutation disrupted a highly energetically favorable hairpin structure, leading to the formation of a loop. Conversely, the rs7208838 mutation caused a structural modification, transforming a poorly structured RNA domain into a more stable hairpin structure.

Based on the structural analysis results, it can be hypothesized that the genetic variations on lncMB1 may affect its secondary structure, potentially influencing its interaction with proteins and other molecules. This, in turn, may impact the putative regulatory effect of lncMB1 on nearby genes, contributing to disease development or progression.

In literature the importance of lncMB1 in RHOT1 expression has already been proved (P. Laneve et al., 2021), even though experiments suggest a role for lncMB1 as a positive regulator of RHOT1 gene at the translational and/or protein stability level, rather than in modulating its mRNA levels. Further studies will be needed to uncover the mechanism behind the regulatory activity of lncMB1 and regulatory networks it may be involved in, and future efforts may be directed to study possible association between rs7208838 and rs13342625 with risk of medulloblastoma cancer, as done by Dan Li et al. (2016) for the lncRNA ZNRD1-AS1.

In conclusion, eQTLs play a crucial role in the functional analysis of lncRNAs. In the context of lncRNAs, eQTLs provide valuable insights into the regulatory mechanisms and functional implications of these non-coding transcripts. By identifying and characterizing eQTLs within lncRNAs, researchers can unravel putative genes that may be under the regulatory activity of the non-coding transcript and characterize their role in diseases.

5. Bibliography

1) Laurent, G. S., Wahlestedt, C., & Kapranov, P. (2015). The Landscape of long noncoding RNA classification. Trends in genetics, 31(5), 239-251.

2) Cipriano, A., & Ballarino, M. (2018). The ever-evolving concept of the gene: the use of RNA/protein experimental techniques to understand genome functions. Frontiers in molecular biosciences, 5, 20.

3) Zhu, J., Fu, H., Wu, Y., & Zheng, X. (2013). Function of lncRNAs and approaches to lncRNA-protein interactions. Science China Life Sciences, 56, 876-885.

4) Gibb, E. A., Brown, C. J., & Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. Molecular cancer, 10(1), 1-17.

5) GTEx Consortium, Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., ... & Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348(6235), 648-660.

6) GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318-1330.

7) Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics, 32(10), 1479-1485.

8) Westra, H. J., & Franke, L. (2014). From genome to function by studying eQTLs. Biochimica et Biophysica Acta (BBA)-molecular basis of Disease, 1842(10), 1896-1902.

9) Rea, J., Carissimo, A., Trisciuoglio, D., Illi, B., Picard, D., Remke, M., ... & Caffarelli, E. (2021). Identification and functional characterization of novel MYC-regulated long noncoding RNAs in group 3 medulloblastoma. Cancers, 13(15), 3853.

10) Northcott, P. A., Robinson, G. W., Kratz, C. P., Mabbott, D. J., Pomeroy, S. L., Clifford, S. C., ... & Pfister, S. M. (2019). Medulloblastoma. Nature reviews Disease primers, 5(1), 11.

11) Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for molecular biology, 6, 1-14.

12) Li, J., Xue, Y., Amin, M. T., Yang, Y., Yang, J., Zhang, W., ... & Gong, J. (2020). ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types. Nucleic acids research, 48(D1), D956-D963.

13) Li, D., Song, L., Wen, Z., Li, X., Jie, J., Wang, Y., & Peng, L. (2016). Strong evidence for LncRNA ZNRD1-AS1, and its functional Cis-eQTL locus contributing more to the susceptibility of lung cancer. Oncotarget, 7(24), 35813.

14) Suvanto, M., Beesley, J., Blomqvist, C., Chenevix-Trench, G., Khan, S., & Nevanlinna, H. (2020). SNPs in lncRNA regions and breast cancer risk. Frontiers in Genetics, 11, 550.