# Annotation and characterization of lncRNAs in Lung Cancer

## Abstract

Lung cancer, a leading global cause of mortality, demands innovative diagnostics and therapeutic strategies. This study focuses on non-small cell lung carcinoma (NSCLC), representing over 80% of cases. Long non-coding RNAs (lncRNAs), emerging as critical players in NSCLC and widely dysregulated in cancer can provide new therapeutic approaches. Two distinct NSCLC A549 cell line populations, parental and paraclone, were explored through RNA sequencing. Parental cells are considered as a less aggressive phenotype than paraclone cells, due to these are characterized as mesenchymal, chemotherapy-resistant and migration and invasion increased activity. **The main purpose** is to identify lncRNAs that may be a potential target to force the conversion to the clinically beneficial phenotype. Initially, a reference-guided transcriptome assembly from all the samples revealed 29,841 novel transcripts, including lncRNAs. Subsequently, a differential expression analysis at the transcript level was performed to find the transcripts and genes whose expression is altered between the two cell types and could be responsible for the more recalcitrant tumor phenotype. Lastly, an integrative analysis based on the intergenic nature of the transcripts, the quality of the annotations on the transcription start sites and end sites and the coding protein potential, was developed to prioritize the potential lncRNA targets.

## Introduction

Lung cancer is a malignant tumor with high mortality and morbidity in the world, being a significant public health concern. The GLOBOCAN 2020 estimation of cancer incidence and mortality produced by the International Agency for Research on Cancer (IARC) shows lung cancer as the leading cause of cancer death, with an estimated 1.8 million deaths in 2020. Especially, non-small cell carcinoma (NSCLC) is the most common type of lung cancer, being more than 80 % of cases of lung cancer. The poor progression of the lung cancer survival rate in the last years and the lack of a full elucidation of the pathogenesis of NSCLC lead to the necessity of finding new diagnosis, prognosis, and therapeutic approaches [1].

Recently, long-chain non-coding RNAs (lncRNAs) have become the focus of considerable interest. LncRNA is a type of non-coding RNA whose transcript length is larger than 200 nucleotides. They also can be spliced, poly-adenylated and capped. Previous studies suggest lncRNA plays an important role in NSCLC through different pathways, including regulation of proliferation, invasion and migration of cancer cells through their interactions with miRNAs [2]. Furthermore, it has been proved that a variety of lncRNAs are abnormally expressed in NSCLC tissues [3]. Unlike mRNA, the expression pattern of lncRNA is highly specific depending on the type of cell, the tissue, and the biological context. Therefore, it can better reflect the disease status as well as indicate the disease diagnosis or classification. Carcinogenic lncRNAs, such as Gm15290 and licl00473 and anti-tumoral lncRNAs, such as IGF2AS and MEG3 have been identified [4]-[5]. Researching these lncRNAs and identifying novel ones could not only open a new window of effective targets for therapeutic drugs but also in predictive biomarkers of radiotherapy and chemotherapy [6].

In this project, two distinct populations of NSCLC human cell line A549 have been explored. The parental population is composed of subpopulations of holo-, mero- and paraclone cells, whereas the paraclone population is associated with mesenchymal phenotype, characterized by therapy resistance, fibroblast-like and elongated morphology and increased migration and invasion capacity, attributes that contribute to rendering the tumor more recalcitrant [7]. The aim is to create a reference-guided transcriptome assembly based on the RNA-sequencing of these

populations and further identify genes that are differentially expressed between them. Particularly, the lncRNAs that may be a potential target to force the conversion to clinically beneficial phenotypes.

## Material and methods

*RNA Sequencing*

Illumina paired-end sequencing reads were obtained from A549 parental cells (P1, P2, P3) and paraclone (3.2, 3.4, 3.7) cells, concretely, three biological replicates per clone type. The type of library used was TrueSeq Stranded mRNA libraries, as explained in [7].

*Quality control*

To evaluate the quality of the reads, FASTQC software (version 0.11.9) was used [8]. Read trimming was not required to improve the quality.

*Alignment*

The reads were mapped to the reference human genome version hg38/GRCh38 [9] using HISAT2 (version 2.2.1) and taking into account the strandedness and the mated-pairs reads [10]. To execute this step the indexed version of the reference human genome was built. For each sample, one SAM file was generated that subsequently was converted into BAM file using SAM tools (version 1.10) [11].

*Transcriptome assembly*

In order to assemble the entire transcriptome of all the samples, the reference-guided approach with the reference human genome (release 45, GRCh38.p14) comprehensive gene annotation (CHR) from Gencode [12] was followed. This annotation contains the reference chromosomes only. Using StringTie (version 1.3.3b) as software [13], firstly, the guided assembly was performed for each BAM file from each sample. Secondly, the assemblies were merged in one meta-assembly (GTF format).

*Quantification of the expression*

To quantify the expression of all the genes, the abundances of transcripts per sample were quantified using Kallisto (version 0.46.0) program [14]. The parameters for attending the strandedness and performing 100 bootstraps were specified. Before the quantification, a reference transcriptome FASTA file (using the tool gffread from Cufflinks version 2.2.1 [15]) and the Kallisto indexed format from the meta-assembly GTF file (generated previously) were constructed. It was required to be used as an annotation file.

*Differential expression analysis*

To find which transcripts are differentially expressed (DE) between the two populations of cells: parental samples (P1, P2 and P3) and paraclone samples (3.2, 3.4 and 3.7) Sleuth (version 0.30.1) software was used [16]. Due to the differential expression analysis was performed at transcript level, the mapping of these transcripts to infer which are the DE genes was required. This information was extracted from the transcriptome meta-assembly generated previously. In this step and on the subsequent downstream analysis only the transcripts on reference chromosomes were included, the ones on assembly patches were excluded to increase the levels of confidence. The test performed was the Wald test, and the cutoffs chosen to considerate a gene as significantly differentially expressed were | 2 | for $\log_2$FC and 0.05 for the false discovery rate (FDR).

*Integrative analysis*

After the DE analysis, a list of transcripts that are differentially expressed between the parental cells and paraclone cells has been identified. However, the main purpose of this project is to select which of the DE transcripts are in first instance novel, and in the second instance, potential lncRNAs capable of enhancing the conversion to a beneficial cell phenotype, in this case, the paraclone cells. Because of that, may be interesting to identify which of the novel transcripts are intergenic, in other words, located between genes, and if their 5' and 3' annotations have enough confidence to claim that they are expressed, as well as their low protein coding potential.

Firstly, to identify the novel intergenic genes, two different BED files were created from the transcriptome meta-assembly. One file contains the novel transcripts, and another one contains the annotated protein coding genes. The intersection of these two files was extracted, reporting only the novel genes that does not overlap with the annotated genes, performed with BED Tools (version 2.29.2) [17].

Secondly, to assess the annotations in 5' and 3' of the novel transcripts two datasets from FANTOM project were used: cluster annotations of transcription start sites (TSS) and cluster annotations of polyA sites, respectively [18]. Both datasets come from the Homo sapiens genome (GRCh38.96). For this purpose, the intersection of the transcription start sites and the dataset of annotations of TSS was performed. In the same way, the intersection of the transcription end sites (TES) and the dataset of annotations of polyA sites was performed using BED Tools (version 2.29.2) [17]. To be less restrictive, a window of 100 nucleotides was created around the TSS and the TES of each novel transcript. The strandedness was also considered.

Thirdly, to distinguish between coding and noncoding novel transcripts, CPAT (version 1.2.4) was used, it is based on a logistic regression model built with the following features: open reading frame size and coverage, Fickett TESTCODE statistic and hexamer usage bias [19]. The human hexamer bias and the human logistic model were required [20], as well as the novel transcripts sequences in FASTA file format. The FASTA file was previously generated using BED Tools (version 2.29.2) [17].

*Potential candidates' selection*

The features mentioned in the integrative analysis, such as the TSS and TES annotations, the intergenic condition and the protein coding potential, and the values of $log_2FC$ and adjusted p-values were compiled in order to find potential novel candidates.

The novel transcripts that were intergenic and were well-annotated in the TSS and TES were first prioritized. Then, they were sorted from greater to smaller $log_2FC$ absolute values and in ascending adjusted p-values to consider if they were significantly DE genes ($log_2FC > | 2 |$ and adjusted p-value $< 0.05$).

## Results

*Quality control*

The quality of the reads from all samples was the required to perform the next steps of the analysis, as shown in Figure 1A. Considering Q30 as the threshold, in other words, 99.9% of base call accuracy, the overall of all positions fulfil this requirement. Other features like the per sequence GC content and per base N content were reported as satisfied. In addition, overrepresented sequences and adapter content was not found. Despite the failure on the per base sequence content feature, it is considered as unmeaningful. This fail is due to the first 10-12 bases result from

hexamer priming that occurs during RNA-seq library preparation, that gives an enrichment in particular bases for these initial nucleotides (Figure 1B).

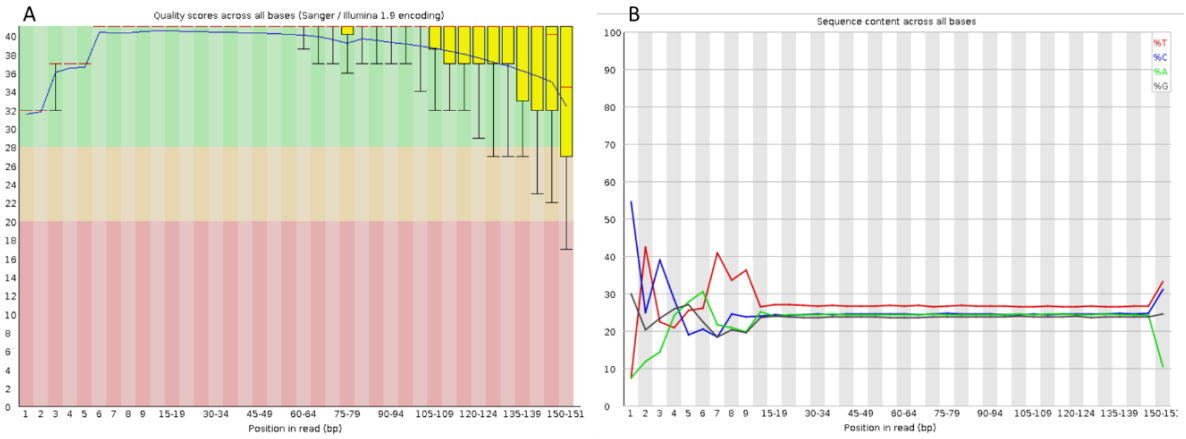


**Figure 1. A) Per base sequence quality of the reads from P1 sample. B) Per base sequence content.** Other samples shown similar results.

*Alignment*

The overall alignment rates were between 97.87% and 98.20% in all the samples, meaning that the read mapping was successful (Table 1). Furthermore, the strandedness was proved through Integrative Genomics Viewer (IGV) visualization [21], comparing the reference human genome to the BAM file from one sample.

**Table 1. Number of reads for each replicate. "P" refers to parental cells and "3" to paraclone cells.**

| Sample | Number of reads per mate | Overall alignment rate (%) |
|---|---|---|
| P1 | 32,840,352 | 97.87 |
| P2 | 32,894,526 | 97.91 |
| P3 | 33,414,876 | 98 |
| 3.2 | 35,438,437 | 97.91 |
| 3.4 | 33,442,327 | 98.14 |
| 3.7 | 33,745,478 | 98.20 |

*Transcriptome assembly*

To assess the transcriptome meta-assembly obtained from merging the assemblies of the six biological samples the number of transcripts, exons and genes were counted considering the ones that were on the reference chromosomes and on the assembly patches. The number of transcripts found was 265,343, and 29,841 of these were classified as novel transcripts due to their absence of Gencode annotation. The number of genes was 60,017, and 29,841 of these genes were novel genes, and the number of exons was 1,779,950.

As a quality check, the number of transcripts and genes composed of just one single exon were counted, finding 22,398 single exon genes and 26,698 single exon transcripts. These structures were considered as lower significant than the multi-exon ones.

*Quantification of expression*

Two different units of expression were used: the estimated counts, that represent the raw count of reads that are assigned to each transcript and transcripts per million (tpm), that is a normalized

measure of transcript abundance taking into account both library size and transcript length. It provides a more comparable measure of expression across different samples. The results expressed in estimated counts can be found in Supplementary 1.

As a quality check, the total number estimated counts for each sample were compared to the number of raw reads, similar values were obtained as expected. Furthermore, the number of transcripts, novel transcripts, genes, and novel genes detected for each sample was counted, defining detected as the transcripts or genes that have a number of estimated counts greater than 0.

*Differential expression analysis*

The enhanced volcano plot (Figure 2) shows which genes are significantly differentially expressed between the parental samples and the paraclone samples. Up-regulated DE genes in the parental samples like C5 or HNF4A, and down-regulated genes in the parental samples like CXCL8 were also found in the previous studies [7], which validates this analysis. The detailed results can be found in Supplementary 2.

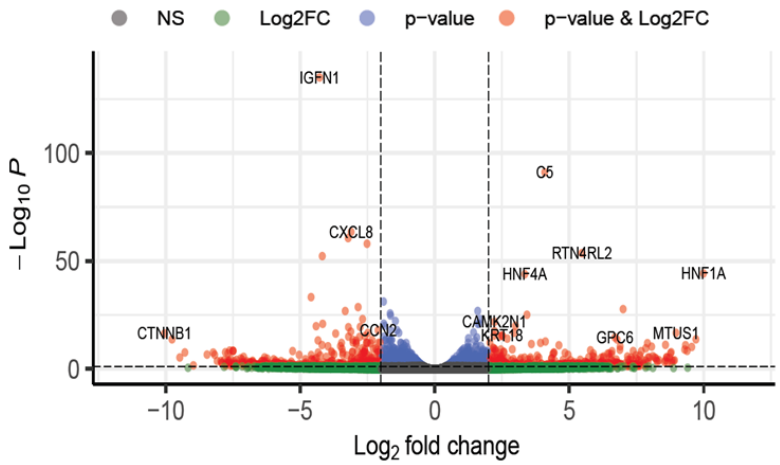


**Figure 2. Enhanced volcano plot of parental samples versus paraclone samples**. The $log_2$FC indicates the mean expression level for each gene. Each dot represents one gene. In red: significantly DE genes (FC > |2|, p-value < 0.05). In blue: non-DE genes with significant p-value (FC < |2|, p-value < 0.05). In green: DE genes without statistical significance (FC > |2|, p-value > 0.05). In grey: non-significant and non-DE genes.

*Integrative analysis*

Regarding the steps of the integrative analysis, the percentage of intergenic novel transcripts, the percentage of transcripts that were 5' well-annotated, the percentage of them that were 3' well-annotated and the percentage of them that were well-annotated in both transcription sites 5' and 3' were calculated, as shown in Figure 3A. In addition, the transcripts with a potential coding estimation greater than 0.364 were classified as protein coding.

The intergenic transcripts are considered more interesting due to their easy manipulability that could facilitate therapeutics approaches such as CRISPR. The same percentages were calculated for the novel intergenic transcripts that fulfil the conditions of 5' and 3' annotations or have a potential coding estimation greater than the threshold, as shown in Figure 3B. The number of novel intergenic transcripts was 141, which 26 of them were well annotated in 5', 34 of them were well annotated in 3' and 6 of them were well annotated in both ends.

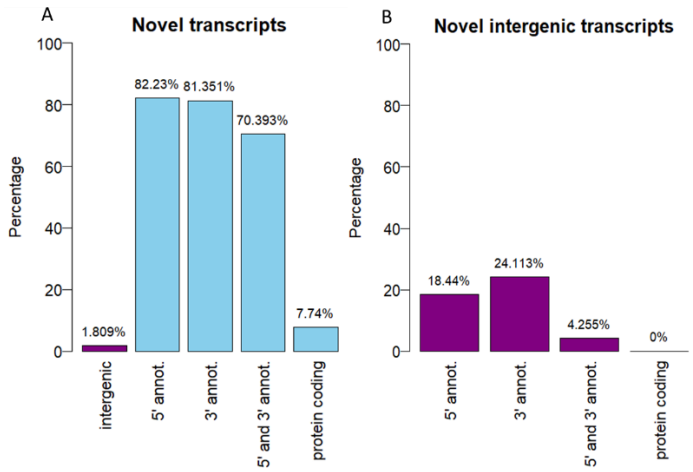


**Figure 3. Percentage of intergenic, 5' well-annotated, 3' well-annotated, 5' and 3' well-annotated and protein coding transcripts above the total of: A) novel transcripts and B) novel intergenic transcripts.**

*Potential candidates' selection*

In the optimum scenario, the novel transcripts that fulfil the three features: intergenic, 5' well-annotated and 3' well-annotated and are significantly DE ($log_2FC > | 2 |$ and adjusted p-value < 0.05) would be chosen. The protein coding potential provides additional information and could be used as evidence to indicate noncoding region (protein coding potential < 0.364). However, is not interpreted as a filter with a stringent threshold. The results detailed can be found in Supplementary 3.

In Figure 4, the visualization in IGV of one possible candidate is shown. It is considered as intergenic because it only overlaps with an uncharacterized locus (XR_930101.4 [22]). The TSS and TES overlaps with the dataset annotations of TSS and TES, respectively, indicating that they are well annotated and increasing the level of confidence that this hypothetic lncRNA is expressed.

However, it is necessary to establish a trade-off between the four main features mentioned above and the differential expression conditions, because not all the transcripts that fulfil the features are considered as DE genes and vice versa.


**Figure 4. Visualization of one candidate (MSTRG.28189.1) in IGV** [21]. The transcription start site (overlap3prime.bed, in red) and transcription end site (overlap5prime.bed, in green) overlap with the FANTOM cage clusters of TSS (refTSS_v4.1_human_coordinate.h38.bed, in dark red) and TES (atlas.clusters.2.0.GRCh38.96.bed, in dark green), respectively. The gene is considered as intergenic, it only overlaps with an uncharacterized locus (XR_930101.4 [22]).

## Discussion and conclusion

In conclusion, this study has successfully revealed a list of novel long non-coding RNAs in non-small lung cancer (NSCLC) through a comprehensive RNA sequencing analysis. This analysis is based on a novel reference-guided transcriptome assembly built by merging the entire transcriptome of the parental and paraclone populations of NSCLC human cell line A549. The differential expression analysis manifests genes significantly expressed between the two populations, providing insight into potential therapeutic targets that could effectively force the conversion to a cancer cell more beneficial phenotype, less resistant and recalcitrant. In addition, the integrative approach prioritizes candidates based on their intergenic nature, well-annotated transcription start and end sites, and non-coding potential.

The intergenic nature allows for relatively specific targeting without affecting nearby protein-coding genes, therefore minimizing off-target effects and reducing the risk of unintended side effects. It also makes the target more accessible for genome editing technologies like CRISPR-Cas9, or other genetic manipulations like antisense oligonucleotides, RNA interference, or small molecules [23]. The well-annotated transcription start and end sites, and non-coding potential are evidences for consideration of the novel transcripts as lncRNAs, due to their natural characteristics of being non-coding regions that are expressed. Nevertheless, the list of potential candidates obtained might be interpreted as a preliminary result for further studies, like functional analysis.

In terms of future perspectives, the validation of the functional roles of the identified lncRNAs through in vitro and in vivo experiments, like knockdown or overexpression studies, is essential for confirming their potential therapeutic relevance. Considering the challenges of performing functional analysis in non-coding regions, the integration of functional genomic data, such as DNAse I hypersensitivity sites and chromatin marks can serve as additional filters to refine the prioritization process, as has been proved in previous studies [6].

Furthermore, machine learning algorithms can be employed. Integrating various features such as sequence conservation, subcellular localization, and interaction networks between lncRNAs and the intracellular miRNAs or proteins, into a predictive model can enhance the accuracy of prioritization. Supervised learning approaches, trained on known functional lncRNAs, can be applied to assign a score to each candidate, aiding researchers in focusing on the most promising targets [24].

## Git Hub repository (code)

https://github.com/lfercer/lncRNA_project

## Supplementary materials

https://github.com/lfercer/lncRNA_project/tree/214c3065ce70f5ac26efc66b33e64ebb94260340/Supplementary

## References

[1] H. Sung *et al.*, 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries', *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] Q. Hu, H. Ma, H. Chen, Z. Zhang, and Q. Xue, 'LncRNA in tumorigenesis of non-small-cell lung cancer: From bench to bedside', *Cell Death Discov.*, vol. 8, no. 1, p. 359, Aug. 2022, doi: 10.1038/s41420-022-01157-4.

[3] R. Chen *et al.*, 'Comprehensive Analysis of lncRNA and mRNA Expression Profiles in Lung Cancer', *Clin. Lab.*, vol. 63, no. 2, pp. 313–320, Feb. 2017, doi: 10.7754/Clin.Lab.2016.160812.

[4] T. L. Kruer *et al.*, 'Expression of the lncRNA Maternally Expressed Gene 3 (MEG3) Contributes to the Control of Lung Cancer Cell Proliferation by the Rb Pathway', *PloS One*, vol. 11, no. 11, p. e0166363, 2016, doi: 10.1371/journal.pone.0166363.

[5] Y. Dong, X. Huo, R. Sun, Z. Liu, M. Huang, and S. Yang, 'lncRNA Gm15290 promotes cell proliferation and invasion in lung cancer through directly interacting with and suppressing the tumor suppressor miR-615-5p', *Biosci. Rep.*, vol. 38, no. 5, p. BSR20181150, Oct. 2018, doi: 10.1042/BSR20181150.

[6] R. Esposito *et al.*, 'Multi-hallmark long noncoding RNA maps reveal non-small cell lung cancer vulnerabilities', *Cell Genomics*, vol. 2, no. 9, p. 100171, Sep. 2022, doi: 10.1016/j.xgen.2022.100171.

[7] C. C. Tièche *et al.*, 'Tumor Initiation Capacity and Therapy Resistance Are Differential Features of EMT-Related Subpopulations in the NSCLC Cell Line A549', *Neoplasia*, vol. 21, no. 2, pp. 185–196, Feb. 2019, doi: 10.1016/j.neo.2018.09.008.

[8] 'Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data'. Accessed: Dec. 19, 2023. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[9] 'Homo sapiens genome assembly GRCh38.p14', NCBI. Accessed: Dec. 24, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000001405.40/

[10] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nat. Biotechnol.*, vol. 37, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[11] H. Li *et al.*, 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[12] A. Frankish *et al.*, 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Res.*, vol. 47, no. D1, pp. D766–D773, Jan. 2019, doi: 10.1093/nar/gky955.

[13] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nat. Biotechnol.*, vol. 33, no. 3, pp. 290–295, Mar. 2015, doi: 10.1038/nbt.3122.

[14] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, 'Near-optimal probabilistic RNA-seq quantification', *Nat. Biotechnol.*, vol. 34, no. 5, Art. no. 5, May 2016, doi: 10.1038/nbt.3519.

[15] C. Trapnell *et al.*, 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–515, May 2010, doi: 10.1038/nbt.1621.

[16] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, 'Differential analysis of RNA-seq incorporating quantification uncertainty', *Nat. Methods*, vol. 14, no. 7, Art. no. 7, Jul. 2017, doi: 10.1038/nmeth.4324.

[17] A. R. Quinlan and I. M. Hall, 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[18] M. Lizio *et al.*, 'Gateways to the FANTOM5 promoter level mammalian expression atlas', *Genome Biol.*, vol. 16, no. 1, p. 22, Jan. 2015, doi: 10.1186/s13059-014-0560-6.

[19] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, 'CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model', *Nucleic Acids Res.*, vol. 41, no. 6, p. e74, Apr. 2013, doi: 10.1093/nar/gkt006.

[20] 'CPAT - Browse /v1.2.2/prebuilt_model at SourceForge.net'. Accessed: Jan. 23, 2024. [Online]. Available: https://sourceforge.net/projects/rna-cpat/files/v1.2.2/prebuilt_model/

[21] J. T. Robinson *et al.*, 'Integrative Genomics Viewer', *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.

[22] 'LOC105376136 uncharacterized LOC105376136 [Homo sapiens (human)] - Gene - NCBI'. Accessed: Jan. 25, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/gene/?term=XR_930101.4

[23] Y. Chen, Z. Li, X. Chen, and S. Zhang, 'Long non-coding RNAs: From disease code to drug role', *Acta Pharm. Sin. B*, vol. 11, no. 2, pp. 340–354, Feb. 2021, doi: 10.1016/j.apsb.2020.10.001.

[24] J. Zhang *et al.*, 'CLING: Candidate Cancer-Related lncRNA Prioritization via Integrating Multiple Biological Networks', *Front. Bioeng. Biotechnol.*, vol. 8, 2020, Accessed: Jan. 24, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00138