RNA-Sequencing Report

**lncRNA lung cancer, project 1 [Holoclonal]**

**Supervisor : Dr. Pangiotis Chouvardas**

# Abstract

The project aimed to investigate and identify novel long non-coding RNAs capable of explaining the phenotypical characteristics of non-small cell lung cancer (NSCLC), drawing inspiration from the publication of Tièche et al.1 This toy-study was conducted on human lung cancer cells (lineage A549) and provides a list of potential novel non-coding genes.

The research explored modern bioinformatics techniques, with a particular focus on de novo reference-guided transcriptome assembly and differential gene expression. Utilizing Next-Generation Sequencing (NGS), biological data were converted into machine-readable files. These files were then analyzed using cloud computing pipelines, enabling the calculation and representation of insightful information. Additionally, the data were processed to construct the transcriptome, facilitating a coherent interpretation of the enormous amount of data at our disposal.

# Introduction

Gene investigation has been the go-to method to answer many biological questions, and this approach has exponentially grown over the last decades, mainly propelled by a deeper and more complete understanding of genes' implication in cellular regulation. A more comprehensive understanding of the genome's structure and regulatory networks highlighted the importance of developing faster and more reliable methods to retrieve DNA sequences.

The advent of Next-Generation Sequencing (NGS) allowed the conversion of genomic data (DNA) into text files (FASTQ files), which can then be parsed and analyzed by computers. Soon after, driven by the novel and exciting discoveries that were (and constantly are) made, new questions arose and pushed the development of new techniques to investigate other biological molecules, such as RNA (DNA transcripts).

Many fields in biology can exploit the properties of high-throughput machines, and one of them is oncology. It has been proven that cancers arise from gene mutations that can dysregulate important checkpoints in many cellular pathways. The (over-)expression of oncogenes and under-regulation of tumor suppressor genes cause healthy cells to become cancerous and develop toxic traits for nearby tissues, other than growing uncontrollably. Understanding the intricate networks developed by cancer cells is key to opening up research to new therapeutics for cancer patients, increasing their life expectancy, or even curing them.

Out of all cancers, lung cancer is the most common cause of cancer-related mortality worldwide; more than 80% of lung tumors are non–small cell lung cancers (NSCLCs)1. These are the main reasons why research on these specific cancer cells is still relevant today. During the last years, protein-coding genes were the main focus of research for this particular disease. Still, because life expectancy has not continued to grow even with new therapeutics, researchers have widened their search for possible causes even beyond coding RNAs, in particular, long non-coding RNAs (lncRNAs), which are known to be involved in the regulation of many processes and, therefore, can be potential oncogenes/tumor suppressor genes.

Talk about HOLOCLONALITY

# Material and methods

For this project, a total of twelve (six pairs) FASTQ files representing RNA-sequenced cells from NSCLC A549 lineage were provided, with half coming from holoclonal cells and the other from control -healthy- cells.

The fastqc pipeline (version 0.11.9) facilitated graphical visualization on HTML files of basic reads information and quality. Each cell sample is represented by two FASTQ files, one for the forward strand and the other for the reverse strand.

Reads mapping and reference indexing were performed with the HISAT2 module (2.2.1) against the comprehensive human gene annotation reference GRCh38 (version November 2024). The resulting SAM files were then converted into their respective binary version (BAM) using samtools (1.10).

The transcriptome was assembled with STRINGTIE (1.3.3b) set with –rf strandness, producing six GTF files representing the transcriptome for each sequenced sample, which were then merged with the option –merge. Assembly was guided by the same human genome annotation reference used for mapping.

The complete transcriptome was prepared for gene quantification through Kallisto (0.46.0) and Cufflinks (2.2.1), aiding in the indexing of the transcriptome into a FASTA file. The Kallisto -quant option allowed for the exact quantification of reads by checking the detection of reads for each pair of FASTQ files, aided by the index FASTA created through a pseudoalignment system.

Differential gene expression was analyzed using R Studio (Version 2023.09.0+463) with the Sleuth package (0.31.0), which was fed custom tables for differentiation between annotated and novel genes. The R analysis provided visual representation and a ranked table for differentially expressed genes on a log2Fold scale.

To perform integrative analysis of transcripts, GTF files needed to be parsed and trimmed into BED files. The conversion was made possible with BEDTools intersect (module version 2.29.2). The intersect option allowed for the detection of intergenic regions, thus identifying proper novel genes that are not non-annotated isoforms of already known genes.

Start and end sites of transcripts were quality assessed with FANTOM 5 CAGE Clusters (5’) and PolyA sites (3’) with a sliding window of ~50nt to account for small errors in nucleotide position.

To assess the probability of novel genes being protein-coding, CPAT (1.2.4) was utilized.

# Results

Read quality and statistics

FASTQ files representing the RNA-sequencing’s data, count approximately 34 millions reads each. Everyone shows great quality, and no trimming or sanity procedures need to be done. For instance, the following figure (Fig. 1) shows the “Per base sequence quality”, which is the quality -probability to detect the right nucleotide- of each base per position.

A clear pattern of Illumina sequencing is visible, where the quality increases as we get closer to the center of the sequence and less accurate at both ends.

A screenshot of a computer screen

Description automatically generated

Figure 1: Hololonal cell 1\_5 RNA-sequencing (R2) per base sequence quality statistic

Read mapping

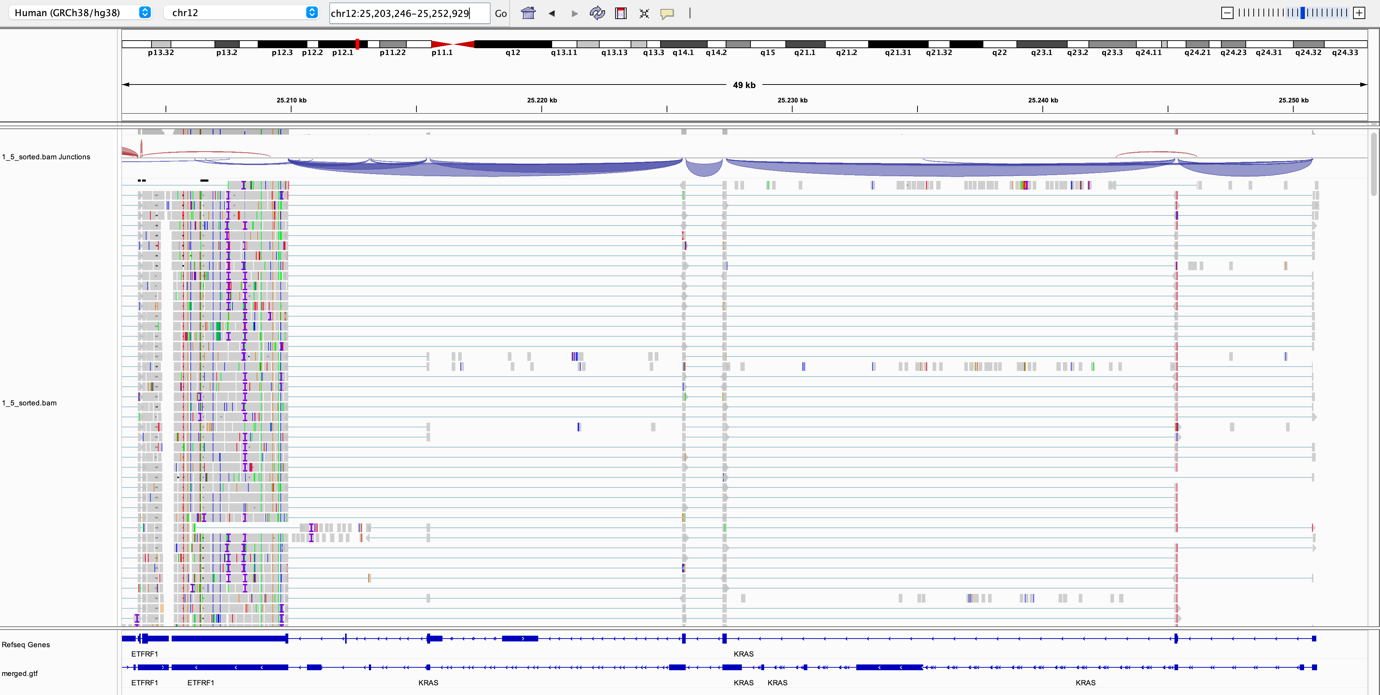
Mapped reads, initially in SAM (Sequence Alignment and Map) format, were subsequently converted to BAM (Binary Alignment and Map) files, complete with their respective indexed files. Overall alignment rates for each file exceed 97% (refer to OutputP1P2.txt for precise information), indicating nearly perfect alignment for each sample. Quality checks were performed using IGV software (2.16.2), where alignments were manually inspected for sense correctness against the installed reference on IGV software (GRCh38/hg38) (Figure 2). The correct orientation of aligned data is crucial for a coherent transcriptomic assembly, essential for assessing differential gene expression and accurate gene position identification. Alignment orientation is visualized by a change in color, with blue arches (genetic gap due to intronic regions) representing the forward sense and red ones representing the reverse sense. Accurate orientation is critical for achieving the research project's end-goal of providing our lab colleagues with a list of the most probable genes that could explain the observed phenotype, motivating our research question. Therefore, an inverted sequence would not yield positive lab results. 

Figure 2: KRAS gene in 1\_5 holoclonal sample portraying correct strandness when compared to GRCh38 reference

Transcriptome assembly

Six GTF files, representing the meta-data assembly, were merged into a single file encompassing the transcriptome data from both control and holoclonal lung cancer conditions. Due to the dynamic nature of genetic transcription, not all genes present in an individual genome are transcribed into RNA. The unified GTF file offers a comprehensive overview of cell activity in both conditions, serving as a baseline for comparing gene activity in cancer data.

General information regarding the number of genes, transcripts, exons, novel transcripts, novel exons, single exon transcripts, and single exon genes is available in scripts/Questions/code-to-answer.txt

Quantification

The unit employed for quantifying transcription is TPM (Transcripts Per Million). This is demonstrated by summing all elements in the fifth column (tpm) of one of the abundance.tsv files generated by Kallisto. The final result yields 106 total transcripts per million.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1\_1 | 1\_2 | 1\_5 | P\_1 | P\_2 | P\_3 |
| Transcripts | 120270 | 121798 | 118512 | 115908 | 116824 | 118036 |
| Genes | 52239 | 52935 | 51537 | 49798 | 49916 | 50814 |
| Novel Trans. | 10534 | 10624 | 10504 | 10432 | 10463 | 10565 |
| Novel Genes | 2669 | 2676 | 2660 | 2601 | 2612 | 2675 |

Differential expression

Due to issues with Biomart in R script, a custom table was manually created to insert information about annotated genes, as detailed in the script scripts/Step5-DE/genesID.sh. Two graphical options, a heat map or a "volcano plot," can be used to visually represent differentially expressed genes. The change in color indicates the effective alteration from the control condition, either up-regulated or down-regulated. The choice of base two in the log-operation is arbitrary for ease of interpretation. Figure 3 illustrates the differentially expressed genes that are already annotated.



Figure 3: Volcano plot representing differentially expressed annotated genes

With these findings, further investigation into the paper's claims is possible. For instance, one of the discoveries was the over-expression of CDH17 in holoclonal cells, which is an epithelial marker, aligning with holoclonal characteristics of being the outer layer of cancers and respecting epithelial-to-mesenchymal transition. The subsequent Figure 4 illustrates the change compared to control cells, confirming the epithelial phenotype and a higher potential for migration and invasion.



Figure 4: CDH17 (ENST00000441892.6) expression in control and holoclonal sample

Interestingly, other mesenchymal markers like THY were not found, further emphasizing the phenotype characteristics of holoclonal cells.

Integrative analysis

Over the entire assembly, 213 novel intergenic genes have been identified. Using CPAT, their probability of being protein-coding has been investigated. According to documentation, the human coding probability (CP) is 0.364. Therefore, novel transcripts with coding probability less than 0.364 will be considered highly potentially non-coding. This results in approximately 4.61% of coding genes among our newly discovered transcripts (95.31% non-coding). The result is encouraging, as a strong majority of the newly discovered transcripts are believed to be non-coding, and therefore potentially lncRNAs.

To assess the quality of the transcripts, 5’ and 3’ annotation accuracy has been calculated, showing 80.656% accuracy for Transcription Start Site (TSS) and 79.914% for Transcription End Site (TES). The similar results imply a balanced assembly, taking into account a window for the correction of false positives and negatives.

Prioritization

Due to time limitations, a prioritization list has not been developed. Nonetheless, the steps I would take to complete such a task would involve merging the shared transcripts from those non-coding genes and the intergenic transcripts. The final table would, therefore, contain transcripts that are not part of other genes and are most probably not protein-coding

# Discussion

Overall, all the steps taken, from the sanity check of reads to the integrative analysis of potentially coding transcripts, were executed with great results and produced a transcriptomic assembly very similar to the reference genome used. Therefore, the accuracy of novel transcripts is higher due to the successful guiding system.

The high number of both intergenic and non-protein coding transcripts suggests the presence of RNAs that could play a regulatory role in cancer development and/or survival and, finally, be targeted by new therapeutics.

Interestingly, the discoveries claimed by the paper of Tièche et al.1 stood their ground by reappearing in our analysis, for instance in Figure 4, but also in many more instances. This result reconfirms once again the central dogma of cancer biology, proving many aspects long studied by scientists, such as the EMT (Epithelial-to-Mesenchymal-Transition) or other specific genetic characteristics evolved by cancer during differentiation.

In the future, a more extensive research could be performed. Already, with the small sample size at our disposal, many novel pieces of information have been extracted. A more comprehensive study could improve the results that have already been obtained and strengthen the probability in favor of truly non-coding RNAs.

# Supplementary material– git hub link (ev documentation)

<https://github.com/michael-jopiti/RNA-seq>

# References