RNA-Sequencing Report

lncRNA lung cancer, project 1 [Holoclonal]
Supervisor : Dr. Pangiotis Chouvardas

Abstract

The project aimed to explore and uncover novel long non-coding RNAs that could elucidate the phenotypical characteristics of non-small cell lung cancer (NSCLC), inspired by the work of Tièche et al. (1). This study focused on human lung cancer cells (lineage A549) and yielded a list of potential novel non-coding genes.

The investigation delved into modern bioinformatics techniques, emphasizing de novo reference-guided transcriptome assembly coupled with differential gene expression analysis. Analysis of Next-Generation transcriptomic sequencing data involved mapping and assembly to identify novel, yet unannotated, intergenic and non-coding RNAs, with a specific focus on long non-coding RNAs (lncRNAs).

These molecules are intriguing due to their recognized roles in regulatory mechanisms, and suspicions regarding their involvement in lung cancers. Among the 60,108 genes identified in the assembly, 12,128 are novel transcripts, with approximately 95.31% believed to be non-coding, and 213 situated in intergenic positions.

Introduction

Gene investigation, a foundational approach to address biological questions, has exponentially advanced in recent decades due to a deeper understanding of genes' roles in cellular regulation. The evolving comprehension of the genome's structure and regulatory networks underscored the demand for faster, more reliable methods in DNA sequence retrieval.

Next-Generation Sequencing (NGS) has enabled the conversion of genomic data (DNA) into machine-readable text files (FASTQ files), enabling computer parsing and analysis. Propelled by ongoing discoveries, new questions have emerged, stimulating the development of techniques to probe various biological molecules, including RNA (DNA transcripts).

Various biological domains leverage high-throughput machines, with oncology emerging as a prominent field. Cancers, rooted in gene mutations disrupting critical cellular checkpoints, showcase the (over-)expression of oncogenes and the under-regulation of tumor suppressor genes, transforming healthy cells into cancerous entities. Understanding the intricate networks altered in cancer cells is crucial for advancing research in new therapeutics, potentially extending the life expectancy or even providing a cure for cancer patients.

Lung cancer, responsible for over 80% of global cancer-related deaths, predominantly comprises non–small cell lung cancers (NSCLCs) (1), emphasizing the sustained relevance of researching these specific cancer cells. While protein-coding genes dominated recent research, the plateauing life expectancy with new therapeutics has prompted researchers to expand exploration beyond coding RNAs. Long non-coding RNAs (lncRNAs), known for diverse regulatory roles, have become a focal point as potential oncogenes or tumor suppressor genes.

Cancer represents a intricate system of malfunctioning pathways enabling the potentially endless growth of tumor masses, eluding immune attacks, creating complex structures, and manipulating conditions to steal nutrients. Simultaneously, cancers exhibit the capacity to invade and proliferate in various tissues.

Adding to the complexity is intratumoral heterogeneity, where cells within the same tumor mass adopt different phenotypes based on their position and role within the cancer. For example, the epithelial-to-mesenchymal transition (EMT) generates holoclonal (epithelial), meroclonal, and paraclonal (mesenchymal and stem-like) cells, creating high differentiation that enhances defense against therapeutics and complicates the fight with chemotherapeutics.

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 (1_1, 1_2, 1_5) and the other from control -healthy- cells (P1, P2, P3).
The RNA sequencing has been performed with Illumina TruSeq (2).

FASTQC pipeline (version 0.11.9) (3) allowed graphical visualization on HTML files of basic reads information and quality. Each cell sample is represented by two FASTQ files (due to their size they are stored in gzipped), one for the forward strand and the other for the reverse strand (paired end sequencing).

Reads mapping and reference indexing were performed with the HISAT2 module (2.2.1) (4) against the comprehensive human genome annotation reference GRCh38 (version November 2024). The resulting SAM files were then converted into their respective binary version (BAM) using Samtools (1.10) (5).

The transcriptome was assembled with STRINGTIE (1.3.3b) (6) set with –rf strandness, producing six GTF files representing the transcriptome for each sequenced sample, which were then later merged with the option –merge. Assembly was guided by human annotation of GENCODE version 44 (7).

The complete transcriptome was prepared for gene quantification with Kallisto (0.46.0) (8) and Cufflinks (2.2.1) (9), 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, through a pseudoalignment system. 600 bootstrap were used, but very similar results can be achieved even with less than 100. RF stranded was specified to keep track of read sense.

Differential gene expression was analyzed using R Studio (Version 2023.09.0+463) with the Sleuth package (0.31.0) (10), 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 (11) 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 FANTOM5 CAGE (12) 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) (13) 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.

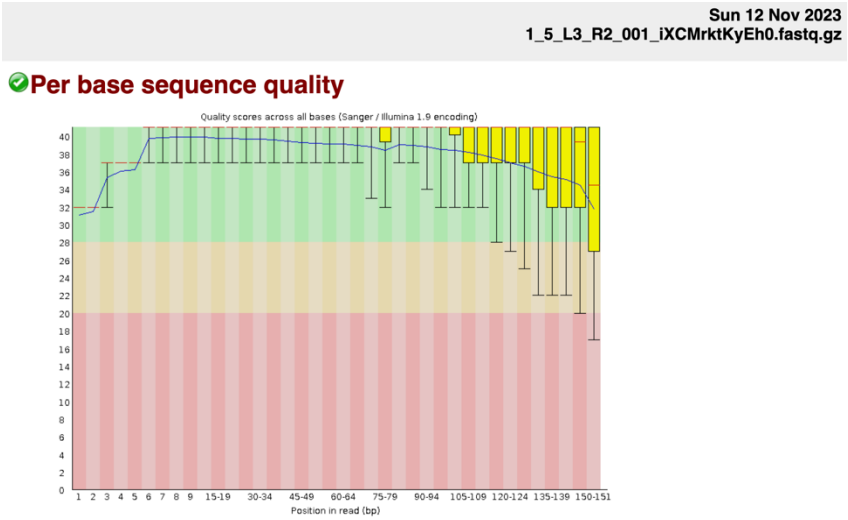

Figure 1: Hololonal cell 1_5 RNA-sequencing (R2) per base sequence quality statistic

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 output example), indicating very good alignment for each sample (70% is considered sufficient). 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 to other colleagues 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.

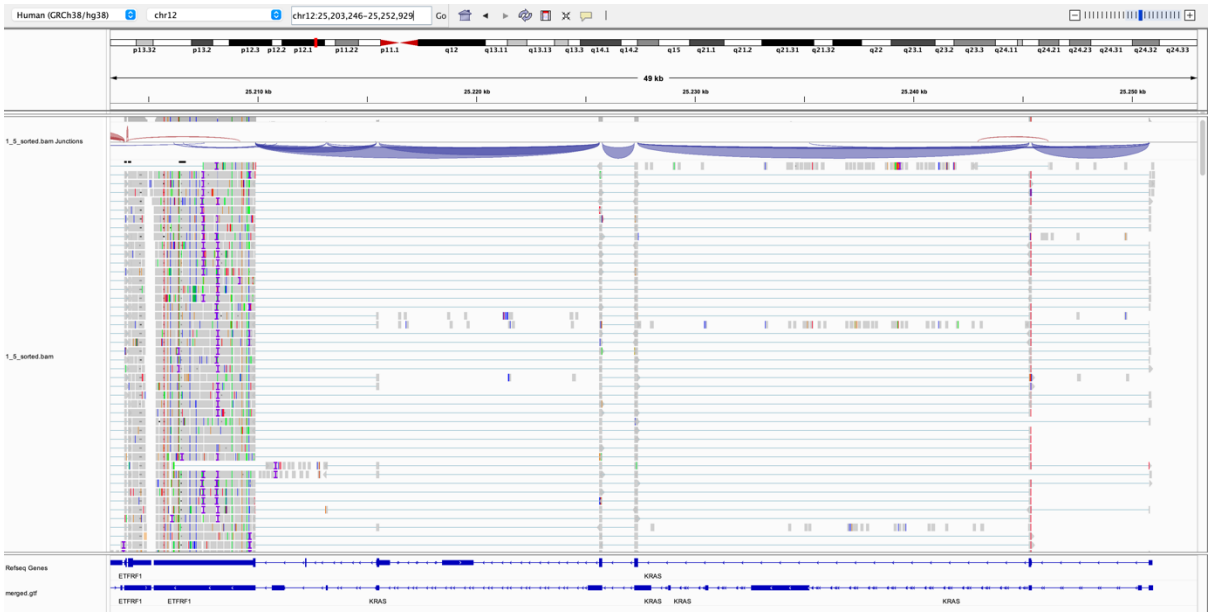

Figure 2: KRAS gene in 1_5 holoclonal sample portraying correct strandness when compared to GRCh38 reference

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, important to quantify the difference in expression between cancer cells and control.

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 can be demonstrated by summing all elements in the fifth column (tpm) of one of the *abundance.tsv* files generated by Kallisto. The final result yields 10^6 total transcripts.

	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.

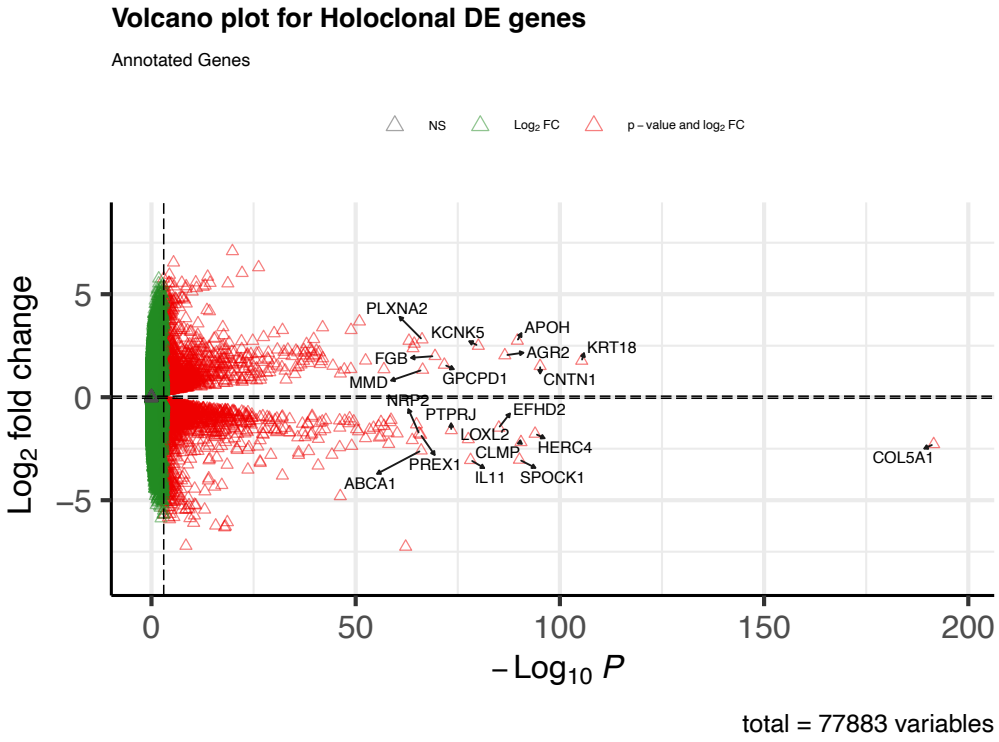

Figure 3: Volcano plot representing differentially expressed annotated genes

With the volcano plot above, the 20 most differentially expressed genes are shown, and interestingly, the majority of those represent in some way cell's molecules involved in tissue and matrix adhesion, typical of epithelial phenotypes.

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.

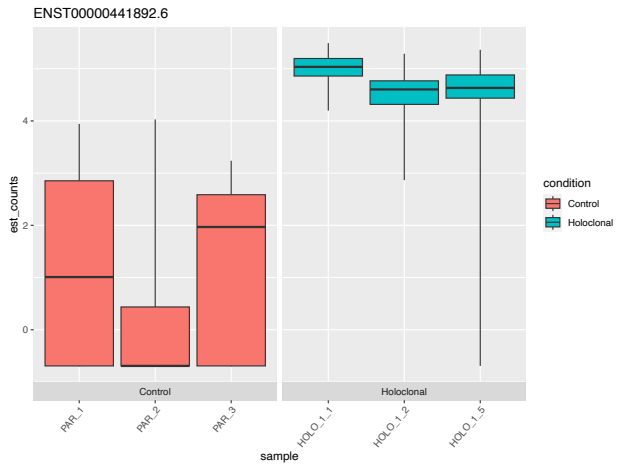

Figure 4: CDH17 (ENST00000441892.6) expression in control and holoclonal sample

Interestingly, other mesenchymal markers such as 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 its 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, 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

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

References

1. Tièche, Colin Charles, Yanyun Gao, Elias Daniel Bührer, Nina Hobi, Sabina Anna Berezowska, Kurt Wyler, Laurène Froment, et al. “Tumor Initiation Capacity and Therapy Resistance Are Differential Features of EMT-Related Subpopulations in the NSCLC Cell Line A549.” Neoplasia 21, no. 2 (February 2019): 185–96. https://doi.org/10.1016/j.neo.2018.09.008.

2. Illumina TruSeq Stranded RNA https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-stranded-total-rna.html

3. FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

4. HISAT2: https://daehwankimlab.github.io/hisat2/manual/

5. Samtools: https://www.htslib.org/doc/

6. STRINGTIE: https://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

7. GENCODE: https://www.gencodegenes.org/human/

8. KALLISTO: https://pachterlab.github.io/kallisto/source

9. Cufflinks: https://cole-trapnell-lab.github.io/cufflinks/manual/

10. Sleuth: https://pachterlab.github.io/sleuth_walkthroughs/trapnell/analysis.html

11. Bedtools: https://bedtools.readthedocs.io/en/latest/

12. FANTOM5 CAGE: https://fantom.gsc.riken.jp/5/

13. CPAT: https://rna-cpat.sourceforge.net/