Annotation and characterization of lncRNAs in Lung Cancer

Mansour Faye, Fall 2023

Abstract

Lung cancer is the first cause of cancer-related mortality, and non-small cell lung cancers (NSCLC) make up more than 80% of them. Three distinct subpopulations were identified within the NSCLC cell line A549, with one in particular showing high expression of cancer stem-cell (CSC)-related mRNAs and protein markers. CSCs have been known to be resistant to most conventional therapeutics, making them promising targets to improve cancer therapy. As the implication of lncRNAs in cancer stemness and tumorigenesis is becoming clearer in recent years, we decided to identify protein-coding and lncRNA genes that were differentially expressed between the holoclonal subpopulation and their parental line.

Introduction

Lung cancer is the first cause of cancer-related mortality, and non-small cell lung cancers (NSCLC)[Ettinger, Tièche] make up more than 80% of them. The A549 cancer cell line, which was isolated from lung carcinoma tissue by Giard et al.[Giard] has been used for nearly 50 years to study NSCLCs. Three distinct subpopulations were identified within the A549 parental cell line[Ye, Tièche]: holoclones, paraclones and meroclones. All showed distinct morphological characteristics, as well as different mRNA expression and DNA methylation profiles, and response to specific therapies[Tièche]. Holoclones in particular gave rise to colonies with high expression of cancer stem-cell (CSC)-related mRNAs and protein markers[Ye]. CSCs have been known to be resistant to most conventional therapeutics[Batlle, Shibue], making them promising targets for the improvement of cancer therapy. In their 2019 study, Tièche et al.[Tièche] investigated the three aforementioned subpopulations at the mRNA expression and protein-level, as well as their migration capacity, and resistance to therapies inducing DNA damage and AXL inhibition. However, lncRNAs, which are known to be involved in tumorigenesis, in the maintenance of stemness during cancer development[Jiang], and many other aspects of cancer were not included in the study. These RNAs are not translated into proteins, but regulate gene transcription, translation and epigenetic modification[Jiang, Ulitsky]. As most of them are absent from high-quality annotations, detecting them requires additional analysis steps. In this here study, we wish to identify differentially expressed lncRNAs in the A549 cell line's parental and holoclonal cells.

Using RNA-seq data from holoclonal and parental colonies sequenced by Tièche et al., we conducted a reference-guided assembly, creating an annotation that included yet unidentified lncRNA transcipts. All transcripts' expression levels could then be quantified, in order to identify differentially expressed protein-coding and lncRNA genes. In order to identify novel lncRNAs, we estimated the protein-coding potential of novel transcripts using CPAT, which is based on open reading frame (ORF) features and hexamer frequencies[Wang].

Materials and Methods

Disclaimer: The scripts used for all data processing and analysis steps are available here.

Experimental data

The RNA-seq data was obtained by Illumina sequencing. Using TruSeq Stranded mRNA libraries prepared from parental and holoclonal cells (3 technical replicates each).

Reference data

The reference genome used here was the Genome Reference Human Consortium Human Build 38 (GRCh38 or hg38). The FASTA file was downloaded from the GENCODE website, along with GENCODE[Frankish]'s full comprehensive gene annotation.

The transcription start sites (TSS) BED file came from the FANTOM5 project's CAGE peaks[Forrest], version hg38_v9 available here.

The polyadenylation sites (PAS) BED file was retrieved from the University of Basel's PolyASite atlas[Herrmann] for Homo sapiens, version 2.0 available here.

Protein-coding potential assessment with CPAT[Wang] required a prebuilt in-frame hexamer frequency table and a prebuilt logistic regression model, which were both downloaded here.

Quality Control

The quality of the raw reads was checked using FastQC[Andrews] (ver. 0.11.9). All twelve outputs (two per-replicate, for paired-end reads) were merged into a single report using MultiQC[Ewels]. In addition to potential flags from fastqc, we paid particular attention to Phred scores and adapter sequences, and that the number of forward and reverse reads were identical. The Phred score graph was downloaded directly from the multiQC report. If need be, reads could be filtered and/or trimmed using fastp[Chen].

Mapping

The reads were mapped to the reference genome using the STAR[Dobin] (ver. 2.7.9a) RNA-seq aligner. We first generated an index for the reference genome using the "genomeGenerate" mode, before running the alignment with the "alignReads" mode. Fastq files were submitted by pairs (paired-end reads), with forward reads first and reverse reads second. The output format was a BAM file which was sorted by coordinate. The sorting was necessary to then index and visualize the BAM file in IGV[Robinson], checking whether reads are on the right strand and whether they align with exons. Here, we only indexed and visualized one of the 6 BAM files, as all runs had identical parameters.
We also extracted the total number of reads, aligned reads and uniquely aligned reads from the log files using awk, allowing us to check alignment rates for each BAM file.

Transcriptome assembly

As we wished to quantify the expression of non-annotated lncRNAs, we needed to generate our own gene annotation. A reference-guided assembly was therefore run once per replicate using StringTie[Pertea] (ver. 1.3.3b), and the six output files merged into one final GTF file. We used the previously mentioned GENCODE gene annotation as a reference, and specified "—rf" strandedness given that the first read comes from the reverse strand in TruSeq Stranded mRNA libraries.
We then checked the number of genes, exons, transcripts, novel transcripts and single-exon transcripts in the merged annotation file. We also generated a mapping of gene name to transcript ID for the differential expression step.

Quantification

In order to quantify the expression of the genes identified in the previous step, Kallisto[Pimentel] (ver. 0.46.0) needs a FASTA file of these target sequences. We generated the FASTA file from our merged GTF file and the reference FASTA, using the "gffread" function from the Cufflinks[Pertea] (ver. 2.2.1) package. Kallisto relies on a pseudo-alignement of the fastq files to a reference sequence for quantification. Using our own annotation to generate this reference is what allowed quantification of the newly identified transcripts (which include the lncRNAs).

This FASTA was then indexed using kallisto and the quantification was run for each pair of fastq files, with the "—rf-stranded" option. Kallisto outputs the abundance estimates (counts and TPM) in plain-text and in HDF5 format. The HDF5 file contains the bootstrap summary and is read by sleuth[Pimentel] for differential expression. Information regarding the number of detected (counts > 0) transcripts was extracted using awk.

Differential Expression

The Differential expression (DE) analysis was done using the R package sleuth[Pimentel] (ver. 0.30.1). The abundance estimates were separated into two groups (holoclonal and parental) with three technical replicates each. Sleuth compared count estimates between the two groups for each transcript, and automatically applied the Benjamini-Hochberg procedure to control the false-discovery rate. The analysis was run at the transcript-level first, and the p-values were then aggregated to also obtain gene-level DE results. Results for all genes and transcripts with a q-value below 0.05 were saved and uploaded to the GitHub repository. The wald test was run to obtain the log2 fold change (β) estimates used in the volcano plot. Plots were made using sleuth's built-in *plot_volcano* and *plot_bootstrap* functions.

lncRNA Transcripts Validation

Not all of the differentially expressed novel transcripts identified in the previous step would make robust lncRNA candidates. We therefore checked if their transcription start (TSS) and/or end sites overlapped with known TSS[Forrest] and/or poly-adenylation sites[Herrmann]. We also used the Coding Potential Assesment Tool[Wang] (CPAT ver. 1.2.4) to identify which of the novel transcripts were likely to be non-coding.

First, two BED files were made from the merged annotation file: one containing a 100 bp window around each of the novel transcripts' start sites and another with the same window around end sites. We then used *bed intersect* from the BEDTools software to identify the transcripts that overlapped with CAGE peaks and PolyA sites.

CPAT was run with the prebuilt hexamer table and logistic regression model, and using the FASTA file we generated during the quantification step. All transcripts whose coding probability fell below the threshold of 0.364 were marked as non-coding.

We then merged the results from these last three steps with the DE results, and selected the 10 best lncRNA candidates according to the following criteria: Non-coding transcripts with validated start and end sites rank the highest. Follow those that only have their start or end site validated. Non-coding transcripts whose start and end site both aren't validated come third, and the coding transcripts are not considered. Within each category, candidate transcripts are ranked according to their q-value.

Results

Quality Control

Between 30 and 35 million pairs of 140-150 bp long reads were sequenced for each replicate (Figure 2). FastQC issued warnings for both sequence content and sequence duplications, but this is common for RNA-seq libraries. FastQC does not take paired-end reads into consideration, and the values themselves are not alarming. No other aspect stands out, and mean quality scores are satisfying too (Annex 2).

Mapping

For each replicate, more than 99% of reads were successfully aligned, with approximately 94% of all reads mapping only once. Visualization of the BAM files in IGV showed that most reads aligned with exons, and were present on the expected strand.

Replicate	Total reads	Total mapped reads	% mapped reads	Uniquely mapped reads	% Uniquely mapped reads
Holoclonal 1	34687742	34408545	99.1951	32657108	94.146
Holoclonal 2	34391514	34108579	99.1773	32449047	94.3519
Holoclonal 3	32059284	31777315	99.1205	30218031	94.2567
Parental 1	32840352	32575015	99.192	30863695	93.981
Parental 2	32894526	32630971	99.1988	30774184	93.5541
Parental 3	33414876	33156636	99.2272	31388986	93.9372

Figure 2: Number or reads and alignment rates. Percentages were calculated with respect to the total number of reads.

Transcriptome assembly and quantification

The merged annotation contains 69'628 genes, 234'943 transcripts and 1'548'076 exons. 26'622 of these transcripts are novel, and only 5.1% of them were single-exon transcripts. This last number rises to 12.7% for non-novel transcripts, in accordance with what can be found in the litterature[Kangueane].

Kallisto estimated counts for all 234'943 transcripts in each replicate, and quantified expression in transcripts per million (TPMs). It detected expression of 107'460 transcripts per replicate on average, and ~9% of them were novel transcripts. The sum of all TPM values equals a million for all replicates.

DE of known genes

Out of the 17'229 known genes analyzed by sleuth, 5791 have a q-value below the 0.05 threshold. In accordance with Tièche et al.'s results, the lung CSC markers ABCG2, EpCAM, SOX2 and KLF5 are overexpressed in holoclonal cells (see Figure 3 for q-values). The immuno-modulators PD-L1 and PD-L2 are also downregulated in holoclones. Regarding known genes, other results align with findings reported in the previously-mentioned study.

gene name	# of aggregated transcripts	p-value	q-value	up/down regulated
PD-L1	2	3.62E-39	3.78E-37	Downregulated
PD-L2	1	5.36E-09	6.55E-08	Downregulated
SOX2	1	1.11E-27	6.40E-26	Upregulated
KLF5	3	5.51E-07	4.88E-06	Upregulated
ABCG2	4	1.22E-09	1.61E-08	Upregulated
EPCAM	3	3.09E-17	8.69E-16	Upregulated

Figure 3: A few annotated differentially expressed genes. The p-values of the genes with multiple transcripts were aggregated by Sleuth. Full table of results can be found on Github along with the corresponding transcript-level results.

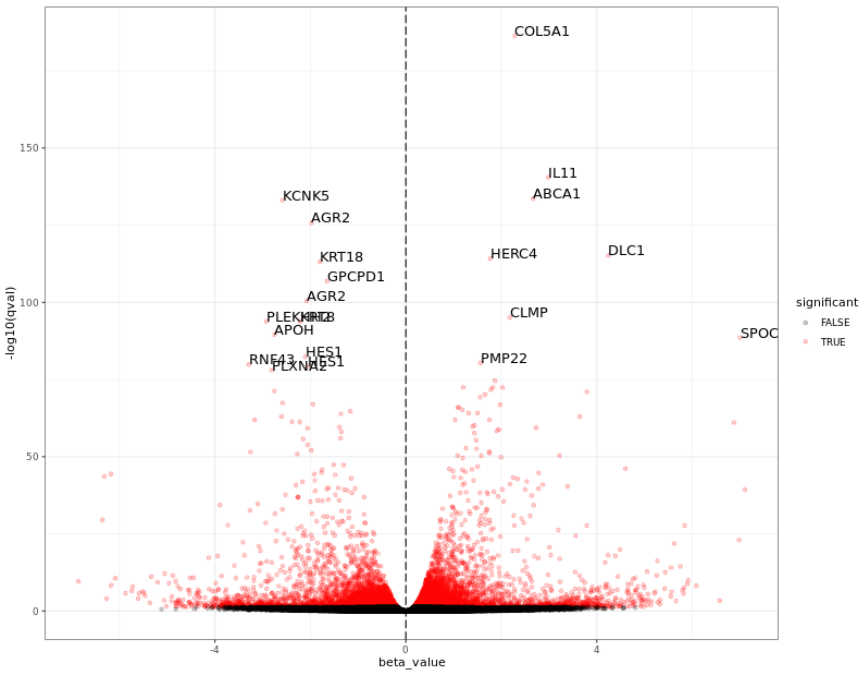


Figure 4: Volcano plot showing transcript-level log2 fold change (β) and q-values. Note that the

Differentially expressed novel transcripts

Out of the 3773 novel transcripts analyzed by sleuth, only 140 have a q-value below the 0.05 threshold. Out of these, 125 are upregulated in holoclonal cells, and 15 are downregulated. The validity of these transcripts is reviewed in the next section.

Validation of lncRNA transcripts

The novel transcripts' 5' and 3' annotations overlapped with CAGE and PolyA sites for 1294 and 1260 transcripts respectively. Meaning that approximately 5% of all novel transcripts have a validated end or start site. Only 22 of them have both valid start and end sites.

CPAT results show that 13'737 (51%) of the novel transcripts are noncoding, while 84'253 (40%) of known genes would be considered noncoding (see Annex 1). Out of the 140 differentially expressed novel transcripts, 54 are noncoding and only three have a validated start or end site. They are listed in Figure 5, with their respective q-value and log2 fold change (β). Note that the model used in Sleuth used the holoclonal cells' expression levels as a reference. This means that a positive value points to a transcript *downregulated* in holoclonal cells.

Transcript ID	q-value	β	PolyA	CAGE	Single exon	Coding
MSTRG.12388.2	0.00438918	-2.90270586	FALSE	TRUE	FALSE	FALSE
MSTRG.10368.2	0.01107733	-0.65385462	TRUE	FALSE	FALSE	FALSE
MSTRG.32577.2	0.04221925	-1.5458053	FALSE	TRUE	FALSE	FALSE
MSTRG.16371.3	4.59E-35	-3.89238993	FALSE	FALSE	FALSE	FALSE
MSTRG.4049.2	3.89E-22	-0.8856563	FALSE	FALSE	FALSE	FALSE
MSTRG.8184.2	1.85E-17	-1.83993487	FALSE	FALSE	FALSE	FALSE

MSTRG.4049.4	2.08E-11	-1.2688271	FALSE	FALSE	FALSE	FALSE
MSTRG.28748.2	1.81E-08	-0.82063082	FALSE	FALSE	FALSE	FALSE
MSTRG.4051.12	1.55E-07	-0.77888071	FALSE	FALSE	FALSE	FALSE
MSTRG.4051.1	1.07E-06	-0.75055607	FALSE	FALSE	FALSE	FALSE

Figure 5: Ten best lncRNA candidates, ranked. The complete list of candidates is available on GitHub.

Discussion

The existence of a previous study[Tièche] partly focusing on mRNA expression in the cells we're analyzing here makes us relatively confident in the quantification and DE part of our analysis. Indeed, the DE results described in said study are in line with what we observed in the known genes' expression levels. The novel lncRNAs' case is a bit more complex, as our final list of candidates depends not only on the quantification and DE, but also the mapping, assembly, and validation of transcripts. Mapping and assembly results seem both satisfying, with high alignment rates and the expected proportion of single-exon transcripts. CPAT results are reliable although small ORFs might slip under the radar. A tool based on evolutionary signatures like PhyloCSF might be relevant if we wanted to increase confidence in the coding potential results we already have. Overlap between the novel transcripts' annotations and CAGE/PolyA sites was calculated on a 100 bp window around the start and end sites. Depending on the accuracy of the CAGE peaks and PolyA sites, it might also be interesting to see if increasing this searching window heavily influences the number of hits. Out of our 54 candidate lncRNAs, we would suggest to functionally investigate at least the best three, as all but one of our five selection criteria are filled for each of them.

References

1. Akunuru, S., James Zhai, Q. & Zheng, Y. Non-small cell lung cancer stem/progenitor cells are enriched in multiple distinct phenotypic subpopulations and exhibit plasticity. *Cell Death Dis* **3**, e352–e352 (2012).

2. Andrews, S. FastQC: A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

3. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat Med* **23**, 1124–1134 (2017).

4. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).

5. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

6. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

7. Ettinger, D. S. *et al.* Non–Small Cell Lung Cancer. *Journal of the National Comprehensive Cancer Network* **8**, 740–801 (2010).

8. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

9. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

10. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

11. Giard, D. J. *et al.* In Vitro Cultivation of Human Tumors: Establishment of Cell Lines Derived From a Series of Solid Tumors2. *JNCI: Journal of the National Cancer Institute* **51**, 1417–1423 (1973).

12. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**, D174–D179 (2020).

13. Jiang, M.-C., Ni, J.-J., Cui, W.-Y., Wang, B.-Y. & Zhuo, W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* **9**, 1354–1366 (2019).

14. Kangueane, P. *et al.* A report on single exon genes (SEG) in eukaryotes. *FBL* **9**, 3262–3267 (2004).

15. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).

16. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687–690 (2017).

17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

18. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).

19. Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat Rev Clin Oncol* **14**, 611–629 (2017).

20. Tièche, C. C. *et al.* Tumor Initiation Capacity and Therapy Resistance Are Differential Features of EMT-Related Subpopulations in the NSCLC Cell Line A549. *Neoplasia* **21**, 185–196 (2019).

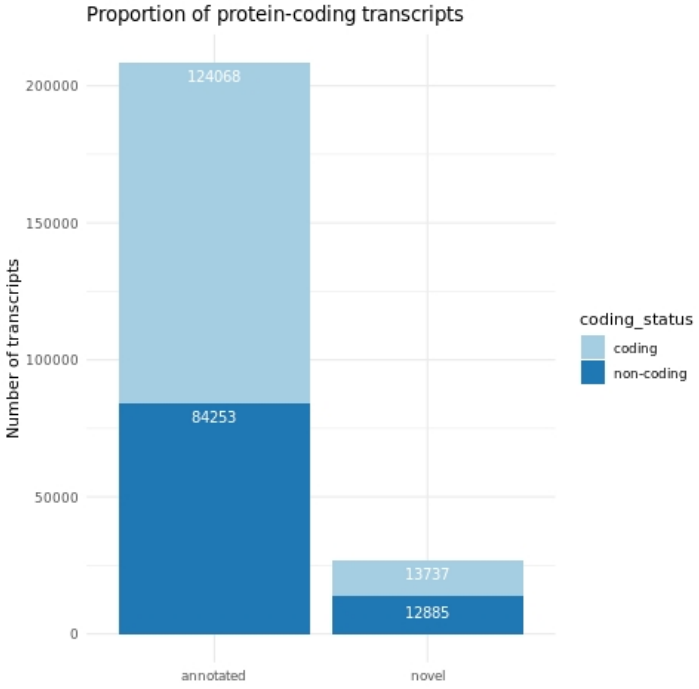
21. Ulitsky, I. & Bartel, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**, 26–46 (2013).

22. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41**, e74 (2013).

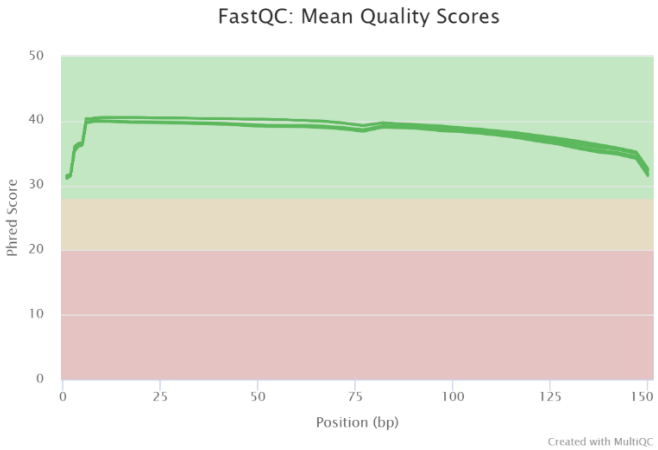
23. Ye, X.-Q. *et al*. Mitochondrial and energy metabolism-related properties as novel indicators of lung cancer stem cells. *International Journal of Cancer* **129**, 820–831 (2011).

Annex

Complete code can be found at: https://github.com/MansFaye/lncRNAseq22



Annex 1: Number of coding vs non-coding transcripts according to CPAT, for novel and known genes.



Annex 2: Mean Phred quality scores for all 12 FASTQ files.