Long Non-Coding RNA in NSCLC Cell Line A549

Kartik Kohli — 19-111-814

Jan 18, 2021

1 Abstract

Long Non-coding RNA (lncRNA) are transcribed from a number of genes in humans and may play a significant role in gene regulation. Mutations in lncRNA can promote tumorigenesis and metastasis [1]. This paper studies the sub-populations in the NSCLC Cell Line A549, namely heroclone, meroclone and paraclone. We further deploy an RNA-Seq pipeline, analyse the differentially expressed genes and deploys various filtering and ranking methods to find lncRNA genes that could serve as potential targets for drug therapy

2 Introduction

Lung cancer is one of the leading causes of cancer-related mortalities worldwide, causing more than two million deaths per annum, according to the latest World Health Organisation estimate. While studying the cell line A549, three subpopulations were thoroughly described and characterized by Tièche et. al[6]. While holoclone cells had a stem-cell like phenotype, meroclones had a mesenchymal phenotype, paraclones had an intermediate phenotype.

The human genome has 3 billion base-pairs, of which only 2% encode proteins. It has been a long held belief, dating as far back as the 1960s, that the rest of the Human DNA that doesn't code for proteins is "Junk DNA". That is, it does not impart any fitness advantage to the organism and the genetic matter was passively accumulated over many millennia - nonsense genes, viruses that integrated into our genomes, none of which was relevant to the survival of the organism. Over the past decade, a lot of scientific research has brought these assumptions into question, including but not limited to microRNAs and long non-coding RNA.

We will focus on lncRNAs, which are perhaps the least well-understood products of transcripts from genomes. Although they don't code for proteins, they play a crucial role in various cellular and physiological processes [3]. LncRNAs have shown to influence cancer metastasis, mediate global gene repression in p53 response [7] and also play a key driver in some Melanomas. Here we present novel lncRNA which could have therapeutic value in targeting non-small cell lung cancer

3 Materials and Methods

The dataset includes 12 samples, 2 replicates for each sample (Forward & Reverse). There are 3 samples for each subclone - Holoclone, Metaclone and Paraclone. Finally, 3 samples for the Parent cell line, serving as control. In total, that makes 24 FASTQ files. Statistical Analysis of the data is present in the following subsections. An overview of the RNA-Seq pipeline is presented in Figure 1

3.1 Quality Control

Summary statistics for the replicates is generated using *stats* command from the *Seqkit* toolset. The data is presented in Table 2. FastQC version 0.11.9 is used for quality control of all replicates. Subsequently, MultiQC version 1.8 is used to aggregate all FastQC results. A snapshot of the Base Quality scores is presented in Figure 2

3.2 Mapping

HISAT2 is a popular tool for mapping of next generation sequencing reads. For the reference genome, GenBank Human reference genome - Hg38 was used. Feeding the FASTQ files to HISAT2 version 2.2.1, the reads were mapped in paired-end mode to the aforementioned reference genome and Binary Alignment Map (BAM) output was obtained. According to the library preparation method, the parameter *–rna-strandness* had to be set to RF

3.3 Transcriptome Assembly

To obtain the transcriptome meta-assembly, StringTie version 2.3 was used. Stringtie assembles RNA-Seq alignments into potential transcripts. For a reference guided assembly, the latest Gencode annotation was used in order to obtain General Transfer Format (GTF) file for each replicate. Individual GTF files were aggregated to obtain one merged meta-assembly GTF file.

3.4 Quantification

For Differential Expression analysis of genes, first, quantification of the reads was required. Kallisto version 0.46.0 was used for this step. The transcriptome assembly files for respective replicate was fed to Kallisto to obtain abundance estimates for the various transcripts. Kallisto uses bootstrap sampling to obtain probabilistic pseudo-alignment estimates. 10 bootstrap samples were used in our analysis. The metric used was Transcript per Million (TPM). The counts for various transcripts were normalized across the replicate

3.5 Differential Expression Analysis

DE Analysis is a crucial step in the RNA-Seq pipeline. Popular tools such as Sleuth can easily be used in conjunction with Kallisto for Differential Expression. Sleuth version 0.30.0 was used in our pipeline. Sleuth builds a generalized linear model to model the true estimates. Wald Test method was used to obtain differentially expressed genes using a FDR adjusted pvalue of 0.05. Volcano plots for the same can be seen in Figure 5

3.6 Integrative Analysis

To filter and prioritize the candidate genes, further analysis was performed. Using FANTOM CAGE (Cap Analysis of Gene Expression) Clusters, quality check of our end annotations was performed. To find non coding transcripts, a protein annotation tool CPAT version 3.0.4 was used to determine protein coding potential of the novel transcripts in our Transcriptome. Lastly, we filter for intergenic transcripts, as they have a higher likelihood of influencing protein coding genes. These metrics and filters were combined to obtain a ranked list of candidates

3.7 Ranking

Correct end annotation of 5' cap and 3' tail gives a higher confidence that the transcript annotation was correct. These transcripts were filtered out and picked. Estimating the coding potential using CPAT, a cutoff of 0.364 (for humans) was used to obtain non-coding candidates. Among these candidates, those that were found to be intergenic formed the candidate list. The candidates were ranked using various metrics such as coding potential, number of exons etc as described in results

4 Results

As can be seen in Table 2, average sequence length in all FASTQ files is roughly 150. The number of sequences in the respective paired reads is the same, as a sanity check. From the MultiQC output in Figure 2, we can see that the base quality score stays in the green zone i.e. > 30 for the entire read length of all the replicates. High sequence duplication was found, which taking into context the exercise of RNA-Seq, is naturally admissible. There was no adapter contamination either. Inferring from all these sanity checks, there was judged to be no need for cleaning up

Mapping our reads to the reference genome using paired-end mode, a mean alignment rate of 97% was achieved. Building a transcriptome assembly using StringTie and Gencode reference annotation, the results are tabulated in Figure 3. A total of 56,096 genes were identified, of which 30,562 were novel genes. Subsequently, 255,932 Transcripts were identified, out of which 19,152 were novel transcripts. As a sanity check, we identified 362 single exon transcripts, which we deprioritize in our analysis. Single exon transcripts are less likely to play a significant role in gene regulation, which are usually impacted by dysregulation in alternative splicing

Using these quant aggregates from Kallisto, Differential Gene expression was performed using Sleuth. Using the Parent cell line as control, expression level was contrasted with the three sub-populations. This experiment design was used to obtain differential expressed genes between each subpopulation vs parent. Using a False Discovery Rate cut-off of 0.05, a filtered list of genes and transcripts was obtained. A sanity check using behaviour of known genes such as VIM and THY, the expression was compared as against the original publication [6]

Predicting the coding potential of our novel transcripts using CPAT software, it was observed that 58% of the ORFs fall below the cut-off of 0.364 for non-coding sequence. 6,269 of the novel transcripts have a correct end annotation and 1,441 are intergenic. Aggregating the results, an inner-join performed from the three outputs gives us a filtered list of 440 novel gene candidates. The venn diagram for this can be seen in Figure 4

Taking the candidate genes from Differential Expression Analysis, filtering the ones that pass the criterions of Step 6, we get a final list of candidates. Using features such as number of exons and beta i.e. log-fold-change value. With these features, the ideal candidates' sequence we blast and come up with a final candidate list that could be used as potential therapeutic targets. Results are summarized in Table 1

5 Discussion

Using the RNA-Seq pipeline, more than 440 non-coding sequences were identified. Prioritizing by exon-structure and differential expression, we proposed ranked list of novel lncRNA gene candidates as described in Table 1. One interesting result would be identification of STRG.5861 gene in our ranked list, which is an annotated intergenic non-coding RNA (LINC02210) in the GRCH38 genome.

This gives us confidence that our ranked candidates successfully meet all the filtering critera and serves as a sanity check for our aggregation methodology.

The top candidate has a > 30 fold change in expression levels in the meroclone subpopulation and has a multi exon structure. The gene bears similarity to sequencing influencing the SRC Gene, which is one of the primordeal candidates in tumorogenesis. Other genes presented here also have a multi-exon structure and $30 - 50$ fold change in expression levels. All of them share similarities with genes which play a role in Chromosomal Rearrangement Disorders or some forms of Cancer.

Gene ID	Exons	Beta	Clone	Blast query results
STRG.12181	2	-5.58	mero	Promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation
STRG.6581	5	4.88	holo	Targeting the Vav1/miR29b axis as a potential approach for treating selected molecular subtypes of triplenegative breast cancer.
STRG.10807	19	-4.93	holo	Functional involvement of RINF, retinoid-inducible nuclear factor (CXX in normal and tumoral human myelopoiesis
STRG.5861	2	-6.56	para	Homo sapiens long intergenic non-protein coding RNA 2210 (LINC02210), transcript variant 5, long non-coding RNA

Table 1: Candidate Genes

6 Bibliography

1. Bhan, A. & Mandal, S. S. LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. Biochim Biophys Acta 1856, 151–164 (2015).

2. Faghihi, M. A. et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14, 723–730 (2008).

3. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227 (2009).

4. Vitiello, M., Tuccoli, A. & Poliseno, L. Long non-coding RNAs in cancer: implications for personalized therapy. Cell Oncol (Dordr) 38, 17–28 (2015).

5. Khalil, A. M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106, 11667–11672 (2009).

6. Tièche CC, Gao Y, Bührer ED, Hobi N, Berezowska SA, Wyler K, Froment L, Weis S, Peng RW, Bruggmann R, Schär P, Amrein MA, Hall SRR, Dorn P, Kocher G, Riether C, Ochsenbein A, Schmid RA, Marti TM. Tumor Initiation Capacity and Therapy Resistance Are Differential Features of EMT-Related Subpopulations in the NSCLC Cell Line A549. Neoplasia. 2019 Feb;21(2):185-196. doi: 10.1016/j.neo.2018.09.008. Epub 2018 Dec 27. PMID: 30591423; PMCID: PMC6309124.

7. Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010;142(3):409-419. doi:10.1016/j.cell.2010.06.040

Appendix

Supplementary code and instructions can be found at github.com/thewayofknowing/lncRNA

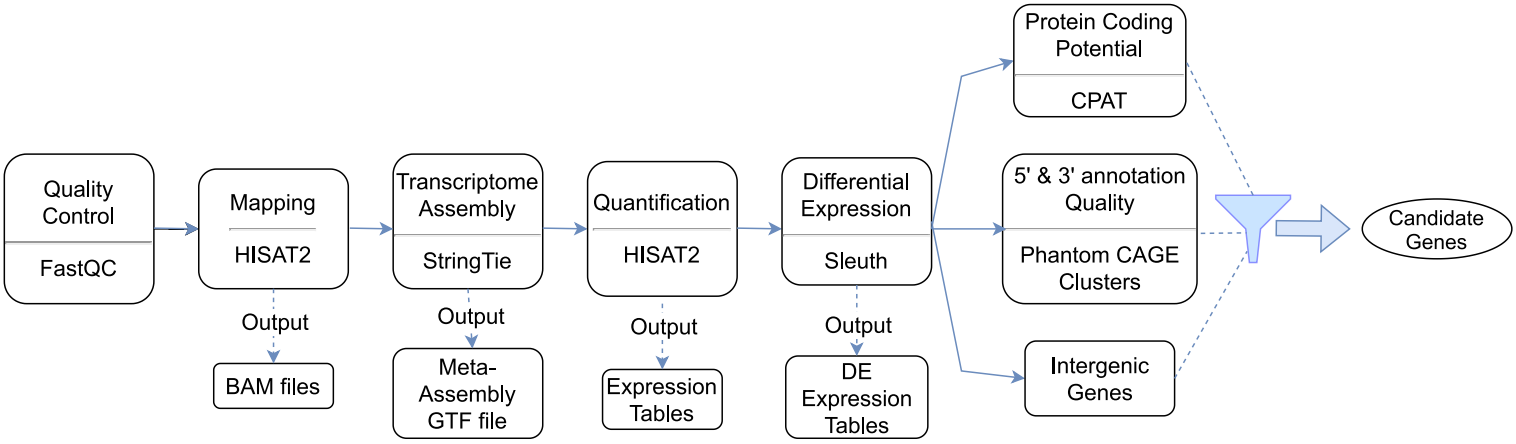


Figure 1: Pipeline

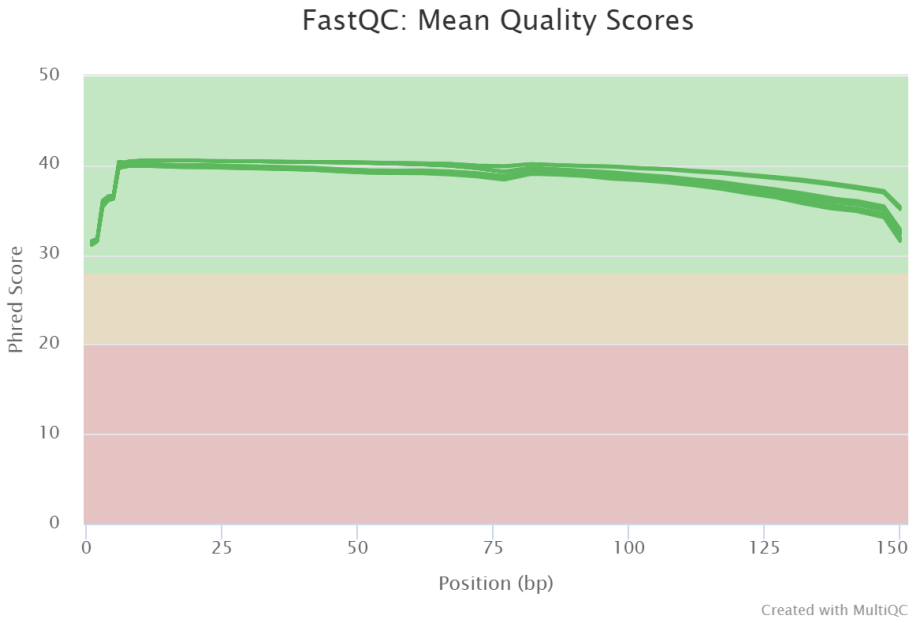


Figure 2: MultiQC Base Quality

filename	format	type	num_seqs	min_len	avg_len	max_len
fastq/1_1_L3_R1_001_ij43KLkHk1vK.fastq	FASTQ	DNA	34,687,742	35	150.4	151
fastq/1_1_L3_R2_001_qyjToP2TB6N7.fastq	FASTQ	DNA	34,687,742	35	150.4	151
fastq/1_2_L3_R1_001_DnNWKUYhfc9S.fastq	FASTQ	DNA	34,391,514	35	150.5	151
fastq/1_2_L3_R2_001_SNLaVsTQ6pwl.fastq	FASTQ	DNA	34,391,514	35	150.4	151
fastq/1_5_L3_R1_001_iXvvRzwmFxF3.fastq	FASTQ	DNA	32,059,284	35	150.3	151
fastq/1_5_L3_R2_001_iXCMrktKyEh0.fastq	FASTQ	DNA	32,059,284	35	150.2	151
fastq/2_2_L3_R1_001_77KSDZXkzpN2.fastq	FASTQ	DNA	31,771,713	35	150.2	151
fastq/2_2_L3_R2_001_2oenLbeyyPvS.fastq	FASTQ	DNA	31,771,713	35	150.1	151
fastq/2_3_L3_R1_001_DZmuiRvA53zD.fastq	FASTQ	DNA	33,140,498	35	150	151
fastq/2_3_L3_R2_001_bW28atsMceL2.fastq	FASTQ	DNA	33,140,498	35	149.9	151
fastq/2_4_L3_R1_001_ezH0ldTDxUdi.fastq	FASTQ	DNA	34,494,035	35	150.3	151
fastq/2_4_L3_R2_001_5qJL43xGflsJ.fastq	FASTQ	DNA	34,494,035	35	150.2	151
fastq/3_2_L3_R1_001_DID218YBevN6.fastq	FASTQ	DNA	35,438,437	35	150.4	151
fastq/3_2_L3_R2_001_UPhWv8AgN1X1.fastq	FASTQ	DNA	35,438,437	35	150.3	151
fastq/3_4_L3_R1_001_QDBZnz0vm8Gd.fastq	FASTQ	DNA	33,442,327	35	150.2	151
fastq/3_4_L3_R2_001_ng3ASMYgDCPQ.fastq	FASTQ	DNA	33,442,327	35	150.1	151
fastq/3_7_L3_R1_001_Tjox96UQtyIc.fastq	FASTQ	DNA	33,745,478	35	150.5	151
fastq/3_7_L3_R2_001_f60CeSASEcgH.fastq	FASTQ	DNA	33,745,478	35	150.4	151
fastq/P1_L3_R1_001_9L0tZ86sF4p8.fastq	FASTQ	DNA	32,840,352	35	150.5	151
fastq/P1_L3_R2_001_yd9NfV9WdvvL.fastq	FASTQ	DNA	32,840,352	35	150.4	151
fastq/P2_L3_R1_001_R82RphLQ2938.fastq	FASTQ	DNA	32,894,526	35	150.4	151
fastq/P2_L3_R2_001_06FRMIIGwpH6.fastq	FASTQ	DNA	32,894,526	35	150.3	151
fastq/P3_L3_R1_001_fjv6hlbFgCST.fastq	FASTQ	DNA	33,414,876	35	150.4	151
fastq/P3_L3_R2_001_xo7RBLLYYqeu.fastq	FASTQ	DNA	33,414,876	35	150.3	151

Table 2: Replicate Statistics

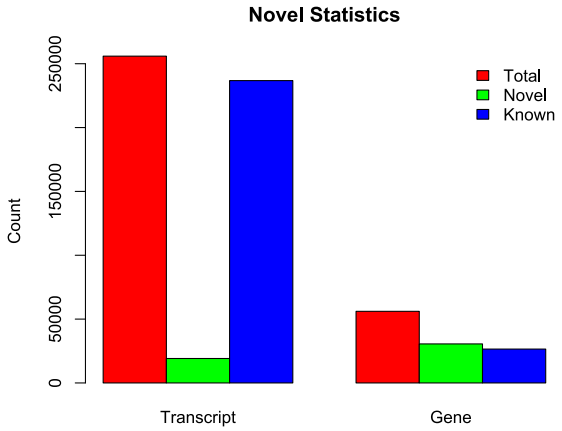


Figure 3: Transcriptome Statistics

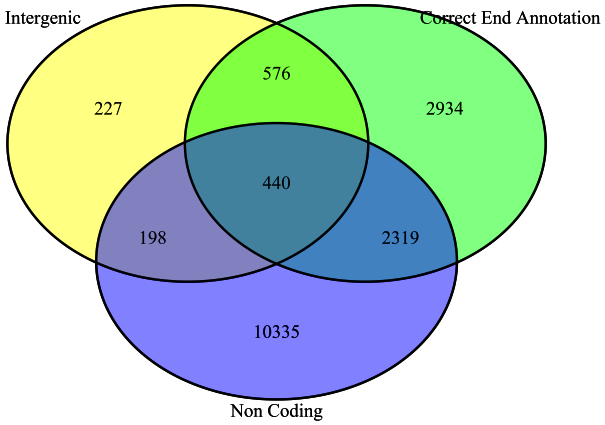
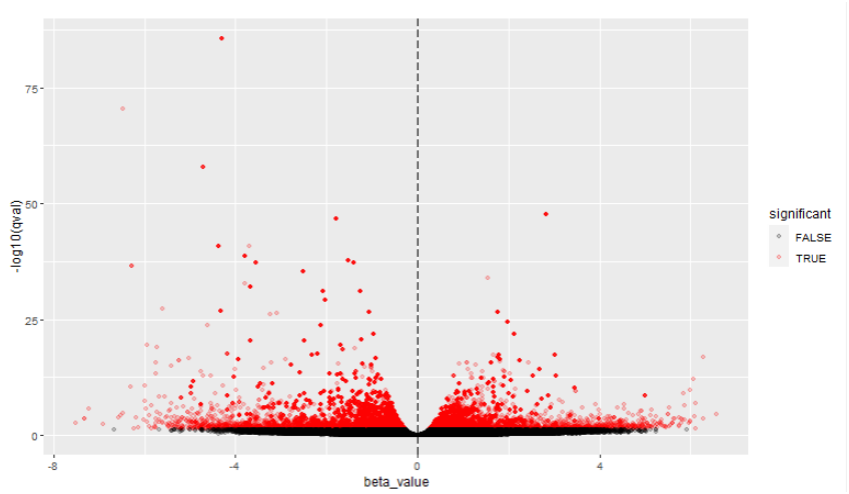
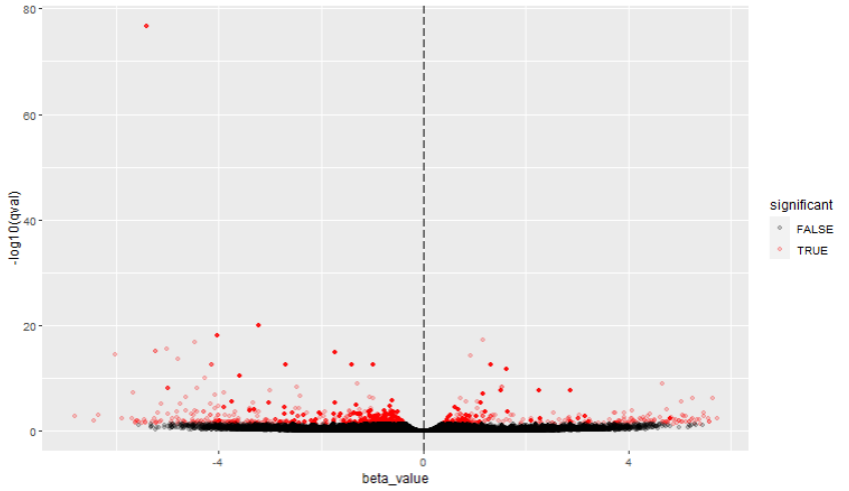


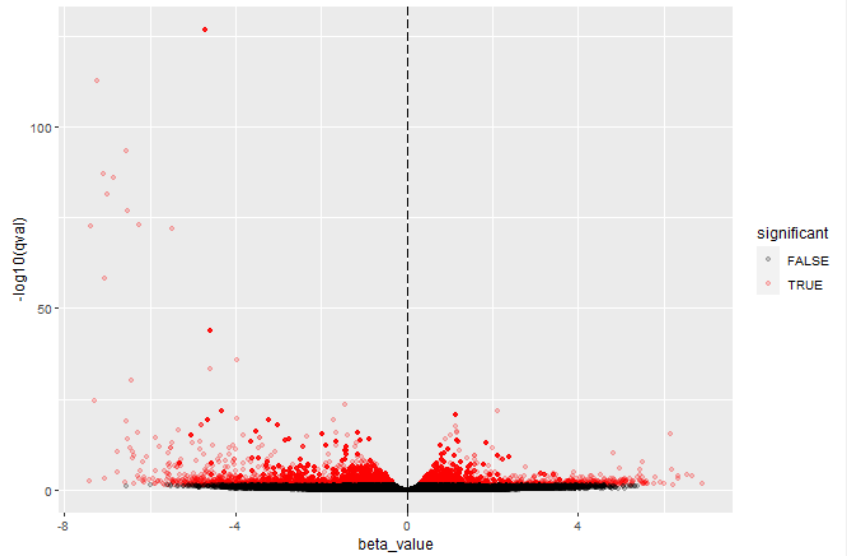
Figure 4: Integrative Analysis : Venn Diagram



(a) Holoclone



(b) Meroclone



(c) Paraclone

Figure 5: Volcano Plots