

# Enhancer RNAs - archive

## Quantifying and evaluating enhancer RNAs (eRNAs) in RiboErase and poly(A) RNA-seq data

### Table of content

<b>Background</b>	<b>2</b>
Broader picture of science in the area	2
Current state of the art (literature review) and its shortcomings	2
Specific issues/shortcomings to be addressed in the study	3
What needs to be done to overcome these shortcomings?	3
Aim of the study	3
Outline the results	3
How do these results enhance the understanding of specific issues	3
Conclusion	3
<b>Results</b>	<b>4</b>
Designing a pipeline for eRNA quantification	4
Comparing detectable eRNA counts from total RNA vs poly(A) RNA-seq datasets	4
Performing exploratory data analysis using PCA, rPCA (robust PCA), and t-SNE	5
<b>Methodology</b>	<b>5</b>
eRNA pipeline design	5
eRNA proportion estimation	6
Exploratory analysis	7
<b>Conclusion</b>	<b>7</b>
<b>Code location on R server</b>	<b>7</b>
To note, all the code related to eRNA project is here: /mctp/share/users/gondal/cptac3-rarercc	7
Code for eRNA quantification	7
Code for PCA, rPCA, t-SNE	7
Code for Heatmaps generation	7
Code for RPKM and proportion estimation	8
<b>Supplementary materials</b>	<b>8</b>
Bioinformatics presentation slides link	8
Potential future directions sheets link	8
Dropbox link	8

*This page is intentionally left blank*

# Background

## 1. Broader picture of science in the area

Enhancers function as distal DNA regions which enhance the transcription of target genes, usually through interaction with the gene promoter site. In 2010, Kim *et al.* reported that active enhancers are also transcribed into non-coding RNAs referred to as enhancer RNAs (eRNAs). eRNAs are reported to be tissue and cell lineage-specific making them excellent markers for enhancer activation. Numerous studies have also reported eRNAs to play a critical role in regulating gene transcription, however, the exact mechanism of eRNA function is still unknown.

According to the Functional Annotation of the Mammalian Genome (FANTOM) Project, almost 10% of total eRNAs are polyadenylated. These polyadenylated (polyA) eRNAs are frequently longer (almost 4 kb) and unidirectionally transcribed, whereas non-polyadenylated eRNAs are shorter (346 bases), unspliced, and bidirectionally transcribed. Many eRNAs are also associated with long non-coding RNAs and are reported to have clinical implications as potential target therapy in diseases such as cancer. Taken together, due to high heterogeneity in literature amongst the structural and functional role of eRNAs they remain a rich topic for exploration and future analysis.

## 2. Current state of the art (literature review) and its shortcomings

Numerous studies have developed eRNA quantification pipelines to investigate eRNA role in complex diseases as well as cellular regulation. Amongst these studies, in 2018 by Cheng *et al.* employed The Cancer Genome Atlas (TCGA) to quantify eRNAs from 9000 tumor samples in 33 cancer types. The study identified 15808 eRNAs and reported eRNAs to be positively associated with cancer aneuploidy. Although this study was instrumental in showing the functional role of eRNAs in cancer across large-scale patient samples, it remained limited in its analysis due to polyA selection and relatively low depth of TCGA RNA seq dataset. Similarly, in 2019 de Lara *et al.* reported the clinical utility of eRNAs in cancer by developing a large-scale computational pipeline for eRNA detection and quantification using multi-omics datasets from TCGA, CCLE, ENCODE, FANTOM, Roadmap Epigenomics, and 4D Nucleome projects as well as other pharmacogenomics data. The study showed a strong correlation between eRNA expression pattern and cancer lineage suggesting eRNA as critical diagnostic as well as prognostic markers for cancer therapy. Although these results can be employed for further functional investigations into eRNA for cancer therapy, it was again limited due to polyA selection which can only capture a subset of all eRNAs present in the cell. Recently, in 2020 Wu *et al.* reported a pipeline for eRNA quantification and expression analysis in lung cancer cell lines and tissues. They reported 15808 eRNAs and suggested the potential role of eRNAs as microproteins.

### **3. Specific issues/shortcomings to be addressed in the study**

Although these reports have demonstrated eRNA importance in cancer, their methods remain limited in quantifying all potential eRNAs in a cell since polyadenylated eRNAs are only a subset of the entire eRNAs population.

### **4. What needs to be done to overcome these shortcomings?**

To overcome this limitation, we employed The Clinical Proteomic Tumor Analysis Consortium (CPTAC) rare-RCC cohort which contains total RNA-seq data that can be useful in eRNA quantification and evaluation. The data can provide more information on eRNAs which could not be possible with the previous eRNA quantification pipelines which relied on only polyA data.

### **5. Aim of the study**

Here, we propose a computational pipeline for eRNA quantification to demonstrate that total RNA seq has more total eRNAs compared to polyA RNAseq data. The results from this study can be investigated further to help provide more insights into eRNA's role in transcription as well other functional aspects in cancer.

### **6. Outline the results**

Towards this aim, we implemented an eRNA quantification pipeline and identified 15159 eRNA thereby reproducing previous results. We also addressed issues of region filtering and depth normalization to correctly quantify eRNAs and showed that there is higher enrichment for eRNAs in the Ribo-Erase dataset (approx. 10-fold). We obtained a significantly higher number of detectable eRNAs in the riboErase dataset (790 vs 170). Our results also demonstrated that eRNA can be used to correctly cluster rare-RCC by histological types.

### **7. How do these results enhance the understanding of specific issues**

We observed that eRNAs represent a very small fraction of the total transcriptome (approx. 0.02%) and expression of the great majority of eRNAs very low (<2 reads) which makes eRNAs hard to disambiguate from other types of non-genic transcription.

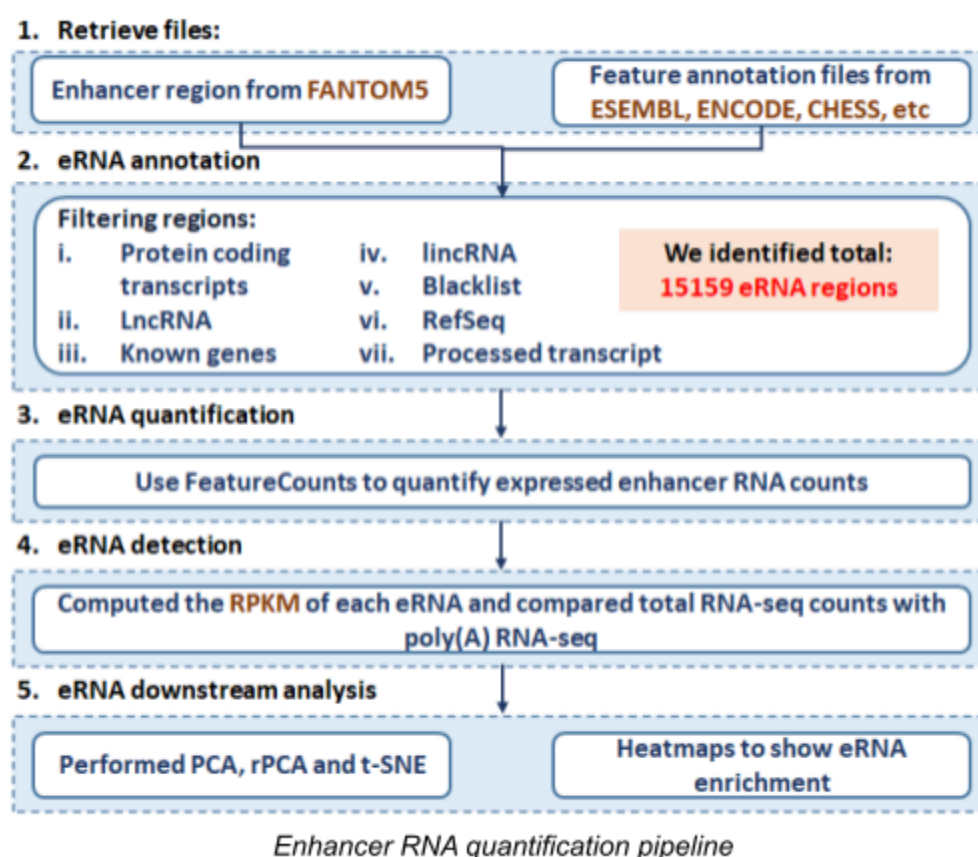
### **8. Conclusion**

In conclusion, our eRNA quantification pipeline is the first of its kind to quantify eRNAs from total RNA seq data detecting more eRNA than the previously reported studies. However, not all of our detectable eRNAs fit the classical definition of eRNAs from literature and might be noise or a by-product of the active gene transcription machinery. Therefore, we conclude that in order to study eRNAs in detail with highly probable eRNAs we would need to do some eRNAs specific assays such as BruSeq to be sure that regions we are quantifying are actually eRNAs.

# Results

## 1. Designing a pipeline for eRNA quantification

In this study, we report a computational pipeline for the identification and quantification of enhancer RNAs (eRNAs) expressions using the CPTAC3 dataset on rare-RCC patients. This pipeline can be divided into three salient stages: (i) eRNA region annotation, (ii) eRNA quantification, and (iii) eRNA detection. We obtained expressed enhancer annotation from The Functional Annotation of the Mammalian Genome (FANTOM) Project. Next, we added additional filters for example for protein-coding, lncRNA, and known gene regions, to remove overlapping regions. We obtained 15159 eRNA regions which we later used for quantification.



## 2. Comparing detectable eRNA counts from total RNA vs poly(A) RNA-seq datasets

To show that there are more detectable eRNAs in total RNAseq data compared to polyadenylated RNAseq we extracted polyA reads from TCGA's KICH (chromophobe patient cohort). To identify detectable eRNAs from polyadenylated as well as total RNAseq data, we mapped our 15159 eRNA region to CPTAC and TCGA's KICH RNAseq reads to characterize the expression landscape of eRNA in the rare-RCC cohort and TCGA's KICH. We observed

that there were a significantly higher number of total eRNAs in the riboErase dataset (790 vs 170).

### 3. Performing exploratory data analysis using PCA, rPCA (robust PCA), and t-SNE

To demonstrate that eRNA can be used to correctly cluster rare-RCC by histological types. We performed clustering based on eRNA expression. Performing classical PCA clusters showed a big papillary RCC cluster, however, after performing a more robust form of PCA we observed two main cluster formations for oncocytomas and papillary PCC. Similarly, we were able to obtain a similar form of clustering with t-SNE.

## Methodology

The objective of this project is to quantify eRNAs from total RNA-seq data from the CPTAC3 project and evaluate if we can get more total eRNAs from total RNAseq compared to polyA. Towards this aim, the following silent tasks were performed (1) pipeline for eRNA quantification, (2) eRNA reads proportion estimation, and (3) exploratory analysis on the resultant eRNA regions.

### 1. eRNA pipeline design

The pipeline can be divided into three stages: (i) eRNA region annotation, (ii) eRNA quantification (iii) eRNA expression detection

**eRNA annotation** - To annotate enhancer RNAs (eRNA) we extracted 63,285 enhancer regions from The Functional Annotation of the Mammalian Genome (FANTOM) Project. These regions were processed using a series of filters to remove overlapping (i) known genes, (ii) introns, (iii) lncRNA, (iv) protein-coding transcripts, and (v) blacklist regions. As a result, we identified 15159 eRNA regions with an average length of ~278 bp.

**eRNA quantification** - Aligned RNA reads for the Chromophobe renal cell carcinoma (chRCC) patient was fetched from the CPTAC project. All annotations were based on the GRCh38 reference genome. FeatureCounts were employed to quantify raw counts of eRNAs.

**eRNA detection** - To normalize the raw read counts we computed the RPKM of each eRNA. For that, we first computed the read depth for our sample using the following formula:

$$total\ reads = echo \$ (cat\ chRCC.fq|wc -l)/4|bc$$

The resultant read depth was 91,607,526 for CPTAC's chRCC patient. We next employed the RPKM formula to calculate normalized expression.

$$rpkm = 10^9 \times \left( \frac{reads\ mapped\ to\ eRNA}{(total\ reads \times eRNA\ length)} \right)$$

In which “reads mapped to transcript” are the raw reads from *featureCounts* result, and “total reads” is what we computed previously i.e read depth or library size, and “transcript length” is the length of our eRNA regions in bp.

In particular,

$$rpkm = 10^9 \times \left( \frac{\text{reads for each eRNA in chRCC patient}}{(91,607,526 \times \text{length of each eRNA})} \right)$$

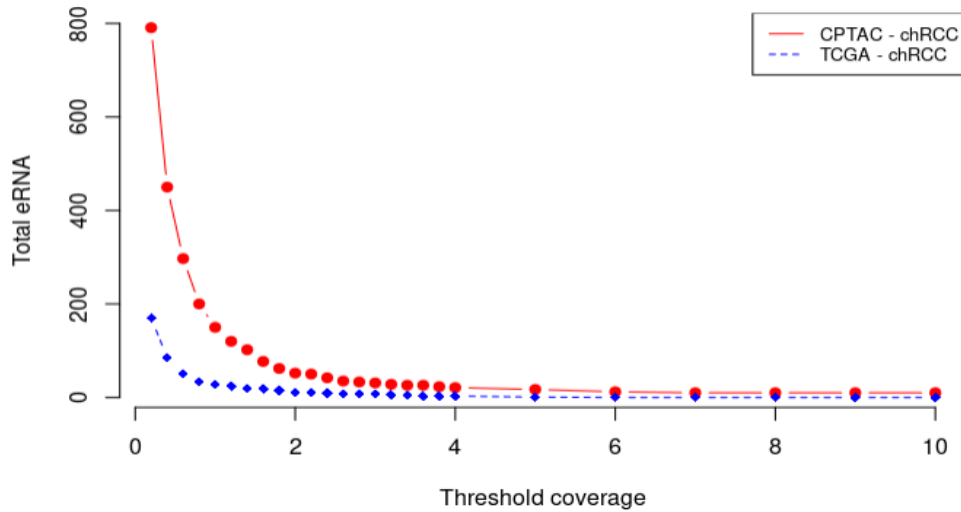
To evaluate how many detectable eRNA can be computed, we used a threshold coverage of 1 till 10. For that, the following computation was undertaken,

$$thresh\_total1 = \text{Normalized\_counts\_total\_1} > 1)$$

We did this for 1, 1.2, 1.4, 1.6 ... 10 to compare the outcome. Next, we evaluated the TRUE eRNA cases and stored them to count the detectable eRNAs.

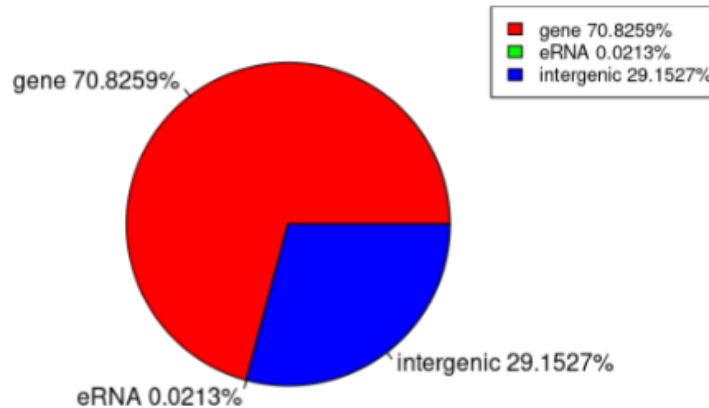
$$CPTAC\_enhancer1 = \text{length}(thresh\_total1[thresh\_total1 == TRUE])$$

A line graph was generated to view results for detectable eRNAs for different threshold coverage for CPTAC and TCGA's chRCC patients.



## 2. eRNA proportion estimation

To estimate gene, eRNA, and intergenic read proportions from total RNA reads, after quantification we summed the reads for the eRNA, gene, and intergenic regions separately and divided them by total reads for the three features. The results showed that only 0.02% of the reads counted towards eRNAs in total RNA seq data.



### 3. Exploratory analysis

To compute principal component analysis for CPTAC rare RCC eRNA expression we employed `rpca` and `tsne` packages to perform PCA and t-SNE analysis.

## Conclusion

In this study, we successfully implemented an eRNA quantification pipeline and detected 15159 eRNA regions which were later used to quantify eRNA reads from the total RNAseq CPTAC rare-RCC dataset. We also addressed issues of region filtering and depth normalization to correctly quantify eRNAs and showed that there is higher enrichment for eRNAs in the Ribo-Erase dataset (approx. 10-fold). We also obtained a significantly higher number of detectable eRNAs in the riboErase dataset (790 vs 170). We also demonstrated that eRNA can be used to correctly cluster rare-RCC by histological types. We also observed that eRNAs represent a very small fraction of the total transcriptome (approx. 0.02%) and expression of the great majority of eRNAs very low (<2 reads) which makes eRNAs hard to disambiguate from other types of non-genic transcription.

Taken together, our eRNA quantification pipeline is the first of its kind to quantify eRNAs from total RNA seq data detecting more eRNA than the previously reported studies. However, not all detectable eRNAs fit the classical definition of eRNAs from literature and might be noise or by-product of the gene transcription machinery. Therefore, we conclude that in order to study eRNAs in detail with highly probable eRNAs we would need to do some eRNAs specific assays such as BruSeq to be sure that regions we are quantifying are actually eRNAs.

## Code location on R server

To note, all the code related to eRNA project is here: </mctp/share/users/gondal/cptac3-rarercc>  
For specific tasks please refer to the paths below



### 1. Code for eRNA quantification

[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01\\_Shell\\_Script/Pipeline\\_all\\_features\\_32.sh](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01_Shell_Script/Pipeline_all_features_32.sh)

### 2. Code for PCA, rPCA, t-SNE

[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/02\\_Rnotebooks/Rnotebook\\_v16\\_updated\\_enh\\_Cancer.Rmd](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/02_Rnotebooks/Rnotebook_v16_updated_enh_Cancer.Rmd)

### 3. Code for Heatmaps generation

SRA and TCGA download, STAR and Heatmaps:

[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01\\_Shell\\_Script/Heatmap\\_SRA\\_08\\_GRCH38.sh](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01_Shell_Script/Heatmap_SRA_08_GRCH38.sh)

CPTAC Heatmap

[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01\\_Shell\\_Script/CPTAC\\_heatmap\\_GRCH38.sh](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01_Shell_Script/CPTAC_heatmap_GRCH38.sh)

BAM to Fastq conversion and read count code:


[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01\\_Shell\\_Script/Pipeline\\_18\\_proteinencoding.sh](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/01_Shell_Script/Pipeline_18_proteinencoding.sh)

### 4. Code for RPKM and proportion estimation


[/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/02\\_Rnotebooks/Rnotebook\\_v26\\_all\\_features\\_CPTAC\\_&\\_TCGA.Rmd](/mctp/share/users/gondal/cptac3-rarercc/rna-seq-v5-GRCh38/02_Rnotebooks/Rnotebook_v26_all_features_CPTAC_&_TCGA.Rmd)

## Supplementary materials

### 1. Bioinformatics presentation slides link

 [Bioinformatics Presentation - Quantifying and evaluating enhancer RNAs \(eRNAs\) in Ri...](#)

### 2. Potential future directions sheets link

 [Potential Future Directions - eRNA project](#)

### 3. Dropbox link

[https://www.dropbox.com/sh/rvz38mqwhizon4q/AACURspppqWuejFSYY\\_a3Mj5a?dl=0](https://www.dropbox.com/sh/rvz38mqwhizon4q/AACURspppqWuejFSYY_a3Mj5a?dl=0)