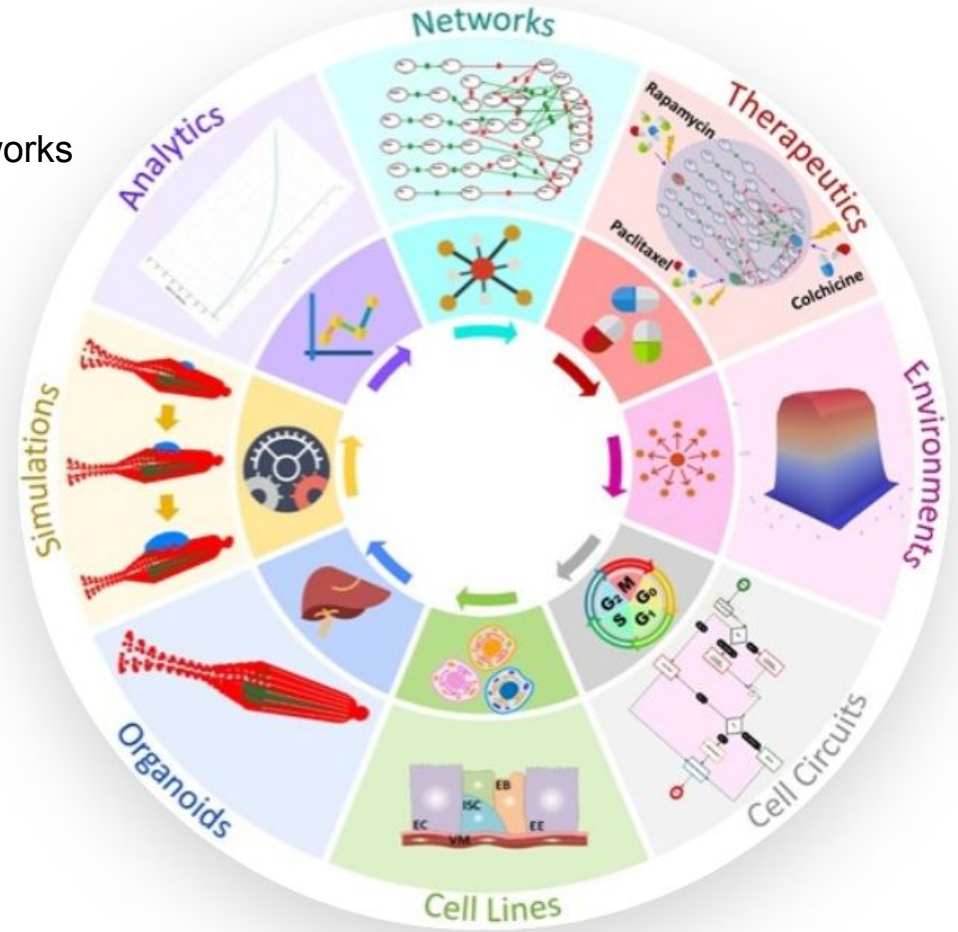
Quantifying and evaluating enhancer RNAs (eRNAs) in RiboErase and poly(A) RNA-seq data

Cieslik Lab, Michigan Medicine,
Department of Computational Medicine and Bioinformatics (DCMB),
University of Michigan (UM)

Disclosure

- Pakistan
- Medical First Responder
- Systems biologists

1. Develop personalized **networks** models
2. Carry out **therapeutic evaluations** on the networks
3. Integrate extracellular **environment**
4. Construct **finite state machines**
5. Formulate *in silico* cell lines and **organoids**
6. **Simulate** organoids
7. Query the **biomolecular regulations**



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

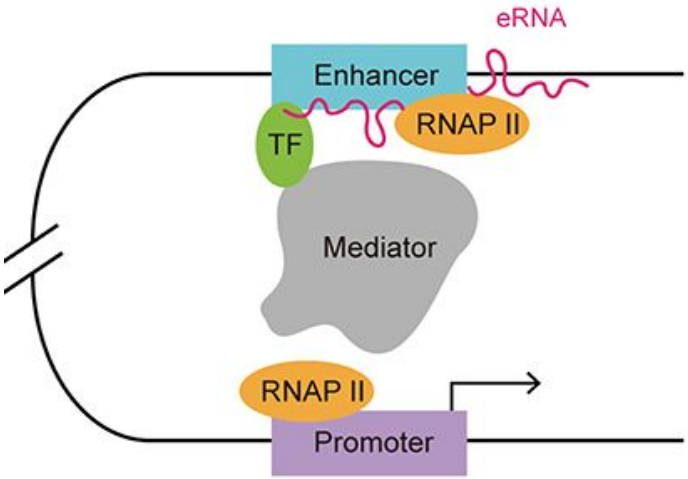
Cieslik Lab, Michigan Medicine,
Department of Computational Medicine and Bioinformatics (DCMB),
University of Michigan (UM)

Outline

- **Background**
  - What are eRNAs and why are we interested in them?
  - Previous literature on eRNA quantification
  - What is the Clinical Proteomic Tumor Analysis Consortium (CPTAC) rare-RCC cohort?
- **Our goals**
  - Implement a pipeline to quantify eRNAs
  - Reproduce previous findings
  - Compare riboErase with polyA RNA-seq
  - Determine the utility of this data in separating rare-RCC subtypes
- **Approach**
  - Pipeline design
- **Preliminary results**
  - Re-analysis of the SRA data
  - CPTAC vs TCGA data
  - Rare-RCC cluster subtypes
- **A closer look, what are we measuring really?**
- **Future directions**

What are eRNAs?

What are eRNAs?



What are the plausible mechanisms of eRNA function/initiation?

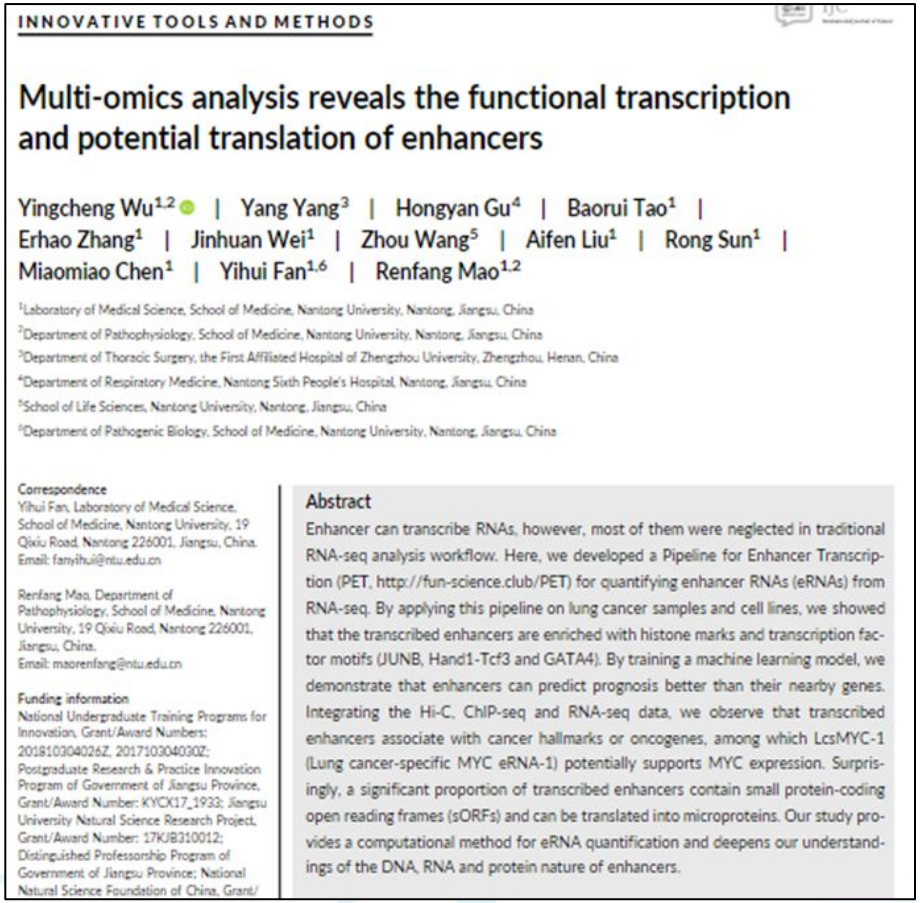
- Can eRNA function cis or trans?
- What comes first transcription of eRNA or target genes?
- Which eRNAs are functional, and how eRNA function is linked to structure and localization?
- Are eRNA different from lncRNAs?

Wu, M. & Shen, J. From Super-Enhancer Non-coding RNA to Immune Checkpoint: Frameworks to Functions . *Frontiers in Oncology* vol. 9 1307 (2019).

Features/role of eRNAs?

1. **eRNA an excellent marker for segregating active versus quiescent enhancers**
2. **eRNAs show tissue and lineage specificity**
3. **eRNAs can serve as markers of cell state and function**
4. **Diverse primary structure heterogeneity:**
    a. Polyadenylated eRNAs are frequently longer (up to 4 kb), unidirectionally transcribed, and transcribed from higher-activity enhancers than are their bidirectional, non-polyadenylated counterparts
    b. Majority of eRNAs are short (median of 346 nucleotides), bidirectionally transcribed, unspliced, and non-polyadenylated
5. **Diverse secondary structure heterogeneity:**
    a. Some eRNA molecules contain different domains with unique functions
6. **Relationship with lncRNA:**
    a. Long polyadenylated eRNAs are structurally similar to lncRNAs, but are transcribed from active enhancers rather than promoters
7. **Mechanism of action:**
    a. eRNA-protein interactions may play roles in protein recruitment, altering protein interactions, and providing scaffolding. (eg eRNA cooperation with YY1)
8. **knockdown of the respective eRNA decreases in promoter-enhancer chromatin looping**
9. **Clinical Implications**

Previous studies to quantify eRNAs



INNOVATIVE TOOLS AND METHODS

Multi-omics analysis reveals the functional transcription and potential translation of enhancers

Yingcheng Wu[1,2] | Yang Yang[3] | Hongyan Gu[4] | Baorui Tao[1] | Erhao Zhang[1] | Jinhuan Wei[1] | Zhou Wang[5] | Aifen Liu[1] | Rong Sun[1] | Miaomiao Chen[1] | Yihui Fan[1,6] | Renfang Mao[1,2]

[1]Laboratory of Medical Science, School of Medicine, Nantong University, Nantong, Jiangsu, China
[2]Department of Pathophysiology, School of Medicine, Nantong University, Nantong, Jiangsu, China
[3]Department of Thoracic Surgery, the First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
[4]Department of Respiratory Medicine, Nantong Sixth People's Hospital, Nantong, Jiangsu, China
[5]School of Life Sciences, Nantong University, Nantong, Jiangsu, China
[6]Department of Pathogenic Biology, School of Medicine, Nantong University, Nantong, Jiangsu, China

Correspondence
Yihui Fan, Laboratory of Medical Science, School of Medicine, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu, China.
Email: fanyihui@ntu.edu.cn

Renfang Mao, Department of Pathophysiology, School of Medicine, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu, China.
Email: maorenfang@ntu.edu.cn

Funding information
National Undergraduate Training Programs for Innovation, Grant/Award Numbers: 201810304026Z, 201710304030Z; Postgraduate Research & Practice Innovation Program of Government of Jiangsu Province, Grant/Award Number: KYCX17_1933; Jiangsu University Natural Science Research Project, Grant/Award Number: 17KJB310012; Distinguished Professorship Program of Government of Jiangsu Province; National Natural Science Foundation of China, Grant/

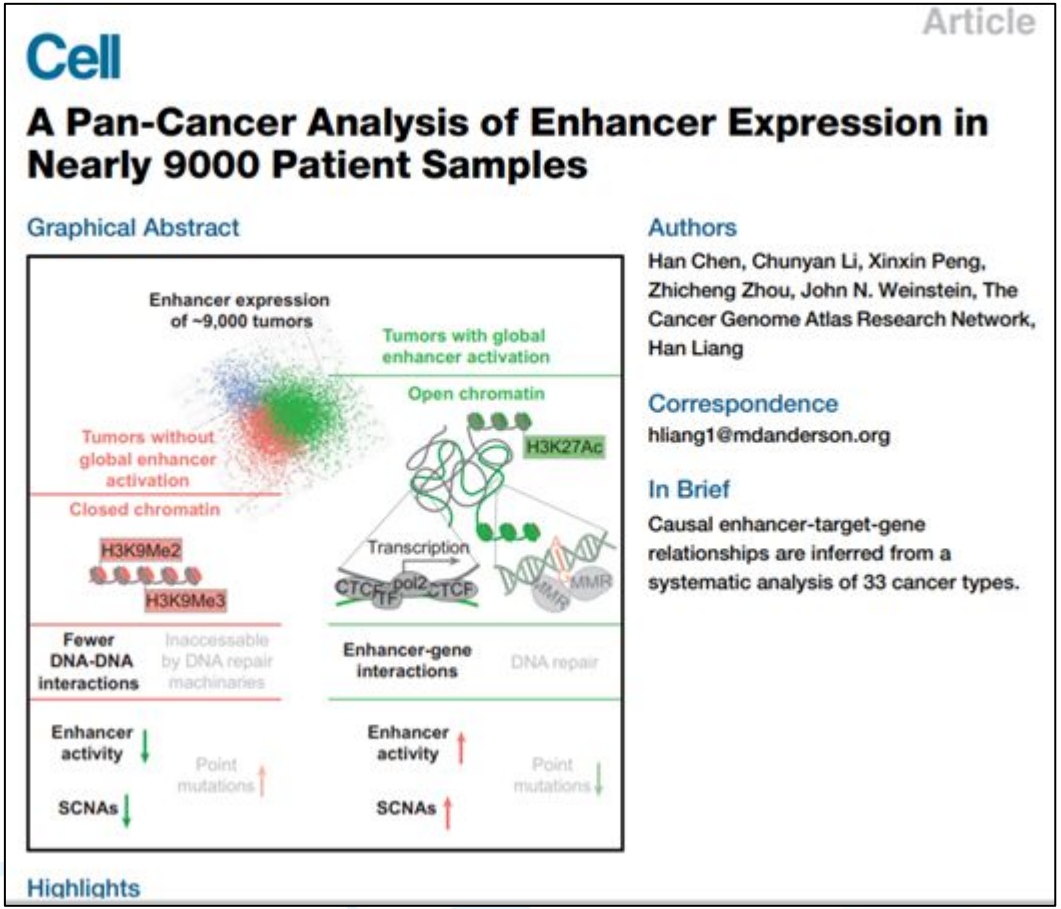
Abstract

Enhancer can transcribe RNAs, however, most of them were neglected in traditional RNA-seq analysis workflow. Here, we developed a Pipeline for Enhancer Transcription (PET, http://fun-science.club/PET) for quantifying enhancer RNAs (eRNAs) from RNA-seq. By applying this pipeline on lung cancer samples and cell lines, we showed that the transcribed enhancers are enriched with histone marks and transcription factor motifs (JUNB, Hand1-Tcf3 and GATA4). By training a machine learning model, we demonstrate that enhancers can predict prognosis better than their nearby genes. Integrating the Hi-C, ChIP-seq and RNA-seq data, we observe that transcribed enhancers associate with cancer hallmarks or oncogenes, among which LcsMYC-1 (Lung cancer-specific MYC eRNA-1) potentially supports MYC expression. Surprisingly, a significant proportion of transcribed enhancers contain small protein-coding open reading frames (sORFs) and can be translated into microproteins. Our study provides a computational method for eRNA quantification and deepens our understandings of the DNA, RNA and protein nature of enhancers.

Key takeaways:

- Identified **15808** eRNAs
- Used K562 cell line's **total RNA-seq** and **polyA RNA-seq** dataset to show that both give the same enrichment of eRNAs
- Explored the function of eRNAs in **lung cancer cell lines and tissues** from SRA
- Defined the transcriptional factor program such as GATA4 controlling eRNAs.
- Demonstrated that a significant proportion of transcribed enhancers contain small protein-coding Open Reading Frames (ORF) can be translated into **microproteins**

Previous studies to quantify eRNAs



Cell

Article

A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples

Graphical Abstract

Authors

Han Chen, Chunyan Li, Xinxin Peng, Zhicheng Zhou, John N. Weinstein, The Cancer Genome Atlas Research Network, Han Liang

Correspondence

hliang1@mdanderson.org

In Brief

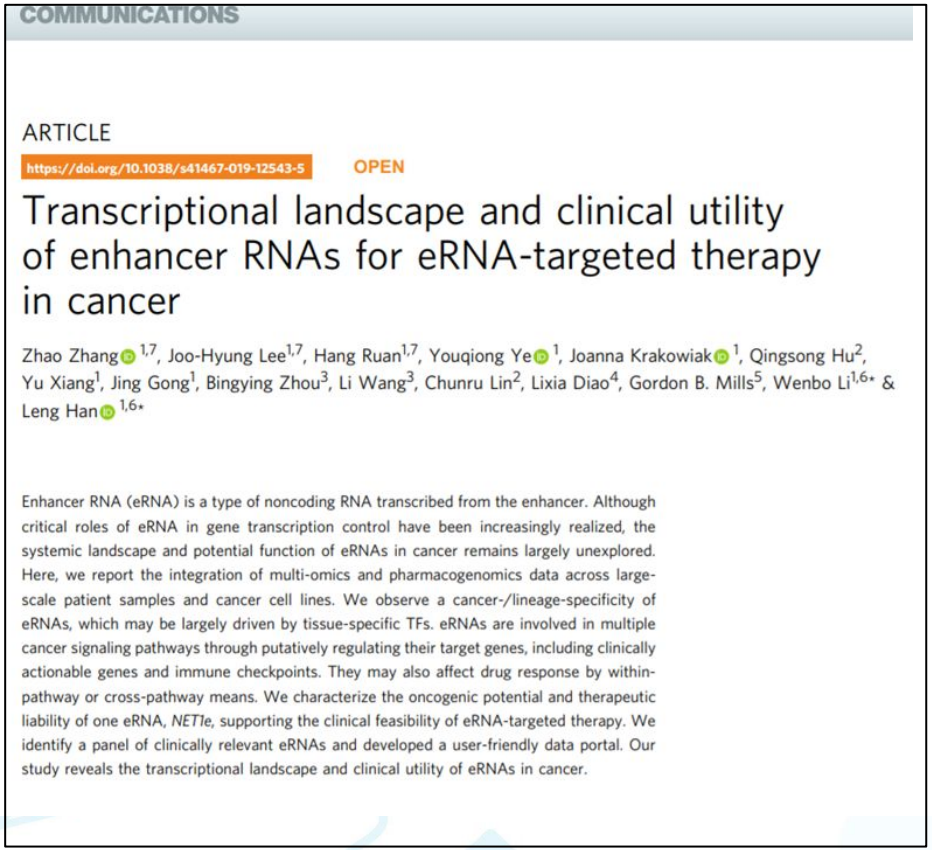
Causal enhancer-target-gene relationships are inferred from a systematic analysis of 33 cancer types.

Highlights

Key takeaways:

- **~9,000** tumor samples across **33** cancer types using **TCGA RNA-seq data**
- Identified **15808 eRNAs**
- Premise: expression of an enhancer approximately reflects its activity.
- Demonstrated that global enhancer activation in cancer is **positively associated with tumor aneuploidy**
- Revealed a considerable number of enhancers, including enhancer-9 for PD-L1, associated with clinically actionable genes

Previous studies to quantify eRNAs



COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12543-5 OPEN

Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer

Zhao Zhang[1,7], Joo-Hyung Lee[1,7], Hang Ruan[1,7], Youqiong Ye[1], Joanna Krakowiak[1], Qingsong Hu[2], Yu Xiang[1], Jing Gong[1], Bingying Zhou[3], Li Wang[3], Chunru Lin[2], Lixia Diao[4], Gordon B. Mills[5], Wenbo Li[1,6*] & Leng Han[1,6*]

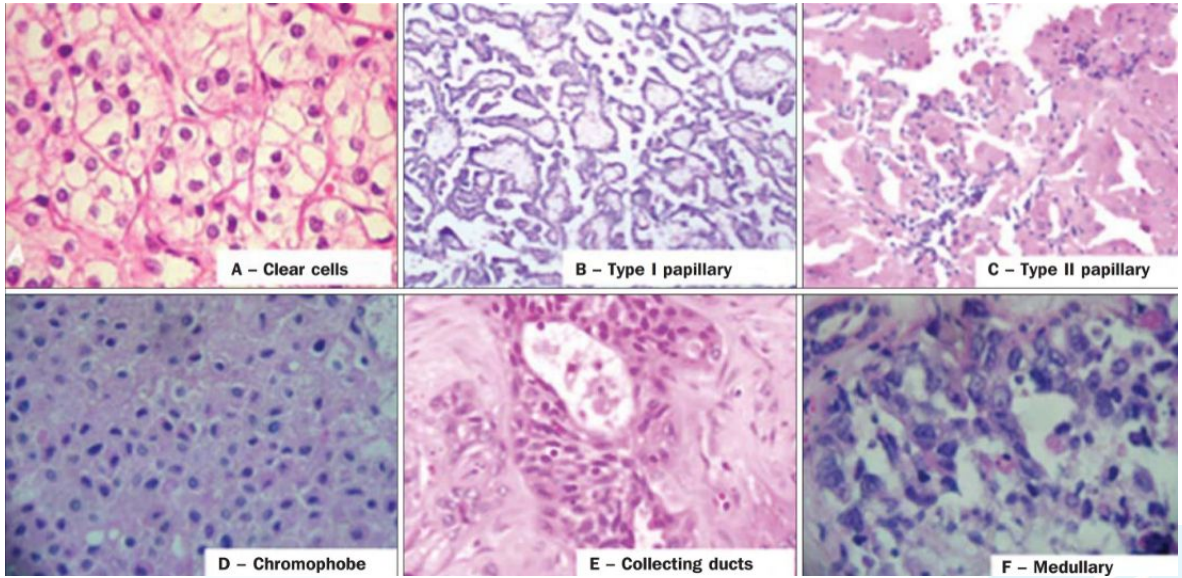
Enhancer RNA (eRNA) is a type of noncoding RNA transcribed from the enhancer. Although critical roles of eRNA in gene transcription control have been increasingly realized, the systemic landscape and potential function of eRNAs in cancer remains largely unexplored. Here, we report the integration of multi-omics and pharmacogenomics data across large-scale patient samples and cancer cell lines. We observe a cancer-/lineage-specificity of eRNAs, which may be largely driven by tissue-specific TFs. eRNAs are involved in multiple cancer signaling pathways through putatively regulating their target genes, including clinically actionable genes and immune checkpoints. They may also affect drug response by within-pathway or cross-pathway means. We characterize the oncogenic potential and therapeutic liability of one eRNA, *NET1e*, supporting the clinical feasibility of eRNA-targeted therapy. We identify a panel of clinically relevant eRNAs and developed a user-friendly data portal. Our study reveals the transcriptional landscape and clinical utility of eRNAs in cancer.

Key takeaways:

- Quantified eRNAs from **poly(A) TCGA RNA-seq**
- Observed a **cancer-/lineage-specificity of eRNAs**, which may be largely driven by tissue-specific TFs.
- Showed eRNAs are involved in **multiple cancer signaling pathways** through putatively regulating their target genes
    - Build a **global eRNA-gene regulatory network** across cancer types based on the physical distance and co-expression between individual eRNA and their target genes
- Demonstrated that **eRNAs may also affect drug response** by within-pathway or cross-pathway means.
    - Correlated eRNA expression level and drug sensitivity of cancer cell lines from CCLE
- Showed NET1 as an oncogenic eRNA in BRCA which may be promising target for eRNA therapy

What is CPTAC-rareRCC cohort?

- Clear cell renal cell carcinoma (ccRCC) represents up to **75–85%** of primary kidney malignancies.
- Other histologies known collectively as **non-clear cell renal cell carcinomas** (non-ccRCC) or rareRCC account for the remaining **15–25%**
- Non-ccRCC encompasses a heterogeneous group of tumors including **papillary, chromophobe, collecting duct**, **translocation**, **medullary** and **unclassified subtypes**
- These histologic subtypes have **pathologic** and **molecular features** distinct from ccRCC and often display different clinical phenotypes



Koshkin *et al*, *Clinical activity of nivolumab in patients with non-clear cell renal cell carcinoma*
Muglia *et al, Renal cell carcinoma: histological classification and correlation with imaging findings*

What is CPTAC-rareRCC cohort?

- Total CPTAC has over 1000 patients and ~10 cancer types
- CPTAC rare-RCC has ~39 patients with 17 paired normal
- CPTAC data is **riboErase**
- One chromophobe renal cell carcinoma (chRCC) patient
- Given this, can we learn something more?

Difference between RiboErase and polyA RNA-seq data?

Total RNA-seq contains all the RNA molecules (**coding** and **noncoding**):

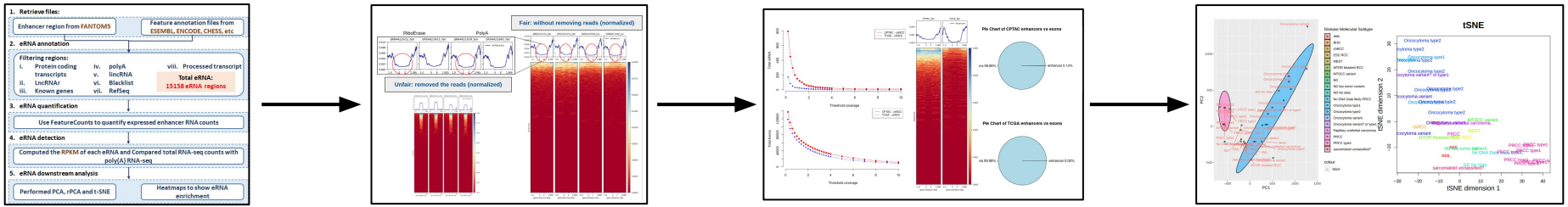
•Enhancer RNA (eRNA),
•precursor messenger RNA (pre-mRNA),
•messenger RNA (mRNA),
•several types of noncoding RNA (ncRNA),
•transfer RNA (tRNA),
•microRNA (miRNA), and
•long ncRNA

mRNA-seq is enriched for **polyadenylated (poly(A))** RNA.

•Primarily for the coding region.
•Enriched for poly(A) tails.

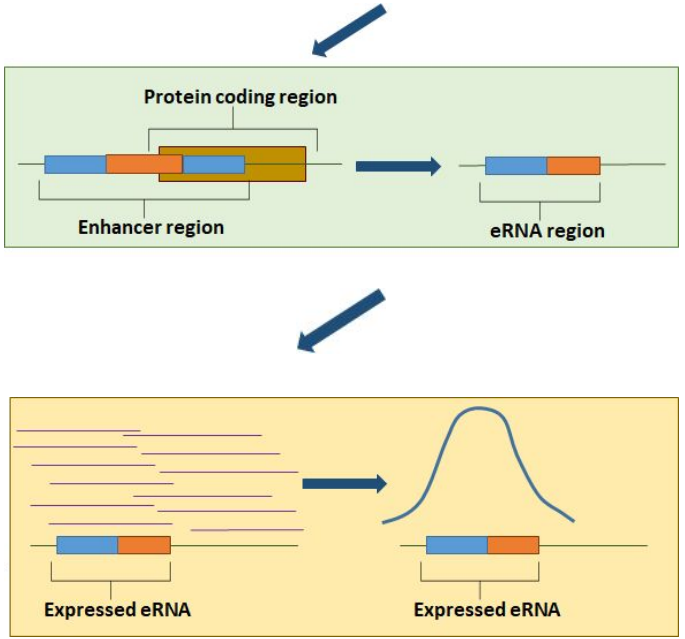
Our goals

1. Implement a pipeline to quantify eRNAs
2. Reproduce previous findings
3. Compare riboErase with polyA RNA seq for the purpose of quantifying eRNAs
4. Determine the utility of this data in separating rare-RCC subtypes

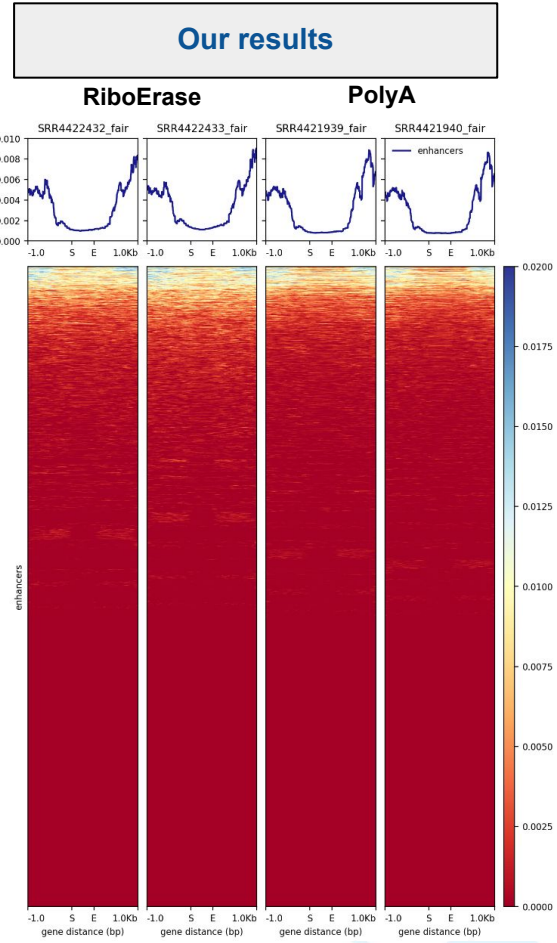


Pipeline design

1. **Retrieve files:**
   - Enhancer region from FANTOM5
   - Feature annotation files from ESEMBL, ENCODE, CHESS, etc
2. **eRNA annotation**
   - Filtering regions:
     - i. Protein coding transcripts
     - ii. LncRNA
     - iii. Known genes
     - iv. lincRNA
     - v. Blacklist
     - vi. RefSeq
     - vii. Processed transcript
   - We identified total: 15159 eRNA regions
3. **eRNA quantification**
   - Use FeatureCounts to quantify expressed enhancer RNA counts
4. **eRNA detection**
   - Computed the RPKM of each eRNA and compared total RNA-seq counts with poly(A) RNA-seq
5. **eRNA downstream analysis**
   - Performed PCA, rPCA and t-SNE
   - Heatmaps to show eRNA enrichment



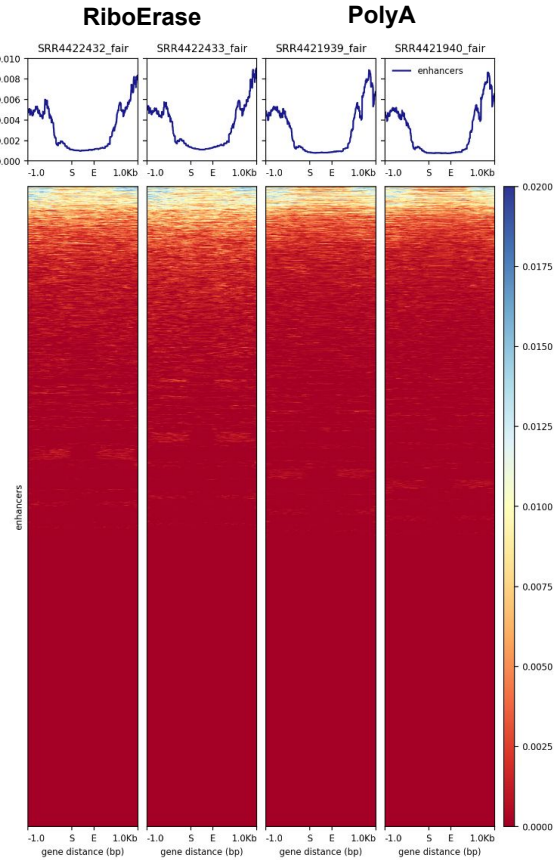
Re-analysis of the SRA data



- We used K562 data from SRA
- K562 is *human immortalised myelogenous leukemia cell line*
- We compared K562 total RNA-seq and polyA RNA-seq reads for eRNAs

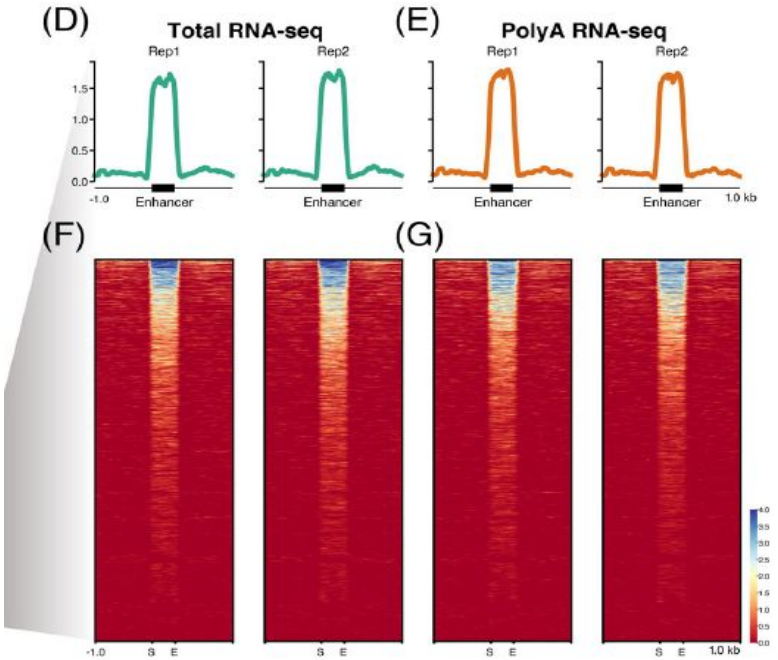
Re-analysis of the SRA data

Our results



Something very wrong.

Published results

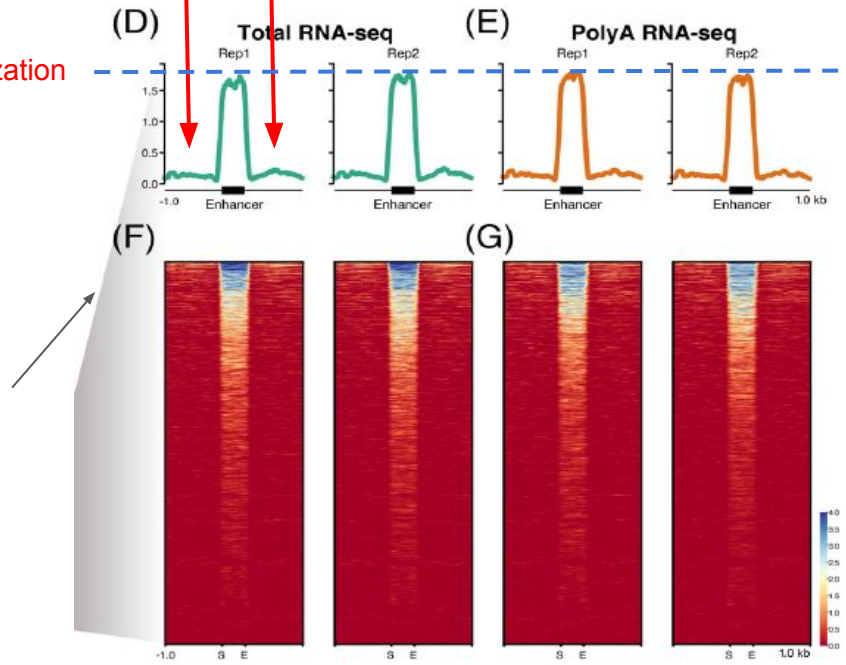


Wu, Y. *et al.* Multi-omics analysis reveals the functional transcription and potential translation of enhancers. *Int. J. Cancer* **147**, 2210–2224 (2020).

Errors in previous analysis

Methods: they removed the reads

Methods: incorrect normalization



Two problems:

1. They removed the reads which they considered as "*noise*"

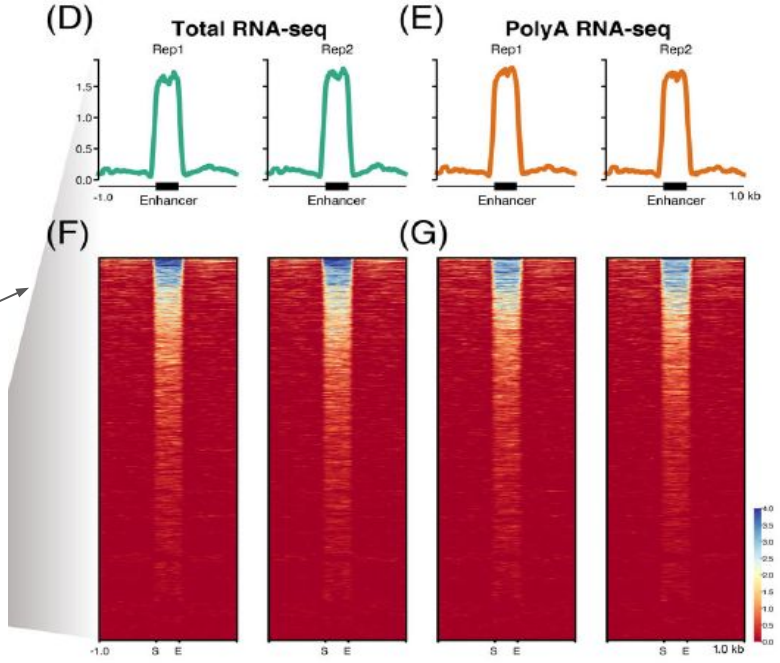
2. Incorrect normalization
   a. They have equal mostly likely means that they removed everything else

Wu, Y. *et al.* Multi-omics analysis reveals the functional transcription and potential translation of enhancers. *Int. J. Cancer* **147**, 2210–2224 (2020).

Re-analysis of the SRA data

Published results

Given that eRNAs might not be fully mature, total RNA-seq and polyA RNA-seq might have different power to identify such new RNA species.[12,40] Thus, we compared the eRNAs generated by total RNA-seq and polyA RNA-seq. To avoid bias from different cell line batches, we used the same K562 cell line but distinct RNA-seq methods (Methods). Unexpectedly, eRNAs detected by total or polyA RNA-seq shows similar expression pattern (Figures 1D-G and S1A,B). In K562 cells, total RNA-seq identified 3809 eRNA and polyA RNA-seq identified 3613 eRNAs (Figure S1B). The signal intensity of eRNAs was similar between total and polyA RNA-seq (Figure 1D-G). There are 2526 eRNAs can be found in both two group. Around 65% of eRNAs can be identified by either total RNA-seq or polyA RNA-seq (Figure 1I). For example, we show the total RNA-seq, polyA RNA-seq, and ChIP-seq signal at the Calmodulin Like 5 (CALML5) locus (Figure 1C). A peak was located ~10 000 bp upstream of CALML5 locus in multiple datasets. In summary, no large difference was observed between polyA RNA-seq and total RNA-seq in eRNA quantification.

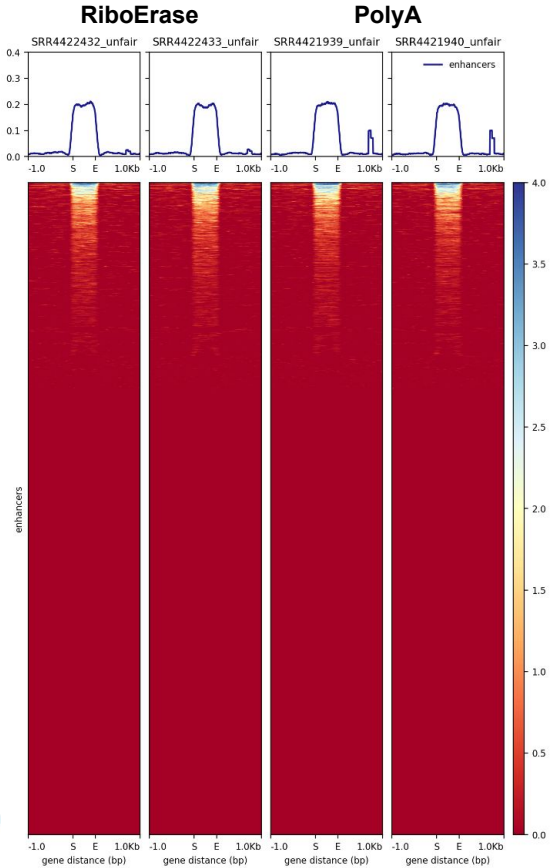


Wu, Y. *et al.* Multi-omics analysis reveals the functional transcription and potential translation of enhancers. *Int. J. Cancer* **147**, 2210–2224 (2020).

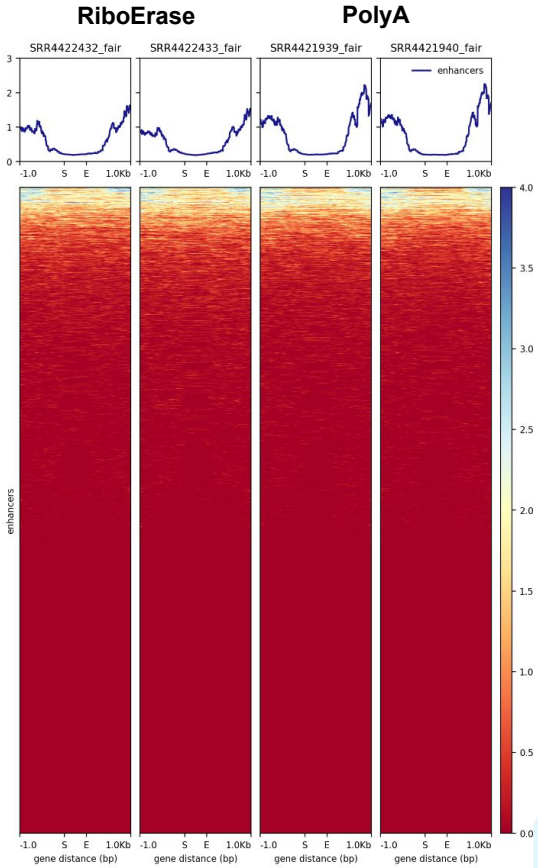
Re-analysis of the SRA data

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Unfair: removed the reads (not normalized)



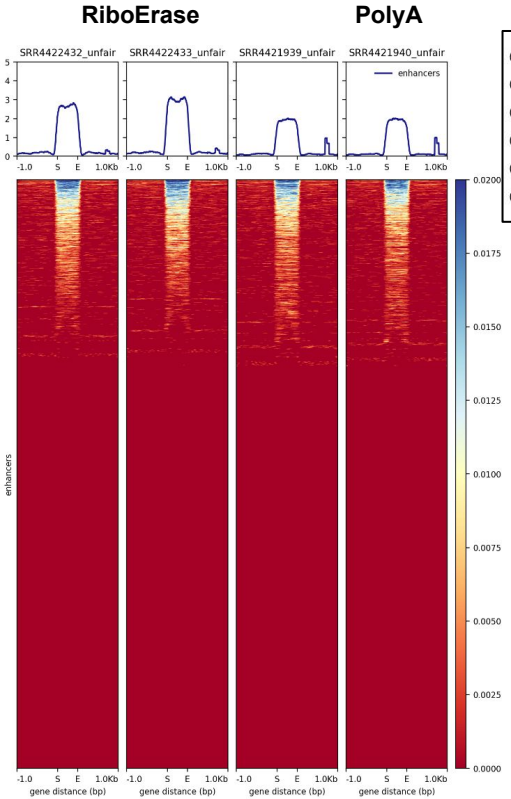
Fair: without removing reads (not normalized)

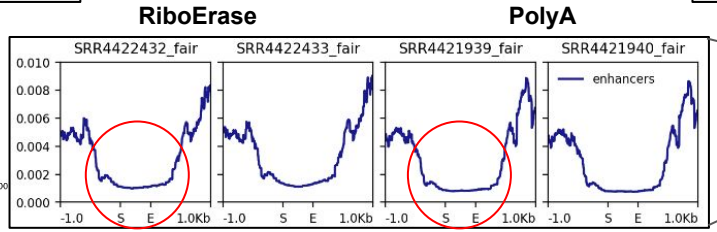


Re-analysis of the SRA data

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

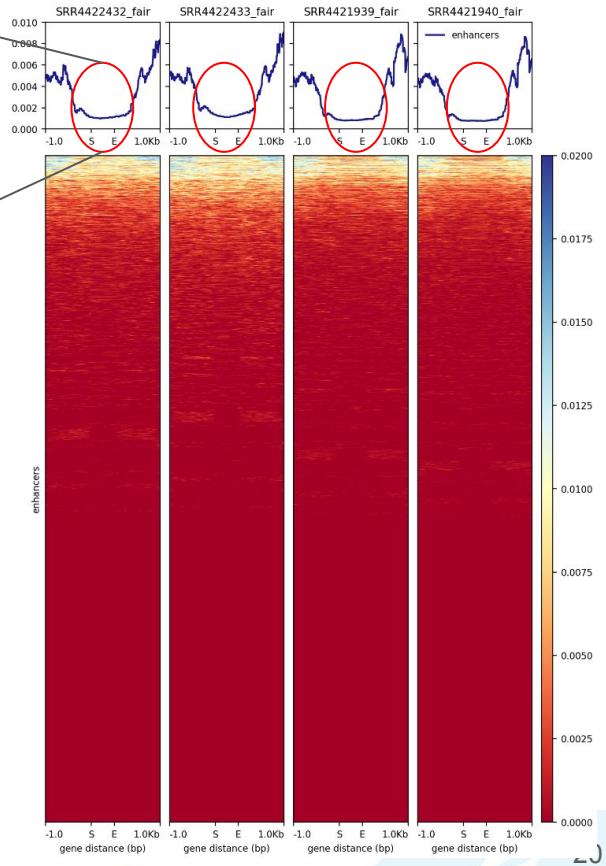
Unfair: removed the reads (normalized)





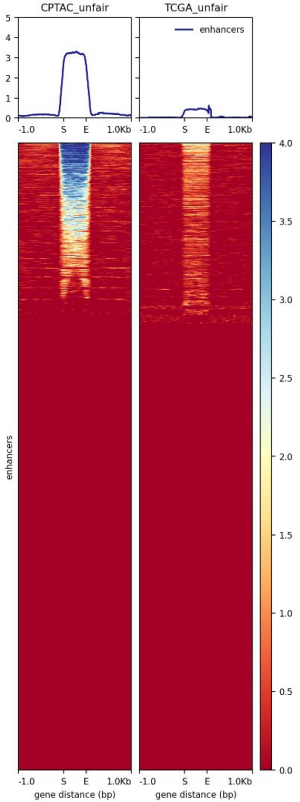
Higher enrichment for RiboErase data

Fair: without removing reads (normalized)

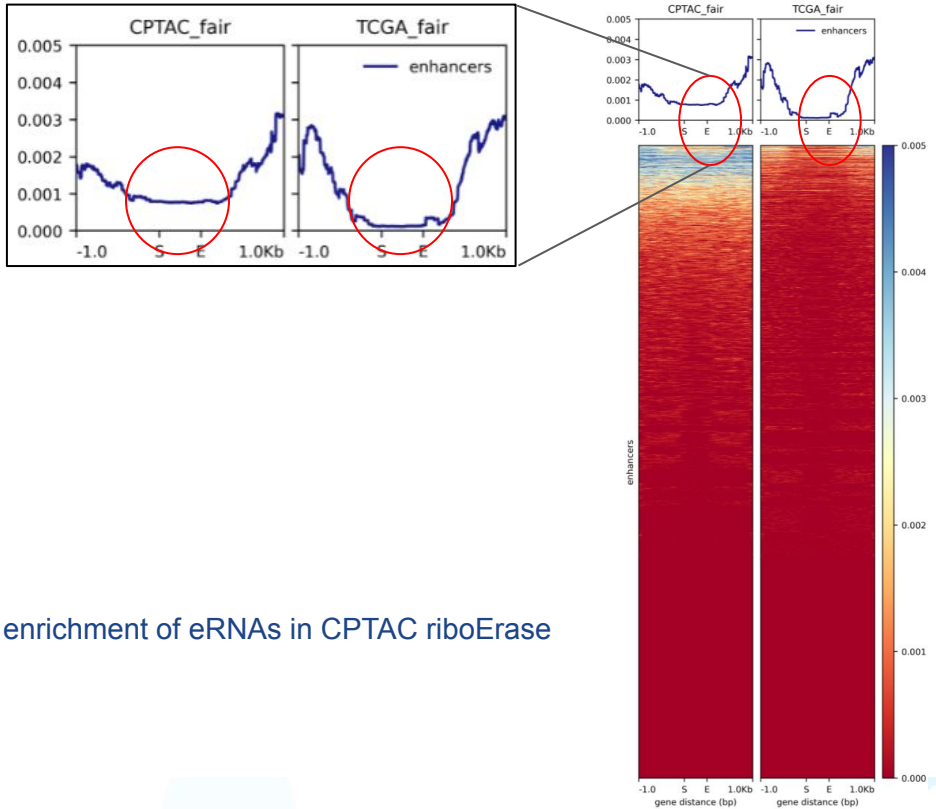


Comparing eRNA enrichment from CPTAC & TCGA data

Unfair: removed the reads (normalized)



Fair: without removing the reads (normalized)



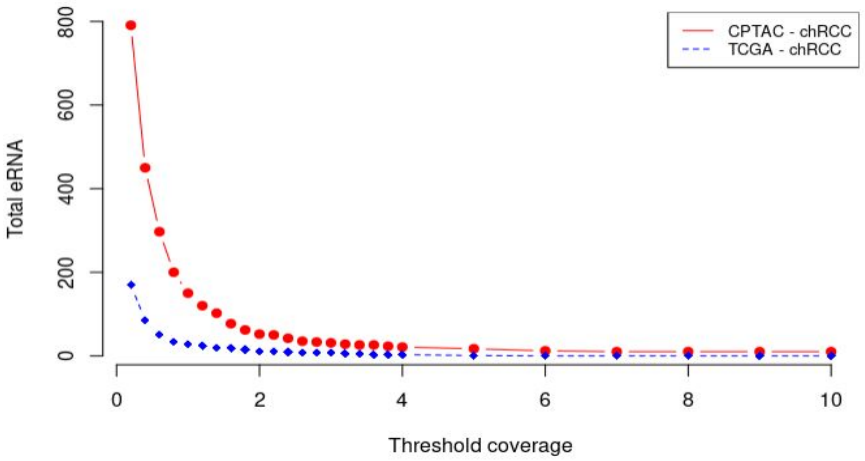
Conclusion:

- Higher enrichment of eRNAs in CPTAC riboErase data

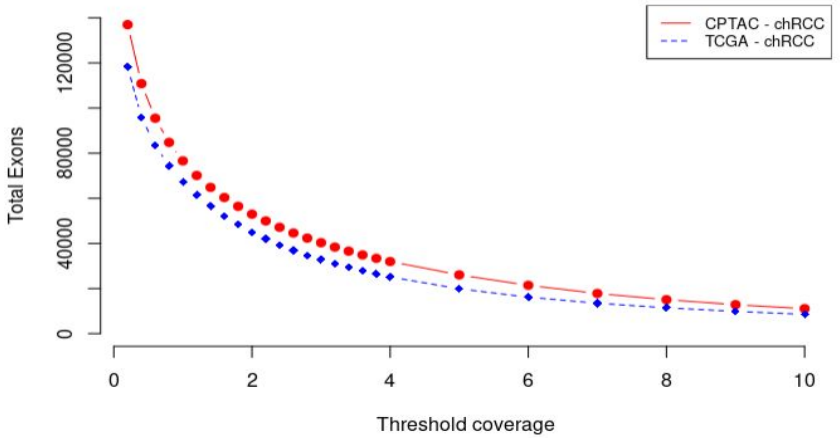
Evaluating number of detected eRNA and exons in RiboErase RNA-seq vs polyA RNA-seq

CPTAC-chRCC vs TCGA KICH-sample

Number of detected eRNAs



Number of detected exons



Conclusion:

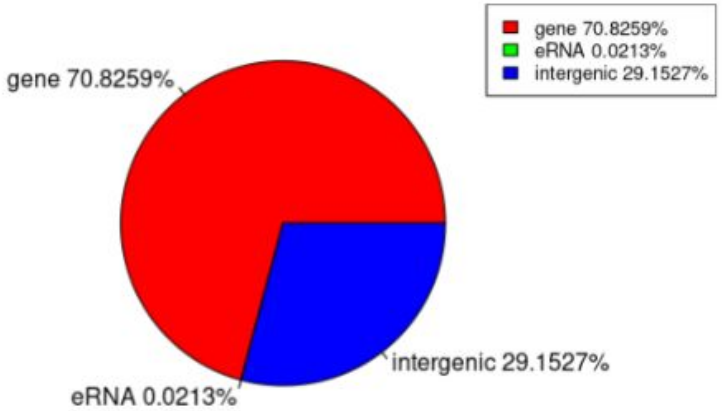
- Higher total number of eRNAs in CPTAC riboErase data

Evaluating proportion of reads for gene, eRNAs, and intergenic regions in RiboErase RNA-seq vs polyA RNA-seq

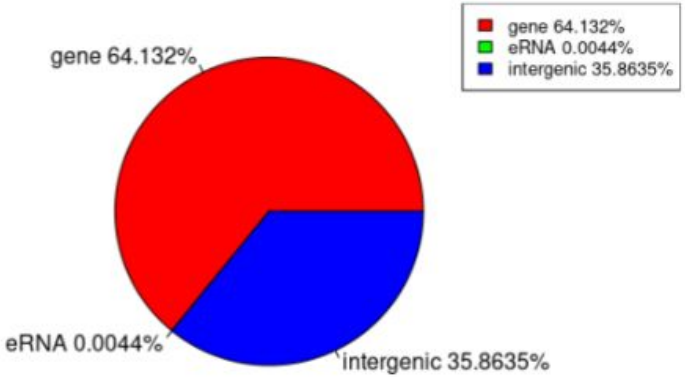
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

CPTAC-chRCC vs TCGA KICH-sample

CPTAC eRNA, exon, and intergenic number



TCGA eRNA, exon, and intergenic number

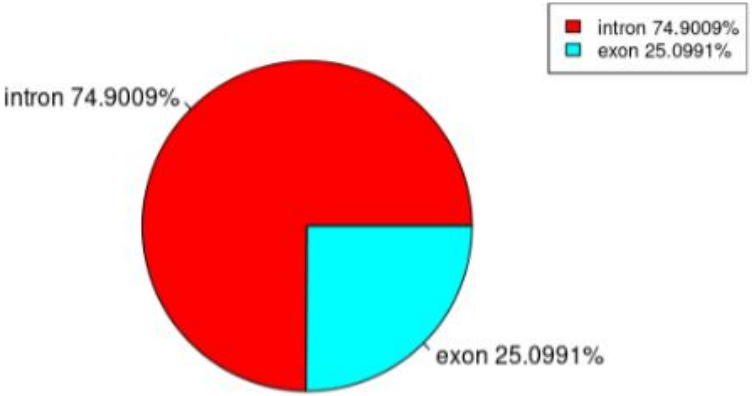


Evaluating proportion of reads for introns and exons regions in RiboErase RNA-seq vs polyA RNA-seq

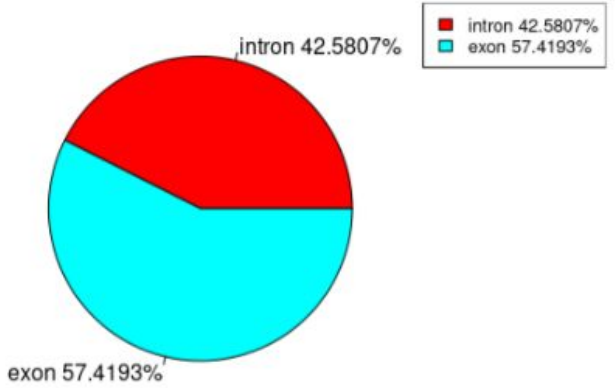
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

CPTAC-chRCC vs TCGA KICH-sample

CPTAC exon vs intron number



TCGA- exon vs intron number

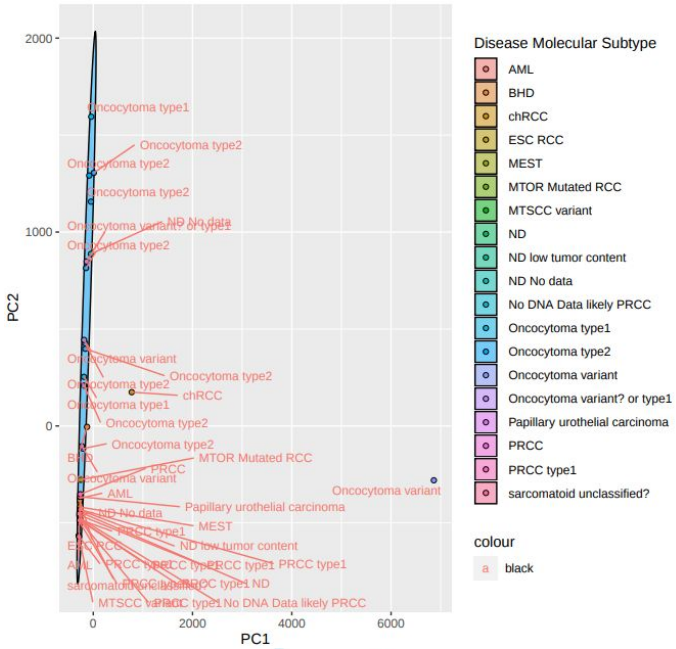


Determining the utility of eRNAs in separating rare-RCC subtypes

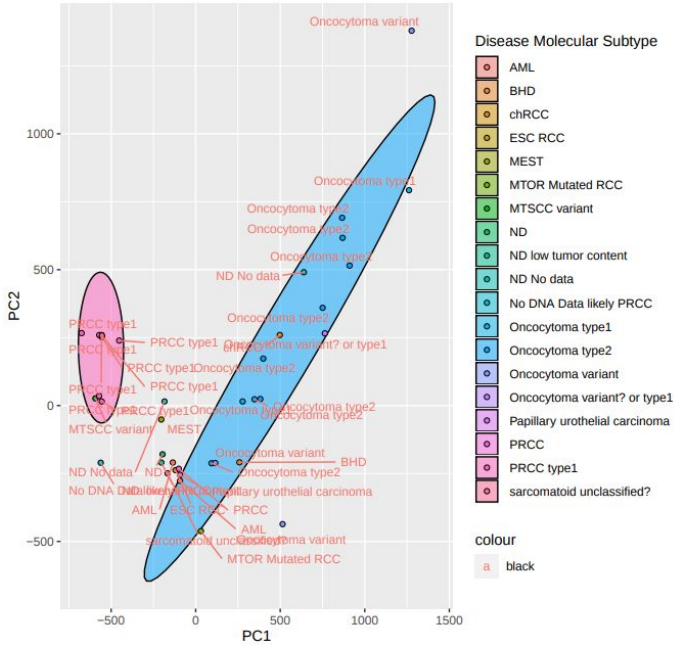
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Principal Component Analysis (PCA)

Classical PCA



Robust PCA



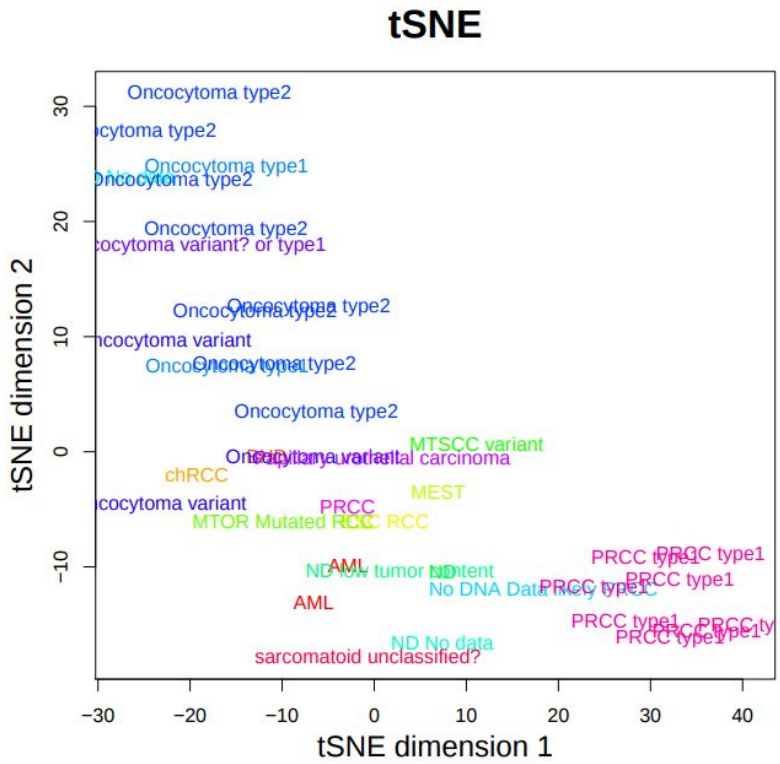
Determining the utility of eRNAs in separating rare-RCC subtypes

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

t-Distributed Stochastic Neighbor Embedding (t-SNE)

Conclusion:

- eRNAs can be used to identify rareRCC clusters

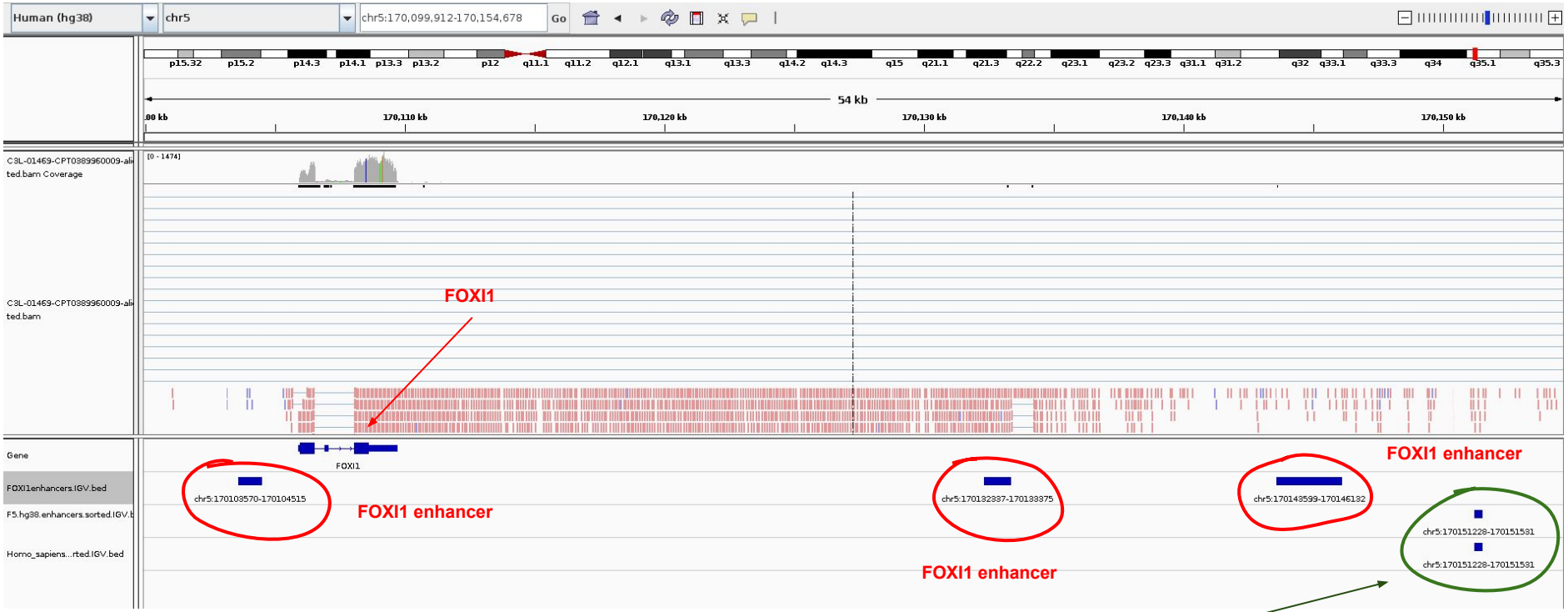


What do we expect from a good eRNA?

- **We expect eRNAs to:**
  - eRNAs are transcribed **bidirectionally** with **predominant production coming from either the negative or positive strand**
  - eRNA **activity-dependent induction** correlates with induction of nearby genes
  - eRNA are **relatively short (median of 346 nucleotides) unspliced RNAs**, the generation of which is strongly related to enhancer activity.
  - Most CAGE-defined enhancers gave rise to **nuclear (>80%)** and **non-polyadenylated (~90%) RNAs**
  - Interestingly, single-cell CAGE sequencing revealed that while on a bulk level (total RNA-seq) enhancers can be described as bidirectionally transcribed, **on a single-cell level enhancers are almost exclusively unidirectionally transcribed from either strand**

Kim *et al*, Widespread transcription at neuronal activity-regulated enhancers
Andersson *et al,* An atlas of active enhancers across human cell types and tissues

eRNAs close to known genes e.g, FOXI1



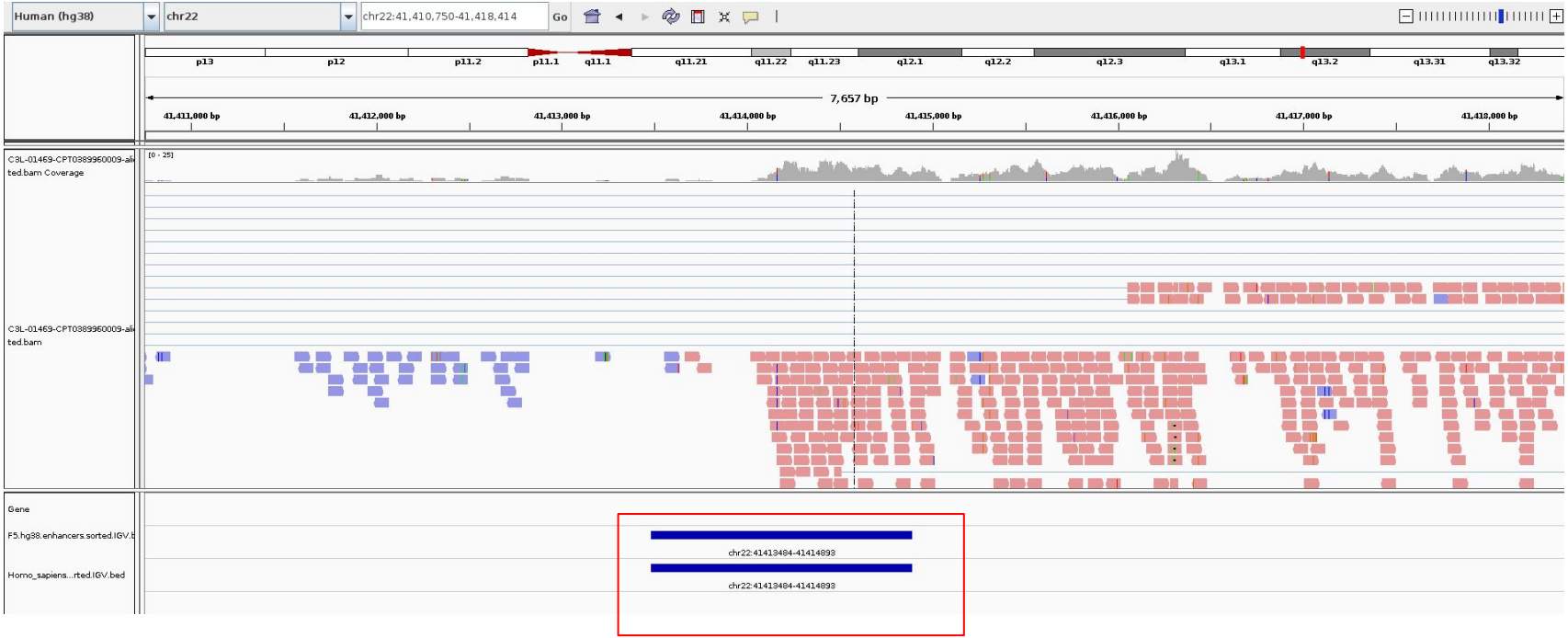
A closer look, what are we measuring really?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many are not eRNA like - Close to strange regions most probably intron retention region:

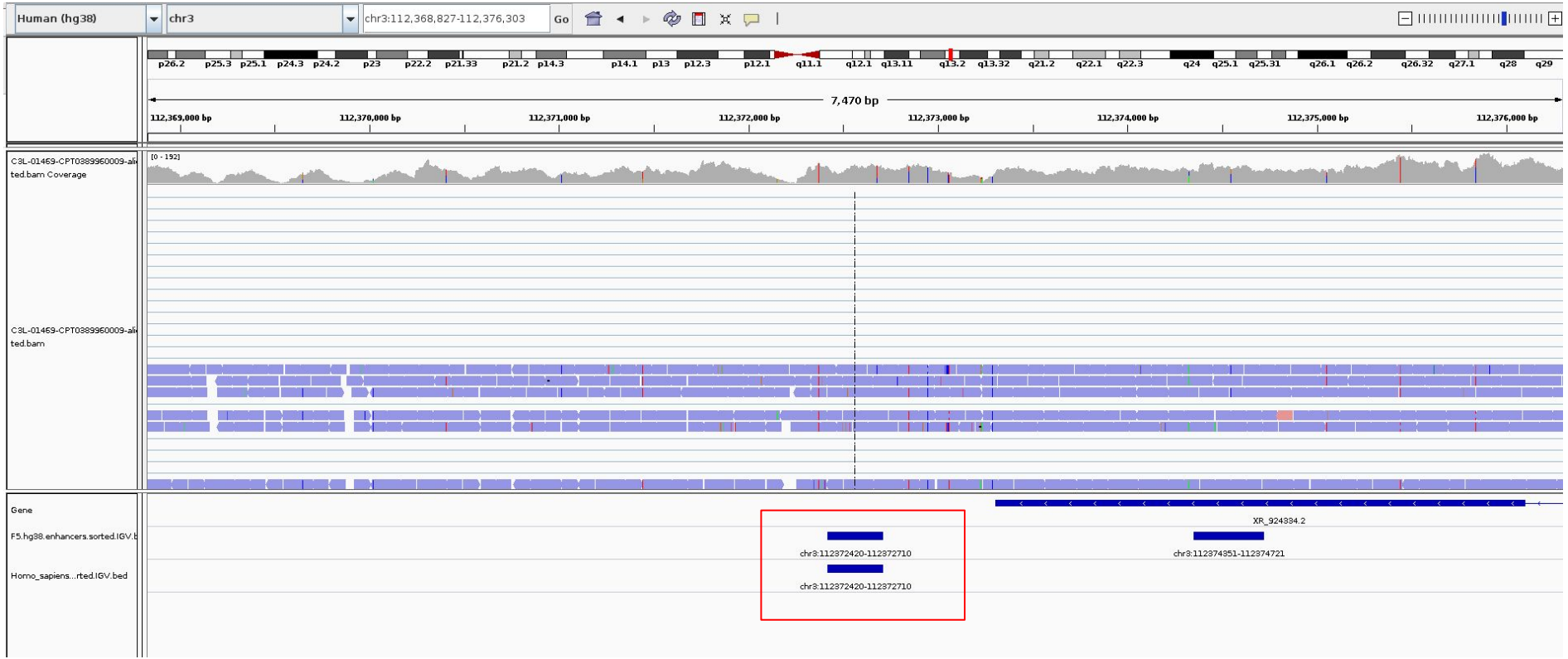
A closer look, what are we measuring really?

Many are not eRNA like - Close to unspliced transcription too large to be eRNAs:



A closer look, what are we measuring really?

Many are not eRNA like - Close to poorly annotated / unknown genes:



A closer look, what are we measuring really?

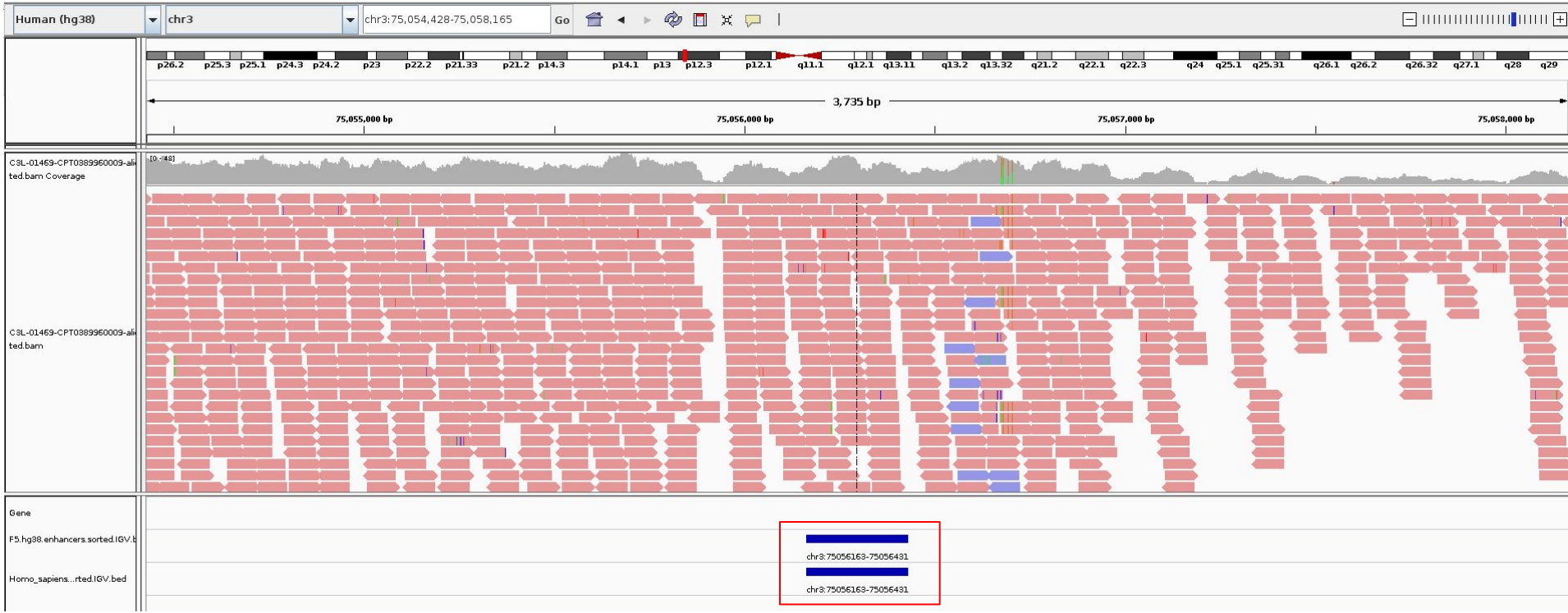
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many are not eRNA like - Close to unknown active regions:

A closer look, what are we measuring really?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many are not eRNA like - Close to unknown active regions with high transcription rate:



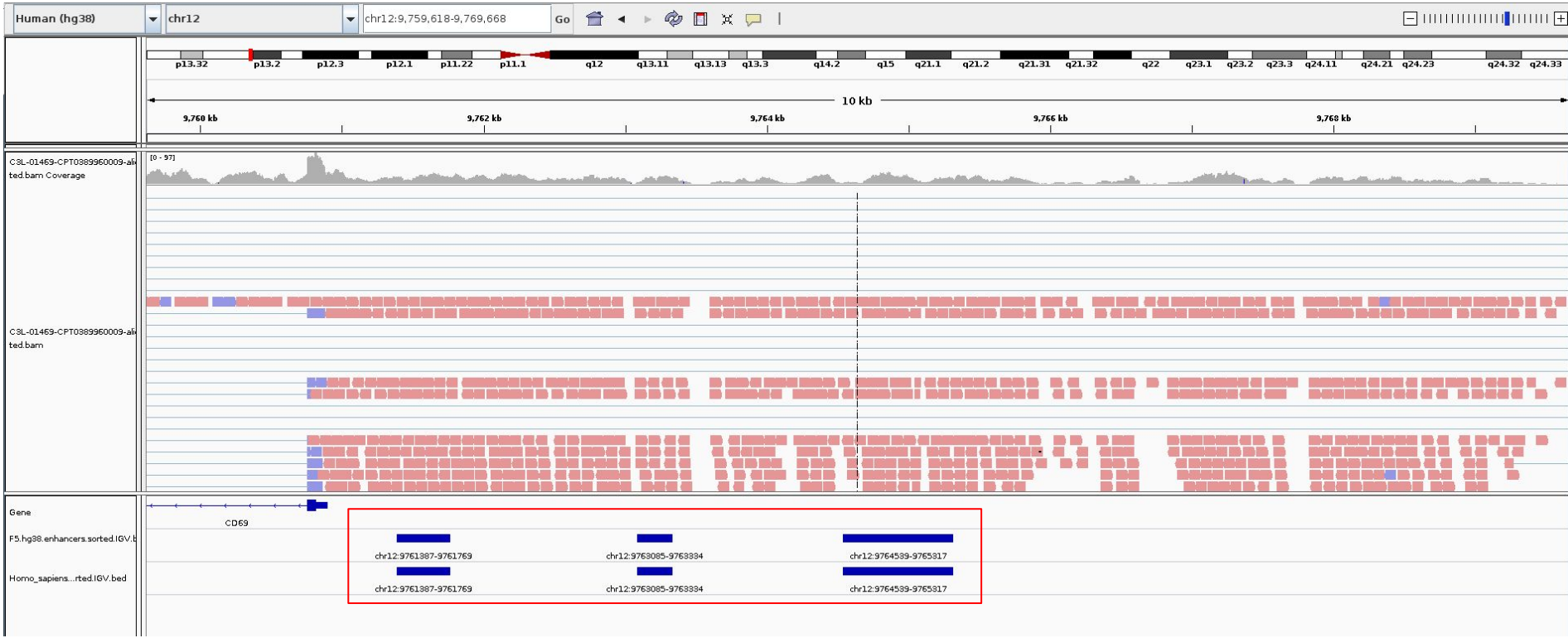
A closer look, what are we measuring really?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many are not eRNA like - eRNA transcription in one way (possible anti-sense tx):

A closer look, what are we measuring really?

Many are not eRNA like - eRNA transcription in one way (anti-sense transcripts?):



A closer look, what are we measuring really?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many are not eRNA like - Close to promoter region:

A closer look, what are we measuring really?

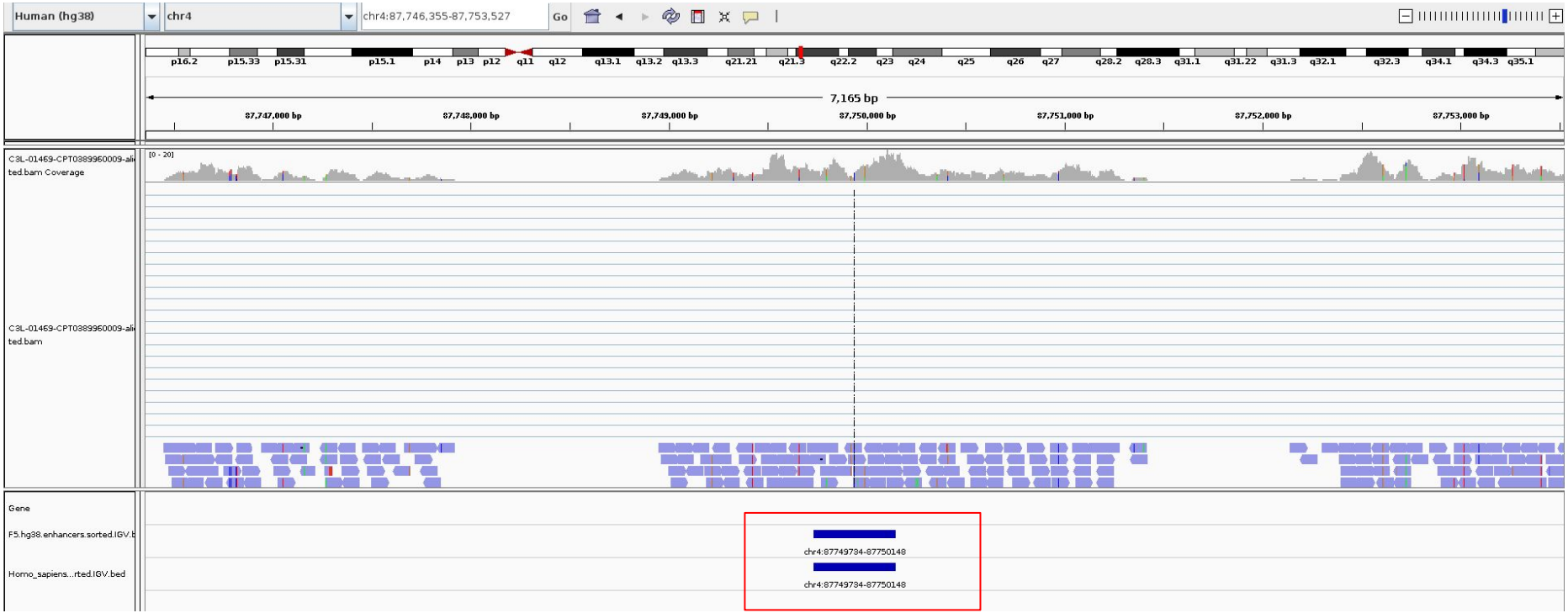
Some good eRNA like plots

A closer look, what are we measuring really?

Some genes nearby

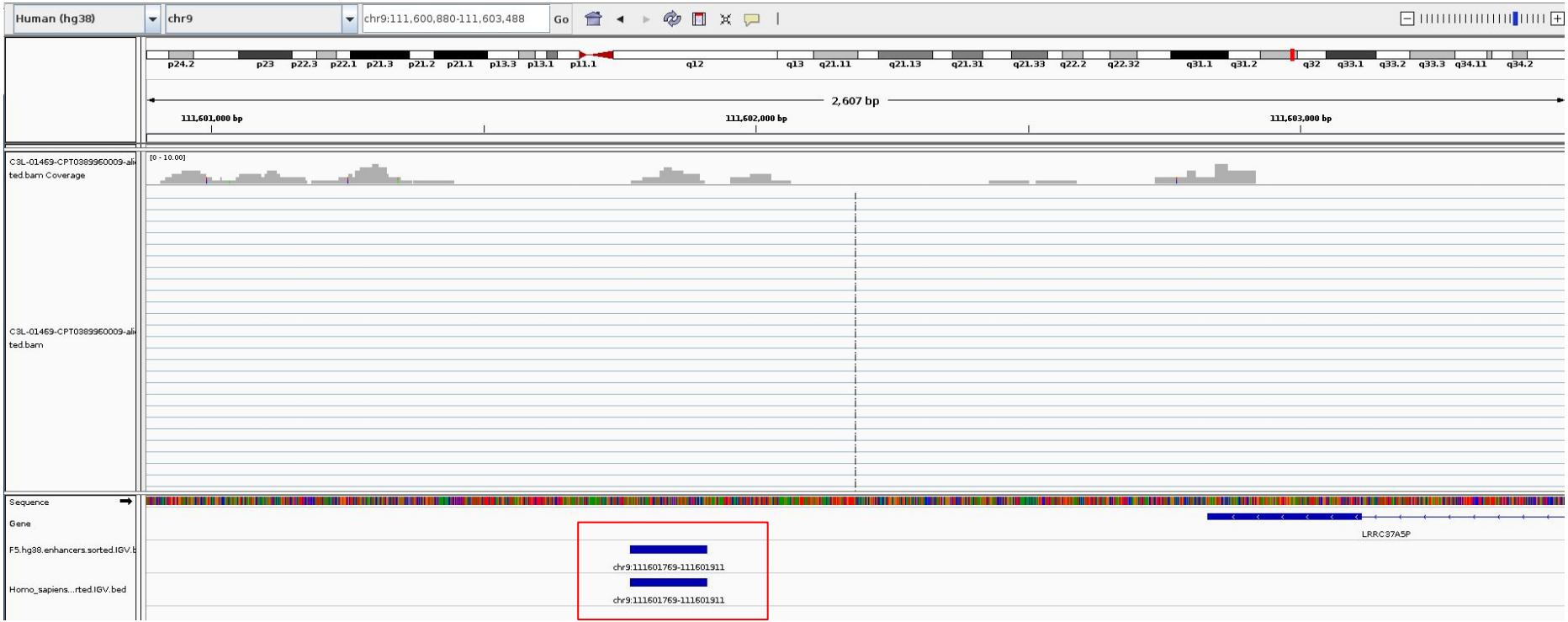
A closer look, what are we measuring really?

Some good eRNA like plots



A closer look, what are we measuring really?

Some good eRNA like plots



Relevant question: is not if eRNAs have a biological role, but which eRNAs are functional, and how eRNA function is linked to structure and localization.

Conclusion

Positive:

- Successfully implemented an eRNA quantification pipeline reproducing previous results
- Addressed issues of region filtering and depth normalization to correctly quantify eRNAs
- Higher enrichment for eRNAs in Ribo-Erase dataset (approx. 10-fold)
- Significantly Higher number of detectable eRNAs in riboErase dataset (790 vs 170)
- eRNA can be used to correctly cluster rare-RCC by histological types

Negative:

- eRNAs represent a very small fraction of the total transcriptome (approx. 0.02%)
- Expression of the great-majority of eRNAs very low (<2 reads)
- eRNAs may be hard to disambiguate from other types of non-genic transcription

Potential future directions

- **To better characterize eRNA**
  - Include epigenetics data using HEK cell line from Encode
  - Characterize these regions in HEK cells
- **Correlate eRNA with target genes using linear regression modelling**
- **Think how systems biology link to cancer genomics**
  - Carry out functional investigation of eRNA-related TFs
- **Identify regions where many enhancers are upregulated and locate nearby oncogenes**
  - Perform motif analysis
  - De Novo motif discovery
  - Analysis of master regulators
- **Correlation between clinically actionable genes and eRNA**

Thank you!

Appendix

Literature

pipeline to quantify eRNAs from RNA-seq. Transcribed enhancers were previously annotated in FANTOM5,[35] which provides a list of ~65 000 enhancers with transcription potential. However, a fraction of those 65 000 enhancers overlap with gene bodies. Thus, it is hard to detect whether the expressed RNAs derive from enhancers or genes. To reduce such noise, enhancers overlapping with known gene bodies, alternative TSS and alternative polyadenylation sites were further removed[15] (Methods and Figure 1A). Through such analysis, we extracted a total of 15 808 enhancers of high confidence.

Dataset	Sample IDs	Total read count	Data type
K562	SRR4422432	39,095,879	total
	SRR4422433	36,989,957	total
	SRR4421939	45,487,792	polyA
	SRR4421940	46,379,810	polyA
TCGA	04c861f3-3130-446c-a226-0dc08f365dc5_gdc_realn_rehead	99,084,357	polyA
	098a03fe-6a1b-4248-834a-9fa70ea02000_gdc_realn_rehead	95,785,353	
	60a27b66-0ba1-4ecc-8ab2-15c187ed7018_gdc_realn_rehead	92,735,127	
	9bc5e4f0-c822-4259-a8dd-c7998064cb8b_gdc_realn_rehead	94,232,189	
	aaaafcf9-1631-49b4-8c1d-c3f6194ef1a9_gdc_realn_rehead	96,312,737	
CPTAC	C3N-02440-CPT0384450006-alig	77,223,879	total
	C3N-02818-CPT0384570006-alig	81,076,602	
	C3N-02818-CPT0384610006-alig	102,178,109	
	C3N-03021-CPT0189110009-alig	103,042,062	
	C3N-03021-CPT0189120006-alig	101,711,038	
	C3L-01469-CPT0389960009-alig	91,607,526	