Pan-cancer discovery and characterization of clinically informative lncRNAs

Keren Isaev[1,2], Jüri Reimand[1,2]

[1]Department of Medical Biophysics, University of Toronto
[2]Ontario Institute for Cancer Research

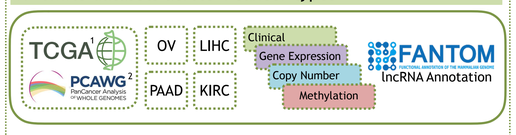
Background

It is becoming increasingly evident that the genome is pervasively transcribed into non-coding RNA including long non-coding RNAs (~20,000 estimated to be functional by FANTOM). Several lncRNAs including *HOTAIR, MALAT1 and SChLAP1* have demonstrated a relationship between gene expression and patient survival outcomes or metastasis in multiple cancers. However, the majority of lncRNAs remain both functionally and clinically undefined while presenting a diverse array of translational opportunities due to their tissue specific expression and potential to regulate cancer pathways. Thus, the purpose of this project is to define a subset of high-confidence lncRNAs for further study and potential translation for diagnostic and prognostic value.

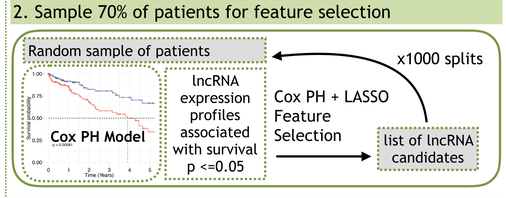
Methods and workflow

Aims:
1. Identify a set of of high-confidence prognostic lncRNA expression profiles
2. Elucidate the underlying genetic or epigenetic differences between the lncRNA annotated risk groups
3. If function is unknown, predict its protein-coding gene (PCG) interactions and associated biological pathways

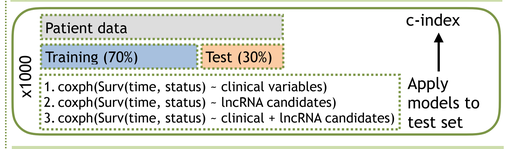
1. *-omics* datasets across 4 cancer types



2. Sample 70% of patients for feature selection



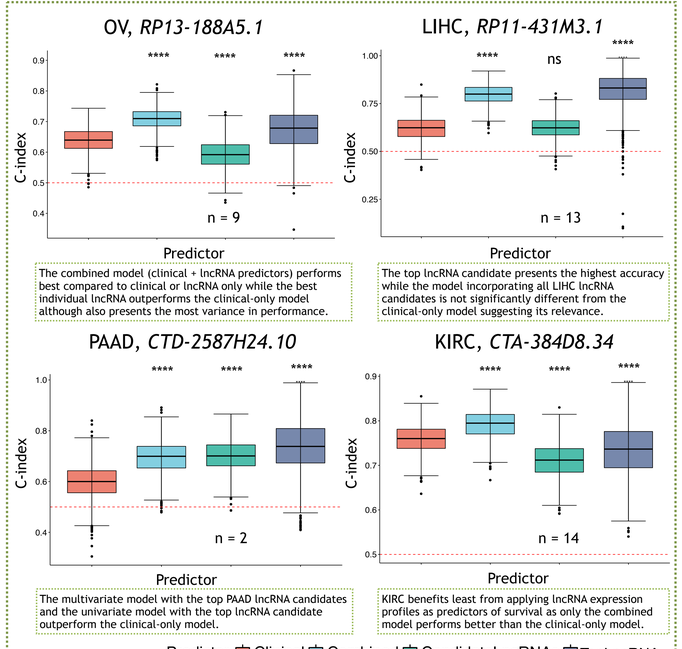
3. Final list of lncRNA candidates consists of those chosen by LASSO in at least 40% of the random splits



4. Identify genetic or epigenetic aberrations driving expression profiles and predict functional associations

1. lm(lncRNA expression ~ promoter methylation)
2. lm(lncRNA expression ~ segment mean)
3. glmnet(lncRNA expression (high/low) ~ PCGs, family = "binomial")

Top lncRNA candidates outperform clinical variables in predicting survival in LIHC, OV and PAAD



The combined model (clinical + lncRNA predictors) performs best compared to clinical or lncRNA only while the best individual lncRNA outperforms the clinical-only model although also presents the most variance in performance.

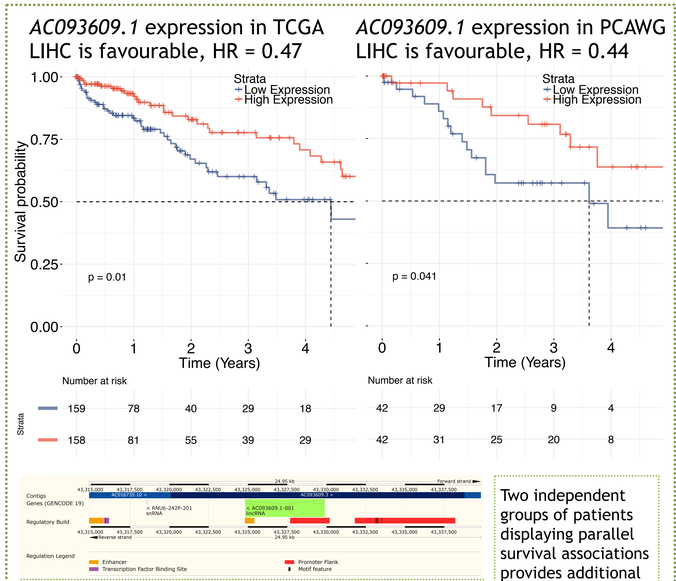
The top lncRNA candidate presents the highest accuracy while the model incorporating all LIHC lncRNA candidates is not significantly different from the clinical-only model suggesting its relevance.

The multivariate model with the top PAAD lncRNA candidates and the univariate model with the top lncRNA candidate outperform the clinical-only model.

KIRC benefits least from applying lncRNA expression profiles as predictors of survival as only the combined model performs better than the clinical-only model.

Predictor Clinical Combined CandidateLncRNAs TopLncRNA

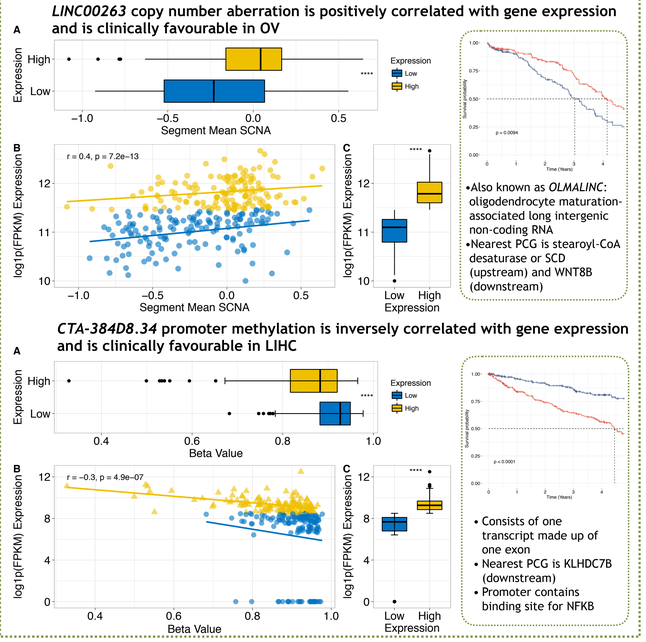
AC093609.1's prognostic expression is validated in an external dataset PCAWG



- *AC093609.1* is located on chromosome 2 with no PCGs nearby while overlapping enhancer and promoter elements
- Co-expression analysis will reveal potential targets and regulators of this lncRNA

Two independent groups of patients displaying parallel survival associations provides additional evidence that the candidate lncRNA is clinically relevant.

Copy number aberrations and methylation influence the expression of candidate lncRNAs



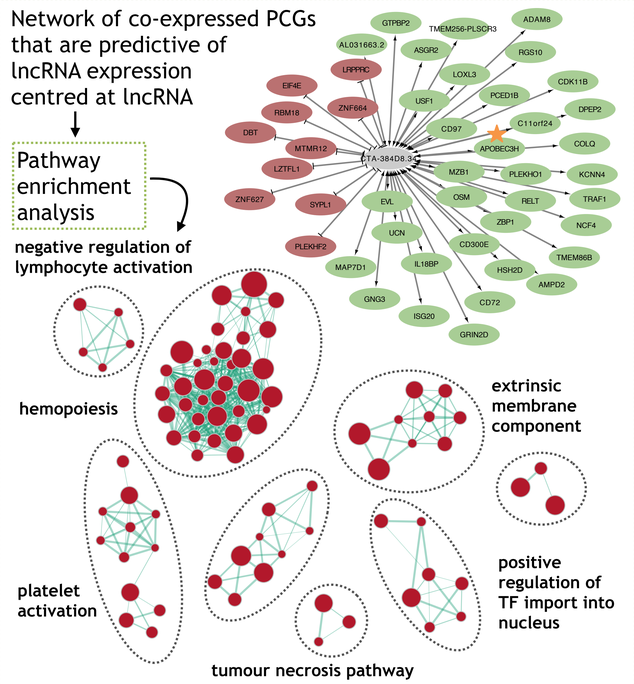
- Also known as *OLMALINC*: oligodendrocyte maturation-associated long intergenic non-coding RNA
- Nearest PCG is stearoyl-CoA desaturase or SCD (upstream) and WNT8B (downstream)

- Consists of one transcript made up of one exon
- Nearest PCG is KLHDC7B (downstream)
- Promoter contains binding site for NFKB

CTA-384D8.34 is associated with immune response in KIRC

Network of co-expressed PCGs that are predictive of lncRNA expression centred at lncRNA

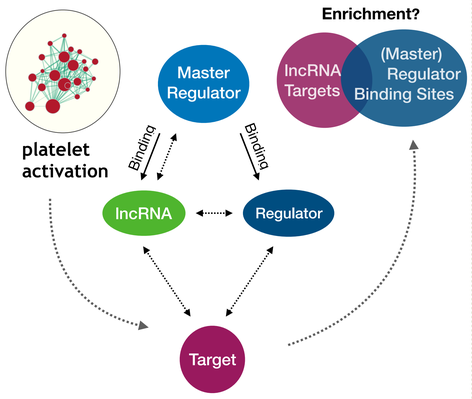
Pathway enrichment analysis



Future directions and applications

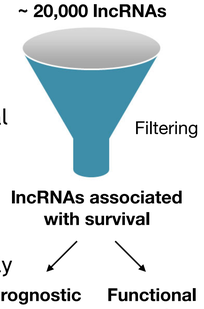
1. Consider **additional clinical variables** including drug response, metastasis and recurrence
   - Combine best performing lncRNA molecular profiles into multivariate model
   - Compare performance of lncRNA predictors with other genetic factors such as the presence of known cancer driver mutations and driver gene expression

2. What is the biological mechanism driving the differences in expression between the different risk groups?
   - Integrate transcription factor binding data



Summary

There is a large group of lncRNAs whose function is mostly unknown. We can focus on a smaller subset that promises more clinical relevance and translation. This sort of filtering can allow us to narrow the list of 1000s of genes to those that are most likely functional while potentially improving patient diagnosis and prognosis.



Citations and Acknowledgements

1. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkologia*, *1A*, 68-77.

2. Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G., & Korbel, J. O. (2015). Data analysis: Create a cloud commons. *Nature*, *523*(7559), 149-151.

Medical Biophysics
UNIVERSITY OF TORONTO

Ontario

Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario.