Interpretable Deep Learning for Chromatin-Informed Inference of Transcriptional Programs Driven by Somatic Alterations Across Cancers

Yifeng Tao¹,†, Xiaojun Ma²,†, Drake Palmer³, Russell Schwartz¹,⁴, Xinghua Lu²,⁵, Hatice Ulku Osmanbeyoglu²,⁶,*

¹Computational Biology Department, School of Computer Science, Carnegie Mellon University
²Department of Biomedical Informatics, School of Medicine, University of Pittsburgh
³Department of Biological Sciences, University of Pittsburgh School of Arts & Sciences
⁴Department of Biological Sciences, Carnegie Mellon University
⁵Department of Pharmaceutical Science, School of Medicine, University of Pittsburgh
⁶Department of Bioengineering, School of Engineering, University of Pittsburgh

†Contributed equally: Y.T., X.M.
Cancers are caused by the perturbations of multiple pathways and transcriptional regulatory programs.
Pan-cancer modeling of regulatory programs

• Similar TFs may be dysregulated across cancers
• Similarities between cancer types can inform new therapies
• Extensive training data from more common tumor types also compensates for smaller sample sizes in similar but rarer cancers (e.g. pheochromocytoma and paraganglioma; PCPG)

Modeling non-linear relationships

- Effects of upstream alterations not equal, e.g., cancer drivers vs. passengers
- Complex interactions between genes, e.g., mutual exclusivity
- Role of genomic alterations is context specific
- Attention mechanism!

https://www.intogen.org
Attention mechanism

- A deep learning method to assign importance weights to input features
  - Widely used in Computer Vision/Natural Language Processing
  - Computed in a contextual manner
Datasets/Approach: Modeling impact of somatic alterations on gene expression programs

Patient somatic alterations

Patient ATAC-seq

Patient RNA-seq

Mutation

Cancer type 1

Mutation

Cancer type 2

Mutation

Cancer type 17

Patients-specific regulatory networks

\[ y_t - X_t \leq s_t + \mu_s + \lambda L_s^2 \]
Approach: interpretable deep learning

- CITRUS
  - Chromatin-informed Inference of Transcriptional Regulators Using Self-attention mechanism
  - Self-attention mechanism
  - Sparse connections constrained by ATAC-seq prior
# Pan-cancer data sources

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAC-seq</td>
<td>410 tumors</td>
</tr>
<tr>
<td>Bladder (BLCA)</td>
<td>371 tumors</td>
</tr>
<tr>
<td>Breast (BRCA)</td>
<td>719 tumors</td>
</tr>
<tr>
<td>Cervical and endocervical (CESC)</td>
<td>267 tumors</td>
</tr>
<tr>
<td>Colon (COAD)</td>
<td>271 tumors</td>
</tr>
<tr>
<td>Esophageal (ESCA)</td>
<td>170 tumors</td>
</tr>
<tr>
<td>Glioblastoma (GBM)</td>
<td>143 tumors</td>
</tr>
<tr>
<td>Head and Neck (HNSC)</td>
<td>475 tumors</td>
</tr>
<tr>
<td>Kidney renal clear cell (KIRC)</td>
<td>357 tumors</td>
</tr>
<tr>
<td>Kidney renal papillary cell (KIRP)</td>
<td>272 tumors</td>
</tr>
<tr>
<td>Liver hepatocellular (LIHC)</td>
<td>336 tumors</td>
</tr>
<tr>
<td>Lung adenocarcinoma (LUAD)</td>
<td>459 tumors</td>
</tr>
<tr>
<td>Lung squamous (LUSC)</td>
<td>430 tumors</td>
</tr>
<tr>
<td>Pheochromocytoma and Paraganglioma (PCPG)</td>
<td>109 tumors</td>
</tr>
<tr>
<td>Prostate (PRAD)</td>
<td>449 tumors</td>
</tr>
<tr>
<td>Stomach (STAD)</td>
<td>373 tumors</td>
</tr>
<tr>
<td>Thyroid (THCA)</td>
<td>216 tumors</td>
</tr>
<tr>
<td>Uterine corpus endometrial (UCEC)</td>
<td>361 tumors</td>
</tr>
</tbody>
</table>


The Cancer Genome Atlas Research Network (TCGA)
ATAC-seq identifies shared and unique epigenetic landscape across cancers

TF motif prediction in ATAC-seq peak regions

CITRUS better predicts gene expression in held-out tumors compared to bilinear models

- Affinity regression (bilinear) vs. CITRUS (deep learning)

Overall attention weights

- Impacts of somatic alterations
Clustering based on inferred TF activity largely recovered the distinction between the major tumor types.
Landscape of mutations and inferred TF activities

CITRUS-inferred TF activities

Somatic mutations

Somatic copy number alterations

Association score := direction * log_{10}(FDR)
Impact of mutations on TFs in breast cancer

• Knock out in silico: different from t-test, simulates the knockout of mutations

<table>
<thead>
<tr>
<th>SM_CASP8</th>
<th>SM_TP53</th>
<th>SM_CDH1</th>
<th>SM_GATA3</th>
<th>SM_MAP2K4</th>
<th>SM_PTN</th>
<th>SM_PIK3CA</th>
<th>SM_PIK3CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM_CDH1</td>
<td>SM_GATA3</td>
<td>...</td>
<td>SM_CDH1</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIF1A</td>
<td></td>
<td></td>
<td>HIF1A↑</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ΔTF

SM_PIK3CA in BRCA

-log10(FDR)

Δ(TF activities)
Impact of PIK3CA mutation on TFs in breast cancer

Impact of TP53 mutation across cancers

Cancers:
- COAD
- HNSC
- UCEC
- LIHC
- STAD
- ESCA
- LUSC
- LUAD
- BLCA
- BRCA
- GBM
- PRAD

TFs:
- ZNF274
- ERG
- TFEB
- HOXC6
- EOMES
- NR5A2
- ID4
- MAFK
- TCF3
- GLIS1
- FLI1
- HOXA13
- AR
- REST
- TP63
- SPDEF
- PPARD
- JDP2
- CREB3L1
- MEOX1
- ARNT2
- LBX2
- USF1
- MSX1
- CEBPZ
- CUX2
- MEFOX2
- ETS2
- E3-Ubiquitin ligase
- mtp53

Comparison of BRCA (P=2.16e-106) and BRCA (P=2.51e-04) activities and expression between WT and SM_TP53.
Conclusion and future work

• CITRUS: deep learning approach modeling transcriptional programs in pan-cancer
• Utilize self-attention mechanism to capture non-linear effects of mutations
• Integrate ATAC-seq as knowledge base

• Further explore potential clinical relevance
Acknowledgments

Dr. Hatice Ulku Osmanbeyoglu
University of Pittsburgh

Dr. Xinghua Lu
University of Pittsburgh

Dr. Russell Schwartz
Carnegie Mellon University

Xiaojun Ma
University of Pittsburgh

Drake Palmer
University of Pittsburgh

Looking for students and postdocs!
Please reach out at osmanbeyogluhu@pitt.edu

NIH
National Institutes of Health

ICI
Innovation in Cancer

Pittsburgh Health Data Alliance
Center for Machine Learning and Health
Carnegie Mellon University

HILLMAN FELLOWS
For Innovative Cancer Research Program

NNCI R00 CA207871

aws