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

Yifeng Tao[1,†], Xiaojun Ma[2,†], Drake Palmer[3], Russell Schwartz[1,4], Xinghua Lu[2,5], Hatice Ulku Osmanbeyoglu[2,6,*]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University

[2]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh

[3]Department of Biological Sciences, University of Pittsburgh School of Arts & Sciences

[4]Department of Biological Sciences, Carnegie Mellon University

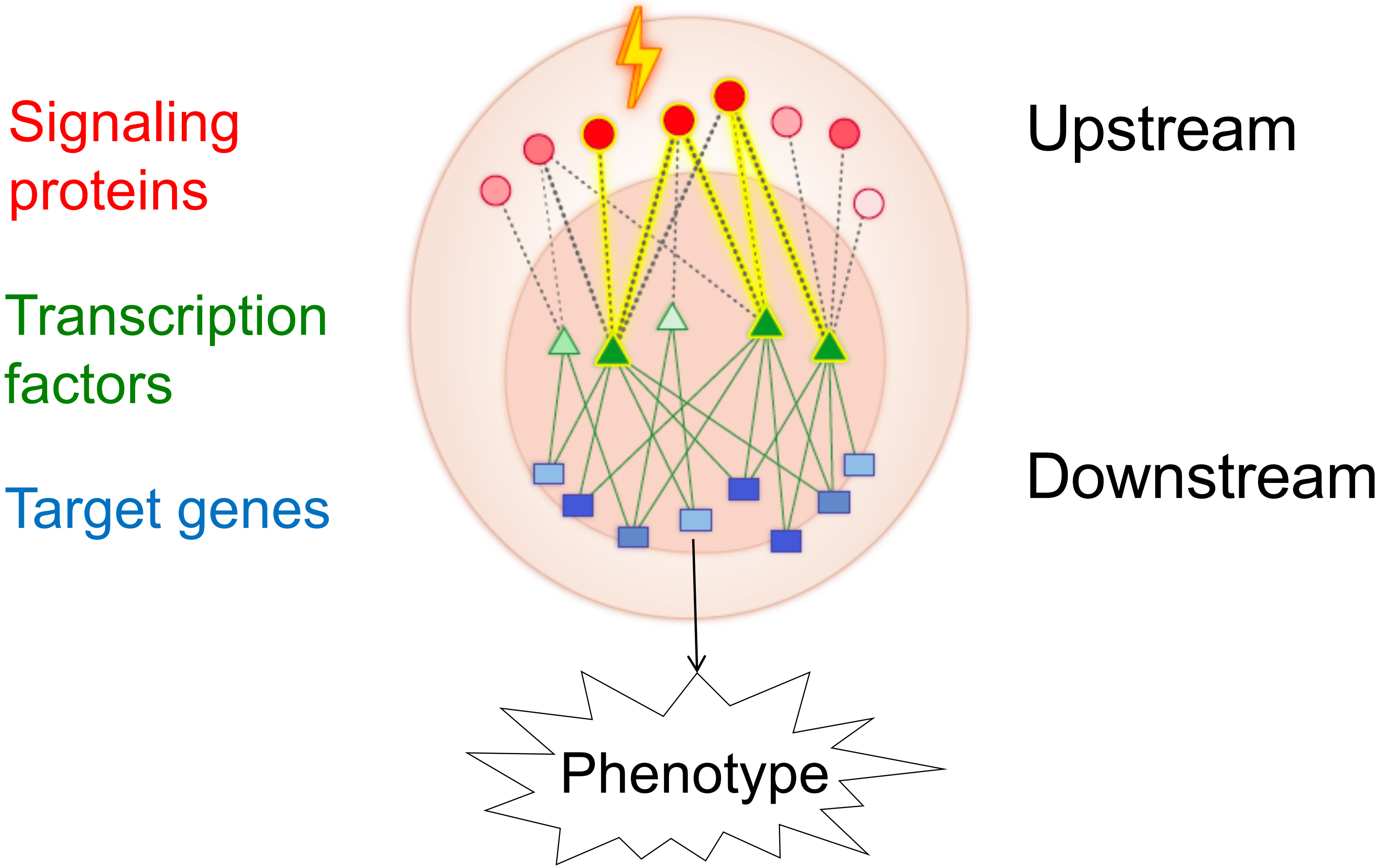
[5]Department of Pharmaceutical Science, School of Medicine, University of Pittsburgh

[6]Department of Bioengineering, School of Engineering, University of Pittsburgh

[†]Contributed equally: Y.T., X.M.

Signaling and transcriptional response

- Cancers are caused by the perturbations of multiple pathways and transcriptional regulatory programs



Signaling proteins

Transcription factors

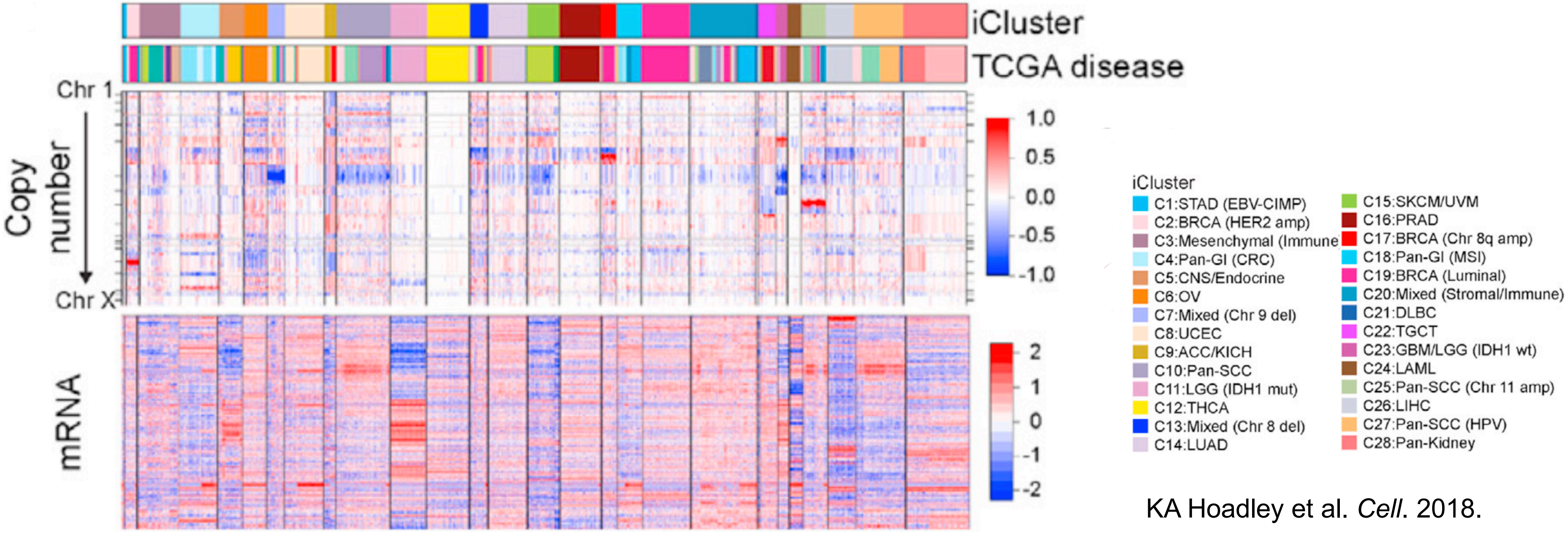
Target genes

Upstream

Downstream

Phenotype

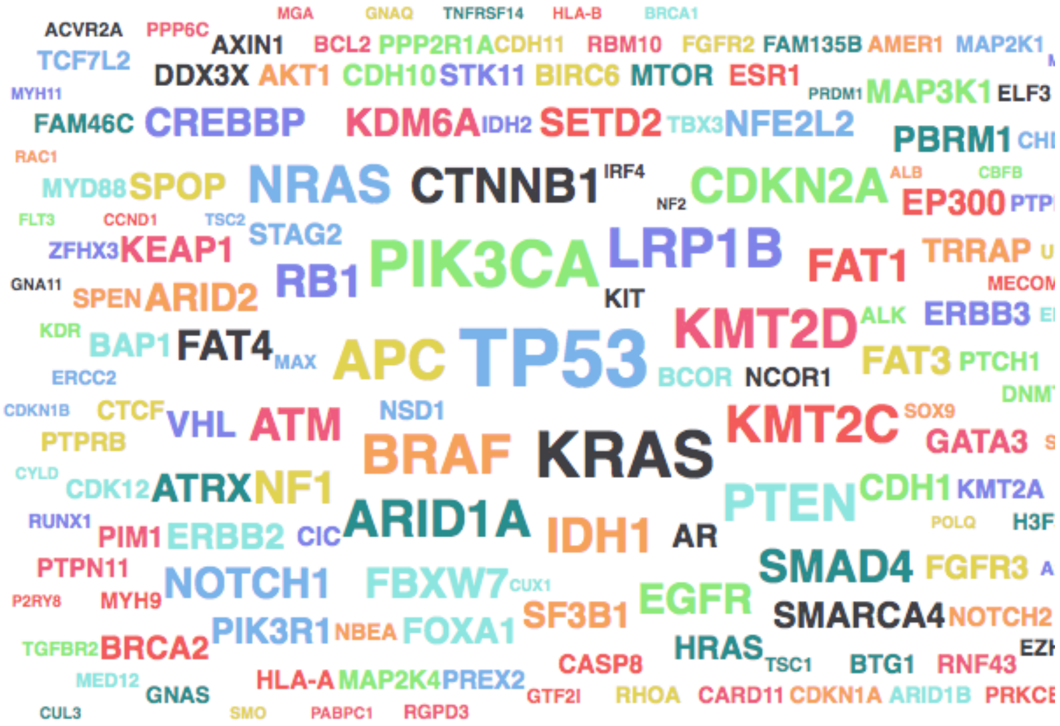
Pan-cancer modeling of regulatory programs



KA Hoadley et al. *Cell*. 2018.

- 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


MGA GNAQ TNFRSF14 HLA-B BRCA1
ACVR2A PPP6C AXIN1 BCL2 PPP2R1A CDH11 RBM10 FGFR2 FAM135B AMER1 MAP2K1
TCF7L2 DDX3X AKT1 CDH10 STK11 BIRC6 MTOR ESR1
MYH11 PRDM1 MAP3K1 ELF3
FAM46C CREBBP KDM6A IDH2 SETD2 TBX3 NFE2L2
RAC1 PBRM1
MYD88 SPOP NRAS CTNNB1 IRF4 NF2 CDKN2A ALB CBFB EP300
FLT3 CCND1 TSC2 STAG2
ZFHX3 KEAP1 PIK3CA LRP1B FAT1 TRRAP
GNA11 SPEN ARID2 RB1 KIT MECOM
KDR BAP1 FAT4 MAX APC TP53 KMT2D ALK ERBB3
ERCC2 BCOR NCOR1 FAT3 PTCH1
CDKN1B CTCF VHL ATM NSD1 KMT2C SOX9
PTPRB BRAF KRAS GATA3
CYLD CDK12 ATRX NF1 PTEN CDH1 KMT2A
RUNX1 PIM1 ERBB2 CIC ARID1A IDH1 AR POLQ
PTPN11 NOTCH1 FBXW7 CUX1 SMAD4 FGFR3
P2RY8 MYH9 EGFR SF3B1 SMARCA4 NOTCH2
TGFBR2 BRCA2 PIK3R1 NBEA FOXA1 HRAS TSC1 BTG1 RNF43
MED12 GNAS HLA-A MAP2K4 PREX2 CASP8
CUL3 SMO PABPC1 RGPD3 GTF2I RHOA CARD11 CDKN1A ARID1B

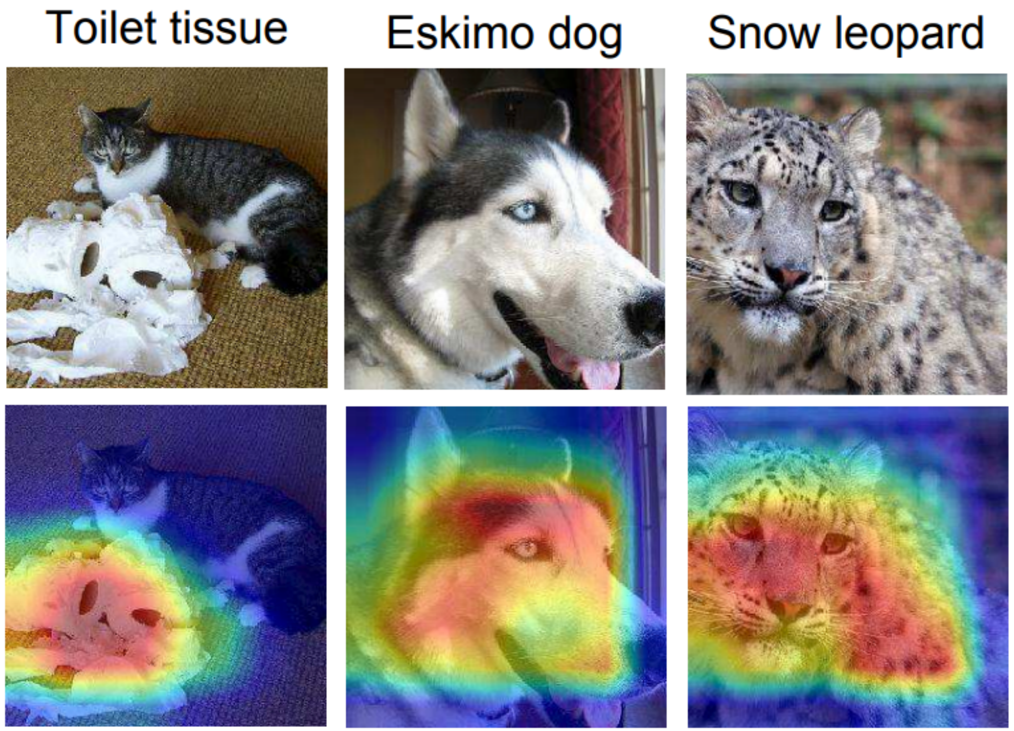
https://www.intogen.org

- 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

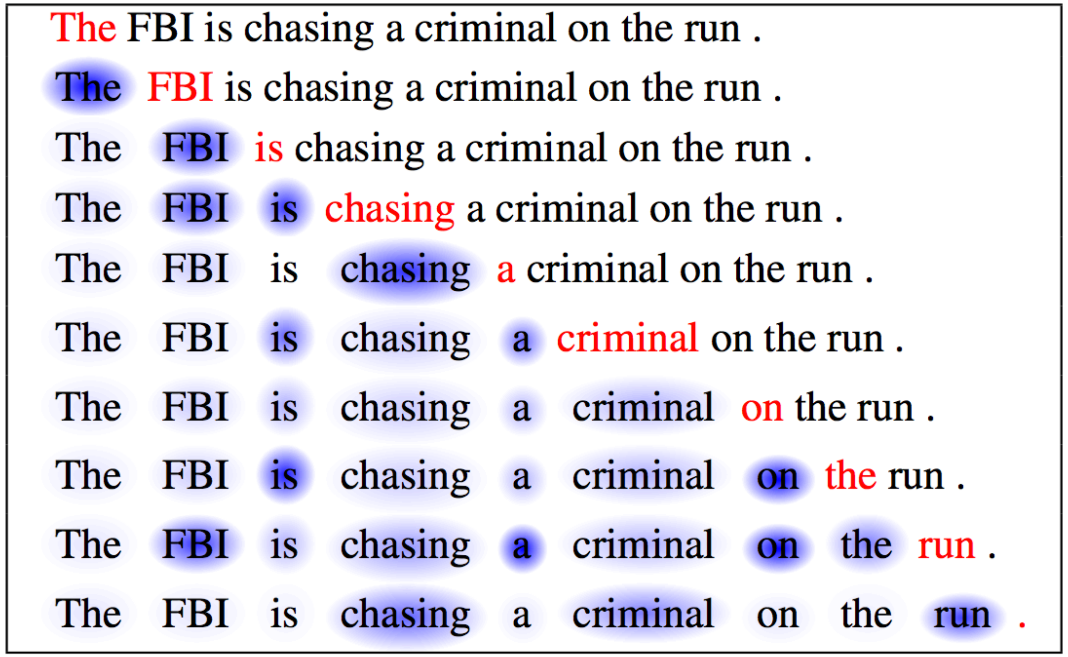
- Attention mechanism!

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



S Woo et al. *ECCV*. 2018.

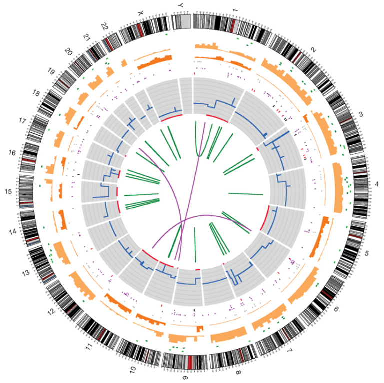


The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .

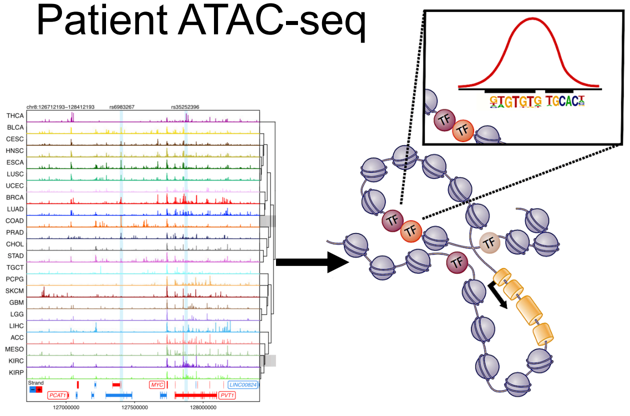
J Cheng et al. *EMNLP*. 2016.

Datasets/Approach: Modeling impact of somatic alterations on gene expression programs

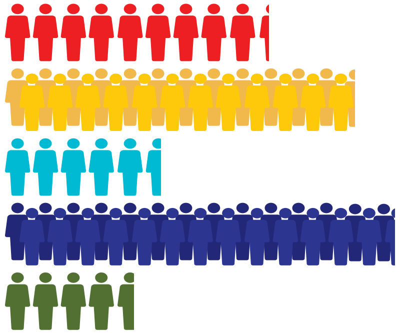
Patient somatic alterations



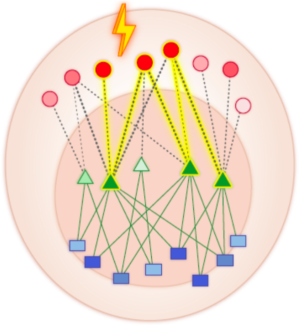
Patient ATAC-seq



Patient RNA-seq

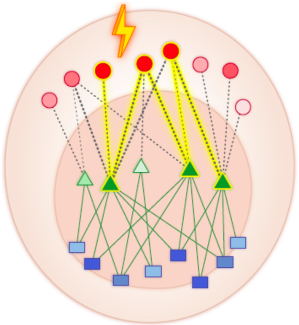


Mutation



Cancer type 1

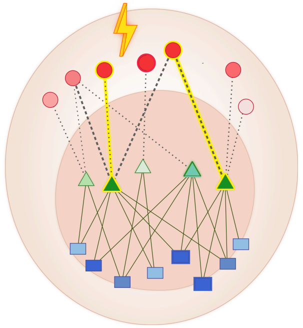
Mutation



Cancer type 2

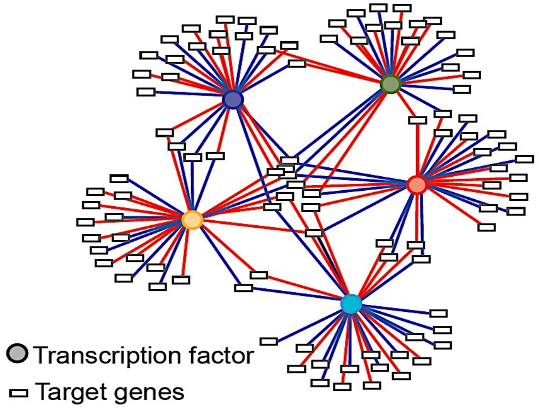
…

Mutation



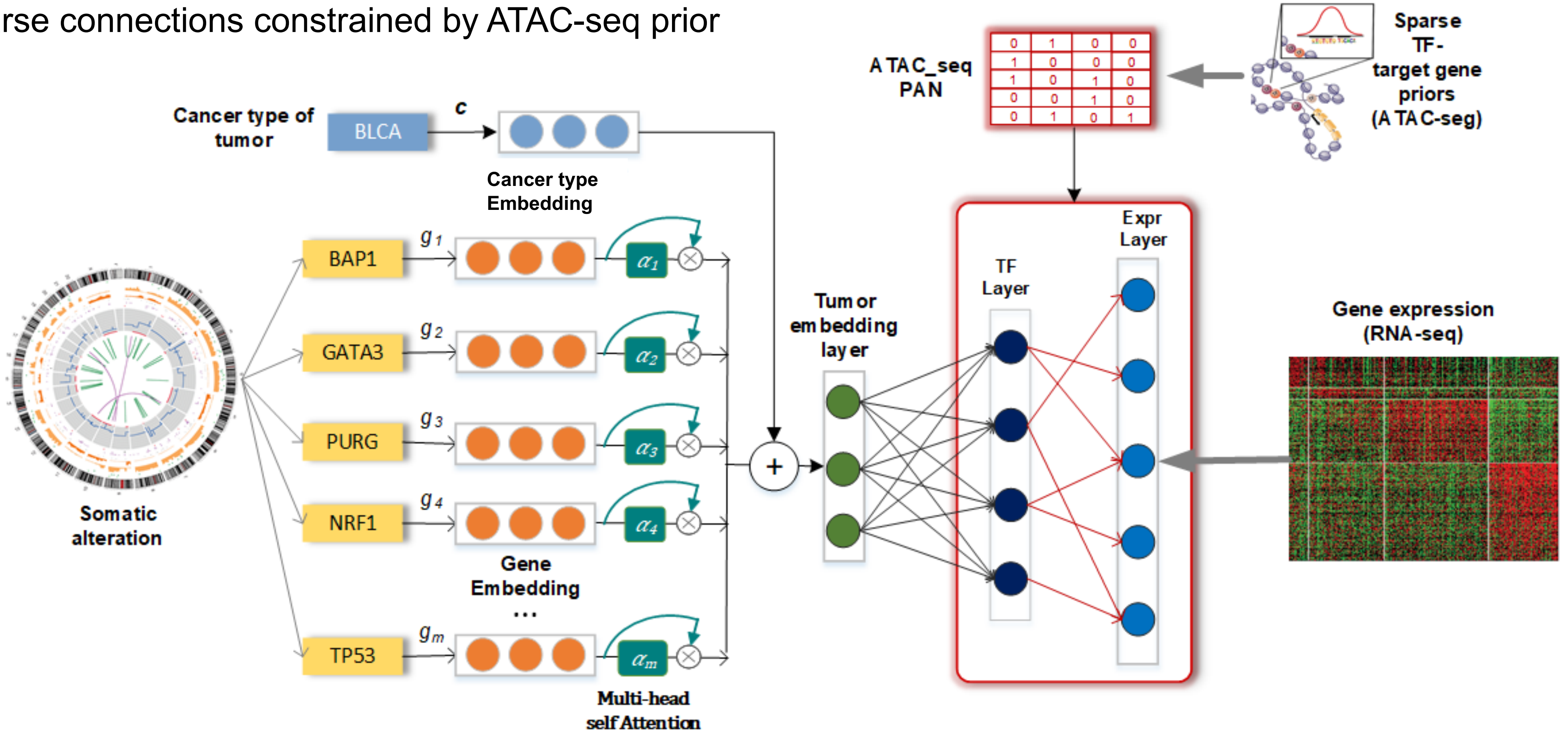
Cancer type 17

Patients-specific regulatory networks



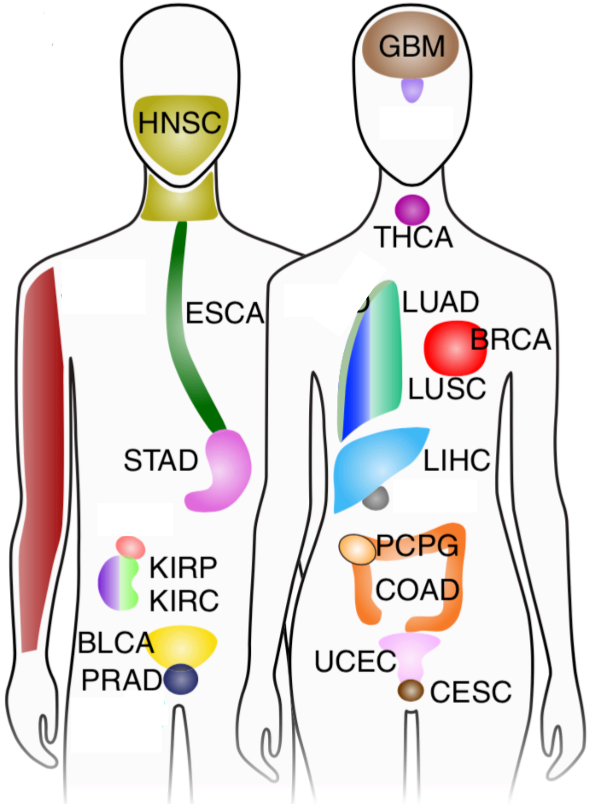
Approach: interpretable deep learning

- CITRUS
    - Chromatin-informed Inference of Transcriptional Regulators Using Self-attention mechanism
    - Self-attention mechanism
    - Sparse connections constrained by ATAC-seq prior



0	1	0	0
1	0	0	0
1	0	1	0
0	0	1	0
0	1	0	1

Pan-cancer data sources

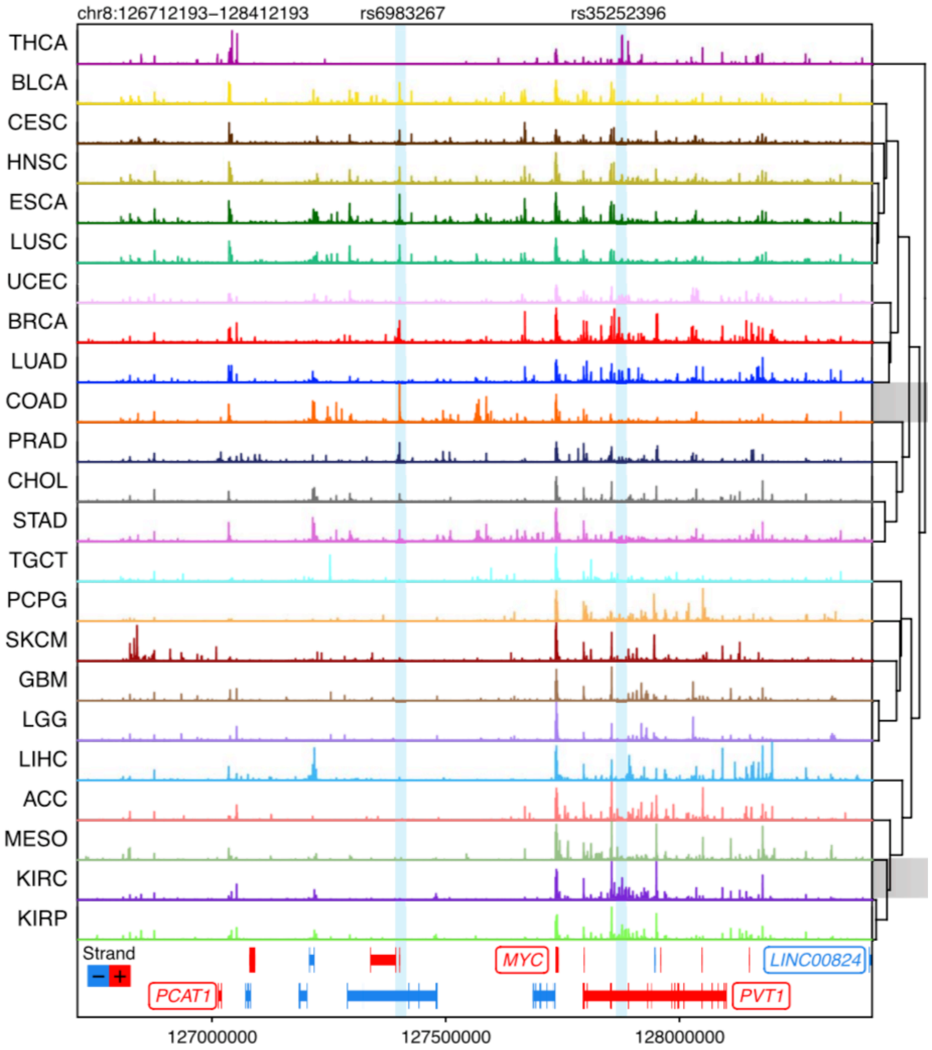


MR Corces et al. *Science*. 2018.

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

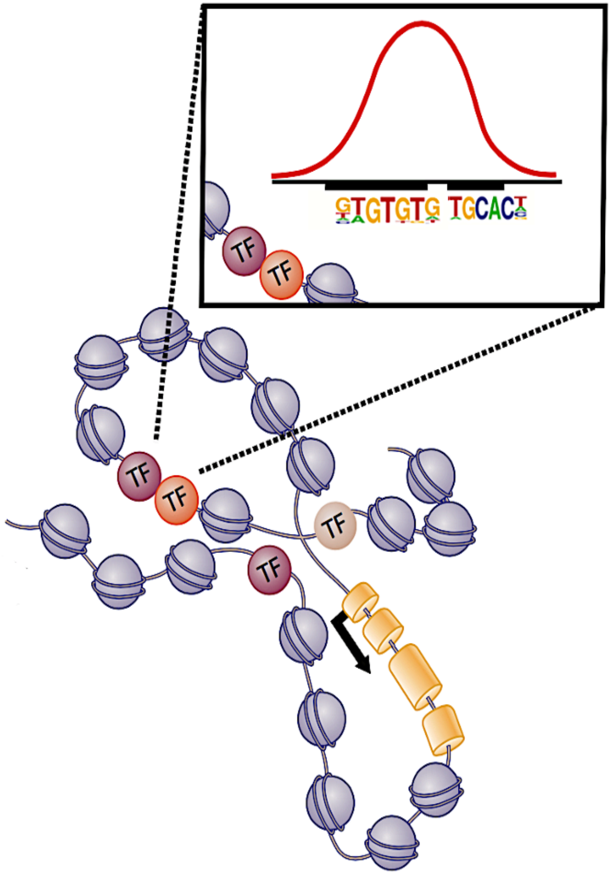
The Cancer Genome Atlas Research Network (TCGA)

ATAC-seq identifies shared and unique epigenetic landscape across cancers



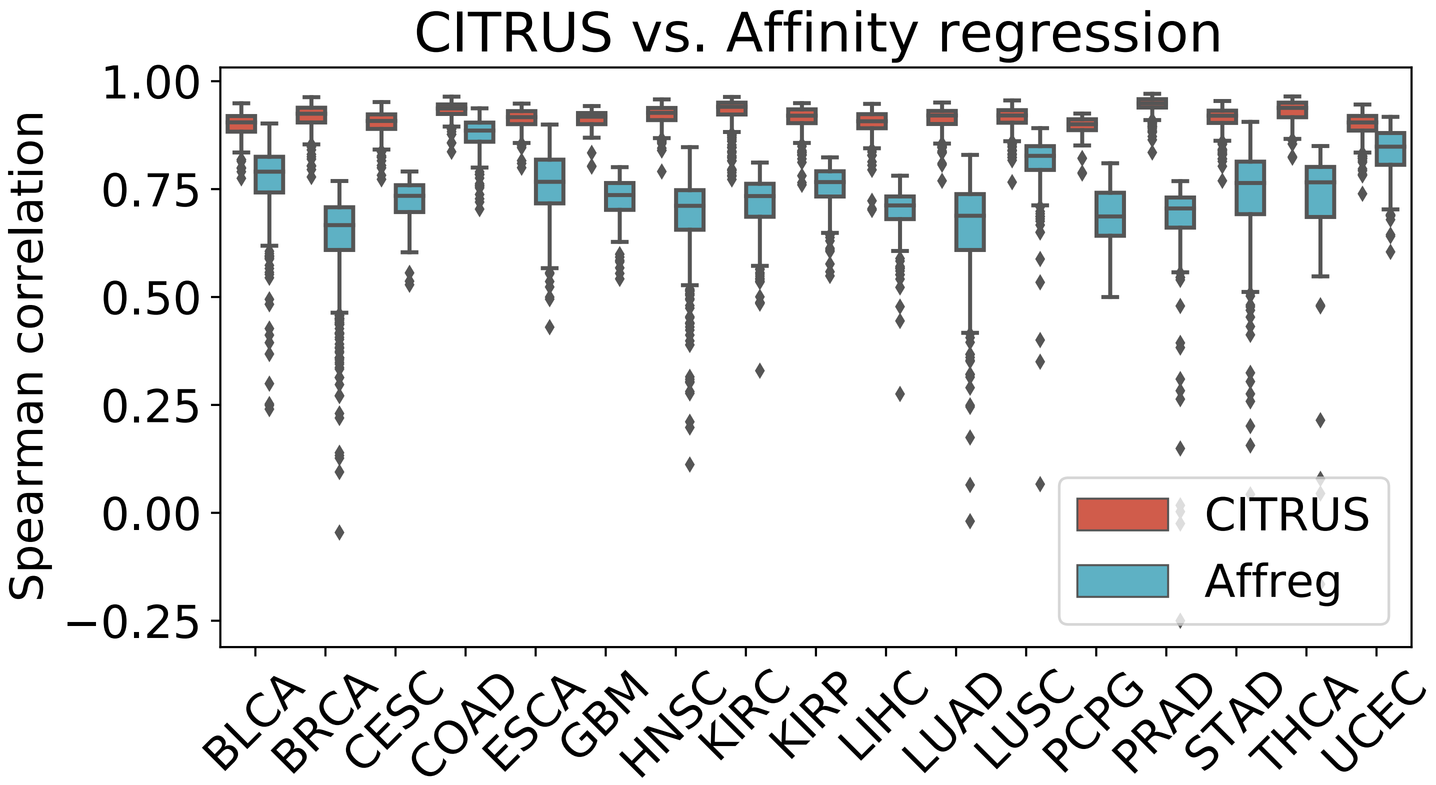
MR Corces et al. *Science*. 2018.

TF motif prediction in ATAC-seq peak regions



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

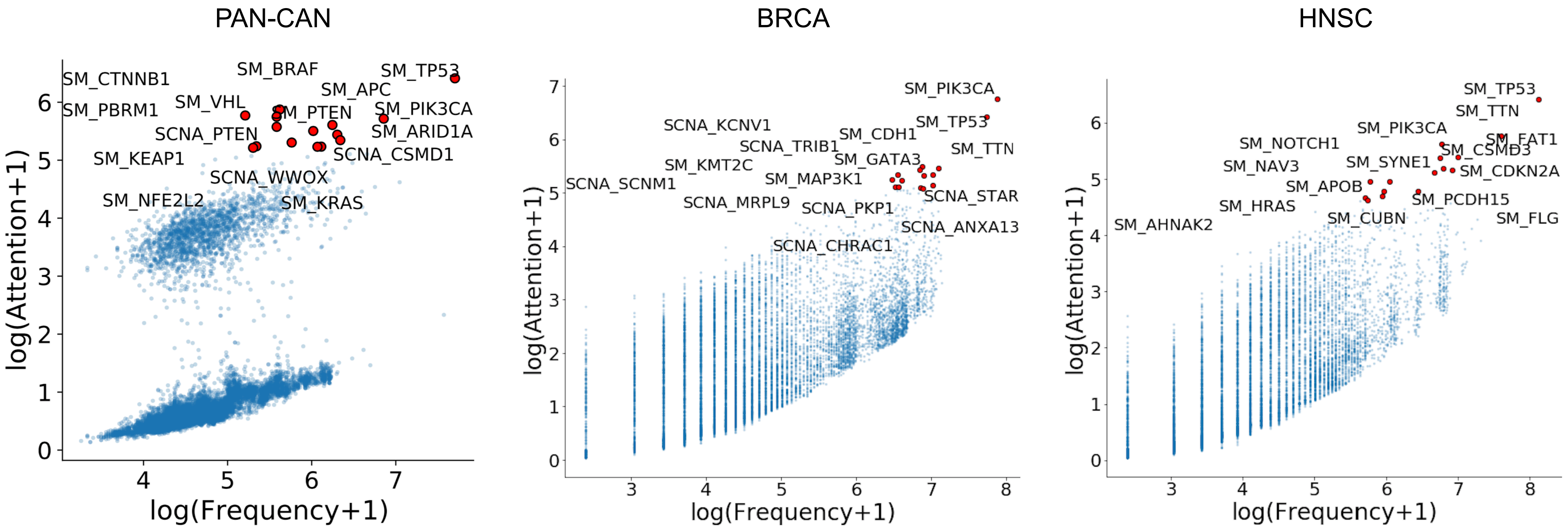
- Affinity regression (bilinear) vs. CITRUS (deep learning)



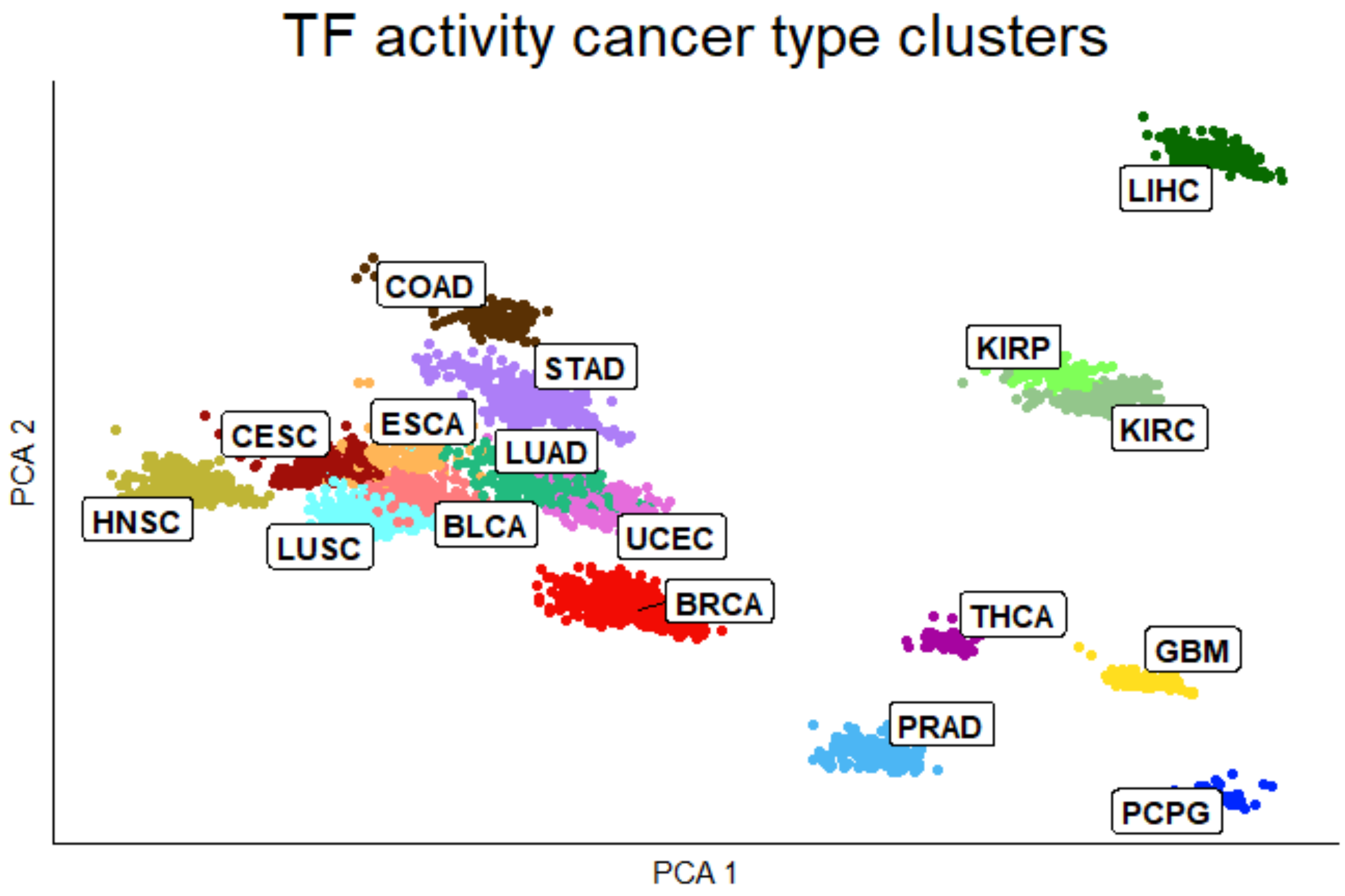
R Pelossof et al. *Nature Biotech*. 2015.
HU Osmanbeyoglu et al. *Nature Comm*. 2017.

Overall attention weights

- Impacts of somatic alterations

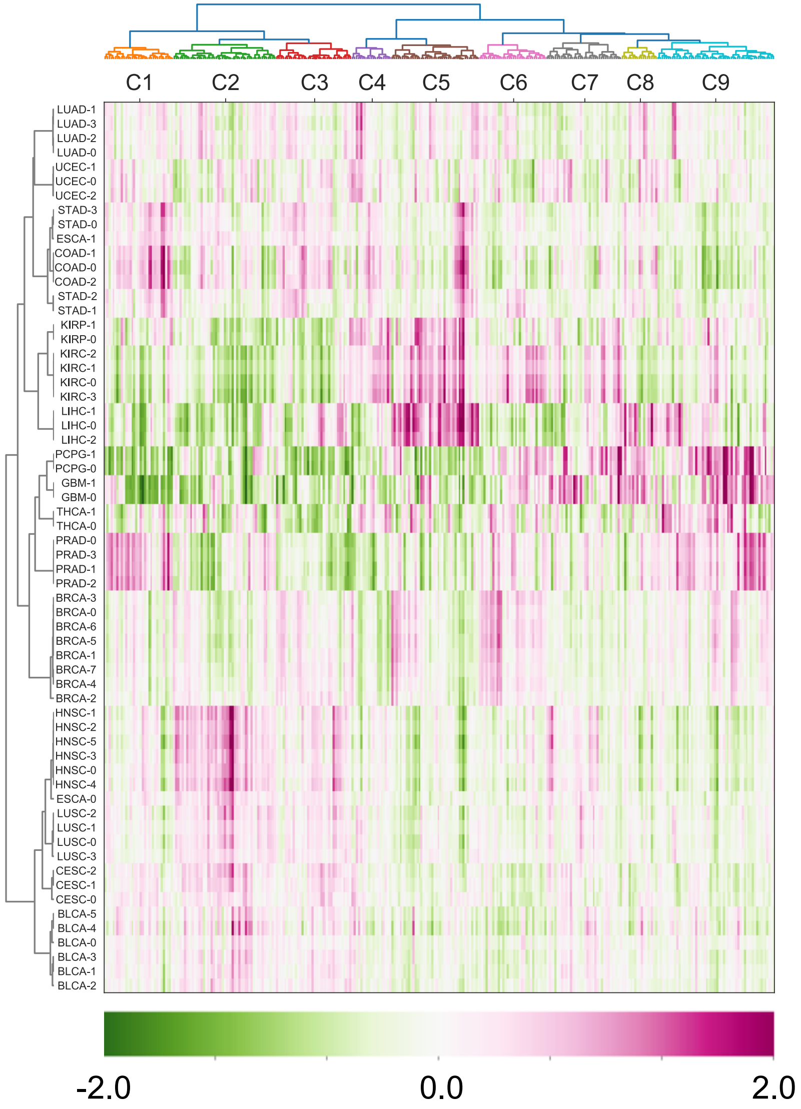


Clustering based on inferred TF activity largely recovered the distinction between the major tumor types



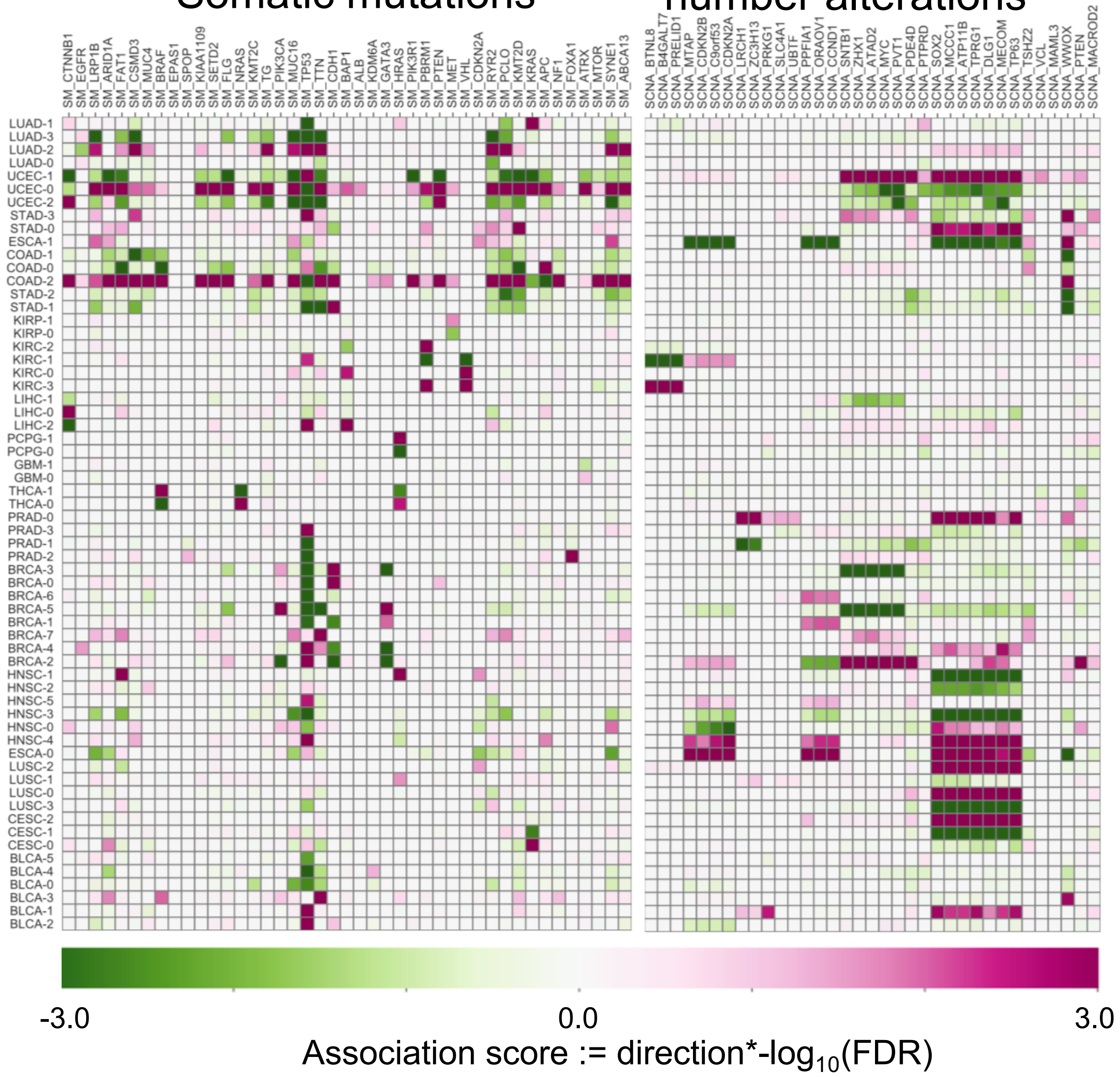
Landscape of mutations and inferred TF activities

CITRUS-inferred TF activities



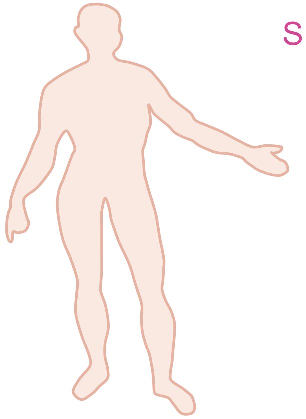
Somatic mutations

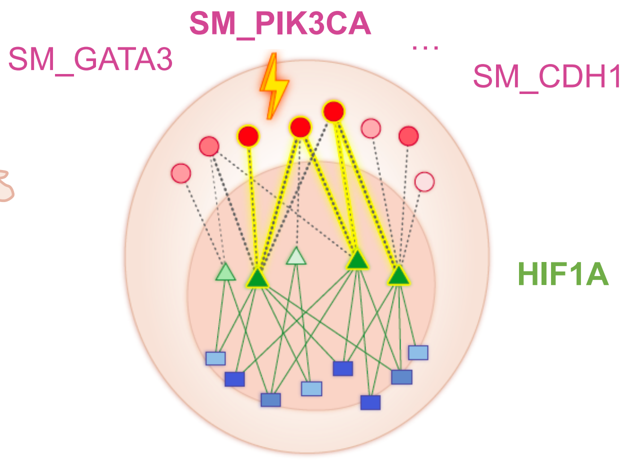
Somatic copy number alterations

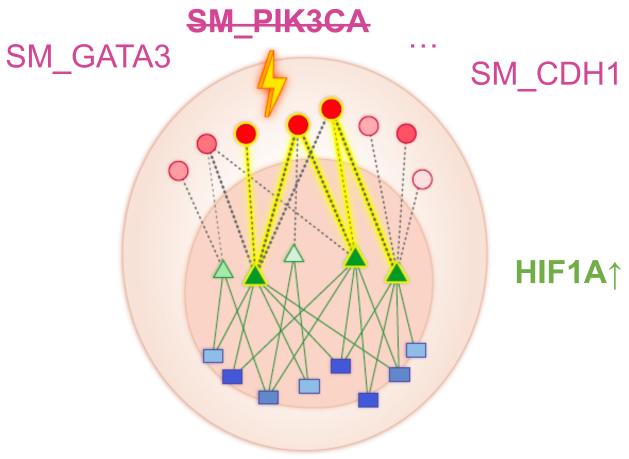


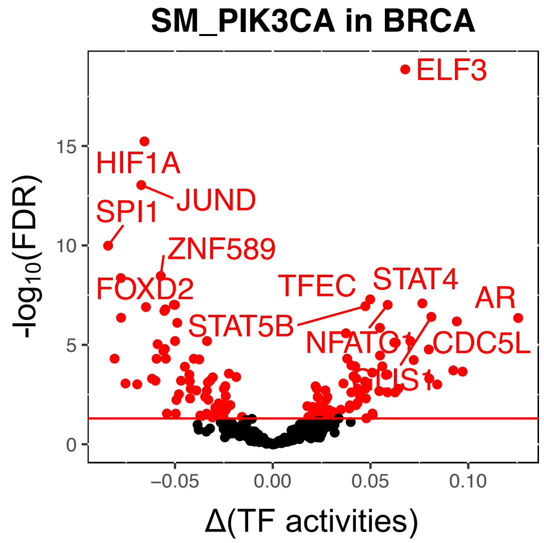
Impact of mutations on TFs in breast cancer

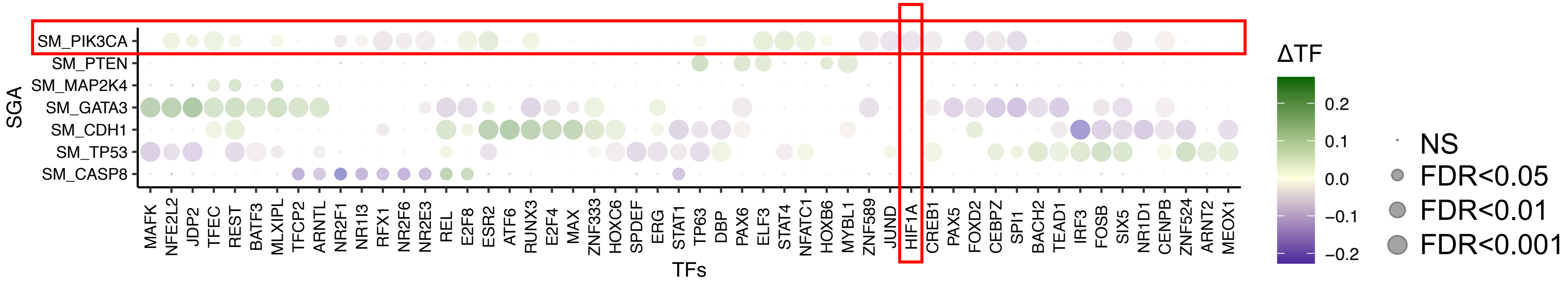
- Knock out *in silico*: different from t-test, simulates the knockout of mutations



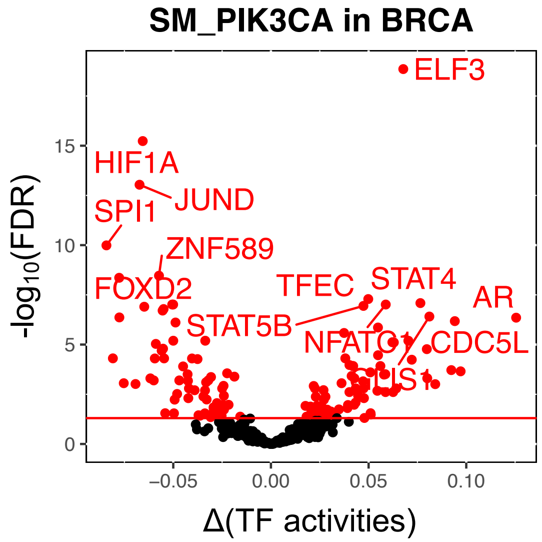


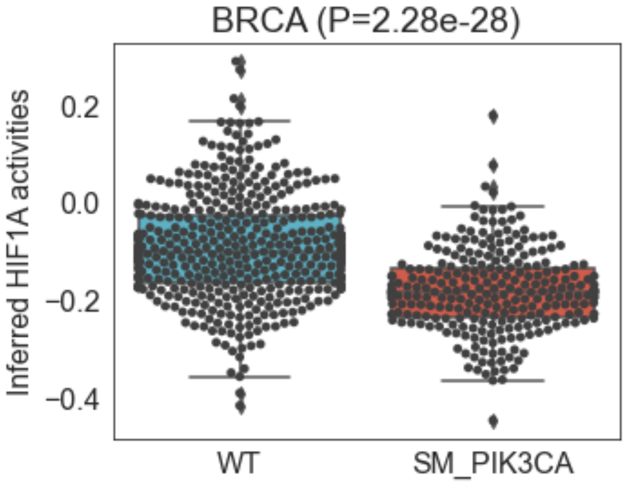


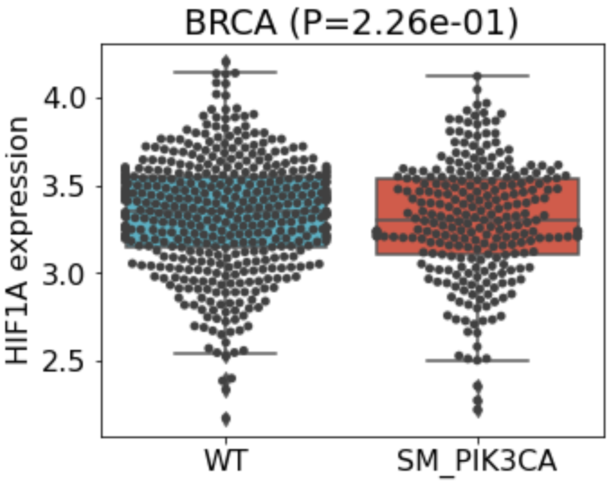


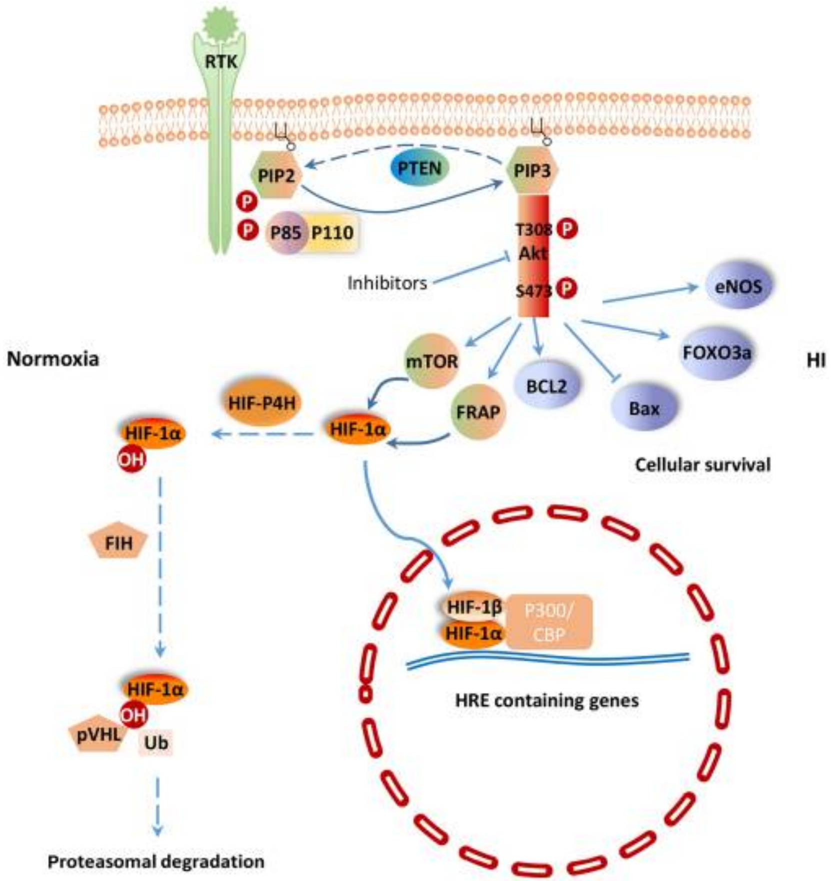


Impact of PIK3CA mutation on TFs in breast cancer



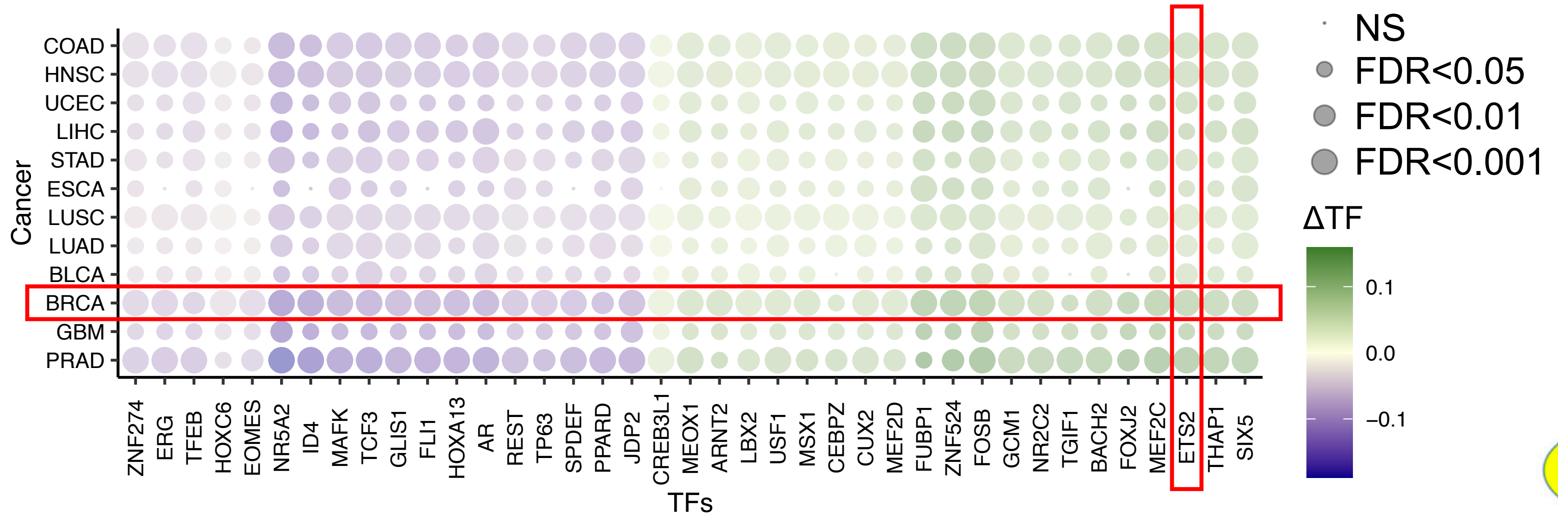


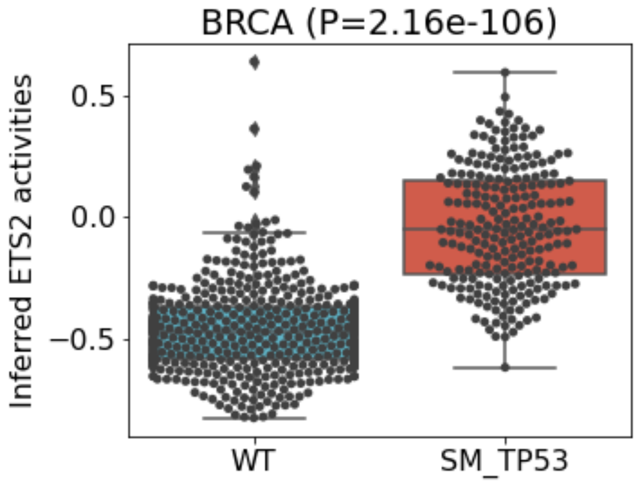


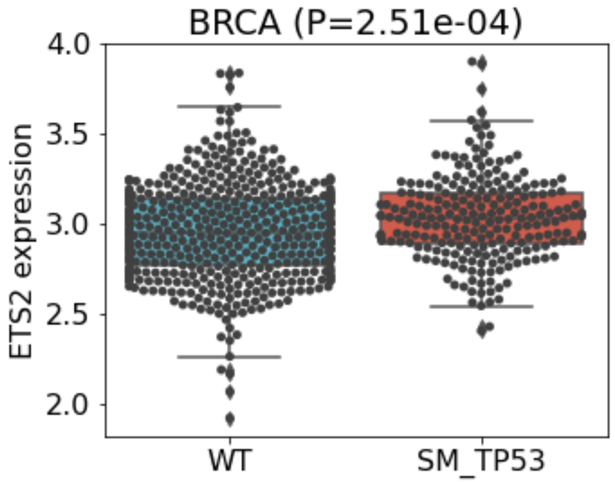


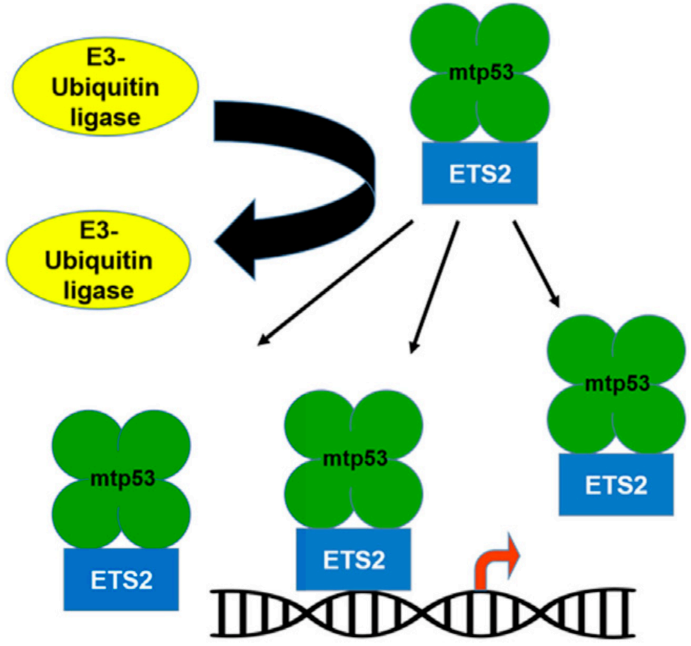
Z Zhang et al. *Mol Med Rep*. 2018.

Impact of TP53 mutation across cancers









LA Martinez. *Front Oncol*. 2016.

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

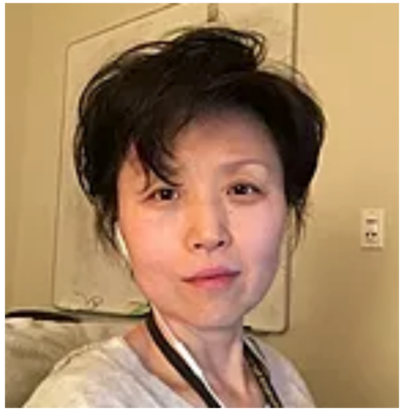
- Further explore potential clinical relevance

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

NNCI R00 CA207871

ICI
INNOVATION IN CANCER INFORMATICS

aws

HILLMAN FELLOWS
For Innovative Cancer Research Program

Pittsburgh Health Data Alliance
Carnegie Mellon University
University of Pittsburgh
UPMC

Center for Machine Learning and Health
Carnegie Mellon University