Interpretable deep learning for chromatin-informed inference of transcriptional programs driven by somatic alterations across cancers

Yifeng Tao[1+], Xiaojun Ma[2,3+], Georgios I. Laliotis[4,5], Adler Guerrero Zuniga[4,5], Drake Palmer[3], Eneda Toska[4,5], Russell Schwartz[1,6], Xinghua Lu[2,7], Hatice Ulku Osmanbeyoglu[2,3,8*]

[1] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
[2] Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, USA
[3] UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, USA
[4] Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[5] Department of Biochemistry and Molecular Biology, Johns Hopkins School of Public Health, Baltimore, MD, USA
[6] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, USA
[7] Department of Pharmaceutical Science, School of Medicine, University of Pittsburgh, Pittsburgh, USA
[8] Department of Bioengineering, School of Engineering, University of Pittsburgh, Pittsburgh, USA

Contributions
These authors contributed equally: Y.T., X.M.

Corresponding author
Correspondence to: Hatice Ulku Osmanbeyoglu (osmanbeyogluhu@pitt.edu)
ORCID ID: 0000-0002-4972-4347

Abstract

Cancer is a disease of gene dysregulation, where cells acquire somatic and epigenetic alterations that drive aberrant cellular signaling. These alterations adversely impact transcriptional programs and cause profound changes in gene expression. Ultimately, interpreting patient somatic alterations within context-specific regulatory programs will facilitate personalized therapeutic decisions for each individual. Towards this goal, we develop a partially interpretable neural network model with encoder-decoder architecture, called **C**hromatin-informed **I**nference of **T**ranscriptional **R**egulators **U**sing **S**elf-attention mechanism (CITRUS), to model the impact of somatic alterations on cellular states and further onto downstream gene expression programs. The encoder module employs a self-attention mechanism to model the contextual impact of somatic alterations in a tumor-specific manner. Furthermore, the model uses a layer of hidden nodes to explicitly represent the state of transcription factors (TFs), and the decoder learns the relationships between TFs and their target genes guided by the sparse prior based on TF binding motifs in the open chromatin regions of tumor samples. We apply CITRUS to genomic, mRNA sequencing and ATAC-seq data from tumors of 17 cancer types profiled by The Cancer Genome Atlas. Our computational framework enables us to share information across tumors to learn patient-specific TF activities, revealing regulatory program similarities and differences between and within tumor types. We show that CITRUS not only outperforms the competing models in predicting RNA expression, but also yields biological insights in delineating TFs associated with somatic alterations in individual tumors. We also validate the differential activity of TFs associated with mutant *PIK3CA* in breast cancer cell line and xenograft models using a panel of PI3K pathway inhibitors.

Introduction

Interplay between complex signaling inputs and genomic transcriptional responses dictates important cellular functions. Dysregulation of this interplay leads to development and progression of disease, most clearly delineated in the context of certain cancers. Cancer cells acquire somatic alterations that drive aberrant signaling which adversely impact transcriptional programs and cause profound changes in gene expression. We still lack a complete understanding of the transcriptional programs and how disruptions in this code affects cellular function in cancer. Interpreting patient somatic alterations within context-specific transcriptional programs can facilitate therapeutic decisions for each individual.

In the last decade, a monumental effort to molecularly profile tumors was undertaken by consortia such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC)[1,2]. These multimodal datasets including gene expression and somatic alterations such as recurrent mutations and copy number variations (CNVs) have enabled the integration of transcriptional states with upstream signaling pathways. Several methods have been developed to connect somatic alterations to a prior network or to gene expression[3-9]. More recently, the Genomic Data Analysis Network generated assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) data for the subset of TCGA samples (~500 patients)[10]. However, so far methods for linking somatic alterations to transcriptional programs across cancers have not incorporated tumor chromatin profiling to encode context-dependent and/or non-linear impacts of transcription factors (TFs) on gene expression. Incorporating DNA sequence information at promoter, intronic and intergenic enhancers using TF motif analysis from tumor ATAC-seq profiles will improve the modeling of transcriptional regulation and delineating the impact of somatic alterations on transcriptional programs.

Deep learning (DL) is a powerful tool for capturing non-linearity. Attention mechanism is a deep learning module that has been widely used in computer vision and natural language processing. In contrast to normal deep learning units, the self-attention mechanism considers the contextual effort of all the input features to each other and assigns different weights of attention to these inputs[11]. In general, the attention mechanism can improve the performance of the DL models or increase the interpretability of the models. More recently, attention mechanisms have also been applied to cancer genomics, including cancer driver detection[12], drug response prediction[13] and predicting base editing outcomes[14]. The genomic impact transformer (GIT) model utilizes the self-attention mechanism to encode the effects of somatic alterations in cancer and uses multi-layer perceptrons to predict differentially expressed genes as the output of the model[12]. The attention mechanism enables it to select the likely driver mutations that lead to downstream phenotypes, such as transcriptome expression levels. However, the GIT model lacks interpretability in the sense it does not model the intermediate TFs during the signaling from somatic alterations to gene expression programs.

In this work, we present **C**hromatin-informed **I**nference of **T**ranscriptional **R**egulators **U**sing **S**elf-attention mechanism (CITRUS), a partially interpretable neural network (NN) model with encoder-decoder architecture, to link somatic alterations to transcriptional programs through modeling the statistical relationships between mutations, CNVs, gene expression and TF-target gene prior information (based on TF binding motif analysis in the open chromatin regions based on tumor ATAC-seq profiling). CITRUS explicitly includes the transcriptional programs in the model, with external knowledge of TF:target-gene priors based on ATAC-seq data. We showed that CITRUS not only outperforms competing models in predicting mRNA expression, but also yields important biological insights in finding dysregulated TFs in individual tumors. We next performed a systematic knock out *in silico* approach to associate frequent somatic alterations with changes in inferred TF activities in each cancer type. This analysis identified key regulators associated with

the major somatic alterations. In particular, we associated *PIK3CA* activating mutations with altered activities of distinct sets of TFs in different cancers. Notably, in cell line and xenograft models of breast cancer, we validated the altered activity of several TFs in the presence of mutant *PIK3CA* with PI3K pathway inhibitors by measuring expression of target genes, confirming the context-specific predictions of our model. These proof-of-principle results suggest a computational strategy for personalized deployment of targeted therapeutics in a pan-cancer setting.

Results

Pan-cancer modeling of regulatory programs

To systematically interpret somatic alterations (SAs) within context-specific transcriptional programs and identify disrupted TFs that drive tumor-specific gene expression patterns across multiple cancer types, we developed CITRUS computational framework (**Fig. 1**). The CITRUS model mimics the biological processing of signaling pathways from SAs to the signaling pathways, to TFs, and finally to the target gene expressions (mRNA levels). Therefore, the model follows an overall encoder-decoder architecture (**Fig. 1**). The encoder module compresses the input SAs into a latent vector variable called a tumor embedding. Then the decoder part first predicts the TF activities from the tumor embedding, and further predicts the TF target-gene expression. We used sparse TF-target gene priors based on tumor ATAC-seq data. Briefly, we started with an atlas of chromatin accessible events derived from the tumor types to be analyzed, using ATAC-seq profiling data ("Methods" section). We represented every gene by its feature vector of TF-binding scores, where motif information was summarized across all promoter, intronic, and intergenic chromatin accessible sites assigned to the gene (see the "Methods" section).

The application of our approach to 17 tumors from TCGA identified key TFs associated with SAs. Our dataset included samples from seventeen different tumor types for which mRNA, somatic mutation, copy number variation and ATAC-seq data were available: bladder urothelial carcinoma (BLCA, n=371), breast cancer (BRCA, n=719), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, n=267), colorectal adenocarcinoma (COAD, n=271), esophageal carcinoma (ESCA, n=170), glioblastoma multiforme (GBM, n=143), head and neck squamous carcinoma (HNSC, n=475), kidney renal cell-clear carcinoma (KIRC, n=357), kidney renal papillary cell carcinoma (KIRP, n=272), liver hepatocellular carcinoma (LIHC, n=336), lung adenocarcinoma (LUAD, n=459), lung squamous cell carcinoma (LUSC, n=430), pheochromocytoma and paraganglioma (PCPG, n=109), prostate cancer (PRAD, n=449), stomach adenocarcinoma (STAD, n=373), thyroid carcinoma (THCA, n=216), and uterine corpus endometrial carcinoma (UCEC, n=361).

For statistical evaluation, we computed the mean Spearman correlation between predicted and measured gene expression profiles on held-out samples (see Methods). We obtained significantly better performance than a regularized bilinear regression algorithm called affinity regression (AR)[15,16] that was trained independently for each cancer type and explains gene expression across tumors in terms of SA status and presence of TF binding sites based on pan-cancer ATAC-seq atlas (**Fig. 2A**).

To identify the SAs that have an impact on gene expression programs, we compared the relationship of overall attention weights (inferred by CITRUS) and the frequencies of somatic alterations (used as the control group) across all cancer types and within a cancer type (**Fig 2B and Supplementary Fig. 1**). In general, the attention weights are correlated with the alteration frequencies of genes. For example, the top altered genes *TP53* and *PIK3CA* had high attention weights. However, our self-attention mechanism assigned low attention weights to many highly

frequently altered genes, indicating these genes can be cancer passengers. Indeed, we found genes with high attention weights were enriched for known cancer drivers from the IntOGen[9] database. We first grouped all the genes into two parts with the threshold of 2 (log(attention+1) >= 2 as the more attended group, and log(attention+1) < 2 as the less attended group). Using the Fisher exact test, we found known cancer drivers were enriched in the highly attended group ($P = 4.48 \times 10^{-41}$) for pan-cancer analysis.

Next, we used CITRUS across tumor types to learn patient-specific TF activities. Clustering of tumors by inferred TF activities as derived from the model largely recovered the distinction between the major tumor types (**Fig. 2C**). In particular, samples with squamous morphology components (BLCA, CESC, ESCA, HNSC, and LUSC) grouped together. Similarly, tumors with tissue or organ similarities or proximity also grouped together. These included neuroendocrine and glioma tumors (GBM and PCPG), clear cell and papillary renal carcinomas (KIRC and KIRP), a gastrointestinal group (COAD, and STAD), breast and endometrial cancer (BRCA and UCEC). We also observed similar clustering with tumor embeddings (**Supplementary Fig 2).**

Next, we assessed TF-tumor type associations by t-test and compare inferred TF activities between samples in a given tumor type vs. those in all other tumor types. We corrected for FDR across TFs and identified significant shared and cancer-specific TFs and the results are shown in **Supplementary Table 1**. **Fig. 3** shows the average TF activity and significance of cancer-specifc TFs across cancer types. For clarity, only the union of 4 top significant TFs per cancer are shown. FUBP1, which regulates *c-Myc* gene transcription, had significantly higher inferred activity in many cancer types including LIHC, HNSC, BLCA, ESCA, CESC, LUSC, PRAD, BRCA, and UCEC. Moreover, in agreement with previous reports, IRF3 activity was significantly higher in GBM[17]; KLF8 had decreased activity in GBM, LIHC and KIRC, consistent with its role in suppressing cell apoptosis during tumor progression[18]; YY1, which regulates various processes of development[19] and had increased activity in CESC and COAD.

CITRUS-inferred TF-activity based cancer subtypes and somatic alteration landscape

Next, we asked whether our method could identify TF activity based subtypes associated with SAs**.** We conducted *k*-means clustering on inferred TF activities for each cancer type to get subtypes, and then conducted hierarchical clustering for both the cancer subtypes and TF activities. **Fig. 4** shows the clustering of subtypes by CITRUS-inferred mean TF activities and corresponding SA associations (see Methods). We observed major variations in mean TF activities across different cancer types, and less but significant variations within each cancer type. These variations within a cancer type may be explained by the distinct mutation or copy number alteration profiles of different subgroups. For example, clustering by TF activities revealed subclasses of endocervical adenocarcinoma (CESC) enriched with *KRAS*; kidney renal cell-clear carcinoma (KIRC) enriched with *VHL*, *BAP1*, *PBRM1* and *TP53*; liver hepatocellular carcinoma (LIHC) enriched with *CTNNB1*, *BAP1* and *TP53*; thyroid carcinoma (THCA) enriched with *NRAS*, *HRAS* and *BRAF* status; pheochromocytoma and paraganglioma (PCPG) enriched with *HRAS* status.

We next developed a systematic statistical approach for modelling the impact of SAs on TF activity, with the eventual goal of deciphering cancer-specific downstream effects of targeted therapies and potentially discovering secondary targets for combination drug strategies. We implemented a knock out *in silico* approach that removes a specific somatic mutation (or copy number variation) *g* from all the tumor samples that carry it to identify a set of TFs predicted to be significantly dysregulated by each SA in each TCGA cancer study (see Methods section). **Fig. 5A** shows the TF activities associated with SAs in UCEC. Our model identified mutations in *PIK3CA*, *PTEN*, *KRAS*, *TP53*, and *CTNNB1* as significantly associated with various TF activities across

UCEC tumors (~66% of tumors have *PTEN* inactivating mutations, ~50% have *PIK3CA* activating mutations, ~38% have *TP53* mutations, ~26% have *CTNNB1*, and ~20% have *KRAS*). UCEC samples with *PTEN* mutations are mutually exclusive with *TP53*, *CTNNB1* and *KRAS* showed distinct patterns of TF activities. Mutations in *PTEN*, which inactivate its phosphatase activity, increase PI3K signaling. TFs associated with *PTEN* mutations involved in cell cycle and differentiation including E2F5, TP63, ELF3, DBP, ZKSCAN3, LHX2, HOXB6, SOX9, DBP, MYLB1, and GLIS1. Whereas, TFs associated with *CTNNB1* mutant status were involved in WNT and TGF-beta signaling including TCF7, TCF7L2, TCF7L1, FOXH1, EMX1, and MYBL1.

Similarly **Fig. 5B** shows the TF activities associated with SAs in BRCA. Our model identified mutations in *PIK3CA*, *PTEN*, *MAP2K4*, *GATA3*, *TP53*, and *CDH1* as significantly associated with various TF activities across tumors. In BRCA, ~36% of tumors have *PIK3CA* activating mutations, ~35% have *TP53* , ~15% have *GATA3*, ~15% have *CDH1*, ~10% have *PTEN*, and ~7% have *MAP2K4* mutations. Activating mutations in *PIK3CA* often occur in one of three hotspot locations (E545K, E542K and H1047R) and promote constitutive signaling though the pathway. TFs associated with *PIK3CA* mutations involved in WNT signaling, epithelial–mesenchymal transition and cancer stem cell transition including ELF3, TFEC, STAT4, STAT5B, NFATC1, GLIS1, CDC5L and AR. BRCA samples with *PIK3CA* and *TP53* mutations are mutually exclusive. Our knock out *in silico* analysis associated different regulators with these mutations. *TP53* mutant tumors are associated with increased activity of TFs that have roles in pro-growth such as ETS2 and FOSB, growth modulatory such as THAP1, CREB3L1, and CEBPZ and development MEF2C/D, MEOX1, MSX1. We also performed similar analyses for other cancer types (**Supplementary Fig. 3**).

We found *TP53* mutation associated with similar TFs across different cancer types (**Supplementary Fig. 4**). *TP53* is one of the most frequently inactivated tumor suppressor genes that suffers from missense mutations in human cancer. These missense mutations express a mutant form of p53 protein. Therefore, the cells retain and express a mutant form of the p53 protein that can either disable other tumor suppressors (e.g., p63 and p73) or enable oncogenes such as ETS2, an ETS family member[20]. Indeed, inferred TF activity of ETS2 was increased in mutant versus WT *TP53* tumors across cancers (**Fig. 5C**); these differences are not as significant at the gene expression level (**Supplementary Fig. 5**).

Experimental validation of oncogenic mutant PI3K-driven TF activity in breast cancer

The PI3K pathway controls proliferation, metabolism, survival and motility and is frequently activated in many cancers, often via mutations in the gene coding for the alpha subunit of the PI3K, *PIK3CA*[24]. The PI3K inhibitor alpelisib was recently approved in metastatic estrogen receptor positive/*PIK3CA* mutant breast cancer[25]. Our analysis associated mutant *PIK3CA* with STAT4, and NFATC1 transcriptional activity in breast cancer patients. To validate the effect of the oncogenic PI3K in the activity of these TFs, we utilized quantitative PCR (qPCR) to measure the expression of canonical target genes in parental and *PIK3CA*^{H1047R} knock-in MCF10a cells treated with a panel of PI3K/AKT inhibitor (the PI3Kα specific inhibitors alpelisib and GDC0077, the PI3Kα/γ/δ inhibitor GDC0032/Taselisib, the pan-AKT inhibitor GDC0068, and the mTOR inhibitor RAD001/Everolimus). Gene expression analysis revealed altered expression of canonical STAT4, and NFATC1 target genes upon mutant *PIK3CA*^{H1047R} compared to parental MCF10a cells. Notably, the expression changes were altered in the opposite direction upon treatment with PI3K/AKT inhibitors (**Fig. 6A, 6B**), but not mTOR, suggesting the robust differential regulation of the transcriptional program of STAT4, and NFATC1 by the PI3K pathway. We also validated these findings in MCF7 (*PIK3CA*^{E545K}) breast cancer cells and in MCF7-derived xenograft tumors treated with vehicle or alpelisib (see Methods section) (**Fig. 6B, 6C**), suggesting

the PI3K-mediated regulation of the transcriptional activity of STAT4, and NFATC1 in cells and tumors.

Discussion

Tumor data sets are a challenging case for regulatory network analysis due to the complexity of cancer genomes (e.g. alterations such as aneuploidy, CNVs, structural variation, and mutations) confounding epigenomic and regulatory sequence analysis. Our method provides a systematic framework for integrating resources on regulatory genomics with tumor expression and mutation and CNV data to better understand expression programs driven by SAs in cancers and infer patient-specific TF activities. Our method uses a deep learning framework called a self-attention mechanism to capture the complex contextual interactions between somatic alterations. For more accurate representation of TF:target-genes relationship, we leveraged ATAC-seq tumor data from patients. Our model is designed to capture flow of information from altered genes (e.g. signaling proteins) to TFs to target genes; the knock out *in silico* analysis is likely to identify causal impacts of SAs. Joint modeling across different tumor types also reveals patient subgroups associated with SAs. We validated CITRUS-predicted TF activity associated with activating *PIK3CA* mutation in BRCA, using vitro and *vivo* models giving a proof-of-principle for the potential therapeutic application of our approach. We showed that for TFs associated with *PIK3CA* mutation, TF target gene expression changed after PI3K inhibitor treatment. In cases where a SA is associated with the activity of a targetable TF or their upstream/downstream component, our analysis may suggest combination therapies.

One limitation of the TF binding motif search approach is that TFs of the same family often share a similar motif and thus are difficult to disambiguate. Therefore, TF motifs encompass the individual activities of multiple TFs. Moreover, co-binding TF binding patterns (e.g., AP-1−IRF complexes) can be biologically more important for fine tuning of gene expression. We will also investigate representing these composite elements as features in our models. Furthermore, we do not represent directionality in the TF:target- gene priors (i.e., whether a gene is activated or repressed by a TF). Hence, negative values of inferred TF activities can be meaningfully interpreted by prior knowledge of whether the TF is acting as an activator or as a repressor. These limitations may confound the interpretation of activities of TFs with context-specific activator and repressor roles. Further, tumor data sets are also a challenging case for regulatory network analysis due to the presence of stromal/immune cells within the tumor and the heterogeneity of cancer cells themselves. However, our framework can be extended to modeling of single-cell RNA-seq or deconvoluted RNA-seq by computational methods as we will report elsewhere.

Despite these limitations, modeling impact of SAs on transcriptional programs may ultimately enable the development of individualized therapies, aid in understanding mechanisms of drug resistance, and allow the identification of biomarkers of response. We anticipate that computational modeling of transcriptional regulation across different tumor types will emerge as an important tool in precision oncology, aiding in the eventual goal of choosing the best therapeutic option for each individual patient.

Methods

Data preprocessing

We downloaded the RNA-seq data for each of the 17 tumor types from the Genomic Data Commons (GDC) portal (https://gdc.cancer.gov/about-data/publications/pancanatlas). The RNA-seq expression data have been log2-transformed into RSEM values. We obtained processed gene-level somatic alterations of each cancer patient from Cai et al.[4]. Briefly, the value in the

tumor for that gene was set to 1 if it hosts a non-synonymous mutation, small insert/deletion, or somatic copy number alteration (deletion or amplification), and otherwise the value was set to 0.

We downloaded the ATAC-seq pancancer peak set from GDC portal (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG)[10]. Using the MEME[21] curated Cis-BP[22] TF-binding motif reference, we scanned pancancer ATAC-seq peak atlas with FIMO[23] to find peaks likely to contain each motif ($P < 10^{-5}$). The final set contained 320 motifs. We associated each peak to its nearest gene in the human genome using the ChIPpeakAnno package[24]. ATAC-seq peaks located in the body of the transcription unit, together with the 100 kb regions upstream of the transcription start site (TSS) and downstream of the 3' end, were assigned to the gene. TF-binding site identification was used to turn each gene's set of assigned ATAC peaks into a feature vector of binding signals by assigning the maximum score of each motif across all peaks to a gene. Then, we created a matrix $C \in \{0,1\}^{kxl}$ that defines a candidate set of associations between TFs and target genes. $C_{i,j} = 1$ when there is a connection from TF j to the gene/RNA i (red lines connecting the TF layer and Exp layer in **Fig. 1**).

CITRUS model

Formally, given a specific tumor t, with the cancer types s, we have a set of SAs in the tumor $\{g_u\}_{u=1}^{m}$, the decoder module first maps each gene g (it is g_u here, but we omit the subscript for the simplicity of notation) into its corresponding gene vector $\boldsymbol{e}_g$. Then the decoder utilizes the multi-head self-attention mechanism to calculate the weighted sum of the both gene embeddings and cancer type embedding:

$$\boldsymbol{e}_t = \boldsymbol{e}_s + \alpha_1 \boldsymbol{e}_1 + \alpha_2 \boldsymbol{e}_2 + \alpha_3 \boldsymbol{e}_3 + \cdots + \alpha_m \boldsymbol{e}_m.$$

The self-attention mechanism takes input of gene embeddings of all the mutated/altered genes, and output the attention weights $\{\alpha_u\}_{u=1}^{m}$ through a sub-neural network. Such attention mechanism captures the contextual impact of co-existing somatic alterations and their complex interactions instead of simpler models. Interested readers can find the mathematical details in the references[12].

The decoder part first infers the TF activities from the encoded tumor embedding $\boldsymbol{e}_t$:

$$\boldsymbol{e}_f = tanh\left(W_f \boldsymbol{e}_t + \boldsymbol{b}_f\right).$$

We used the tanh activation instead of ReLU operation, which is more widely used in deep learning, because it has similar performance to that of ReLU in our model and generates more biologically meaningful results, e.g., distribution of TFs $\boldsymbol{e}_f$. Finally CITRUS predicts the cancer type specific mRNA expressions from the TF activities:

$$\hat{y} = \sigma\left(W e_f + b_r\right),$$

where W corresponds to the sparse TF:target-gene matrix constrained by the prior $C \in \{0,1\}^{k \times l}$. More specifically, in order to integrate priors into our model, W share the same shape with prior C, and $W_{i,j}$ is allowed to be nonzero only when $C_{i,j} = 1$, and $W_{i,j}$ is constrained to be non-negative value. The loss function to be optimized is thus:

$$MSE(y, \hat{y})$$

One might use other common approaches to integrate the priors of C into the W, i.e., by applying a Gaussian prior to the W, which is equivalent to adding an additional penalty to the loss function $\sum_{i,j:C_{i,j}=0}(W)_{i,j}^2$. However, this "soft" constraint tends to generate less stable TF layers across different runs of training compared to the "hard" constraints shown in our present work.

We introduced additional dropout operations with dropout rate of 0.2 after the input layer, activated tumor embedding layer, and activated TF layer to increase the model robustness to noise and prevent overfitting.

Training and evaluation: We implemented the CITRUS through the PyTorch package (https://pytorch.org/) and trained through Adam optimizer with default parameters except the learning rate[15] and weight decay. We set learning rate to be 1×10^{-3}, and weight decay to be 1×10^{-5}. For each fold of training, we used early stopping with patience of 30 steps to stop training.

For statistical evaluation, we computed the mean Spearman correlation (ρ) between predicted and measured gene expression profiles on held-out patients for each tumor type. We splitted the dateset into training (40%), validation (20%) and test sets (20%). For CITRUS model, we utilized the training and validation sets to tune hyperparameters such as learning rate and training steps, and then evaluated on the held-out test sets. For affinity regression (see below), we seperated datasets by cancer type, and conducted 5-fold cross-validation to tune hyperparameters for each type on training and validation sets, and then applied the trained model with selected hyperparameters to the test set for performance evaluation. In order to increase the stability for the analysis of inferred TF activities, we ensembled multiple CITRUS models with different random seeds, by bootstrapping the model for 10 times, and integrate the TF layer by taking the average of 10 trials to increase the stability of inference.

Training the affinity regression models

AR is an algorithm for efficiently solving a regularized bilinear regression problem[15,25] , defined here as follows. For a data set of *M* tumor samples profiled using RNA-seq with *N* genes, we let $\mathbf{Y} \in R^{NxM}$ be the log 10 gene expression profiles of tumor samples. Each column of **Y** corresponds to an RNA-seq experiment for a cancer type. We define each gene's TF attributes in a matrix $\mathbf{D} \in R^{NxQ}$, where each row represents a gene and each column represent the hit vector for a TF, that is, the bit vector indicating whether there is binding site for the TF of each gene based on ATAC-seq data. We define the SA attributes of tumor samples as a matrix $\mathbf{P} \in R^{MxS}$ where each row represents a tumor sample and each column represents the somatic alteration status for the tumor sample. We set up a bilinear regression problem to learn the weight matrix $\mathbf{W} \in R^{QxS}$ on paired of TF and SA features:

$$\mathbf{DWP}^{\mathrm{T}} \sim \mathbf{Y}$$

We can transform the system to an equivalent system of equations by reformulating the matrix products as Kronecker products

$$\mathbf{DWP}^{\mathrm{T}} \approx \mathbf{Y} \;\Leftrightarrow (\mathbf{P} \otimes \mathbf{D})\, \mathrm{vec}(\mathbf{W}) \approx \mathrm{vec}(\mathbf{Y})$$

where $\otimes$ is a Kronecker product and vec(.) is a vectorizing operator that stacks a matrix and produces a vector, yielding a standard (if large-scale) regression problem. Full details and a derivation of the reduced optimization problem are provided elsewhere[15].

Contextual impact of somatic alterations with knock out *in silico* analysis

We implemented a knock out *in silico* approach that removes a specific somatic mutation (or copy number variation) *g* from all the tumor samples that carry it. The new knocked-out SA profiles and CITRUS-inferred TF activities generate the "wild type" corpus that does not contain this alteration *g*. In contrast, all the original samples containing the alteration *g* serve as the "mutant/altered" group. We finally conducted the t-test between the mutant group and wild type group to evaluate the contextual impact of mutation *g*. The knockout *in silico* is different from the normal t-test, since it captures contextual effects of mutations through the non-linear attention module of CITRUS, and provides a perfect experiment/control setting where all mutations are the

same but mutation *g*. For a complex genotype, the model explains TF regulator activity across tumors. We then corrected for multiple hypotheses across regulator models, treating inferred TF activities as separate groups of tests.

Selection of TF targets for validation experiments

The selection of the canonical target genes for each TF, was performed using the Cistrome Cancer Transcription Factor targets tool for BRCA, in the Cistrome project browser[26]

Cell lines and PI3K/AKT/mTOR inhibitors

MCF10A Isogenic parental and *PIK3CA*^{H1047R} heterozygous mutants were purchased from Horizon. MCF-10A cells were maintained in DF-12 media supplemented with 5% filtered horse serum (Invitrogen), EGF (20 ng/μL) (Sigma), hydrocortisone (0.5 mg/mL) (Sigma), cholera toxin (100 mg/mL) (Sigma), insulin (10 μg/mL) (Sigma), and 1% penicillin/streptomycin. Cells were used at low passages and were incubated at 37°C in 5% CO2. MCF10A parental and mutant cells were seeded in 6-multiwell plates in regular culture conditions to allow correct attachment and ensure ~75% confluency at harvesting day. 24 hours after seeding, cells were washed twice with PBS before adding the starvation media (without serum, EGF and insulin). Where indicated, cells were treated with DMSO as control or alpelisib (1μM), taselisib (100nM), GDC0077 (100nM), GDC0068/ipatasertib (1μM) or RAD001/everolimus (100nM) for 4h.

MCF7 were purchased from ATCC (ATCC HTB-22) and grown in DMEM/F12 supplemented with 10% FBS, penicillin/ streptomycin 1% under standard conditions.

The PI3Ka-specific inhibitors alpelisib and GDC0077, the PI3Kα/γ/δ taselisib, the pan-AKT inhibitor GDC0068/ipatasertib, the mTORC1 inhibitor RAD001/everolimus were purchased (Selleckchem). All the cells were tested regularly for mycoplasma, to ensure experiments in mycoplasma-free cultures.

In vivo studies

For the MCF7 xenograft study, 0.18 mg/90d-release oestrogen pellets were implanted into 6-week-old female NOD *scid* gamma mice 3 days prior to the tumor cell transplantation. Ten million MCF7 cells per mouse were subcutaneously transplanted.

RNA extraction and RT–qPCR

RNA was isolated using the QIAGEN RNeasy Kit and retrotranscription was performed using the iScript cDNA synthesis kit from Bio-Rad, following manufacturer's instructions. cDNA was amplified by real time quantitative PCR in a Applied Biosystems Real-Time PCR system, using SYBR Select Master Mix from Applied Biosystems. Each sample was run in technical triplicates and each experiment was performed in triplicate.

Statistical analysis

Statistical tests were performed with the R statistical environment and *Python*. For population comparisons of inferred TF activities, we performed Student's t-test and determined the direction of shifts by comparing the mean of two populations. We corrected raw P-values for multiple hypothesis testing based on two methods: Bonferroni and false discovery rate (BH method).

Association score between TF activity subtypes and frequent SAs. For each somatic mutation or copy number variation, we calculated the *p*-value of its frequency in a cancer subtype is different from that in other subtypes using Fisher's exact test. The *p*-value was further adjusted through

FDR across subtypes. To identify the relative frequency of a SA in a subtype, we defined the association score, which is the product of relative frequency direction and $-\log_{10}$FDR.

Data Availability

ATAC-seq data is available in a public repository from Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG). RNA-seq gene expression data, somatic mutation, copy number variation data and clinical data are available in a public repository from TCGA's Firehose data run (https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata). Only the samples 'whitelisted' by TCGA for the Pan-Cancer Analysis Working Group were used in the study. For our analysis, we restricted to samples with parallel RNA-seq, somatic mutation and GISTIC copy number data.

Code Availability

The software for CITRUS is available from https://github.com/osmanbeyoglulab/CITRUS

References

1 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).

2 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

3 Wang, Z. *et al.* Cancer driver mutation prediction through Bayesian integration of multi-omic data. *PLoS One* **13**, e0196939, doi:10.1371/journal.pone.0196939 (2018).

4 Cai, C. *et al.* Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLoS Comput Biol* **15**, e1007088, doi:10.1371/journal.pcbi.1007088 (2019).

5 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).

6 Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757-2764, doi:10.1093/bioinformatics/btt471 (2013).

7 Basha, O., Mauer, O., Simonovsky, E., Shpringer, R. & Yeger-Lotem, E. ResponseNet v.3: revealing signaling and regulatory pathways connecting your proteins and genes across human tissues. *Nucleic Acids Res* **47**, W242-W247, doi:10.1093/nar/gkz421 (2019).

8 Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* **13**, R124, doi:10.1186/gb-2012-13-12-r124 (2012).

9 Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640-i646, doi:10.1093/bioinformatics/bts402 (2012).

10 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, doi:10.1126/science.aav1898 (2018).

11 Vaswani, A. *et al.* in *Advances in neural information processing systems* 5998-6008 (2017).

12 Tao, Y., Cai, C., Cohen, W. W. & Lu, X. From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. *Pac Symp Biocomput* **25**, 79-90 (2020).

13 Cadow, J., Born, J., Manica, M., Oskooei, A. & Rodriguez Martinez, M. PaccMann: a web service for interpretable anticancer compound sensitivity prediction. *Nucleic Acids Res* **48**, W502-W508, doi:10.1093/nar/gkaa327 (2020).
14 Marquart, K. F. *et al.* Predicting base editing outcomes with an attention-based deep learning algorithm trained on high-throughput target library screens. *Nat Commun* **12**, 5114, doi:10.1038/s41467-021-25375-z (2021).
15 Pelossof, R. *et al.* Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat Biotechnol* **33**, 1242-1249, doi:10.1038/nbt.3343 (2015).
16 Osmanbeyoglu, H. U., Toska, E., Chan, C., Baselga, J. & Leslie, C. S. Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat Commun* **8**, 14249, doi:10.1038/ncomms14249 (2017).
17 Tarassishin, L. & Lee, S. C. Interferon regulatory factor 3 alters glioma inflammatory and invasive properties. *J Neurooncol* **113**, 185-194, doi:10.1007/s11060-013-1109-3 (2013).
18 Wang, M. D. *et al.* A novel role of Kruppel-like factor 8 as an apoptosis repressor in hepatocellular carcinoma. *Cancer Cell Int* **20**, 422, doi:10.1186/s12935-020-01513-3 (2020).
19 Zhang, Q., Stovall, D. B., Inoue, K. & Sui, G. The oncogenic role of Yin Yang 1. *Crit Rev Oncog* **16**, 163-197, doi:10.1615/critrevoncog.v16.i3-4.30 (2011).
20 Martinez, L. A. Mutant p53 and ETS2, a Tale of Reciprocity. *Front Oncol* **6**, 35, doi:10.3389/fonc.2016.00035 (2016).
21 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).
22 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).
23 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
24 Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237, doi:10.1186/1471-2105-11-237 (2010).
25 Osmanbeyoglu, H. U., Pelossof, R., Bromberg, J. F. & Leslie, C. S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res* **24**, 1869-1880, doi:10.1101/gr.173039.114 (2014).
26 Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* **47**, D729-D735, doi:10.1093/nar/gky1094 (2019).

Acknowledgements

The results published here are in whole or part based on data generated by The Cancer Genome Atlas project established by the NCI and NHGRI (accession number: phs000178.v7p6). Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/. We thank Jacob Stewart-Ornstein for helpful discussions. This work was supported by NIH award from R00CA207871. Y.T. was partially supported by Center for Machine Learning and Health Fellowship in Digital Health from Carnegie Mellon University. R.S. was partially supported by NIH award R21CA216452, by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG010589, by the Pennsylvania Dept. of Health under award FP00003273, and by the Mario Lemieux Foundation. This work was was also partially supported by the AWS Machine Learning Research Awards granted to R.S. The content does not necessarily represent the official views of the above funding agencies. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

Ethics declarations

Competing interests

The authors declare no competing financial interests.

Author contributions

H.U.O. conceived general ideas, supervised implementation, planned validation, and interpreted results. X.M. developed novel machine learning models and implemented validation experiments. Y.T. developed novel machine learning models and planned validation experiments. D.P. collected and preprocessed ATAC-seq data of the study. X.L. and R.S. helped to conceive general ideas, and interprete results. G.L. and A.G.Z performed the experimental validation and wrote the experimental validation section. E.T. supervised the experimental validation. H.U.O. and Y.T. wrote the manuscript. X.M., X.L. and R.S. contributed to reviewing, and editing the manuscript.

Figures

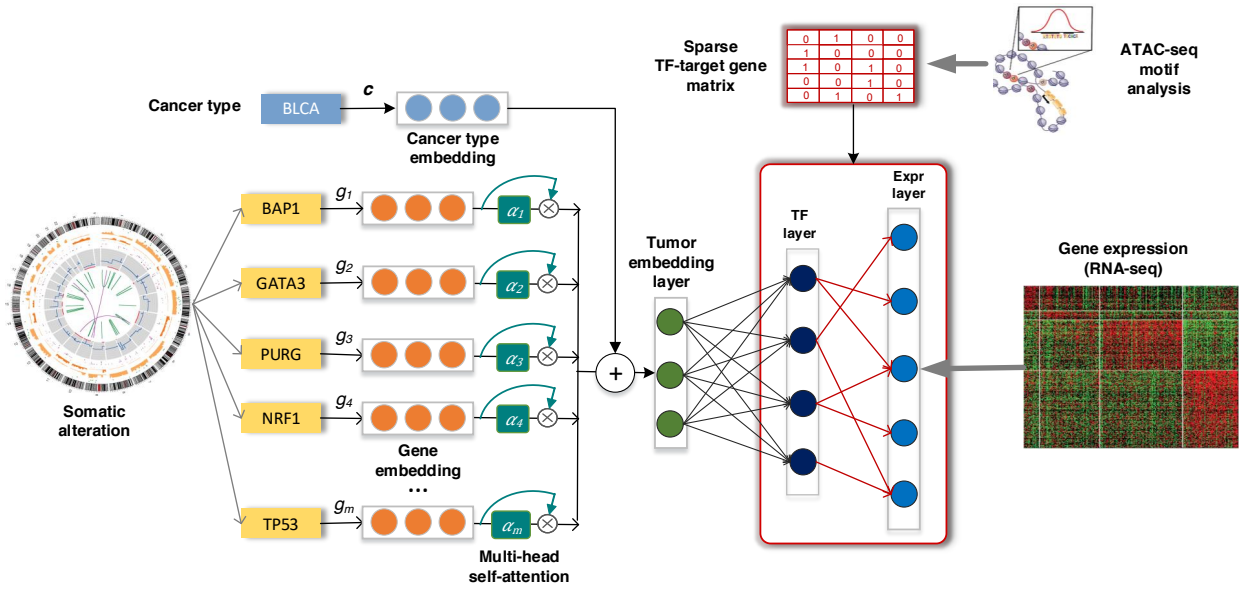

Fig. 1: Overview of CITRUS algorithm: the attention-based model with TF:target-gene priors. The input to our framework includes somatic alteration and copy number variation, assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), tumor expression datasets and TF recognition motifs. CITRUS takes somatic alteration and copy number variation data as input and encodes them as a tumor embedding using a self-attention mechanism. Additional cancer type information is used for stratifying the confounding factor of tissue type. The middle layer further transforms the tumor embeddings into TF layer, which represents the inferred activities of 320 TFs. Finally, the gene expression levels are predicted from the TF activities through a TF:target-gene priors constrained sparse layer based on ATAC-seq.

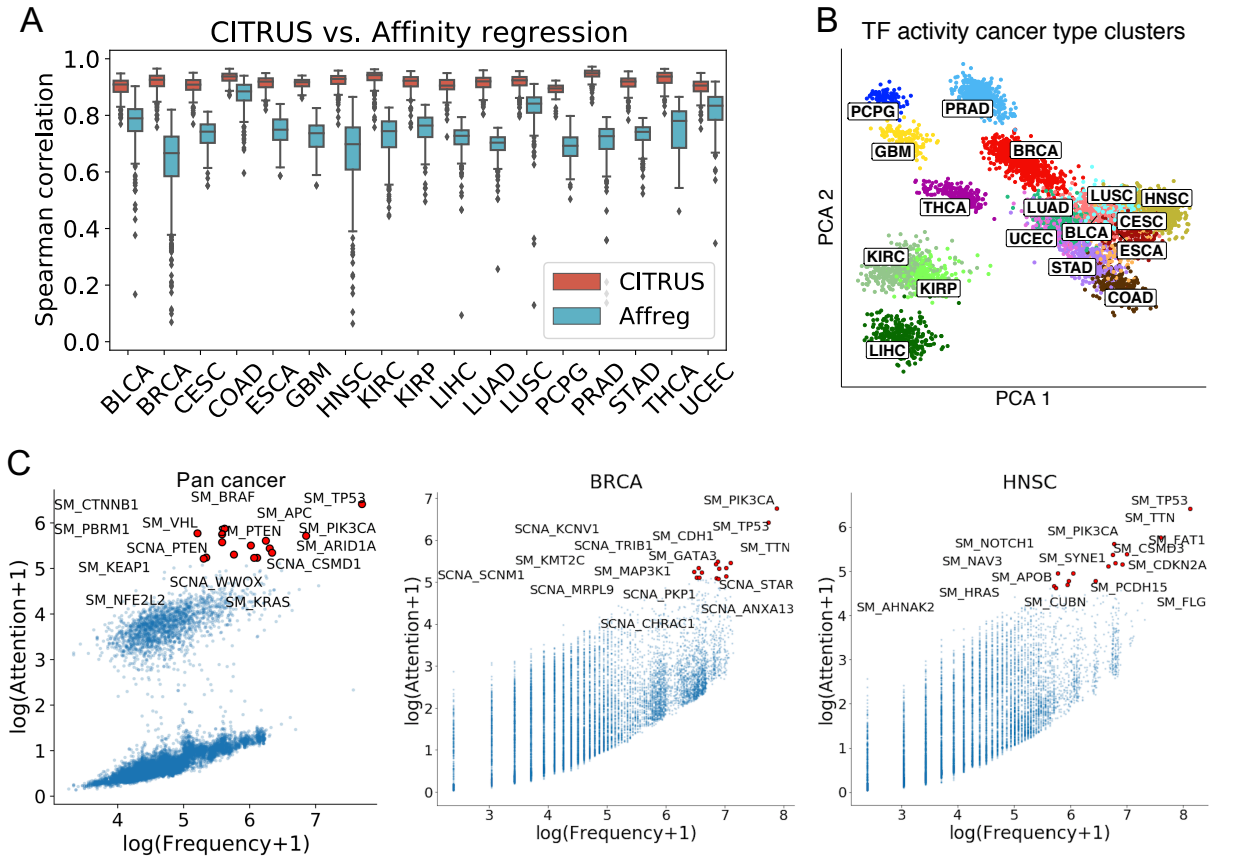

Fig. 2: CITRUS models impact of somatic alterations on gene expression programs. (A) Performance of the CITRUS models for each cancer type compared to regularized bilinear regression method, affinity regression (Affreg). Boxplots showing mean Spearman correlations between predicted and actual gene expression using the CITRUS model (orange) and Affreg (light blue) for TCGA data each cancer-type. Both CITRUS and Affreg are tuned on the training and validation sets, and evaluated on the same held-out test set. **(B)** Principal components analysis (PCA) of TF activity colored by cancer type. **(C)** Mutation frequencies and CITRUS-inferred attention weights of genes. We show cumulated results in Pan-cancer and individual BRCA, and HNSC. See **Supplementary Fig. 1** for full compilation of each cancer type.

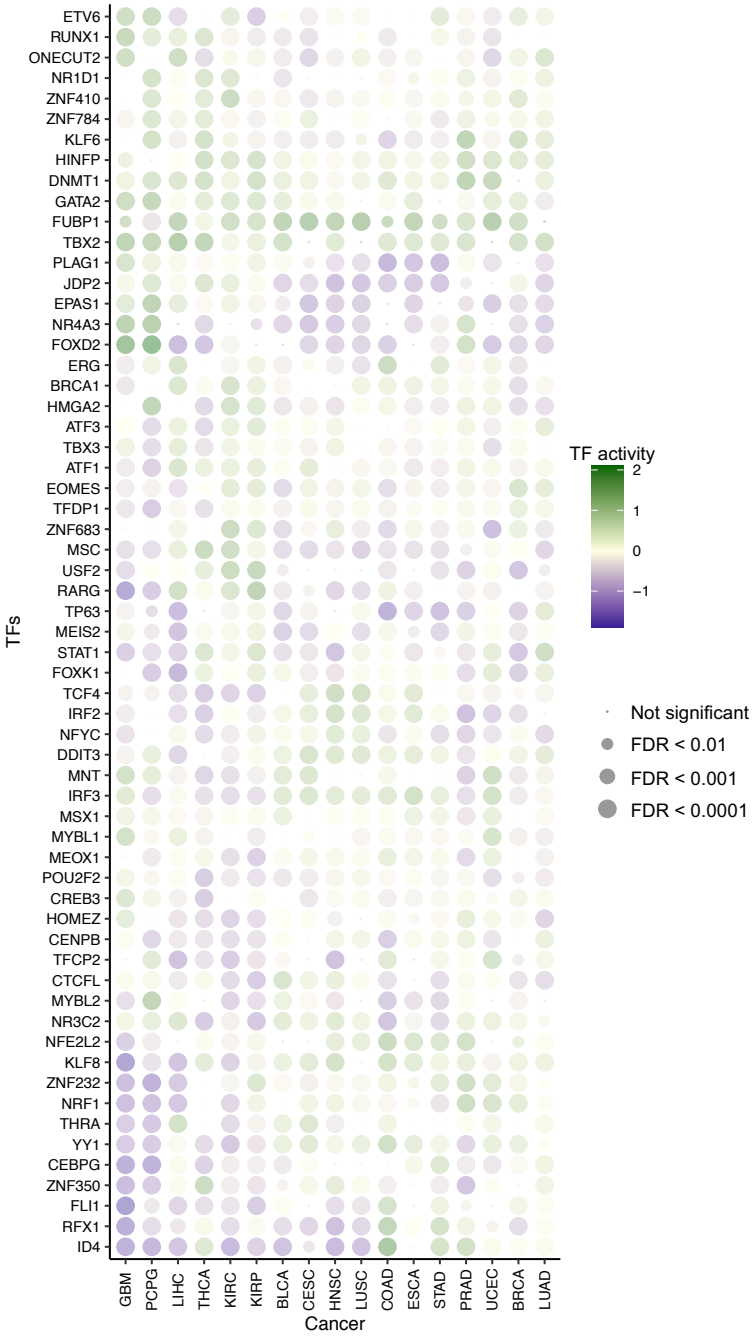

Fig. 3: CITRUS identifies regulatory features of tumor types. Dotplot shows the mean inferred TF activity differences between samples in a given tumor type vs. those in all other tumor types by t-test. We corrected for FDR across TFs for each such pairwise comparison and identified significant TF regulators and the results are shown in **Supplementary Table 1**.The dot size indicates -log10(FDR). For clarity, the union of the top 4 significant TFs in each cancer type is shown.

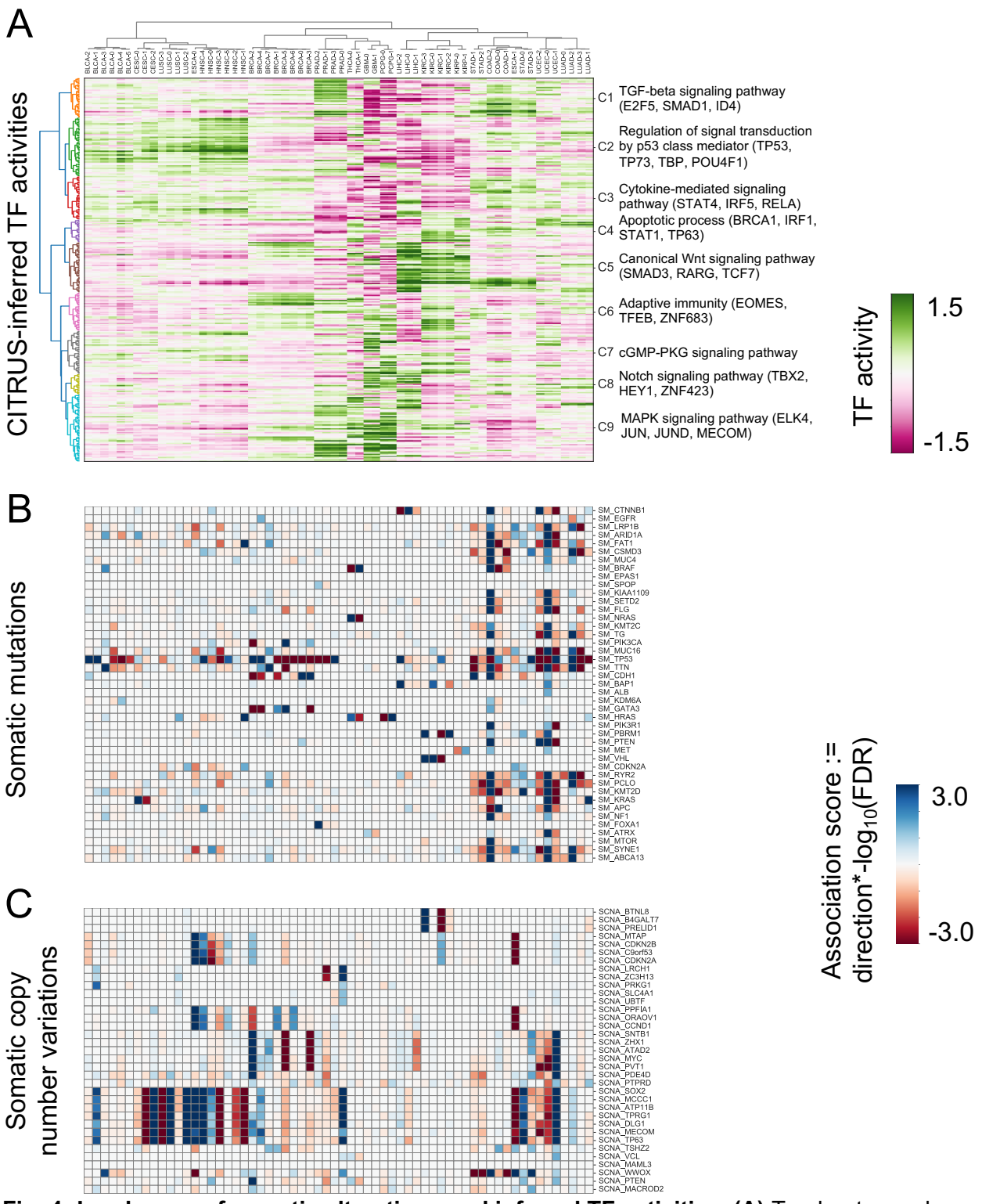

Fig. 4: Landscape of somatic alterations and inferred TF activities. (A) Top heatmap shows tumor subtypes clustered by the mean TF inferred activity. Color scale is proportional to TF activity. The heat map shows -$\log_{10}$ FDR values multipled by the direction derived by Fisher exact test for **(B)** mutations and **(C)** copy number variations.

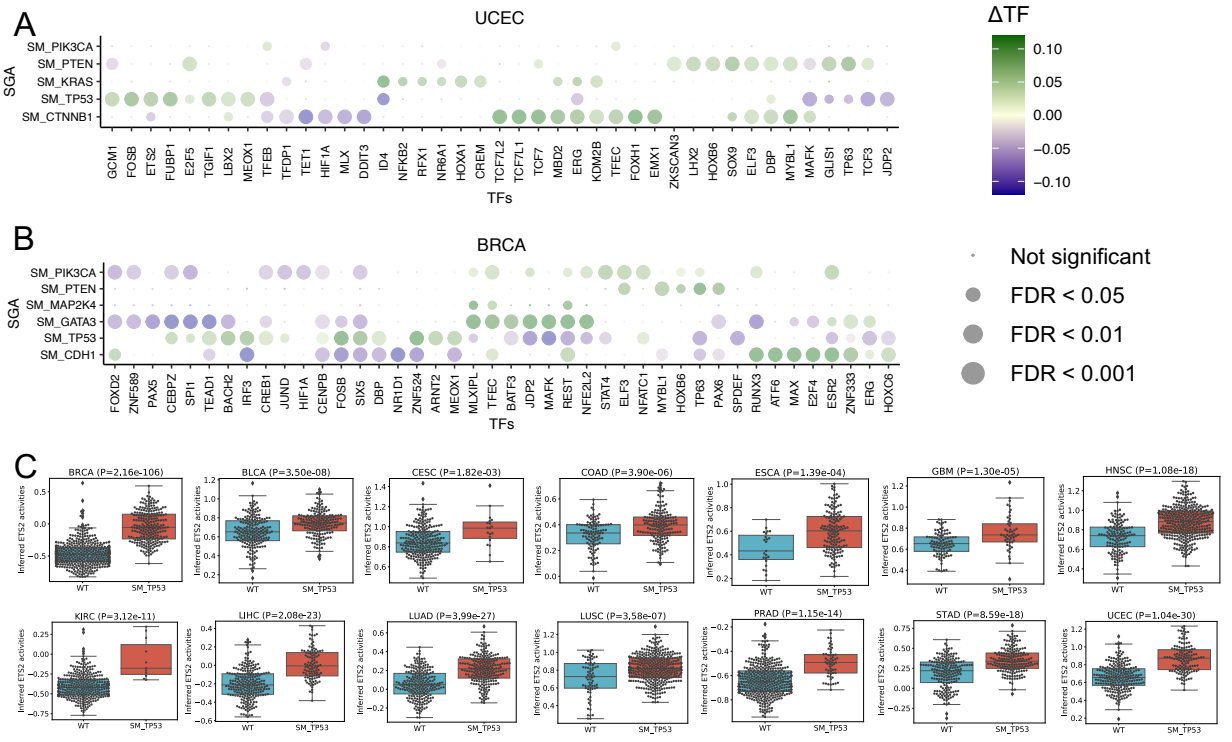

Fig. 5: Somatic alterations are associated with dysregulated TF activity. Impact of SAs on individual TFs based on knock out *in silico* experiments in TCGA **(A)** UCEC and **(B)** BRCA. The dotplot shows mean TF activity and dot size indicates -log10(FDR). See **Supplementary Fig. 3** for full list of cancer types. **(C)** Inferred ETS2 activity in TCGA studies and impact of *TP53* mutations. Tumors with mutant *TP53* have significantly higher activity of ETS2 than WT tumors ($P < 0.01$, t-test). This association is not significant using mRNA levels of ETS2 (**Supplementary Fig. 5**). Box edges represent the upper and lower quantile with median value shown as bold line in the middle of the box. Whiskers represent 1.5 times the quantile.

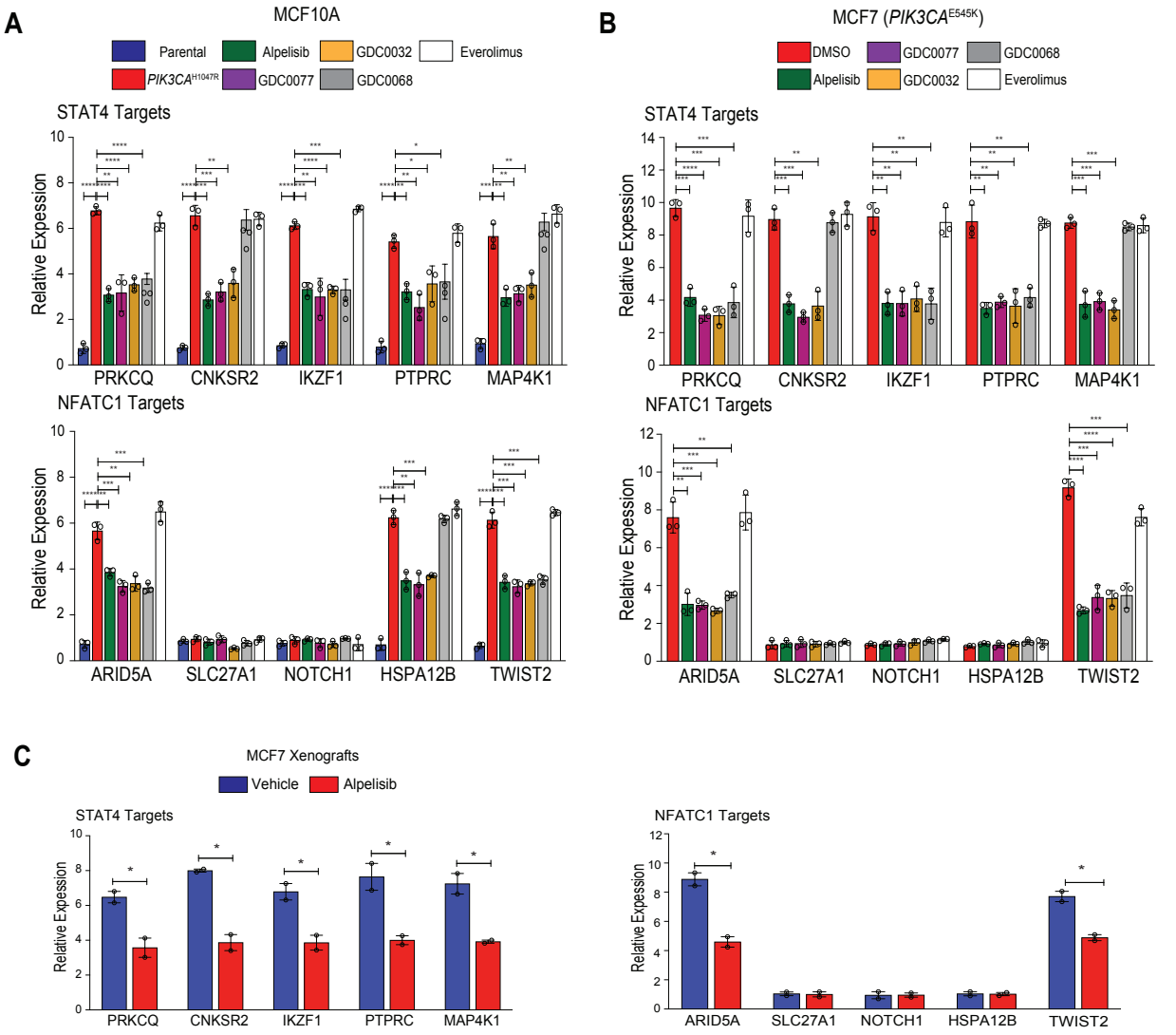

Fig. 6: Experimental validation of the PIK3CA-driven TF in breast cancer. (A) Validation of canonical target genes of STAT4, and NFATC1 in MCF10A parental and PIK3CAH1047R cells treated with DMSO or a panel of PI3K/AKT inhibitors (alpelisib 1μM, GDC0077 100nM, GDC0032 100nM, GDC0068 1μM, Everolimus 100nM) in starvation media for 4 hours, using qPCR. Expression levels were normalized to ACTIN. Circles represent independent experiments. Error bars show SD (n=3). *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. (one-side unpaired t-test). **(B)** Similar analysis of expression of target genes in MCF7 (PIK3CAE545K) was performed as in A. **(C)** Validation of the same target genes as in **A**, in MCF7-derived xenograft tumors treated with Vehicle or Alpelisib ((for details see Methods). Expression levels were normalized on ACTIN. Circles represent independent experiments. Error bars show SED (n=2). *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. (one-side unpaired t-test).