






ARTICLE

<https://doi.org/10.1038/s41467-019-12291-6>

OPEN

# Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers

Hatice U. Osmanbeyoglu <sup>1,2</sup>, Fumiko Shimizu <sup>3,9</sup>, Angela Rynne-Vidal <sup>4,9</sup>, Direna Alonso-Curbelo<sup>5,9</sup>, Hsuan-An Chen <sup>5</sup>, Hannah Y. Wen<sup>6</sup>, Tsz-Lun Yeung<sup>4</sup>, Petar Jelinic<sup>7</sup>, Pedram Razavi<sup>8</sup>, Scott W. Lowe<sup>5</sup>, Samuel C. Mok<sup>4</sup>, Gabriela Chiosis<sup>3</sup>, Douglas A. Levine <sup>7</sup> & Christina S. Leslie <sup>2</sup>

Chromatin accessibility data can elucidate the developmental origin of cancer cells and reveal the enhancer landscape of key oncogenic transcriptional regulators. We develop a computational strategy called PSIONIC (patient-specific inference of networks informed by chromatin) to combine chromatin accessibility data with large tumor expression data and model the effect of enhancers on transcriptional programs in multiple cancers. We generate a new ATAC-seq data profiling chromatin accessibility in gynecologic and basal breast cancer cell lines and apply PSIONIC to 723 patient and 96 cell line RNA-seq profiles from ovarian, uterine, and basal breast cancers. Our computational framework enables us to share information across tumors to learn patient-specific TF activities, revealing regulatory differences between and within tumor types. PSIONIC-predicted activity for MTF1 in cell line models correlates with sensitivity to MTF1 inhibition, showing the potential of our approach for personalized therapy. Many identified TFs are significantly associated with survival outcome. To validate PSIONIC-derived prognostic TFs, we perform immunohistochemical analyses in 31 uterine serous tumors for ETV6 and 45 basal breast tumors for MITF and confirm that the corresponding protein expression patterns are also significantly associated with prognosis.

<sup>1</sup>Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. <sup>2</sup>Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>3</sup>Chemical Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>4</sup>Department of Gynecologic Oncology and Reproductive Medicine—Research, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>5</sup>Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>6</sup>Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>7</sup>Laura and Isaac Perlmutter Cancer Center, New York University Langone Medical Center, New York, NY, USA. <sup>8</sup>Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>9</sup>These authors contributed equally: Fumiko Shimizu, Angela Rynne-Vidal, Direna Alonso-Curbelo. Correspondence and requests for materials should be addressed to H.U.O. (email: [osmanbeyoglu@pitt.edu](mailto:osmanbeyoglu@pitt.edu)) or to C.S.L. (email: [cleslie@cbio.mskcc.org](mailto:cleslie@cbio.mskcc.org))

Cancers arise through the accumulation of genetic and epigenetic alterations that lead to widespread gene expression changes. Transcription factors (TFs) are instrumental in driving these gene expression programs, and the aberrant activity of TFs—induced downstream of activated oncogenic signaling or in concert with epigenetic modifiers—often underlies the altered developmental state of cancer cells and acquisition of cancer-related cellular phenotypes. Data-driven computational strategies may help to infer patient-specific transcriptional regulatory programs and to identify and therapeutically target the TFs that lead to cancer phenotypes. Ultimately, such strategies could be used to personalize therapy and improve patient outcomes.

While several successful methods have been proposed for learning patient-specific regulatory programs, most regulatory network inference approaches in cancer use expression data only<sup>1</sup> or at best rely on analysis of TF motifs in annotated promoter regions<sup>2–4</sup>. However, in a few cancers—notably luminal breast and prostate cancer—ChIP-seq analyses of key transcriptional regulators, estrogen receptor (ER), and androgen receptor (AR) respectively, in both cell line models<sup>5,6</sup> and tumors<sup>7,8</sup> have revealed the importance of enhancers distal to gene promoters in gene regulatory programs. Incorporating DNA sequence information at intronic and intergenic enhancers should therefore improve the modeling of transcriptional regulation in tumors. Leveraging epigenomic data from cell line models, while imperfect, provides a feasible means to make a potentially large advance in the computational dissection of dysregulated gene expression programs in tumors.

Extensive pan-cancer genomic analyses have shown that the same genes and pathways are targeted by somatic alterations across multiple tumor types. These results suggest that pan-cancer modeling of regulatory programs could also be informative, as similar TFs may be dysregulated across cancers. So far, however, methods for inferring patient-specific regulatory programs have been applied to one cancer type at a time<sup>1,9</sup>. Multitask learning (MTL) refers to machine-learning algorithms that learn models for different problems that share information and/or parameters and provides a statistical framework for learning patient-specific regulatory models across multiple cancers<sup>10</sup>. MTL can improve accuracy by making use of limited data (small sample sizes) in each task by sharing information through the common model. This is especially important when reconstructing regulatory networks from high-throughput data because the number of parameters to fit is very large relative to the number of samples. In addition, extensive training data from more common tumor types may be able to compensate for smaller sample sizes in similar but rarer cancers.

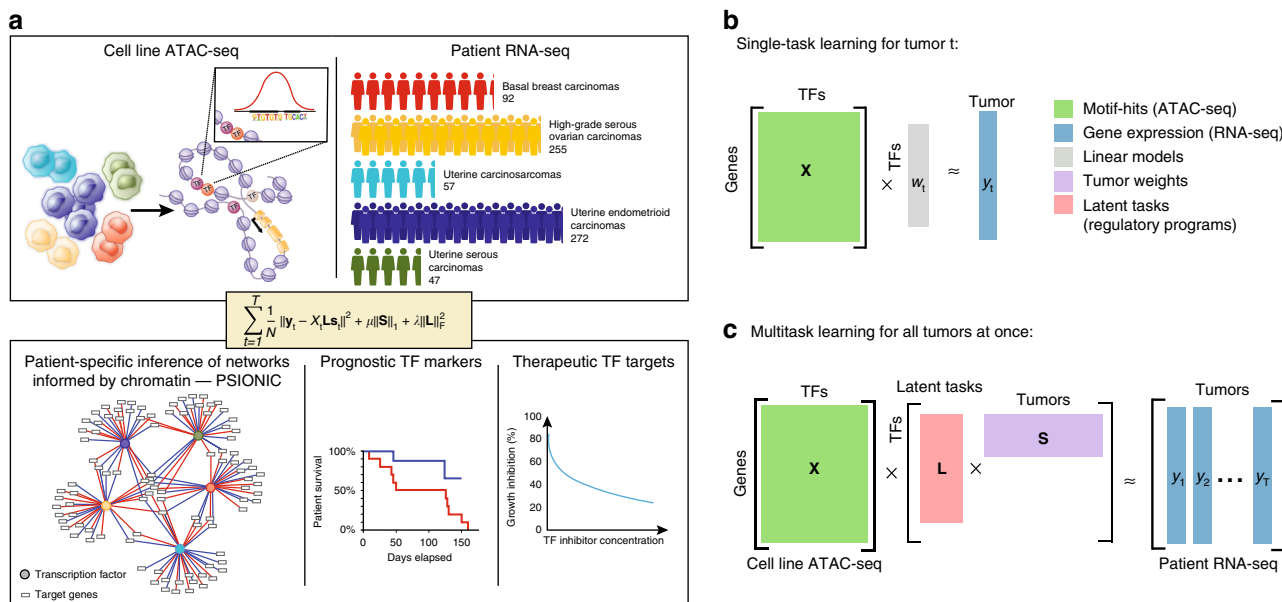
Large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA) and others have suggested molecular similarities between gynecologic cancers from different sites of pelvic origin and breast cancers<sup>11</sup>. Specifically, uterine serous carcinomas (UCS), high-grade serous ovarian carcinomas (HGSOCs), and triple negative breast cancers (TNBCs) share frequent *TP53* somatic mutations and widespread somatic copy number alterations<sup>11</sup>. HGSOCs and TNBCs also both display inactivation of similar DNA repair pathways. Though each gynecologic disease has a variety of histologic subtypes, the most common and aggressive tumors including HGSOCs (OV)<sup>12</sup>, UCS<sup>13</sup>, and the serous-like subset of uterine (UCEC)<sup>14</sup>, as well as basal breast cancer<sup>15</sup> were studied by TCGA. These tumors all lack adequate treatment options for recurrent disease and accurate predictors of response and resistance. Inferring patient-specific transcriptional regulatory programs may identify and eventually enable therapeutic targeting of transcriptional mechanisms underlying gynecologic malignancies for individualized treatment.

To improve inferring regulatory programs across cancer types, we developed patient-specific inference of networks incorporating chromatin (PSIONIC), a MTL method that jointly models transcriptional networks for several related cancer types by leveraging chromatin accessibility data in representative cancer cell lines. More specifically, PSIONIC integrates regulatory sequence from ATAC-mapped promoters and enhancers from a panel of cancer cell lines with RNA-seq data from patient tumors in order to infer patient-specific TF regulatory activities. We apply our approach to 723 RNA-seq experiments from gynecologic and basal breast cancer tumors<sup>12–15</sup> as well as 96 cell lines<sup>16</sup>, using a novel ATAC-seq data set for cell line models of these cancers. ATAC-seq data from cell lines allows us to incorporate DNA sequence information at intronic and intergenic enhancers to improve the modeling of transcriptional regulation from tumor data. Although much work has been done in regression-based inference of transcriptional regulation from cis-regulatory information in a single tumor type, we use MTL across different tumor types to jointly learn patient-specific regulatory models. Our analysis identifies key transcriptional regulators as well as new prognostic markers and therapeutic targets.

## Results

**Pan-cancer modeling of regulatory programs.** To systematically identify TFs that drive tumor-specific gene expression patterns across multiple cancer types, we developed the PSIONIC computational framework (Fig. 1a). We started with an atlas of chromatin accessible events derived from cell line models of 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). Single-task learning (STL) of a patient-specific regulatory model simply learns the TF activities that predict normalized gene expression levels in each tumor independently, using regularized regression (Fig. 1b, see the “Methods” section). In PSIONIC, we instead adopted a MTL approach called GO-MTL<sup>17</sup> to represent patient-specific TF activity model vectors across multiple tumor types as linear combinations of latent regulatory programs, where both the coefficients in the linear combination and the latent models were learned jointly by regression against all the normalized tumor expression profiles (Fig. 1c, see the “Methods” section). The latent regulatory programs capture common TF-gene regulatory relationships across patients both within and between tumor types.

**Gynecologic and basal breast cancer ATAC-seq analysis.** To enable PSIONIC modeling for gynecologic and basal breast tumors, we first generated a reference chromatin accessibility atlas for uterine (endometrioid, serous, carcinosarcoma), ovarian serous, and basal breast cancers using a panel of 12 cancer cell lines representing these five tumor types using the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq). We assembled an atlas of ~282K reproducible accessibility regions for all cell lines, as well as tumor type-specific atlases ranging from ~93 to ~153K reproducible regions (Supplementary Table 1, see the “Methods” section). Principal component analysis (PCA) identified heterogeneity in the chromatin accessibility landscape in these gynecologic and basal breast cancer cell lines (Fig. 2a, Supplementary Data 1). Notably, ovarian and basal breast cancer cell lines displayed more similar chromatin accessibility profiles than most of the uterine cancer cell lines. Interestingly, for the two uterine carcinosarcoma cell lines, the copy number high SNU685 cell line clustered with ovarian



**Fig. 1** Overview of PSIONIC algorithm. **a** The input to our framework includes assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) profiles, TF recognition motifs, and tumor expression datasets. PSIONIC integrates regulatory information for each gene based on motifs in ATAC-mapped promoters and enhancers from cancer cell lines ( $X$ ) with RNA-seq data from patient tumors ( $Y$ ) in order to infer patient-specific TF activities ( $W = LS$ ). Here, columns in the matrix  $L$  represent these latent regulatory programs, while  $S$ , the tumor weight matrix, captures the grouping structure and specifies the coefficients of the linear combination of latent regulatory programs for each tumor. A schematic comparison of **b** single task learning (STL) and **c** multitask learning (MTL) models

and basal breast cancer cell lines, whereas JHUCS1 clustered with uterine endometrioid cell lines.

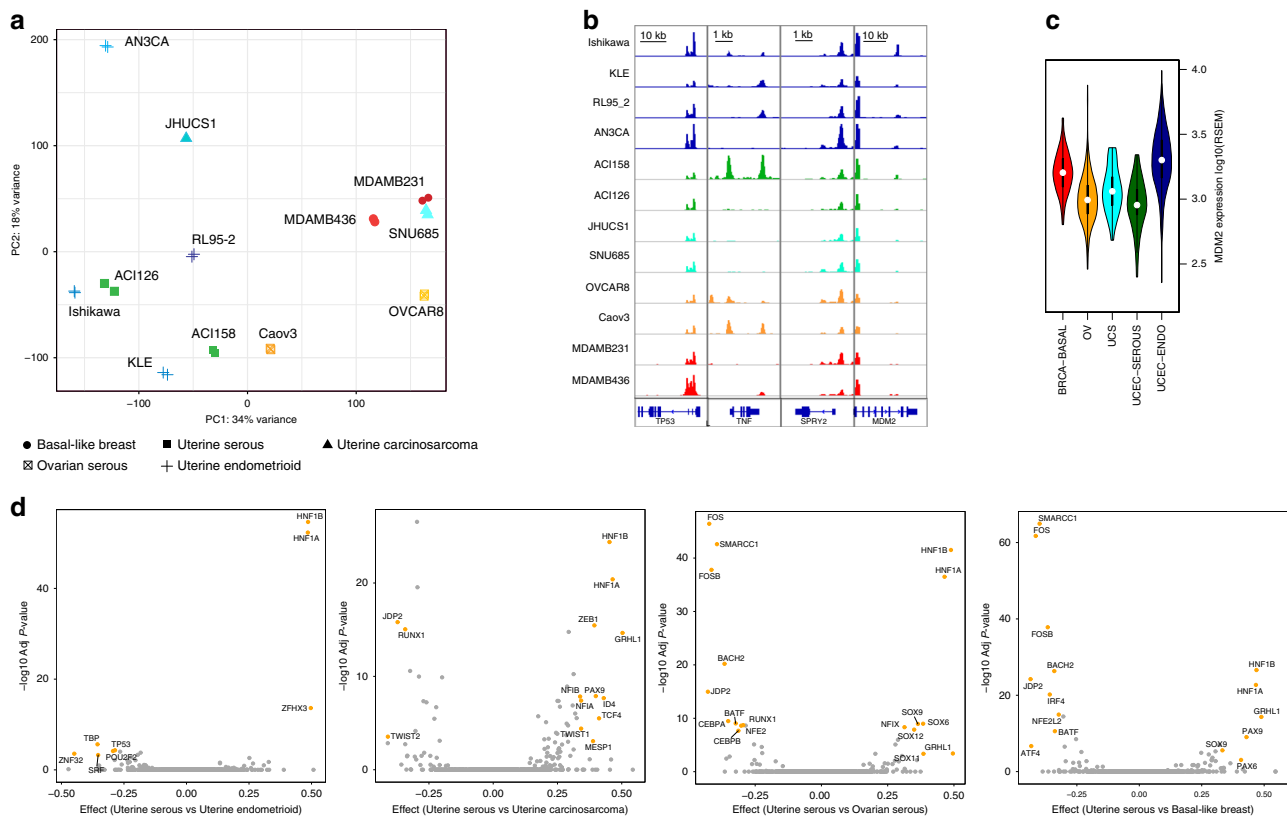
Next, we assigned each accessible region in the tumor type specific atlas to the nearest gene (Fig. 2b), and we defined the regulatory locus complexity of a gene<sup>18</sup> as the total number of accessible regions within the tumor type. We grouped genes into three equally sized classes (tertiles) based on their regulatory complexity in tumor type specific atlases. Complexity classes were defined by dividing genes at the 33rd and 66th percentiles of the distribution of the number of accessible regions to produce groups with similar numbers of genes. We found that the normalized expression levels of low-complexity genes were lower than high-complexity and medium-complexity genes in tumor samples from TCGA for each tumor type ( $P < 1 \times 10^{-16}$ , one-sided Kolmogorov–Smirnov (KS) test for all comparisons). The importance of enhancers is illustrated by the region surrounding the *MDM2* gene. Despite the ubiquitous accessibility of the *MDM2* promoter, nearby distal regulatory elements of *MDM2* were more accessible in uterine endometrioid cell lines, consistent with higher *MDM2* gene expression observed in corresponding tumor samples from the TCGA cohort (Fig. 2c).

We also compared the cell line accessibility patterns with those of primary tumors using recently published ATAC-seq signal data for tumor samples from TCGA<sup>19</sup> including 13 UCEC-ENDO (24 with replicates) and 15 BRCA-BASAL (30 with replicates). Differential analysis of endometrial and basal breast cancer cell lines identified 366 endometrial-specific peaks and 368 basal breast-specific peaks ( $FDR < 10^{-4}$ ,  $\log_2(FC) > 3$ ). Consistent with our cell line data, high accessibility regions in breast cancer cell lines displayed significantly higher accessibility in BRCA-BASAL patients than in UCEC-ENDO ( $P < 10^{-4}$ , one-sided Wilcoxon signed-rank test, see the “Methods” section), while high accessibility regions in uterine endometrioid cell lines showed significantly higher accessibility in UCEC-ENDO patients than in BRCA-BASAL patients ( $P = 0.00016$ , one-sided Wilcoxon signed-rank test), as shown in Supplementary Fig. 1.

**Motifs underlying differential accessibility in cell lines.** Next, we determined the TFs that are most associated with open

chromatin for each tumor type through motif analyses and differential accessibility (see the “Methods” section). We examined the patterns of gain or loss of chromatin accessible regions between each pair of tumor types by performing pairwise differential read count analysis on accessible regions. The heatmap in Supplementary Fig. 2 shows the patterns of differential accessibility found among ~40,000 peaks across cell lines. Many TFs whose motifs were identified at differentially accessible regions between pairs of tumor types have roles in tumorigenesis (Fig. 2d, Supplementary Fig. 3, Supplementary Data 2). For example, chromatin peaks with HNF1 family motifs were more accessible in the endometrioid subset of uterine cell lines than in other cell types ( $P < 10^{-16}$ , one-sided KS test). HNF1 $\beta$  is associated with cancer risk in several tumors, including hepatocellular carcinoma, pancreatic carcinoma, renal cancer, ovarian cancer, endometrial cancer, and prostate cancer<sup>20</sup>. KLF and ETS family motifs were more accessible in endometrioid uterine and ovarian serous cell lines than in other cell types ( $P < 10^{-16}$ , one-sided KS test). These TFs have been implicated in the pathogenesis of these endocrine-responsive cancers of female reproductive tissues<sup>21,22</sup>. Chromatin peaks with FOS family motifs were more accessible in basal breast, ovarian serous and uterine carcinosarcoma and less accessible in uterine endometrioid cell lines than in other cell types ( $P < 10^{-16}$ , one-sided KS test). FOS family TFs have been implicated as regulators of cell proliferation, differentiation, and transformation.

In some cases the TF signal between cell lines might be due to the tissue of origin. To look more closely at this issue, we examined publicly available chromatin accessibility data in relevant normal tissues. We generated a reference chromatin accessibility atlas for normal uterine ( $n = 1$ ), ovarian ( $n = 3$ ), and breast ( $n = 1$ ) tissue using DNase-seq data by the Roadmap Epigenomics project<sup>23</sup> and assembled an atlas of ~397K accessibility regions. We performed motif analysis in each chromatin accessible regions in the common atlas. Then, we examined the patterns of gain or loss of chromatin accessible regions between each pair of tumor types by performing pairwise differential read count analysis on accessible regions.



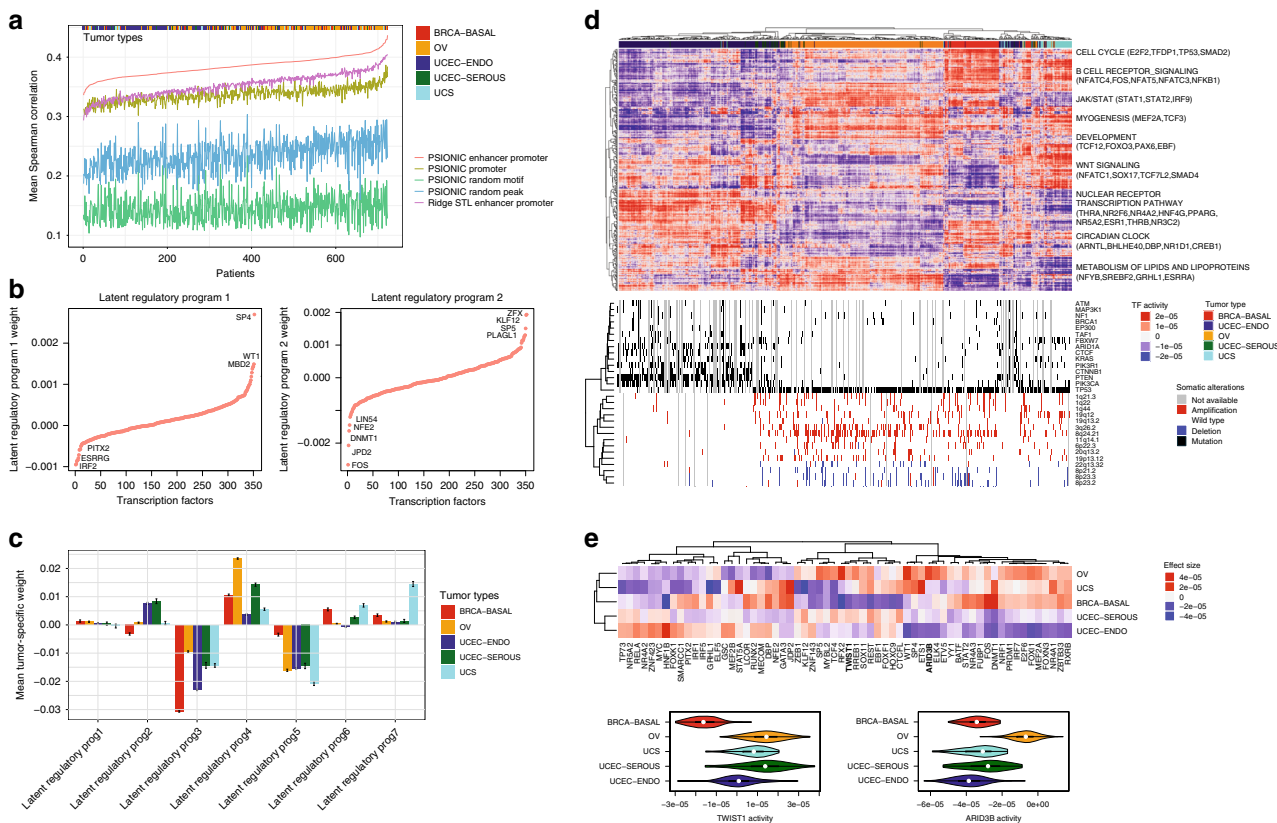
**Fig. 2** ATAC-seq analysis identifies key TFs in gynecologic and basal breast cancer cells. **a** Unsupervised principal component analysis based on the chromatin accessibility for all 12 cell lines at each of the 10K most variable chromatin accessible regions in the cell line panel. Samples are color coded according to the cell line type. Each symbol represents a single biological replicate and different symbols represent the tumor type of origin. Source data are provided as a Source Data file. **b** Normalized ATAC-seq profiles at important genes. Profiles represent the union of all biological replicates for each cell type. Genomic coordinates for the loci: *TP53*, chr17:7571720–7590868; *TNF*, chr6:31543344–31546112; *SPRY2*, chr13:80910112–80915086; *MDM2*, chr12:69201952–69239324. All y-axis scales range from 0 to 235 in normalized arbitrary units. The x-axis scale is indicated by the scale bars. **c** Violin plots indicate the distribution of *MDM2* gene expression across tumor types. *MDM2* gene expression is higher in uterine endometrioid carcinomas (UCEC-ENDO) compared to other tumor types. **d** Pairwise comparison of transcription factor motifs enriched in differentially accessible regions in cell lines. Volcano plot showing effect size versus  $-\log_{10}$ (adjusted *P*), using a Bonferroni correction to adjust *P* values for each plot. TF symbol annotations are written where the absolute value of the effect size is in at least 30 and adjusted *P* <  $10^{-3}$ . The foreground occurrence is the number of peaks containing a particular TF motif within the group of 5000 upregulated or 5000 downregulated *P* peaks according to  $\log_2$ -fold-change read counts, respectively. The background occurrence is the number of peaks containing a particular TF motif found among all the differentially accessible peaks. Remaining pairwise comparisons are shown in Supplementary Fig. 3. Source data are provided as a Source Data file

While several FOS family motifs and SMARCC1 are enriched both in normal uterus vs. ovary as well as in the comparison of uterine serous vs. ovarian serous, in most cases the motifs identified by differential accessibility in cancer cell lines did not arise from the tissue of origin based on available normal tissue accessibility data (Supplementary Fig. 4). While many identified TFs are known to play a role in other cancers, their impact on gene regulation has not been characterized in gynecologic and basal breast cancers. We therefore developed a regression framework to model the regulatory role of TFs on gene expression in tumor samples.

**Multitask regression explains tumor expression profiles.** We next used a MTL strategy across tumor types to learn patient-specific regression models to predict tumor gene expression from gene regulatory sequence derived from cell line ATAC-seq data. Our method assumes that observed gene expression levels in each tumor can largely be explained by the unobserved activities of a smaller number of TF regulatory proteins through correlation with TF-binding motif scores. Moreover, our approach shares information across tumor samples and tumor types by representing each patient-specific regulatory model as a linear combination of a latent regulatory models.

Formally, we developed PSIONIC, a multitask-learning framework for integrating regulatory elements for each gene based on motifs in ATAC-mapped promoters and enhancers from cancer cell lines (*X*) with RNA-seq data from patient tumors (*Y*) to infer patient-specific TF regulatory activities ( $W = LS$ ) (Fig. 1c). We adopted an algorithm for learning grouping and overlap structure in MTL (GO-MTL)<sup>17</sup>; here, the model does not assume a disjoint assignment of tasks (patients) to different groups (e.g. tumor type) but rather allows patient-specific models to overlap with each other by sharing one or more latent basis tasks, or latent regulatory programs. Further, the matrix *L* represents these latent regulatory programs, while *S*, the tumor weight matrix, captures the grouping structure and specifies the coefficients of the linear combination of latent regulatory programs for each tumor. MTL enables selective sharing of information across other tumors, while standard STL trains a regression model for each tumor independently.

The application of our approach to 723 uterine, ovarian, and basal breast tumors from TCGA identified key TFs as potential common or cancer-specific drivers of expression changes. Our expression dataset included samples from five different tumor types, namely basal breast (BRCA-BASAL, *n* = 92), high-grade



**Fig. 3** PSIONIC identifies regulatory features of tumor types. **a** PSIONIC and STL regression models predict differential expression of held-out genes and subtypes of tumor samples. Plot showing Spearman correlations between predicted and actual gene expression changes for all samples, sorted based on performance of the PSIONIC model using enhancer and promoter TF-binding sites. For each method and each sample, the Spearman correlation is computed using 10-fold cross-validation on held-out genes. Using TF-binding sites from enhancer promoter as features (mean  $\rho = 0.384 \pm 0.016$ ) is significantly better than if we randomized motif hits for each chromatin accessible region across all motifs (mean  $\rho = 0.144 \pm 0.022$ ;  $P < 10^{-32}$ , one-sided Wilcoxon signed-rank test), or if we randomized accessible regions for each motif, then assigned to the nearest gene (mean  $\rho = 0.235 \pm 0.025$ ;  $P < 10^{-32}$ , one-sided Wilcoxon signed-rank test). PSIONIC models with motif data from promoter and enhancer regions outperformed models where only motif hits in promoter regions were used (mean  $\rho = 0.337 \pm 0.012$ ;  $P < 10^{-16}$ , one-sided Wilcoxon signed-rank test) and STL approach based on ridge regression (mean  $\rho = 0.352 \pm 0.019$ ;  $P < 10^{-21}$ , one-sided Wilcoxon signed-rank test). TCGA tumor types are shown in the top bar. **b** Example of latent regulatory programs. TFs are ranked based on the magnitude of coefficients. Remaining latent regulatory programs are shown in Supplementary Fig. 6. **c** Mean ( $\pm$  SE, standard error) tumor weight matrix (denoted by **S**) grouped according to tumor type for each latent regulatory program. **d** Hierarchical clustering mean-centered model vectors (denoted by **W**) on the TCGA tumor data sets. Source data are provided as a Source Data file. **e** Heatmap shows the mean inferred TF activity differences between samples in a given tumor type vs. those in all other tumor types. For each comparison, the absolute value of the mean inferred TF activity differences (effect sizes) are ranked and the union of top 20 TFs for each comparison are shown in the heatmap. Violin plots indicate the distribution of inferred ARID3B and TWIST1 TF activities across tumor types. Source data are provided as a Source Data file

serous ovarian (OV,  $n = 255$ ), uterine carcinosarcoma (UCS,  $n = 57$ ), uterine endometrioid carcinomas (UCEC-ENDO,  $n = 272$ ), and uterine serous carcinomas (UCEC-SEROUS,  $n = 47$ ). These results were obtained using binding site predictions for 352 human sequence-specific TFs based on motif hits from the CisBP database as motif data (see the “Methods” section).

Performance of PSIONIC and STL based on ridge regression for each tumor type using 10-fold cross-validation is shown in Fig. 3a. For statistical evaluation, we computed the mean Spearman correlation ( $\rho$ ) between predicted and measured gene expression profiles on held-out genes for each tumor type and obtained mean  $\rho = 0.384 \pm 0.016$  for PSIONIC, a highly significant result ( $P < 10^{-16}$ , one-sided Wilcoxon signed-rank test). This regression performance was significantly better than STL ( $P < 10^{-21}$ , one-sided Wilcoxon signed-rank test). Similarly, our models with motif data from promoter and enhancer regions outperformed models where only motif hits in promoter regions were used ( $P < 10^{-16}$ , one-sided Wilcoxon signed-rank test). By contrast, if we randomized motif hits for each chromatin

accessible region across all motifs, or if we randomized accessible regions for each motif, then assigned to the nearest gene, the prediction performance also significantly decreased ( $P < 10^{-32}$ , one-sided Wilcoxon signed-rank test).

When we compared 10-fold cross-validation results with different values of  $K$ , we found that prediction performance was stable after  $K = 4$ , with no sign of overfitting with higher  $K$ . However, a higher number of regulatory programs did allow PSIONIC-inferred models to better distinguish between tumors of distinct subtypes (Supplementary Fig. 5). Therefore,  $K = 7$  seemed to be a reasonable choice for optimizing both overall prediction performance and capturing tumor-type-specific components of the regulatory models (Fig. 3b, Supplementary Fig. 6). Figure 3c shows a summary of mean tumor weights (**S**) across each tumor type for each latent regulatory program. For example, latent regulatory program 1 appeared to capture a common gene regulatory program shared across all cancer types, whereas latent regulatory program 2 captures TF-gene-regulatory relationships shared by uterine serous and endometrioid tumors.

Hierarchical clustering of tumors by inferred TF activities,  $\mathbf{W} = \mathbf{L}\mathbf{S}$ , as derived from the model largely recovered the distinction between the major tumor types (Fig. 3d, Supplementary Data 3). In particular, clustering based on inferred TF activity mostly stratified patients by *TP53* mutation status. Uterine endometrioid tumors have distinct patterns of TF activities, consistent with their differing expression and mutational patterns.

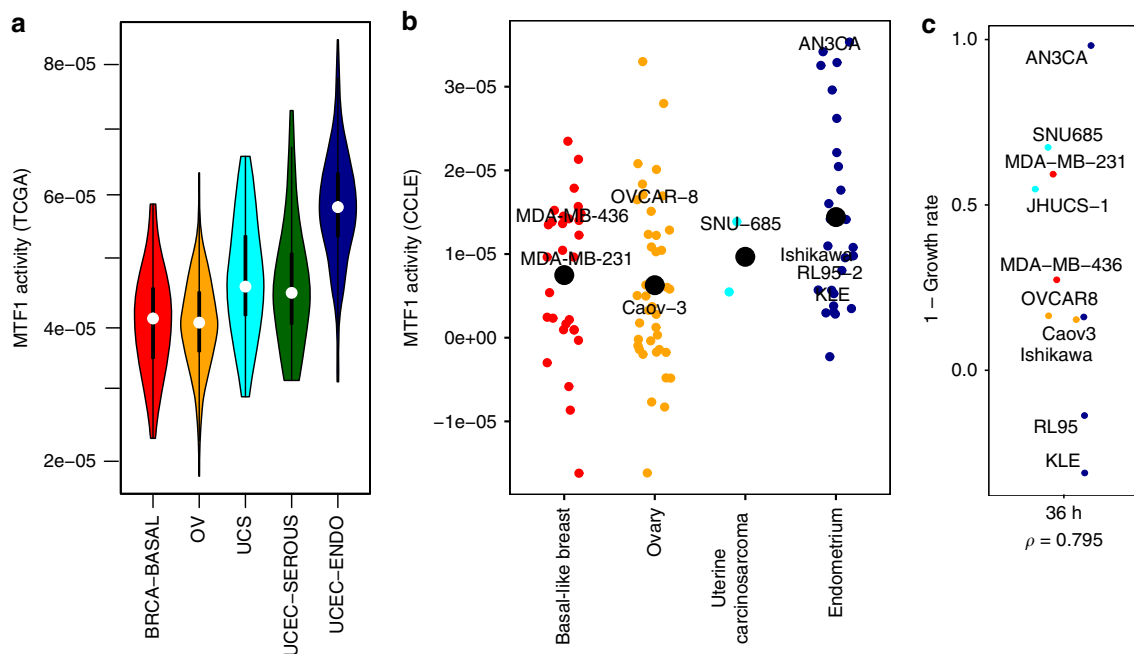
**Multitask regression identifies tumor type-specific TFs.** Next, we assessed TF-tumor type associations by *t*-test to compare inferred TF activity between samples in a given tumor type vs. those in all other tumor types. We corrected for FDR across TFs for each such pairwise comparison and identified significant TF regulators and the results are shown in Supplementary Data 4 and Fig. 3e. FUBP1, which regulates *c-Myc* gene transcription, had significantly higher inferred activity in BASAL-BRCA than in gynecologic tumors, whereas ARID3B activity was significantly higher in OV, consistent with its role in promoting ovarian tumor development, in part by regulating stem cell genes<sup>24</sup>. NR5A2 (also known as liver receptor homolog-1, LRH-1) was significantly higher in uterine endometrioid tumors, consistent with its function in regulating metabolism and hormone synthesis. Moreover, in agreement with previous reports, WT1 activity was significantly higher in ovarian serous<sup>25</sup> and uterine sarcoma<sup>26</sup>; TWIST1, a central player in the EMT, had increased activity in ovarian serous<sup>27</sup> and uterine serous cancers; YY1, which regulates various processes of development and differentiation and is involved in tumorigenesis of breast and ovarian cancer<sup>28</sup>, had increased activity in these cancers.

In addition to confirming key TFs from previous studies, our analysis also predicted novel TF regulators in gynecologic and basal breast cancers. For example, MEF2A, a transcriptional regulator implicated in muscle development, cell growth control, and apoptosis, had significantly higher activity in OV, BRCA-

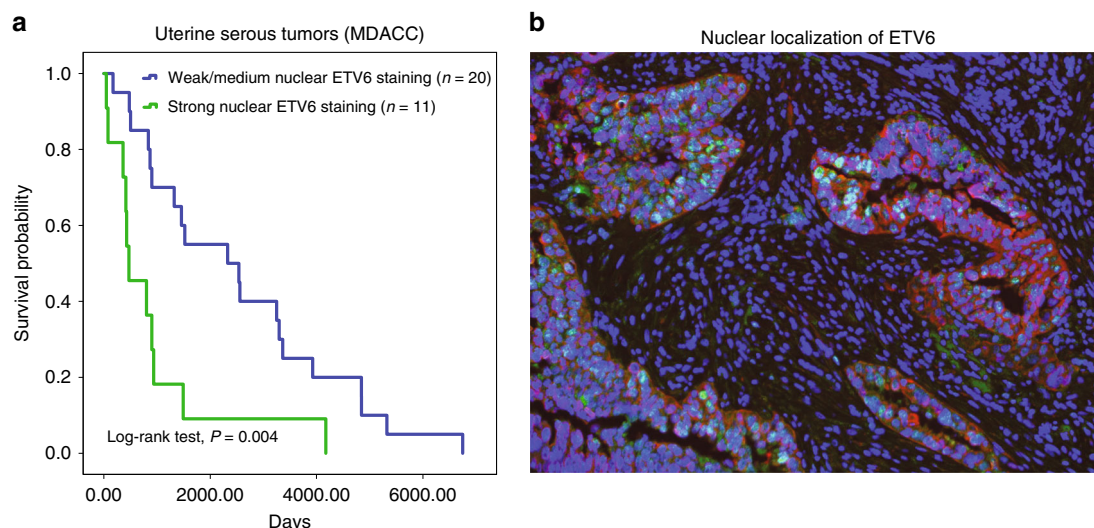
BASAL, and UCS; the activity of microphthalmia-associated transcription factor (MITF), was significantly higher in uterine carcinosarcoma than in other cancers and displayed high variation across patients. Indeed, UCSs are characterized by an admixture of at least two histologically distinct components, one resembling carcinoma and another resembling sarcoma<sup>13</sup>. The roles of MEF2A and MITF have not been previously characterized in these cancers and may present promising targets for study and potentially for therapeutic intervention.

To investigate the potential of using PSIONIC-inferred TF activities to predict sensitivity to TF-targeted therapeutics, we decided to translate our model into cancer cell lines where drug sensitivity can be experimentally determined. Therefore, we first assembled a collection of basal-like, ovary and endometrium transcriptional profiles of immortalized human cancer cell lines from the CCLE<sup>16</sup>, trained a PSIONIC model on this data set, and hence inferred cell line-specific TF regulatory activities. Similar to tumor models, we obtained significantly better regression performance with PSIONIC than with STL based on ridge regression in cell lines using 10-fold cross-validation (Supplementary Fig. 7). Regulatory models for cell lines to some extent recapitulated patient-specific tumor regulatory models (Supplementary Fig. 8). Importantly, cell line models as well as tumor models clustered mostly by cancer type.

While few drugs directly target TFs, we were able to use the metal-regulatory transcription factor-1 (MTF1) inhibitor LOR-253 for a proof of principle analysis. MTF1 is a ubiquitously expressed TF that is activated by heavy metals, redox stresses, growth factors, and cytokines<sup>29</sup>. We assessed our original panel of 10 cell lines for sensitivity to LOR-253 by measuring growth rate inhibition. Consistent with expectation, cell lines with higher inferred MTF1 activity showed a greater decrease in growth rate after the treatment with LOR-253 (Fig. 4). Overall, MTF1 inferred activity was significantly associated with growth rate inhibition by Spearman correlation analysis ( $\rho = 0.795$  for these cell lines).



**Fig. 4** PSIONIC predicts cell line sensitivity to TF-targeted therapy. **a** Violin plots indicate the distribution of inferred MTF1 TF activities across tumor types. **b** We trained a PSIONIC model on 96 cell lines from the CCLE study. The dot plots show inferred MTF1 activities in basal-like breast, endometrium, ovary, and uterine carcinosarcoma cell lines. Black dots indicate mean inferred MTF1 activity for each tumor type. **c** 1—growth rate (GR) values<sup>60</sup> (growth inhibition) for MTF1 inhibitor LOR-253 36 h after the treatment in gynecologic and basal breast cancer cell lines (for these cell lines spearman correlation  $\rho = 0.795$ ). Source data are provided as a Source Data file



**Fig. 5** Clinical validation of ETV6 in uterine serous cancer. **a** Kaplan–Meier plot for uterine serous patients stratified by ETV6-staining score. Patient samples ( $N = 31$ ) were divided into two groups based on intensity of ETV6 and positivity in nuclei or cytoplasm (patients with cytoplasmic, negative or weak/medium [score = 1 or 2] nuclear ETV6 staining,  $N = 20$ ; and patients with strong [score = 3] nuclear ETV6 staining,  $N = 11$ ). A significant difference in survival was observed between the groups ( $P = 0.004$ , log-rank test). The median survival was 2330 days (95% CI: 104–4556 days) for the cytoplasmic or weak nuclear group and 214 days (95% CI: 53–891) for the strong nuclear group. Source data are provided as a Source Data file. **b** Representative image of immunofluorescence staining on a primary uterine serous tumor. Double staining for ETV6 (green) and cytokeratin (red), shows nuclear localization of ETV6 in tumor cells; DAPI (4',6-diamidino-2-phenylindole): blue. Source data are provided as a Source Data file

**Clinical outcome based on inferred TF activities.** To investigate the clinical relevance of TF activities, we examined whether inferred TF activities were associated with therapeutic response. The standard of care for ovarian serous patients is aggressive surgery followed by platinum/taxane chemotherapy. After therapy, platinum-resistant cancer recurs in ~25% of patients within 6 months<sup>30</sup>. The clinical significance of recurrence following current standard of care for ovarian serous patients prompted us to determine TFs linked to platinum resistance in OV. Inferred TF activities of seven TFs were significantly associated with platinum response, including HIF-1 $\alpha$  and ZNF423 ( $t$ -test,  $P < 0.05$ , Supplementary Fig. 9). Consistent with our findings, HIF-1 $\alpha$  has been associated with platinum resistance in a variety of cancers, including ovarian<sup>31</sup>. Moreover, ESR1 and ZNF423 have a role in cancer cell proliferation<sup>32,33</sup> and were significantly associated with platinum-sensitive tumors.

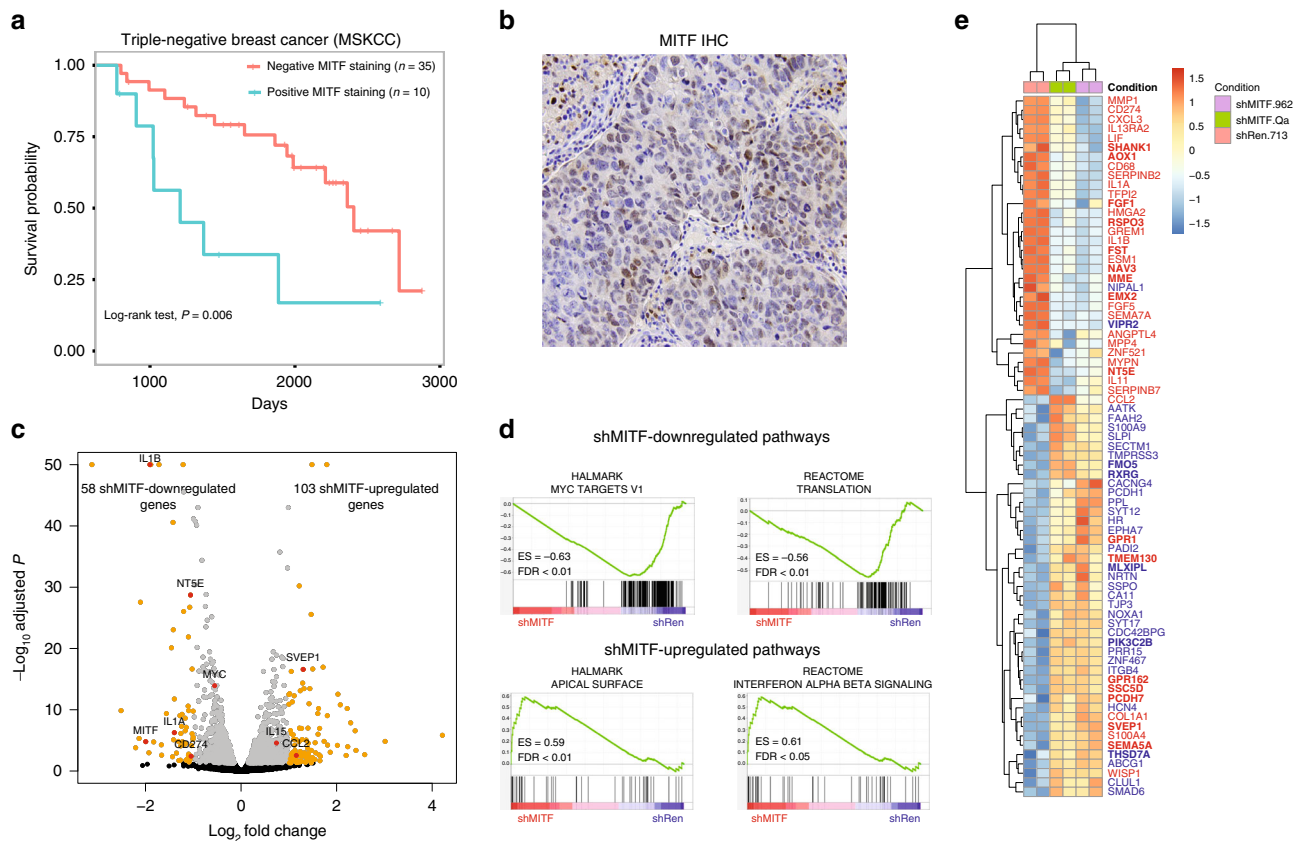
Next, we examined whether inferred TF activities were linked to survival data from the TCGA. We fit Cox proportional hazards regression models for each TF activity using clinical stage and age as additional covariates. The patient survival data and matched TF activities enabled us to perform TF-centric survival analyses to identify prognostic TFs within tumor type (TFs with FDR-adjusted  $P < 0.02$ , Cox analysis). Numerous TFs were significantly associated with survival outcome in BRCA-BASAL, UCEC-SEROUS, and UCEC-ENDO (Supplementary Tables 2–4). For some TFs, the prognostic value has been reported previously; for example, PGR<sup>34</sup> has been associated with survival in uterine cancer. However, most of the identified prognostic TFs lack prior reports of a link to survival in these cancers, making them potential candidates for follow-up studies.

For example, ETV6 inferred activity separated patients into high-risk and low-risk groups in UCEC-SEROUS (FDR  $< 0.02$ , Cox analysis). ETV6 exhibits antitumor effects suppressing proliferation and metastatic progression in prostate cancer<sup>35</sup>. However, its role in uterine serous cancer has not been studied. To further investigate whether prognostic TFs identified through inferred activity analyses could be verified at the protein level, we

performed immunohistochemical analyses in primary tumor samples from patients with uterine serous cancer ( $n = 31$ ) for ETV6. Our analysis of the two groups of patient samples divided based on intensity of ETV6 and positivity in nuclei or cytoplasm showed a significant difference in survival between the groups ( $P < 0.004$ , log-rank test), with median survival of 2330 and 214 days for the weak or medium nuclear and the strong nuclear groups, respectively. The Kaplan–Meier survival curve based on ETV6 staining is shown in Fig. 5a. A representative image of immunofluorescence staining of a primary uterine serous tumor shows protein level nuclear localization of ETV6 in tumor cells (Fig. 5b).

Similarly, MITF inferred activity separated patients into high-risk and low-risk groups in BASAL-BRCA (FDR = 0.011, Cox analysis). Indeed, tissue microarray analyses in clinically annotated primary basal breast tumor samples ( $n = 45$ ) validated MITF positivity in tumor cells and revealed a significant association between MITF expression and patient survival ( $P < 0.006$ , log-rank test), with median survival of 1208 and 2406 days for the positive and negative staining groups, respectively (see Kaplan–Meier survival curve and representative MITF-positive staining in basal breast cancer patients in Fig. 6a, b). MITF is a key TF in melanocyte development and differentiation and a diagnostic biomarker for metastatic melanoma<sup>36</sup>. However, the role of MITF in non-melanoma cancer cells, including basal breast cancer, is largely undefined. Thus, we next sought to functionally validate PSIONIC-predicted MITF activity in basal breast cancer cells.

To this end, we generated inducible shRNA vectors<sup>37</sup> targeting MITF and evaluated their impact on basal breast cancer gene expression. Potent shRNA-driven MITF downregulation was confirmed in both MDA-MB-436 basal breast cancer cells and SK-Mel-28 melanoma cells with known high MITF levels (Supplementary Fig. 10A–C). RNA-seq following MITF silencing revealed an effect on gene expression with 58 consistently down-regulated and 103 consistently up-regulated genes (adjusted  $P < 0.05$  and fold change  $> 2$ ) in MDA-MB-436 cells transduced with



**Fig. 6** Clinical and in vitro validation of MITF in basal breast cancer. **a** Kaplan-Meier plot for basal breast cancer patients stratified by MITF staining score. Patient samples ( $n = 45$ ) were divided into two groups based on MITF positivity ( $n = 10$ ) and negativity ( $n = 35$ ) in nuclei or cytoplasm staining. A significant difference in survival was observed between the groups (log-rank test,  $P = 0.006$ ). The median survival was 1208 days for the positive staining group and 2406 days for the negative staining group. Source data are provided as a Source Data file. **b** Representative image of IHC staining with MITF antibody on a primary basal breast cancer tumor. **c** Volcano plot depicting the changes in representation ( $\log_2$ -fold change,  $x$ -axis) and significance ( $-\log_{10}$  adjusted  $P$ ,  $y$ -axis) of shRNA expressing MDA-MB-436 cells targeting Mitf vs. Ren at day 17. qPCR-validated genes (Supplementary Fig. 10) labeled. Source data are provided as a Source Data file. **d** Hallmarks of cancer and REACTOME gene sets analyzed from the transcriptome analysis comparing MDA-MB-436 cells transduced with two independent MITF shRNAs and control. Enrichment score (ES) is shown. **e** Heat map based on the subset of differentially expressed genes where gene expression correlated with PSIONIC-inferred MITF activity across breast cancer cell lines (target genes with  $|\rho| > 0.4$  shown). Expression values were transformed (VST) and corrected. The color coding in each cell reflects the deviation from the gene's average across all samples. Red labels indicate positive correlation with inferred MITF activity, blue labels indicate negative correlation with inferred MITF activity. Bold labels indicate the existence of correlation in TCGA BASAL-BRCA tumors

two independent MITF shRNAs (Fig. 6c; Supplementary Table 5). Interestingly, commonly downregulated genes included *c-Myc* and *c-Myc* target genes, as well as additional pro-oncogenic factors, such as *IL1B*, *NT5E* (*CD73*), and other molecules with functions in tumor immune escape (Fig. 6d, Supplementary Tables 6 and 7)<sup>38,39</sup>, which were validated by qPCR (Supplementary Fig. 10C). Commonly upregulated genes were enriched in ontology terms associated with immune activation (defensins, complement, IFN, *IL15*, *CCL2*) and cell adhesion (e.g. *SVEP1*) (Fig. 6d, Supplementary Tables 6 and 7, Supplementary Fig. 10D). These effects were not associated with changes in the proliferation rate of MDA-MB-436 cells in vitro yet are suggestive of an in vivo role for MITF in the regulation of cancer—microenvironment crosstalk in basal breast cancer. Importantly, most differentially expressed genes (DEGs) identified in MDA-MB-436 upon MITF suppression correlated with PSIONIC-inferred MITF activity across multiple basal breast cancer cell lines ( $n = 29$ ; 75 out of 161 DEG,  $\sim 47\%$ ,  $|\rho| > 0.4$ , Fig. 6e) as well as across patient samples ( $n = 92$ ; 43 out of 161 DEG,  $\sim 27\%$ ,  $|\rho| > 0.4$ ). Together, these results validate the predictions made by PSIONIC on MITF activity and gene regulation in basal breast cancer.

## Discussion

With the development of high-throughput sequencing technologies, transcriptomic, proteomic, genomic profiles of tumor samples have been rapidly generated for diverse cancer types. Identifying differentially expressed genes or recurring mutations does not always clarify the molecular pathways that actually regulate tumor state and survival. There is still a large methodological gap between generating molecular profiles of tumor samples and understanding the molecular mechanisms underlying tumorigenesis and response to therapy.

Our PSIONIC method provides a systematic framework for integrating resources on regulatory genomics with tumor expression data to better understand gene regulation in cancers and infer patient-specific TF networks. PSIONIC uses a reduced rank representation model based on latent tasks, which helps regularize patient-specific regression models in light of noisy tumor gene expression data while sharing information between tumors and tumor types. Joint inference of TF activities across different tumor types may also reveal clinically relevant patient subgroups common to multiple cancers. As new ATAC-seq technologies for frozen tissue are developed<sup>40</sup>, ATAC-seq will



become feasible in clinical samples, and then TF-binding site signals from tumor-specific ATAC-seq mapped regions can be incorporated to our framework.

One limitation of our approach is the multiplicity of inferred effects, which is biologically reasonable but complicates interpretation. Our model also currently makes the assumption that a TF either induces or represses its targets, but some TFs may play either role depending on coordination with co-factors. These limitations may confound the interpretation of inferred TFs with dual activator/repressor roles. 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, the PSIONIC framework can be extended modeling of single-cell RNA-seq, as we will report elsewhere.

We used PSIONIC to perform a comprehensive transcriptional network analysis of gynecologic and basal breast cancer tumors. These tumors have not previously been subject to extensive epigenetic or computational analyses, and they all lack accurate predictors of response and treatment strategies for recurrent disease. PSIONIC can identify transcriptional processes that are active across otherwise very different tumors, such as MEF2A activity in the OV, BRCA-BASAL, and UCS cohorts. Applying our method to other pan-cancer cohorts such as squamous carcinomas or pediatric cancers might again find biological processes that are activated in a large number of tumor types and provide insight into common regulatory programs in tumors of different origin.

We demonstrated that PSIONIC-predicted activity for TFs in cell line models correlated with sensitivity to inhibition of a targetable TF, MTF1, giving a proof-of-principle for the potential therapeutic application of our approach. MTF1 target genes including PGF, HIF-1, and TGF $\beta$ 1 are involved in apoptosis, resistance, invasion, metastasis, and angiogenesis. Under normal conditions, MTF1 localizes both to the nucleus and the cytoplasm but accumulates in the nucleus upon these diverse stresses. After binding DNA, MTF1 recruits different co-regulators and often relies on other TFs, such as p300/CBP, Sp1, and HIF1 $\alpha$  for coordinated target gene expression. Its established targets have important roles in metal homeostasis, embryonic development, tumor progression, and oxidative stress or hypoxia signaling. Importantly, inhibition of MTF1 induces the expression of tumor suppressor factor Kruppel like factor 4 (KLF4)<sup>41</sup>. This leads to the downregulation of cyclin D1, blocking cell cycle progression and proliferation. The MTF1 inhibitor LOR-253 enhances apoptosis induced by cisplatin in both SKOV3 and OVCAR3 cells<sup>42</sup>, is cytotoxic to Raji and Raji/253R lymphoma cell lines<sup>43</sup>, and suppresses the proliferation of acute myeloid leukemia (AML) cell lines<sup>44</sup>. A clinical trial testing LOR-253 in patients with AML and myelodysplastic syndrome is currently ongoing (ClinicalTrials.gov: NCT02267863). Our results suggest that the potential role of MTF1 in gynecologic and basal breast cancers merits further investigation.

We also showed that several PSIONIC-predicted TF activities were significantly associated with survival outcome in basal breast, uterine serous and endometrioid carcinomas. We validated two prognostic TFs, MITF and ETV6, in independent patient cohorts, giving a proof-of-principle for the potential prognostic application of our approach.

ETV6 encodes an ETS family transcription factor that is essential for hematopoietic processes<sup>45</sup>. Indeed, immunolocalization of ETV6 on tissue samples from uterine serous cancer patients demonstrated that strong nuclear ETV6 expression is significantly associated with poor disease prognosis. Possible future validation experiments to confirm the tumor-promoting roles of ETV6 in uterine serous cancer, expression levels of ETV6

can be manipulated in cultured cells through overexpression or silencing and the effects of ETV6 on cancer cell proliferation, survival, motility, and invasion potential can be evaluated. Ultimately, in vivo validation of the roles of ETV6 expression on uterine cancer progression can be studied using uterine serous cancer-bearing mouse models through in vivo silencing of ETV6 using siRNAs.

Encouraged by MITF's prognostic value in basal breast cancer patients, we directly examined the functional impact of loss of MITF in basal breast cancer cells by transducing MDA-MB-436 cells with inducible *MITF* shRNAs followed by RNA-seq. Although MITF shRNA did not compromise tumor cell proliferation in vitro, we found many cancer-relevant genes to be regulated by MITF, including cell-surface and secreted factors repressing immune-mediated anti-tumor responses in triple-negative breast cancer (e.g. NT5E/CD73)<sup>38,39</sup>. Moreover, many of the factors de-repressed upon MITF knockdown are important players that activate anti-tumor immunity (e.g. IL15, CCL2), which suggests a potential role of MITF in evasion of immune surveillance. Hence, our data suggest MITF might play tumor-promoting roles in vivo by regulating the crosstalk of basal breast cancer cells with their tumor microenvironment. More globally, these analyses validate PSIONIC as a predictive tool to predict TF activity in specific tumor settings, as shown for MITF in basal breast cancer, that expands its role in cancer beyond its known lineage-specific functions in melanoma.

Patient-specific inference of TF networks 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

**Datasets.** RNA-seq data for each of the five tumor types were downloaded from TCGA's Firehose data run [<https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata>]. Log<sub>10</sub>-transformed RNA-seq RSEM gene expression values were unit-normalized by tumor sample. Cancer cell lines RNA-seq data were downloaded from the CCLE website [<http://www.broadinstitute.org/ccle>]. Log<sub>10</sub>-transformed RNA-seq TPM gene expression values were unit-normalized by cell line.

Bigwig files of ATAC-seq profiles of tumor samples from TCGA<sup>19</sup> including 13 UCEC-ENDO (24 with replicates) and 15 BRCA-BASAL (30 with replicates) were downloaded from <https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>.

**Cell line selection for ATAC-seq.** In this study, we chose cell lines widely used as representative of corresponding tumor types depending on availability to our group. In several cases, we are providing the first epigenomic characterization of these cell line models. ATAC-seq libraries generated from basal breast (MDA-MB-231, MDA-MB-436) high-grade serous ovarian (OVCAR8, Caov3), uterine carcinosarcoma (JHUCS, SNU685), endometrial endometrioid (AN3-CA, KLE, Ishikawa, RL95-2), and serous carcinoma (ACI-126, ACI-158) cell lines. Gynecologic cell lines OVCAR8, Caov3, JHUCS, SNU685, AN3-CA, KLE, Ishikawa, and RL95-2 were supplied by Douglas A. Levine. Uterine serous cell lines ACI-126 and ACI-158 were kindly supplied by John I. Risinger from Michigan State University. Basal breast cancer cell lines MDA-MB-231 and MDA-MB-436 were acquired from ATCC. The cell lines have been tested negative for mycoplasma contamination.

Briefly, Ishikawa and RL-95-2 derived from type I and KLE and AN3CA derived from type II endometrial carcinomas tumors have been widely used as models to investigate molecular genetics and mechanisms underlying their development, progression, and response to therapeutics<sup>46</sup>. KLE and AN3CA cells, originating from peritoneal and lymph node metastases, respectively, and RL-95-2 cells derived from a moderately differentiated (Grade 2) endometrial adenocarcinoma. Ishikawa cells were established from the epithelial component of a moderately differentiated, stage 2, endometrial adenocarcinoma. CAOV3 and OVCAR8 have been widely used as representatives of high-grade serous cancer. CAOV3 and OVCAR8 possess *TP53* mutations and substantial copy-number changes, key characteristics of high grade serous ovarian cancer (HGSOC). ACI-158 and ACI-126 are the main uterine serous (UPSC) cell lines. JHUCS-1 was established from a carcinosarcoma (malignant mixed

mesodermal tumor) of the uterus that was surgically removed from a 57-year-old Japanese woman<sup>47</sup>. SNU-685 was derived from uterine malignant mixed müllerian tumor<sup>48</sup>.

**Sample preparation for ATAC-seq.** Cell lines were re-suspended in cold PBS according to ATAC-Seq protocol<sup>49</sup>. Chromatin was extracted and processed for Tn5-mediated tagmentation and adapter incorporation, according to the manufacturer's protocol (Nextera DNA sample preparation kit, Illumina®) at 37 °C for 30 min. Reduced-cycle amplification was carried out in the presence of compatible indexed sequencing adapters. The quality of the libraries was assessed by a DNA-based fluorometric assay (Thermo Fisher Scientific™) and automated capillary electrophoresis (Agilent Technologies, Inc.). Sample preparation and sequencing for ATAC-seq was performed by Epinomics. For each sample, ATAC-seq was performed on two biological replicates.

**ATAC data analysis.** Starting from fastq files containing ATAC-seq paired-end reads, sequencing adaptors were removed using Trimmomatic<sup>50</sup>. Trimmed reads were mapped to the hg19 human genome using Bowtie2<sup>51</sup> allowing at most 1 seed mismatch and keeping only uniquely aligned reads. Duplicates were removed using Picard (<http://picard.sourceforge.net>). For peak calling the read start sites were adjusted (reads aligning to the +/- strand were offset by +4 bp/-5 bp, respectively) to represent the center of the transposase binding-event<sup>49</sup>.

BigWig files were generated using bamCoverage from the deepTools suite with options—binSize 10—normalizeTo 1 × 2451960000—ignoreForNormalization chrX. The log<sub>2</sub>-transformed ATAC-seq signal were calculated using bamCompare from deepTools<sup>52</sup>. Resulting normalized BigWig files were used as input to computeMatrix to calculate scores for regions of interest (using either scale-regions or reference-point mode) and visualized using plotHeatmap tool from deepTools.

Peak calling was performed on each cell type individually: first, the reads from different replicates were pooled, and the MACS2.0 peak caller<sup>53</sup> was then used to identify peaks with a permissive threshold ( $P < 2 \times 10^{-3}$ ). Finally, IDR was used to identify reproducible peaks using two biological replicates for each cell type (IDR <  $1 \times 10^{-2}$ ). Peaks found reproducibly in each cancer cell subtype were combined to create a genome-wide atlas of accessible chromatin regions. Reproducible peaks from different samples were merged if they overlapped by more than 75%. The atlas of chromatin accessibility across 12 gynecologic and basal breast cancer cell lines contains 282,248 peaks. The number of reproducible peaks for each cell line and number of peaks in each cancer type specific atlas are listed in Supplementary Table 1.

We associated each peak to its nearest gene in the human genome using the ChIPpeakAnno package<sup>54</sup>. ATAC-seq peaks located in the body of the transcription unit, together with the 100 kb regions upstream of the TSS and downstream of the 3' end, were assigned to the gene.

Using the MEME<sup>55</sup> curated Cis-BP<sup>56</sup> TF-binding motif reference, we scanned each ATAC-seq tumor type peak atlas and common atlas with FIMO<sup>57</sup> to find peaks likely to contain each motif ( $P < 10^{-5}$ ). We filtered TFs that were not expressed in at least 50% of samples in at least one of the five tumor types. Further, similarity of predicted target peak sets was measured using the Jaccard index (size of intersection/size of union). If two TFs had a high Jaccard index (>0.5), we looked at the mean Jaccard index of each TF with all other TFs, and we removed the TF with the largest mean Jaccard index. The final set contained 352 motifs.

We created a matrix that defines a candidate set of associations between TFs and target genes: 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.

**Differential peak accessibility.** Reads aligning to atlas peak regions were counted using the countOverlaps function of the R packages GenomicAlignments and GenomicRanges<sup>58</sup>. Differential accessibility of these peaks was then calculated for all pairwise comparisons of cancer types using DESeq2<sup>59</sup>.

**Motifs underlying differential accessibility in cell lines.** The shift in the cumulative distribution of chromatin accessibility changes (log<sub>2</sub>-fold change) of the subset of the atlas occupied by each TF, compared to that of the background atlas, was measured by a one-sided KS test in either direction. The foreground occurrence is the number of peaks containing a particular TF motif within the group of 5000 differentially open or 5000 differentially closed peaks according to log<sub>2</sub>-fold change read counts. The background occurrence is the number of peaks containing a particular TF motif among all the differentially accessible peaks.

**TCGA ATAC-seq analysis.** Currently, only bigwig files are publicly available for TCGA ATAC-seq but not raw data. We extracted the sum of ATAC-seq signals ±0.5 kb from the peak center for differentially accessible cell line peak regions for patients from these bigwig files and used these values for statistical analyses.

**Multitask learning.** For MTL, we trained regression models jointly for all tumors using *grouping and overlap in MTL* (GO-MTL)<sup>17</sup>. In this approach, prediction of each gene expression  $y_i$  is considered one task, and we wish to solve  $T$  tasks jointly

so that information is “shared” between them. Let  $X$  be the data matrix of size  $d \times N$  where each row represents a gene and each column is a motif hit score representing the target genes of a TF. We assume there are  $K (< T)$  latent basis tasks and each observed task can be represented as linear combination of a subset of these basis tasks.

Briefly, we jointly learn regression models  $w_i$  that predict gene expression as linear combinations of latent regulatory programs in tumors. Formally, we learn a model matrix  $W = LS$ , where every column of matrix  $W$  is a model vector for a tumor transcriptional regulatory network,  $w_i$ ;  $L$  is matrix of latent tasks  $L = (l_1 \dots l_K)$ ; and  $S$  expresses the tumor transcriptional regulatory network models as linear combinations of the latent tasks,  $w_i = Ls_i$ . The model vector  $w_i$  represents the inferred global role of these TFs in driving gene expression; the event's true gene expression is denoted by  $y_i$ ; and the predicted gene expression is given by  $w_i X_i$  (treating both as row vectors for notational convenience). The matrix  $L$  captures the predictive structure of the tasks and the grouping structure is determined by matrix  $S$ . Tasks that have same sparsity pattern can be seen as belonging to the same group, while tasks whose sparsity patterns are orthogonal to each other can be seen as belonging to different groups. The partial sharing of latent basis tasks allows us to do away with the concept of disjoint groups. Any task that does not share latent bases with any other task in the pool can be seen as outlier task. Our learning cost function takes the following form:

$$\min_{L, S} \sum_{i=1}^T \frac{1}{N} \|y_i - X_i L s_i\|^2 + \mu \|S\|_1 + \lambda \|L\|_F^2 \quad (1)$$

The parameter  $\mu$  controls the sparsity in  $S$ . The penalty on the Frobenius norm of  $L$  regularizes the predictor weights to have low  $\ell_2$  norm and avoids overfitting.

To assess single task learning (STL) performance, we trained ridge regression models for each tumor (task) independently. We fit the ridge regression models using the SLEP MATLAB package and evaluated performance on held-out genes.

**In vitro drug-sensitivity analysis.** Detailed information on cell culture media is provided in the Supplementary Table 8. All cell lines were cultured under standard conditions at 37 °C and 5% CO<sub>2</sub>. Cells were plated at 10–20% confluency (with the exception of JHUCS-1, MDA-MB-436, and RL-95 which were plated at ~50%) in 24-well plates in complete medium, and incubated inside an IncuCyte ZOOM system (Essen BioScience, Inc., MI, USA). The following day (22–24 h later), cells were exposed to LOR-253 (MedChemExpress, NJ, USA) at 0, 50, 250, or 1250 nM. To monitor cell growth, phase contrast images of the cell cultures in the presence or absence of the drug were captured automatically at 2-h intervals for up to 36 h, and occupied area of the cells (% confluency) was calculated using the IncuCyte image analysis software. We analyzed drug response data using a recently developed growth rate inhibition (GR) metric that corrects for differences in cell proliferation rates<sup>60</sup>.

**Immunohistochemistry for ETV6 in UPSC.** The population considered for this study consisted of 31 patients diagnosed with uterine papillary serous carcinoma (UPSC) in stage III or IV, who underwent salpingo-oophorectomy at University of Texas MD Anderson Cancer Center (MDACC) and did not receive neoadjuvant therapy. Patients were divided in two groups based on survival: <2 years (eight patients) and more than 10 years (six patients). This study was approved by the Institutional Review Board at the MDACC. Informed consent was obtained from all patients. Formalin-fixed paraffin-embedded (FFPE) tumor blocks of archived UPSC were obtained from the repository of the Department of Gynecologic Oncology and Reproductive Medicine at MDACC. Clinical information was obtained from the electronic medical records.

FFPE 4 μm sections from patient tissues were deparaffinized and fixed in methanol prior to antigen retrieval in heated citrate buffer (pH 6.0, Poly Scientific R&D Corp.) at 120 °C for 7 min, followed by 10 min at 90 °C. Endogenous peroxidase was blocked with hydrogen peroxide solution (Millipore Sigma) 3% for 10 min. Protein blocking was performed using PBS-Tween 3%, BSA 1% donkey serum (Millipore Sigma) for 30 min. Anti-ETV6 antibody from Sigma (catalog number: HPA000264) at the titer of 1:500 was used. Samples were incubated with ETV6 for 1 h (1:75, polyclonal, Millipore Sigma) followed by use of MACH 3 rabbit HRP polymer detection (Biocare Medical). Antibodies were visualized by means of a dextran-polymer conjugate technique (EnVision+, Dako) using 3,3'-diaminobenzidine (DAB) (Dako) as chromogen. Tissue sections were counterstained with haematoxylin. Images were captured with a Leica DM LB microscope (Leica Microsystems).

Intensity of ETV6 stain was graded separately in nuclei and cytoplasm as 0 (negative), 1 (weak), 2 (moderate), and 3 (strong). Patient samples were divided into two groups based on this scoring: weak and medium nuclear (scores 0, 1 or 2) ETV6 staining,  $n = 20$ ; and patient samples with strong (score = 3) nuclear ETV6 staining,  $n = 11$ . Statistical analysis studied the association between ETV6 staining intensity scores and survival time by using Kaplan–Meier curves and Log-rank test ( $P < 0.004$ ).

**Immunohistochemistry for MIF in TNBC.** Immunohistochemical stain for MIF was performed on tissue microarrays (TMAs) containing triple negative breast carcinoma (TNBC). TNBC was defined as invasive breast carcinoma with

ER and PR staining in <1% of the tumor cells by immunohistochemistry and no HER2 overexpression by immunohistochemistry and no HER2 amplification by fluorescence in situ hybridization. Assessment of ER, PR, and HER2 follows the ASCO/CAP guidelines. Triplicate 0.6 mm diameter core from formalin-fixed paraffin-embedded TNBC blocks were used to construct the TMAs. MITF (D5) clone Dako Ab (catalog number: M3621) was used on Leica platform with ER2 pre-treatment for 40 min. Pr Ab dilution: 1:50. Standard DAB Kit was used. MITF staining of any percentage and any intensity was considered positive. Some positivity is seen in the tumor-infiltrating lymphocytes. Clinical information was obtained from the electronic medical records. The association between MITF staining and survival was analyzed using Kaplan–Meier curves and log-rank test. This study was approved by the Institutional Review Board at MSKCC.

**Sample preparation for RNA-Seq and data analysis.** shRNA vector cloning: shRNA sequences for targeting human MITF were designed using the Splash algorithm prediction tool<sup>61</sup>. The shRNA was cloned into the LT3GEPiR miR-E backbone<sup>37</sup> enabling inducible shRNA expression in transduced cells upon doxycycline treatment. Two independent shRNA were used to target MITF (sh962 or shQa), and a previously described shRNA-targeting Renilla luciferase<sup>62</sup> was used as control. The sequences of all shRNAs can be found in Supplementary Table 9.

For protein lysates cells were incubated with RIPA buffer supplemented with protease inhibitors (Protease inhibitor tablets, Roche) for 30 min and cleared by centrifugation (15 min 14,000 rpm 4 C). Protein was quantified using the Bio-Rad protein assay (Cat. 500006). Primary antibody incubation was performed overnight at 4 °C in Tris-buffered saline containing 5% milk and 0.05% Tween-20. The following primary antibodies were used for immunoblotting: Mitf (ab12039, Abcam), Actin-HRP (A3854, Sigma). Mouse HRP-linked secondary antibody (GE Healthcare) was used and blots were developed with Lumi-Light Western Blotting Substrate (Roche).

MDA-MB-436 cells were maintained in RPMI 1640, supplemented with 10% FBS (Gemini, Cat. 900-208), 1X Glutamax (Gibco, Cat. 35050061), and penicillin–streptomycin (1%). MDA-MB-436 cells were transduced with a lentiviral construct (LT3GEPiR<sup>37</sup>) and infected cells were then selected with puromycin (1 µg/ml, Sigma-Aldrich) and harvested at 5 or 17 days after culture in growth media containing doxycycline (1 µg/ml, Sigma-Aldrich). Total RNA was extracted using TRIzol (Thermal Fisher Scientific, Cat. 15596018), according to manufacturer's instructions. For RT-qPCR experiments, cDNA was obtained using Transcriptor First Strand cDNA Synthesis Kit (Roche, Cat. 04896866001). Gene-specific primer sets for human sequences were designed using PrimerBank [<https://pga.mgh.harvard.edu/primerbank/>] (see Supplementary Table 10 for qPCR primer sequences). *HPRT1* served as endogenous normalization controls. RT-qPCR was carried out in triplicate using PerfeCTa SYBR Green FastMix (QuantaBio, Cat. 95072-012) on the ViiA 7 Real-Time PCR System (Life technologies). For RNA-Seq, 500 µg of RNA was used, and PolyA mRNA was selected using beads coated with polyT oligonucleotides. Purified polyA mRNA was subsequently fragmented, and first and second strand cDNA synthesis performed using standard Illumina mRNA library preparation protocols (TruSeq RNA Sample Prep Kit v.2). Double-stranded cDNA was subsequently processed for TruSeq dual-index Illumina library generation. For sequencing ~30–40 million 80 bp single-end reads were acquired per replicate condition in a NextSeq Illumina system at the integrated genomics operation (IGO) Core at MSKCC.

Raw RNA-Seq reads were trimmed and filtered for quality using Trimmomatic<sup>50</sup>. Reads were aligned using STAR<sup>63</sup> against GRCh37.75(hg19). The RefSeq transcript annotations of the hg19 version of the human genome was used for the genomic location of transcription units. Genome-wide transcript counting was performed by HTSeq<sup>64</sup> to generate a matrix of raw counts. Differential expression of genes across cell types was calculated using DESeq2<sup>59</sup>. FDR threshold of 0.05 was imposed unless otherwise stated. A log<sub>2</sub>-fold change cutoff of 1 was used. We functionally annotated our expression profiling and performed gene set enrichment analysis<sup>65</sup> on all curated gene sets in the Molecular Signatures Database.

Pooled data is presented as mean ± standard deviation (SD) values of duplicate or triplicate biological replicates, as indicated in corresponding figure legends. Statistical significance differences compared to shMITF/shRen (for day 17) or shMITF +Dox/–Dox (for day 5) were determined by an unpaired one-tailed Student's *t*-test. In figures, \* stands for *P* < 0.05, \*\* for *P* < 0.01, and \*\*\* for *P* < 0.001.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The ATAC-seq and RNA-seq data have been deposited in the Gene Expression Omnibus accession number GSE114964 and GSE129337, respectively. The source data underlying Figs. 2, 3A–E, 4A–C, 5A, 6 and Supplementary Figs. 1, 2, 3, 6–10 are provided as Source Data files.

## Code availability

The software for PSIONIC is available from <https://github.com/osmanbeyoglu/PSIONIC>.

Received: 25 July 2018 Accepted: 2 September 2019

Published online: 25 September 2019

## References

- Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
- Balwiercz, P. J. et al. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24**, 869–884 (2014).
- Osmanbeyoglu, H. U., Pelosof, R., Bromberg, J. F. & Leslie, C. S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res.* **24**, 1869–1880 (2014).
- Setty, M. et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol. Syst. Biol.* **8**, 605 (2012).
- Joseph, R. et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor- $\alpha$ . *Mol. Syst. Biol.* **6**, 456 (2010).
- Yu, J. et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
- Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
- Sharma, N. L. et al. The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer Cell* **23**, 35–47 (2013).
- Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
- Evgeniou, T., Michelli, C. A., Pontil, M. & Shawe-Taylor, J. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.* **6**, 615–637 (2005).
- Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705 e699 (2018).
- Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
- Cherniack, A. D. et al. Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell* **31**, 411–423 (2017).
- Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
- Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- Kumar, A., Daumé III, H. Learning task grouping and overlap in multitask learning. in *Proceedings of the 29th International Conference on Machine Learning* (eds. Langford, J. & Pineau, J.). (Omnipress, USA, 2012).
- Gonzalez, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.* **47**, 1249–1259 (2015).
- Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, <https://doi.org/10.1126/science.aav1898> (2018).
- Yu, D. D., Guo, S. W., Jing, Y. Y., Dong, Y. L. & Wei, L. X. A review on hepatocyte nuclear factor-1 $\beta$  and tumor. *Cell Biosci.* **5**, 58 (2015).
- Tetreault, M. P., Yang, Y. & Katz, J. P. Kruppel-like factors in cancer. *Nat. Rev. Cancer* **13**, 701–713 (2013).
- Gutierrez-Hartmann, A., Duval, D. L. & Bradford, A. P. ETS transcription factors in endocrine systems. *Trends Endocrinol. Metab.* **18**, 150–158 (2007).
- Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Roy, L. et al. ARID3B increases ovarian tumor burden and is associated with a cancer stem cell gene signature. *Oncotarget* **5**, 8355–8366 (2014).
- Taube, E. T. et al. Wilms tumor protein 1 (WT1)—not only a diagnostic but also a prognostic marker in high-grade serous ovarian carcinoma. *Gynecol. Oncol.* **140**, 494–502 (2016).
- Coosemans, A. et al. Wilms tumor gene 1 (WT1) is a prognostic marker in high-grade uterine sarcoma. *Int. J. Gynecol. Cancer* **21**, 302–308 (2011).
- Hosono, S. et al. Expression of Twist increases the risk for recurrence and for poor survival in epithelial ovarian carcinoma patients. *Br. J. Cancer* **96**, 314–320 (2007).
- Zhang, Q., Stovall, D. B., Inoue, K. & Sui, G. The oncogenic role of Yin Yang 1. *Crit. Rev. Oncol.* **16**, 163–197 (2011).
- Gunther, V., Lindert, U. & Schaffner, W. The taste of heavy metals: gene regulation by MTF-1. *Biochim. Biophys. Acta* **1823**, 1416–1425 (2012).
- Miller, D. S. et al. Phase II evaluation of pemetrexed in the treatment of recurrent or persistent platinum-resistant ovarian or primary peritoneal carcinoma: a study of the Gynecologic Oncology Group. *J. Clin. Oncol.* **27**, 2686–2691 (2009).

31. Selvendiran, K. et al. Hypoxia induces chemoresistance in ovarian cancer cells by activation of signal transducer and activator of transcription 3. *Int. J. Cancer* **125**, 2198–2204 (2009).
32. Harder, L., Puller, A. C. & Horstmann, M. A. ZNF423: transcriptional modulation in development and cancer. *Mol. Cell. Oncol.* **1**, e969655 (2014).
33. Ciocca, D. R. & Fanelli, M. A. Estrogen receptors and cell proliferation in breast cancer. *Trends Endocrinol. Metab.* **8**, 313–321 (1997).
34. Tangen, I. L. et al. Loss of progesterone receptor links to high proliferation and increases from primary to metastatic endometrial cancer lesions. *Eur. J. Cancer* **50**, 3003–3010 (2014).
35. Tsai, Y. C. et al. Epidermal growth factor receptor signaling promotes metastatic prostate cancer through microRNA-96-mediated downregulation of the tumor suppressor ETV6. *Cancer Lett.* **384**, 1–8 (2017).
36. Hartman, M. L. & Czyz, M. MITF in melanoma: mechanisms behind its expression and activity. *Cell. Mol. Life Sci.* **72**, 1249–1260 (2015).
37. Fellmann, C. et al. An optimized microRNA backbone for effective single-copy RNAi. *Cell Rep.* **5**, 1704–1713 (2013).
38. Buisseret, L. et al. Clinical significance of CD73 in triple-negative breast cancer: multiplex analysis of a phase III clinical trial. *Ann. Oncol.* **29**, 1056–1062 (2018).
39. Loi, S. et al. CD73 promotes anthracycline resistance and poor prognosis in triple negative breast cancer. *Proc. Natl Acad. Sci. USA* **110**, 11091–11096 (2013).
40. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
41. Kindermann, B., Doring, F., Budczies, J. & Daniel, H. Zinc-sensitive genes as potential new target genes of the metal transcription factor-1 (MTF-1). *Biochem. Cell Biol.* **83**, 221–229 (2005).
42. Wang, B. et al. KLF4 expression enhances the efficacy of chemotherapy drugs in ovarian cancer cells. *Biochem. Biophys. Res. Commun.* **484**, 486–492 (2017).
43. Tsai, C. Y. et al. APTO-253 is a new addition to the repertoire of drugs that can exploit DNA BRCA1/2 deficiency. *Mol. Cancer Ther.* **17**, 1167–1176 (2018).
44. Local, A. et al. APTO-253 stabilizes G-quadruplex DNA, inhibits MYC expression, and induces DNA damage in acute myeloid leukemia cells. *Mol. Cancer Ther.* **17**, 1177–1186 (2018).
45. Eguchi-Ishimae, M. et al. Leukemia-related transcription factor TEL/ETV6 expands erythroid precursors and stimulates hemoglobin synthesis. *Cancer Sci.* **100**, 689–697 (2009).
46. Korch, C. et al. DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. *Gynecol. Oncol.* **127**, 241–248 (2012).
47. Yamada, K. et al. Establishment and characterization of JHUCS-1 cell line derived from carcinosarcoma of the human uterus. *Hum. Cell* **17**, 139–144 (2004).
48. Yuan, Y. et al. Establishment and characterization of cell lines derived from uterine malignant mixed Mullerian tumor. *Gynecol. Oncol.* **66**, 464–474 (1997).
49. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Ramirez, F., Dunder, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
53. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
54. Zhu, L. J. et al. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinform.* **11**, 237 (2010).
55. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
56. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
57. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
58. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
60. Hafner, M., Niepel, M., Chung, M. & Sorger, P. K. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* **13**, 521–527 (2016).
61. Pelossof, R. et al. Prediction of potent shRNAs with a sequential classification algorithm. *Nat. Biotechnol.* **35**, 350–353 (2017).
62. Tasdemir, N. et al. BRD4 connects enhancer remodeling to senescence immune surveillance. *Cancer Discov.* **6**, 612–629 (2016).
63. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
65. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

## Acknowledgements

We would like to thank John I. Risinger for sending us the ACI158 and ACI126 cell lines, Sha Tian for her technical assistance in the RNA-seq library preparation, and Irina Linkov for performing MITF staining. 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/>. This work was supported by NCI R21 award CA205819. H.U.O. is supported by NCI K99/R00 award CA207871.

## Author contributions

H.U.O. and C.S.L. conceived and designed the study. H.U.O. carried out the model training and computational validation. H.U.O. and C.S.L. analyzed data and wrote the manuscript. F.S. performed the experimental validation for in vitro drug sensitivity and under supervision of G.C. and helped to write the experimental validation section. A.R.-V. and T.-L.Y. performed the ETV6 immunohistochemical staining under the supervision of S.C.M. and helped to write the experimental validation section. D.A.-C. and H.-A.C. performed MITF knock-down under supervision of S.W.L. and helped to write the experimental validation section. P.R. gathered clinical data for basal breast cancer patients. H.Y.W. analyzed MITF tissue microarrays. P.J. and D.A.L. assisted with the study design.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-12291-6>.

**Competing interests:** P.R. reports consulting/advisory board for Novartis and Institutional Research support from Illumina and GRAIL, Inc. The remaining authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Peer review information** *Nature Communications* thanks Lucia Peixoto and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019