SUPPLEMENTARY DATA

(Including Supplementary Methods, 6 Supplementary Figures, and 1 Supplementary Table)

Supplementary Methods

Data collection and preprocessing

All ChIP-seq data were from Cistrome DB (1) and were processed using ChiLin, a ChIP-seq and DNase-seq analysis and quality control pipeline (2). Sequence reads were mapped to the reference human genome hg38 using BWA (3), and TF binding sites were called using MACS2 (4). Peaks meeting a minimum fold enrichment of 5 were selected for the analysis. For each TF ChIP-seq dataset, a regulatory potential (RP) score was calculated for each gene in the genome using BETA (5), as the weighted sum of peaks within 100kb from the transcription start site (TSS) of that gene.

Molecular profiling data for over 10,000 primary cancer samples from 31 cancer types were collected from TCGA, including RNA-seq transcriptomic profiles, gene-based copy number variation (CNV) profiles, DNA methylation profiles on gene promoters, and clinical survival information for each patient. Tumor purity score for each sample was calculated using CHAT (6) based on CNA profiles, and infiltrating immune cell concentration for each sample was estimated using TIMER (7) based on RNA-seq transcriptomic profiling data.

Cancer type reannotation based on TCGA transcriptomic profile clustering

In order to obtain robust and consistent cancer type annotations, 9,637 tumor samples from 31 cancer types (originally annotated by TCGA) were re-clustered based on their transcriptomic profiles (RNA-seq) (Fig. S1). Cancer types were re-annotated based on this re-clustering, to reflect the actual molecular features of each sample. Some unambiguous cancer subtypes were annotated separately as different cancer types (e.g. BRCA_1 for mainly luminal breast cancer and BRCA_2 for mainly basal breast cancer (8) (Fig. S2a), while some samples from different TCGA annotations but having similar transcriptomic profiles were merged as one cluster and re-annotated as the representative cancer type with an asterisk (e.g. COAD and READ merged as

COAD_READ*, having less expression differences than the 2 breast cancer subtypes, Fig. S2b). We also confirmed that such reclustering was not confounded with sample batch effect, as shown in Fig. S1c. Cancer type abbreviations used here are the same as TCGA cancer abbreviations (https://tcga-data.nci.nih.gov/tcga/). For simplicity, we refer to the reannotated cancer clusters as "cancer types". As a result, 29 reannotated cancer types were retained and used to construct the Cistrome Cancer database (Supplementary Table 1).

Functional enhancer prediction

For each cancer type, cancer-specific genes, defined as up-regulated genes in the cancer samples compared with normal samples, were first identified by analyzing the TCGA transcriptomic RNA-seq data using VOOM-LIMMA (9) with a p-value cutoff of 0.01 and fold-change cutoff of 2. Eliminating the cancer types with less than 15 available normal samples for robust differential expression analysis, we obtained cancer-specific gene sets for 15 cancer types. Note that cancer-specific genes are not specific to each cancer type. Genes can repetitively occur in multiple cancer types.

The MARGE-express module (4) was used to generate functional enhancer profiles for each cancer type, integrating information from over 1,200 H3K27ac ChIP-seq datasets collected in the Cistrome DB database. For each H3K27ac ChIP-seq dataset, MARGE defines a regulatory potential (RP) score, P_i, for each gene i by summarizing H3K27ac signals within a 200kb region surrounding its transcription start site (TSS).

$$P_i = \sum_{\Delta_{ik} \in [-100\text{kb}, 100\text{kb}]} w_{ik} z_k$$

where z_k is the H3K27ac ChIP-seq read count at genomic location k, Δ_{ik} is the distance from the TSS of gene i to genomic location k, w_{ik} is the weight as a function of Δ_{ik}, defined as

$$w_{ik} = \frac{2e^{-\mu|\Delta_{ik}|}}{1 + e^{-\mu|\Delta_{ik}|}}$$

where the decay rate parameter μ is set so that a H3K27ac read 10kb from the TSS contributes one-half of a read at the TSS, determined empirically. After RP scores for all genes were calculated for each ChIP-seq dataset j in the compendium, a logistic regression was performed to

retrieve relevant H3K27ac profiles that accurately model the given cancer-specific gene set for each cancer type.

$$y_i \sim \alpha_0 + \sum_j \alpha_j P_{ij}$$

where y_i is the indicator of whether a gene belongs to the cancer-specific gene set (y_i =1) or not (y_i =0), P_{ij} is the RP score for gene i in H3K27ac dataset j, and α is the vector of regression coefficients. Using 5-fold cross validation, 10 relevant H3K27ac datasets from the over 1,200 dataset compendium were retrieved by forward stepwise regression. These 10 H3K27ac profiles that best model the cancer-specific gene set mark functional enhancers regulating this gene set. Then MARGE adopts a semi-supervised learning approach to identify the relative weights between these 10 H3K27ac datasets, and uses the union DNaseI hypersensitive sites (UDHS) ranked by the weighted integration of H3K27ac signals as the predicted genomic profile of the enhancers functional to regulate these cancer-specific genes. In a predicted enhancer profile, the MARGE-predicted enhancer score S_l on UDHS site l is calculated as

$$S_l = \sum_{m=1}^{10} \lambda_m u_{lm}$$

where u_{lm} is the normalized H3K27ac read count within a 1kb region centered around UDHS site l from retrieved H3K27ac dataset m (m = 1,…,10), and λ_m is the weight for dataset m imputed from the semi-supervised learining. In order to get optimized results, MARGE was run 5 times for each cancer type and the prediction with the highest ROC-AUC score was selected as the output result.

MARGE-predicted enhancer profiles are available for download or direct visualization using WashU genome browser (10) or UCSC genome browser (11). Cancer-specific genes and MARGE retrieved integrative RP score for each gene are also available for download, where genes with high-ranked RP scores are “super-enhancer” target cancer type specific genes.

Transcription factor activity and target gene prediction

Target genes for each active TF in each cancer type were predicted from integrative analysis of TCGA molecular profiling data and public TF ChIP-seq data.

1) **TF activity by cancer type.** TF activity in each cancer type was assessed by 2 measurements: average expression level (RPKM) and the expression ratio score. The average expression level (RPKM) was calculated for all tumor samples for each cancer type. The expression ratio score of a TF in a cancer type was defined as the fraction of tumor samples in which the TF expression level passes a threshold. The threshold is TF-dependent and defined as the larger value between 1 RPKM and the median expression level of this TF across all tumor samples from all cancer types (Fig. S3a). The expression ratio score reflects cancer type specific TF expression. Target genes were predicted only for the active TFs in any cancer type. A TF is defined as active if its expression ratio score passes a threshold of 0.25.

2) **Candidate target identification by expression correlation.** Candidate target genes of each TF in a cancer type were identified as having a high expression correlation with the TF across all tumor samples in this cancer type. Using the expression correlations of 1 million randomly selected pairs of genes in a cancer type as a background null distribution, a gene is a candidate target of an active TF if their expression correlation coefficient's absolute value is ranked in the top 5% in the null distribution (Fig. S3b). To eliminate potential confounding effects of CNV, tumor purity, and promoter DNA methylation level on the expression variation across tumor samples, a multiple regression model was constructed for each gene *i*

$$\text{Expression}(i) \sim \text{Expression}(\text{TF}) + \text{CNV}(i) + \text{TumorPurityScore}(i) + \text{DNAme}(i)$$

and genes with significant regression coefficient (under a p-value threshold of 0.01) against the TF expression were further selected as candidate target genes.

3) **Target predictability of TF ChIP-seq data.** The RP scores on all genes for each ChIP-seq dataset were used to generate a prediction model of the candidate target genes. In order to evaluate whether the TF ChIP-seq data is predictive of candidate target genes, a likelihood ratio test between the prediction model using TF ChIP-seq and chromatin input sequencing data (L1) and the prediction model using only chromatin input sequencing data (L2) was performed.

$$L1:\ Y \sim \beta_0 + \beta_1 X_{bg} + \beta_2 X_{ChIP}$$

$$L2:\ Y \sim \beta_0 + \beta_1 X_{bg}$$

where Y is the expression correlation between the TF and candidate target genes, X_{ChIP} represents TF ChIP-seq RP scores, X_{bg} represents chromatin input RP scores. 500 chromatin

input sequencing datasets were collected in different cell types as the genomic background. For each cancer type and each TF, the most correlated ChIP-seq sample (X_{ChIP}) with expression correlation (Y) was chosen and TF ChIP-seq was matched with the most similar input sample (X_{bg}). Two prediction models were built based on the selected TF ChIP-seq sample and imput sample. The likelihood ratio test was used to compare the goodness of fit of two models, to determine whether the TF ChIP-seq (X_{ChIP}) had a higher prediction power than ChIP-seq input. The likelihood ratio score is also reported to reflect the prediction power of the TF ChIP-seq data.

4) **Target gene identification.** Putative target genes of a TF were predicted as showing high expression correlation as well as being supported by ChIP-seq derived TF binding information. Using candidate target genes with a high expression correlation and RP scores from TF ChIP-seq data, a generalized multiple regression model was used to select informative ChIP-seq datasets. At most 10 informative ChIP-seq datasets were selected in the model, adjusted RP scores were generated as a linear combination using regression coefficients. The final target genes were identified as genes passing a threshold on this model defined RP score. The threshold was determined as maximizing the Youden index (sensitivity+specificity−1) (12). The final target gene list was ranked by the rank product between expression correlation and adjusted RP score by default, with the expression correlation and the adjusted RP score percentile represented as the color density and the size of the square, respectively.

Web interface

The Cistrome Cancer website displays cancer enhancer predictions and TF target predictions on 2 separate webpages. *A video introduction is available on the home page.*

On the Cancer enhancer prediction webpage, given a gene symbol in the search field, information about whether this gene is up-regulated in each cancer type or not can be displayed. Predicted regulatory profile near this gene locus can also be displayed by being redirected to genome browser. Cancer-specific gene lists, MARGE-retrieved integrative RP scores, and MARGE predicted enhancer profiles are available for download. Genes with high integrative RP scores can be used as “super-enhancer”-like target genes.

On the TF target prediction web page, a list of available TFs is available for search, and prediction results for each TF is displayed as an individual page (Fig. S4). TF activities and putative target genes are displayed in heatmaps, in which each column represents a cancer type. TF activity information, including TF average expression (RPKM), the expression ratio, and the likelihood ratio for ChIP-seq predictability are shown on the top three rows, in green, purple, and cyan, respectively. Predicted target genes are listed below with colored squares representing the prediction scores for each gene and each cancer type where the TF is active. The color of the square represents the expression correlation between the target gene and the TF across tumor samples. The size of the square represents the adjusted RP score (as a percentile) from TF ChIP-seq data. Predicted target genes can be ranked by the rank product of expression correlation and adjusted RP score in each cancer type.

Auxiliary analysis functions are also provided in the TF target prediction webpage. These functions include:

1) **Advanced search:** For a given TF, selected target genes in a query cancer type can be displayed with any given expression correlation and adjusted RP score cutoffs.

2) **Survival and expression:** Given a query TF and a cancer type, a Kaplain-Meier survival plot can be generated using the top 25% and bottom 25% of tumor samples based on the TF expression level. Boxplots of expression levels of the TF in all tumor samples and all normal samples can also be generated and displayed.

3) **Target gene overlap:** A Venn diagram showing the number of shared target genes between two TFs in any two cancer types can be generated.

4) **Download:** For each TF, the bulk datasets for expression correlation with all genes in all cancer types and adjusted RP scores on all genes in all cancer types are available for download.

Supplementary References

1. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic acids research. 2017;45:D658-D62.
2. Qin Q, Mei S, Wu Q, Sun H, Li L, Taing L, et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC bioinformatics. 2016;17:404.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60.
4. Wang S, Zang C, Xiao T, Fan J, Mei S, Qin Q, et al. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome research. 2016;26:1417-29.
5. Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nature protocols. 2013;8:2502-15.
6. Li B, Li JZ. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. Genome biology. 2014;15:473.
7. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome biology. 2016;17:174.
8. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61-70.
9. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology. 2014;15:R29.
10. Zhou X, Wang T. Using the Wash U Epigenome Browser to examine genome-wide sequencing data. Current protocols in bioinformatics. 2012;Chapter 10:Unit10
11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002;12:996-1006.
12. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. Biometrical journal Biometrische Zeitschrift. 2008;50:419-30.

Supplementary Figure Legends

Figure S1: Cancer sample reclustering. Heatmap of the k-means clustering of 9,637 TCGA tumor samples over 31 cancer types. Rows and columns are clustered according to cancer type and k-means cluster center, with each row representing a cancer type and each column a k-means cluster.

Figure S2: Cancer type reannotation.
a. Pie charts show the breast cancer molecular subtypes of tumors in BRCA_1 (mostly luminal) and BRCA_2 (mostly basal). Breast cancer subtype information was derived from PAM50 subtype assignments.
b. Density plots show the distributions of log2 expression fold-change (FC) in BRCA_1 vs. BRCA_2 and COAD vs. READ, supporting the merging of COAD with READ and the splitting of BRCA into BRCA_1 and BRCA2. The x-axis represents log2 expression fold-change (FC) and the y-axis represents frequency.
c. Percentage of BRCA_1 and BRCA_2 samples by batches. Each dot represents one batch, and the x- and y-axes represent the percentage of samples classified to BRCA_1 and BRCA_2 in that batch. Most batches have similar proportion of samples assigned to the two batches, suggesting minimal batch effect.

Figure S3: TF target prediction in cancer.
a. Schematic of the TF expression ratio calculation. For a given TF an expression baseline threshold was set as the larger value between 1 RPKM (green dotted line) and the median expression level of the TF across tumor samples of all cancer types (black dotted line, chosen). For each cancer type, the TF expression ratio is calculated as the fraction of tumor samples (purple and white) in which the TF expression level is above the baseline (purple). A TF is considered to be active if its expression ratio is greater than 0.25.
b. Distribution of gene expression Spearman correlation coefficients between gene pairs in LUSC (blue) and TGCT (red). The solid lines represent the background distributions calculated from one million randomly selected gene pairs. The dashed lines represent expression correlation distribution between FOXM1 and every other gene.

Figure S4: Snapshot of the Cistrome Cancer TF target prediction webpage. For each TF (scrollable TF list on the left, with AR displayed as an example on the right), predicted target genes are displayed for all cancer type where the TF is active. Top rows represent the TF average expression level (RPKM), percent of tumors expressing the TF above baseline (expression ratio), and TF ChIP-seq predictability (likelihood ratio), in green, purple, and cyan, respectively. A red/blue square is displayed if a gene (row) is predicted as a target of this TF in a particular cancer type (column). The color of the square represents the gene expression correlation between the TF and the predicted target gene. The size of the square represents the TF binding score for the predicted target gene, calculated as the percentile of integrative regulatory potential score from selected TF ChIP-seq profiles.

Figure S5: FOXM1, a pan-cancer active TF.

a. FOXM1 expression levels in tumor and normal samples in different cancer types.

b. Significantly different survival curves from the top 25% (red) and bottom 25% (black) of BRCA_1 patients according to FOXM1 expression level. The p-value was calculated using the Kaplain-Meier log-rank test.

c. Active TFs ranked by proportion of overlapping target genes with FOXM1 in BRCA_1.

d. Venn diagram showing ChIP-seq peak overlap between FOXM1, MYBL2 and E2F1.

Figure S6: Distinct patterns of immune cell infiltration in kidney and colorectal cancers.

a. STAT4 expression in KIRC tumor and normal samples. The p-value was calculated by t-test.

b. Survival curves from the top 25% (red) and bottom 25% (black) KIRC patients according to STAT4 expression level. The p-value was calculated using the Kaplain-Meier log-rank test.

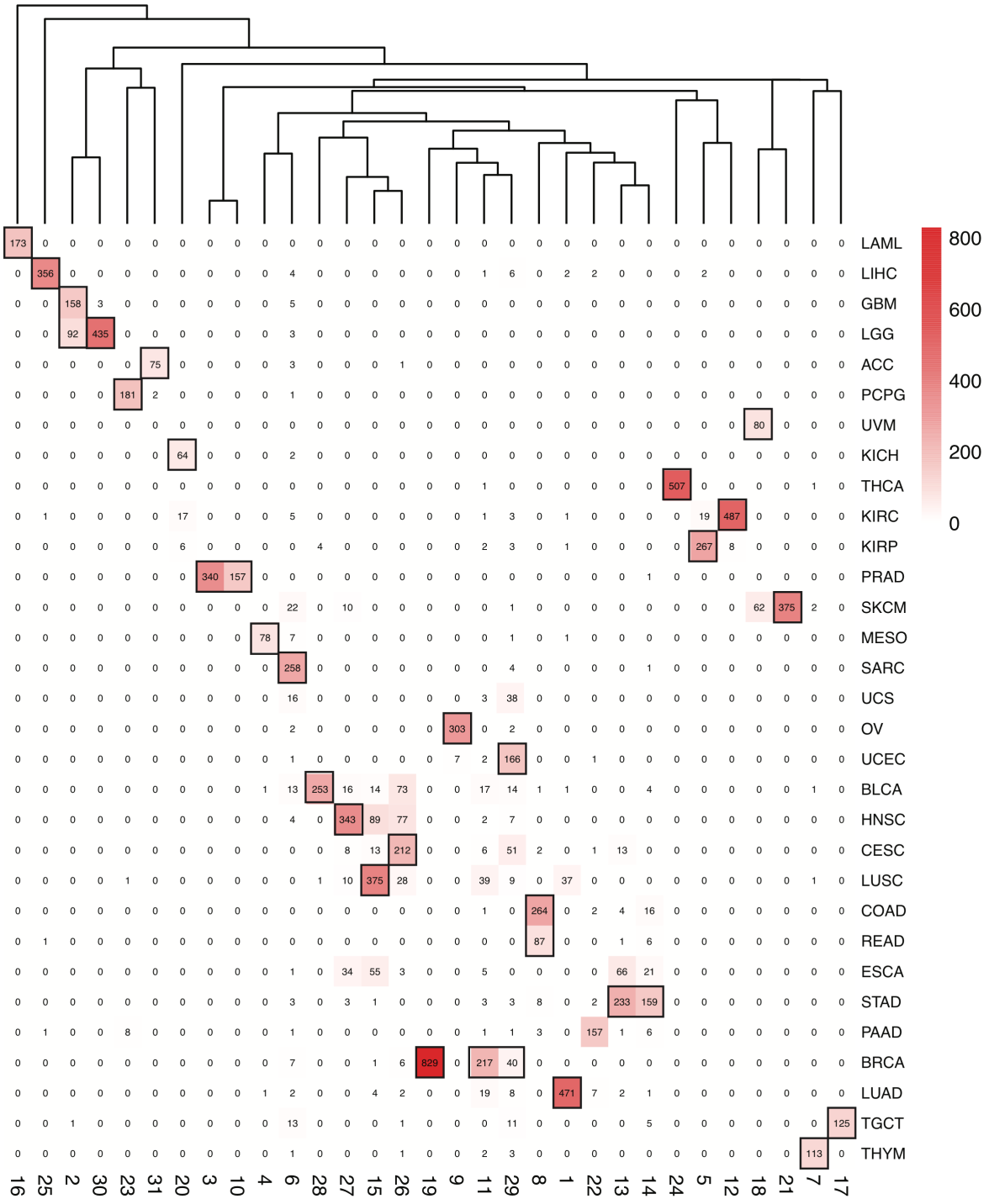
c. Waterfall plot of the differential expression of predicted STAT4 target genes in KIRC between tumor and normal samples. Inset: Gene Ontology enrichment of STAT4 target genes.

d. IRF4 expression in COAD_READ* tumor and normal samples. The p-value was calculated from t-test.

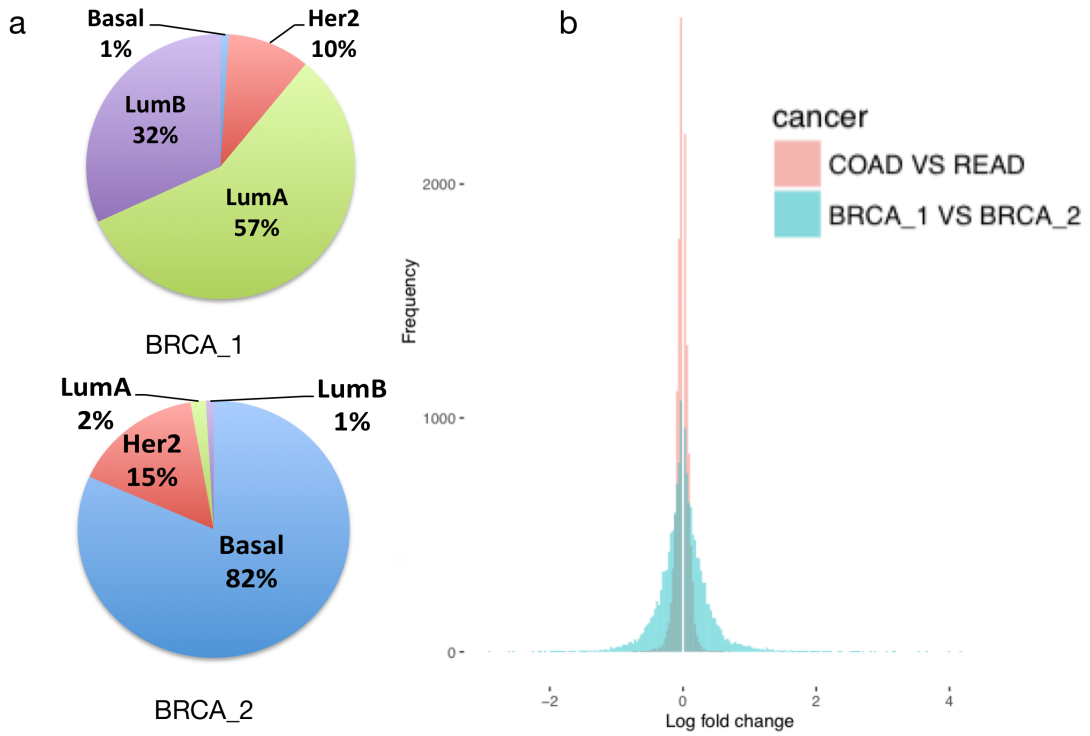
e. Survival curves from the top 25% (red) and bottom 25% (black) COAD_READ* patients according to IRF4 expression levels. The p-value was calculated using the Kaplain-Meier log-rank test.

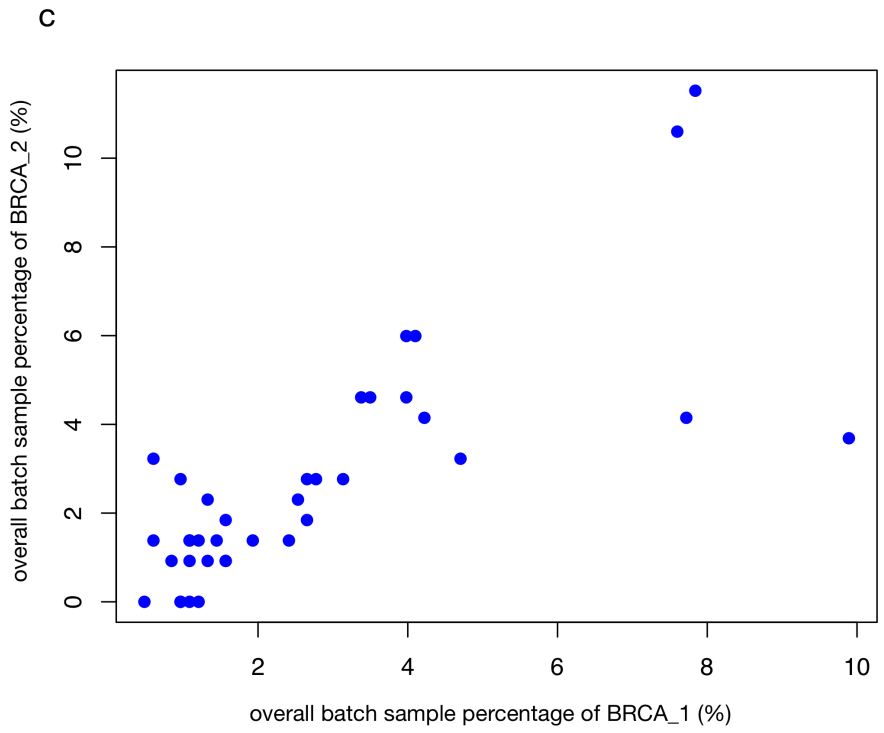
f. Waterfall plot of the differential expression of predicted IRF4 target genes in COAD_READ* between tumor and normal samples. Inset: Gene Ontology enrichment of IRF4 target genes.
g. The abundance of tumor infiltrating CD8+ T-cells in KIRC and COAD_READ primary tumors and adjacent normal tissues.
h. Scatter plot showing the STAT4 expression level and CD8+ T-cell concentration in each KIRC sample.
i. Scatter plot showing the IRF4 expression level and CD8+ T-cell concentration in each COAD_READ* sample.

Supplementary Figure S1

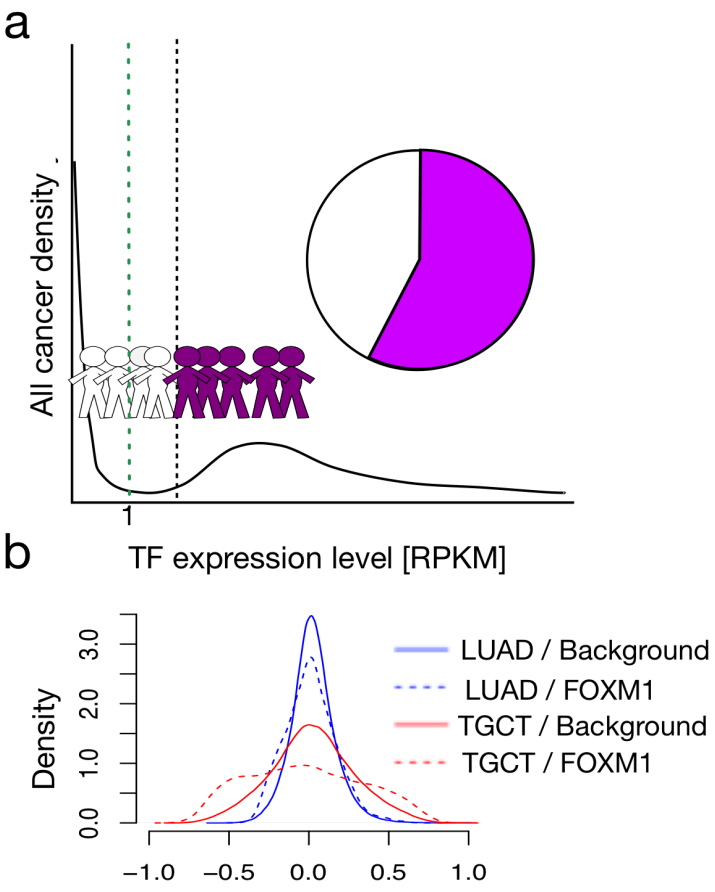


Supplementary Figure S2



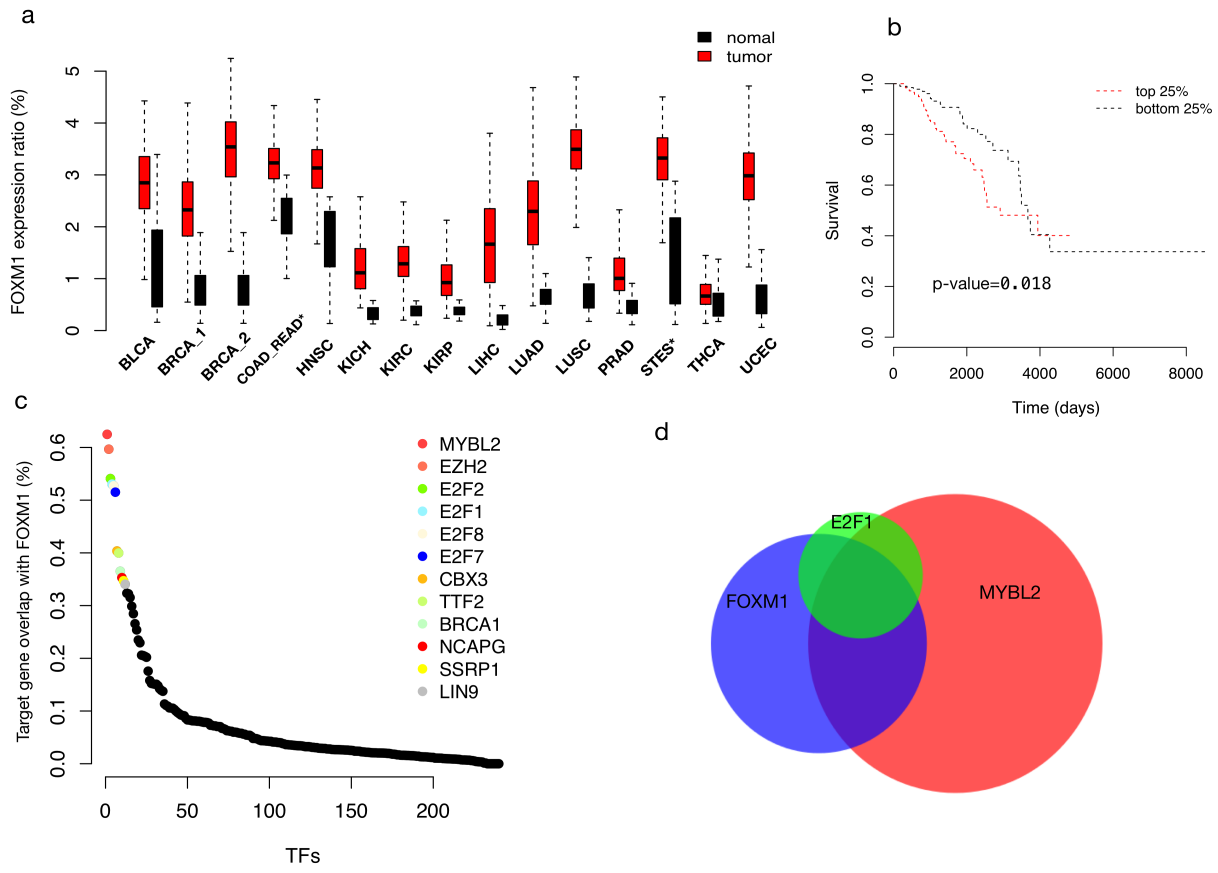


Supplementary Figure S3

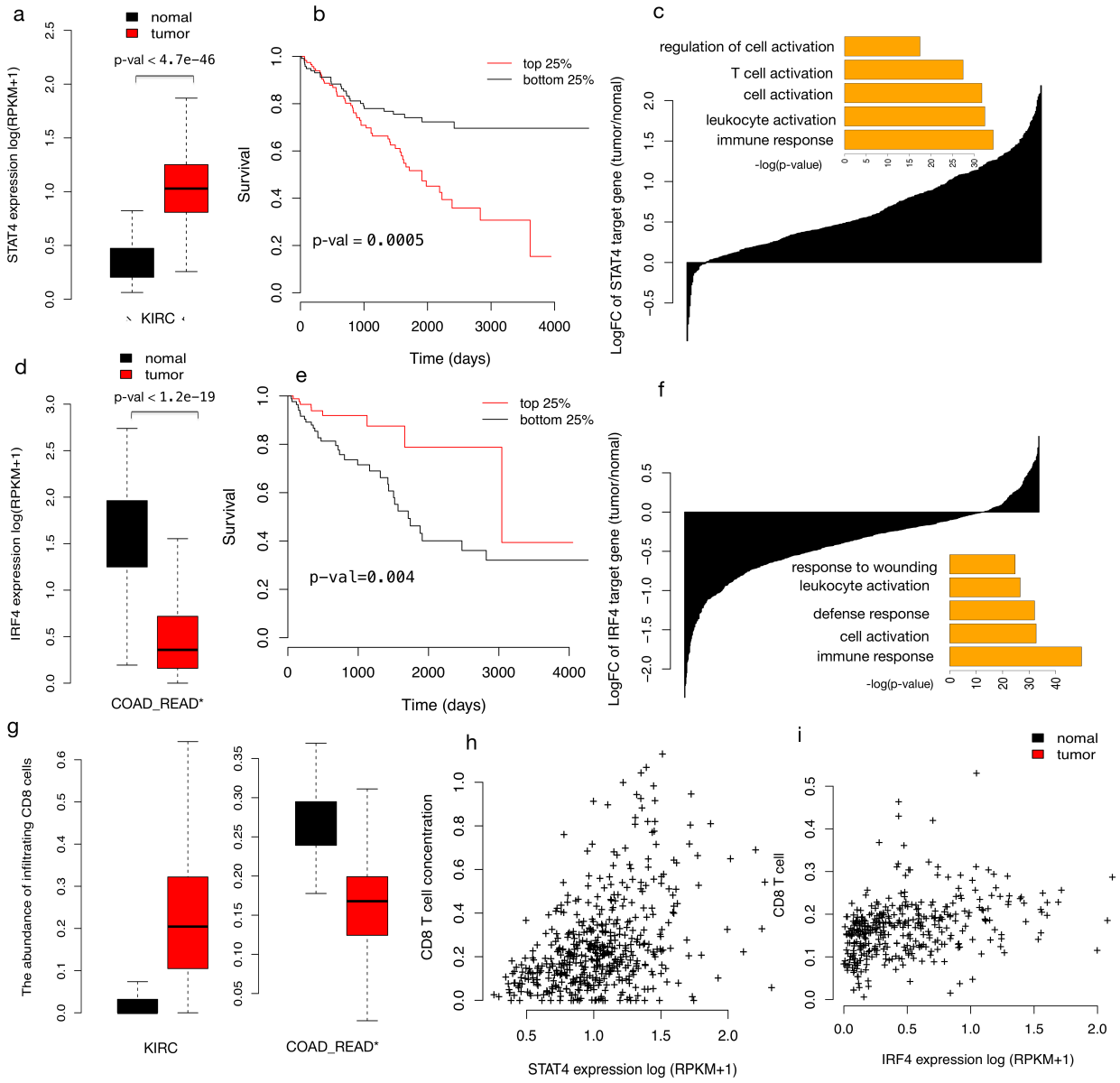


Supplementary Figure S4

Supplementary Figure S5



Supplementary Figure S6



Supplementary Table 1: Cancer Type Abbreviations

Abbreviation	Cancer Type
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA_1	Breast invasive carcinoma type 1 (luminal)
BRCA_2	Breast invasive carcinoma type 2 (basal)
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
COAD_READ	Colon and colorectal adenocarcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid leukemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STES	Stomach and esophageal carcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UVM	Uveal melanoma