ISCI, Volume *20*

Supplemental Information

Deep Learning Implicitly Handles Tissue Specific Phenomena to Predict Tumor DNA Accessibility and Immune Activity

Kamil Wnuk, Jeremi Sudol, Kevin B. Givechian, Patrick Soon-Shiong, Shahrooz Rabizadeh, Christopher Szeto, and Charles Vaske

Transparent Methods

Baseline tissue-specific dataset

All tissue-specific models described in this work were trained and evaluated following the exact procedure of the Basset network (Kelley et al., 2016), using DNase-seq peak data from 164 sample types obtained from ENCODE (Consortium, 2012) and Roadmap Epigenomics (Kundaje et al., 2015) projects.

Greedy merging of overlapping peaks across all DNase-seq data samples allowed us to create a universal set of potential accessibility sites. For each site, a binary vector was used to label its accessibility state in each of the 164 cell types. Data was then split by genomic site so that 70,000 peak locations were held out for validation, 71,886 for testing, and the remaining 1.8 million sites were used for training. The model input was a 600 base pair window of the DNA sequence centered at a site of interest, represented as one-hot encoding.

Baseline tissue-specific model implementation

We used TensorFlow (Abadi et al., 2015) to implement the Basset architecture for our baseline. We used Adam (Kingma and Ba, 2014) instead of RMSProp (Tieleman and Hinton, 2012) to optimize network parameters. We also found that use of a dynamic decay rate (that increased over the course of training) for updating moving averages in batch normalization (Ioffe and Szegedy, 2015) led to a model with competitive performance more quickly than when using a fixed decay. No other significant deviations from the original implementation were included.

When improving on the baseline model with convolutional layer factorizations, we focused experiments on factorizations that maintained the effective region of influence of the original layers and did not significantly increase the overall number of network parameters.

ENCODE DNase-seq and RNA-seq dataset

Data from the ENCODE project was initially collected at the start of 2017 for all cell or tissue types for which RNA-seq and DNase-seq measurements were both available. In order to capture a greater diversity, gene quantifications from RNA-seq files with the following ENCODE labels were collected: “RNA-seq”, “polyA mRNA”, “polyA depleted”, “single cell”. All files with “ERROR” audit flags were rejected. We kept files with “insufficient read depth,” and “insufficient read length” warnings. Despite being below ENCODE project standards, we believe the available read depths and lengths in warning situations were likely to be less of an issue when it comes to differentiating cell types (Conesa et al., 2016), and preferred to accept more potential noise in favor of a larger diversity of sample types.

The final step of data preparation involved assigning associations between specific RNA-seq and DNase-seq files within the same tissue type. In cases where there existed multiple exact matches of "biosample accession” identifiers between the two file types, associations were restricted to such exact matches. If exact accessions did not match, two file types were associated if it could be verified that they originated from the same tissue sample, cell line, or patient. This eliminated several tissue types for which no such correspondences existed. Both technical and biological replicates were treated as independent samples of the same tissue since we wanted to put the burden of learning non-invertible aspects of noise due to the measurement process on the neural network model.

The dataset was refined in late 2017, as several samples that had been part of our training and testing data were revoked by the ENCODE consortium due to quality concerns and updates. The final dataset consisted of 74 unique tissue types, distributed among partitions as discussed earlier (Table 2). The validation set was held constant, while the training and test sets included two variations.

We utilized the same greedy merge methodology described in Basset (Kelley et al., 2016) on all DNase-seq samples in our training sets to obtain a set of all potential sites of accessible DNA along the whole genome.

We used a fixed length of 600 base pairs (bp) centered at DHS peaks to define each site. Blacklisted sites at which measurements were suggested to be unreliable were excluded (Kundaje, 2016). This led to a total of 1.71 million sites of interest in the case of the held-out tissue data partition, and 1.75 million sites in the tissue overlap data partition. Using all sites across all available DNase-seq files, this produced a total 338.7 million training examples in the held-out tissue split.

As in other recent work on DNA-based prediction tasks (Alipanahi et al., 2015, Kelley et al., 2016, Singh et al., 2016, Quang and Xie, 2016) the sequence for each genomic site was obtained from human genome assembly hg19/GRCh37.

Training the expression informed model

During training data was balanced per batch due to a 14:1 ratio of negative to positive examples. Each batch sampled an equal amount of accessible and non-accessible sites without replacement, such that one pass through all available negative training examples constituted multiple randomly permuted passes through all positive training examples. In situations where a DNase-seq file had more than a single matching RNA-seq file, sites from that DNase-seq file were randomly assigned to one of the multitude of corresponding RNA-seq expression vectors each time they were selected for a training batch.

To generate a validation set that was manageable to evaluate frequently we selected 40,000 random samples from each of accessible and non-accessible sites per validation DNase-seq file. This resulted in a set of 440,000 validation examples that were used to estimate ROC AUC throughout training.

However, upon stopping we also evaluated prediction performance across whole genomes (all potential DHSs) of all validation samples (Supplemental Table S2). In cases where multiple RNA-seq file matches existed, predictions across the entire genome were evaluated once for every possible DNase-seq and RNA-seq file pair. Whole genome evaluation gave a better characterization of performance on the intended application, especially as captured by PR AUC, which is less misleading in the presence of data imbalance. Results on the test sets were evaluated across whole genomes following the same procedure.

The total number of examples (all sites across all samples) for validation was 20.5 million and 22.2 million for testing in the held-out tissue partition.

All RNA-seq expression data used to train and test models was in units of log(TPM + 1). There were many possible strategies for selecting the subset of genes for our input signature, but to initially avoid optimizing in this space, we relied on the prior work of the Library of Integrated Network-based Cellular Signatures (LINCS) and used their curated L1000 list of genes (LINCS, 2018). To ensure that the models could be applied later to cancer genomes in TCGA we converted all L1000 gene names into Ensembl gene identifiers and kept only those genes that were available in both ENCODE and TCGA TOIL RNA-seq files. After this refinement, our final input L1000 gene list consisted of 978 genes.

Expression informed model architecture and hyperparameters

We trained several alternative versions of our model and reported validation results over the course of training in Supplemental Figure S3 and Supplemental Figure S4.

The tissue-specific models demonstrated that multi-task outputs could share common convolutional layers and provide an accurate prediction of DNA accessibility across distinct sample types. Thus, we expected that if an input vector was discriminative of cell type it was likely to be sufficient to integrate it into the network after the convolutional layers. We evaluated adding a fully connected layer (depth = 500) before concatenating the vector of L1000 gene RNA-seq data to output from the convolutional layers, but found that it performed consistently worse (Supplemental Figure S3) than direct concatenation without the fully connected layer (Figure 1D).

Transfer learning consistently shortened the training time across model variants, and we found that using weights learned from the corresponding data partitions before final cleanup of revoked files was more effective on the validation set than was transfer of convolutional layer weights from the best tissue-specific model. However, our most impactful changes were increasing the batch size (from 128 to 512, and finally to 2048), and decreasing the learning rate (from 0.001 to 0.0001).

The tissue-specific models had multi-task outputs so that each training sample provided an information-rich gradient based on multiple labels for backpropagation. Since using RNA-seq inputs eliminated the need for multi-task outputs, each sample now only provided gradient feedback based on a single output. The batch size increase was intended to compensate for this change in output dimension to produce a more useful gradient for each batch.

The learning rate decrease, on the other hand, was guided by the observation that training was reaching a point of slow improvement before even a single full pass through all negative training examples. Our new dataset was also significantly larger than that used to train tissue-specific models.

We initialized our final expression-informed model (Figure 1D) with weights learned from the first iteration of the dataset, before erroneous revoked files were removed. In turn, those models were initialized with convolutional layer parameters from our best performing tissue-specific factorized convolutions model (Figure 1C). An effective batch size of 2048 was used for training (2 GPUs processing distinct batches of 1024), with an Adam (Kingma and Ba, 2014) learning rate of 0.0001 and a 0.25 fraction of positive to negative samples in every batch.

Expression informed model evaluation on ENCODE and genomic site annotations

ENCODE test set results were summarized in two ways: as a mean of AUC scores computed per whole genome sample (mean tissue type AUC in Tables 3 and 4), and as a single AUC score computed by considering predictions for all sites across all whole genome samples together (overall AUC in Tables 3 and 4). Only the latter was reported for performance analysis by genomic site type.

Two key sources were used to assign functional labels to accessibility prediction sites for performance breakdown. Exon, intragenic, and intergenic regions were derived from annotations defined by GENCODE v19 (Harrow et al., 2012). Promoter and promoter flank, and enhancer region annotations were obtained from the Ensembl Regulatory Build (Zerbino et al., 2015).

When investigating correlation of training similarity to test sample performance, since the modulating factor between predictions applied to different tissues is the input RNA-seq data, distance between test samples, t, and the training set, T, was computed as $\mathrm{d}(t,T) = \min_{i \in T} \|\boldsymbol{r}_i - \boldsymbol{r}_t\|$, where $\boldsymbol{r}_t$ is a test sample's vector of $\log_2$(TPM + 1) expression levels for all L1000 genes.

Predicting DNA accessibility in TCGA

We applied our best expression informed model trained on the held-out tissue ENCODE partition to predict accessibility in TCGA. We restricted our predictions to promoter and promoter flank sites, since performance at those sites was high across all tests.

TOIL RNA-seq transcripts per million (TPM) gene expression data was used to obtain L1000 input gene signatures for all processed TCGA samples (Vivian et al., 2017, TOIL RNA-seq recompute, 2016). All expression values were converted from $\log_2$(TPM + 0.001) to $\log_2$(TPM + 1) before use.

For landscape views of accessibility and mutation impact analysis (Figure 3) we considered only samples with WGS available, and used mutation calls from an internal tool. For each sample site affected by at least one mutation, the change in predicted accessibility was computed before and after each mutation was applied, independently for SNPs and INDELs (Figure 3B). In order to apply t-SNE to generate the per-site landscape view (Figure 3D) we represented each site by a vector of binary accessibility decisions at that position across all selected TCGA samples with all mutations applied. All mutations were also applied when generating the per-patient t-SNE visualization (Figure 3F).

Accessibility in LUAD

To assess the uniqueness in perspective of accessibility versus RNA-seq, all LUAD samples for which we had WGS data were clustered into two groups via K-means. For this, four data sources were used: accessibility predictions for promoter and promoter flank sites, TOIL $\log_2$(TPM + 1) RNA-seq gene expression data, HiSeqV2 $\log_2$(normalized count + 1) RNA-seq gene expression data (TCGA Genome

Characterization Center, 2017) and TOIL $\log_2$(TPM + 1) RNA-seq gene expression data for all genes in the L1000 gene set used as inputs to our expression informed model. For the first three datatypes, we clustered samples based on the top 5% most variable sites (for accessibility) or genes (for TOIL and HiSeqV2) across the LUAD cohort, following the logic that the most highly variable sites may highlight the most dramatic differentially active pathways. For the L1000 genes, clustering was based on the entire set of gene expression levels. To show the difference quantitatively between cluster assignments across data types we used adjusted mutual information (Vinh et al., 2010) (Figure 4B).

Exploration of pathway enrichment between the accessibility clusters was performed using Enrichr (Kuleshov et al., 2016). Genes for enrichment analysis were selected by first eliminating all genes below a standard deviation threshold of 0.33 (in TOIL data) across the LUAD cohort (in HiSeqV2 data the equivalently selected standard deviation threshold was 1.0). This threshold was selected to include the main peak of gene standard deviation and exclude the peak around zero (Supplemental Figure S5), comprised of genes with little change or very low levels of expression. All remaining genes were then compared with a two-sided t-test between the two clusters and p-values were adjusted with Benjamini Hochberg (BH) correction. Due to the low number of WGS samples in either cluster (21 samples in C0 and 20 samples in C1) a more permissive false discovery rate of 0.25 was chosen as the cutoff for differential expression. In TOIL data, this procedure returned 512 genes upregulated in C0 and 857 genes upregulated in C1. In HiSeqV2 data, the same process yielded 344 upregulated genes in C0 and 339 in C1.

For comparison of tumor mutational burden (TMB) across clusters, TMB was computed as the total count of missense and nonsense mutations in each WGS sample.

When the patient analysis set was expanded to include all LUAD samples without WGS mutation information, clustering based on promoter and promoter flank accessibility predictions was repeated with the same procedure as before (Figure 4D).

To investigate whether the accessibility space appeared continuous along the dimensions of most variance across LUAD we used Principal Component Analysis (PCA) applied to all promoter and flank accessibility predictions to project each sample onto the first three principal components (Figure 4F).

Correlating accessibility count with gene expression

Total accessibility count used to investigate gene correlations was computed as the total number of promoter and promoter flank sites predicted to be accessible after applying the binary decision threshold (at 80% precision) defined on ENCODE data. Again, only genes whose standard deviation was above 0.33 were considered for correlation analysis. Both Pearson and Spearman measures were evaluated, and the threshold for both measures was an absolute value above 0.4. All genes satisfying the threshold were analyzed for KEGG pathway enrichment with Enrichr (666 genes for Pearson correlation, and 418 genes for Spearman) (Supplemental Table S4 and S5)

Accessibility analysis in immune cell driven clusters

LUAD samples were clustered into two groups using K-means on vectors of lymphoid (21 cell types) and myeloid (13 cell types) xCell estimates (Aran et al., 2017), revealing a survival difference (Supplemental Figure S6C). We noticed that a plane orthogonal to the first principal component (PC1) partitioned cluster labels when xCell vectors were reduced to three dimensions with PCA (Supplemental Figure S6A and B). To exclude cases of near ambiguous label assignment and focus on more prominent differences we removed samples within a small margin at the midpoint between clusters (in PC1). Margin size was equal to the standard deviation of the smallest cluster in the PC1 dimension (Supplemental Figure S6D and E). After ignoring margin samples, the survival difference of patients between clusters increased in significance (logrank test $p = 6.7e-4$) (Supplemental Figure S6F).

Total methylation for all LUAD samples was computed as the sum of values at all sites measured by the Infinium HumanMethylation450 BeadChip, available from TCGA (TCGA, 2016). Total accessible site

count considered all promoter and promoter flank sites, with binary class assignment based on the 80% precision threshold (Figure 5B).

For further analysis of accessibility, only sites previously determined as facultative were considered and all with low standard deviation (< 0.135) across LUAD (N = 512) were eliminated, to ignore cohort specific constitutive sites with some tolerance for noise. The threshold was selected so that at minimum 10 accessibility values at a site had to be distinct from the site's values across the whole cohort (Supplemental Figure S7A).

Each accessibility prediction site was assigned to its nearest gene, according to distance in base pairs, as defined by GENCODE v19 (Harrow et al., 2012). We considered only accessibility sites within 50,000 base pairs as having a valid correspondence to a gene (Supplemental Figure S7B). Significantly differentiated accessibility sites were then used to vote for candidate upregulated genes in each cluster. A very conservative significance threshold (two sided t-test BH adj. $p < 1.0e-5$) was selected so as to only focus on the most striking accessibility differences. Each site was allowed to contribute a single vote to its corresponding gene according to the cluster in which the site was more accessible.

Genes with a consistent direction of upregulation votes were considered cluster-specific candidate genes to test for differential expression (532 genes in X0 and 2250 genes in X1). From the candidate genes for each cluster that also had significant (two sided t-test BH adj. $p < 0.01$) differential expression (190 in X0 and 835 in X1) we identified the group in each cluster that was consistent (123 in X0, and 536 in X1) and inconsistent (67 in X0, and 299 in X1) with the direction of increased accessibility. All four sets were then tested for KEGG pathway enrichment via Enrichr (Supplemental Table S6 and S7).

Predicting immune state from promoter and promoter flank accessibility

To train an ensemble of distinct models to discriminate immune hot from immune cold we used three fold cross validation; independently partitioning hot (X0 from immune cell based clustering) and cold (X1 from immune cell clustering) samples randomly to maintain an equal ratio across each fold. Training on

different random subsets of data enhanced robustness when dealing with training label uncertainty. Each classifier was an RBF kernel SVM with $C = 3.5$, and $\gamma = \frac{1}{N\sigma}$, where N was the number of features and σ was the standard deviation of feature values across the training set. Additionally, training samples were balanced by weights inversely proportional to class frequency, and Platt scaling (Platt, 1999) was used to obtain probability estimates from SVM classification. During ensemble classifier application we excluded all samples that did not have a mean probability of at least 0.5 for the ensemble's majority class prediction.

Input features to the classifier were binary accessibility predictions for a set of 484 sites comprised from the union of all immune hot (X0) sites consistent with gene expression and immune cold (X1) sites inconsistent with expression, as obtained from analysis of the xCell driven LUAD clusters. These sites were chosen both for their association with significant differences in expression of corresponding genes and the enrichment of those gene sets for immune relevant pathways.

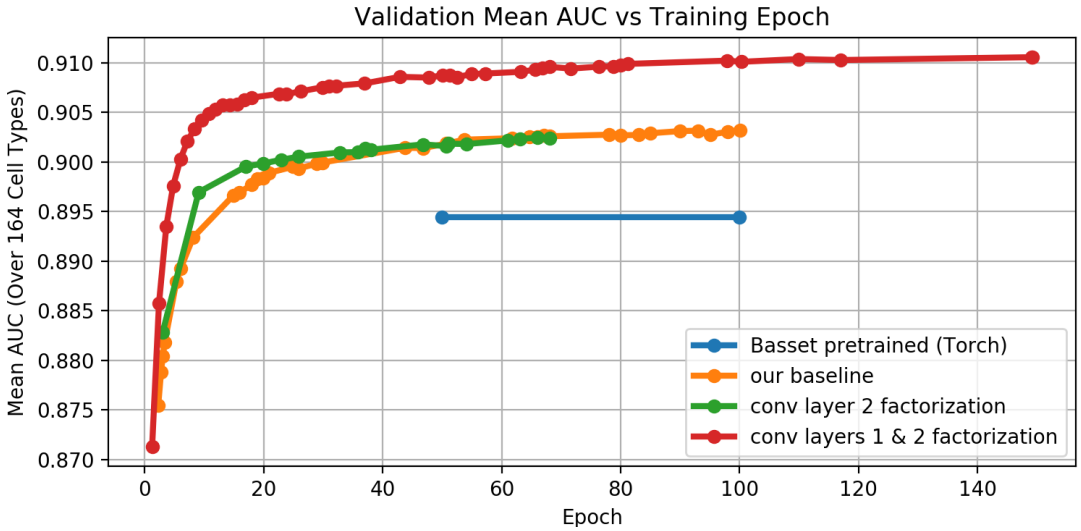
Expanding the application domain of the immune activity classifier to previously unprocessed TCGA cohorts involved first applying our expression informed convolutional neural network model to all promoter and promoter flank sites in the new data. As previously, when expanding our LUAD sample size to non-WGS data, we used only the reference genome (hg19/GRCh37) and TOIL $\log_2(\text{TPM} + 1)$ RNA-seq gene expression data for all predictions. Predictions that incorporated mutation information were included only for samples in our original six cohorts for which WGS was available.

Validating TCGA predictions with measured ATAC-seq peaks

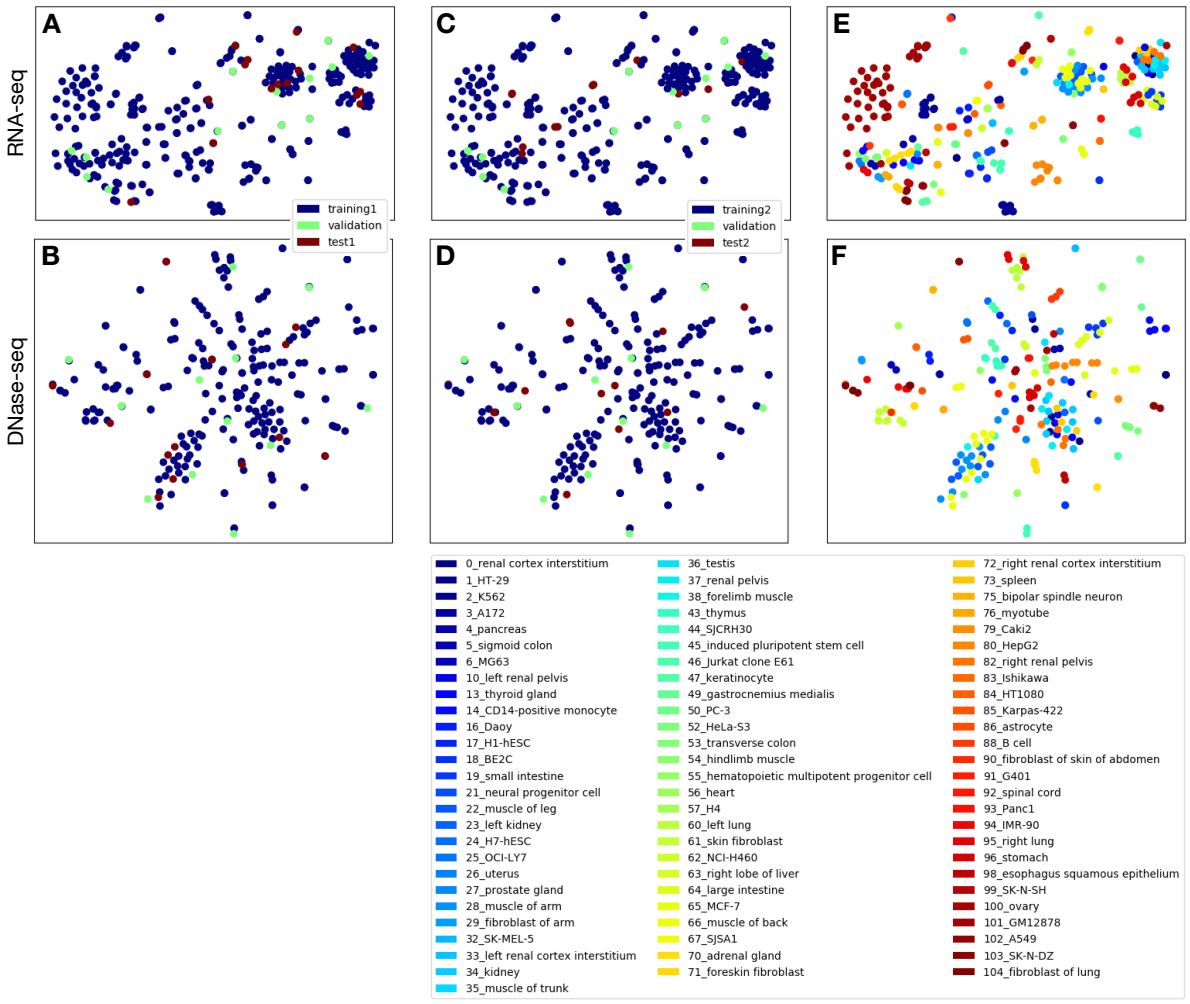
The list of pan-cancer peaks, TCGA sample identifiers, and the normalized ATAC-seq insertion counts within the pan-cancer peak set were obtained from supplemental material of the empirical investigation of chromatin accessibility in TCGA (Corces et al., 2018) at: https://gdc.cancer.gov/about-data/publications/ATACseq-AWG . To visualize distributions of ATAC-seq peaks for our clustering-based constitutive and facultative site labels, we computed the mean values within KIRC, KIRP, LUAD,

and LUSC cohorts individually for each pan-cancer peak that corresponded uniquely to our promoter and promoter flank sites with at least 70% overlap (Figure 7A). This was not restricted to TCGA samples for which we had also made predictions. But when looking at normalized count distributions for prediction decisions made by our accessibility classifier (Figure 7B,C) we did restrict analysis only to samples where both ATAC-seq was performed and our predictions were available. For every TCGA sample, each accessibility prediction site with a matching pan-cancer ATAC-seq peak contributed a single datum to the distribution plots. So every matched TCGA sample whose numbers are listed in Figure 7B,C contributed 61,342 data points. We validated matched samples for all cohorts to which we had previously applied our immune activity classifier with the exception of SARC, for which no ATAC-seq measurements were available, and GBM, for which none of the TCGA samples measured matched those for which we had run predictions. As in all previous TCGA analyses the classification decision threshold for binarizing accessibility predictions was based on an 80% precision (20% false discovery rate) threshold on the ENCODE held out tissue test set.

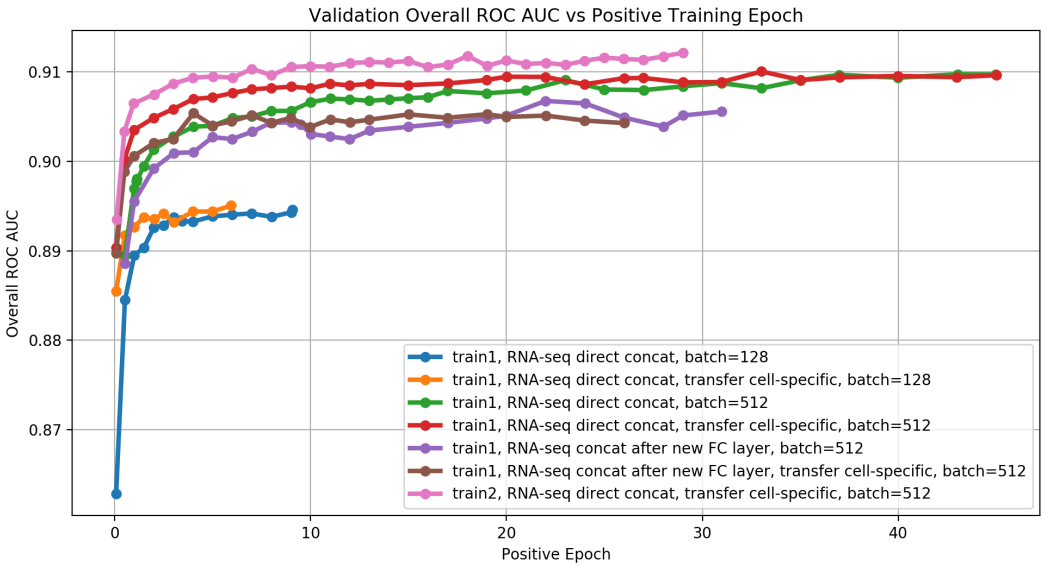
Supplemental Figures



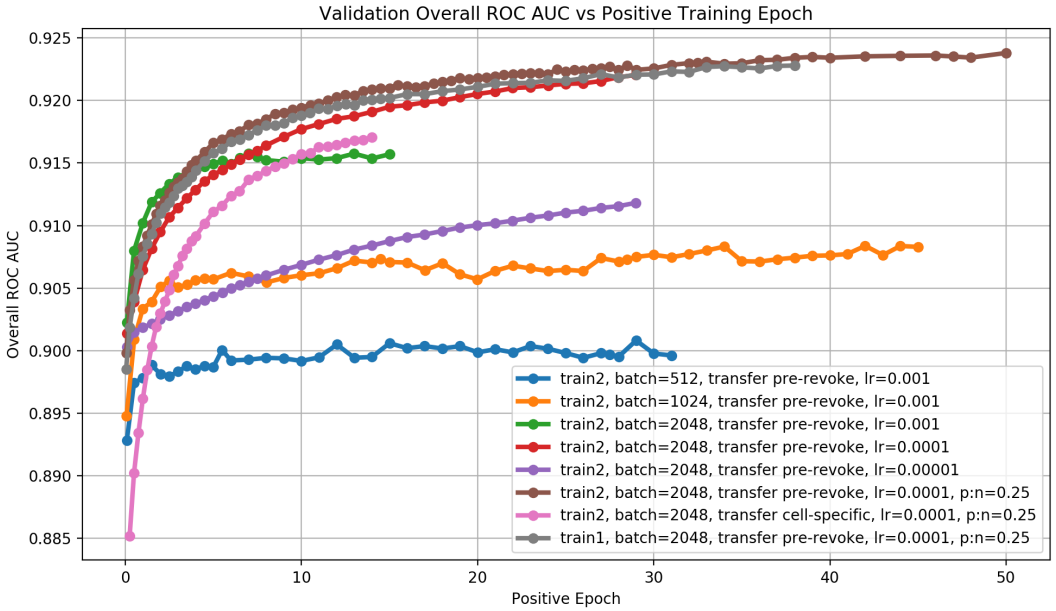
Supplemental Figure S1. Training of tissue-type-specific model architectures. Related to Figure 1 and Table 1. The mean ROC AUC across 164 cell types in the validation set versus training epoch is shown. The result obtained by the pre-trained model provided by the authors of Basset is shown for reference, but since the number of training epochs was not reported, an arbitrary range was selected for display. We explored independent factorization of the second convolutional layer of the baseline model, and achieved the best performance when both the first and second convolutional layers were factorized.



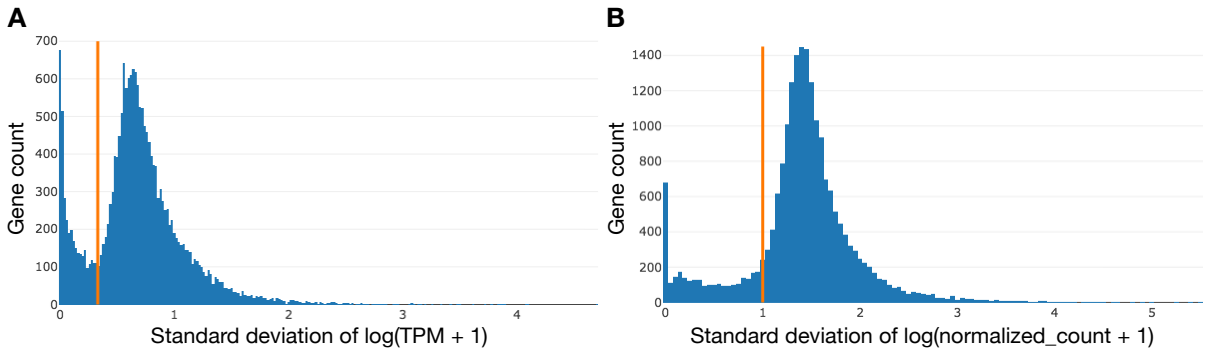
Supplemental Figure S2. t-SNE embedding of ENCODE dataset partitioning in RNA-seq and DNase-seq space. Related to Tables 2, 3, 4, and Figure 2. Sample distribution is illustrated by t-SNE embedding of the tissue overlap data partitions (training1 and test1) based on (A) RNA-seq log2(TPM + 1) expression data and (B) DNase-seq peaks, as well as the held-out tissues data partitions (training2 and test2) based on (C) RNA-seq and (D) DNase-seq. The original ENCODE sample type labels are also shown for t-SNE embedded (E) RNA-seq and (F) DNase-seq samples, illustrating that samples of similar tissue or function often appear in proximity to each other across both data types.



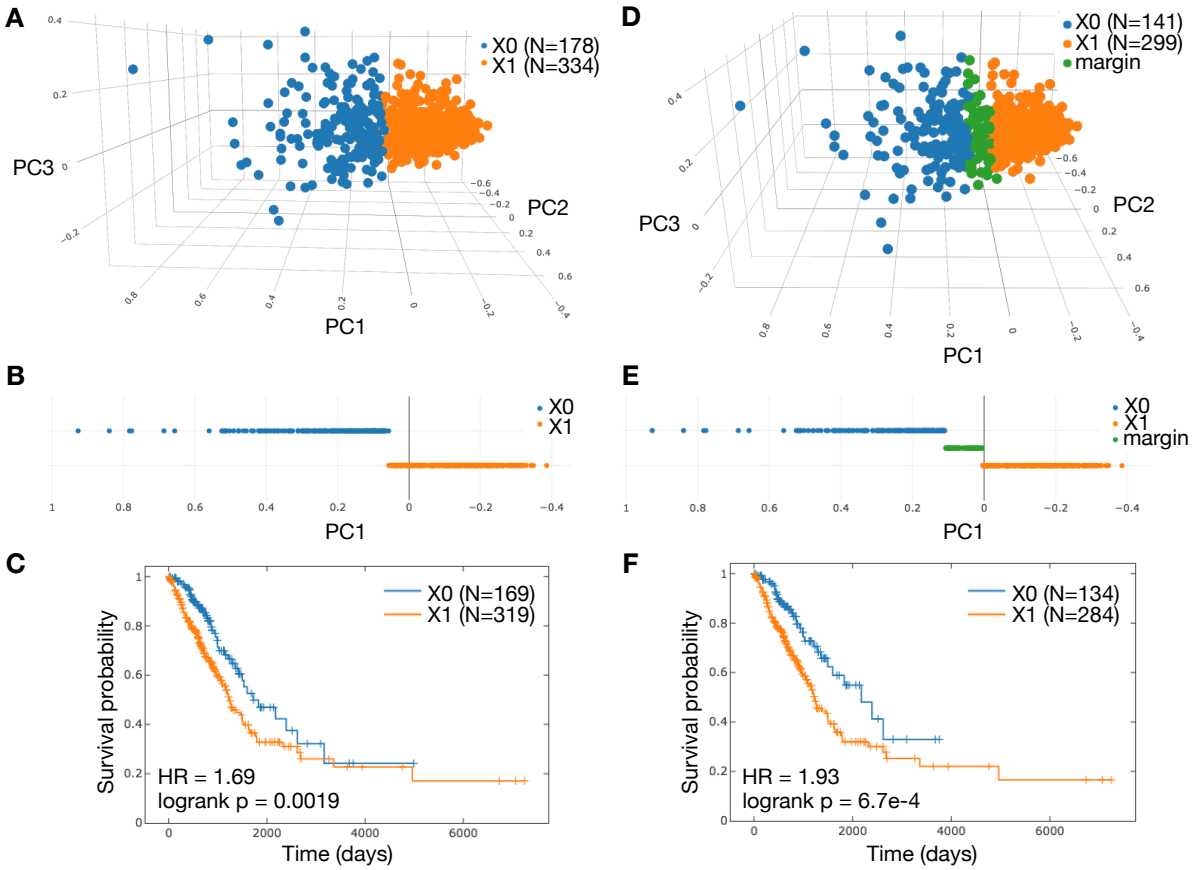
Supplemental Figure S3. Overall ROC AUC for the small validation set. Related to Figures 1, 2, and Tables 3, 4. The ROC AUC for the small validation set over number of passes through all positive examples (positive epochs) for several expression informed model architectures is shown. We experimented with adding a fully connected (FC) layer of depth 500 before concatenating (concat) gene expressions with outputs from the convolutional (conv) layers. However, increasing the batch size and initializing the convolutional layers with weights from our final tissue-specific model (transfer) improved performance most. Models trained on the tissue overlap set (train1) showed similar validation performance as those trained on the held-out tissue set (train2) with the same hyperparameters. This evaluation was done before the final dataset revision which revoked several suspected low quality samples, yet still provided valuable feedback for model selection.



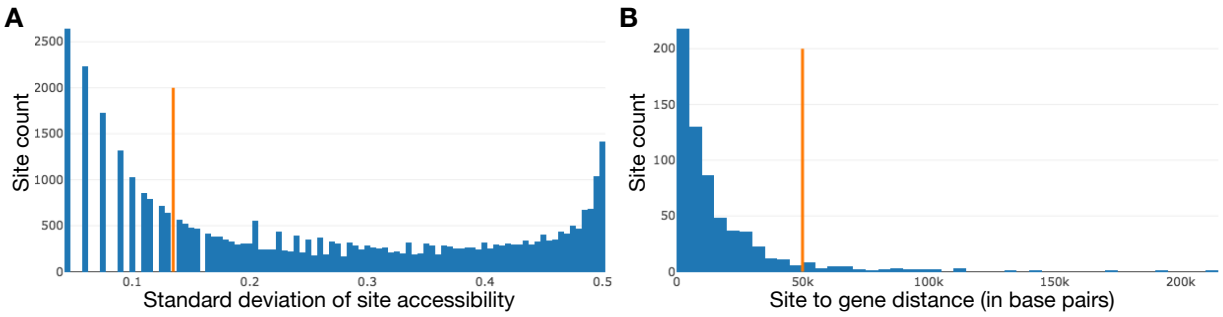
Supplemental Figure S4. Overall ROC AUC for the small validation set over positive training epochs for models trained after the final dataset revision. Related to Figures 1, 2, and Tables 3, 4. A further increase in batch size as well as a decreased learning rate (lr) led to additional significant improvements. Changing the fraction of positive samples per training batch (from p:n=0.5 to p:n=0.25) also slightly improved both ROC AUC as well as PR AUC in whole genome validation. Transfer of weights learned before final revoking of data (Figure S3) was a more effective initialization than weight transfer from our final tissue-specific model. Finally, we again confirmed that the same hyperparameters led to good validation performance across both training partitions: tissue overlap (train1) and held-out tissue (train2).



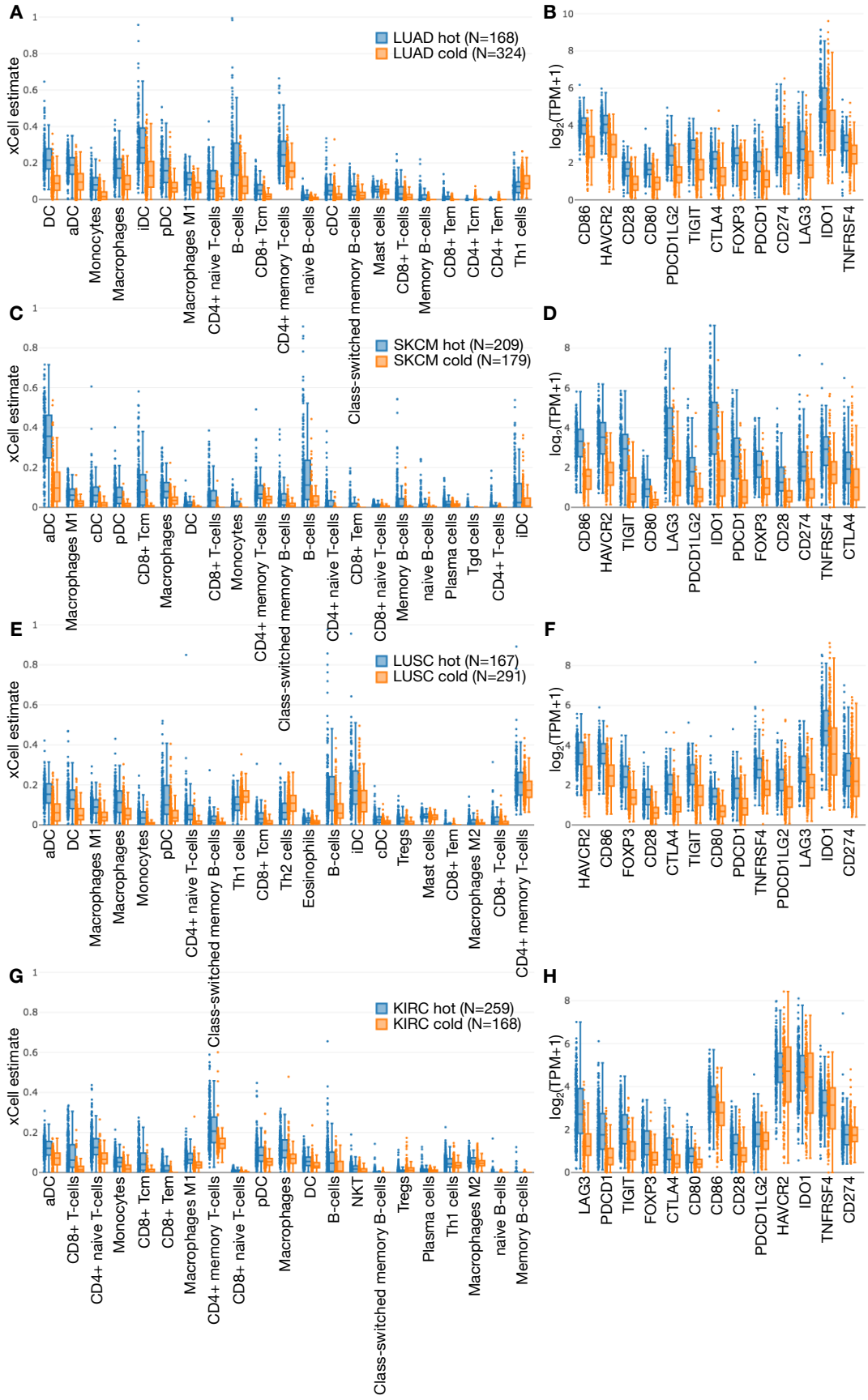
Supplemental Figure S5. Histograms of gene expression standard deviations. Related to Figure 4. (A) Histogram of gene expression standard deviations across LUAD WGS samples in TOIL RNA-seq log2(TPM + 1) data along with the selected threshold (0.33) in orange is shown. (B) Gene expression standard deviations across the same samples in HiSeqV2 log2(normalized count+1) data and the selected threshold (1.0) in orange (B) are also shown. In both cases the thresholds eliminate genes with little change across samples or very low levels of expression, and keep all genes that constitute the main peak of values.



Supplemental Figure S6. Adding a margin between immune cell based clusters in LUAD samples. Related to Figure 5. (A) All LUAD samples are shown with respect to the first three principal components (PC1-3) of their lymphoid and myeloid xCell estimates, colored according to labels assigned from k-means clustering. (B) Plotting the labeled data according to only the first principal component clearly shows the location of a separating plane between the clusters. The points excluded by introducing a margin at the scale of the smaller cluster's standard deviation along PC1 are shown (D) and (E). The impact on survival between X0 and X1 is shown by Kaplan-Meier plots before (C) and after (F) the margin was introduced. Kaplan-Meier plots are annotated with group size (N), logrank test p-values and hazard ratio (HR) based on a Cox proportional hazards (CoxPH) model regression using class assignment as the only explanatory variable.



Supplemental Figure S7. Histograms of accessibility and site to gene distance standard deviations. Related to Figure 5. (A) A histogram of the standard deviation of accessibility classifications in LUAD of promoter and promoter flank sites previously identified as facultative (40,823 sites) based on t-SNE across our initial set of six TCGA cohorts is shown. The threshold (< 0.135), in orange, identifies a subset of 25,093 sites facultative in LUAD. (B) The second histogram shows the site to nearest gene distances for all accessibility sites that also satisfy BH adjusted $p < 1.0e\text{-}5$ from a two-sided t-test between the LUAD immune cell driven clusters (3246 sites). Only sites within 50k base pairs (orange) were considered when voting for gene accessibility.


A
LUAD hot (N=168)
LUAD cold (N=324)
xCell estimate
DC
aDC
Monocytes
Macrophages
iDC
pDC
Macrophages M1
CD4+ naive T-cells
B-cells
CD8+ Tcm
CD4+ memory T-cells
naive B-cells
cDC
Class-switched memory B-cells
Mast cells
CD8+ T-cells
Memory B-cells
CD8+ Tem
CD4+ Tcm
CD4+ Tem
Th1 cells
B
log2(TPM+1)
CD86
HAVCR2
CD28
CD80
PDCD1LG2
TIGIT
CTLA4
FOXP3
PDCD1
CD274
LAG3
IDO1
TNFRSF4
C
SKCM hot (N=209)
SKCM cold (N=179)
xCell estimate
aDC
Macrophages M1
cDC
pDC
CD8+ Tcm
Macrophages
DC
CD8+ T-cells
Monocytes
CD4+ memory T-cells
Class-switched memory B-cells
B-cells
CD4+ naive T-cells
CD8+ Tem
CD8+ naive T-cells
Memory B-cells
naive B-cells
Plasma cells
Tgd cells
CD4+ T-cells
iDC
D
log2(TPM+1)
CD86
HAVCR2
TIGIT
CD80
LAG3
PDCD1LG2
IDO1
PDCD1
FOXP3
CD28
CD274
TNFRSF4
CTLA4
E
LUSC hot (N=167)
LUSC cold (N=291)
xCell estimate
aDC
DC
Macrophages M1
Macrophages
Monocytes
pDC
CD4+ naive T-cells
Class-switched memory B-cells
Th1 cells
CD8+ Tcm
Th2 cells
Eosinophils
B-cells
iDC
cDC
Tregs
Mast cells
CD8+ Tem
Macrophages M2
CD8+ T-cells
CD4+ memory T-cells
F
log2(TPM+1)
HAVCR2
CD86
FOXP3
CD28
CTLA4
TIGIT
CD80
PDCD1
TNFRSF4
PDCD1LG2
LAG3
IDO1
CD274
G
KIRC hot (N=259)
KIRC cold (N=168)
xCell estimate
aDC
CD8+ T-cells
CD4+ naive T-cells
Monocytes
CD8+ Tcm
CD8+ Tem
Macrophages M1
CD4+ memory T-cells
CD8+ naive T-cells
pDC
Macrophages
DC
B-cells
NKT
Class-switched memory B-cells
Tregs
Plasma cells
Th1 cells
Macrophages M2
naive B-cells
Memory B-cells
H
log2(TPM+1)
LAG3
PDCD1
TIGIT
FOXP3
CTLA4
CD80
CD86
CD28
PDCD1LG2
HAVCR2
IDO1
TNFRSF4
CD274

Supplemental Figure S8. Example immune cell and checkpoint gene distributions across predicted hot and cold tumors. Related to Figure 6. Examples of xCell estimates and checkpoint gene expression levels compared across multiple cohorts for tumors classified by our 3 SVM ensemble as hot or cold, with group size shown (N). All x-axis labels are ordered by significance based on a two-sided t-test between tumors classified as hot and cold. Only the top 21 most significant xCell estimates from lymphoid and myeloid cell categories are shown. (A) The LUAD xCell estimate and (B) checkpoint gene distributions demonstrate how application of the classifier affected significance ordering from the raw training data illustrated in Figure 5. (C,D) SKCM provides an example of a distinct cohort in which immune hot patients exhibited longer survival. (E,F) LUSC demonstrates one case where immune activity appears to have no effect. (G,H) Finally, KIRC shows a case where immune active patients have significantly worse survival, and also exhibits a curious lack of difference between levels of CD274 (also called PD-L1) and PDCD1LG2 (also called PD-L2) compared to other cohorts split by our accessibility based immune activity classifier.

Supplemental Tables

Supplemental Table S1. Number of DNase-seq files by tissue name per dataset partition, including tissue overlap (TO) and held-out tissue (HOT) sets. Related to Tables 2, 3, 4, and Figure 2.

ENCODE Tissue Name	Validation	Train (TO)	Test (TO)	Train (HOT)	Test (HOT)
renal cortex interstitium	0	3	0	3	0
HT-29	0	2	0	2	0
K562	0	2	0	2	0
A172	0	1	1	2	0
pancreas	0	1	0	1	0
MG63	0	2	0	2	0
left renal pelvis	0	3	1	4	0
thyroid gland	0	6	0	6	0
CD14-positive monocyte	1	1	0	1	0
Daoy	0	2	0	2	0
H1-hESC	0	2	0	2	0
BE2C	0	2	0	2	0
small intestine	0	3	2	5	0
neural progenitor cell	0	2	0	2	0
muscle of leg	0	7	0	7	0
left kidney	0	1	0	0	1
H7-hESC	0	3	0	3	0
OCI-LY7	0	2	0	0	2
uterus	0	1	0	1	0
prostate gland	0	1	0	0	1
muscle of arm	2	7	1	8	0
fibroblast of arm	1	1	0	1	0
SK-MEL-5	0	2	0	2	0
left renal cortex interstitium	0	4	0	4	0
kidney	0	4	0	4	0
muscle of trunk	0	1	0	1	0
testis	0	2	0	2	0
renal pelvis	0	3	0	3	0
forelimb muscle	0	0	1	1	0
thymus	0	6	0	6	0
SJCRH30	1	2	0	2	0
induced pluripotent stem cell	0	2	0	2	0
Jurkat clone E61	0	2	0	2	0
keratinocyte	0	1	1	2	0
PC-3	0	2	0	2	0
HeLa-S3	0	4	0	4	0
hindlimb muscle	0	1	0	0	1
hematopoietic multipotent progenitor cell	0	1	0	1	0
heart	0	2	0	2	0
H4	0	2	0	2	0
left lung	0	6	0	6	0
skin fibroblast	0	8	1	9	0
NCI-H460	0	2	0	2	0
large intestine	0	5	1	6	0
MCF-7	0	4	0	4	0
muscle of back	0	8	2	10	0
SJSA1	0	2	0	2	0
adrenal gland	0	5	1	6	0
foreskin fibroblast	1	1	0	1	0
right renal cortex interstitium	0	3	0	3	0
spleen	0	1	0	0	1
bipolar spindle neuron	0	2	0	2	0
myotube	0	2	0	2	0
Caki2	0	2	0	2	0
HepG2	0	4	0	4	0
right renal pelvis	1	3	0	3	0
Ishikawa	0	1	0	1	0
HT1080	0	2	0	2	0
Karpas-422	0	2	0	2	0
astrocyte	0	2	0	0	2
B cell	0	4	0	4	0
fibroblast of skin of abdomen	0	1	0	0	1
G401	0	2	0	0	2
spinal cord	1	1	0	1	0
Panc1	1	1	0	1	0
IMR-90	0	4	0	4	0
right lung	1	3	0	3	0
stomach	0	5	0	5	0
SK-N-SH	0	2	0	2	0
ovary	0	1	0	1	0
GM12878	0	2	0	2	0
A549	1	2	0	2	0
SK-N-DZ	0	1	1	2	0
fibroblast of lung	0	5	1	6	0

Supplemental Table S2. Whole genome validation results for our expression-informed model trained on tissue overlap (TO) and held-out tissue (HOT) sets. Related to Tables 2, 3, and 4.

Sample tissue type	ROC AUC (TO)	PR AUC (TO)	ROC AUC (HOT)	PR AUC (HOT)
CD14-positive monocyte	0.888	0.559	0.889	0.563
muscle of arm	0.774, 0.959	0.654, 0.808	0.783, 0.960	0.671, 0.811
fibroblast of arm	0.898, 0.900	0.806, 0.809	0.898, 0.900	0.808, 0.811
SJCRH30	0.875	0.727	0.875	0.730
foreskin fibroblast	0.953	0.774	0.953	0.771
right renal pelvis	0.967	0.833	0.968	0.836
spinal cord	0.947	0.714	0.946	0.713
Panc1	0.957	0.713	0.958	0.711
right lung	0.958	0.781	0.958	0.782
A549	0.902	0.735	0.900	0.734
mean tissue type AUC	0.915	0.743	0.916	0.745
overall AUC	**0.912**	**0.721**	**0.913**	**0.725**

Supplemental Table S3. Enhancer results across held-out tissue test set whole genomes. Related to Table 6 and Figure 2.

Sample tissue type	ROC AUC	PR AUC
left kidney	0.934	0.737
OCI-LY7	0.845, 0.845, 0.850, 0.850	0.645, 0.643, 0.606, 0.605
prostate gland	0.817	0.490
hindlimb muscle	0.933	0.908
spleen	0.809	0.471
astrocyte	0.931, 0.898	0.967, 0.833
fibroblast of skin of abdomen	0.953	0.940
G401	0.640, 0.694	0.432, 0.270
overall AUC	**0.870**	**0.732**

Supplemental Table S4. Pathway enrichment (Enrichr) results with adjusted p < 1.0e-4 for all 418 genes correlated with total promoter and promoter flank accessibility in LUAD with |Spearman correlation| > 0.4. Related to Figure 4.

KEGG pathway	p	Adj. p	Z-score	Combined score
Osteoclast differentiation (hsa04380)	3.74e-17	7.45e-15	-1.86	70.51
TNF signaling pathway (hsa04668)	1.31e-10	1.30e-8	-1.89	42.96
Amoebiasis (hsa05146)	2.48e-9	1.64e-7	-1.82	36.03
Pathways in cancer (hsa05200)	2.33e-8	6.50e-7	-1.95	34.27
Tuberculosis (hsa05152)	6.68e-9	2.66e-7	-1.71	32.13
Pertussis (hsa05133)	4.72e-9	2.35e-7	-1.64	31.46
Regulation of actin cytoskeleton (hsa04810)	2.61e-8	6.50e-7	-1.75	30.52
NF-kappa B signaling pathway (hsa04064)	8.08e-9	2.68e-7	-1.62	30.28
Epstein-Barr virus infection (hsa05169)	1.28e-6	2.83e-5	-1.75	23.75
PI3K-Akt signaling pathway (hsa04151)	3.31e-6	5.21e-5	-1.78	22.48
Chemokine signaling pathway (hsa04062)	2.11e-6	4.01e-5	-1.69	22.15
Influenza A (hsa05164)	4.29e-6	6.10e-5	-1.63	20.18
Cytokine-cytokine receptor interaction (hsa04060)	3.40e-6	5.21e-5	-1.59	20.00
AGE-RAGE signaling pathway in diabetic complications (hsa04933)	8.58e-6	1.00e-4	-1.60	18.72
Jak-STAT signaling pathway (hsa04630)	6.09e-6	7.58e-5	-1.55	18.61
Staphylococcus aureus infection (hsa05150)	2.22e-6	4.01e-5	-1.43	18.56
Measles (hsa05162)	5.70e-6	7.56e-5	-1.50	18.17

Supplemental Table S5. Pathway enrichment (Enrichr) results with adjusted p < 1.0e-6 for all 666 genes correlated with total promoter and promoter flank accessibility in LUAD with |Pearson correlation| > 0.4. Related to Figure 4.

KEGG pathway	p	Adj. p	Z-score	Combined score
Osteoclast differentiation (hsa04380)	4.35e-22	9.66e-20	-1.86	91.69
Influenza A (hsa05164)	8.68e-13	9.63e-11	-1.94	53.97
TNF signaling pathway (hsa04668)	1.37e-12	1.01e-10	-1.86	50.83
Regulation of actin cytoskeleton (hsa04810)	5.54e-12	3.08e-10	-1.85	47.86
Hepatitis B (hsa05161)	9.68e-11	4.30e-9	-1.83	42.30
HTLV-I infection (hsa05166)	1.61e-10	5.95e-9	-1.82	41.00
Jak-STAT signaling pathway (hsa04630)	5.20e-10	1.44e-8	-1.75	37.47
Pathways in cancer (hsa05200)	1.76e-9	3.26e-8	-1.82	36.61
Chemokine signaling pathway (hsa04062)	7.17e-10	1.77e-8	-1.72	36.23
Epstein-Barr virus infection (hsa05169)	8.26e-10	1.83e-8	-1.73	36.09
Leukocyte transendothelial migration (hsa04670)	3.16e-10	1.00e-8	-1.57	34.40
Tuberculosis (hsa05152)	1.23e-9	2.48e-8	-1.56	32.02
Acute myeloid leukemia (hsa05221)	3.69e-9	6.29e-8	-1.55	30.04
Measles (hsa05162)	4.63e-9	7.15e-8	-1.53	29.37
Amoebiasis (hsa05146)	4.83e-9	7.15e-8	-1.48	28.35
Viral carcinogenesis (hsa05203)	2.30e-8	2.83e-7	-1.55	27.26
MAPK signaling pathway (hsa04010)	3.37e-8	3.94e-7	-1.54	26.49
B cell receptor signaling pathway (hsa04662)	1.42e-8	1.86e-7	-1.46	26.36
AGE-RAGE signaling pathway in diabetic complications (hsa04933)	3.67e-8	4.07e-7	-1.52	26.00
NF-kappa B signaling pathway (hsa04064)	1.01e-8	1.41e-7	-1.34	24.68
Focal adhesion (hsa04510)	7.25e-8	7.31e-7	-1.39	22.92
Fc gamma R-mediated phagocytosis (hsa04666)	6.71e-8	7.09e-7	-1.29	21.35

Supplemental Table S6. Top pathway enrichment (Enrichr) results for genes whose expression was consistent with increased accessibility in the immune active (X0) group of LUAD patients identified by xCell clustering. Related to Figure 5.

KEGG pathway	p	Adj. p	Z-score	Combined score
Focal adhesion (hsa04510	0.000255	0.0355	-1.92	15.90
Osteoclast differentiation (hsa04380)	0.00135	0.0936	-1.84	12.15
PI3K-Akt signaling pathway (hsa04151)	0.00518	0.114	-1.99	10.47
Amoebiasis (hsa05146)	0.00339	0.114	-1.82	10.34
Acute myeloid leukemia (hsa05221)	0.00523	0.114	-1.88	9.86
Toxoplasmosis (hsa05145)	0.00610	0.114	-1.79	9.14
Proteoglycans in cancer (hsa05205)	0.00844	0.114	-1.81	8.62
Fc epsilon RI signaling pathway (hsa04664)	0.00851	0.114	-1.74	8.28
Renal cell carcinoma (hsa05211)	0.00784	0.114	-1.67	8.09
AMPK signaling pathway (hsa04152)	0.00725	0.114	-1.64	8.07
Rap1 signaling pathway (hsa04015)	0.00987	0.114	-1.68	7.78
Melanoma (hsa05218)	0.00957	0.114	-1.62	7.52

Supplemental Table S7. Pathway enrichment (Enrichr) results (adj. $p < 0.05$) for genes whose expression was inconsistent with increased accessibility in the immune cold (X1) group of LUAD patients identified by xCell clustering. Related to Figure 5.

KEGG pathway	p	Adj. p	Z-score	Combined score
Platelet activation (hsa04611)	2.24e-6	4.38e-4	-1.92	24.93
Inflammatory mediator regulation of TRP channels (hsa04750)	0.000112	0.0109	-1.99	18.07
Vascular smooth muscle contraction (hsa04270)	0.000449	0.0235	-1.82	14.00
Chemokine signaling pathway (hsa04062)	0.000529	0.0235	-1.85	13.96
cGMP-PKG signaling pathway (hsa04022)	0.000944	0.0235	-1.81	12.63
Focal adhesion (hsa04510)	0.000960	0.0235	-1.77	12.31
Intestinal immune network for IgA production (hsa04672)	0.000731	0.0235	-1.70	12.26
Cholinergic synapse (hsa04725)	0.00141	0.0247	-1.83	12.04
T cell receptor signaling pathway (hsa04660)	0.000960	0.0235	-1.69	11.72
Calcium signaling pathway (hsa04020)	0.00159	0.0247	-1.68	10.83
PI3K-Akt signaling pathway (hsa04151)	0.00197	0.0263	-1.73	10.78
Phospholipase D signaling pathway (hsa04072)	0.00148	0.0247	-1.64	10.70
Glutamatergic synapse (hsa04724)	0.00164	0.0247	-1.61	10.33
Pathways in cancer (hsa05200)	0.00275	0.0299	-1.66	9.77
ECM-receptor interaction (hsa04512)	0.00144	0.0247	-1.44	9.42
Long-term depression (hsa04730)	0.00202	0.0263	-1.42	8.79
Fc gamma R-mediated phagocytosis (hsa04666)	0.00273	0.0299	-1.41	8.30
Rap1 signaling pathway (hsa04015)	0.00461	0.0442	-1.51	8.12
Renin secretion (hsa04924)	0.00268	0.0299	-1.34	7.92
B cell receptor signaling pathway (hsa04662)	0.00474	0.0442	-1.36	7.27
Tuberculosis (hsa05152)	0.00544	0.0484	-1.29	6.74
Leishmaniasis (hsa05140)	0.00474	0.0442	-1.21	6.49