T-cell receptor sequences correlate with and predict gene expression levels in T cells

Hao Wang[1] **and Zhicheng Ji**[2,†]

[1]Department of Statistical Science, Duke University, Durham, NC, USA.
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
[†]Corresponding author. E-mail: zhicheng.ji@duke.edu

ABSTRACT

T cells exhibit high heterogeneity in both their gene expression profiles and antigen specificities. We analyzed fifteen single-cell immune profiling datasets to systematically investigate the association between T-cell receptor (TCR) sequences and the gene expression profiles of T cells. Our findings reveal that T cells sharing identical or similar TCR sequences tend to have highly similar gene expression profiles. Additionally, we developed a foundational model that leverages TCR information to predict gene expression levels in T cells.

Main

As a crucial component of the adaptive immune system, T cells represent a highly heterogeneous cell population with diverse phenotypes and antigen specificities, enabling them to detect and combat a wide array of antigens. Single-cell immune profiling, which measures both gene expression (GEX) and T-cell receptor (TCR) sequences in individual cells, is a groundbreaking method for identifying and understanding T cell diversity. Prior studies have indicated that T cells grouped by clones often exhibit similar gene expression profiles in various conditions, including lung, liver, and colorectal cancers[1–4], breast tumors[5], yellow fever[6], and homeostasis[7]. Computational methods like CoNGA[8] and tessa[9] have been developed to analyze GEX and TCR data concurrently. However, a comprehensive characterization of the association between TCR and GEX across various tissues and disease conditions is still missing. Such an understanding is vital for unraveling the coordinated immune response of T cells and could illuminate potential therapeutic approaches, including the design of TCR-engineered T (TCR-T) cell therapies[10].

In this study, we compiled paired GEX and TCR data for 846,168 T cells from 395 samples, derived from fifteen published single-cell immune profiling datasets[2,5,11–22] (Figure 1a, Supplementary Table S1). These datasets encompass a broad spectrum of diseases, including breast tumor, clear cell renal cell carcinoma, lung cancer, melanoma, esophageal squamous cell carcinoma, COVID-19, HIV-1, HBV, flu-like illness, Kawasaki disease, Multisystem Inflammatory Syndrome in Children (MIS-C), and samples from healthy donors. Utilizing GEX data, we categorized T cells into two primary functional subtypes: CD4+ T cells and CD8+ T cells (Methods). We identified a total of 550,830 CD4+ T cells (65.1%) and 295,338 CD8+ T cells (34.9%) (Figure 1a). Analyzing TCR information, we identified 594,158 unique clones (Methods), where a clone is defined as a group of T cells sharing identical CDR3 amino acid sequences of both TCRα and TCRβ chains within each sample. Among these, 64.2% of cells belong to clones with only one cell, 17.0% to small-sized clones (2 to 10 cells), 12.0% to medium-sized clones (11 to 100 cells), and 6.9% to large-sized clones (more than 100 cells).

We first investigated the GEX similarity among T cells belonging to the same clone. We define a clone's purity as the higher of two proportions: the proportion of CD4+ T cells and the proportion of CD8+ T cells within the clone. Remarkably, 95.1% of clones with at least two cells exhibited 100% clone purity, indicating that these clones consist exclusively of either CD4+ T cells or CD8+ T cells. We observed that clone purities calculated from real data were significantly higher than those derived from randomly permuting the identities of CD4+ and CD8+ T cells in clones of various sizes (Figure 1b). Focusing on clones with at least two cells that consist solely of either CD4+ T cells or CD8+ T cells, we define the GEX dissimilarity within each such clone as the average GEX distance across all possible cell pairs (Methods). We then randomly permuted the cell assignments to the clones for CD4+ and CD8+ T cells separately and recalculated the GEX dissimilarities. Figure 1c and Supplementary Figure 1 demonstrate that GEX dissimilarities calculated from real data are significantly lower than those obtained after random permutations in both CD4+ and CD8+ T cell subsets. These findings indicate that T cells within the same clone exhibit highly similar GEX profiles.

We next investigated the association between GEX similarities and TCR similarities across T cell populations with different

TCRs. We defined the TCR dissimilarity of two T cell clones as the Levenshtein distance between their TCRs (Methods). The GEX dissimilarity of two T cell clones is defined similarly to the single clone case (Methods). Figure 1d and Supplementary Figure 2 illustrate that GEX dissimilarity follows an increasing and then constant pattern as TCR dissimilarity increases. This finding contradicts the linear relationship proposed in a previous study[9]. We observed that GEX dissimilarity ceases to increase and is no longer significantly smaller than GEX dissimilarity calculated using randomly selected clone pairs when TCR dissimilarity exceeds 1 or 4 in CD4+ T cells or CD8+ T cells, respectively. These results suggest that T cell clones with similar TCRs tend to have similar gene expression profiles.

These analyses indicate a substantial correlation between TCR and GEX. Building on this, we hypothesize that TCR information alone can act as a predictor for GEX information. To test this hypothesis, we developed a deep learning framework based on Transformers[23] to predict GEX from TCR (Methods, Fig. 2a). The model comprises a convolutional block and a Pre-Layer Normalization (Pre-LN) Transformer block. The convolutional block, commonly used in other studies[24–26], extracts local features such as motifs within sequences. These features are then input into the Transformer block, which captures complex dependencies between them by modeling positional relationships through positional encoding and attention mechanism. The Pre-LN configuration stabilizes the training process, eliminating the need for a learning rate warm-up stage required by the original Post-Layer Normalization (Post-LN) setting[23,27]. Pre-trained with vast amounts of data, this model can function as a foundational model that can be adapted to various new diseases and tissues through fine-tuning.

We tested the foundational model without fine-tuning using a cross-validation (CV) procedure, where one study was left out for testing and all other studies were used for training in each CV round (Methods). The performance was evaluated by Pearson Correlation Coefficients (PCCs) between the predicted and observed gene expression values in the test set. Figure 2b shows the predicted and observed gene expression values for certain example genes, including CD8A, KLRB1, CD4, and NKG7. These are well-known marker genes for determining T cell phenotypes[28–30] and are among the genes with the best prediction performance (Figure 2c). We identified two distinct types of genes for which the TCR information either has or lacks predictive power (Figure 2d). Genes where the TCR model has predictive power are enriched in GO terms such as immune response (Methods, Figure 2e). We then evaluated the performance of the foundational model with fine-tuning, wherein the model was allowed to utilize information from the left-out study (Methods). After fine-tuning, the model showed improved performance for almost all genes, with an average increase of 10.4% in PCC (Figure 2f). Given that each study represents a distinct disease type and tissue, these results imply that fine-tuning enables the model to better generalize to diseases and tissues not encountered during its pre-training stage.

In summary, we have demonstrated that T cells with identical or similar TCR sequences exhibit highly similar gene expression profiles. This finding supports the design of a set of gene expression markers to identify T cells that recognize specific types of antigens, such as neoantigens[2]. Furthermore, it allows for the prediction of specific gene expression levels based on TCR sequences using a foundational model and fine-tuning. The performance of this approach can be further enhanced with single-cell immune profiling across a broader range of samples, tissues, and diseases, which will be generated in future studies.

Methods

Single-cell data preprocessing

Gene expression count matrices and TCR CDR3 amino acid sequences were downloaded from fifteen single-cell immune profiling datasets (Supplementary Table S1). For single-cell gene expression data, cells with fewer than 200 expressed genes, more than 2500 expressed genes, or with over 20% of reads assigned to the mitochondrial genome were filtered out. For single-cell TCR sequencing data, only cells with high-confidence, full-length sequences, and CDR3 sequences with exactly one TCRα chain and one TCRβ chain were retained. Cells passing filtering criteria for both single-cell gene expression and single-cell TCR sequencing were used in downstream analysis. Seurat (version 4.3.0.1) was used to[31] process the gene expression data for each sample using default settings. The gene expression count matrix was normalized using the NormalizeData function. Highly variable genes were identified using the FindVariableFeatures function with "vst" method. Data were scaled using the ScaleData function. Principal Component Analysis (PCA) was conducted on the 2000 most variable features using the RunPCA function. To address dropout issues in the single-cell RNA-seq data, SAVER (version 1.1.2)[32] imputation was performed on the normalized gene expression values, and the imputed values were log2 transformed.

CD4+ and CD8+ cells were identified in each dataset using a procedure similar to that described in a previous study[2]. Briefly, a density curve was fitted to the log2-transformed and SAVER-imputed CD8A expression values of all cells within each dataset. The trough of the bimodal density curve was used as the cutoff value. Cells with log2-transformed and SAVER-imputed CD8A expression values higher than this cutoff were classified as CD8+ cells, while the rest were categorized as CD8- cells. CD4+ and CD4- cells were identified using a similar approach.

For training and testing the Transformer model, samples lacking information on any of the following T cell marker genes were filtered out: CD4, CD8A, CD8B, NKG7, CST7, CCL5, GZMA, CTSW, KLRB1, CMC1, and CCR7. Genes expressed in

at least 5% of cells in all samples (except those from GSE187515[20], which include only CD4+ T cells) were retained.

TCR and GEX dissimilarity

TCR dissimilarity between two clones was defined as the Levenshtein distance between their CDR3β amino acid sequences. GEX dissimilarity between two T cells i and j, denoted as $d_{i,j}$, was defined as the Euclidean distance between their first 10 principal components (PCs) obtained from their gene expression profiles.

Denote the set of T cells in clone k as $\mathbb{C}_k$. GEX dissimilarity within clone k is defined as $\frac{\sum_{i,j\in\mathbb{C}_k, i\neq j} d_{i,j}}{\binom{|\mathbb{C}_k|}{2}}$. GEX dissimilarity between clone k and clone k' is defined as $\frac{\sum_{i\in\mathbb{C}_k, j\in\mathbb{C}_{k'}} d_{i,j}}{|\mathbb{C}_k||\mathbb{C}_{k'}|}$.

Transformer model architecture

The Transformer model was trained and tested for each gene separately. For each clone within a sample, the log2-transformed and SAVER-imputed gene expression levels were averaged across all T cells belonging to that clone and sample. The gene expression levels across clones and samples were then standardized to have a mean of zero and unit variance. For TCR sequences across clones and samples, both CDR3α and CDR3β amino acid sequences were tokenized by converting each amino acid into a unique integer index. Zero padding was added to ensure that all encoded CDR3α and CDR3β sequences have fixed lengths of L_a and L_b, respectively. Consequently, the inputs to the Transformer model are encoded CDR3α and CDR3β arrays with shapes $n \times L_a$ and $n \times L_b$, respectively, where n is the training sample size. These input arrays were then mapped into dense vectors with dimensions $n \times L_a \times 512$ and $n \times L_b \times 512$ through an embedding layer.

Six convolutional blocks, each consisting of two (BN - GeLU - 1×1 Conv) layers connected via a residual connection and followed by 1×1 max-pooling, were used to extract local features within sequences. The number of filters in these blocks, which represents the dimension of the output space in the convolution, progressively increases from 256 to 512. This approach is similar to that used in a previous study for predicting gene expression using DNA sequences[24]. The kernel size and stride length of the convolutional blocks are 5 and 1, respectively.

A Pre-Layer Normalization (Pre-LN) Transformer block, incorporating positional encoding based on fixed sine and cosine functions[23], was then utilized to capture complex dependencies. This block consists of two sub-layers: a multi-head self-attention mechanism and a point-wise fully connected feed-forward network[23]. Layer normalization was applied before each sub-layer to stabilize the training process[27]. Additionally, dropout and residual addition were implemented after each sub-layer to reduce the risk of overfitting and mitigate the vanishing gradient problem. We used the same hyperparameters as those in the original Transformer paper[23]. For multi-head attention, eight attention heads (h = 8) were employed, with the sizes for the query, key, and value in each head being 64. The input and output dimensions for the feed-forward network are 512, and the hidden layer dimension is 2048.

After employing GlobalAveragePooling1D to compute the average of features along the sequence dimensions, both CDR3α and CDR3β sequences result in output sizes of 512. These are then concatenated, and a 1D dense layer with a linear activation function is used to calculate the expression level for each clone and gene.

Foundation model pre-training

10% of the training data was set aside as the validation set, and mean squared error (MSE) was employed as the loss function. The batch size for training was set to 32. We used the Adam optimizer[33] to minimize the training loss, with an initial learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e-07$. The learning rate was reduced by a multiplicative factor of 0.5 when the validation MSE ceased to improve after 8 epochs. Early stopping was implemented to prevent overfitting, triggered if the validation MSE did not improve for 20 epochs.

Foundation model fine-tuning

During fine-tuning, the model was initialized with the weights obtained from the pre-training. We retrained the linear output layer and froze all preceding layers. For the held-out study, we randomly sampled 80% and 20% of the data for training and testing, respectively. 10% of training data was set aside as the validation set. The loss function used was mean squared error (MSE), and the batch size for training was set to 16. The Adam optimizer[33] was employed to optimize the training loss, with an adjusted initial learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e-07$. The learning rate was reduced by a multiplicative factor of 0.5 when the validation MSE ceased to improve after 50 epochs. Early stopping was implemented to prevent overfitting, triggered if the validation MSE did not show improvement for 20 epochs. All other settings were consistent with those of the pre-trained foundational model.

Gene Ontology Analysis

DAVID[34] with default parameters was used to identify the enriched Gene Ontology terms by comparing the genes with > 0.1 median correlation to all predicted genes as background.

Acknowledgments

Z.J. was supported by the National Institutes of Health under Award Number 1U54AG075936-01.

Author contributions

All authors conceived the study, conducted the analysis, and wrote the manuscript.

Competing interests

All authors declare no competing interests.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

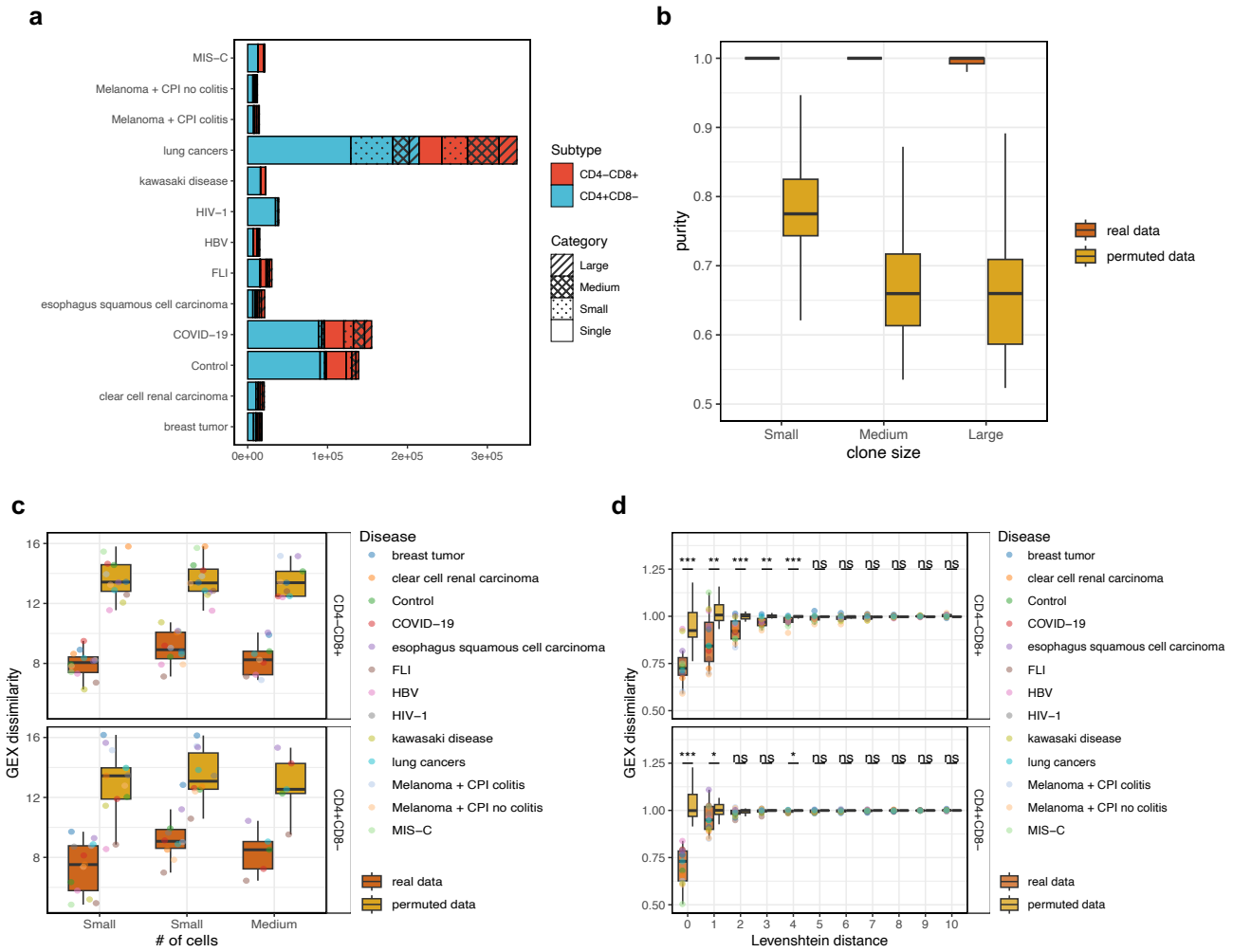

Figure 1. T cells with similar TCR sequences have similar gene expression profiles. **a,** Number of T cells in each study by cell subtypes and clone sizes. **b,** The estimated and permuted purity for clones in small, medium, and large size. Permuted purity is the average over 100 times of permutations. **c,** Real and permuted GEX dissimilarity within CD4+ and CD8+ T cell clones for each disease. Median GEX dissimilarity was first calculated within each sample across all clones. Each data point in the plot is the median of sample-level median GEX dissimilarity across samples with a disease. Permuted GEX dissimilarity is the average over 100 times of permutations. **d,** Relationship between GEX dissimilarity and TCR dissimilarity within CD4+ and CD8+ T cell clones for each disease. For each sample, its median GEX dissimilarity across pairs of clones with a certain Levenshtein distance was divided by the median of median GEX dissimilarity with Levenshtein distance > 5 to allow for comparisons across samples and studies. Each data point in the plot is the median of such normalized sample-level GEX dissimilarity for a disease. Wilcoxon test was performed to compare GEX dissimilarity obtained from real and permuted data. "*" means 0.01 < p-value < 0.05. "**" means 0.001 < p-value < 0.01. "***" means p-value < 0.001. "ns" means non-significant. Permuted GEX dissimilarity is calculated based on only one time of permutation.

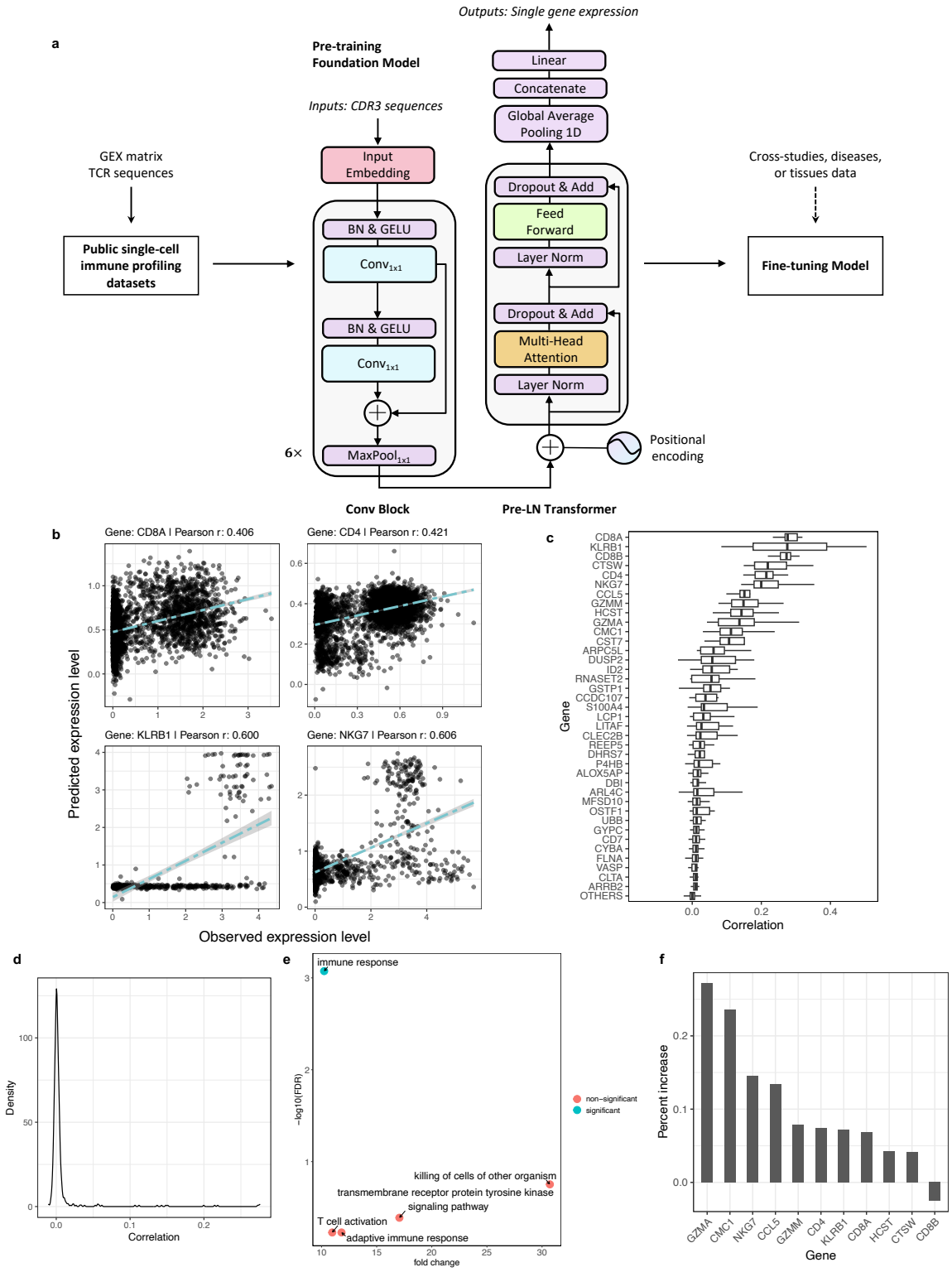

Figure 2. A foundation model for predicting gene expression levels using TCR information. **a,** Architecture of the foundation model. **b,** Example scatterplots comparing real and predicted expression level by the foundation model for CD8A, CD4, KLRB1, and NKG7. **c,** PCCs comparing real and predicted expression levels for each gene. Each data point is a study used as testing. OTHERS represent the remaining genes that have less than 0.01 median PCCs. **d,** Density plot of median PCCs across CV rounds for each gene. **e,** Gene ontology analysis for genes with median PCCs larger than 0.1. Top 5 Gene Ontology terms with the largest -log10(FDR) are shown. **f,** Percent increase of median PCCs comparing the foundation model with or without fine-tuning.

References

1. Guo, X. *et al.* Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing. *Nat. medicine* **24**, 978–985 (2018).
2. Caushi, J. X. *et al.* Transcriptional programs of neoantigen-specific til in anti-pd-1-treated lung cancers. *Nature* **596**, 126–132 (2021).
3. Zheng, C. *et al.* Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356 (2017).
4. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of t cells in colorectal cancer. *Nature* **564**, 268–272 (2018).
5. Azizi, E. *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
6. Minervina, A. A. *et al.* Primary and secondary anti-viral response captured by the dynamics and phenotype of individual t cell clones. *Elife* **9**, e53704 (2020).
7. Zemmour, D. *et al.* Single-cell gene expression reveals a landscape of regulatory t cell phenotypes shaped by the tcr. *Nat. immunology* **19**, 291–301 (2018).
8. Schattgen, S. A. *et al.* Integrating t cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (conga). *Nat. biotechnology* **40**, 54–63 (2022).
9. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of t cell receptor repertoires by single-t cell transcriptomics. *Nat. methods* **18**, 92–99 (2021).
10. Li, D. *et al.* Genetically engineered t cells for cancer immunotherapy. *Signal Transduct. Target. Ther.* **4**, 35 (2019).
11. Su, Y. *et al.* Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. *Cell* **183**, 1479–1495 (2020).
12. Neal, J. T. *et al.* Organoid modeling of the tumor immune microenvironment. *Cell* **175**, 1972–1988 (2018).
13. Borcherding, N. *et al.* Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. *Commun. biology* **4**, 122 (2021).
14. Luoma, A. M. *et al.* Molecular pathways of colon inflammation induced by cancer immunotherapy. *Cell* **182**, 655–671 (2020).
15. Zheng, Y. *et al.* Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment. *Nat. communications* **11**, 6268 (2020).
16. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nat. medicine* **26**, 842–844 (2020).
17. Ramaswamy, A. *et al.* Immune dysregulation and autoreactivity correlate with disease severity in sars-cov-2-associated multisystem inflammatory syndrome in children. *Immunity* **54**, 1083–1095 (2021).
18. Wang, Z. *et al.* Single-cell rna sequencing of peripheral blood mononuclear cells from acute kawasaki disease patients. *Nat. communications* **12**, 5444 (2021).
19. Georg, P. *et al.* Complement activation induces excessive t cell cytotoxicity in severe covid-19. *Cell* **185**, 493–512 (2022).
20. Collora, J. A. *et al.* Single-cell multiomics reveals persistence of hiv-1 in expanded cytotoxic t cell clones. *Immunity* **55**, 1013–1031 (2022).
21. Wen, W. *et al.* Immune cell profiling of covid-19 patients in the recovery stage by single-cell sequencing. *Cell discovery* **6**, 31 (2020).
22. Zhang, F. *et al.* Adaptive immune responses to sars-cov-2 infection in severe versus mild individuals. *Signal transduction targeted therapy* **5**, 156 (2020).
23. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
24. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. methods* **18**, 1196–1203 (2021).
25. Agarwal, V. & Shendure, J. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell reports* **31** (2020).
26. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS computational biology* **16**, e1008050 (2020).

27. Xiong, R. *et al.* On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524–10533 (PMLR, 2020).

28. Kioussis, D. & Ellmeier, W. Chromatin and cd4, cd8a and cd8b gene expression during thymic differentiation. *Nat. Rev. Immunol.* **2**, 909–919 (2002).

29. Truong, K.-L. *et al.* Killer-like receptors and gpr56 progressive expression defines cytokine production of human cd4+ memory t cells. *Nat. communications* **10**, 2263 (2019).

30. Turman, M. A., Yabe, T., McSherry, C., Bach, F. H. & Houchins, J. P. Characterization of a novel gene (nkg7) on human chromosome 19 that is expressed in natural killer cells and t cells. *Hum. immunology* **36**, 34–40 (1993).

31. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

32. Huang, M. *et al.* Saver: gene expression recovery for single-cell rna sequencing. *Nat. methods* **15**, 539–542 (2018).

33. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

34. Sherman, B. T. *et al.* David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic acids research* **50**, W216–W221 (2022).