CANCER IMMUNOTHERAPY DATA SCIENCE CHALLENGE

REPROGRAMMING T CELLS TO COMBAT TUMORS

Organized by the Eric and Wendy Schmidt Center at the Broad Institute

January 9 - February 10, 2023

Challenge 1

Dataset

We consider a single-cell Perturb-Seq dataset consisting of 71,388 T cells taken from a mouse tumor. After quality control, we are left with 4,978 unperturbed cells and 26,031 perturbed cells, where each perturbation is one of 73 different CRISPR mediated single gene knockouts. The training dataset contains all 4,978 unperturbed cells and 23,719 of the perturbed cells, namely the subset of perturbed cells corresponding to 66 out of the 73 gene knockouts. The cells for the remaining 7 knockouts are held out for validation (3 knockouts) and for test (4 knockouts).

For each cell, a 15,077-dimensional gene expression vector and its condition ('unperturbed' or the name of the gene targeted in the knockout) are given. In addition, each cell is classified into one of 5 cell states ('progenitor', 'effector', 'terminal exhausted', 'cycling', 'other') defined by Leiden clustering followed by expert annotation. For the unperturbed cells, the proportion of cells in each cell state is

$$\begin{aligned} & (\text{'progenitor'}, \text{'effector'}, \text{'terminal exhausted'}, \text{'cycling'}, \text{'other'}) \\ = \ & (0.0675, 0.2097, 0.3134, 0.3921, 0.0173). \end{aligned}$$

Note that these proportion vectors can also be thought of as probabilistic masses and will always add up to 1. Similarly, for each gene knockout, the proportion of cells in each cell state can be represented by a 5-dimensional vector of probabilities that add up to 1 and can be computed from the dataset. This proportion vector is related to the effectiveness of a knockout for cancer immunotherapy.

Dataset Structure

- The dataset is hosted by Saturn Cloud. Please see the forum for how to access the data. Each participant also has the option to use the Saturn Cloud computing environment, which provides 100 free hours of compute time per participant and a python environment. Finally, we are providing a jupyter notebook demonstrating how to load and work with the data. This notebook can be found in both Saturn Cloud and in the forum.

- The dataset is provided in the format anndata in the file `sc_training.h5ad`. We provided the gene expression values after normalizing by total counts per cell followed by $\log(1+x)$-transformation in `.X`. Such standard normalization for single-cell gene expression data is described in lecture 2 of the crash course. For completeness we also provided the raw gene expression values before normalization in `.layers['rawcounts']`.
- The condition information ('Unperturbed' or the target gene name if perturbed) is provided in `.obs['condition']`.
- The cell state of each single cell is stored in `.obs['state']` and denoted by 'progenitor', 'effector', 'cycling', 'terminal exhausted', 'other'.
- The cell state of each single cell was classified by experts based on the expression of particular genes. For a description of this procedure, see this tutorial [1] as well as lectures 2 and 3 in the crash course.
- During the pre-processing step over half of the cells were removed; these may be doublets or cells without a single confident guide call. In our pre-processing pipeline, we used a stringent approach, which combines the 10X Genomics Cell Ranger gRNA calling pipeline, i.e., removing any cells that are estimated to have less than or more than 1 guide, and filters based on the ratio of highest abundant to second highest abundant guide, and the ratio of highest abundant guide to total number of guide reads.

Challenge Specifics

For each of the 7 held-out knockouts (targeting genes 'Aqr', 'Bach2', 'Bhlhe40', 'Ets1', 'Fosb', 'Mafk', 'Stat3'), predict the resulting 5-dimensional vector of cell state proportions (a, b, c, d, e), where

$$\begin{aligned} a &= \text{predicted proportion of progenitor cells} \\ b &= \text{predicted proportion of effector cells} \\ c &= \text{predicted proportion of terminal exhausted cells} \\ d &= \text{predicted proportion of cycling cells} \\ e &= \text{predicted proportion of other cells} \end{aligned}$$

and $a + b + c + d + e = 1$. In addition, provide a short maximum 1-page write-up describing your approach.

Evaluation: The predicted cell state proportion vector for each of the 7 held-out knockouts will be evaluated based on the total variation distance, i.e., the l_1-loss, to the experimentally determined ground truth proportion vector. For example, if the ground truth cell state proportion vector is $(0.2, 0.1, 0.3, 0, 0.4)$ and the predicted cell state proportion vector is $(0.19, 0.11, 0.28, 0, 0.42)$, then the loss is $|0.2-0.19|+|0.1-0.11|+|0.3-0.28|+|0-0|+|0.4-0.42| = 0.06$.

Validation and Test: All teams have three chances (by January 13, 20, and 27) to submit the predicted cell state proportion vectors for the 3 held-out knockouts in the validation set (targeting genes 'Aqr', 'Bach2', 'Bhlhe40'). In each submission, the average loss on the 3 knockouts will be computed and used for the leaderboard. To avoid overfitting to the test set, we only offer one successful submission opportunity on the test set. The final ranking of teams will depend on the average loss on the 4 held-out knockouts in the test set (targeting genes 'Ets1', 'Fosb', 'Mafk', 'Stat3').

[1] Andrews, T. S. et al. (2021), Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data, Nat. Protoc., 16(1), pp. 1-9.

Note: 71 out of the 73 targeted genes (66 in training set and 7 in held-out set) are among the 15,077 measured genes in the expression vectors. Two of the targeted genes in the training set ('Fzd1', 'P2rx7') did not pass quality control and thus are not in the expression vectors. Participants are encouraged to use the expression data to build their prediction model, but the expression vectors of the held-out validation and test set will **not** be supplied for predicting the resulting cell state proprtion vectors. Prediction input is limited to **only** the targeted gene names ('Aqr', 'Bach2', 'Bhlhe40', 'Ets1', 'Fosb', 'Mafk', 'Stat3').

External Resources: The application of external resources (e.g., outside transcriptional datasets, gene ontology, pretrained embeddings, etc.) is allowed; however, all external resources must be published or in the public domain and properly credited. In addition, you can optionally use the Saturn Cloud computing environment, which provides 100 free hours of compute time per participant and a python environment.

Challenge 2

Challenge Specifics

Considering the space of potential perturbation experiments, Challenge 1 only explored a small number of genes (73 out of all possible 15,077 genes that passed quality control). The natural continuation in Challenge 2 is to elect further genes to study. Using the dataset described in Challenge 1, participants are tasked to identify novel knockout perturbations for inducing desired cell state proportions, and participant submissions will be experimentally validated. Since this challenge requires experimental validation, the winners will only be announced in June 2023.

Constructing appropriate objective functions that score the efficacy of a perturbation for cancer immunotherapy is still a very active area of research. Challenge 2 (A) and (B) will score participant submissions against pre-defined objective functions, while Challenge 2 (C) will score participant submissions against a participant-defined objective function resulting from Challenge 3.

Participants are encouraged to generalize their algorithms from Challenge 1. Algorithms must take in a gene name and output the predicted cell state proportion vector induced by knocking out said gene. Specifically, submitted algorithms are expected to predict over the set of 15,077 possible genes that passed quality control.

(A) **Immune Checkpoint Blockade Therapy:** A strategy identified by experts to improve the effectiveness of checkpoint therapy is to transiently increase the proportion of progenitor cells [1]. The first goal is thus to identify the gene (out of all 15,077 possible genes that passed quality control) that when knocked out would lead to the largest proportion of progenitor cells (as well as a large enough number of T cells in the tumor).

More specifically, we define the desired cell state proportion vector to be $P = (1, 0, 0, 0, 0)$, where as in Challenge 1 the cell states are ordered as ('progenitor', 'effector', 'terminal exhausted', 'cycling', 'other'). Provide an ordered list of the 15,077 measured genes whose knockout (single-gene knockout) would induce a cell state proportion vector that is closest to the desired vector P in l_1-loss, i.e., it maximizes the proportion of progenitor cells. Provide for each gene the predicted proportion of progenitor cells and indicate for each gene whether its knockout would give rise to **at least** 5% of all cells being in the cycling state (1 if this constraint is satisfied and 0 otherwise). This additional minimum cycling percentage constraint is to ensure that the number of cells that enter the tumor is large enough. For example, the perturbation targeting 'Klf2' would not be of interest for immune checkpoint blockade therapy although it only produces progenitor cells, since the total number of cells is small.

Evaluation: The top 20 submissions of Challenge 1 will each have their top predicted genes (exact number depending on experimental capacity) experimentally validated in a mouse model. For each of these gene knockouts, we compute the resulting cell state proportion vector based on the performed experiment; e.g., knocking out gene i in the experiment results in a cell state proportion vector of $P_i = (a_i, b_i, c_i, d_i, e_i)$. The experimentally validated genes that satisfy the minimum cycling percentage constraint (i.e., $d_i \geq 0.05$) will be ordered by the same objective function (namely maximizing the proportion of progenitor cells a_i). In addition, the experimentally validated genes will also be ordered based on the submitted list. These two orders will be compared as follows. Given a total of K experimentally validated genes that satisfy the minimum cycling percentage constraint, we consider the top k genes as ordered by experimental validation and plot the percentage of these genes that appeared in the top k genes as ordered by participant prediction for $k = 1, 2, ..., K$. Figure 1 shows an example, where the x-coordinate of a point, k/K, equals to the size of each gene list we considered normalized by K and the

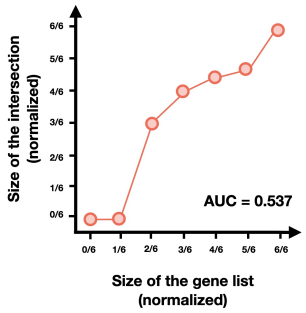

Figure 1: Example where $K = 6$ and the ranking of these genes that satisfy the minimum cycling percentage constraint from the experimental validation is $[g_1, g_2, g_3, g_4, g_5, g_6]$. The AUC score of a submitted (predicted) list where these 6 genes are ordered as $[g_2, g_4, g_1, g_5, g_6, g_3]$ is 0.537 (rounded to the first three decimal points).

y-coordinate is the size of the intersection (i.e., overlap) of the two top k gene lists (ranked by experimental validation and participant prediction) normalized by k. The area under the curve (AUC) is used to score the submitted ordered list[2].

(B) **CAR T-Cell Therapy:** A strategy identified by experts to improve the effectiveness of CAR T-cell therapy is to reduce the proportion of terminal exhausted cells and increase the proportion of progenitor, effector, and cycling cells [2]. The goal is thus to identify the gene (out of all 15,077 possible genes that passed quality control) that when knocked out would lead to a small proportion of terminal exhausted cells and a large proportion of progenitor, effector, and cycling cells.

More specifically, let $P_i = (a_i, b_i, c_i, d_i, e_i)$ be the predicted cell state proportion vector for knocking out gene i. We define the objective function to be

$$\frac{a_i}{0.0675} + \frac{b_i}{0.2097} - \frac{c_i}{0.3134} + \frac{d_i}{0.3921},$$

where each denominator corresponds to the proportion of cells in the respective state as computed among the unperturbed cells, so as to weigh each state similarly. Provide an ordered list (together with the value of the objective function) of the 15,077 measured genes whose knockout would induce a cell state proportion vector that maximizes the objective function. In addition, we will use the same minimum cycling percentage constraint (and your predictions of whether a perturbation satisfies it or not) from Challenge 2 (A).

Evaluation: The top 20 submissions of Challenge 1 will have their top predicted genes (exact number depending on experimental capacity) experimentally validated in a mouse model. For each of these gene knockouts, we compute the resulting cell state proportion vector based on the performed experiment; e.g., knocking out gene i in the experiment results in a cell state proportion vector of $P_i = (a_i, b_i, c_i, d_i, e_i)$. The experimentally validated genes will be ordered by the same objective function, where we replace the denominators with the corresponding cell state proportions of the unperturbed cells obtained from the new experiment. Finally, each group is scored using the AUC computed similarly as in (A) but using the predicted and experimental rankings obtained based on the objective function used in (B).

[2]Note that (1) this scoring scheme measures the similarity between two ordered gene lists using a quantitative analysis similar to a receiver operating characteristic (ROC) curve; (2) a randomly ordered gene list would result in points on the diagonal in expectation, thus giving rise to an AUC of 0.5; (3) we do not penalize participants based on where these top K genes appeared within the whole submitted (predicted) list of 15,077 genes based on the rationale that the perturbation results of the other 15,077-K genes are unknown.

(C) **Other Therapeutic Strategies:** While in (A) and (B) we provided two specific objective functions to score perturbations, these objective functions are not set in stone and a lot of research remains to be done in this area. In fact, the goal of Challenge 3 is to identify functions that satisfy different desired properties for scoring perturbations. It is thus desirable to develop algorithms to predict the effect of novel perturbations (i.e., as of yet untested perturbations outside of the 73 gene knockouts established in Challenge 1), where the predictions could be plugged into a new objective function resulting in an ordered list of all genes.

Provide an algorithm that takes in a gene name (out of the 15,077 possible genes that passed quality control) and outputs the predicted cell state proportion vector induced by knocking out that gene. Provide also for each of the 15,077 genes the corresponding predicted cell state proportion vector.

Evaluation: An objective function based on the winning strategies of Challenge 3 will be used to rank all genes based on the predicted cell state proportion vectors output by the submitted algorithm. The top 20 submissions of Challenge 1 will have their top predicted genes (based on the new objective function from Challenge 3) experimentally validated in a mouse model. For each of these gene knockouts, we compute the resulting cell state proportion vector based on the performed experiment; e.g., knocking out gene i in the experiment results in a cell state proportion vector of $P_i = (a_i, b_i, c_i, d_i, e_i)$. The objective function based on the winning strategies of Challenge 3 will be used to rank these genes both using the predicted cell state proportion vectors output by the submitted algorithm and the cell state proportion vectors computed from the experimental results. Finally, each group is scored using the AUC computed similarly as in (A) but using the predicted and experimental rankings as described above[3].

The sum of the scores in (A), (B) and (C) will be used as the final score for each participant for Challenge 2. In order to be considered as one of the winners, your submitted algorithm in (C) has to output cell state proportion vectors that can reproduce the submitted two ordered lists in (A) and (B). Additionally all participants must provide a short maximum 1-page write-up describing their approach.

External Resources: The application of external resources (e.g., outside transcriptional datasets, gene ontology, pretrained embeddings, etc.) is allowed; however, all external resources must be published or in the public domain and properly credited. In addition, you can optionally use the Saturn Cloud computing environment, which provides 100 free hours of compute time per participant and a python environment.

References

[1] Miller, B. C. et al. (2019), Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade, Nat. Immunol., 20(3), pp. 326-336.

[2] Feucht, J. et al. (2019), Calibration of CAR activation potential directs alternative T cell fates and therapeutic potency, Nat. Med., 25(1), pp. 82-88.

[3] Note that you are not required to know the winning objective function to complete this challenge.

Challenge 3

Challenge Specifics

While the problem of how to score different perturbations is critical for any down-stream task, there has been little work on identifying strong metrics, and simple metrics like the ones described in Challenge 2 (A) and (B) are currently being used. The goal of this challenge is to develop new metrics for ranking perturbations in terms of their capacity to move cells from an undesired to a desired state. While solving Challenges 1 and 2 is not a prerequisite for participating in Challenge 3, it is recommended to read through Challenges 1 and 2 to familiarize yourself more with the context.

Let P_0 denote the empirical gene expression distribution of the unperturbed cells, i.e., P_0 is a distribution in 15,077-dimensional space. Similarly, let P_i denote the gene expression distribution of the cells obtained by knocking out gene i. Let Q denote the desired cell state proportion vector, i.e., Q is a 5-dimensional vector of probabilities that add up to 1. As an optional task, you are invited to submit your proposal for how to choose Q for cancer immunotherapy; see below.

Propose a statistic $s(\cdot)$ that summarizes the gene expression distribution P_i obtained from knocking out gene i as well as a scoring function that takes in P_0, Q and the predicted statistic $\hat{s}(P_i)$ and outputs the score of knocking out gene i (where a larger score indicates a better perturbation). For example, Challenge 2 (A) and (B) use simple scoring functions that do not depend on P_0 and use as the statistic to summarize P_i the cell state proportion vector.

For experimental design purposes, the statistic $s(\cdot)$ should be predictable for unseen perturbations. Also, keep in mind the trade-off between estimating a higher-dimensional and more informative statistic and the requirements in terms of sample size. For example, using the identity map as a statistic would require predicting the full gene expression distribution obtained by knocking out gene i, which would require a large sample size to get an accurate estimate and may not be necessary for identifying optimal perturbations with respect to the desired 5-dimensional cell state proportion vector Q.

Additional desiderata of the proposed statistic and scoring function are:

(a) scoring function should ideally depend on P_0;

(b) scoring function should ideally take uncertainty into account given by the different sample sizes for the perturbations in the training dataset;

(c) different perturbations lead to different growth rates and thus result in a different number of cells; the scoring function may also take into account the predicted number of cells resulting from a guide and favor perturbations with a large growth rate;

(d) scoring function could take into account the classification boundaries of each cell state;

(e) it may be helpful to use a more informative statistic than the cell state proportion vector that could for example take into account the classification boundaries of each cell state.

Evaluation: All teams participating in this challenge will submit a write-up of a maximum of 3 pages (excluding figures and references, for which an additional 2 pages can be used maximally). After the submission deadline, all participants get a chance to evaluate the submissions; namely, they get a total of 3 votes that they can distribute among the assigned submissions. The top 20 submissions based on participant ranking will be evaluated by an expert panel, and 5 out

of these 20 submissions will be selected and used to evaluate the perturbations proposed in Challenge 2 (C); see Challenge 2 for more details.

Optional: As an optional task, you are invited to submit your proposal for how to choose Q for cancer immunotherapy. Provide a proposal for Q together with a short justification (1 page maximum including figures and references). For the evaluation of Challenge 2 (C), Q will be selected by an expert panel, taking into consideration the optional submissions from participants.

Optional Material

Optional Auxiliary Datasets

- For each genetic knockout, 3 different guides were used to ensure the gene was knocked out. Note that some guides do not appear in the dataset since all corresponding cells were removed in the preprocessing step. This guide information is stored in `.obs['gRNA_maxID']`. You are encouraged to investigate guide differences in order to assess the robustness of perturbing the targeted gene.
- For each guide, we provide the abundance of that guide when infecting the cells (obtained using plasmid pool sequencing to compare each guide's representation in the plasmid pool). This could be used to estimate the viability of the perturbation by this guide in the tumor microenvironment. This information can be found in `guide_abundance.csv`.
- An extra barcode can be used to read out the clone information for each cell. This data can be found in `clone_information.csv`. This could be helpful in order to take into account temporal aspects of how cells move between different cell states by analyzing the proportion of cells in the same clone across different cell states.
- The cells were loaded into 4 different wells on a 10x Genomics Chip. This may introduce a batch effect, in part because sequencing depth may vary across the 4 wells. Information on this is provided in `.obs['lane']`, which may be useful as a covariate in the prediction tasks.
- In addition to the Perturb-Seq data collected from T cells in the mouse melanoma tumor, we also performed joint profiling of gene expression and chromatin accessibility in unperturbed T cells. As with the Perturb-Seq data, this dataset is hosted by Saturn Cloud and can be found in `scRNA_ATAC.h5`. To learn more about this experiment and how the data may inform your work on the challenges, please read the following section Epigenetics review.

Optional Review Articles

This challenge draws on many different subject areas, which are covered in the three introductory crash course lectures. To supplement this, we provide scientific review articles on these subject areas, which can give you a more detailed perspective and point you to other relevant datasets and data modalities. **Reading these articles is not necessary to complete the challenge, but we believe these can be a helpful resource**.

T cell basic science and T cell exhaustion review

- A guide to cancer immunotherapy: from T cell basic science to clinical practice
- 'Stem-like' precursors are the fount to sustain persistent CD8+ T cell responses
- CD8 T Cell Exhaustion During Chronic Viral Infection and Cancer
- Defining 'T cell exhaustion'

Other immunology resources These resources give an overview of immunology. More focused units on T cell exhaustion can be found in the review articles above.

- Fundamentals of Immunology: T Cells and Signaling

- Immune: A Journey into the Mysterious System That Keeps You Alive

Single cell transcriptomics review

- Single-cell transcriptomics to explore the immune system in health and disease

Single cell transcriptomics courses and data repositories

- Single-cell best practices
- Orchestrating Single-Cell Analysis with Bioconductor
- Analysis of single cell RNA-seq data

Below are repositories of single cell transcriptomic datasets. These may be combined with the Perturb-Seq data from the challenge to make better models. There are thousands of datasets here, and we note that these datasets are collected from different organisms (human, mouse, ...), tissue types (skin, lung, ...) and disease states (healthy, cancer, infection, ...). Therefore if using these additional datasets, you will need to choose datasets that have a meaningful biological relationship to the Perturb-Seq data in the challenge. For example, T cells collected from other cancers like breast or lung cancer may be relevant, whereas data collected from neurons in brain tissue would be less relevant.

- Human Cell Atlas Data Portal
- Tabula Sapiens Human transcriptome reference at single cell resolution
- Tabula Muris Mouse transcriptome reference at single cell resolution
- Single cell studies database
- Jingle Bells: A repository of standardized single cell RNA-Seq datasets for analysis and visualization at the single cell level

Also as a technical note, the gene names (or gene symbols) used for mouse and human genes are unfortunately different from one another. The genes in mice and in humans are evolutionarily related to one another, and often carry out similar function, but the nomenclature differs. As an example, the Pdcd1 gene in mice is named PDCD1 in humans. This is important to know if you decide to incorporate data from human studies, since the Perturb-Seq data in the challenge is collected from mice. To find the mapping between corresponding mouse and human genes (often referred to as homologs), you can use the BiomaRt resource. Also here is one potential implementation of this mapping in both the R and python languages.

Gene ontology review In natural language processing we often use embeddings, where words with similar meanings have similar representations. These embeddings can make models more generalizable and also help when training data is limited. Similarly, in biology we can learn gene embeddings, where genes with similar function have similar representations. These representations may be learned from Gene Ontology, which incorporates decades of biological knowledge on gene function for various organisms (including mouse). Gene ontology describes three aspects of gene function: molecular function, cellular component, and biological process. **Incorporating gene ontology may lead to better models, and again we stress it is up to you whether to experiment with this!**

- The Gene Ontology Resource: 20 years and still GOing strong

Epigenetics review In the same way that we can describe the physical world using different modalities such as video, audio, and text, the state of a biological cell can be described with modalities other than gene expression. In biology, the hope is that incorporating additional modalities will result in more predictive and interpretable models, but this remains an open question. In this challenge, you work with a Perturb-Seq dataset where the T cell states are quantified based on gene expression levels in the T cells. However, what in the cell regulates gene expression, and determines which genes are expressed and which are not? The study of epigenetics largely focuses on the modalities that define the causal regulatory relationships among genes, and building these regulatory relationships provides a principled view of how the cell state is programmed and may be shifted from one state to another.

In epigenetics, we measure data modalities beyond gene expression, including chromatin accessibility, DNA methylation, and histone modification. For example, DNA in the nucleus is wrapped around proteins called nucleosomes and packaged into a complex called chromatin. This chromatin packing varies greatly across our chromosomes. When this packing is very tight, the genes encoded in the DNA are transcriptionally inactive and not expressed. When the packing is more loose, the genes are transcriptionally active and expressed. This packing can be measured with the ATAC-Seq assay. We know that the packing varies across T cell states and is important in regulating T cell exhaustion.

In addition to the Perturb-Seq data in this challenge, we have also collected ATAC-Seq data that you may incorporate into your models. This epigenetic data can provide information on how chromatin accessibility relates to gene expression in the different exhausted T cell states, and to build regulatory relationships from this. In particular from this data we may be able to identify specific transcription factors, special regulatory genes that can turn gene expression on or off, that drive T cell differentiation towards one state over another. Transcription factors bind specific DNA sequences, or motifs, in regions of open chromatin accessibility, and thereby direct gene expression. Transcription factor genes are expected to have especially strong effects in controlling T cell states and the overall proportions of different T cell states in tumors.

The epigenetic experiment followed the same setup as the Perturb-Seq experiment, except that mice with melanoma were treated with unperturbed T cells (T cells receiving non-targeting control sgRNA) instead of T cells with gene knockouts. After collecting T cells from the tumor, we jointly measured both chromatin accessibility and gene expression from the same single T cells using the Chromium Single Cell Multiome assay (10x Genomics), and we pre-processed the data using Cell Ranger ARC v2.0.2. We provide you with the resulting filtered feature-barcode matrix, where the features are both genes and peaks of chromatin accessibility. The dataset consists of 4,208 cells that can be found in `scRNA_ATAC.h5`. You can obtain the cell state annotations by clustering the gene expression data alone in the same manner as we had annotated the T cell states for the Pertub-Seq data (see code provided in `sc_training_visualization.ipynb`. To work with the data, we suggest using either `MuData` or `muon`, python packages that are extensions to the `AnnData` and `scanpy` frameworks that allow you to work with multimodal data (both RNA-Seq and ATAC-Seq), or ArchR, a popular package in R for working with ATAC-Seq data. To help infer which transcription factors are active in a T cell state, we also suggest using `JASPAR`, which is a database of transcription factor binding profiles. Within a T cell state, the presence and enrichment of a transcription factor DNA-binding motif in an open chromatin (i.e., accessible) region may indicate that a transcription factor regulates that T cell state.

Below are review articles that introduce you to epigenetics and how it relates to T cell exhaustion. Sensibly combining epigenetic data with the Perturb-Seq data may result in better models. **We stress that it is an open question whether incorporating different modalities will lead to better models, and it is up to you whether to experiment with this in the challenge!**

- Epigenetic regulation of T cell exhaustion
- Divergent clonal differentiation trajectories of T cell exhaustion
- Characterizing cis-regulatory elements using single-cell epigenomics
- Assessment of computational methods for the analysis of single-cell ATAC-seq data
- MUON: multimodal omics analysis framework
- ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis
- JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles
- chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data

Optional Related Papers

Below we provide a set of papers on T cell datasets, genetic perturbations, and modeling approaches. The chosen papers are not meant to be an endorsement of specific datasets and models, and are not meant to be a comprehensive overview. **Reading these articles is not necessary to complete the challenge. Treat these as a resource for learning more about these incredibly active fields, if you are interested.**

Single cell transcriptomic T cell datasets

- A unified atlas of CD8 T cell dysfunctional states in cancer and infection
- Shared and distinct biological circuits in effector, memory and exhausted CD8+ T cells revealed by temporal single-cell transcriptomics and epigenetics
- Divergent clonal differentiation trajectories of T cell exhaustion
- A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade
- Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma
- Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma
- Deciphering the transcriptomic landscape of tumor-infiltrating CD8 lymphocytes in B16 melanoma tumors with single-cell RNA-Seq
- Dynamic chromatin regulatory landscape of human CAR T cell exhaustion

CRISPR/Cas9 genetic perturbation screens

- Genome-wide CRISPR screens of T cell exhaustion identify chromatin remodeling factors that limit T cell persistence
- Enhanced T cell effector activity by targeting the Mediator kinase module
- Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq

Modeling perturbations and causality

- Machine learning for perturbational single-cell omics
- Elements of Causal Inference: Foundations and Learning Algorithms
- GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations
- Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling
- Learning interpretable cellular responses to complex perturbations in high-throughput screens
- PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations
- Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution
- Active Learning for Optimal Intervention Design in Causal Models
- Control of cell state transitions
- GeneDisco: A Benchmark for Experimental Design in Drug Discovery
- scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data