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