

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Biophysics
have examined a dissertation entitled

**Methods and applications of single cell
transcriptomics**

presented by **Tessa D. Green**

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Martha L. Bulyk
Signature _____

Typed name: Prof. Martha L. Bulyk

Allon M. Klein
Signature _____

Typed name: Prof. Allon M. Klein

Alex K. Shalek
Signature _____

Typed name: Prof. Alex K. Shalek

Jonathan S. Weissman
Signature _____

Typed name: Prof. Jonathan S. Weissman

Date: _____ October 2, 2023 _____

Methods and applications of single-cell transcriptomics

A DISSERTATION PRESENTED
BY
TESSA D. GREEN
TO
THE COMMITTEE ON HIGHER DEGREES IN BIOPHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOPHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
OCTOBER 2023

©2023 – TESSA D. GREEN
ALL RIGHTS RESERVED.

Methods and applications of single-cell transcriptomics

ABSTRACT

Single-cell transcriptomics has transformed biology by enabling deep interrogation of the RNA contents of individual cells. This has led to in-depth study of cellular heterogeneity and the role of cell state transitions in disease. Observational single-cell studies have moved towards hypothesis generation about complex cellular processes; interventional studies enable mechanistic insights. Here, we use single cell transcriptomics to identify cell states underlying nasal polyp formation. We also interrogate how gene expression in the sinus changes in response to asthma treatment, combining single cell and bulk analyses for a more complete view. We then move beyond changes in individual cell types to uncover how cells relate to each other, using matrix decomposition to reveal multi-cell-type changes in gene expression in breast cancer, suggesting interaction signatures specific to breast cancer subtypes, and interactions predicting response to treatment. Our findings on drug response in these two disease cases were limited by a lack of robust statistical tools; to improve tools for interrogating perturbation response in single cells, we created an annotation-harmonized collection of single cell perturbation studies, then used this data resource to characterize the performance of E-statistics for evaluating perturbation similarity and efficacy. In total, this thesis contains two stories of using perturbation in patients to study disease, and one example of using a collection of datasets to improve methods used for interrogating biological systems.

Contents

TITLE PAGE	i
COPYRIGHT	ii
ABSTRACT	iii
o INTRODUCTION	1
1 THE SINUS TRANSCRIPTOME IN ASPIRIN-EXACERBATED RESPIRATORY DISEASE	5
1.1 Abstract	6
1.2 Introduction	6
1.3 Proliferating B cells in nasal polyps	11
1.4 The inferior nasal turbinate as a marker tissue for drug response	17
1.5 Discussion	30
2 CELL STATE AND CELL-CELL COMMUNICATION IN TRIPLE-NEGATIVE BREAST CAN-CER	32
2.1 Abstract	32
2.2 Introduction	33
2.3 Methods	39
2.4 Results	43
2.5 Discussion	59
3 POINT CLOUD DISTANCE METRICS FOR SINGLE CELL PERTURBATION DATA	64
3.1 Abstract	65
3.2 Introduction	66
3.3 Methods	69
3.4 Results	81
3.5 Discussion	97
4 CONCLUSION	100

APPENDIX A AN UNBIASED ESTIMATOR FOR THE E-DISTANCE	106
APPENDIX B ROBUSTNESS ANALYSIS OF E-STATISTICS	111
APPENDIX C COMPUTATIONAL COMPLEXITY OF THE E-TEST	119
REFERENCES	122

List of Figures

1.1	Arachidonic acid metabolism is dysregulated in aspirin-exacerbated respiratory disease (AERD). Red arrows denote changes in levels or activity in AERD. Receptors and ligand-receptor interactions are denoted in blue. COX-1 and COX-2 refer to cyclooxygenase 1 and 2 respectively. LTA ₄ is an unstable product. Simplified from ¹¹¹	8
1.2	Quality control plots for B cells from surgical samples. nFeature_RNA is the number of genes detected in each cell. nCount_RNA is the number of counts detected in each cell. percent.mt is the percentage of counts from mitochondrial RNA. These plots were used to establish cutoffs of 200 — 2000 for the number of features per cell and < 10% mitochondrial reads. Cells outside of these ranges were excluded from further analysis.	13
1.3	UMAP projection of scRNA-seq from ¹⁵⁰ . Cells are colored by diagnosis. The inset shows B cells re-analyzed here. κ and λ class-switched cells form two separate clusters; a third cluster is characterized by higher levels of proliferation markers.	13
1.4	Louvain clustering of B cells, displayed on UMAP coordinates. 5 clusters are apparent.	14
1.5	Major clusters are separated by κ and λ identity. IGK and IGL fractions are the percentages of total reads originating from genes starting with IGK or IGL respectively.	15
1.6	KL ratio for each sample (left) and diagnosis (right). Although the ratio varies across patients, it does not change with diagnosis.	15
1.7	$\kappa-\lambda$ ratio for each sample (left) and diagnosis (right). Although the ratio varies across patients, it does not change with diagnosis.	16
1.8	Expression of IL-5 receptor alpha (IL ₅ R α) in plasma cells from patients with AERD and chronic rhinosinusitis with nasal polyps (CRSwNP).	16
1.9	multidimensional scaling (MDS) plot for all samples shows read depth predominates. Points are scaled by the total number of reads and colored by condition. Number of reads is represented by the scale of the dots; small, less, sequenced dots are separated by MDS dimension 1. A: control, B: acute, C: 8 weeks, D: 10 weeks.	18
1.10	MDS plot for samples with at least 30M reads shows that outliers were successfully removed. A: control, B: acute, C: 8 weeks, D: 10 weeks.	19

1.11	Scrapes consisted predominantly of epithelial cells. Cell type fractions predicted using CIBERSORTx ¹⁴⁸ . Samples are labeled as 'patient ID_timepoint'.	21
1.12	IL-22 counts and predicted neutrophil level are somewhat correlated across all samples (Pearson = 0.68).	22
1.13	Although IL-22 is significantly different during the acute reaction, this finding is likely not meaningful.	23
1.14	RAMP1 is downregulated during acute aspirin reaction (adjusted p=0.02).	23
1.15	Ajuba is upregulated during acute aspirin reaction (adjusted p=0.04).	24
1.16	Three different alpha-amylases are decreased during acute aspirin reaction.	25
1.17	Growth hormones GH1 and GH2 have matching decreases during aspirin reaction.	26
1.18	Genes which had increased expression after aspirin desensitization. Adjusted p-values for individual comparison from control using DESeq are denoted using asterisks (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$).	27
1.19	Genes which had decreased expression after aspirin desensitization. Adjusted p-values for individual comparison from control using DESeq is denoted using asterisks (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$).	28
1.20	Some genes with only unadjusted significant p-values with known relevant functions may be playing a functional role in desensitization response.	29
2.1	uniform manifold approximation and projection (UMAP) plots produced using PCA and single-cell variational inference (scVI) of breast cancer data from ¹⁵¹ . A circle highlights a subpopulation of ER+ tumor cells which are visibly distinct only after scVI reduction.	44
2.2	multicellular program (MCP) membership for PCA and scVI reductions of data in ¹⁵¹ . Black squares denote cell types included in the MCP on the x-axis.	45
2.3	Pair plot for the first MCP for the scVI dimensionality reduction. Along the diagonal is a histogram of the average score for each MCP by sample for the listed cell type. Significance and Pearson correlations for each pair are displayed in the upper triangle. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). Points marked in red are samples from patients with TNBC.	46
2.4	Pair plot for the first MCP for the PCA dimensionality reduction. Along the diagonal is a histogram of the average score for each MCP by sample for the listed cell type. Pearson correlations for each pair and associated significance are displayed in the upper triangle. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). Points marked in red are samples from patients with TNBC.	48
2.5	MCP score for MCP ₃ in breast cancer tumors from ¹⁵¹ . triple-negative breast cancer (TNBC) tumors are shown in blue; ER+ and HER2+ tumors are analyzed together and shown in red. MCP ₃ score is relatively similar across breast cancer types aside from in pericytes; where a high-MCP ₃ subpopulation is seen only in TNBC.	49

2.6	Pair plot for MCP ₄ from analysis of ²²⁰ . Along the diagonal is a kernel density estimate of the average score for each MCP by sample for the listed cell type. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). All patients in the partial response category had tumors which shrank after treatment but were not entirely eliminated.	51
2.7	Single-cell MCP scores for cells from patients who did or did not respond to treatment in ²²⁰ . Cells from patients who received either treatment are pooled here.	52
2.8	MCP loadings for a response-predictive MCP in memory B cells and CD4 central memory T cells. The top ten positive and negative contributory genes are shown, with their associated component contribution on the y-axis.	55
2.9	The proportion of MCP genes identified via the DIALOGUE multilevel marketing procedure that are also significant (adjusted p-value <0.01) according to differential expression testing of high-MCP and low-MCP cells. Cell type abbreviation expansions are in Table 2.1. The t at the start of each name refers to tumor residence. Along the x-axis are genes increased (up) or decreased (down) for each MCP.	56
2.10	Jaccard indices between top MCP genes as determined by DIALOGUE and the MCP loadings (left) and as determined by multilevel modeling or by direct testing (right). The multilevel modeling genes (DIALOGUE genes as determined by the method in ⁹³) have minimal overlap with the MCP loadings, and moderate overlap with genes differentially expressed between top and bottom MCP	57
2.11	Top genes with increased expression in each of the most treatment-response associated cell types in MCP ₄ of ²²⁰ . Rankings are z-scores from t-tests comparing gene expression in the 10% of cells with the highest and lowest MCP ₄ scores for each cell type.	58
2.12	An example imaging mass cytometry (IMC) region of interest with cell type labeling from AnnoSpat. This is from a slice of a core needle biopsy of a pre-treatment TNBC patient.	61
2.13	Example neighborhood enrichment analysis. In the image shown in Fig. 6, fibroblasts are over-represented in the 100-300 distance range. Neighborhood enrichment statistics were calculated using squidpy ¹⁵²	63
3.1	Effect of subsampling UMI counts per cell and number of cells per perturbation on E-statistics. (A) E-distance of each perturbation to unperturbed in ¹⁴⁹ while subsampling the number of cells per perturbation; Color indicates E-test results; “significance lost”: perturbation significant when all cells are considered, but not significant after subsampling. The E-test loses significance with lower cell numbers while the E-distance actually increases. (B) Overall number of perturbations with significant E-test decreases when subsampling cells. (C) As in Subfigure A but subsampling UMI counts per cell while keeping the number of cells constant. Loss of E-test significance and dropping E-distance to unperturbed as overall signal gets deteriorated with removal of counts. (D) As in Subfigure B but subsampling UMI counts per cell while keeping the number of cells constant.	79

3.2	Single-cell perturbation-response datasets are diverse in type, size, and quality. (A) The majority of included datasets result from CRISPR (DNA cut, inhibition or activation) perturbations using cell lines derived from various cancers. (B) Sequencing and cell count metrics across scPerturb perturbation datasets (rows), colored by perturbation type. From left to right: distribution of total RNA counts per cell (left); distribution of the number of genes with at least one count in a cell (middle); distribution of number of cells with at least one count of a gene per gene (right). Most datasets have on average approximately 3000 genes measured per cell, though some outlier datasets have significantly sparser coverage of genes. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.	82
3.3	(A) Definition of E-distance, relating the width of cell distributions of high-dimensional molecular profiles to their distance from each other. A large E-distance of perturbed cells from unperturbed indicates a strong change in molecular profile induced by the perturbation. (B) Distribution of E-distances (plus 1 for log scale) between perturbed and unperturbed cells across datasets. The number of perturbations per dataset is displayed along the bottom. Note that this plot is best used to compare the shape of the E-distance distribution rather than the magnitude; the mean E-distance will vary significantly with other dataset properties. (C-E) Analysis based on E-statistics for one selected dataset ¹⁴⁹ : (C) Distribution of E-distances between perturbed cells and unperturbed cells as in subfigure B. Each circled point is a perturbation, i.e., represents a set of cell profiles. Each perturbation was tested for significant E-distance to unperturbed (E-test). (D) Pairwise E-distance matrix across the top and bottom 3 perturbations of Figure 3C and the unperturbed cells. (E) UMAP of single cells of the weakest (left, bottom 3) and strongest (right, top 3) perturbations.	84
3.4	Behaviour of E-distance and E-test in simulated data. Varying L ₂ FC of DEGs and proportions of genes that are simulated as differentially expressed. E-test significance (adjusted p-value < 0.01) marked with “*”.	85
3.5	Relationship between E-test significance and visual separation in two UMAP dimensions. UMAPs for simulated data corresponding to Figure 3.4. Note that we forced 1/3 of cells in the non-control group (orange) to have no phenotype change in order to mimic CRISPR-cas9 data. Significance of E-test marked in top right corner of each UMAP as in Figure 3.4 * indicates significant E-test results, n.s. not significant. Genes DE is the number of genes differentially expressed. The red line denotes the boundary of significance loss.	87
3.6	E-distance dissects perturbation hierarchy in data from Papalex et al. (A) E-distance between cells of all pairs of perturbations in Papaplexi et al. dataset ¹⁵³ Hierarchical clustering of this matrix reveals two groups, one which is more similar to unperturbed cells (green) and one which has a stronger transcriptional change (orange). (B) Signaling pathway downstream of IFNG receptor. Permutations of nodes upstream of IRF1 induce similar phenotypes.	89

3.7	Histogram of cells per perturbation in ⁷³ . Many perturbations had fewer cells than our standard recommendation.	90
3.8	Background doses are separated by an underscore. For (a), the units are BMP4 (ng/mL); Scriptaid and decitabine(μL); retinoic acid (ng/mL). For (b), the units are BMP4 (ng/mL); EGF and bFGF (ng/mL), Scriptaid and decitabine(μL). E-distance is measured from the lowest dose of the x-axis drug on the indicated background set of perturbations. Data from ⁷³	91
3.9	(A, B) Pairwise E-distances for two datasets ^{119,141} between perturbations and control in different feature spaces defined by four different methods (ChromVar, latent semantic indexing (LSI)-embedding, gene scores, and peaks). (C, D) Corresponding pairwise Pearson correlations of E-distances to control across feature spaces.	93
3.10	Visualization of cell type relationships in full multimodal dataset after batch correction. Coordinates and cell type annotations from ⁸¹	95
3.11	Hierarchical clustering of pairwise E-distances computed using RNA matches prior knowledge of transcriptome-defined cell types. Dendrogram and heatmap use the same distances. Data from ⁸¹	95
3.12	As in Figure 3.11 but using antibody-tagged surface proteins instead of RNA.	96
A.1	Sample correcting the calculation for σ removes count bias of the E-distance with respect to sample size. The naive estimator uses Equation A.2, and the bias corrected uses Equation A.4. Vertical lines show standard deviation across 50 simulation runs. The mean of σ and E-distance remain confident as fewer data points are sampled. Each of the 30 dimensions used in this example is an independent Gaussian with the same mean and variance.	110
A.2	PCA prior to E-distance biases results at low n. Both sigma and delta change due to PCA. Convergence at around 100 cells. Dimensions were independent.	110
B.1	Asymmetric sample sizes: as long as both conditions have at least 200 cells, E-distances and E-tests are well-behaved. For varying number of cells in the control group and fixed (n=200) number of cells per perturbed group we recorded for each perturbation the (A) E-distance to control cells and (B) the number of perturbations with a significant E-test (p-value < 0.05). Black line corresponds to average over all perturbations, all colored lines to different perturbations. RNA data from ¹⁵³	112
B.2	(A),(B) E-distance is largely stable when at least 2000 (gray line) genes are used to compute PCA. The dotted line is at 2000 highly variable genes. (C),(D) Modification of feature selection to specifically use genes which are differentially expressed under perturbations minimally effects the E-distance. HVG: highly variable genes; DEG: differentially expressed genes; hybrid: the union of HVGs and DEGs.	114
B.3	Overlap of genes from different feature selection methods in the two datasets considered.	115

List of Tables

1.1	Marker genes for B cell substates as supplied by Tanya Laidlaw. HI refers to high expression, MID to medium expression, LO to low expression. Cells are left blank where expression is not clearly defined for the relevant cell type.	12
1.2	High-quality samples available for each patient	19
2.1	Cell type abbreviations and names after filtering for DIALOGUE analysis of ²²⁰	41
2.2	Cell and sample counts used for DIALOGUE analysis after filtering ¹⁵¹	43
2.3	Cell type associations with treatment response. This is testing whether the average MCP score for each cell type for each patient's pre-treatment tumor sample was significantly different in responding and non-responding patients. The p-values were calculated using an independent t-test. Adjusted p-values were adjusted for the number of cell types tested using a Benjamini-Hochberg correction factor.	53
3.1	Dataset information for experiments included in the scPerturb database. *: perturbation total treats perturbations A, B, and (A and B) as three unique perturbations †: T-cell receptor (TCR) stimulation	70

LIST OF ACRONYMS

- AERD** aspirin-exacerbated respiratory disease
AJUBA Ajuba lim protein
ANXA₃ annexin A₃
bFGF basic fibroblast growth factor
BIRC₅ baculoviral IAP repeat containing 5
BRCA₁ breast cancer gene 1
CAF cancer-associated fibroblast
CCA canonical correlation analysis
CCDC85A coiled-coil domain containing 85A
CFAP47 cilia and flagella associated protein 47
CITE-seq cellular indexing of transcriptomes and epitopes
CLR centered log ratio
COPD chronic obstructive pulmonary disease
CRABP₂ cellular retinoic acid binding protein 2
CRSwNP chronic rhinosinusitis with nasal polyps
CysLT cysteinyl leukotriene
CysLT₁R cysteinyl leukotriene receptor 1
CyTOF cytometry by time of flight
DC dendritic cell
DEG differentially expressed gene
EGF epidermal growth factor
ER estrogen receptor
EP₂ PGE₂ receptor 2 subtype
FLACC₁ flagellum associated containing coiled-coil domains 1
FN₁ fibronectin 1
FOXD₂ forkhead box D2
GREB₁ growth regulating estrogen receptor binding 1
GWAS genome-wide association study
HIST₁H₂AJ histone Cluster 1 H₂A Family Member J
HER₂ human epidermal growth factor receptor 2

HSP heat shock protein
HVG highly variable gene
IgE immunglobulin E
IGLL1 immunoglobulin lambda like polypeptide 1
IL5R α IL-5 receptor alpha
ILC innate lymphoid cell
ILC2 type 2 innate lymphoid cell
IMC imaging mass cytometry
iPSC induced pluripotent stem cell
JUN Jun proto-oncogene
KRT5 keratin 5
L2FC log-2-fold change
LFC log-fold change
LSI latent semantic indexing
LTC4S leukotriene C₄ synthase
LTE4 leukotriene E₄
LXA4 lipoxin A₄
MAIT cell mucosal-associated invariant T cell
MCPT Monte Carlo permutation test
MCP multicellular program
MDS multidimensional scaling
MKI67 marker of proliferation Ki-67
miRNA microRNA
NACT neoadjuvant chemotherapy
NB negative binomial
NK natural killer
NOD1 nucleotide-binding oligomerization domain
PBMC peripheral blood mono-nuclear cell
PCA principle component analysis
PC principle component
PD-1 programmed cell death protein 1

PD-L1 programmed death-ligand 2
peak bc peak barcode
PGE₂ prostaglandin E₂
PGD₂ prostaglandin G₂
PGH₂ prostaglandin H₂
pDC plasmacytoid dendritic cell
PR progesterone receptor
RAMP1 receptor activity modifying protein 1
S100A13 S100A13
scATAC-seq single cell assay for transposase-accessible chromatin using sequencing
scRNA-seq single-cell RNA sequencing
scVI single-cell variational inference
SLC solute carrier
sgRNA single guide RNA
SNP single nucleotide polymorphism
SPHK1 sphingosine kinase 1
TAM tumor-associated macrophage
T_{cm} cell central memory T cell
Th2 cell type 2 T helper cell
TCF7L2 transcription factor 7-like 2
TCR T-cell receptor
TME tumor micro-environment
TNBC triple-negative breast cancer
tSNE t-distributed stochastic neighbor embedding
TTLL9 tubulin tyrosine ligase-like 9
TXA₂ thromboxane A₂
UMAP uniform manifold approximation and projection
UMI unique molecular identifier
VAE variational autoencoder
ZINB zero-inflated negative binomial

Author List

The following authors contributed to Chapter 1: Stefan Peidli, Ci Yue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, Chris Sander.

S.P. and T.D.G. jointly wrote the manuscript with contributions from T.G., C.Sh. and J.P.T.-K., and consultation from N.B., C.Sa and L.S. T.G. performed ATAC-seq data processing. S.P., T.D.G., and C.Sh. performed all other data processing. Analysis, statistics, and mathematical development was performed jointly by S.P. and T.D.G. J.M. developed the initial website housing the database. S.G. unified annotations for drug perturbations. C.Sa., N.M., A.L., D.S.M., and N.B. conceived of the project, with initial idea development by B.Y. A.L., N.B., and C.Sa. supervised the work with consultations from D.S.M.

The following authors contributed to Chapter 2: Kathleen Buccheit, Daniel Dwyer, Jose Ordovas-Montanes, Debora S. Marks, Tanya Laidlaw.

T.L. and K.B. conceived of the project. J.O.-M. advised on data analysis. D.D. supplied annotated single cell data objects. K.B. collected patient samples and advised on interpretation. T.D.G. organized cross-institutional sequencing, performed data analysis, and wrote the manuscript. D.M. and T.L. supervised the work.

The following authors contributed to Chapter 3: Giuseppe Torrisi, Yuge Ji, Luciana M. Luque, Debora S. Marks, Linus J. Schumacher, Chris Sander.

T.D.G. performed transcriptomics analysis and wrote the manuscript. Y.J. re-implemented the matrix decomposition algorithm in Python and advised on interpretation. G.T. was involved in extensive ongoing discussion throughout and performed analysis of spatial data with L.M.L. C.S. conceived of the project. L.J.S. and C.S. supervised the work with consultation from D.S.M.

TO MY PARENTS, WILLIAM GREEN AND LINDA DURAKIS, WHO HAVE SPENT MY LIFETIME
NURTURING MY CURIOSITY.

Acknowledgments

THIS DISSERTATION TOOK SEVEN YEARS, and I wouldn't be here finishing it without the support of many, many people, certainly more than I'm able to thank here.

First, academically: thank you to Debbie, for believing in me from the very beginning, and thank you Chris, for guiding me towards scientific maturity in the final years. Thanks also go to the many other faculty members who guided me, formally and informally, including past and present committee members: Allon Klein, Stefani Spranger, Alex Shalek, Caroline Uhler, and Galit Alter. Thank you to the other mentors I've had along the way: Tanya Laidlaw, Linus Schumacher, Nils Blüthgen, Suzanne Gaudet, and Allessandra Welker, and I'm sure others I'm forgetting to name. Copious thanks to Miao-Ping Chien for teaching me to think like a biologist. Thank you to my many wonderful labmates past and present, particularly my closest-cohort of Eli Weinstein, June Shin, Ada Shaw, Aaron Kollasch, Alan Amin, and Rose Orenbuch—you all made the lab such a wonderful place to be. Thank you to Stefan Peidli, my partner in single-cell analysis and always a true joy to work with. Thanks to Kelly Brock, who's continued to support me even after she left the lab, and the many other post-doctoral fellows, particularly Mafalda Dias, Jonny Frazer, Nikki Thadani and Benni Schubert. I joined the lab because I knew from the beginning that it was a safe place to take intellectual risks, and I never had to be embarrassed about asking questions—everyone in the lab helped make that true.

Thank you to everyone in the Biophysics program. Jim Hogle, Martha Bulyk, Venky Murthy, and Rachelle Gaudet: you made a Biophysics program that was such an open place for me to grow. And thanks always to Michele, who was there keeping the gears of the program turning and also always so deeply emotionally supportive of every one of us. Thank you to everyone in my cohort, especially Duluxan and Kevin—I really couldn't have done it without you. Beyond the Biophysics program, I found my scientific home in the Systems Biology department, and I'm grateful to everyone there for welcoming me in. The SysDevBio journal club—David, Debra, Hailey, Kalki, Gemma, Anna—really helped me through the middle years, and was a huge force in my thinking.

The deepest of thanks to everyone who made the Harvard Graduate Student Union and the finance and benefits committee possible: Lisa, Matt, Rachel, Ashley, Steffanie, Jenni, Belle, Cory, Bridger, Denish...I could so easily keep going: I can only hope that in the next stage of my life I can accomplish something half as meaningful as what we've built here. Thank you also to my many medical providers who believed me, supported me, and did everything they could to help, especially

everyone at Spaulding, and particularly Joseph Hanak, who sat down with me in 2018 and said he'd get me better and then did.

Thank you to all of my wonderful friends, particularly Grace, who's stayed within 5 miles of me since 2003, and Aina and Michael, who have kept close through their many moves. I feel so lucky to have gone through my PhD alongside Jinghui, Adam, Tej, and innumerable others. Thank you to Chris, for supporting me through the indecision of my early years, and Xuan, for signing up to be there through the final ones.

Thank you to my family: my parents, who supported me in every possible way through all of the ups and downs. My sisters: Amelia, thank you for following me to Boston, being my friend always, and helping me so, so much over the years. Sophia, thank you for trusting me to take care of you, and sharing in the hardest parts with me. And thank you too especially to my grandparents: Grandma Mary, for loving me no matter what; Papa, for making sure I knew he was proud of me; and Grandma Rhoda, for the mountain of books, letting me win at Scrabble and for the many, many phone calls. I wouldn't be the person and thinker I am today without you, and I wish so deeply that you were still here to see me in this moment of accomplishment.

O

Introduction

ORGANISMS ARE COMPOSED OF TISSUES, which are made of cells, which themselves are fundamentally bags of proteins, amino acids, and other miscellaneous chemicals. The biologist's task is to make sense of these complex, lipid-bilayer-wrapped objects—why they act the way they do, how they interact with each other, and how those interactions give rise to tissue and organ-level behavior.

This work uses single-cell transcriptomics to move towards understanding cellular function in

context by measuring the RNA in individual cells and using computational methods to contextualize those measurements. We use single cell transcriptomes to define cell states, and then study how those cell states change in response to interventions. High throughput single-cell RNA sequencing (scRNA-seq) relies on droplet microfluidics to physically separate, label, and then rejoin and sequence the transcriptome of individual cells^{105,134}. In general, only 10-20% of the transcriptome is amplified for sequencing⁸⁸. The resultant cell-by-gene matrices are very sparse, and interpreting which zeros are true absence of expression is complicated; similarly identifying to what extent observed heterogeneity is true difference between cells is challenging.

Even with perfect knowledge of the RNA in every cell, we still wouldn't have a perfect measure. The transcriptome is a snapshot of what cells are producing at a given point in time, and, in the case of single-cell, a very messy one. There is significant variation across experiments and even more across experimental protocols¹⁹¹. Ambient contamination due to dead cells is also common; this is most obvious when cells of a given type seem to be expressing conflicting markers, such as immune cells from a given sample showing epithelial markers only in a subset of batches. Methods for correcting this such as SoupX estimate background contamination using empty droplets, and correct for that contamination in the remainder of the dataset²¹⁷. Layering on corrections like this moves data further from the theoretical negative binomial generating function formed naturally from counts-based sampling, complicating the statistical assumptions underlying downstream analysis methods.

Despite these concerns, single-cell -omics methods have transformed scientific understanding of cell state, enabling the discovery of novel cell types and cell state dynamics¹³¹. Moving from these sparse matrices to interpretable findings about single cells has been the work of countless bioinformaticians; over the course of my PhD the field has matured significantly, with increasing numbers and kinds of analysis pipelines performing inference tasks that had not yet been devised when I started^{190,212,71}. This thesis is my small piece of this tidal wave of analytic methods and applications.

In this work, I provide two examples of transcriptomics applied to the study of disease. In Chapter 1 we consider aspirin-exacerbated respiratory disease (AERD), a condition historically called eosinophilic asthma which is actually a complex inflammatory syndrome driven by metabolic changes across many cell types²⁰⁹. We use scRNA-seq to look closely at a single cell type, discovering a proliferating plasma cell state with a possible role in nasal polyp formation in AERD. We also use bulk transcriptomics to investigate drug response in the sinus, and we bring scRNA-seq to that analysis to infer cell type ratios and gene expression in the observed cell types. This work shows that RNA-seq is capable of hypothesis-free cell state discovery and drug response mechanism determination. It also exemplifies some of its limitations: the layers of computation involved in coaxing meaningful results from sparse 20,000-dimensional matrices mean that some of those supposed findings do not experimentally replicate, illustrating the importance of robust statistical measures for evaluating cell states.

In Chapter 2 we move from defining states and their changes to interpreting when those cell states matter and how they interact with each other. Cancer is conceptualized as a monoclonal sub-organism but in truth has extensive genetic heterogeneity⁴⁷. Even cells in a tumor with the same genome can form a diversity of cell states¹¹⁴. There is yet more complexity within the tumor micro-environment (TME), where different immune cells act in concert to help or hinder tumor growth¹³⁵. In Chapter 2, we explore how cells of different types work together in triple-negative breast cancer (TNBC). Specifically, we interrogate changes in gene expression that correlate across multiple cell types, identifying a TNBC-specific pericyte sub-state that has a unique correlation with local immune cells. We also identify a correlated gene set in T cell and B cell subtypes in the TME which might result from an interaction mediated by IL-7; this interaction predicts poor response to neoadjuvant chemotherapy (NACT). The mathematical method used to detect these correlations, canonical correlation analysis (CCA), is a method for finding interpretable patterns in complex data. Even the CCA outputs require extensive interpretation, and going from those results

to statistically rigorous findings is highly nontrivial. At the end of the chapter, we have some interesting predictions, but we are unable to confirm them without additional data, and we also cannot provide clear confidence bounds of how likely they are to hold. Forthcoming experimental results and additional sequencing data from different but similar patients will be a clear opportunity to test these findings; however, the limitation of the initial exploration is clear.

This brings us, then, to the methods development work described in Chapter 3. The drug response examples in Chapters 1 and 2 are both highly heterogeneous; separating drug response from patient-patient variability isn't fully possible without considerably more data. We require statistical tools for quantifying when a cell has entered a new cell state, rather than just asserting a new cell type based on a higher-resolution clustering method. To develop and benchmark these tools, we needed data with less noise and fewer cell types. Thus, we looked specifically at single-cell perturbation datasets, unifying annotations and creating a data resource of broad use to the scientific community. Using this resource, we investigated a statistical measure for perturbation efficacy which operates as an effective distance between cell states. This method has been made available both as a standalone R and Python package and has also been incorporated into a larger perturbation-specific package as part of a multi-institution collaborative effort⁸⁶.

Although the primary focus of Chapter 3 is on whether perturbed cells are distinguishable, this mathematical tool is also useful for defining whether any cell state significantly differs from another. This method aims to improve the statistical rigor of cell state definition more broadly and enable clearer definitions of cell state transitions and cell subtypes. Together with the applications described in Chapters 1 and 2, it is a step forward in the analysis of single-cell transcriptomic data, and a step towards a more complete understanding of cell state, cell state transitions, and the role of cell-cell interaction in disease.

*When my attacks of breathlessness went on inexplicably,
long after my pleurisy had cleared up, my parents called
in Dr. Cottard. A doctor consulted in a case like this must
be more than just well versed. In the face of symptoms
which may be those of three or four different illnesses, the
thing that enables him to decide which of them he is most
likely to be dealing with, behind appearances that are very
similar, is ultimately his flair, the sharpness of his eye.*

Marcel Proust, In the Shadow of Young Girls in Flower

1

The sinus transcriptome in aspirin-exacerbated respiratory disease

1.1 ABSTRACT

ASPIRIN-EXACERBATED RESPIRATORY DISEASE (AERD) is a subtype of asthma characterized by aspirin intolerance, nasal polyp growth, and chronic rhinosinusitis. The immunological mechanism underlying the connection between these symptoms and response to AERD-specific treatments is not fully understood. Here, we used single-cell transcriptomics to investigate the role of B cells in nasal polyps, identifying a previously unobserved population of proliferating plasma cells. We also use bulk RNA-seq to study changes in the sinus transcriptome during and after aspirin desensitization, finding changes in expression of genes involved in barrier maintenance and immune cell function. In so doing, we show that the easily accessible inferior turbinate can be used as a marker tissue for drug response even though polyps are only formed in the less accessible ethmoid sinus.

1.2 INTRODUCTION

AERD is distinguished by the combination of asthma, chronic rhinosinusitis, nasal polypsosis, and aspirin hypersensitivity¹⁸⁴. It generally appears first as adult-onset persistent rhinitis (29.7 ± 12.5 years) followed by other AERD symptoms¹⁹². The condition is sometimes referred to as eosinophilic asthma due to infiltration of the lungs and sinuses with eosinophils¹⁸⁴. However,

most AERD patients are not predisposed to atopy, and of the 30% who are, asthma and rhinitis sometimes appear without associated aspirin sensitivity earlier in life¹⁹². Although the symptoms of an aspirin reaction mimic an allergic response, unlike in atopic disease the reaction is not immunoglobulin E (IgE) mediated¹⁸⁴. AERD progression also differs from that of atopic asthma: nasal polyps recur rapidly after surgery, sometimes within a few weeks²⁰⁹. Asthma in AERD patients is also more likely to be severe than in other asthmatics²⁰⁹.

Medical record reviews show that 3-5% of asthmatics are known to have aspirin hypersensitivity; 3-15% are predicted to have hypersensitivity reactions if challenged¹⁵⁴. An estimated 1.3 million patients in the US have AERD²⁰⁹. The tendency toward atopy and associated high risk of asthma is highly heritable, but AERD is not; a family member with aspirin hypersensitivity is only seen in 6% of AERD patients¹⁵⁴. Nonetheless, genome-wide association study (GWAS) analysis has identified several genes associated with increased risk of developing AERD, though results across studies have been inconsistent⁴⁸. Many of the identified genes are involved in arachidonic acid metabolism and signaling⁴⁸.

This hints at the arachidonic acid metabolism dysregulation at the heart of AERD (Figure 1.1). Lipoxin A₄ (LXA₄) acts to decrease leukotriene production, reduce pulmonary eosinophils, and is an antagonist of cysteinyl leukotriene receptor 1 (CysLT₁R). Cysteinyl leukotrienes (CysLTs), which are elevated in AERD patients due to increased leukotriene C₄ synthase (LTC₄S) and decreased LXA₄, trigger bronchoconstriction, airway mucous production, and eosinophil migration. PGE₂ receptor 2 subtype (EP₂), which has decreased activation in AERD due to the reduction in prostaglandin E₂ (PGE₂), acts to reduce leukotriene production, eosinophil migration and fibroblast proliferation. Thromboxane A₂ (TXA₂), increased in AERD, acts to increase bronchoconstriction and decrease LTC₄S activity. The binding partners of prostaglandin H₂ (PGH₂) then drive bronchoconstriction and chemotaxis of eosinophils, basophils, and type 2 innate lymphoid cells (ILC2s), triggering swelling and edema of respiratory tissues¹¹¹.

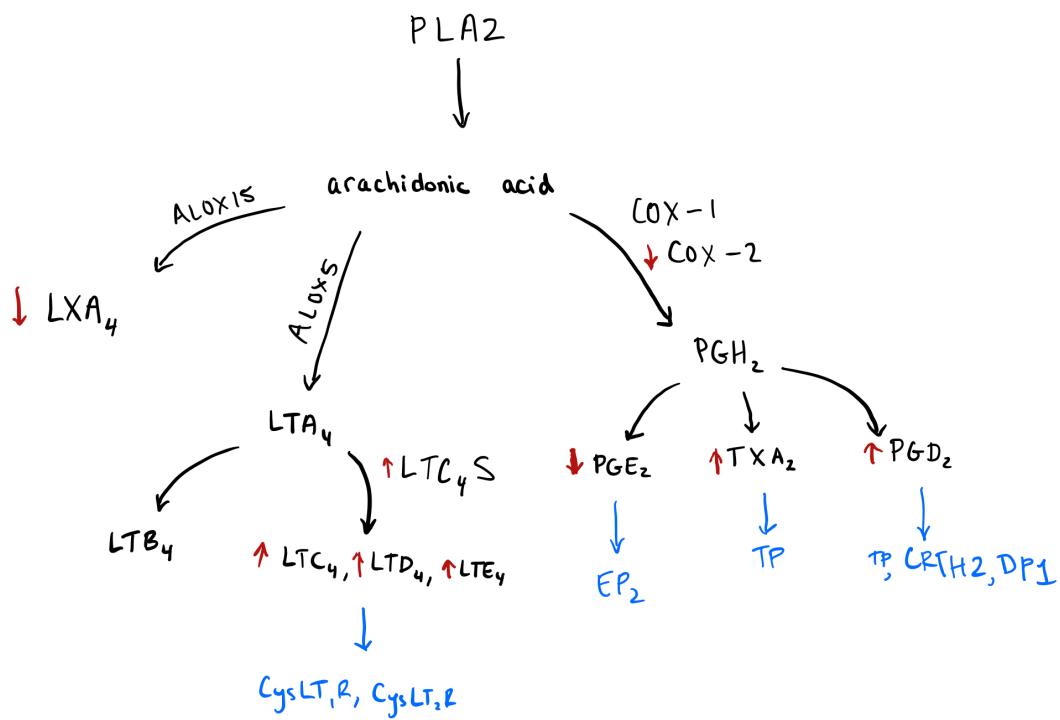


Figure 1.1: Arachidonic acid metabolism is dysregulated in AERD. Red arrows denote changes in levels or activity in AERD. Receptors and ligand-receptor interactions are denoted in blue. COX-1 and COX-2 refer to cyclooxygenase 1 and 2 respectively. LTA₄ is an unstable product. Simplified from ¹¹¹.

In total, the metabolic changes in AERD result in increased tissue eosinophils and, to a lesser extent, other immune cells, as well as increased bronchoconstriction and airway mucous production—in short, eosinophilic asthma. Even outside of respiratory tissues, blood eosinophils are also significantly elevated relative to healthy people and other asthmatics⁴⁴. The leukotriene elevation is similarly not isolated to the respiratory system; urinary leukotriene E₄ (LTE₄) in particular is higher in AERD patients than in those with aspirin-tolerant asthma⁴⁴. Why these metabolic changes result in nasal polyp growth, and how aspirin desensitization acts to reduce polyp growth and asthma severity, is a subject of ongoing research, including the work presented here.

AERD, like many immune diseases, is characterized by irregular interactions across a variety of cell types. The arachidonic acid metabolism changes described above are well established, but the roles of individual cell types in AERD disease processes are the subject of considerable ongoing work. It has been established that during aspirin reaction and at baseline, AERD patients have increased levels of effector cells expressing LTC₄S¹¹². These cells release leukotrienes, which trigger an innate type II immune response in epithelial cells. These epithelial cells release mast cell activating mediators which activate mast cells. Factors released by the mast cells signal smooth-muscle constriction and airflow obstruction, as well as recruitment of more eosinophils and basophils, type 2 T helper cells (Th₂ cells), and ILC2s. Single-cell sequencing and other recent research have added additional layers to this picture. Prior work revealed AERD-specific plasma cells that appeared to be long-lived in nasal polyps¹⁵⁰. Among other unique transcriptional signatures, AERD long-lived plasma cells expressed IL-5 receptor alpha (IL₅R α), which was hypothesized to drive the long-term survival of these cells in nasal polyps²⁷. Here, we perform further analysis of those plasma cells, with the aim of more deeply characterizing their functional role in maintaining aberrant inflammation in the AERD sinus.

The aspirin hypersensitivity reaction mimics allergic response symptomatically, but AERD patients do not have antibodies to aspirin, and the molecular mechanisms are distinct¹⁸⁴. As can be

seen in Figure 1.9, aspirin depletes the mast cell stabilizer PGE₂, which is believed to trigger the hypersensitivity reaction³⁷. The hypersensitivity reaction is also driven by increased CysLTs relative to non-AERD individuals¹⁸⁴. The rapid rise in CysLT levels triggers bronchoconstriction, airway mucous production, and eosinophil migration: congestion and an asthma attack. However, why aspirin's cyclo-oxygenase inhibition triggers CysLT production is unknown, as is what causes the irregular response in AERD³⁷.

A single instance of aspirin reaction protects against future ones: when patients were challenged with 325 mg of aspirin (inducing a reaction), the same challenge on the subsequent day did not have an effect¹⁸⁶. This protection can be harnessed via aspirin desensitization, in which patients continue to take daily aspirin. As long as the treatment is maintained, patients can avert hypersensitivity reactions. If patients take a higher dose than is required to maintain desensitization (referred to as high-dose aspirin therapy), polyp growth is slowed and overall AERD symptoms reduced²⁰⁹. This treatment triples the mean interval to repeat surgery from 3 years to 9 years and reduces the frequency of sinus infection and of hospitalization for asthma¹⁸⁵. The mechanism by which high-dose aspirin slows polyp growth is not fully elucidated, but there is some explanatory evidence. Long-term treatment with aspirin has been shown to reduce the level of prostaglandin G2 (PGD₂) relative to a fixed LTC₄ to PGE₂ ratio¹⁹. Long-term treatment is associated with reduced PGD₂ and LTC₄, as well as lower levels of eosinophils and basophils in tissue, with no associated change in CysLTs^{36,209}. As can be seen in Figure 1.1, these changes move arachidonic acid metabolism closer to healthy, non-AERD behavior.

Although high-dose aspirin therapy is highly efficacious, many patients are unable to tolerate it due to gastrointestinal side effects¹¹⁶. Identifying alternative therapies for AERD that maintain the improved arachidonic acid metabolism while reducing side effects is an area of active research. Our work here investigating the mechanism of aspirin desensitization will help point toward future, more targeted treatments. This study of the transcriptional response to aspirin desensitization

was also a pilot for using bulk RNA-seq time course samples of the inferior turbinate as a marker for drug response in AERD. Samples of the ethmoid sinus, where nasal polyps develop, are only available during surgery. By using the more easily accessible inferior turbinate, we can collect more samples per patient, and achieve a higher-resolution view of epithelial and immune function in the sinus. Bulk RNA-seq lacks the single-cell or single-cell-type resolution of single-cell RNA sequencing (scRNA-seq), but by using existing scRNA-seq from the inferior turbinate we can still say something about cell type fractions and cell type-specific gene expression ¹⁴⁸. In total, the work presented here shows the scope of applying transcriptomics to hypothesis-free investigation of a complex immune disease.

1.3 PROLIFERATING B CELLS IN NASAL POLYPS

1.3.1 METHODS

Data was collected as described in ¹⁵⁰. An R object of the cell-by-gene matrix for B cells was provided by the authors of ²⁷. Data was processed and normalized using Seurat v3.1.3. Plasma cells from 6 patients were used: 3 with AERD and 3 with chronic rhinosinusitis with nasal polyps (CRSwNP). Cells were kept with at least 200 – 2000 nFeature_RNA and < 10 % mitochondrial reads (Figure 1.2). Counts were log normalized and 1500 variable features were computed using `FindVariableFeatures` with selection method “vst”. Data was scaled with `ScaleData` prior to principle component analysis (PCA). The first 15 principle components (PCs) were used for neighbor identification. Louvain clustering was performed using resolution 0.25. Cell cycle state identification was performed using Seurat `CellCycleScoring` with default gene lists. Differential expression testing used Seurat `FindMarkers` with default parameters. To assign κ or λ class, the fractions of observed reads with names beginning with IGK and IGL were computed, and then the larger fraction was used to assign class identity. Genes used to identify B cell states are in Table 1.1.

	B cell progenitor	Pre B	Mature B cell	Activated B cell	Plasma cell
CCND ₃	HI	HI	LO		LO
OCA ₂	HI	LO	LO	LO	
CD69	LO	LO	HI	MID	
IGLL ₁	MID	HI	VERY LO	VERY LO	VERY LO
PRDM ₁	MID	MID	LO	LO	HI
RELA	LO	LO	MID	MID	HI
XBP ₁	LO	MID	LO	MID	VERY HI
IGKC	LO	LO	MID	MID	VERY HI
SPIB	HI	HI	MID	MID	VERY LO
SDC ₁	LO	LO	LO	LO	VERY HI
PAX ₅	HI	HI	HI	HI	VERY LO
IRF ₄	LO	LO	LO	LO	HI
CD8 ₃	MID	MID	HI	HI	LO
BCL6	MID	MID	HI	LO	VERY LO
BACH ₂	MID	MID	HI	MID-HI	LO
IRF8	MID	MID	HI	MID	LO
CD19	MID	MID	VERY HI	LO	VERY LO
IL5RA	MID	MID	LO	LO	HI

Table 1.1: Marker genes for B cell substates as supplied by Tanya Laidlaw. HI refers to high expression, MID to medium expression, LO to low expression. Cells are left blank where expression is not clearly defined for the relevant cell type.

1.3.2 RESULTS

Louvain clustering was used to identify 5 clusters (Figure 1.4). Clusters 0 and 1 are a κ -switched subpopulation; clusters 2 and 3 are λ -switched (Figure 1.5). The kappa-lambda ratio varied from patient to patient but did not differ by diagnosis (Figure 1.6). This suggests that although κ or λ identity is a significant source of variation in the data, it is not a feature relevant to AERD pathology.

The λ cluster has notably high levels of immunoglobulin lambda like polypeptide 1 (IGLL₁), generally recognized as a marker of early B cell development⁴. While this would suggest that these cells were pre-B cells, IGLL₁ is actually expressed, albeit at a lower level than in pre-B cells, in plas-

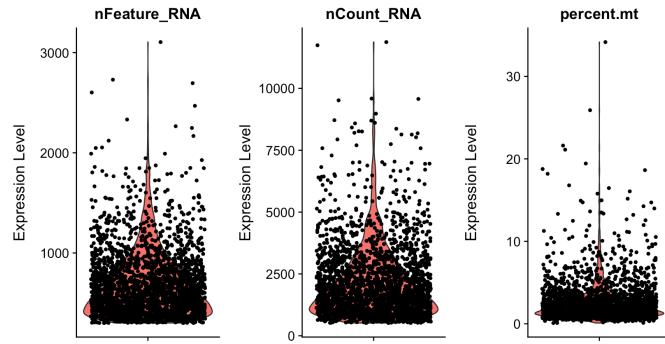


Figure 1.2: Quality control plots for B cells from surgical samples. nFeature_RNA is the number of genes detected in each cell. nCount_RNA is the number of counts detected in each cell. percent.mt is the percentage of counts from mitochondrial RNA. These plots were used to establish cutoffs of 200 – 2000 for the number of features per cell and < 10% mitochondrial reads. Cells outside of these ranges were excluded from further analysis.

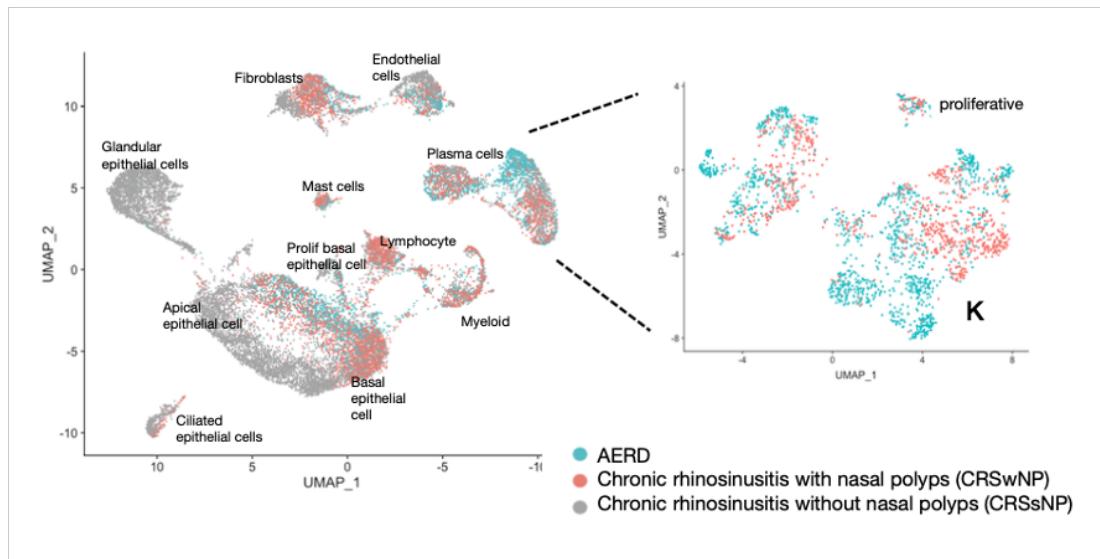


Figure 1.3: UMAP projection of scRNA-seq from ¹⁵⁰. Cells are colored by diagnosis. The inset shows B cells re-analyzed here. κ and λ class-switched cells form two separate clusters; a third cluster is characterized by higher levels of proliferation markers.

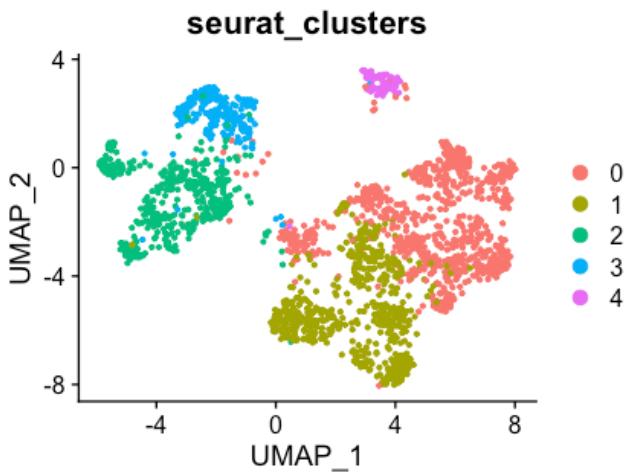


Figure 1.4: Louvain clustering of B cells, displayed on UMAP coordinates. 5 clusters are apparent.

mablasts¹⁴².

The small cluster (Cluster 4) was characterized by high expression of baculoviral IAP repeat containing 5 (BIRC5), marker of proliferation Ki-67 (MKI67) and histone Cluster 1 H2A Family Member J (HIST1H2AJ), present in nearly half the cells in the proliferating cluster and fewer than 1% of other cells (Supplemental Table 1.1). These genes are all involved in cell proliferation. Moreover, cell cycle calling on all B cells revealed that the cells in the proliferating cluster were much less likely to be in G1 state, suggesting active proliferation (Figure 1.7). These proliferating cells were common across all polyps and were not specific to AERD.

Of the markers of interest suggested by our collaborators (Table 1.1), only IL5R α was significantly different across diagnoses (Figure 1.8). This gene had a log-2-fold change (L2FC) of 0.52 and was expressed in 24% of cells from AERD patients and only 11% of cells from CRSwNP patients. Detailed analysis of the features unique to AERD plasma cells in these samples, as well as additional experimental findings, is available in²⁷.

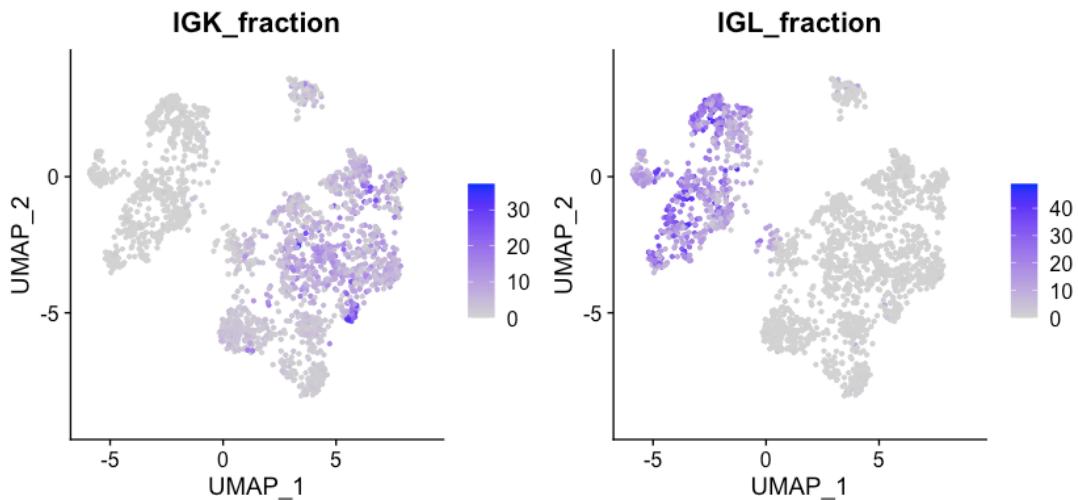


Figure 1.5: Major clusters are separated by κ and λ identity. IGK and IGL fractions are the percentages of total reads originating from genes starting with IGK or IGL respectively.

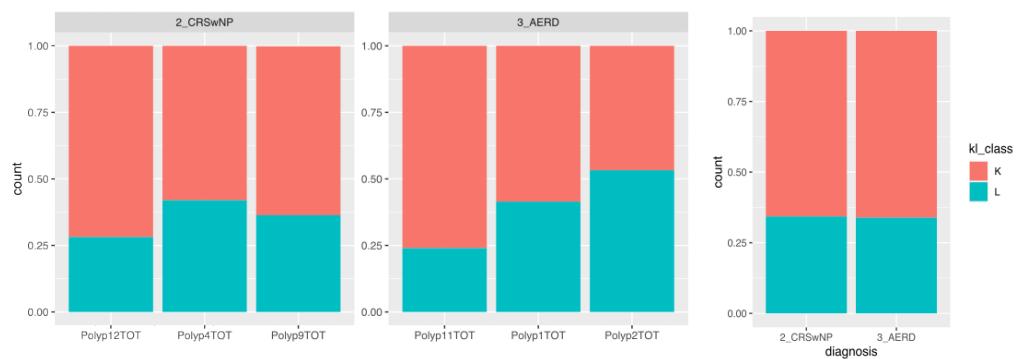


Figure 1.6: KL ratio for each sample (left) and diagnosis (right). Although the ratio varies across patients, it does not change with diagnosis.

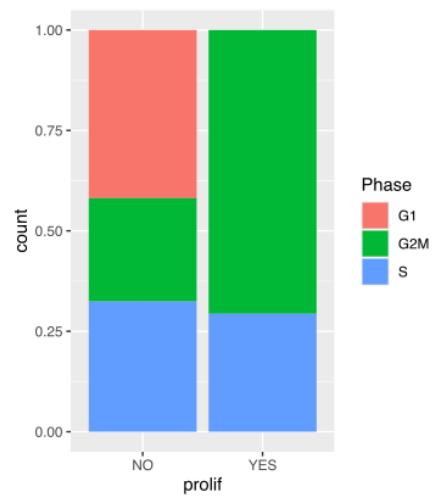


Figure 1.7: $\kappa - \lambda$ ratio for each sample (left) and diagnosis (right). Although the ratio varies across patients, it does not change with diagnosis.

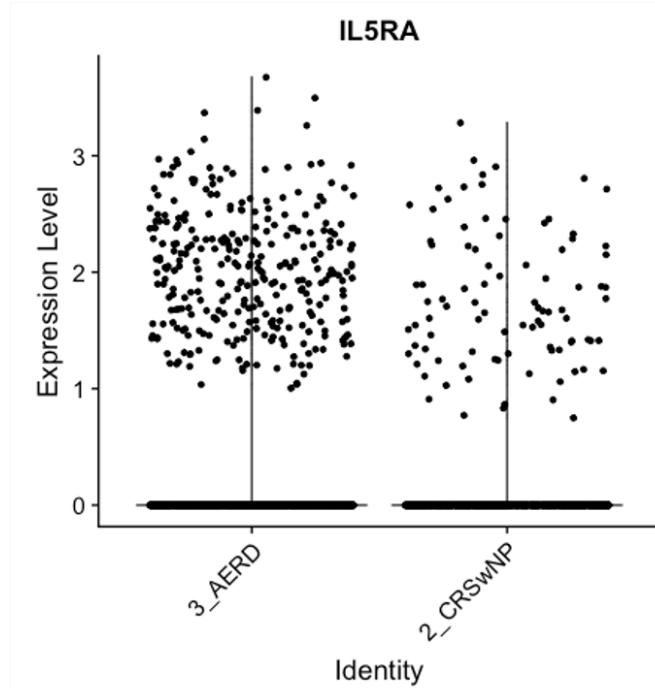


Figure 1.8: Expression of IL₅R α in plasma cells from patients with AERD and CRSwNP.

1.4 THE INFERIOR NASAL TURBinate AS A MARKER TISSUE FOR DRUG RESPONSE

1.4.1 METHODS

Nasal swabs of the inferior turbinate were collected from 9 patients and frozen as previously described¹⁵⁰. Samples were taken prior to aspirin desensitization (“control”); one hour into an aspirin hypersensitivity reaction (“acute”); after 8 weeks of high-dose aspirin, 625 mg 2x/day (“8 weeks”); and two weeks into switching from high-dose to medium-dose aspirin, 325 mg 2x/day (“10 weeks”). Strand-specific whole transcriptome sequencing was performed at the Broad Institute Sequencing Core. The sequencing facility provided .bam files with reads STAR-aligned to GRCh37. Due to facility error, a significant portion of samples were damaged and had very low reads. Quality control metrics for all samples are provided in Supplemental Table 1.2. Samples with fewer than 30M reads were excluded from analysis.

Gene calling was performed using GenomicAlignments (v. 1.28.0) SummarizeOverlaps with mode “Union”, strand specificity True, and fragments True. Differential expression testing used DESeq2 (v 1.32.0) with default parameters⁵. Genes with fewer than 50 total reads were removed prior to DESeq for the all-sample analysis. The design formula used was ~ timepoint + patientID. Olfactory receptor family 51 subfamily P member 1 pseudogene (OR51P1P) did not converge under negative binomial Wald testing and was excluded from further analysis.

Cell type fraction and cell-type-specific gene expression prediction were performed using the CIBERSORTx web platform with default parameters¹⁴⁸. The single-cell reference was created using publically available sinus scrape samples¹⁵⁰, labeled following the procedure described. The single-cell reference and the bulk counts were both normalized to 10k counts per cell and per sample respectively prior to upload. The bulk RNA-seq samples were subset to genes included in the single-cell data (named genes only).

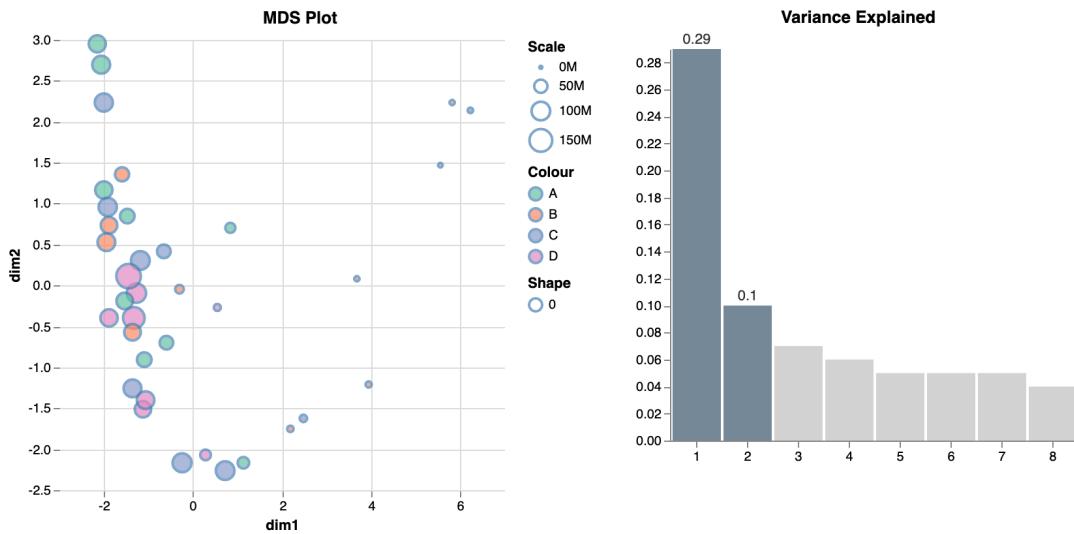


Figure 1.9: multidimensional scaling (MDS) plot for all samples shows read depth predominates. Points are scaled by the total number of reads and colored by condition. Number of reads is represented by the scale of the dots; small, less, sequenced dots are separated by MDS dimension 1. A: control, B: acute, C: 8 weeks, D: 10 weeks.

1.4.2 RESULTS

Variance in the data was dominated by the total number of reads (Figure 1.9), with 29% of the variance of the data explicable by the first MDS component, which, as can be seen in the figure, is higher for samples with fewer reads. Based on this and the splitting of MA plots after DESeq, data was subset to samples with at least 30M reads prior to differential expression testing. This number was selected by sequentially removing the lowest-depth sample and examining the MDS plot; the lowest possible cutoff which did not expose a sequencing-based outlier was selected (Figure 1.10). The time points and patient IDs for samples used for analysis are shown in Table 1.2.

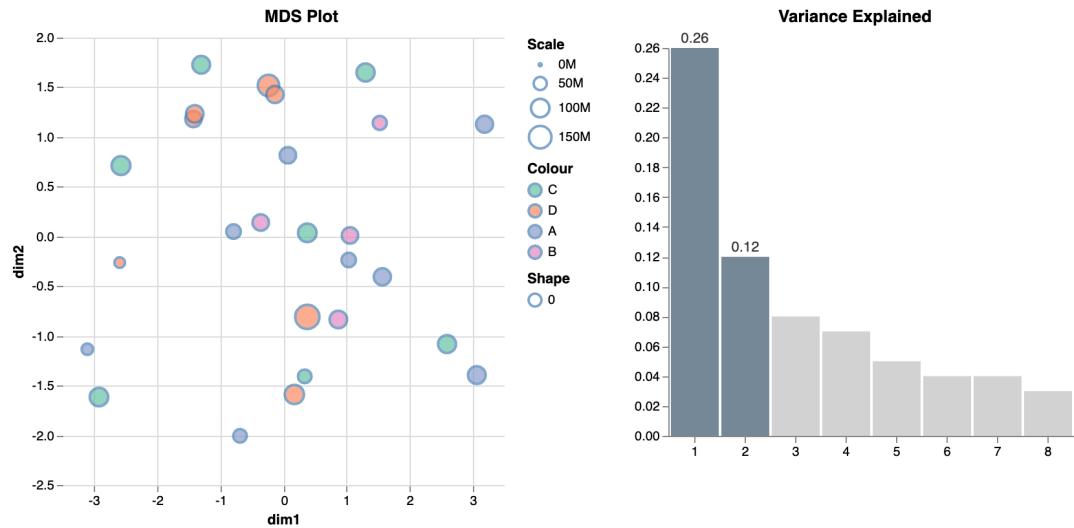


Figure 1.10: MDS plot for samples with at least 30M reads shows that outliers were successfully removed. A: control, B: acute, C: 8 weeks, D: 10 weeks.

Patient ID	Control	Acute	8 weeks	10 weeks
913			X	X
914	X	X		X
916	X			
917	X		X	
918	X		X	X
920	X		X	X
921	X	X	X	X
922	X	X	X	X
923	X	X	X	X

Table 1.2: High-quality samples available for each patient

1.4.3 DECONVOLUTION OF CELL TYPES

To investigate how immune infiltration changes with aspirin exposure, we used CIBERSORTx to predict cell type fractions. Unsurprisingly, the vast majority of samples were dominated by differentiating and secretory epithelial cells (Figure 1.11). The fraction of immune cells present did not change significantly with aspirin treatment. Two samples had significant neutrophil presence predicted, but these two samples were collected from different patients at different time points (Patient 918 control and Patient 920 8 weeks). We also used CIBERSORTx digital cytometry to predict which cell types expressed the observed genes; this was not possible for all genes discussed, but results are noted where relevant and are available in full in Supplemental Table 1.3. For less prevalent cell types, the genes predicted as originating from these cell types appeared sometimes inaccurate; for example, the gene predicted as most predominant in mast cells and eosinophils was keratin 5 (KRT5), which is in truth expressed almost entirely by basal epithelial cells¹⁰¹.

ACUTE ASPIRIN HYPERSENSITIVITY REACTION

Due to the small number of high-quality samples from the acute reaction, our ability to identify differentially expressed genes was limited and interpretation challenging. Our analysis identified 21 significantly upregulated and 5 significantly downregulated genes (see Supplemental Table 1.4). Some of these are likely artifacts of the low patient counts; IL-22 has an adjusted p-value of 0.002 but is only appreciably present in one sample (Figure 1.13). CIBERSORTx predicts that IL-22 presence is predominantly due to expression by neutrophils, which varied in their observed fraction between samples (Figure 1.12). There is some evidence of expression of IL-22 in neutrophils in other studies⁵⁹. It's also a known regulator of neutrophil recruitment, so it's possible that the observed correlation with neutrophil levels stems from epithelial expression driving neutrophil recruitment¹⁵⁵. With the data available here, we aren't able to evaluate whether the relationship between IL-22 and

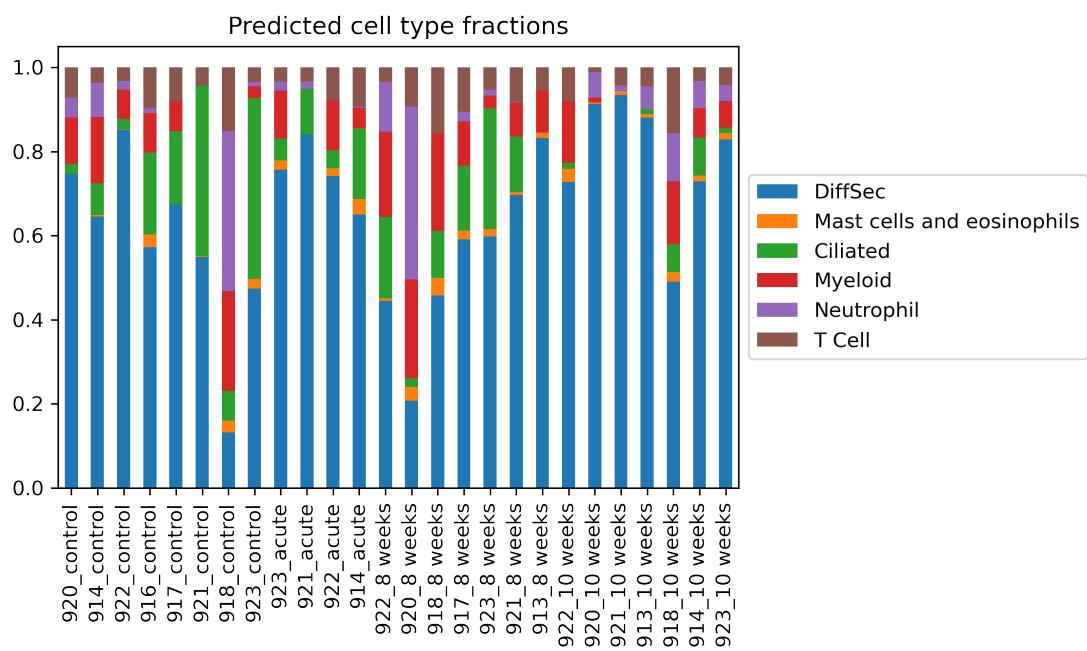


Figure 1.11: Scrapes consisted predominantly of epithelial cells. Cell type fractions predicted using CIBER-SORTx¹⁴⁸. Samples are labeled as 'patient ID_timepoint'.

IL22 counts are higher in samples with larger predicted neutrophil fraction

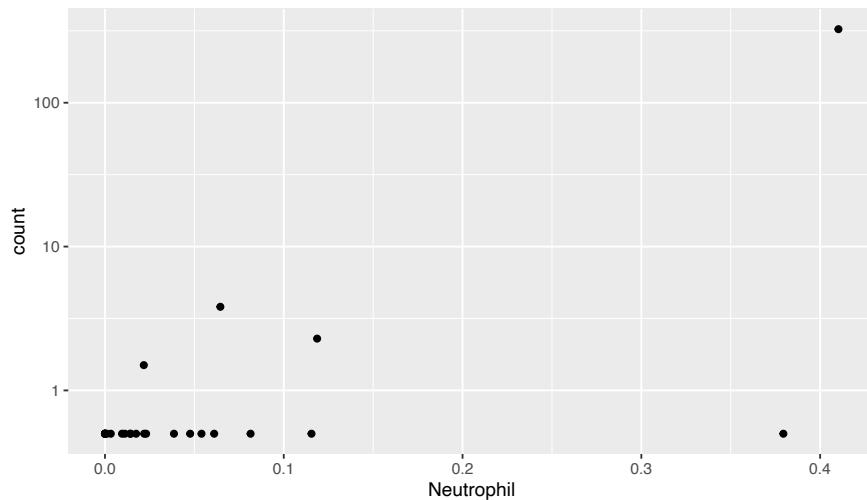


Figure 1.12: IL-22 counts and predicted neutrophil level are somewhat correlated across all samples (Pearson = 0.68).

neutrophil presence involves neutrophil expression of IL-22. Regardless, it's clear that the observed reduction in IL-22 during hypersensitivity reaction is likely an artifact of low patient counts and high patient-patient variability.

We observe a 4-fold decrease in receptor activity modifying protein 1 (RAMP1) during the acute reaction (Figure 1.14, adjusted p=0.02). This protein has been previously shown to be dysregulated in asthmatic epithelium²⁰. Ajuba lim protein (AJUBA) is increased by a factor of 2.23 (Figure 1.15, adjusted p=0.04). This protein is a key regulator of response to hypoxia and may be indicative of an immunological response to aspirin-induced hypoxia²².

In some cases, related genes or subunits of the same gene are changed similarly, indicating an effect that is likely reproducible despite the limited available data. Growth hormones GH₁ and GH₂ are both decreased by a factor of 30 during the acute reaction (Figure 1.17, adjusted p= 0.01, 0.01). CIBERSORTx predicts GH₁ and GH₂ are predominantly expressed by neutrophils across all samples; relative expression by neutrophils:T cells:myeloid cells was 4:3:2 for GH₁ and 4:3:1 for GH₂.

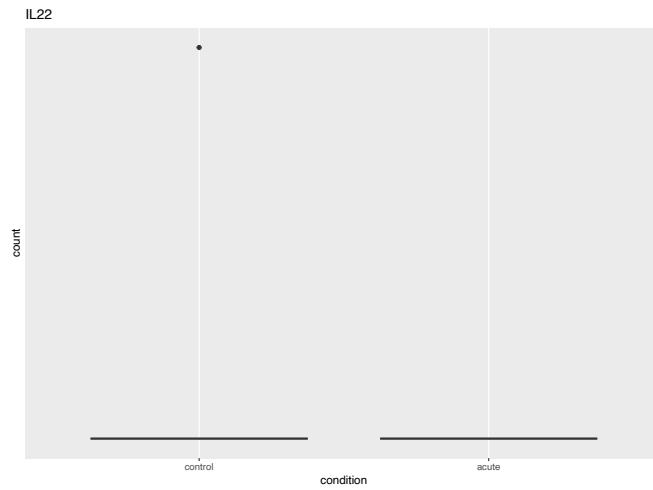


Figure 1.13: Although IL-22 is significantly different during the acute reaction, this finding is likely not meaningful.

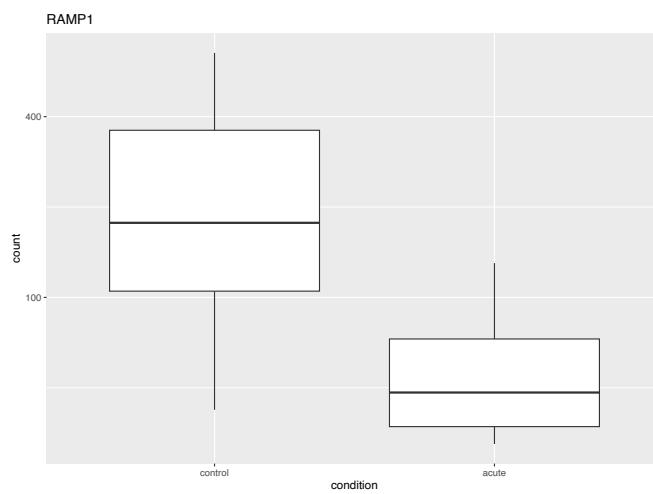


Figure 1.14: RAMP1 is downregulated during acute aspirin reaction (adjusted $p=0.02$).

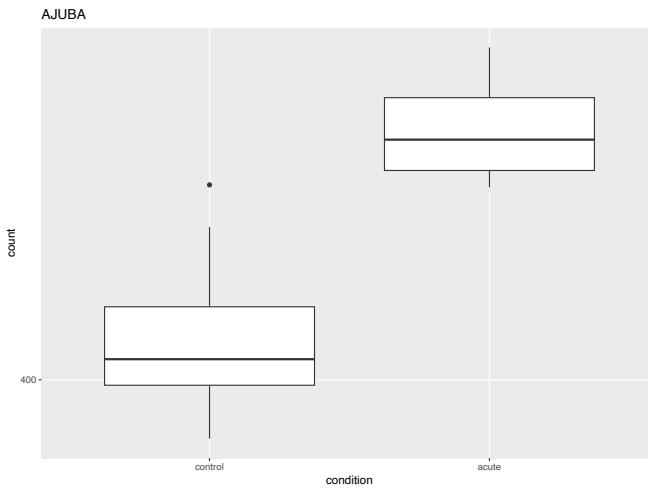


Figure 1.15: Ajuba is upregulated during acute aspirin reaction (adjusted p=0.04).

While expression by these cell types is possible, it also may be that GH₁ and GH₂ are acting as chemoattractors and stimulating the local immune response¹³⁶, which then results in the observation of more of these cells in our samples. Single-cell sequencing, sequencing of immune cell subsets or proteomic investigation would be required to confirm these findings.

Amylase alphas AMY₁A, AMY₁B, and AMY₁C all decreased by a factor of 4 (Figure 1.16, adjusted p= 0.09, 0.14, 0.12). Although these p-values are above our preferred significance threshold, the shared change is suggestive of a real shift in the level of amylase alpha 1 proteins. While the known function of amylase is in the digestive system, one prior study found that amylases are also present in the nasal mucosa¹⁹⁴; the function of these proteins in this tissue is not known¹¹⁰, but they are involved in cell proliferation and differentiation in intestinal mucosa⁵⁰. CIBERSORTx predicts that all three are being produced by ciliated epithelial cells, which is likely inaccurate, but may indicate that secretory cells are only producing amylase when proximal to ciliated epithelial cells; it may also indicate misclassification of rare serous cells in the tissue. Regardless of the cell type of origin, the consistent reduction in all three proteins suggests a role for alpha-amylases in the AERD aspirin hypersensitivity reaction. Determining whether that role is contributing to symptoms or

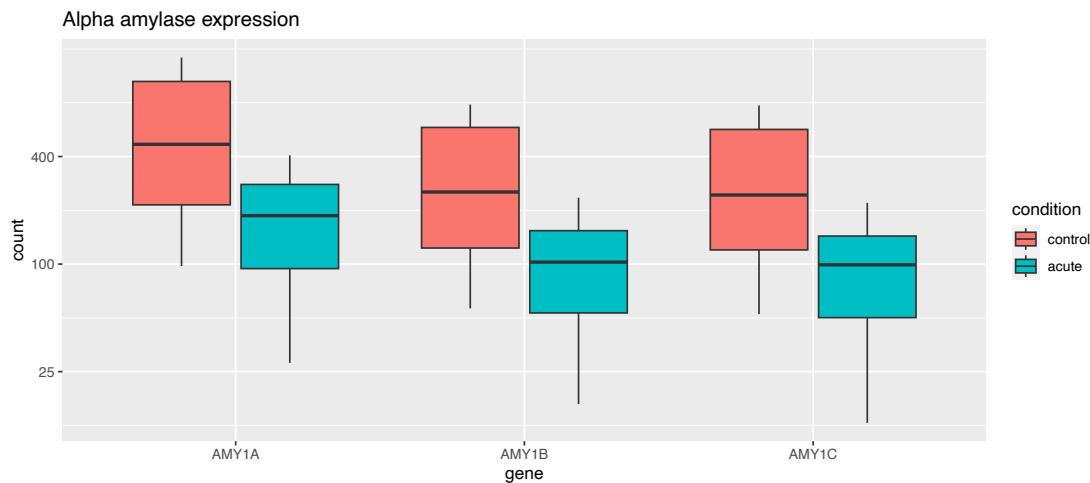


Figure 1.16: Three different alpha-amylases are decreased during acute aspirin reaction.

merely reflects a shift in cellular priorities requires experimental follow-up.

In total, these findings suggest that it is possible to observe aspirin hypersensitivity reaction changes to the transcriptome of the inferior turbinate. Additional sequencing of more patients would be necessary to characterize this change and identify robust findings on reaction pathology.

1.4.4 ASPIRIN DESENSITIZATION

8 weeks of high-dose aspirin therapy had only a small observable effect on the inferior turbinate transcriptome. At an adjusted p-value threshold of 0.10, there was one down-regulated gene (VSTM2L, $p=0.02$), and 5 upregulated ones (ENSG00000226698, $p=1.87E-15$; IGHV1-3, $p=1.27E-13$; ENSG00000257142, $p=2.31E-05$; C1orf68, $p=0.03$; TSHZ3, $p=0.05$). After 8 weeks of treatment, patients were switched from high-dose to medium-dose aspirin, and after two weeks on the medium dose samples were collected again (10 week time point). At an adjusted p-value threshold of 0.10, there were 22 significantly differentially expressed genes at the 10 week timepoint relative to pre-treatment samples. The full results are available in Supplemental Table 1.5.

Genes that significantly increased at one or both time points are shown in Figure 1.18. Some

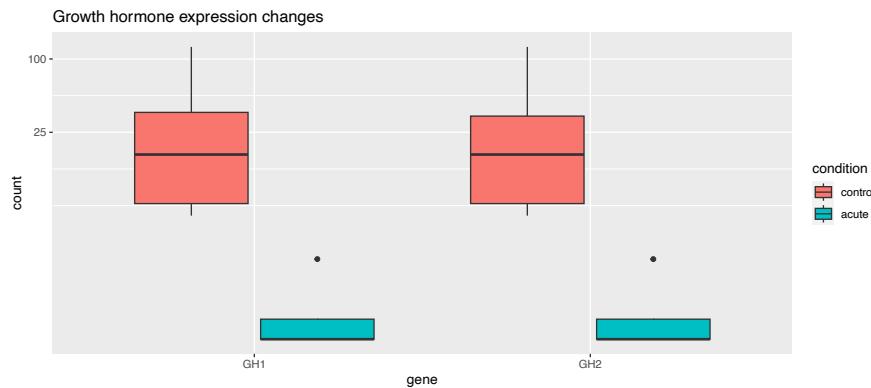


Figure 1.17: Growth hormones GH₁ and GH₂ have matching decreases during aspirin reaction.

of these genes suggest improved health after desensitization. C1orf68, also known as KPLCE, is a skin-specific protein believed to be involved in maintaining barrier function⁶⁰. Function-decreasing mutations in C1orf68 have been shown to significantly increase the risk of candidaemia¹⁰⁹. Its increase here suggests successful barrier improvement due to aspirin therapy and may implicate barrier dysfunction in the pathogenesis of AERD. Others are less clear: immunoglobulin variable regions IGHV₁₋₃ and IGKV₂₋₂₈ suggest increased levels of B cells, but are difficult to interpret without additional experiments; moreover, mapping of non-targeted RNA-seq to these regions is often unreliable. Prior work found that the solute carrier (SLC) protein SLC₅A₅ is elevated in nasal samples from atopic asthma patients relative to healthy individuals⁶⁶, and that SLC₂₆A₄ is increased in nasal polyps of CRSwNP patients¹⁷⁸. Given this, the fact that these proteins are elevated after desensitization is surprising, but may indicate an AERD-specific gene expression profile in the inferior turbinate. Both of these SLC genes are iodide transporters, and their function is more extensively studied in the thyroid²¹⁵. CIBERSORTx predicts that SLC₅A₅ is expressed by differentiating and secretory cells; expression of SLC₂₆A₄ was not sufficient in the single-cell data to predict cell type here, but its antisense RNA (SLC₂₆A₄-AS1) is also predicted to be expressed by differentiating and secretory cells. Why the expression of these transporters changes, and whether this has a meaningful

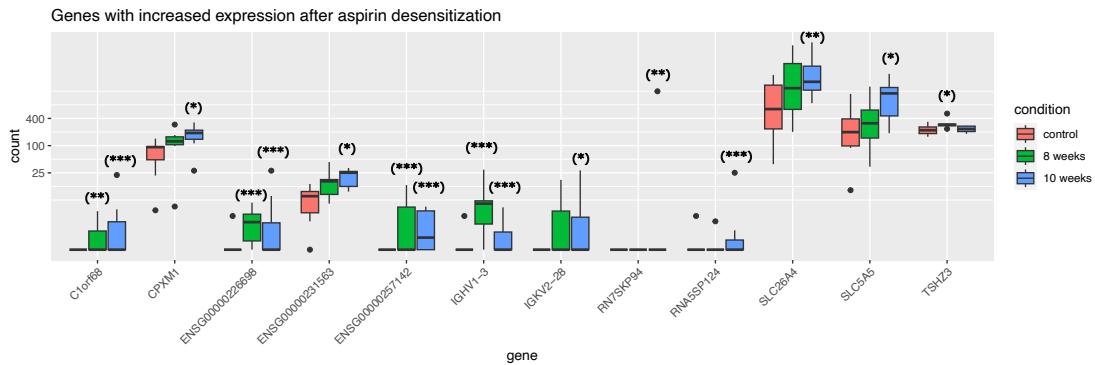


Figure 1.18: Genes which had increased expression after aspirin desensitization. Adjusted p-values for individual comparison from control using DESeq are denoted using asterisks (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$)

effect on pathology, requires further investigation.

Genes that decreased at one or both time points are shown in Figure 1.19. While the exact function of ALS2CR12, also known as flagellum associated containing coiled-coil domains 1 (FLACC1), in the sinus is not known, mutations to ALS2CR12 in skin cells conveys increased risk of basal cell carcinoma and cutaneous squamous cell carcinoma^{172,182}; its decrease here may be suggestive of a role in reduced cell growth in the ethmoid sinus. CXorf22, also known as cilia and flagella associated protein 47 (CFAP47), is a ciliated flagellar protein¹²¹ and is correctly predicted by CIBERSORTx to originate in ciliated cells. Tubulin tyrosine ligase-like 9 (TTLL9) similarly is predicted to originate in ciliated cells and has a microtubule regulating function in cilia¹⁰⁸. The decrease in these proteins may indicate reduced activity of ciliated cells after desensitization. DNER is a notch ligand shown to modulate IFN γ levels in the lung in models of chronic obstructive pulmonary disease (COPD)¹¹. Its decrease here may be connected to healthy changes in inflammation and inflammatory signaling cascades associated with desensitization. IL-5R α is expressed by many cells in AERD; here, CIBERSORTx predicts it is primarily expressed by ciliated cells. Prior work has shown that IL-5 inhibition can successfully treat AERD²⁹; the decrease in levels of the receptor here may indicate reduced IL-5 signaling contributes to improved symptoms after desensitization.

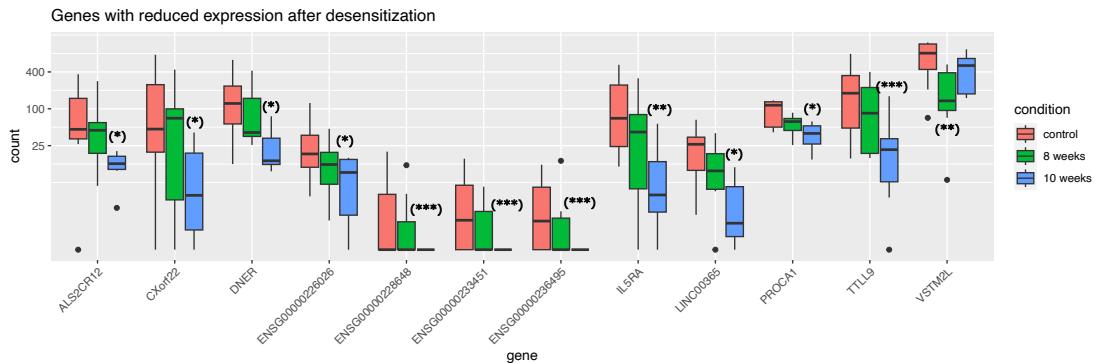


Figure 1.19: Genes which had decreased expression after aspirin desensitization. Adjusted p-values for individual comparison from control using DESeq is denoted using asterisks (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$)

DESeq considers each gene independently and does not have built-in functionality for testing changes across multiple time points. There were a number of additional genes with significant unadjusted p-values at both 8 weeks and 10 weeks, with annotations suggesting possible involvement in desensitization. A few of these genes are highlighted in Figure 1.20. Significantly, prior work found that expression of transcription factor 7-like 2 (TCF7L2) in nasal brushes is associated with asthma remission¹⁶²; our observed increase in TCF7L2 after aspirin desensitization concurs with that finding. Another gene with increased expression after desensitization is ATG16L1, an autophagy-associated gene that is a known regulator of intestinal inflammation¹⁰⁰; it may also be regulating the reduced inflammation here. Nucleotide-binding oligomerization domain (NOD1) initiates inflammation responses and has previously been shown to be involved in asthma⁸⁹; it's plausible that this gene is also regulating inflammation here, though less clear why its level would increase. SMARCD1, a chromatin remodeling protein increased after desensitization, is known to be involved in steroid response in asthma¹³⁸. A single nucleotide polymorphism (SNP) in coiled-coil domain containing 85A (CCDC85A) is associated with asthma exacerbations despite corticosteroid use⁸⁵, and a variant in LRRC8D, a volume-regulation anion channel component, is associated with atopic asthma¹⁰⁴; their increased expression here suggests the published variants may be disrupting

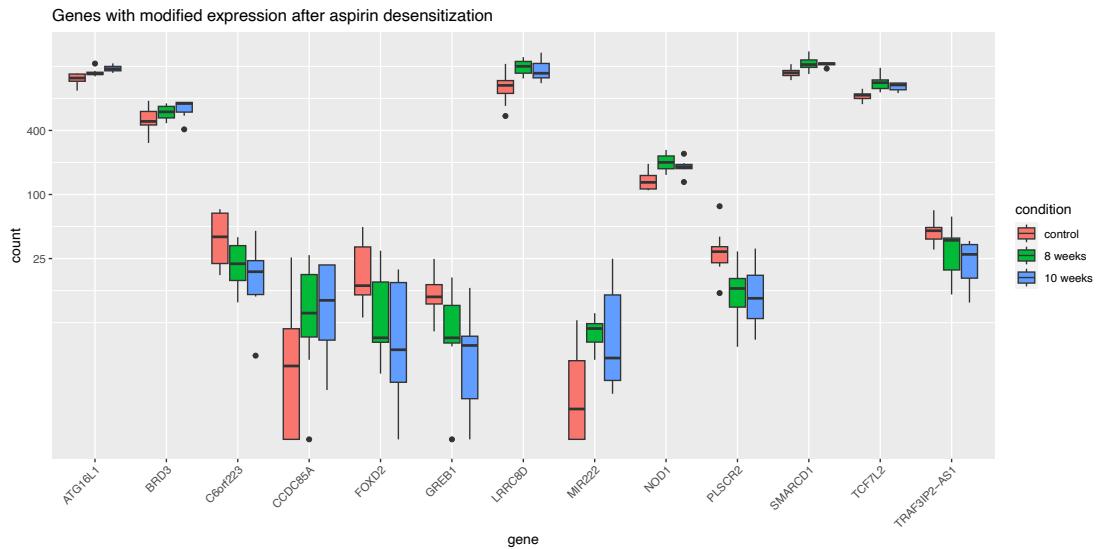


Figure 1.20: Some genes with only unadjusted significant p-values with known relevant functions may be playing a functional role in desensitization response.

a role in tissue repair or inflammation reduction. CIBERSORTx predicts that CCDC85A is predicted to be expressed by mast cells and/or eosinophils, indicating its relevance to the inflammatory processes underlying AERD.

A number of genes with consistently reduced expression were also of interest. This includes forkhead box D2 (FOXD2), a regulatory gene previously shown to have varied methylation in response to treatment of asthma¹⁷, and with methylation associated with childhood asthma²¹⁹. CIBERSORTx predicts that FOXD2 is expressed by ciliated cells and T cells in these samples. Growth regulating estrogen receptor binding 1 (GREB1), which is also decreased after desensitization, has been previously shown to be expressed at a lower level in IL-10+ innate lymphoid cells (ILCs) from nasal tissue than IL-10- ILCs¹⁴⁴. The resolution of our cell type prediction is too low to determine if ILCs are present, but CIBERSORTx does predict that the GREB1 in our sample was expressed by myeloid cells. MicroRNA (miRNA) MIR222 is increased after desensitization; previous work has found that one isoform of this miRNA, MIR222-3p, is elevated in allergic rhinitis²²². Using

CIBERSORTx, we predict that this miRNA originates in eosinophils and mast cells; why it would be increased under the ideally inflammation-reduced desensitized state is not clear. The reduction in phospholipid scramblase PLSCR2 suggests a change in the regulation of interferon response²⁰⁰. TRAF3IP2-AS1 is an antisense-RNA of TNF-receptor-associated factor 3 interacting protein 2. This antisense RNA, which decreases after desensitization, is a known regulator of IL-17⁸⁴. While these genes are potentially compelling, further study or additional samples are needed to validate and more fully interpret these findings.

1.5 DISCUSSION

1.5.1 PROLIFERATING B CELLS IN NASAL POLYPS

Of the markers of interest suggested by our collaborators (see Table 1.1), only IL5R α was significantly different across diagnoses (Figure 1.8). This finding agrees with other published work on this dataset²⁷. While this is interesting, IL5R α expression and cell counts in the scRNA-seq dataset were both relatively low, and additional experimental data was needed to reach conclusions. This reflects the fact that IL5R α is a surface receptor, and surface receptors tend to be relatively stable proteins with low RNA⁹. Follow-up experiments using cytometry by time of flight (CyTOF) were performed to investigate the IL5R α finding and explore the novel proliferative subtype. So far, we have not found any additional experimental evidence confirming the existence of this subtype; further work may include more sequencing and other experiments focused on these cells. Other studies of plasma cells in this tissue have also not identified proliferation²²¹. Repeating this analysis from raw data using read callers optimized for improved performance on immunoglobulin genes may also be helpful. Our findings on class switching and proliferative cell presence have not been directly published, but have contributed to our collaborator's thinking on this issue, as described in recent work²⁸.

1.5.2 THE INFERIOR NASAL TURBinate AS A MARKER TISSUE FOR DRUG RESPONSE

The results on the inferior turbinate are potentially compelling, but due to the small sample size and many degraded samples concrete results were limited. The observed transcriptional changes are modest, and larger studies would be needed to confirm and contextualize findings. Work following this study successfully used the inferior turbinate as a marker tissue of response to IL-5 inhibitor mepolizumab²⁹. Similar to what we observe here, drug treatment improved barrier integrity, in that case by upregulating genes associated with tight junction maintenance. Comparative analysis of gene expression in the inferior turbinate in healthy individuals and affected ones, as well as comparison of transcriptional responses to different drugs, may be interesting in the future.

*One autumn evening in a train
catching the diamond-flash of sunset*

in puddles along the Hudson

I thought:—I understand

life and death now, the choices

I didn't know your choice

or how by then you had no choice

how the body tells the truth in its rush of cells

Adrienne Rich, A woman dead in her forties

2

Cell state and cell-cell communication in triple-negative breast cancer

2.1 ABSTRACT

Understanding cell-cell communication is key to determining the role of the immune system and the tumor micro-environment in cancer progression. We use penalized matrix decomposi-

tion of single-cell RNA sequencing (scRNA-seq) data to interrogate multi-cell-type gene expression changes in triple-negative breast cancer (TNBC) tumors. We find that dimensionality reduction method choice strongly influences detected correlations, and identify reduction method and cancer-type specific multi-cell-type signatures. Furthermore, we explore gene set identification techniques to investigate a treatment-response predictive multicellular program in which increased interleukin 7 (IL-7) signaling by memory B cells is associated with increased expression of heat shock proteins in memory and naive T cells. In total, we demonstrate the power of statistical tools for deriving meaning from complex high-dimensional data. Nonetheless, cell-cell communication prediction from scRNA-seq is limited by the absence of ground truth data, and significant experimental studies are still needed to substantiate our predictive findings.

2.2 INTRODUCTION

CELL-CELL COMMUNICATION IS FUNDAMENTAL TO MULTICELLULAR LIFE and underlies many disease processes. This cell-cell communication then drives immune dysregulation at the cell level, the tissue level, or at the individual level, causing a whole-body syndrome or disease². Particularly in the immune system, cell signaling is extremely dependent on context, with ligand and receptor levels dynamically adapting to extracellular cues¹⁶⁷. Even within cells, the gene expression changes induced by a signaling molecule depend on the transcriptomic state of the receiving cell⁹⁰. As a result, the study and development of therapeutics for the immune system is inextricably linked to understanding and interpreting cells' interactions with their environmental. Building mathematical tools that can go from deep -omics data to interpretable biological mechanisms is key to understanding the role of the immune system.

Historically, cell-cell communication was investigated by exquisitely detailed experiments, one

interaction at a time. Even today deep investigation of the interactions between just two cell types relies on hundreds of experiments⁷⁶. Targeted experimental methods can only investigate known interactions, requiring foreknowledge of the interactions of interest¹⁵. Hypothesis-free experimental modalities such as transcriptomics have moved towards discovering cell-cell communication without the requirement of predefined interactions. Even before the advent of scRNA-seq, bulk RNA-seq of sorted cells and databases of known ligand-receptor interactions could be used to discover the role of cell-cell communication in development and disease. For example, cell-type specific ligand production and comparatively non-specific receptors were demonstrated to underlay the niche-composition dependence of differentiation in bone marrow¹⁶⁴.

More recently, scRNA-seq has been applied in a broad body of work that has identified cell-cell communication in immunology, development, and other fields⁶. These methods infer likely interactions in single-cell transcriptomic data using models built from interaction databases. For instance, one sequencing study identified ligand-receptor pairs between maternal and fetal cells at the placenta and decidua²⁰². The method developed as part of that work, CellPhoneDB, uses assembled biological knowledge of ligand-receptor interactions along with scRNA-seq to identify frequently occurring cell-cell interactions⁶². Two other state-of-the-art approaches are RNA-Magnet and Nich-eNet, both of which, like CellPhoneDB, use models built from interaction databases to analyze interactions in single-cell transcriptomic data^{9,24}. The latest developments have extended this analysis further, into quantifying intercellular interactions between individual cells in a dataset without pooling over defined cell types²¹⁰. Comparative analysis of diverse published cell-cell communication inference tools found that although the routes of communication highlighted as significant by different methods varied, most did seem to be identifying true features of the underlying data⁵⁶.

We aim to uncover interactions that are not yet known, as well as correlations due to shared response to external stimuli. As such, rather than working from an existing cell-cell communication identification method, we work from a method for identifying cell-cell interactions using multi-

cell-type correlations discovered via matrix decomposition⁹³. These correlations are not necessarily due to interactions, as they can also represent a shared response to external stimuli. However, they nonetheless define features shared across cell types, and can be used to identify gene sets with context-dependent connections.

2.2.1 TRIPLE-NEGATIVE BREAST CANCER

We are specifically interested in interactions between cell types which underlie treatment response in TNBC. Triple-negative breast cancer is a relatively rare form of breast cancer, accounting for 15-20% of diagnosed breast cancers²¹⁶. The triple-negative portion of the name refers to the lack of expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), and a corresponding lack of response to therapeutic agents targeting those proteins¹⁹⁸. Compared to other breast cancers TNBC has a higher rate of distant recurrence and worse 5-year prognosis⁹³. TNBC is highly heterogeneous and is thought to be composed of multiple subtypes which may respond differently to treatments³⁴. At present, first line treatment for TNBC is typically neoadjuvant chemotherapy (NACT), sometimes with additional immunotherapy drugs targeted to known tumor-immune interactions²²⁰.

Single-cell profiling of breast cancer tumors and the tumor micro-environment (TME) has revealed the complexity and diversity of tumor infiltrating lymphocytes²¹⁸. The TME consists of immune cells surrounding and infiltrating the tumor, some of which are trying to fight the cancer, others of which have been co-opted to protect it. Tumor cells generate immunosuppressive metabolites, driving T cell exhaustion in the TME and protecting the tumor¹⁷⁰. Single-cell sequencing of breast cancer tumors has revealed T-cell subtypes not discoverable from surface protein measurement alone; TNBC tumors were specifically found to be enriched for plasmacytoid dendritic cells (pDCs)¹⁶³. These pDCs recruit regulatory T cells, inducing immunosuppression and protecting the tumor²²⁴. Tumor-immune crosstalk is believed to be a significant driver of treatment

resistance, as reviewed in ¹³. Treatment response has also been investigated using single-cell data. One study looked at multiple time points of two patients undergoing treatment with NACT and adjunctive immunotherapy, finding that programmed cell death protein 1 (PD-1) expressing T cells decreased after treatment in a responding patient, but not in a non-responding patient ⁵³. Here, we use data from an atlas of breast cancers of different types to examine TNBC-specific features of the TME ¹⁵¹. We also investigate treatment response using a larger published dataset of TNBC patients treated with NACT with or without augmentation with programmed death-ligand 2 (PD-L1) inhibitor paclitaxel ²²⁰.

Predicting treatment response can minimize time patients are treated with toxic therapies, reducing side effect burden, and, in the future, hopefully point towards when adjunctive therapies will improve outcomes ¹²³. Prior work using single-cell transcriptomics to predict breast cancer prognosis as of 2021 is reviewed in detail in ¹⁶⁵. Here, we aim to use single-cell transcriptomics to identify cross-cell type interactions which can predict treatment response in TNBC.

2.2.2 DIMENSIONALITY REDUCTIONS AND MATRIX DECOMPOSITIONS

Virtually all analysis of scRNA-seq data relies on dimensionality reductions of some form, as the full data, with more than 20,000 genes and considerable dropout, does not obey the requisite assumptions of most standard statistical tests. This is most commonly applied dimensionality reduction method is principle component analysis (PCA), which is used to remove unwanted noise by projecting the dataset onto the axes of highest variation prior to downstream analysis such as clustering. Alternatives to PCA specifically designed for scRNA-seq data can improve performance, for example by improving cluster distinguishability or detection of nonlinear effects. These include latent variable models with noise correction such as scLVM, which corrects for both known sources of noise, like the cell cycle, and unknown noise factors ³¹. Variational autoencoders (VAEs) have also been used extensively for modeling scRNA-seq data ^{122,78}. Such VAEs represent the data using negative

binomial (NB) or zero-inflated negative binomial (ZINB) distributions for each gene, then infer the parameters of those distributions as part of learning an embedding. Some methods like single-cell variational inference (scVI) use negative binomial distributions while simultaneously explicitly modeling dropout¹²²; others use ZINB distributions directly^{158,168}. With any of these methods, one of the main challenges is how to interpret the data once embedded.

Further dimensionality reduction to 2D for visualization is typically performed using t-distributed stochastic neighbor embedding (tSNE)^{106,133} or uniform manifold approximation and projection (UMAP)^{139,14}. While it is tempting to use these visualizations to assist in interpretation, both can misleadingly distort high dimensional structures⁴⁰. As such, these visualizations cannot be reliably used for data interpretation.

How, then, might one go from scRNA-seq data to an interpretation or story of behavior across cell types? Some have approached this challenge by specifically designing dimensionality reduction techniques to improve interpretability. For example, β -VAEs modify VAEs by adding a β parameter which forces the latent space to be independent of user-defined nuisance factors, resulting in readily interpretable latent dimensions³². This has been applied to scRNA-seq data to predict response to perturbation across cell types¹²⁴. Similarly, latent space constructed using separate autoencoders for genes, perturbations, and other covariates enabled improved prediction of drug responses across cell types¹²⁵. While compelling, understanding what the methods themselves are actually doing and evaluating performance to prevent artifactual findings remains the subject of active research.

2.2.3 PENALIZED MATRIX DECOMPOSITION APPLIED TO BREAST CANCER

While these directly interpretable dimensionality reduction methods are undoubtedly a compelling area for future research, here we focus instead on readily-interpretable PCA and the relatively simple scVI VAE. We apply the DIALOGUE algorithm⁹³ on dimension-reduced data to extract multi-cell-type correlations and use them to interpret the observed transcripts. This algorithm uses penalized

matrix decomposition²¹¹ to discover multicellular programs (MCPs) across multiple cell types after latent space embedding. The penalized matrix decomposition underlying DIALOGUE is a direct analog of sparse PCA⁹⁷. By applying this decomposition to a dimension reduced space for each cell type, we extract components of those reduced dimensions which maximally correlate with each other. The resulting MCPs are sets of changes in gene expression in one cell type that correlates with changes in gene expression in other cell types across samples. MCPs have been found to predict response to therapy in other diseases such as ulcerative colitis and correlate with immunotherapy resistance in melanoma⁹³, but have not been applied to TNBC, and has not been used to investigate interactions in scRNA-seq data reduced with something other than PCA.

Because these calculations are performed on a space of reduced dimensionality, we expect the correlation structure discovered to vary with the dimensionality reduction method. In other words, the choice of latent space may influence which correlations between cell types one can detect. As such, in this study, we first investigate the effect of latent space choices on correlation structure learned using DIALOGUE in scRNA-seq samples from breast cancer patients. Our focus is on identifying genes that have correlated expression across multiple cell types. These correlations may be useful for identification of cell-cell communication without the restriction to known ligand-receptor pairs⁶. We specifically compare two MCP analyses with different dimensionality reduction methods: a standard PCA-based analysis with sample integration from Seurat v3¹⁸⁸ and the scVI variational autoencoder with batch correction¹²². We perform this comparative analysis on a scRNA-seq breast cancer atlas of different cancer types, finding features unique to TNBC and to specific dimensionality reductions.

We also investigate MCP gene signature identification methods using a TNBC dataset with treatment response information²²⁰. Gene signature identification for single cells and differential expression testing for single cells remain unsettled questions. In general, single-cell differentially expressed genes (DEGs) are not a consistent measure of effect size¹⁸⁰. There is poor overlap of DEGs across

studies, including in the case of pseudobulk methods⁹¹. The DIALOGUE algorithm includes a multilevel-modeling based gene identification scheme which is very computationally intensive, and does not provide evidence that this method outperforms existing gene identification techniques⁹³. Using the treatment response dataset as a test case, we explore alternative methods for identifying MCP-associated genes. We use the discovered gene sets to characterize a treatment response predictive MCP in pre-treatment tumor resident B cells and T cells.

2.3 METHODS

2.3.1 LATENT SPACE CHOICE AFFECTS INFERRED MULTI-CELL TYPE RELATIONSHIPS

Single-cell data was obtained from the NCBI Gene Expression Omnibus at GSE161529^{151,12}. Data was subset to samples from breast cancer gene 1 (BRCA1) positive TNBC tumors, ER positive tumors, HER2 positive tumors and TNBC tumors. The cell type labels were used as provided by the original study. Seurat analysis was performed using Seurat v4.0.3⁸¹. To ensure that all patients had all cell types present (necessary for DIALOGUE analysis), the following cell type label changes were made: vascular endothelial cells (“vascEndo”) and lymphatic endothelial cells (“lymphEndo”) were combined into a single endothelial cell type; dendritic cells (DCs) were combined with other myeloid cells to form a single myeloid category; cells labeled with fibroblast, CAF, and CAFs were combined into a single cancer-associated fibroblast (CAF) group; plasma cells and B cells were combined and labeled “B cells”. The sample “ER_ooo1” was removed from analysis due to insufficient cell type diversity.

Cell cycle status was not provided for the BRCA1+ TNBC tumor samples in the original study. To determine cycling status, all BRCA1+ TNBC epithelial tumor cells were used to create a Seurat object. This object was then split by group and normalized. 1000 variable features selected using ‘vst’ for each group then combined using Seurat’s anchor integration method. The combined ob-

ject was scaled, reduced to 20 principle components (PCs), and clustered using `FindClusters` with resolution 0.1. Cells in clusters with expression of marker of proliferation Ki-67 (MKI67) were annotated as cycling epithelial tumor cells; all other cells were annotated as epithelial. Labeled, unnormalized cells from BRCA1+ TNBC tumor samples were then combined with BRCA1- TNBC tumor samples in downstream analysis.

Unnormalized data from all tumor types was combined into a single Seurat object prior to normalization. Variable features were selected using Seurat `FindVariableFeatures` with selection method “vst” and 1000 features. Data was normalized using `NormalizeData` with default parameters and `ScaleData` with the 1000 variable features. PCA was calculated using default parameters. To select the number of PCs we used the minimum of the PC at which 90% of the variation of the data is explained and the individual PC is less than 5% of the variation of the data, and the last PC at which the percentage change in variation explained between that PC and the subsequent PC is less than 0.1%. In this case, that was 18 PCs. Normalized and unscaled input was used as input “tpm”. scVI analysis used `scvi-tools` v. 0.14.4. 5000 highly variable genes (HVGs) were selected using `sc.pp.highly_variable_genes` with flavor `seurat_v3` and batch key set to the sample ID. The scVI model was trained using default parameters: 128 hidden dimensions, 10 latent dimensions, 1 layer, and a dropout rate of 0.1, dispersion fixed by gene across cells, and zero-inflated negative binomial gene likelihood with normal latent distribution. DIALOGUE analysis using the scVI latent space used all 10 latent dimensions. For DIALOGUE analysis, the number of output MCPs was set to 10 and the number of genes detected per cell was used as a confounder.

2.3.2 MULTICELLULAR PROGRAM PREDICTS TREATMENT RESPONSE

Single-cell data was obtained from ²²⁰. Cell type labeling was used as supplied by the original authors. Genes were subset using Seurat `FindVariableFeatures` with `selection.method` set to “vst” and `nfeatures` set to 4000. Data was normalized using Seurat `NormalizeData`. Genes were scaled

Cell type ID	Cell type name
t_Tn-LEF1	Naive T cells
t_Bmem-CD27	Memory B cells
t_CD8_Tem-GZMK	CD8 effector memory T cells
t_pB-IGHG1	Plasma B cells
t_CD4_Treg-FOXP3	CD4 regulatory T cells
t_CD4_Tcm-LMNA	CD4 central memory T cells
t_CD8_Trm-ZNF683	CD8 tissue-resident memory T cells
t_CD8_MAIT-KLRB1	CD8 mucosal-associated invariant T cells
t_mono-FCN1	classical monocytes

Table 2.1: Cell type abbreviations and names after filtering for DIALOGUE analysis of²²⁰.

using `ScaleData` prior to `RunPCA`.

DIALOGUE decomposition was performed on pre-treatment tumor samples only. The sample “Pre_Po10_t” was removed due to low cell type diversity. Cell types were subset to only cell types with at least 3 cells per sample in the remaining patient samples, leaving the cell types shown in Table 2.1. The DIALOGUE decomposition was performed using a Python-based re-implementation of DIALOGUE which has been incorporated into the `pertpy` package⁸⁶. DIALOGUE had `n_mcps` set to 10, `normalize` set to `True` and `solver` set to `bs`, which performs identically to the solver used in⁹³.

A treatment-response predictive MCP was identified by testing each cell type in each MCP separately using `stats.ttset_ind` from SciPy 1.10.1. Benjamini-Hochberg correction was used to correct for the number of cell types tested.

The multilevel modeling method used in⁹³ was re-implemented in Python. Verification of matching results for cell type pairs from the R and Python implementations will be available in an upcoming publication⁸⁶. This method tests pairs of cell types to identify MCP-correlated genes. To create unified MCP gene sets from the set of pairs analyses, we kept only genes which appeared in at least the floor of 0.7*(the total number of cell types). We made this adaptation because the adaptive

thresh-holding method used in ⁹³ is vulnerable to removal of some cell types from the analysis as a result of a perturbation-based testing procedure for MCP associations; as such, the results described here are slightly different from what would be returned by the original DIALOGUE algorithm. On the dataset explored here, this resulted in an average of 100 MCP genes per cell type per MCP.

The second gene identification method looked directly at the MCP loadings. To do this, we matrix multiplied the MCP loadings as provided by Seurat by the PC-to-MCP transformation vector (w) provided by DIALOGUE. For gene set comparisons, we selected the 50 genes with the largest positive contribution component, and the 50 genes with the largest negative contribution component.

For the third gene identification method, referred to as extrema MCP genes, we selected cells which were at the extreme values of the MCP (cells with the top 10% and bottom 10% MCP scores in each cell type), then used the `rank_genes_groups` function from scanpy with default parameters. This function performs a t-test on the two groups of cells to identify differentially expressed genes, and provides an adjusted p-value based on the number of genes tested.

To check if identified genes were involved in a changed cell-cell interaction, genes from each cell type were compared against the set of protein-protein interactions labeled as gene names from ²⁴ using the database supplied by ⁶. An interaction was declared MCP-associated if both the matched receptor and its ligand appeared across two different cell types significant genes. An interaction was MCP-ligand-associated if the ligand was MCP associated for one cell type and the receiver had normalized mean expression greater than 1 in the second cell type. An interaction was MCP-receptor-associated if the receptor was MCP associated for one cell type and the ligand had normalized mean expression greater than 1 in the other cell type.

Condition	Samples	Cells
ER+	12	53k
TNBC	8	87k
HER2+	6	30k

Table 2.2: Cell and sample counts used for DIALOGUE analysis after filtering¹⁵¹.

2.4 RESULTS

2.4.1 LATENT SPACE CHOICE AFFECTS INFERRED MULTI-CELL TYPE RELATIONSHIPS

As described in the introduction of this chapter, DIALOGUE uses penalized matrix decomposition in latent space with the aim of discovering multicellular programs⁹³. Each MCP is a set of changes in gene expression in one cell type that correlates with changes in gene expression in other cell types across samples. DIALOGUE identifies MCPs by computing a penalized matrix decomposition of the average location in a dimension-reduced space for each cell type across samples. The cell and patient totals for each cancer type are shown in Table 2.2.

Because these calculations are performed on a space of reduced dimensionality, we expect the correlation structure discovered to vary with the dimensionality reduction method used. In other words, the choice of latent space may influence which correlations between cell types one can detect. To explore the effect of latent space on MCP analysis, we compare two MCP analyses with different dimensionality reduction methods: a standard PCA-based analysis with sample integration from Seurat v3³⁵ and the scVI variational autoencoder with batch correction¹²². In general, we expect that the autoencoder based method will enable discovery of nonlinear correlations between genes that are not discoverable after PCA.

The two dimensionality reduction methods produce visibly distinct UMAPS (Figure 2.1). The most obvious difference in structure between the two is that scVI separates ER positive epithelial tumor cells (circled) whereas after PCA the two overlap. The cell types which had significant compo-

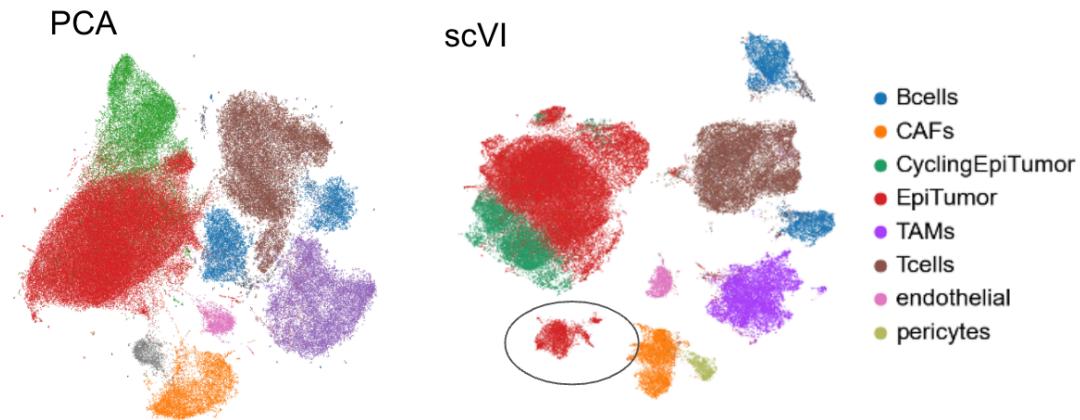


Figure 2.1: UMAP plots produced using PCA and scVI of breast cancer data from¹⁵¹. A circle highlights a subpopulation of ER+ tumor cells which are visibly distinct only after scVI reduction.

nents in each MCP varied significantly between the different dimensionality reductions (Figure 2.2). Most notably, T cells and pericytes had significant contributions to most MCPs in the scVI reduction, but fewer in the PCA reduction; this suggests that the interactions underlying the connection between these cell types may be nonlinear, and thus eliminated by linear dimensionality reduction in PCA.

MCP1 for scVI denotes the greatest contributor to multicellular variation in the dataset for that dimensionality reduction (Figure 2.3). The full list of genes associated with both MCPs via standard DIALOGUE analysis is available in Supplemental Table 2.1. Particularly of note is the relationship between increased expression of proliferation-related (Jun proto-oncogene (JUN), sphingosine kinase 1 (SPHK1), fibronectin 1 (FN1)) and glucocorticoid response (zinc finger proteins ZFP36L2, ZFP36, and SMYD3) genes in CAFs and the expression of calcium channel regulators in tumor-associated macrophages (TAMs). This connection between TAMs and CAFs is in line with well-established findings on the function of these cell types in the TME⁹⁸. Interestingly, scVI-MCP1 most closely resembles PCA-MCP4 rather than PCA-MCP1, a clear indication of the significant impact that dimensionality reduction has on the identified MCPs. 28 of the genes in scVI-MCP1 ex-

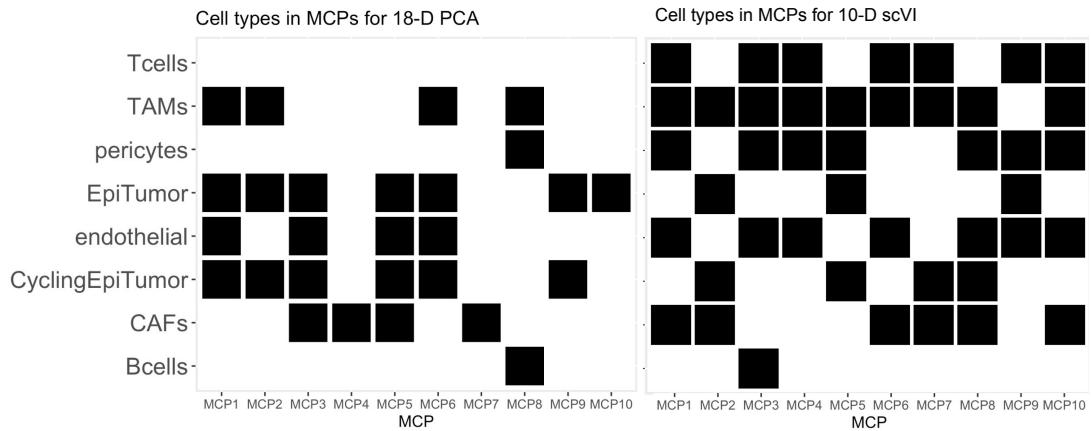


Figure 2.2: MCP membership for PCA and scVI reductions of data in ¹⁵. Black squares denote cell types included in the MCP on the x-axis.

pressed in CAFs are also associated with PCA-MCP₄. This difference may be an artifact of how few associated genes the DIALOGUE method for identifying MCP genes finds after PCA dimensionality reduction. Across all the MCPs, there were 526 individual gene-cell type-MCP associations in the PCA reduction, whereas for scVI there were 906. However, the fact that all the shared genes are from CAFs demonstrates the extent to which dimensionality reduction influences the discovered cross-cell-type relationships.

In contrast, MCP₁ for the PCA reduction has fewer associated genes, fewer significantly related cell types, and overall lower correlations (Figure 2.4). This MCP captures the largest amount of cross-cell-type covariation, and the distinct set of cell types and associated genes from the scVI-MCP₁ demonstrates that this covariation is highly dependent on dimensionality reduction method. Although the MCP contains many cell types, only epithelial tumor cells and cycling epithelial tumor cells had genes which were significantly associated according to DIALOGUE's multilevel modeling-based significant testing. Notably, despite both cell types being nominally similar, the MCP-associated genes were mostly distinct, with only annexin A₃ (ANXA₃) shared. ANXA₃ is a known regulator of cell proliferation; higher expression of ANXA₃ is associated with tumorogene-

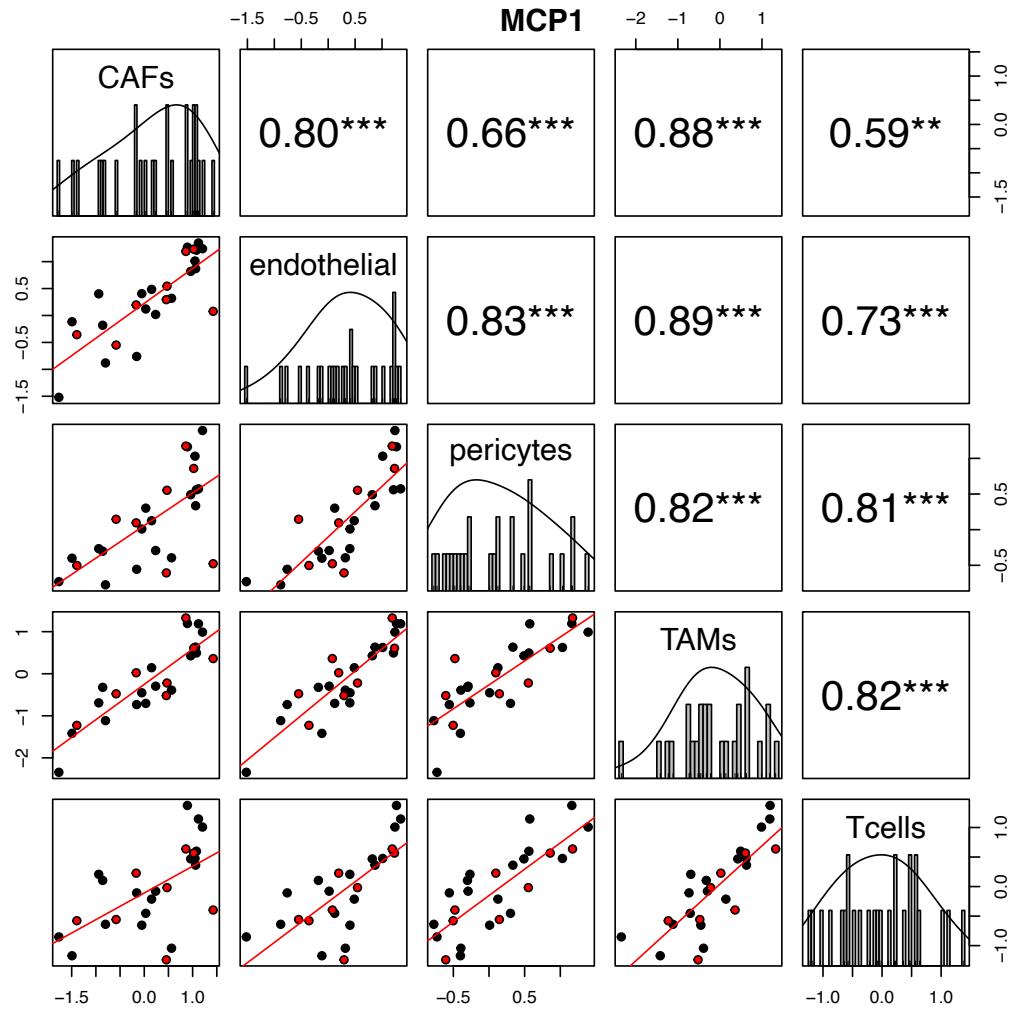


Figure 2.3: Pair plot for the first MCP for the scVI dimensionality reduction. Along the diagonal is a histogram of the average score for each MCP by sample for the listed cell type. Significance and Pearson correlations for each pair are displayed in the upper triangle. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). Points marked in red are samples from patients with TNBC.

sis and progression¹²⁰. This dataset did not provide information on tumor progression or survival time, so we are unable to test if this MCP is indicative of a tumor-promoting interaction. Moreover, interpretability is further limited by the dearth of associated genes from the DIALOGUE gene modeling.

MULTICELLULAR PROGRAM SPECIFIC TO TRIPLE NEGATIVE BREAST CANCER

In the scVI analysis we observe a TNBC-specific pericyte subpopulation present in samples that had high expression of S100A13 (S100A13) in T cells (Figure 2.5). S100A13 is expressed in a wide variety of T cell subtypes¹⁴². It is a known regulator of cell senescence, particularly via action of the non-classical secretory pathway of IL-1 α ¹⁸⁹. There is growing evidence that pericytes can act as regulators of T cells and other adaptive immune cells¹⁴⁷. The full list of genes associated with this MCP in pericytes is available in Supplemental Table 2.2. This list includes cellular retinoic acid binding protein 2 (CRABP2), which suppresses invasion and metastasis in ER+ breast cancer but promotes invasion and metastasis in ER- breast cancers⁶³. These findings hint at a possible regulator role for pericytes and T cells in TNBC, and perhaps point towards a new direction for therapeutic development. However, direct experimental proof of causality is still needed, as all findings from DIALOGUE are purely correlative. It is possible that the pericyte gene signature is playing a functional role in rapid angiogenesis of TNBC tumors, which coincidentally tend to be more immunogenic than other breast cancer subtypes; this immunogenicity may be behind lower levels of T cell senescence, and thus the observed increased expression of S100A13 in this MCP.

This re-analysis of a published single-cell atlas of different types of breast cancer identified an MCP component unique to TNBC only in the scVI reduction. It is clear from this analysis that scVI functions analogously to PCA, and that interesting patterns can be uncovered regardless of the dimensionality reduction used. One major advantage of PCA, however, is the ability to project MCPs to identify contributory genes. PCA, like DIALOGUE, is a dimensionality reduc-

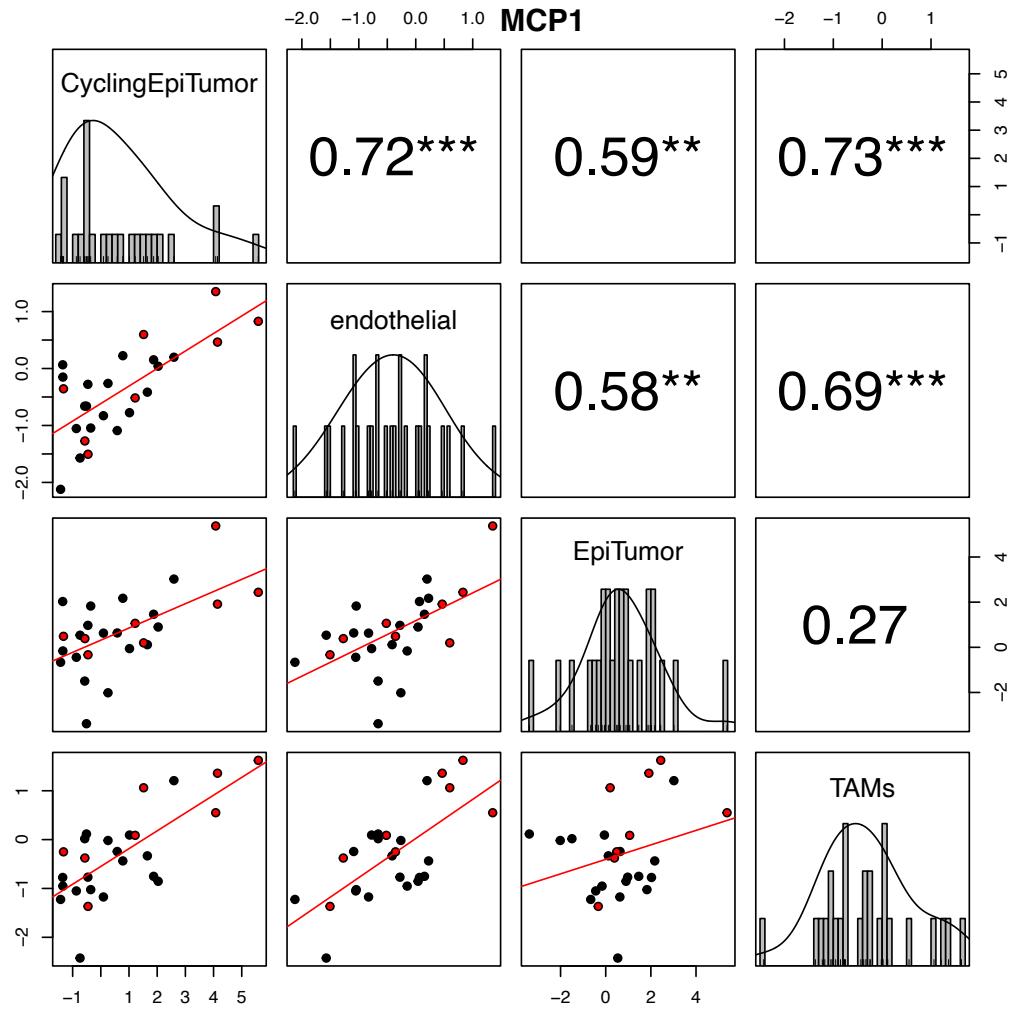


Figure 2.4: Pair plot for the first MCP for the PCA dimensionality reduction. Along the diagonal is a histogram of the average score for each MCP by sample for the listed cell type. Pearson correlations for each pair and associated significance are displayed in the upper triangle. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). Points marked in red are samples from patients with TNBC.

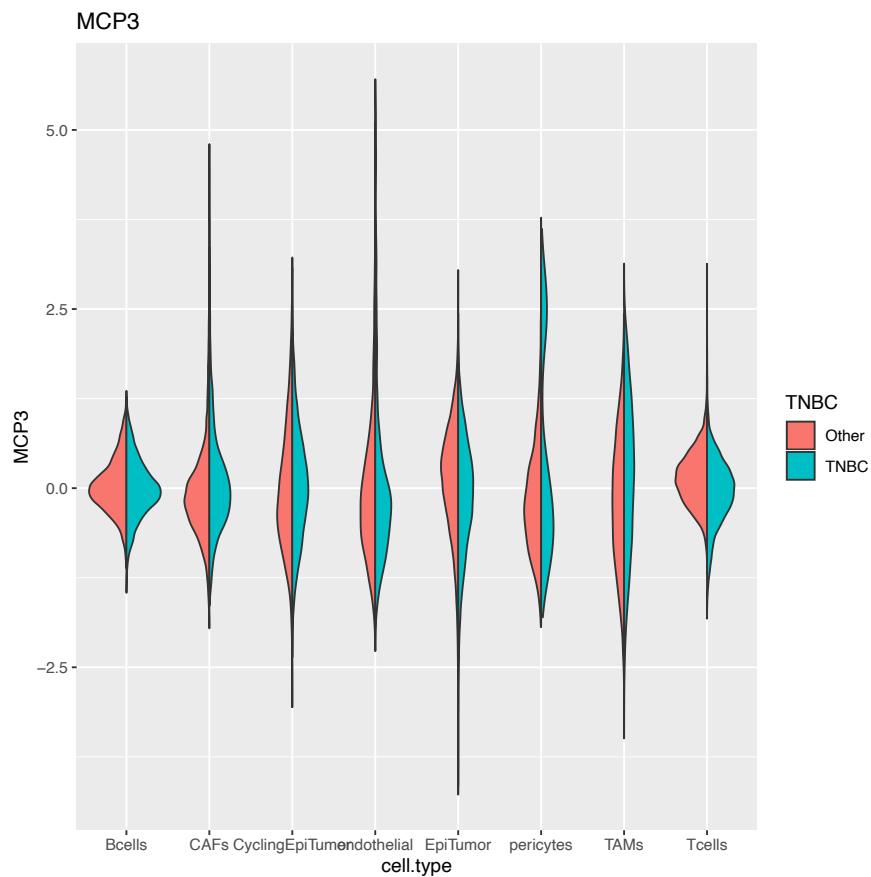


Figure 2.5: MCP score for MCP3 in breast cancer tumors from ¹⁵¹. TNBC tumors are shown in blue; ER+ and HER2+ tumors are analyzed together and shown in red. MCP3 score is relatively similar across breast cancer types aside from in pericytes; where a high-MCP3 subpopulation is seen only in TNBC.

tion method, transforming data from gene space to PCA space to maximize the variation contained in each component. By combining the two transformations, we can directly rank genes by their contribution to each MCP. We thus use PCA for analysis of a second dataset, but note that scVI can also provide valuable insights.

2.4.2 MULTICELLULAR PROGRAM PREDICTS TREATMENT RESPONSE

We are specifically interested in cases of cell-cell communication which are predictive of treatment response. To that end, we also explored the tumor scRNA-seq data available in ²²⁰. This dataset includes pre- and post-treatment samples from TNBC patients being treated with either NACT or NACT with paclitaxel. Due to the relatively small number of pre-treatment tumor samples we pooled both treatment categories, and treatment type was set as a confounder during DIALOGUE analysis; as such, we anticipate any identified MCPs will predict response to chemotherapy in the absence of paclitaxel treatment.

Of the 10 MCPs calculated, MCP₄ was the most correlated with treatment response (Table 2.3). Due to low sample counts, the choice of treatment was not accounted for in the statistical testing. The sample averages for the cell types with the smallest adjusted p-value are shown in Figure 2.6; violin plots of individual cell scores are in Figure 2.7. These include memory B cells, CD4 central memory T cells (T_{cm} cells), CD8 mucosal-associated invariant T cells (MAIT cells), naive T cells, and plasma B cells. It is immediately clear that this MCP separates responding and non-responding patients, though the degree of separation varies; patients who responded to either treatment modality had uniformly low MCP₄ scores. Some non-responding patients had low scores as well, particularly in plasma B cells for patients who received only NACT. This suggests that some NACT-non-responder patients may have responded if treated with anti-PD-L1 in addition to NACT.

Now that we have a putative predictive MCP, we want to identify what changes in transcriptomic space underlay that MCP and use those to propose interaction mechanisms. The original

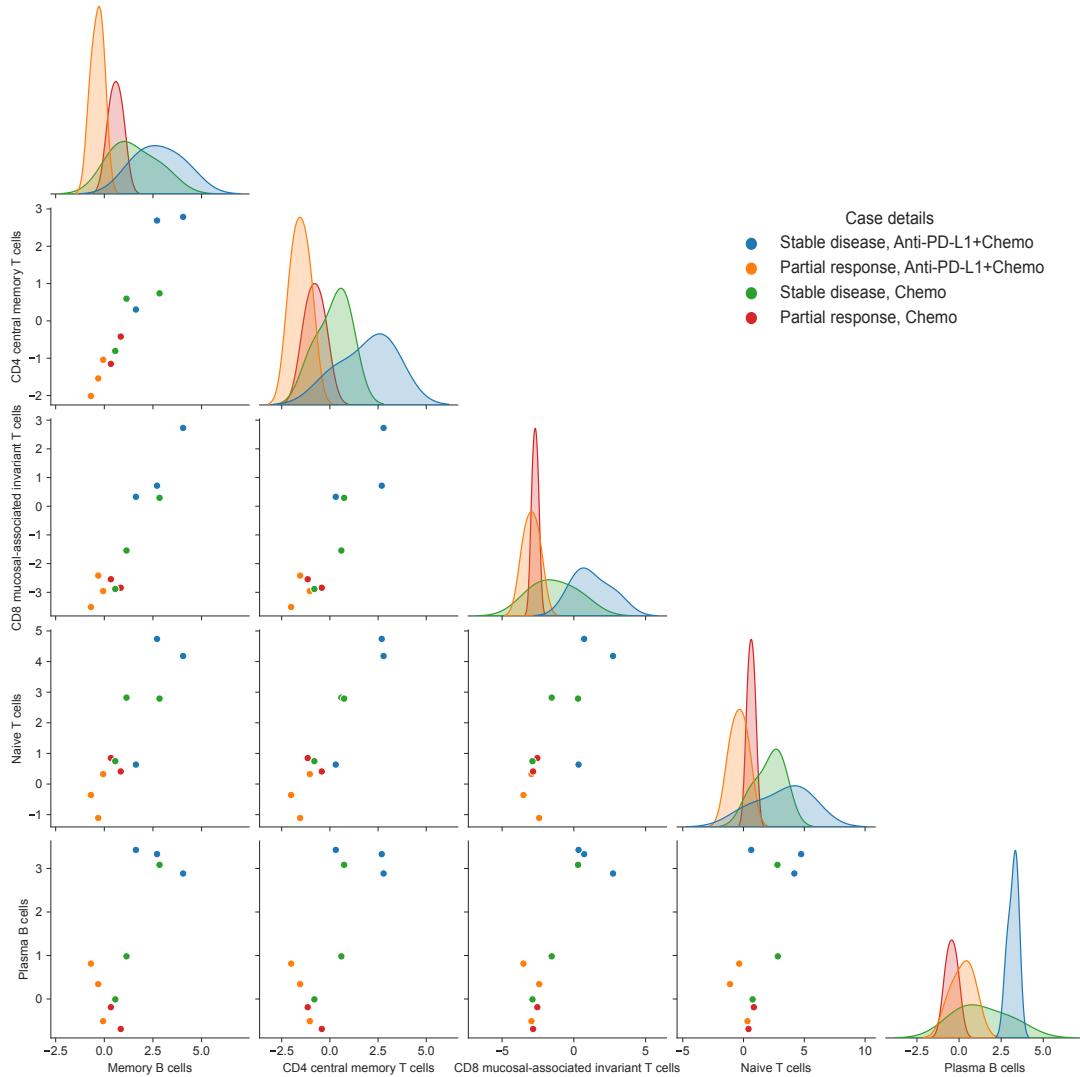


Figure 2.6: Pair plot for MCP4 from analysis of²²⁰. Along the diagonal is a kernel density estimate of the average score for each MCP by sample for the listed cell type. For the scatter plots in the lower triangle, each dot represents a patient average for the cell types listed on the given row (x-axis) and column (y-axis). All patients in the partial response category had tumors which shrank after treatment but were not entirely eliminated.

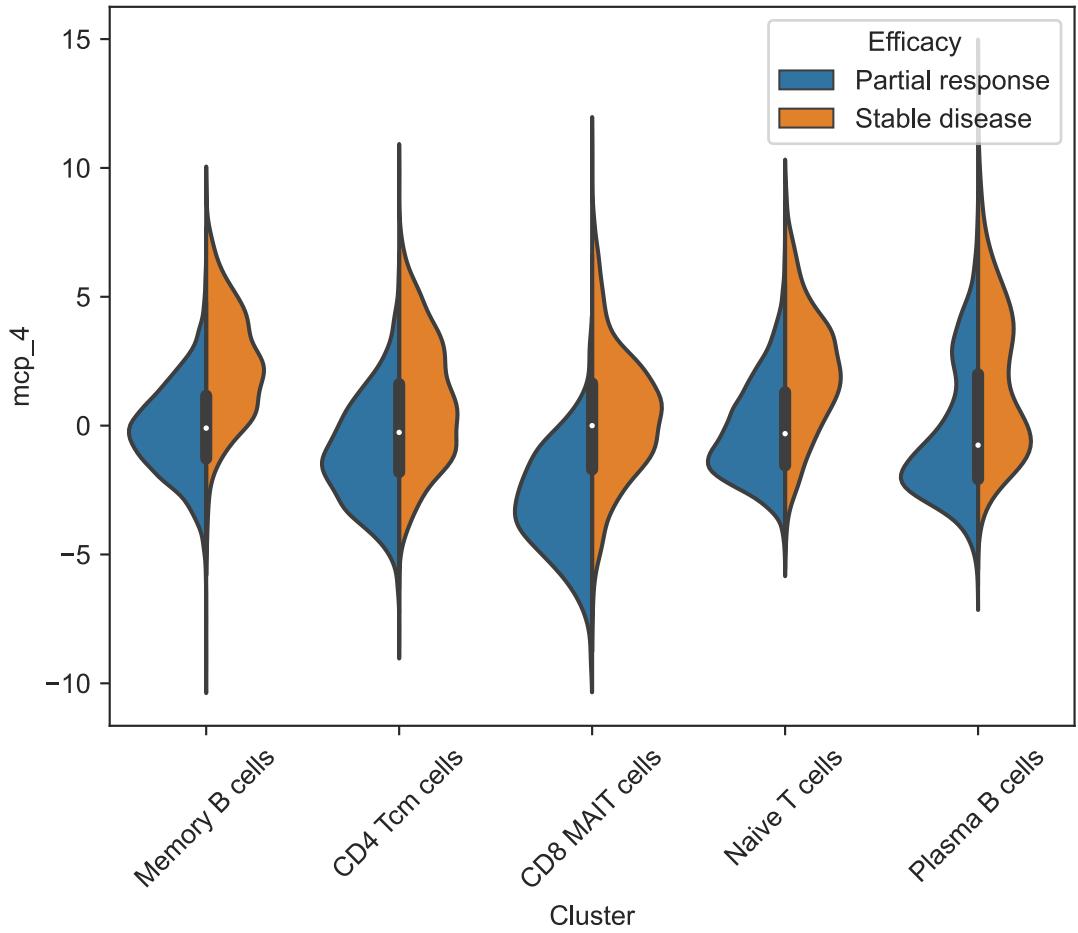


Figure 2.7: Single-cell MCP scores for cells from patients who did or did not respond to treatment in ²²⁰. Cells from patients who received either treatment are pooled here.

Cell type	p-value	Adjusted p-value
Naive T cells	0.011209	0.021844
Memory B cells	0.007831	0.021844
CD8 effector memory T cells	0.028376	0.031923
Plasma B cells	0.008574	0.021844
CD4 regulatory T cells	0.024958	0.031923
CD4 central memory T cells	0.008501	0.021844
CD8 tissue-resident memory T cells	0.025198	0.031923
CD8 mucosal-associated invariant T cells	0.012136	0.021844
classical monocytes	0.079806	0.079806

Table 2.3: Cell type associations with treatment response. This is testing whether the average MCP score for each cell type for each patient’s pre-treatment tumor sample was significantly different in responding and non-responding patients. The p-values were calculated using an independent t-test. Adjusted p-values were adjusted for the number of cell types tested using a Benjamini-Hochberg correction factor.

DIALOGUE paper uses what is described as a multilevel-modeling schema⁹³. However, the code supplied by the authors has only one level, and corrects multiple times for the same set of factors. The procedure described in the paper is also designed to enrich for genes which are MCP-associated in both the cell type of interest and other cell types under investigation—while potentially useful, the MCP itself was already calculated to enforce this cross-cell-type similarity. The multiple stages of optimization also make it extremely slow to compute, and despite all the underlying calculations, it still does not return any sort final statistic for ranking the different genes—statistics are only available for each individual pair of cell types in the MCP.

As such, we wanted to explore alternative approaches, ideally ones which are more conceptually straightforward. We reimplemented the pairwise multilevel modeling scheme from DIALOGUE in Python, and developed a simple heuristic to aggregate MCP genes from pairs to full MCP genes: any gene calculated as significant in the pairwise comparison with at least 70% of the remaining cell types was considered an MCP gene. This method resulted in a set of MCP genes for each cell type and each MCP, which we then compare against two original MCP gene selection methods.

The first of these directly uses the MCP loadings. The DIALOGUE algorithm is, at its core, a matrix decomposition with extensive external corrections. The algorithm returns scores for each cell and sample (these are what is plotted in Figure 2.6) but it also returns the weights used to decompose the PCs into MCPs. By multiplying these weight vectors by the vectors used to move from gene space to PC space, we can directly access the genes contributing to each MCP. Two of the loadings are visualized in Figure 2.8. Concerningly, the most highly associated genes in memory B cells are T cell specific genes (*TRDV2*, *KLRB1*); the most negatively associated are macrophage specific (*F13A1*, *FOLR2*)¹⁰¹. This suggests that some observed correlation may be due to shifts in cell type abundance affecting background expression, for lysed macrophages contributing to ambient sequencing, with larger numbers of these macrophages seen in responding patients.

While this method is useful, it also has a significant downside: only genes that were used to compute the PCs can have an association discovered using this method. While the excluded genes may explain minimal variation across the entirety of the original dataset, that is not necessarily true within each single cell type, particularly in a case like this one where the annotated cell types are very fine-grained—there were more than 50 tumor-resident immune cell types annotated in the original dataset. Of the 1741 genes associated with at least one of the ten MCPs in at least one cell type according to DIALOGUE’s method, only 876 were among the highly variable genes used to compute the PCs. Thus while the loading genes are a meaningful representation of the computation origin of the MCP, using this method means losing access to potentially meaningful biological signal.

As an alternative statistical testing method, we implemented an extrema-based gene testing procedure. Here, we select cells which had very high or very low MCP scores, then use gene differential expression testing to identify genes which are significantly different between the two groups. This method dramatically more computationally efficient than DIALOGUE: the multilevel modeling method run on a personal computer takes a few hours, whereas the extrema testing procedure takes only a few seconds. Unlike DIALOGUE, however, this method does not account for confounders

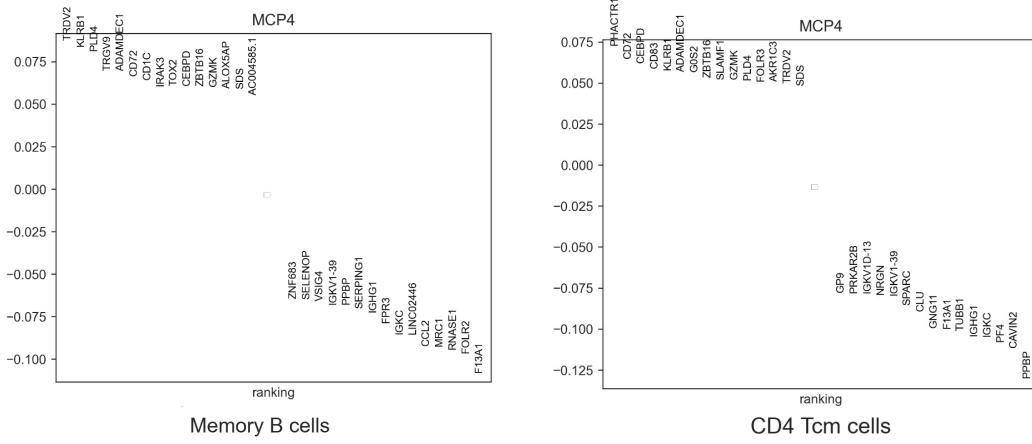


Figure 2.8: MCP loadings for a response-predictive MCP in memory B cells and CD4 central memory T cells. The top ten positive and negative contributory genes are shown, with their associated component contribution on the y-axis.

at the gene set determination stage; though we note that the confounders are still incorporated when solving for the MCPs. In the case considered here, we expect that some genes identified by the extrema method may be specific to response to the PDL1 inhibitor rather than chemotherapy, but for the exploratory work we're doing here that's ok.

The extent to which these three methods overlap varies across MCPs. The fraction of multilevel modeling MCP genes which are marked as significant according to the extrema approach is shown in Figure 2.9. Some, but certainly not all, genes appear in both methods—the average percentage of significant genes was 53%. For all three methods we also considered the Jaccard index for the comparison of the gene sets produced by each method. We subset to the 100 most significant genes (lowest p_{adj}) for the extrema method and the 50 genes with the highest positive and 50 genes with the lowest negative loadings (Figure 2.10). The MCP loadings had minimal overlap with the multilevel modeling gene set, but there was some overlap between multilevel modeling genes and the genes from testing extrema cells, particularly in Memory B cells.

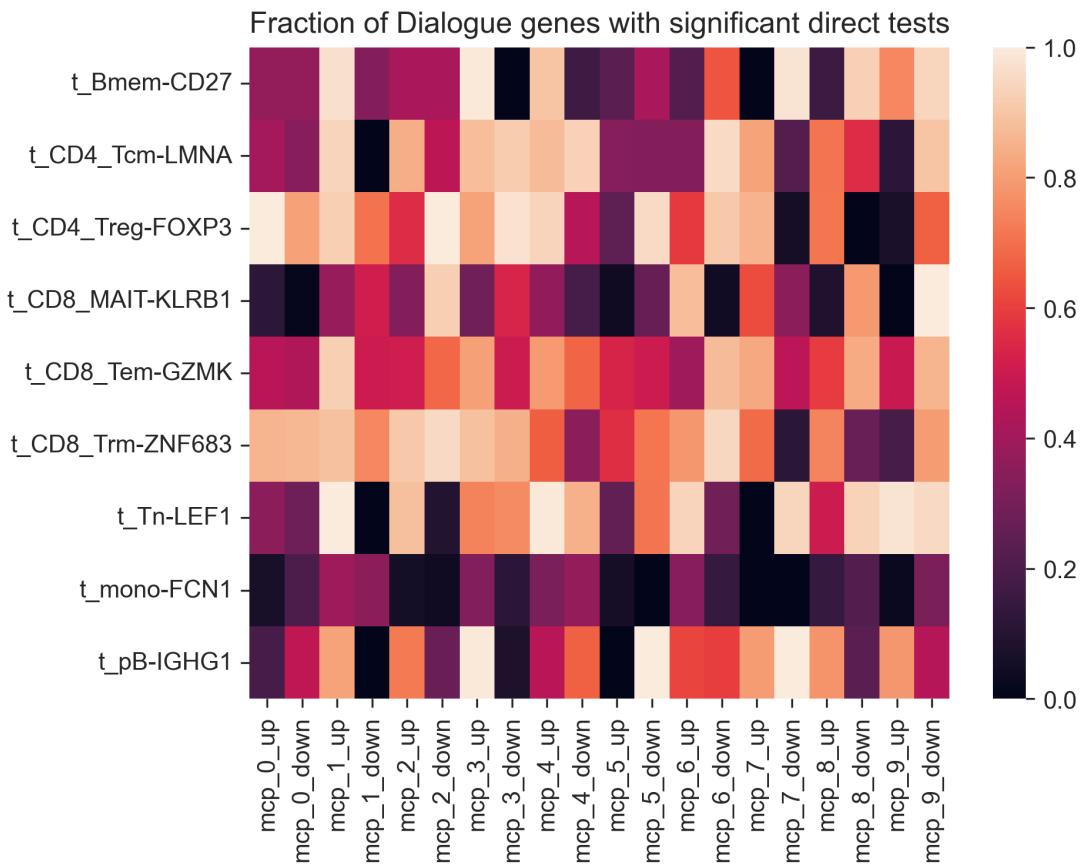


Figure 2.9: The proportion of MCP genes identified via the DIALOGUE multilevel marketing procedure that are also significant (adjusted p-value <0.01) according to differential expression testing of high-MCP and low-MCP cells. Cell type abbreviation expansions are in Table 2.1. The t at the start of each name refers to tumor residence. Along the x-axis are genes increased (up) or decreased (down) for each MCP.

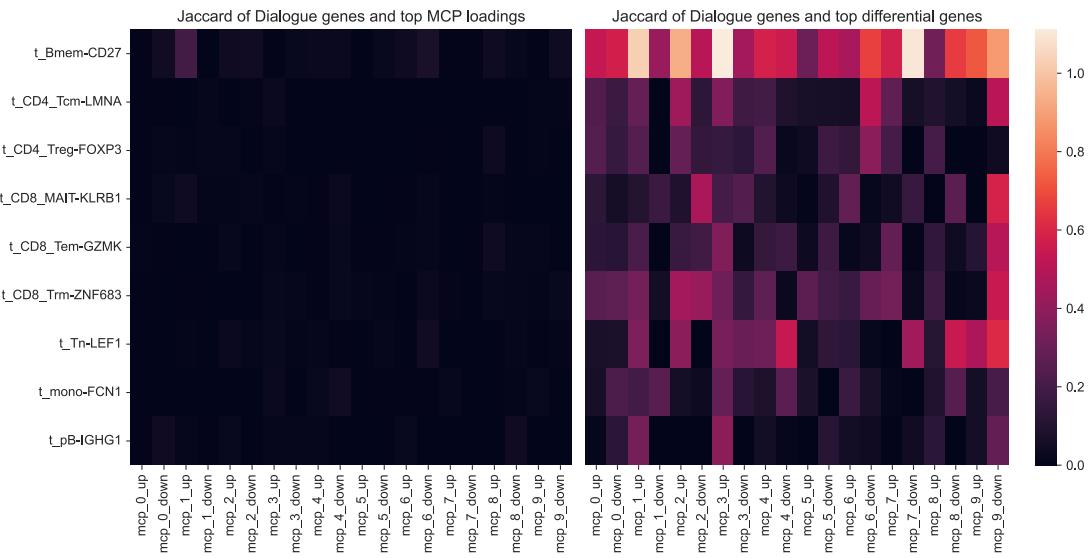


Figure 2.10: Jaccard indices between top MCP genes as determined by DIALOGUE and the MCP loadings (left) and as determined by multilevel modeling or by direct testing (right). The multilevel modeling genes (DIALOGUE genes as determined by the method in ⁹³) have minimal overlap with the MCP loadings, and moderate overlap with genes differentially expressed between top and bottom MCP

Once a set of MCP-associated genes has been established, the resulting gene sets still must be interpreted. For the treatment response-linked MCP₄, we can see the genes with highest increased expression according to extrema testing in Figure 2.11. The full list of MCP-associated genes is provided in Supplemental Table 2.3. The prevalence of heat shock proteins (HSPs) in multiple cell types suggests a role for these proteins in immune cells in cancer progression. HSP1AB, which is significantly increased in this MCP for all five cell types, has been previously identified as a prognostic biomarker in breast cancer⁸³. The connection between HSPs in the tumor micro-environment, in cancer cells, and their corresponding roles in tumor progression, likely merits additional research.

The observed correlation in HSP levels may be due to an external influence, but it could also be related to communication between these cell types. We directly investigated cell-cell communication using a ligand-receptor gene database²⁴. Although there were no ligand-receptor pairs in which both ligand and receptor were included in the MCP genes, there were many cases in which,

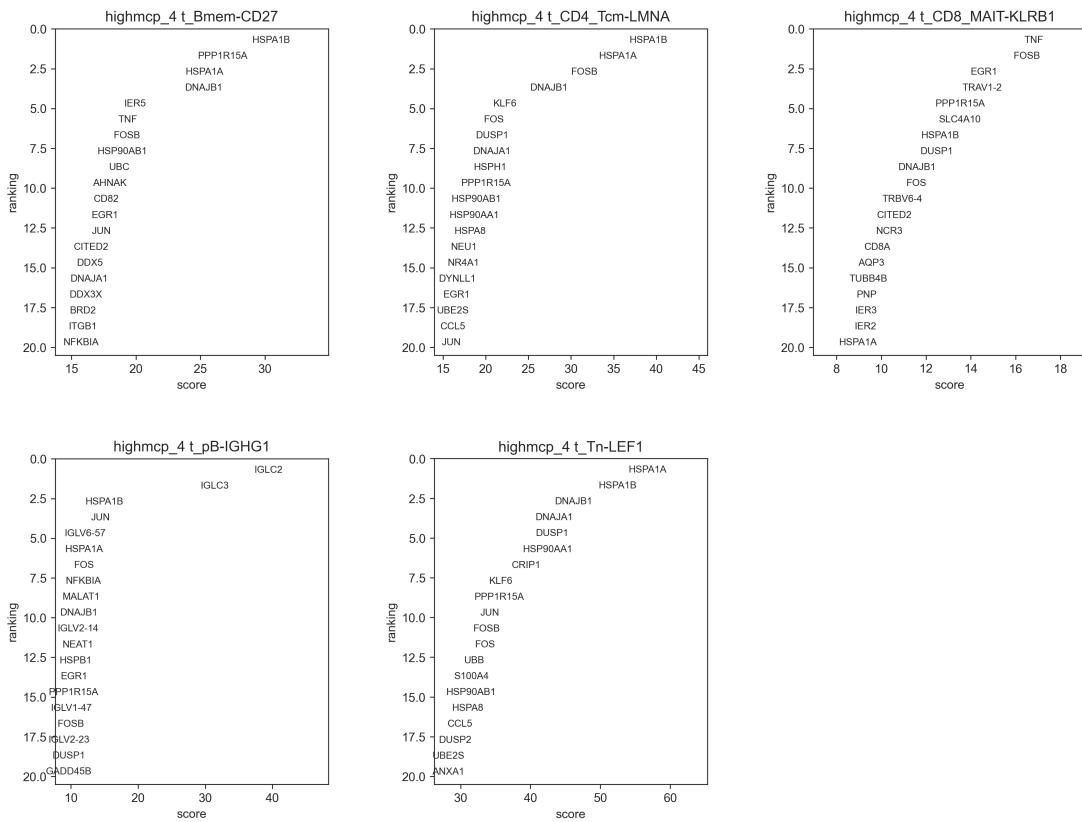


Figure 2.11: Top genes with increased expression in each of the most treatment-response associated cell types in MCP4 of ²²⁰. Rankings are z-scores from t-tests comparing gene expression in the 10% of cells with the highest and lowest MCP4 scores for each cell type.

for example, the ligand expression was MCP associated, and the receptor had a constant, nonzero expression level in another MCP cell type. Of these, the sole observed interaction change between proteins with an experimentally validated interaction is between IL-7 and its receptor IL₇R. There were numerous other possibly enriched interactions between proteins which are predicted to interact, but significant additional experimental work would be needed to validate these findings. IL-7 is a cytokine with well established antitumor role across diverse cancers, though there is also some evidence of pro-tumor activity in lung cancers¹¹⁸. We found that the gene encoding IL-7 is an MCP-associated gene in memory B cells, and its receptor is expressed by central memory T cells and naive T cells. It is plausible that increased IL-7 activity, and the resultant changes in T cells, may also be driving poor treatment response here. Both types of T cells show increased JUN, FOS, and FOSB, all of which are components of the AP-1 transcription factor complex. Intriguingly, AP-1 has been shown to have a complex role in tumor development, with a tumor suppressive or growth enhancing role depending on the context^{61,18}. Lower levels of AP-1 are associated with exhaustion in T cells⁸—the opposite of what one might intuitively expect, in which exhaustion seems to be more prominent in patients who responded to treatment. Whether the identified IL-7 signaling is the causal mechanism underlying the observed correlation is not answerable from the available data; it is possible that all of these adaptive immune cells are responding to some shared signal, which just happens to also trigger communication via IL-7.

2.5 DISCUSSION

By applying and extending a matrix decomposition method, we have proposed disease-relevant cell-cell interactions in the context of triple negative breast cancer, and explored the possibility of a treatment response predictive interaction between B cells and T cells. The clearest limitation of all of this work is that, like scRNA-seq itself, it is best applied for hypothesis generation. We observed poten-

tially interesting findings: a subpopulation of pericytes unique to TNBC, and a multi-cell-type gene expression program involving IL-7 signaling that was increased in those who did not respond to treatment. Deeper analysis of the MCP genes described here, including proposing which predicted cell-cell interactions may be driving behavior, similarly requires additional experimental input. Additional studies are thus needed to confirm these findings; we are actively working with experimental collaborators, both to find additional evidence for these predictions, and to make further discoveries using the data they gather.

Our collaborators' forthcoming experimental results include scRNA-seq of pre- and post-treatment TNBC core needle biopsies. Multimodal imaging-based analysis is also being performed on pre- and post-treatment tumor samples, including both imaging mass cytometry (IMC) and GeoMX spatial transcriptomics. This spatial data can be used to test inferred cell-cell interactions by looking for cells with enriched proximity, and can be used to test the validity of our predictions. One could also confirm specific interactions by staining for relevant proteins, as was done to validate the CytoTalk algorithm⁸⁷. Spatial mapping of breast cancer transcriptomes is still relatively new, but so far indicates that the spatial landscape is highly heterogeneous both within individual tumors and across patients²¹⁸.

The TNBC-specific pericyte subpopulation with high expression of CRABP2 was not observable from the public dataset which included treatment response, as that dataset only sequenced immune cells. Although direct comparison of snRNA-seq and scRNA-seq is not trivial, it may be possible to identify the TNBC-specific pericyte subpopulation using our upcoming snRNA-seq data, which will also include treatment response information. However, the comparison to non-TNBC samples will not be possible, as the current study is only sequencing TNBC patients. We will however, be able to observe how pericyte localization changes in response to treatment, and whether our observed gene signature is present in these tissues.

The proposed multi-cell-type treatment response predictive MCP can be explored using both

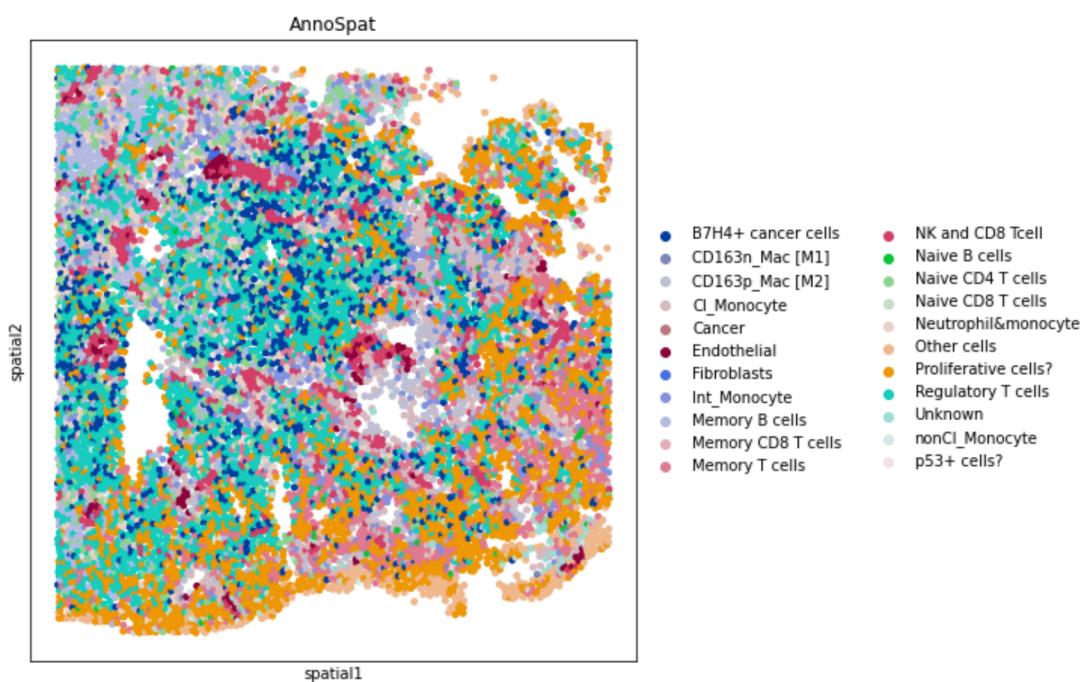


Figure 2.12: An example IMC region of interest with cell type labeling from AnnoSpat. This is from a slice of a core needle biopsy of a pre-treatment TNBC patient.

the scRNA-seq data and the imaging data. We also plan to use the data to potentially discover additional MCPs, as that dataset also includes patients with complete resolution of tumor after treatment. Validating a causal role for cell-cell communication is always difficult, but we may be able to use cell type proximity to suggest cells which are interacting, and observe how those interaction potentials change with respect to treatment efficacy. Using the 35 protein markers measured using IMC, we can label immune cell subtypes in the resulting images. Cell type labeling of IMC data is not a settled question; methods developed for non-spatial CyTOF can be applied with modification, and there are also a few algorithms developed specifically for IMC. Initial labeling of segmented cells in the IMC images was performed using a recently published cell type labeling algorithm which uses a machine learning classifier and a predefined set of positive and negative markers to identify the most likely cell type for each pre-segmented cell¹⁴³. Using just these provisional labels and the example region of interest pictured in Figure 2.12, we find that fibroblasts are enriched at a middle distance from memory B cells (Figure 2.13). This may indicate that cancer-associated fibroblasts are attracting B cells in this tissue, a known CAF-B cell interaction¹⁰⁷. Work is ongoing to optimize the cell segmentation and cell type labeling methods used; we anticipate changes to this result as analysis continues.

Through this work, we have developed a Python re-implementation of DIALOGUE with improved modularity and usability relative to the published R algorithm. This work will become part of the upcoming pertpy package; a developmental release is already available on GitHub⁸⁶. Single-cell transcriptomics methods generate huge amounts of data; increasingly prevalent multi-omics and imaging-based methods generate even more. Interpreting that data and turning it into testable scientific predictions is the job of the bioinformatician in collaboration with the immunologist; this chapter provides an example of using matrix decomposition methods to move from a complex dataset to an interpretable scientific story. Connecting these cell-cell interactions and gene networks to cell state transitions remains an ongoing question, and will depend on our ability to robustly

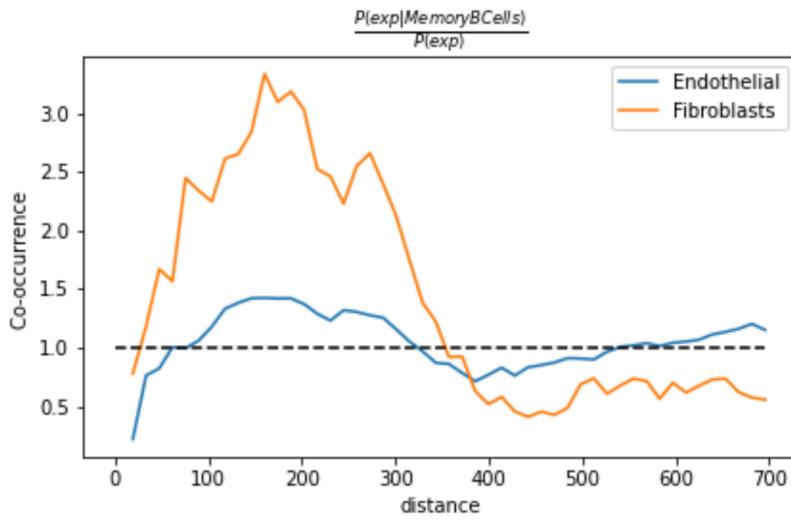


Figure 2.13: Example neighborhood enrichment analysis. In the image shown in Fig. 6, fibroblasts are over-represented in the 100–300 distance range. Neighborhood enrichment statistics were calculated using squidpy¹⁵²

define unified cell states from the upcoming multimodal data.

*Kinds of Light. B.S. Meniscus Films, Ltd. No cast; 16
mm.; 3 minutes; color; silent. 4,444 individual frames,
each of which photo depicts lights of different source,
wavelength, and candle power, each reflected off the same
unpolished tin plate and rendered disorienting at normal
projection speeds by the hyperretinal speed at which they
pass. CELLULOID, LIMITED METROPOLITAN
BOSTON RELEASE, REQUIRES PROJECTION AT
.25 NORMAL SPROCKET DRIVE*

David Foster Wallace, Infinite Jest

3

Point cloud distance metrics for single cell perturbation data

3.1 ABSTRACT

RECENT BIOTECHNOLOGICAL ADVANCES have led to growing numbers of single-cell perturbation studies, which reveal molecular and phenotypic responses to large numbers of perturbations. However, analysis across diverse datasets is typically hampered by differences in format, naming conventions, and data filtering. To facilitate development and benchmarking of computational methods in systems biology, we collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including transcriptomics, proteomics and epigenomics. We apply uniform pre-processing and quality control pipelines and harmonize feature annotations. The resulting information resource enables efficient development and testing of computational analysis methods and facilitates direct comparison and integration across datasets. In addition, we describe E-statistics for perturbation effect quantification and significance testing, and we demonstrate E-distance as a general distance measure for single-cell data. Using these datasets, we illustrate the application of E-statistics for quantifying perturbation similarity and efficacy. The data and a package for computing E-statistics is publicly available at scperturb.org. This work provides an information resource and guide for researchers working with single-cell perturbation data, highlights conceptual considerations for new experiments, and makes concrete recommendations for optimal cell counts and read

depth.

3.2 INTRODUCTION

Perturbation experiments probe response of cells or cellular systems to changes in conditions. These changes traditionally acted equally on all cells by modifying temperature or adding drugs. Nowadays, with the latest functional genomics techniques, single-cell genetic perturbations acting on individual cellular components are available. Perturbations using different technologies target different layers of the hierarchy of protein production. At the lowest layer, CRISPR-cas9 acts directly on the genome, using indels to induce frameshift mutations which effectively knock out one or multiple specified genes^{51,92,57}. Newer CRISPRi and CRISPRa technologies inhibit or activate transcription respectively⁷⁴. CRISPR-cas13 acts on the next layer in the hierarchy of protein production to promote RNA degradation²⁰⁸. Small molecule drugs, in contrast, act directly on protein products like enzymes and receptors. When these techniques are applied to large-scale screens they create a map between genotype, transcriptome, protein, chromatin accessibility, and in some cases phenotype⁶⁷. Barcodes associated with unique CRISPR guide perturbations are read alongside single-cell RNA sequencing (scRNA-seq), cellular indexing of transcriptomes and epitopes (CITE-seq) or single cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) to identify each cell's perturbation condition^{57,67,1,171,153}.

Such large-scale single-cell perturbation-response screens enable exploration of complex cellular behavior inaccessible from bulk measurements. Directionality in regulatory network models cannot be inferred without interventional or time-series data⁸⁰. Experiments with targeted perturbations can be modeled as affecting individual nodes of a regulatory network model, enabling investigation of mechanistic processes and inference of regulatory interactions and their directionality¹⁵⁹. Typically, however, perturbation datasets have been too small to elucidate the complexity of cellular

systems; thus, accurately predictive models of regulatory interactions remain difficult to infer⁷⁹.

This limitation will be reduced as dataset size continues to increase. More directly, drug screens have been used to suggest therapeutic interventions by analyzing detailed molecular effects of targeted drugs, and designing new single or combinations of perturbations^{16,68,160}.

Reliable analysis of increasingly large perturbation datasets requires efficient and powerful statistical tools to harness both massive numbers of cells and perturbations. The inherently high dimensionality of the data complicates calculation of distances between perturbations, as does cell-cell variation and data sparsity¹⁰³. There is presently no convention for statistical comparison in perturbation studies. Some studies calculate pseudo-bulk by combining all cells in a given perturbation^{1,52}. This means losing information about the variation within each condition. Studies with mixtures of cell types have developed complex methods for quantifying similarity between heterogeneous cell populations^{33,49,73}. Cell-based statistical measures can be used to identify successfully perturbed cells but do not currently quantify perturbation similarity^{57,153}. Ideally, statistical comparisons between perturbations and quantification of perturbation strength should be based on a multivariate distance measure between sets of cells. Such a distance measure describes the difference or similarity between cells treated with distinct perturbations, thus inferring unique or shared mechanisms or perturbation targets, which tend to produce similar shifts in molecular profiles^{166,196}. Multiple distance measures for scRNA-seq have been explored by the single-cell community in recent years, including Wasserstein distance^{43,175}, maximum mean discrepancy¹²⁶, neighborhood-based measures^{33,49}, and E-distance¹⁶⁶. Here we exclusively use the E-distance, a fundamental statistical measure of distances between point clouds that can be used in a statistical test to identify strong or weak perturbations as well as to distinguish between perturbations affecting distinct cellular sub-processes. The associated E-test is a statistically reliable tool for computational diagnostics of information content of a specific perturbation and can inform design of experiments and data selection for training models.

Large perturbation screens are specifically designed to study a particular system under a set of perturbations of interest. Over time, the field has thus accumulated a heterogeneous assortment of single-cell perturbation-response data from a wide range of different cell types, such as immortalized cell lines and induced pluripotent stem cell (iPSC)-derived models, and different perturbation technologies, including knockouts, activation, interference, base editing, and prime editing¹⁶¹. Novel computational methods to efficiently harmonize these different perturbation datasets are needed. Such integrative analysis is complicated by batch effects and biological differences between primary tissue and cell culture^{65,129}. Published computational methods for perturbation data are primarily focused on individual datasets^{58,96,127}. Moving from single-dataset to multi-dataset analysis will require development of principled quantitative approaches to perturbation biology; the scPerturb data resource can serve as a foundation for building and testing these models.

While several large databases of perturbations with bulk readouts exist, single-cell perturbation technologies are newer and data not unified^{183,201}. Existing collections of datasets are primarily a means for filtering and do not supply a unified format for perturbations^{113,190,23}. Yet, unified datasets are key to developing generalizable machine learning methods and establishing multimodal data integration. A recent review and repository of single-cell perturbation data for machine learning lists 22 datasets, but supplies cleaned and format-unified data for only six⁹⁴. Unified frameworks for accessing single-cell data are in active development, but do not currently support perturbation datasets or standardize perturbation annotations^{64,38}.

To facilitate the development and benchmarking of computational approaches in systems biology, we provide a resource of standardized datasets reporting targeted perturbations with single-cell readouts. We collected 44 publicly available perturbation-response datasets from 25 papers (Table 3.1). Our perturbation strength quantification and comparison of perturbation-specific variables, such as the number of perturbations and the number of cells per perturbation, across experiments may serve as a reference for optimal experimental design of future single-cell perturbation

experiments. We also describe the E-distance and E-test as tools for statistical comparisons of sets of cells and benchmark their robustness and applicability for distinguishing perturbations across datasets and modalities. A web interface is accessible at scperturb.org, and packages for single-cell E-statistics are publicly available for both Python (PyPI: `scperturb`) and R (CRAN: `scperturbR`).

3.3 METHODS

3.3.1 scATAC-SEQ

DATA ACQUISITION

We included scATAC-seq data from three different sources: Spear-ATAC¹⁵⁷, CRISPR-sciATAC¹¹⁹, and ASAP-seq¹⁴⁰. All data that was used in our analysis can be programmatically downloaded with scripts that are provided in our code repository (<https://github.com/sanderlab/scPerturb>).

scATAC-seq is a biomolecular technique to assess chromatin accessibility within single cells^{30,46}. The starting point of our data processing pipeline are BED-like tabular fragment files, in which each line represents a unique ATAC-seq fragment captured by the assay. Each fragment is mapped to a genomic interval and a cell barcode. The goal of our pipeline is to extract standardized features from this information.

Those are:

- Embeddings derived from latent semantic indexing (LSI)⁴⁶ with 30 dimensions for each cell (a dimensionality reduction method that is well-suited for the sparsity of the data)
- Gene scores that measure the chromatin accessibility around each gene for each cell (the weighted sum of fragment counts around the neighborhood of a gene's transcription start site where more distant counts contribute less)

Source Paper	Modality	Perturbation type	Perturbations
Adamson ¹	RNA	CRISPRi	9, 20, 114
Aissa ³	RNA	drugs	4
Chang ³⁹	RNA	drugs	4
Datlanger ⁵²	RNA	CRISPR-cas9+TCR [†]	97
Datlanger ⁵¹	RNA	CRISPR-cas9+TCR [†]	48
Dixit ⁵⁷	RNA	CRISPR-cas9	31
Frangieh ⁶⁷	RNA,protein	CRISPR-cas9	249
Gasperini ⁶⁹	RNA	CRISPRi	43314*, 39087*, 16531*
Gehring ⁷³	RNA	drugs	4
Liscovitch-Brauer ¹¹⁹	ATAC	CRISPR-cas9	22,84
McFarland ¹³⁷	RNA	drugs, CRISPR-cas9	18
Mimitou ¹⁴¹	ATAC,protein	CRISPR-cas9	6
Norman ¹⁴⁹	RNA	CRISPRa	237
Papalex ¹⁵³	RNA,protein	CRISPR-cas9	11,99
Pierce ¹⁵⁷	ATAC	CRISPRi	41,41,41
Replogle ¹⁶⁶	RNA	CRISPRi	2058, 2394, 9867
Schiebinger ¹⁷⁵	RNA	cytokines	2,3
Schraivogel ¹⁷⁶	RNA	CRISPR-cas9	3105*, 4115*
Shifrut ¹⁷⁹	RNA	CRISPR-cas9+TCR [†]	49
Srivatsan ¹⁸¹	RNA	drugs	5, 8, 189
Tian ¹⁹⁷	RNA	CRISPRi	27
Tian ¹⁹⁶	RNA	CRISPRa, CRISPRi	101, 185
Weinreb ²⁰⁷	RNA	cytokines	5
Xie ²¹⁴	RNA	CRISPR-cas9	229
Zhao ²²³	RNA	drugs	7

Table 3.1: Dataset information for experiments included in the scPerturb database. *: perturbation total treats perturbations A, B, and (A and B) as three unique perturbations

[†]: T-cell receptor (TCR) stimulation

- A peak-barcode matrix that quantifies the chromatin accessibility at (data-set specific) consensus peaks (genomic intervals) for each cell
- chromVAR scores¹⁷⁴, which quantify the activity of a set of transcription factors for each cell, using transcription factor footprints as defined in²⁰³
- Marker-peaks per perturbation target, quantifying the differential regulation of highly variable peaks for each type of perturbation

These features were computed using the ArchR framework version 1.0.1⁷⁷ with standard parameters unless otherwise stated. We provide each feature set as a dedicated h5ad file on scperturb.org, and our analysis roughly follows the pipeline proposed in Spear-ATAC¹⁵⁷, as detailed below.

Note that these features were originally developed for scATAC-seq data on non-perturbed cells, with goals such as the identification of cell types, discovery of cell type-specific regulatory elements, or reconstruction of cellular differentiation trajectories^{30,173}.

PRE-PROCESSING

Filtering out cells of low quality: To ensure a consistent and homogenous quality throughout the different data sets, we filtered out cells with fewer than 1000 and more than 100,000 mapped fragments. We further required a minimum transcription start site enrichment score of 4 to ensure a sufficient signal-to-noise ratio. See ArchR’s `createArrowFile` function for details.

For the Spear-ATAC data set we ran ArchR’s `getValidBarcodes` function on processed 10x Cell Ranger files to subset the data set to valid barcodes. For the other datasets these files were unavailable, and we relied on the original authors’ pre-processing of barcodes.

Assigning single guide RNA (sgRNA) to barcodes: For the Spear-ATAC and CRISPR_sciATAC datasets we had access to cell barcode-sgRNA count matrices (see original publications for details). We assigned the sgRNA with the highest counts to a cell barcode if the sgRNA count exceeded 20

and if that sgRNA combined at least 80% of all sgRNA counts. Cells that could not be assigned a sgRNA were left in the data set. For the ASAP-seq dataset a barcode-sgRNA matrix was not available. Instead, we relied on a sgRNA assignment downloaded from the study's GitHub repository¹⁴¹.

FEATURE COMPUTATION

All features described in the overview above were computed with ArchR functions. For details inspect the `fragments2outputs.R` script in our code repository.

COMPARATIVE ANALYSIS

Processing prior to comparison was performed partially specific to the four non-perturbation-specific analysis methods [marked in square brackets], based on the shape and range of the corresponding data. Log1p refers to transforming X to $\log(1 + X)$. PCA refers to principle component analysis (PCA). HVG refers to subsetting to 2000 highly variable genes (HVGs):

- **chromVAR**: quantification of transcription factor activity. [Log1p, PCA]
- **LSI**: 30 reduced dimensions per cell. These dimensions were used directly as input to E-distance calculations. [no pre-processing]
- **Gene scores**: the weighted sum of fragment counts around the neighborhood of a gene's transcription start site; [Log1p, HVG, PCA]
- **Peak barcode (peak bc)**: peak locations shared across cells are learned from the data and then quantified for each cell. [Log1p, HVG, PCA]

E-distances and Pearson correlations were calculated as described for scRNA-seq.

3.3.2 scRNA-seq

DATA ACQUISITION

Datasets were downloaded from public databases following data availability directions in the source papers. When available from the authors, unnormalized pre-processed cell-by-gene matrices were used. Supplemental information from the papers were used in data analysis when applicable.

DATA PROCESSING

Analysis started from unfiltered, unnormalized cell-by-gene matrices as provided by source papers. For one dataset, preprocessed cell-by-gene matrices were unavailable; pre-processing was performed following the procedure outlined in the original paper, directly using supplied code⁷³. For datasets with cell barcodes, barcode assignments for cells were taken from the original paper when available; when not available, barcode assignment was performed as described in the methods section of the relevant paper. If multiple guides were assigned to the same cell, the guides were listed in decreasing order of counts in the final data object. The code used for processing each individual dataset, including barcode assignment, is available in our code repository.

Datasets were imported into AnnData objects using Scanpy (versions 1.7.2–1.9.1)^{2,12}. Metadata was taken from the original papers when available. For cell lines, information on sex, age, disease, and origin were taken from Cellosaurus¹⁰. Metadata columns are described in Supplemental Table 3.1. Items listed in bold are included for all datasets.

Datasets are stored and supplied as .h5ad files.

3.3.3 SIMULATED DATA

scRNA-seq Simulations used powsimR, a simulator which allows specification of the number of differentially expressed genes (DEGs) (width) and the log₂-fold change (L₂FC) of those DEGs (depth)²⁰⁴. The proportion of cells within each group that are actually perturbed can also be varied. We set this percentage to 2/3 in order to mimic the fact that not all cells in CRISPR-cas9 screens with a guide assigned will have received an effective perturbation.

DATA ANALYSIS

Before calculating E-distances, cells and genes were filtered using Scanpy (versions 1.7.2–1.9.1)²¹². All .h5ad objects published on the resource were saved using Scanpy 1.9.1. Cells were kept if they had a minimum of 1000 unique molecular identifier (UMI) counts, and genes with a minimum of 50 cells. 2000 highly variable genes were selected using `scanpy.pp.find_variable_genes` with flavor `seurat_v3`. We normalized the count matrix using `scanpy.pp.normalize_total` and log-transformed the data using `scanpy.pp.log1p`; We did not z-scale the data. Next, we computed PCA based on the highly variable genes. The E-distances were computed in that PCA space using 50 components and Euclidean distance. To avoid problems due to different numbers of cells per perturbation, we subsampled each dataset such that all perturbations had the same number of cells. We removed all perturbations with fewer than 50 cells and then subsampled to the number of cells in the smallest perturbation left after filtering. Large parts of our analysis were parallelized as workflows using snakemake¹⁴⁶. For applications of E-distance to datasets with confounding factors such as batch effect, we recommend correcting for these factors prior to PCA.

For the example application to CITE-seq, cell type annotations `celltype.l2` were used as provided by⁸¹. Doublets were removed and data was subset to 91 cells per remaining cell type, which is the largest number such that all key cell types had at least that many cells. After subsetting, the

data was processed as for other RNA datasets. Protein data was centered log ratio (CLR) normalized using Muon 0.1.2 and log-transformed prior to PCA²¹. The hierarchy was computed using `scipy.cluster.hierarchy.linkage` from `scipy 1.8.0` with method “single”. The distance metric is “squared_euclidean” and the E-distance was not bias corrected. “2000 HVG E-distance” refers to highly variable genes selected as described above, and is a typical method for gene selection. We also used this HVG-based approach prior to PCA by default in all other E-statistics calculations.

When comparing gene selection methods, we used both default HVGs as described above and an augmented HVG set. To calculate the “union of DEGs”, we used t-tests (as implemented by `scanpy.tl.rank_genes_groups`²¹²) for each perturbation in the dataset to extract the top 50 perturbation-wise DEGs relative to unperturbed cells, then took the set of the union of those genes as features for PCA.

In robustness analysis, at each subsampling point we computed detailed E-statistics (E-distances, δ , σ , E-test results) from each perturbation to the corresponding unperturbed cells of that dataset using PCA with 50 components based on 2000 highly variable genes unless otherwise specified. We downsampled raw UMI counts using the function `scanpy.pp.downsample_counts` on raw counts, then preprocessed (normalized, log1p-transformed) the data as previously described. Cells were downsampled to the same number at each subsampling step across all perturbations to avoid comparability issues. If possible, we recalculated PCA while keeping the highly variable genes originally obtained from the complete dataset. Loss of significance was computed as a running loss of E-test significance ($p\text{-value} < 0.05$) of formerly—i.e. prior to any subsampling—significant perturbations while subsampling, then normalized across datasets through division by the total number of formally significant perturbations in that datasets.

3.3.4 E-DISTANCE

The E-distance is a statistical distance between high-dimensional distributions and has been used to define a multivariate two-sample test, called the E-test¹⁶⁹. It is more commonly known as energy distance, stemming from the original interpretation using gravitational energy in physics. Formally, it contextualizes the notion that two distributions of points in a high-dimensional space are distinguishable if they are far apart compared to the width of both distributions. More specifically,

Let $x_1, \dots, x_N \in \mathbb{R}^d$ and $y_1, \dots, y_M \in \mathbb{R}^d$ be samples from two distributions X, Y corresponding to two sets of N and M cells respectively.

We define

$$\delta_{XY} = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \|x_i - y_j\| \quad (3.1)$$

$$\sigma_X = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^M \|x_i - x_j\| \quad (3.2)$$

and σ_Y defined accordingly. We used the squared euclidean distance when calculating cell-wise distances. Intuitively, δ_{XY} is the mean distance between cells from the two distributions, while σ_X describes the mean distance between a cell from X to another cell from X . The energy distance between X and Y is defined as:

$$E(X, Y) = 2\delta_{XY} - \sigma_X - \sigma_Y \quad (3.3)$$

For the bias-corrected E-distance, we define

$$\sigma_X = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^{M-1} \|x_i - x_j\| \quad (3.4)$$

All calculations use the bias-corrected form of the E-distance unless otherwise noted.

3.3.5 E-TEST

The E-test was performed as a Monte Carlo permutation test using the E-distance as test statistic. For each dataset and each perturbation within that dataset, we took the cells and combined them with the unperturbed cells. Then, we shuffled the perturbation labels and computed the E-distance between the two resulting groups. We repeated this process 10000 times. The number of times that this shuffled E-distance to unperturbed was larger than the unshuffled distance divided by 10000 yields a p-value, which we report for almost all datasets in our resource (Supplemental Table 3.2). We corrected for multiple testing using the Holm-Sidak method in each dataset.

3.3.6 SUBSAMPLING ANALYSIS

We investigate the robustness of E-distance and E-test scores to experimental and computational parameters using our extensive collection of harmonized single-cell perturbation datasets. We subsampled the number of cells per perturbation to create artificially smaller datasets, then examined how the E-distance and E-test results change. We introduce a novel bias correction to the E-distance which improves performance in low cell count regimes (details in Appendix A). Even after bias correction the E-distance increases as the number of cells per perturbation decreases, indicating that cells per perturbation should be standardized via subsampling prior to calculating E-distances (Figure 3.1A). This is due to the inability of PCA to adequately represent data in the low sample regime⁹⁹. Despite the increase in E-distance with falling cell numbers, the number of significant perturbations correctly decreases with fewer cells, and only some datasets have saturated significance at full number of cells in that dataset (Figure 3.1B). This saturation point depends on perturbation strength and on dataset heterogeneity; if all cells are similar, a small set of cells will sufficiently describe every possible response to a perturbation. This suggests that, unsurprisingly, increasing sample size enables discovery of significant perturbations with smaller magnitude.

Similarly, we subset the number of UMI counts per cell, finding that E-distance increases as the number of UMI counts per cell increases (Figure 3.1C). The number of significant perturbations under the E-test, though, saturates around 500 counts per cell, with most perturbations that were significant at the full measured read depth maintaining that significance even with far fewer counts per cell (Figure 3.1D). The stability of E-test results with respect to UMI counts, in contrast to the actual E-distance value, exemplifies the necessity of the E-test as the appropriate statistical measure to evaluate perturbation effects. The optimal UMI and cell counts for a given experiment depend on downstream specific modeling tasks, as discussed in more detail elsewhere⁷⁹. As a baseline for significant perturbations, as defined by the E-test, we suggest at least 300 cells per perturbation (Figure 3.1B) and 1000 average UMI counts per cell (Figure 3.1D) as an experimental guideline for distinguishable perturbations.

More detailed robustness analysis of E-statistics is available in Appendix B.

3.3.7 ADVICE FOR SINGLE CELL PERTURBATION ANALYSIS

Resource users should be aware that memory requirements quickly become a limiting factor, especially with the newer, larger datasets, such as ReplogleWeissman2022 with > 2.5 million cells across more than 9000 perturbations¹⁶⁶. For example, the E-distance presented here for calculating distances between perturbed sets of cells relies on PCA, but computing PCA for all data in this dataset was not possible with 500GB of memory without modifications to accelerate computation. Going forward, computational methods will need to be modified as in⁵⁴ to reduce memory load, or datasets will need to be subsampled. Additionally, the .h5ad datasets shared in this resource can be programmatically accessed using Python package h5py, and perturbations of interest extracted without requiring full dataset access.

To our knowledge, there are not yet established best practices for analysis of single-cell perturbation data. DESeq2 is frequently used for differential expression testing, as it can be applied to

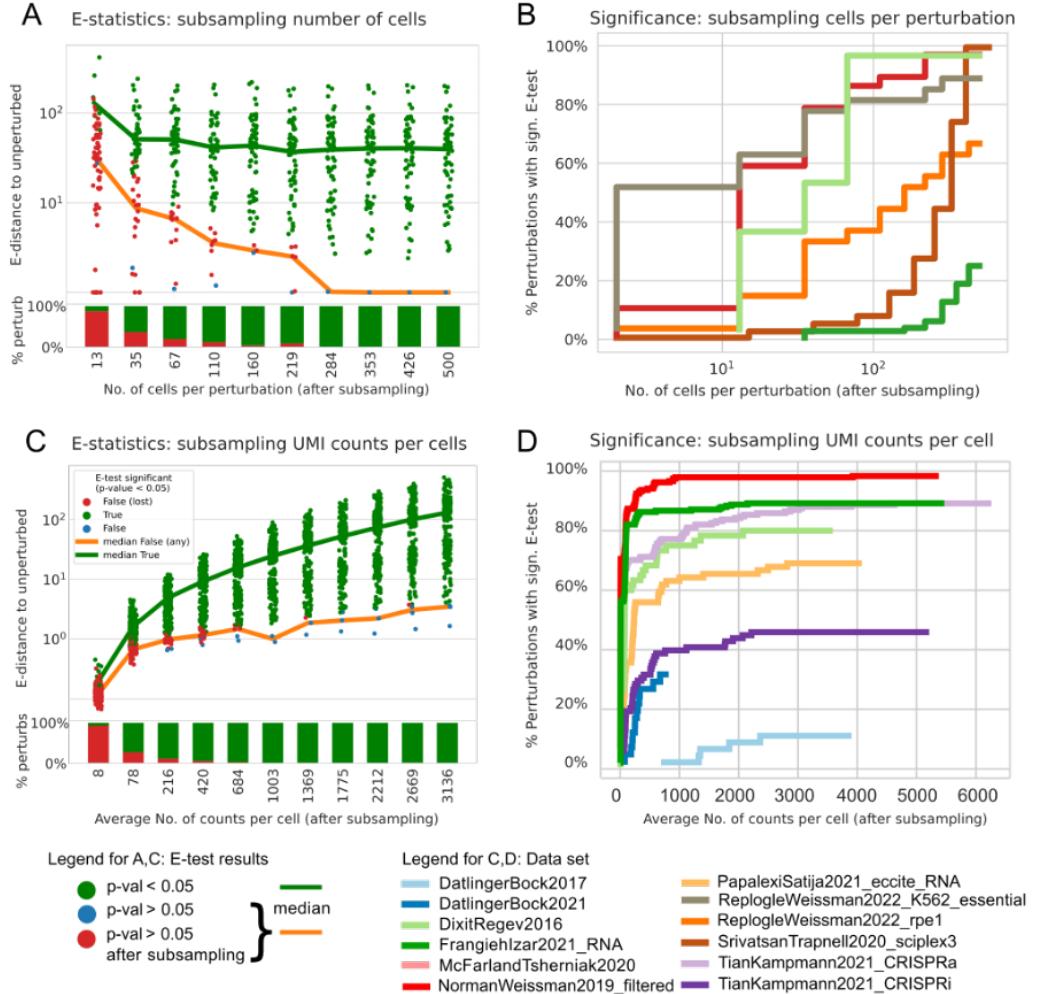


Figure 3.1: Effect of subsampling UMI counts per cell and number of cells per perturbation on E-statistics. (A) E-distance of each perturbation to unperturbed in ¹⁴⁹ while subsampling the number of cells per perturbation; Color indicates E-test results; “significance lost”: perturbation significant when all cells are considered, but not significant after subsampling. The E-test loses significance with lower cell numbers while the E-distance actually increases. (B) Overall number of perturbations with significant E-test decreases when subsampling cells. (C) As in Subfigure A but subsampling UMI counts per cell while keeping the number of cells constant. Loss of E-test significance and dropping E-distance to unperturbed as overall signal gets deteriorated with removal of counts. (D) As in Subfigure B but subsampling UMI counts per cell while keeping the number of cells constant.

pseudo-bulk profiles of each perturbation¹²⁸. An optional next step would be enrichment analysis of the resulting genes. Averaging single-cell measurements over cells per perturbation simplifies analysis and reduces the effect of measurement noise significantly but comes at the cost of removing all system-intrinsic biologically relevant information in cell-to-cell variation. In many studies, these average profiles are then embedded using a dimensionality reduction method of choice and subsequently clustered to reveal groups of perturbations with potentially similar targets^{149,166}.

3.3.8 DATA AVAILABILITY

The website scperturb.org stores harmonized datasets with the following:

- scRNA-seq and antibody-based protein datasets: .h5ad files.
- scATAC-seq: multiple different feature matrix definitions as separate download options.
- Access details for the original publication for each dataset
- Filtering, e.g., by readout or type of perturbation
- RNA data at <https://zenodo.org/record/7041849> and ATAC data at <https://zenodo.org/record/7058382>

3.3.9 CODE AVAILABILITY

Open access source code is at <https://github.com/sanderlab/scPerturb/>. We compiled a corresponding Python package called `scperturb` for performing E-statistics (E-distance and E-testing) in single-cell data, published on PyPI under <https://pypi.org/project/scperturb/> and on CRAN under <https://CRAN.R-project.org/package=scperturbR>.

3.4 RESULTS

3.4.1 DATA CHARACTERISTICS

Molecular readouts for the 44 single-cell perturbation response datasets include transcriptomes, proteins and epigenomes (Table 3.1, Figure 3.2A). Metadata was harmonized across datasets (Supplemental Table 3.3). 32 datasets were perturbed using CRISPR and 9 using drugs. While 32 datasets measure scRNA-seq exclusively, we also include scATAC-seq from three papers, one with simultaneous protein measurements¹⁴⁰. For each scRNA-seq dataset we supply count matrices, where each cell has a perturbation annotation as well as quality control metrics. Three CITE-seq datasets have protein and RNA counts separately downloadable^{67,153}.

In contrast to scRNA-seq data, which can be represented naturally as counts per gene, there is no consensus feature set for scATAC-seq. In its raw form, scATAC-seq provides a noisy and very sparse description of chromatin accessibility over the entire genome. Following prior studies, we generated five feature sets of scATAC-seq data, each of which address different biological questions^{41,77,157}. These attempt to summarize chromatin accessibility information over different types of biologically relevant genomic intervals (e.g. gene neighborhood), or represent dense low-dimensional embeddings of the original data^{30,46,174}.

Sample quality measures vary significantly across datasets (Figure 3.2B). The total number of cells per dataset is usually restricted by experimental limitations, though has increased over time. Total UMI counts per cell and number of genes per cell are calculated as described in¹³⁰. These values are used for quality control in data analysis. The average sequencing depth, i.e. the mean number of reads per cell, in each study affects the number of lowly expressed genes observed. Increasing sequencing depth increases the number of UMI counts measured even for lowly expressed genes, reducing the uncertainty associated with zero counts^{82,190}. The overall number of recoverable UMI counts, usually estimated by the sequencing saturation, also depends on the quality of the experi-

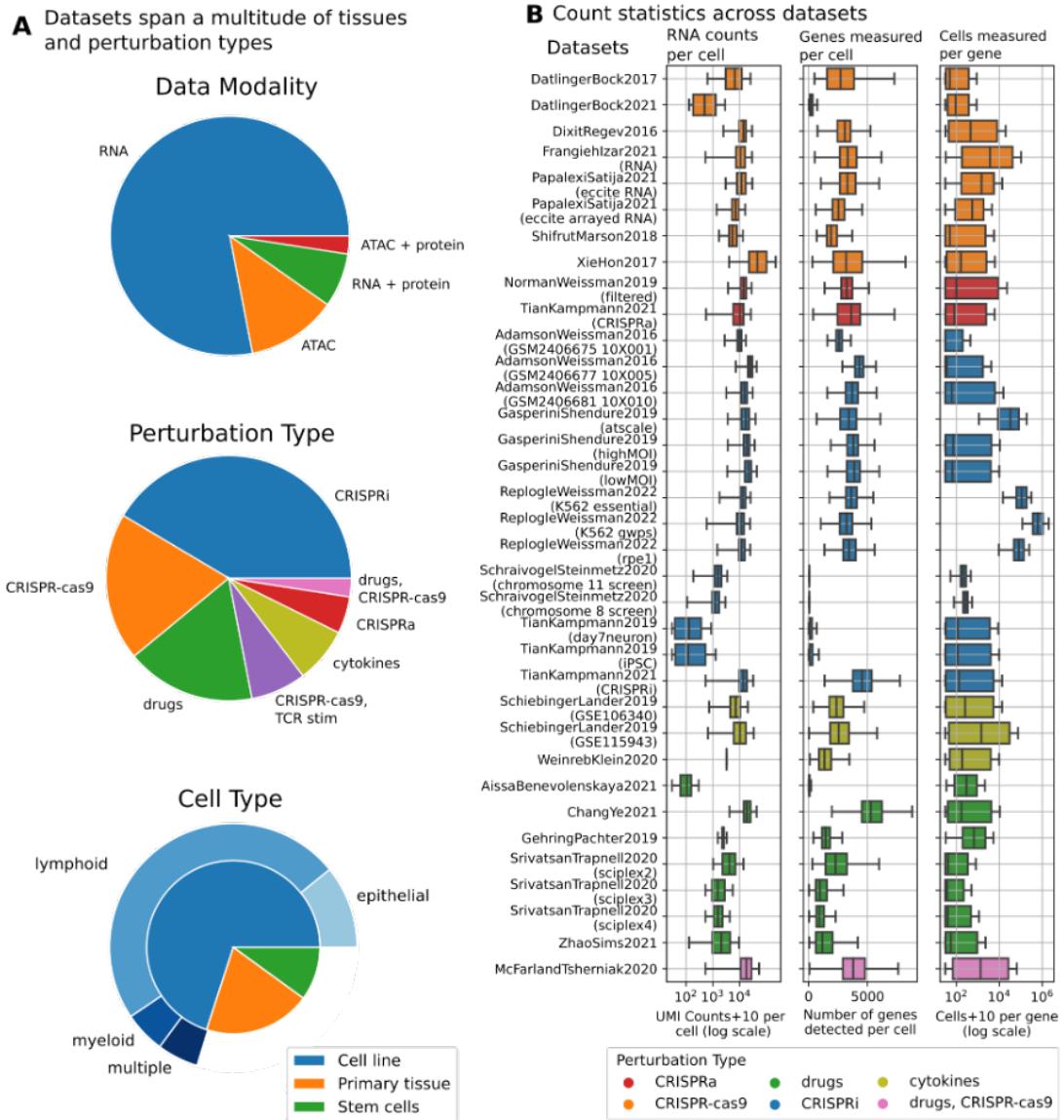


Figure 3.2: Single-cell perturbation-response datasets are diverse in type, size, and quality. (A) The majority of included datasets result from CRISPR (DNA cut, inhibition or activation) perturbations using cell lines derived from various cancers. (B) Sequencing and cell count metrics across scPerturb perturbation datasets (rows), colored by perturbation type. From left to right: distribution of total RNA counts per cell (left); distribution of the number of genes with at least one count in a cell (middle); distribution of number of cells with at least one count of a gene per gene (right). Most datasets have on average approximately 3000 genes measured per cell, though some outlier datasets have significantly sparser coverage of genes. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.

ment; increasing sequencing depth alone can not cover for loss of RNA due to degradation. These differences can affect the distinguishability of perturbations and performance of downstream analysis methods.

3.4.2 E-STATISTICS

E-DISTANCE

To compare and evaluate perturbations within each dataset we utilized the E-distance, a statistical distance measure between two distributions, which provides intuition about the signal-to-noise ratio in a dataset¹⁹³. For two groups of cells, it relates the distance between cells across the groups (“signal”), to the width of each distribution (“noise”) by comparing the mean pairwise distance of cells across two different perturbations to the mean pairwise distance of cells within the two distributions (Figure 3.3A). If the former is much larger than the latter, the two distributions are distinct. A low E-distance indicates that a perturbation did not induce a large shift in expression profiles, reflecting technical problems in the experiment, ineffectiveness of the perturbation, or perturbation resistance. Similar to¹⁶⁶, we compute the E-distance after PCA for dimensionality reduction. The standard E-distance as described in¹⁶⁹ is a biased estimator and increases for low cell counts. Here, we introduce a novel bias-correction to E-distance calculation, analogous to Bessels’ correction to the sample variance. The bias correction factor and a demonstration of unbiasedness of the new estimator is described in detail in Appendix A.

E-TEST

The E-distance can also be used as a test statistic to assess whether cells after a perturbation are significantly different from unperturbed cells. The E-test is a Monte Carlo permutation test that uses the E-distance as a test statistic¹⁹³. The exact value of the E-distance depends on dataset-specific pa-

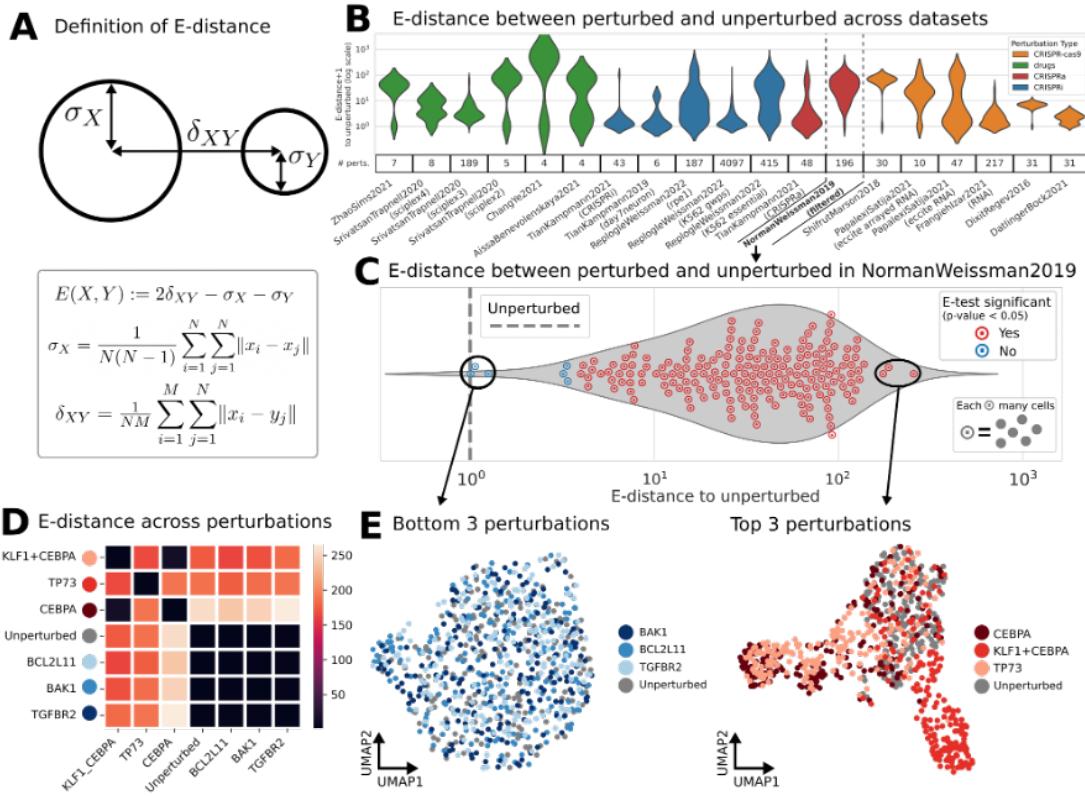


Figure 3.3: (A) Definition of E-distance, relating the width of cell distributions of high-dimensional molecular profiles to their distance from each other. A large E-distance of perturbed cells from unperturbed indicates a strong change in molecular profile induced by the perturbation. (B) Distribution of E-distances (plus 1 for log scale) between perturbed and unperturbed cells across datasets. The number of perturbations per dataset is displayed along the bottom. Note that this plot is best used to compare the shape of the E-distance distribution rather than the magnitude; the mean E-distance will vary significantly with other dataset properties. (C-E) Analysis based on E-statistics for one selected dataset¹⁴⁹: (C) Distribution of E-distances between perturbed cells and unperturbed cells as in subfigure B. Each circled point is a perturbation, i.e., represents a set of cell profiles. Each perturbation was tested for significant E-distance to unperturbed (E-test). (D) Pairwise E-distance matrix across the top and bottom 3 perturbations of Figure 3C and the unperturbed cells. (E) uniform manifold approximation and projection (UMAP) of single cells of the weakest (left, bottom 3) and strongest (right, top 3) perturbations.

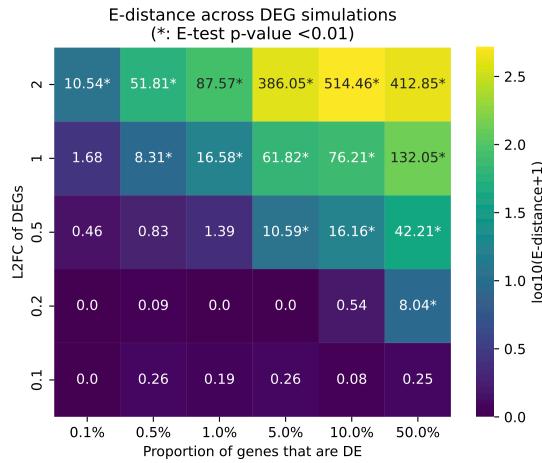


Figure 3.4: Behaviour of E-distance and E-test in simulated data. Varying L₂FC of DEGs and proportions of genes that are simulated as differentially expressed. E-test significance (adjusted p-value < 0.01) marked with “*”.

parameters such as sequencing depth and how the cells are distributed; the E-test accounts for these differences by creating a null distribution using permutations of the data. The computational complexity of the E-test procedure is described in Appendix C.

A SIGNAL PROCESSING PERSPECTIVE ON PERTURBATION WIDTH AND DEPTH

Perturbations of central nodes in signaling networks such as key transcription factors or signaling hubs like p53 can affect the expression of many other factors (a “wide” perturbation). On the other hand, perturbing a relatively isolated pathway or a node with few other nodes downstream may only affect the expression of a low number of features (a “narrow” perturbation). The magnitude of these changes is then the “depth” or “amplitude”. Width can be roughly quantified by the number of DEGs, and depth by the average log-fold change (LFC) of the affected genes. Using powsimmmR ²⁰⁴, we simulated scRNA-seq data while varying these factors to systematically evaluate the E-distance and the sensitivity of the E-test.

It is important to note that while the maximum L₂FC of DEGs can be fairly high, the average

L_2 FCs of all genes affected by a perturbation will rarely be as high as in our simulations. That said, we observe a stronger effect of perturbation depth on E-statistics compared to width (Figure 3.4). This means that according to the E-test, a small number of genes with a large L_2 FC in expression is more impactful than a perturbation which affects a large number of genes by a small amount.

Visualizing these same sets of points using UMAP, we see that all perturbations which are visually distinguishable in the UMAP are also E-test significant (Figure 3.5). Some perturbations which visually look fairly similar, such as 50% DEGs with an average L_2 FC of 0.2, are also significant. UMAP projection is a useful tool for visualizing high dimensional data like this, but can result in spurious patterns and is not advised for use in clustering and other bioinformatics tasks⁴⁰. E-statistics are a tool which can do the same distinguishability check in the full PCA space rather than just by eye in the UMAP.

3.4.3 APPLICATIONS OF E-STATISTICS

DATASET EXPLORATION

Interestingly, we found that E-distances between perturbed and unperturbed cells vary significantly across datasets (Figure 3B). The dataset labeled with “NormanWeissman2019”¹⁴⁹ had the largest mean E-distance between all perturbations compared to datasets of similar size. In fact, expression profiles of most perturbations in this dataset were significantly different from those of unperturbed cells according to the E-test (Figure 3.3C). Plausibly, this is in part caused by two-target perturbations using CRISPRa in that dataset: targeting the same gene with two single guides increases the chances of causing an observable change in the transcript profile. Indeed, the three perturbations with highest E-distance are double perturbations while the three closest in E-distance are not. The corresponding UMAPs for these perturbations, computed using the same principle components (PCs) as the E-distance, provide a confirmatory visual intuition for high and low E-distances

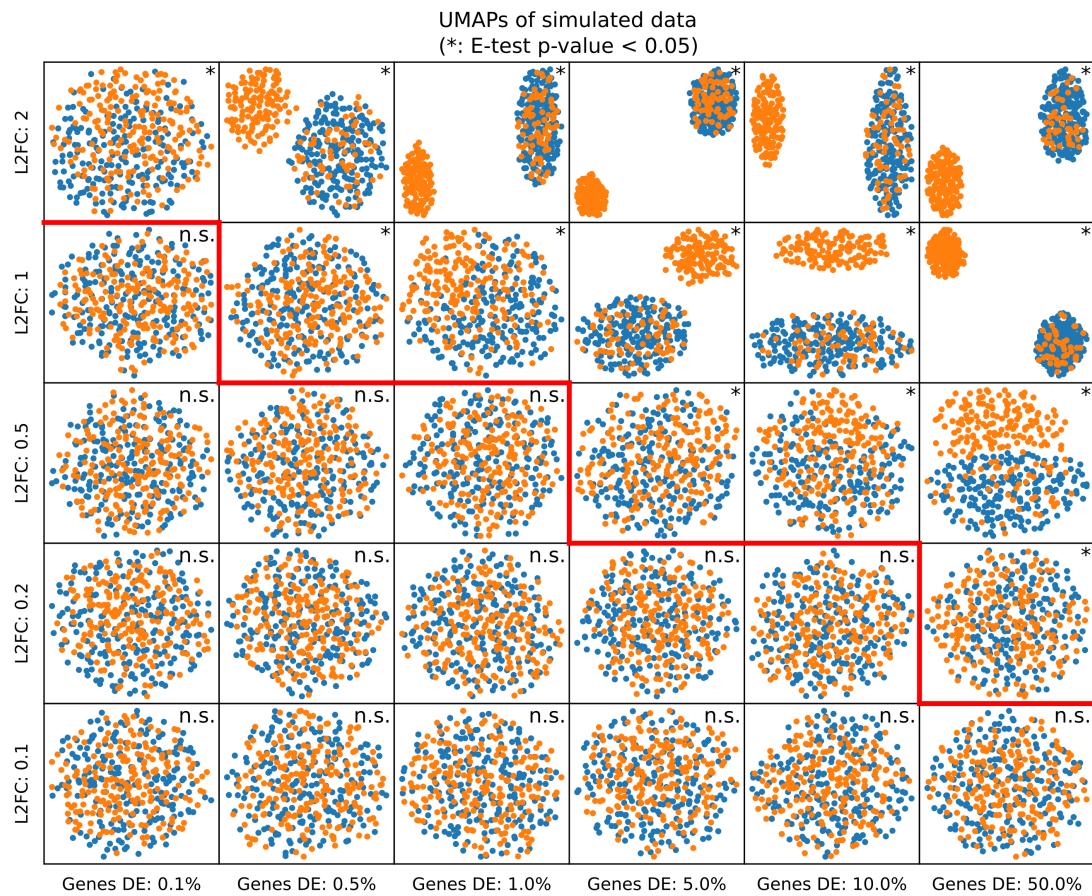


Figure 3.5: Relationship between E-test significance and visual separation in two UMAP dimensions. UMAPs for simulated data corresponding to Figure 3.4. Note that we forced 1/3 of cells in the non-control group (orange) to have no phenotype change in order to mimic CRISPR-cas9 data. Significance of E-test marked in top right corner of each UMAP as in Figure 3.4. * indicates significant E-test results, n.s. not significant. Genes DE is the number of genes differentially expressed. The red line denotes the boundary of significance loss.

(Figure 3.3E). The top three perturbations causing the largest E-distance to unperturbed are easily distinguishable from the gray unperturbed cells, while the bottom three weakest perturbations are part of a single, uniform cloud virtually indistinguishable from the unperturbed cells. The smallest E-distance thus results from perturbations which have the least effect on the distribution of cells.

The E-distance can also be used to measure similarity between different perturbations. For instance, there is a clear overlap of CEBPA and KLF1+CEBPA perturbed cells in the UMAP (Figure 3.3E). This overlap is captured by the low E-distance between the two perturbations; these two perturbations are closer to each other than they are to unperturbed cells or to other perturbations (Figure 3.3D). We envision that the E-distance can be used as a suitable distance for other downstream tasks such as drug embeddings and clustering of perturbations, which could allow inference of functional similarity of perturbations by similarity in their induced molecular responses measured by the E-distance.

Using E-statistics, we can nominate a few particularly notable datasets in the resource. The most extensive drug dataset is sci-Plex 3, which includes 188 drugs tested across three cell lines¹⁸¹; 107 of those perturbations were significant according to E-test analysis (Supplemental Table 3.3). Five drugs in this dataset also appear in other drug perturbation datasets (Supplemental Table 3.4). We hope that future large-scale drug screens will enable more detailed analysis of drug response across different cell types and conditions. Another drug dataset applies combinations of three drug perturbations at varying concentrations across samples⁷³. The most detailed CRISPR dataset is from a recently published study which perturbed 9867 genes in human cells¹⁶⁶. Containing >2.5 million cells, this dataset is the largest in our database, with the number of cells each gene is detected in significantly higher than in other datasets. Notably, 138 CRISPR perturbations are seen in both scRNA-seq and scATAC-seq datasets (Supplemental Table 3.5). More than 100 genes perturbed with CRISPRa in one dataset are perturbed with CRISPRi perturbations in another dataset of the same cell line, either in one paper¹⁹⁶ or across multiple studies^{149,166}. The most frequently per-

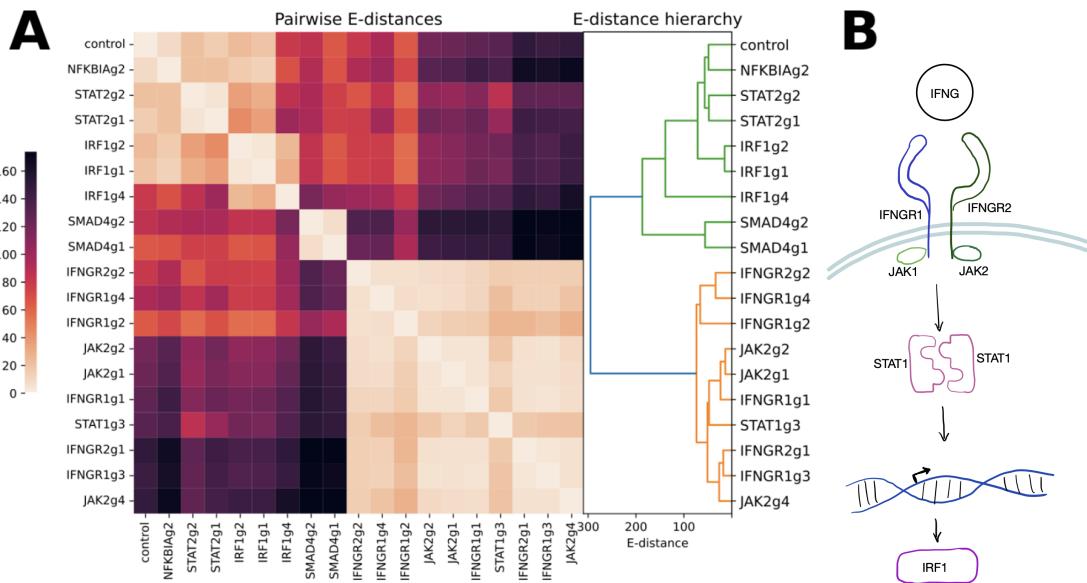


Figure 3.6: E-distance dissects perturbation hierarchy in data from Papalex et al. (A) E-distance between cells of all pairs of perturbations in Papalex et al. dataset¹⁵³. Hierarchical clustering of this matrix reveals two groups, one which is more similar to unperturbed cells (green) and one which has a stronger transcriptional change (orange). (B) Signaling pathway downstream of IFNG receptor. Permutations of nodes upstream of IRF1 induce similar phenotypes.

turbed gene, MYC, is perturbed in 9 datasets from 3 papers. Protein, RNA and ATAC readouts for CRISPRi perturbation of MYC are all available for K562 cells^{67,157,166}.

IDENTIFICATION OF SIMILAR PERTURBATIONS

As a detailed demonstration of the power of E-distance for analyzing perturbation datasets, we calculated pairwise E-distance between all pairs of perturbations in a dataset characterizing inhibitory immune checkpoints (Figure 3.6). This study perturbed genes involved in regulation of PD-L1 using CRISPR-cas9¹⁵³. Hierarchical clustering of the resulting distance matrix revealed two distinct groups of perturbations. The perturbations in the group more dissimilar to unperturbed cells have a low E-distance to each other, suggesting a similar phenotype induced by perturbation of these genes (IFNGR1, IFNGR2, JAK2, STAT1). Indeed, these genes are all part of a signaling cascade upstream

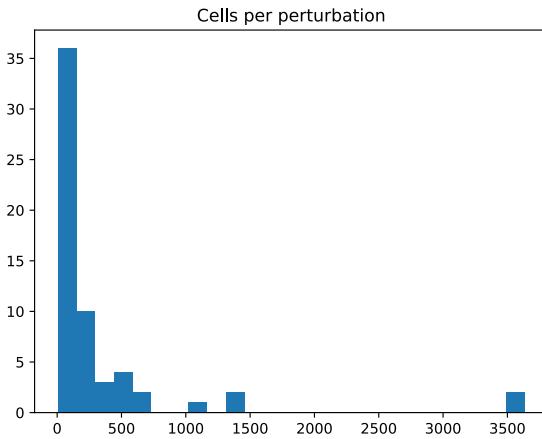
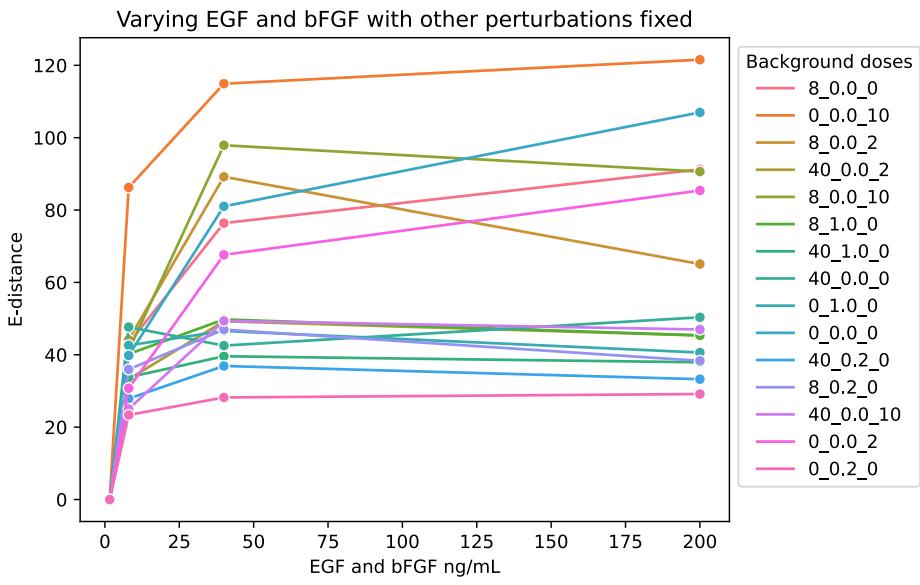


Figure 3.7: Histogram of cells per perturbation in⁷³. Many perturbations had fewer cells than our standard recommendation.

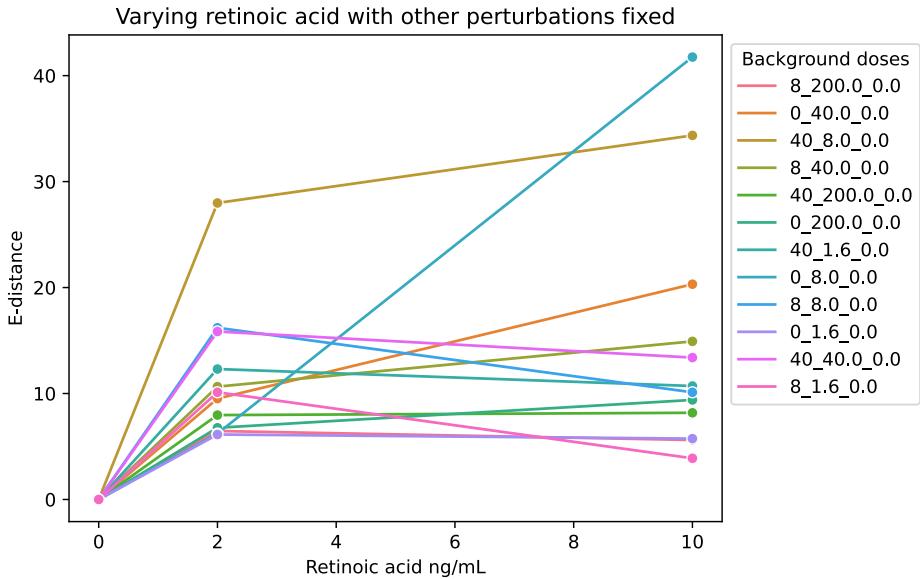
of IRF1 and downstream of IFNg¹⁷⁷. Thus, we observe that any disruption of a gene in this cascade leads to a similar outcome and therefore a similar transcriptome profile.

COMPLEX DATA STRUCTURES

The snakemake pipeline used in the main manuscript cannot be directly applied to datasets with unique perturbation structures. The GehringPachter2019 dataset is a 96-plex tensor of combinations of doses of four different drugs⁷³. For this demonstration, we used all perturbations regardless of how many cells were present; we used the bias-corrected E-distance and do not subsample any conditions. Many of the perturbations had fewer cells than the recommended threshold (Figure 3.7), but results still broadly agreed with the findings from the source paper. Four different doses of epidermal growth factor (EGF) and acbFGF were compared (200 ng/ml, 40 ng/ml, 8 ng/ml and 1.6 ng/ml) against a tensor of different background doses (Figure 3.8a). For each background dose, we take the lowest dose of EGF and basic fibroblast growth factor (bFGF) as a “unperturbed” state, then plot the E-distance from that state as the EGF and bFGF dose is titrated upwards. The



(a)



(b)

Figure 3.8: Background doses are separated by an underscore. For (a), the units are BMP4 (ng/mL); Scriptaid and decitabine(μ L); retinoic acid (ng/mL). For (b), the units are BMP4 (ng/mL); EGF and bFGF (ng/mL), Scriptaid and decitabine(μ L). E-distance is measured from the lowest dose of the x-axis drug on the indicated background set of perturbations. Data from⁷³.

resulting values are clearly interpretable as dose-response curves, where response to EGF and bFGF saturates quickly. This agrees with what the authors of the original paper found, that “Absence of EGF and bFGF has a drastic effect, yielding an isolated group of samples in PCA space”⁷³.

We took the same data and instead projected it along the axis of retinoic acid doses (Figure 3.8b). Here, for each line the “unperturbed” state is the state without added retinoic acid for each experimental background. In most cases, increasing retinoic acid from 2 ng/mL to 10 ng/mL doesn’t significantly effect the E-distance. However, in the case when only retinoic acid and low EGF & bFGF, addition additional retinoic acid continues to increase E-distance. This set of conditions is flagged in Figure 2C of⁷³, where the authors note a “strong conditional dependence” on retinoic acid in this condition. This is explored in more detail in the original work, including cluster analysis and differential gene expression. Here, E-distance analysis quickly reveals that this particular drug combination merits further investigation.

A Jupyter notebook demonstrating this analysis is available on scPerturb Git repo.

COMPARISON OF scATAC-SEQ FEATURE DEFINITION METHODS

Optimal feature definition for scATAC-seq remains an unsettled question, and will depend on plans for downstream analysis^{41,77,157}. For the four feature spaces provided on scPerturb, we examined how feature choice affects the perturbation-to-perturbation distances in two datasets, one with relatively few perturbations (Liscovitch-BrauerSanjana2021-K562-1)¹¹⁹ and one with more (MimitouSmibert2021)¹⁴¹.

On a qualitative level, the resulting pairwise E-distance matrices look very similar in the case of Liscovitch-BrauerSanjana2021-K562-1 (Figure 3.9A) and comparable in the case of the larger dataset MimitouSmibert2021 (Figure 3.9B). To assess potential similarities quantitatively, we computed pairwise Pearson correlations between the E-distances to control across the different feature spaces. While this yields high correlations across all feature spaces for the smaller dataset (Fig-

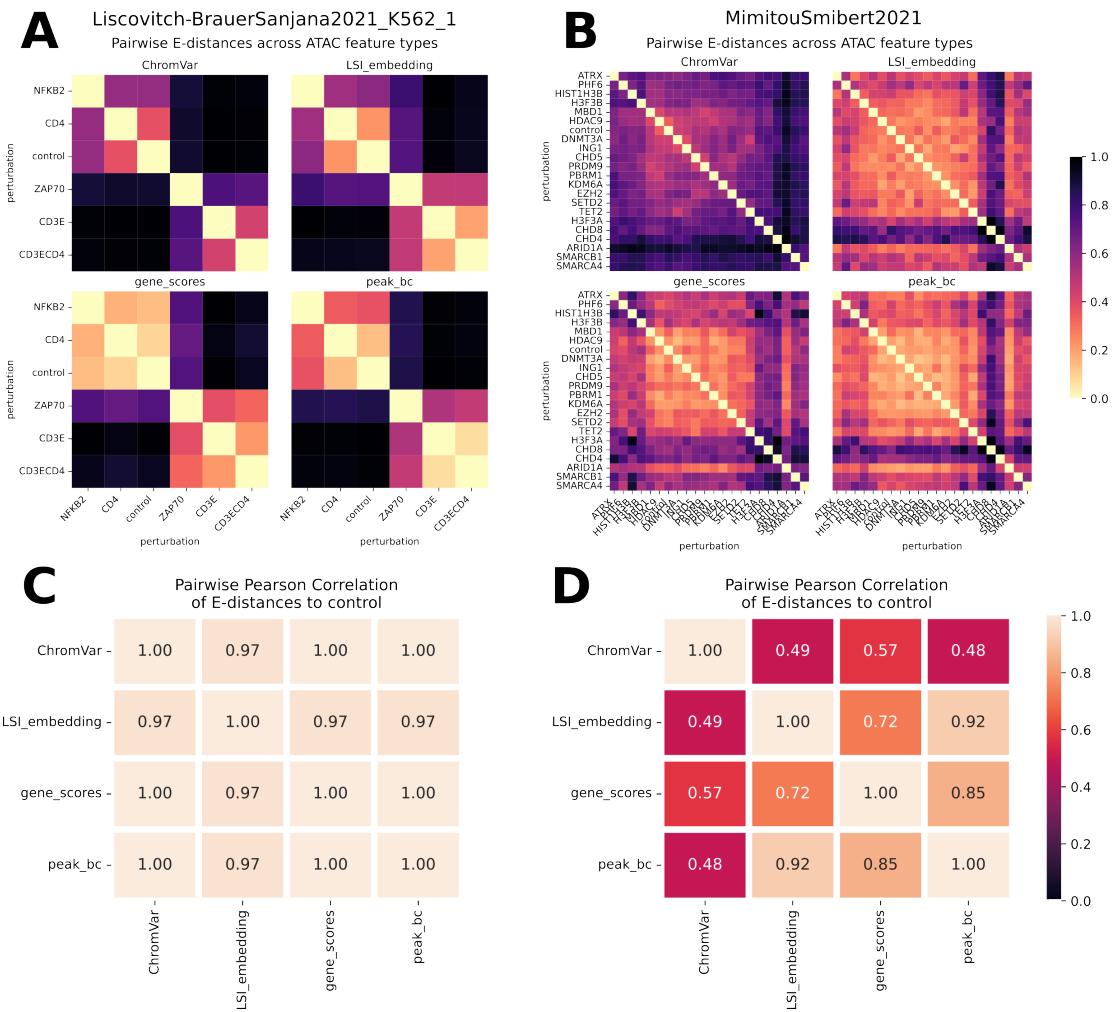


Figure 3.9: (A, B) Pairwise E-distances for two datasets^{119,141} between perturbations and control in different feature spaces defined by four different methods (ChromVar, LSI-embedding, gene scores, and peaks). (C, D) Corresponding pairwise Pearson correlations of E-distances to control across feature spaces.

ure 3.9C), we see less correlation in some cases for the larger dataset: ChromVar, as already evident in the pairwise E-distance heatmap (Figure 3.9B), is the most uncorrelated to the other feature sets. The remaining feature definition methods seem to correlate well with each other, with correlation coefficients between 0.72 and 0.92.

In summary, the correct choice in scATAC-seq feature definition method should be guided by the biological question. For instance, perturbations that affect non-coding regions will be less adequately captured by e.g. gene scores which primarily focuses on regions around transcription start sites of coding regions (genes). Furthermore, the resulting feature spaces serve different purposes and are therefore only partially comparable. For example, the LSI embedding firstly is not meant to provide directly interpretable features and secondly is already so low-dimensional (30 dimensions) such that a PCA is not required nor useful, making it harder to compare E-distances to the other, high-dimensional feature spaces. For a full comparison of scATAC-seq feature definition methods—where different analysis aims were explicitly considered in a more rigorous way—we once again refer to⁴¹.

CELL TYPE DISTINGUISHABILITY

To test whether E-distance values replicate differences between well-known cell types, we applied the E-distance to a CITE-seq human peripheral blood mono-nuclear cell (PBMC) dataset with existing cell type annotations⁸¹. Separately for RNA and protein (from antibody-derived tags), we computed PCA-based E-distances between all pairs of cell types, equivalent to how perturbation E-distance is computed. The resulting pairwise distance matrices were used to compute cell type hierarchies (Figure 3.11, 3.12). We compare this hierarchy to known cell type relationships⁵⁵. The UMAP projection from the original study is in Figure 3.10 for comparison.

In both data modalities, B cell subtypes are clustered together, and platelets are the most distinct from any other cell type. Lymphoid and myeloid cells form two separate groups in the E-distance

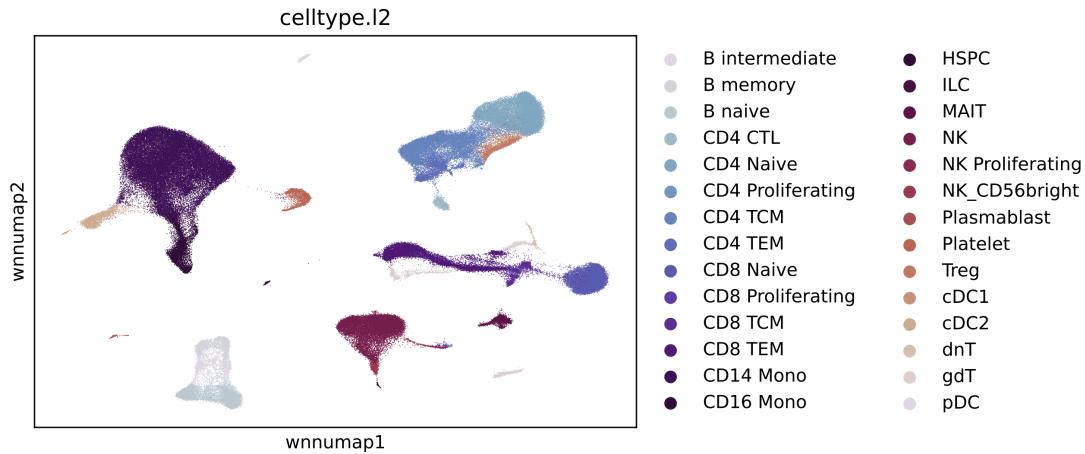


Figure 3.10: Visualization of cell type relationships in full multimodal dataset after batch correction. Coordinates and cell type annotations from ⁸¹

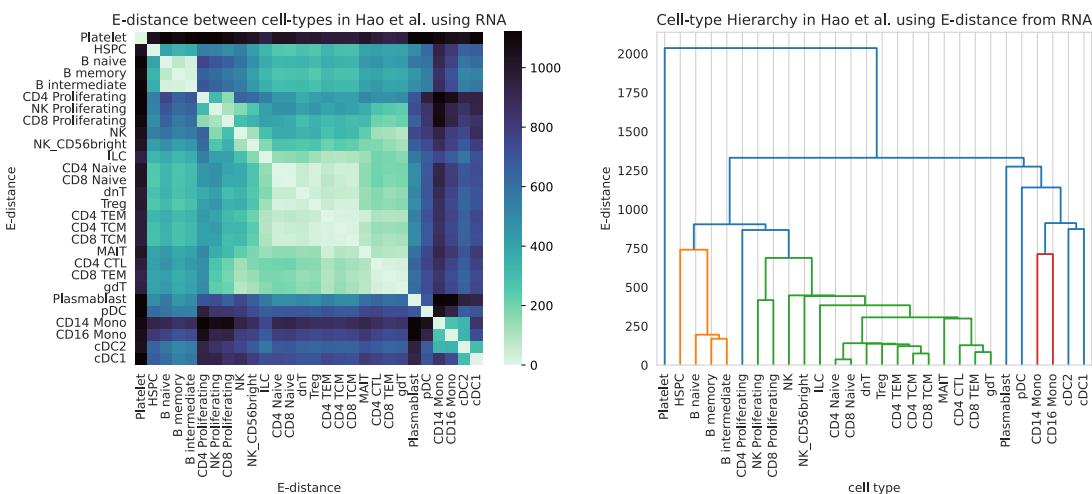


Figure 3.11: Hierarchical clustering of pairwise E-distances computed using RNA matches prior knowledge of transcriptome-defined cell types. Dendrogram and heatmap use the same distances. Data from ⁸¹

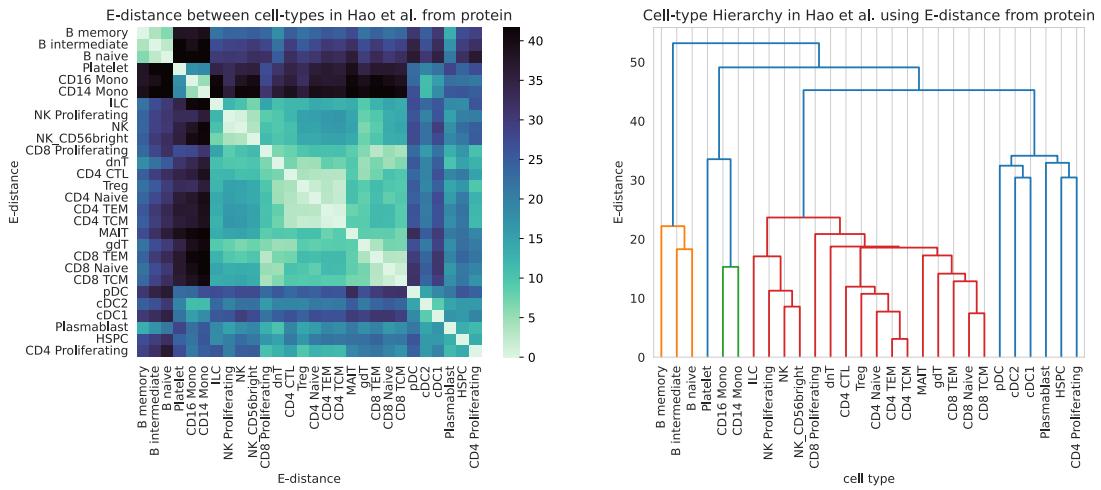


Figure 3.12: As in Figure 3.11 but using antibody-tagged surface proteins instead of RNA.

hierarchy. Notably, innate lymphoid cell (ILC) and natural killer (NK) cell clusters form a distinct group as well. ILCs are innate immune cells that functionally correspond to specific types of classical lymphocytes properly expressing diversified antigen receptors; NK cells are a type of ILC also known as ILC₄ and are functionally similar to cytotoxic T cells^{7,205}. This functional similarity translates to strong similarities in transcriptional profiles, which often leads to difficulties in distinguishing NK cells and cytotoxic T cells in scRNA-seq data. These cell types are more easily disentangled by protein marker based distances, exemplifying the usefulness of CITE-seq as a method for identifying immune cell types. Likewise, when using protein, T cells are clustered primarily by CD4/CD8 type, whereas, when using RNA, they are clustered by functional phenotype (naive, proliferating, memory). For instance, clustering the cells with the E-distance in RNA-space separates proliferating cells of many types into a single cluster, likely due to shared expression of cell-cycle related genes; the cell cycle is known to have a strong effect on the transcriptome profile of cells and is not captured by surface protein measurements.

We conclude by comparing RNA and protein representations that the protein modality more accurately represents cell type differences traditionally defined by immunologists on the basis of

surface proteins, whereas the RNA representation primarily reflects functional programs of the cells such as cytotoxicity or proliferation. In both cases, the E-distance accurately captures known characteristics of each measurement modality.

In this example, E-distances are relatively large, and all cell types are distinguishable. In the case of a smaller sample or more similar cell types, E-tests could be used to identify clusters which might be combined for downstream analysis.

3.5 DISCUSSION

We present a dataset resource and an intuitive method for quantifying and analyzing single-cell perturbation datasets. Processed data with added quality control metrics is available on scPerturb.org. The uniform annotations in this resource enable data integration and benchmarking as well as exploration of shared perturbations across datasets. We introduce a bias-corrected E-distance for quantitatively comparing perturbations. We also investigate the effect of dataset specific parameters on E-statistics, showing that E-statistics stabilize above 1000 counts per cell and 200–500 cells per perturbation.

While this work simplifies dataset access, joint analysis is limited by the complexity of data integration¹²⁹. Across the eight drug datasets examined in this study, only 5 chemical agents occurred in more than one dataset (Supplemental Table 3.4). Shared gene targets are found more often across the CRISPR datasets (Supplemental Table 3.5). However, multiplicity of infection and other conditions also frequently differ; comparisons are further complicated by distinct perturbation methods. Considerable overlap of perturbations across studies makes this a useful resource for benchmarking model generalizability. With more datasets anticipated, we will have the unique opportunity to integrate datasets with more overlapping perturbations and nominate machine learning benchmarks for data integration.

Lack of standardization in data sharing and processing hampered the creation of this resource.

Although many processed datasets were available on the NCBI Gene Expression Omnibus (GEO)^{1,2}, there is no standard format for sharing CRISPR barcode assignments and other metadata. Starting analysis from sequencing reads may have improved interoperability of datasets in this resource, but guide assignment procedures and demultiplexing algorithms are experimental setup specific. For scATAC-seq, data comparison is hindered by the lack of a standard method for feature assignment. In particular, for scATAC-seq feature assignments specific to CRISPR perturbations, known loci-of-action could be used to improve feature calls^{4,1}. In all modalities, many datasets only supplied processed data, or raw data was only available after institutional clearance. Adding more datasets to this resource, or the creation of similar resources in the future, would be easier if there were standard formats for sharing perturbation data, and, more generally, standard formats for sharing single-cell annotations. A community-wide discussion on standardization of such data is urgently needed, as was done for proteomic data^{7,8}.

Experimental design choices such as the recommended minimal number of cells per perturbation and the required sequencing depth for each cell depend on the questions the dataset is intended to answer, and on the strength and uniqueness of the gene expression changes caused by the perturbations. Unfortunately, it is difficult to ascertain to what extent low E-distances between perturbed and unperturbed cells are caused by technical noise. Increasing dose or varying time between perturbation start and harvesting of the cells may be advisable to increase the signal to noise ratio without sequencing more cells. For perturbation distinguishability as defined by the E-test, regardless of experimental parameters, we find that one should have at least 300 cells per perturbation and an average of 1000 UMIs per cell.

We envision that the scPerturb collection of datasets and suggested E-statistics analytic framework will be valuable starting points for analysis of single-cell perturbation data. The unified annotations and perturbation significance testing should prove especially useful to the machine learning

community for training models on this data. We expect new datasets and experimental perturbation methods in the future will enable the community to develop novel computational approaches which exploit the richness of single-cell perturbation data, aiming at the development of increasingly accurate and quantitatively predictive models of cell biological processes and the design of targeted interventions for investigational or therapeutic purposes.

4

Conclusion

This thesis demonstrates the utility of single cell transcriptomics for the study of the immune system in disease; it also makes clear some limitations of this technology. In Chapter 1, we apply transcriptomics to the study of aspirin-exacerbated respiratory disease (AERD). This includes identifying a previously unobserved cell subtype; this form of hypothesis-free discovery of cell states has been one of the most significant applications of single-cell RNA sequencing (scRNA-seq) since its development¹⁹⁹. While this cell state remains of interest in AERD, follow-up experimental study has not

observed any of these cells in surgical samples; whether they might eventually replicate in another study remains to be seen. Separately, we use scRNA-seq to enhance our analysis of bulk RNA-seq, deconvolving bulk samples by cell type and predicting which genes are predominantly expressed by which cell types. These time course samples were the first study to look sequentially at how aspirin desensitization impacts the nasal transcriptome. Despite relatively low sample counts, we were able to observe some consistent shifts in gene expression, including a reduction in expression levels of alpha amylases during the acute aspirin reaction, and a reduction in IL₅RA expression following long-term desensitization. Application of a deconvolution algorithm indicated that these genes are all primarily expressed by ciliated epithelial cells, indicating that even though the nasal epithelium in the inferior turbinate does not produce polyps, polyp-preventing treatment also affects the transcriptome of this tissue. In short, this work suggests that the inferior turbinate is a useful tissue for evaluating drug response in AERD. A separate study by our collaborators used the inferior turbinate as a marker of treatment response to Nucala, an IL-5 inhibitor, and similarly found it a useful marker tissue²⁹.

In Chapter 2, we apply a matrix decomposition algorithm for discovery of multi-cell-type gene expression signatures to predict cell-cell interaction in triple-negative breast cancer (TNBC). We use a breast cancer atlas¹⁵¹ to explore the effects of dimensionality reduction method choice on correlation structure, finding multicellular programs (MCPs) which is reflected in existing literature on TNBC. We also identify a pericyte substate specific to TNBC which is only discoverable when cells were embedded in a latent space using a variational autoencoder (VAE), demonstrating the importance of latent space choice in any scRNA-seq analysis. Using data from a study of TNBC examining treatment response²²⁰, we find an MCP gene expression signature across B cell and T cell subtypes could predict treatment response. This signature includes IL-7 signaling from memory B cells to naive and central memory T cells, where the T cells shared an increase in expression levels of heat shock proteins (HSPs) and multiple subunits of the AP-1 transcription factor complex. In

both cases, the findings are limited by the difficulty associated with confirming cell-cell communication predictions. The matrix decomposition algorithm is able to take complex data and extract out stories; this sense-making is a key feature of any analysis of ultra-high-dimensional data. However, this decomposition does not have any statistical guarantees, and there's no clear means of quantifying robustness of the resulting MCPs.

Seeking more robust statistical measures for moving from single cell data to interpretable biology, in Chapter 3 we investigate a point cloud distance metric which can be used to quantify similarity between cell states and statistically test cell state distinguishability within embeddings. To do so robustly, we create a database of annotation-harmonized single cell perturbation datasets. Using that database, we investigate performance of a distance metric and associated statistical test across a variety of datasets and parameters, identifying the experimental parameters necessary for robustly distinguishable perturbations. This work is already having an impact on the scientific community: in the year since the database was made publicly available we have had more than 4000 data downloads. Our statistical analysis tools are also finding purchase: a recent paper investigating heterogeneity in treatment response among clonal cancer cells used our package to quantify distinguishability between response states⁷⁵. Our statistical work on cell state distinguishability is undoubtedly significant, but it barely scratches the surface of the ongoing need for improving robustness of statistical tools for scRNA-seq.

Moreover, scRNA-seq datasets are growing, and new dataset sizes bring new computational challenges. Standard analysis such as principle component analysis (PCA) is intractable in the ultra-large dataset limit, requiring computational innovations^{115,117}. One of the greatest challenges in producing the work in Chapter 3 was lack of standardized formats for data sharing. Even the basic analytic differential expression pipelines give gene expression differences that vary wildly with small differences in normalization procedures¹⁴⁵. Supposedly standardized analytic steps can also result in divergent count matrices depending on alignment methodologies, though those differences are

fortunately less drastic²⁵. In most cases, studies share only processed data publicly; raw data access often requires extensive bureaucratic processes, and, even when it is shared, basic processing code frequently is not. When fully processed objects are shared, there is no single conventional file type or metadata format; there is no standard method for transferring annotated data objects between R and Python, and existing tools for this transfer can induce surprising errors or result in loss of metadata. Lack of interoperability significantly hampers data access, slowing research and limiting efforts towards multi-dataset analysis.

Even as scRNA-seq analysis remains incompletely solved, experimental techniques connecting scRNA-seq to other forms of single cell observation are increasing prevalent. One limitation we found when using scRNA-seq to study cell-cell communication is that RNA levels for receptors tend to be low, making it difficult to ascertain their presence^{6,9}. In measurements of mammalian tissues, variability in mRNA levels explains only 40% of the variability in protein levels; moreover, this percentage varies widely across different cell types and biological contexts²⁶. This discrepancy is particularly pronounced when examining receptors in scRNA-seq data; low expression of receptors, coupled with high dropout in scRNA-seq, means that receptors are often unobserved^{6,9}. These receptors are the primary means of defining immune cell states in the literature and via prior experimental methods, so computational analysis that enables matching to literature-defined cell states is key—but even now remains an active area of work²⁰⁶. Augmentation of scRNA-seq have been developed in which surface protein levels are assayed using RNA-tagged antibodies which are encapsulated along with their bound cells^{156,187}. A more recently developed method, INS-seq, permeabilizes cells prior to scRNA-seq, and demonstrated the ability to bind fluorescent antibodies and perform FACS pre-sequencing¹⁰². These experimental methods are still new, and computational methods for analyzing their outputs are not fully established. Most methods treat the data types separately during analysis, and focus on using the protein to identify cell types in the scRNA-seq data¹³². One method, totalVI, uses a deep generative model of both RNA and protein measure-

ments to create a joint representation for downstream analysis⁷². Another method learns relative weights for different experimental modalities for producing a kNN graph, which is then used as the basis for further analysis⁸¹. Connecting these joint embedding methods to methods for inferring cell-cell communication is an exciting area for future work, as methods for this analysis that incorporate predictions of single cell protein levels drawn from external datasets are increasingly prevalent^{72,213,225,188}. Innovation will be needed to create computational methods which fully utilize multiple modalities to draw inferences about cellular behavior.

Spatial transcriptomics is another increasingly active experimental space open for innovative computational techniques. Because single cell measurement methods are destructive, such approaches are limited in what they can learn about cellular context⁶². Spatial data maintains this context, allowing for stronger statements about cell-cell interactions and the ability to learn the spatial structure of otherwise-indistinguishable cell states in tissue²⁴. Unfortunately, this is also a space where large file sizes and data sharing procedures can pose a significant issue. Analytic pipelines must extract cells from images, then map expressed genes to cells. Building scientific understanding from these images requires mathematical tools for geometric analysis. Even if one proceeds from processed cell-by-gene matrices as in scRNA-seq analysis, algorithms must use different measurement models and error assumptions than are used when analyzing scRNA-seq data.

As discussed briefly in the conclusion of Chapter 2, we have begun analysis of a spatial proteomics dataset to try to confirm some findings made with published scRNA-seq data. Tasks which have become relatively straightforward in the scRNA-seq world such as cell type annotation must be carried out anew in this space. While some methods exist for automated cell typing¹⁴³, as in scRNA-seq the methods perform relatively poorly—in our hands, half of the observed cells were classed as unknown types, likely because the true biology is imperfectly reflected in the pre-defined list of target cell types. Work is ongoing to improve our cell type assignments, but this is still only considering one of the modalities being captured from these samples. Connecting different imaging modalities

across tissue slices is every bit as important and complex as connecting measurement modalities across modes. There is so much biology that can be learned from these methodologies that was previously completely inaccessible—but getting there is going to require intensive innovations in bioinformatics methods.

I was lucky enough to complete my PhD during the rise of scRNA-seq and the transformation from small niche datasets into rich atlases. Looking forward, methods for sense-making from larger and more complex datasets will be key, as well as methods for combining information across datasets. There are tantalizing hints of the future of computation in this space already. Foundation models trained across mega-datasets of tens of millions of cells appear to be the future of cross dataset analysis^{45,42,195}; fine-tuning of the models to address specific tasks is likely to dramatically alter the field in the years to come. Determining how to evaluate what is learned from those models, and bringing statistical rigor to any deep learning work. It is my hope that work presented here on using single cell transcriptomics to produce interpretable and meaningful biological findings will remain of use as the field moves forward.

A

An unbiased estimator for the E-distance

The standard E-distance, as described by¹⁹³ uses the following formula:

$$\delta_{XY} = \frac{1}{NM} \sum_{i=1}^M \sum_{j=1}^N ||x_i - y_j|| \quad (\text{A.1})$$

$$\sigma_X = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N ||x_i - x_j|| \quad (\text{A.2})$$

and σ_Y defined accordingly. Intuitively, δ_{XY} is the mean distance between cells from the two distributions, while σ_X describes the mean distance between a cell from X to another cell from X . The E-distance between X and Y is defined as:

$$E(X, Y) := 2\delta_{XY} - \sigma_X - \sigma_Y \quad (\text{A.3})$$

Notably, when these summations are used to estimate the E-distance of distributions, the estimates are biased, increasing in the small sample regime. This is because σ decreases for lower number of samples, even though small sample counts should correspond to high uncertainty and therefore should be weighted with high dispersion values. To demonstrate this bias, we simulated samples from a standard normal distribution in 30 dimensions, where all dimensions had the same mean and variance. We used two different distributions corresponding to a control and a perturbed group of cells. Their means differed by 4 and their standard deviations were either 3 and 1 (upper row) or 5 and 3 (lower row). As can be seen in Fig A.1, the naive estimator increases in value for small N. We thus introduce here a sample-corrected estimator for σ :

$$\sigma_X = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N \|x_i - x_j\| \quad (\text{A.4})$$

This sample correction is sufficient to remove bias from the estimator in the case of a squared Euclidean metric, and the corrected E-distance remains stable at low sample numbers. The calculation of δ was not biased to begin with so is not shown. The sample correction accounts for the fact that the entries of the summation with $i = j$, i.e. the diagonal entries of the underlying pairwise distance matrix, are uniformly 0.

In the case where the metric used to compute the E-distance is the squared Euclidian distance, this estimator is also theoretically unbiased, as shown by the following:

Let $x_1, \dots, x_N \in \mathbb{R}^D$ be i.i.d. samples with mean $\mu = (\mu_1, \dots, \mu_D)^T$ and sample variance $S^2 = (S_1^2, \dots, S_D^2)^T$. We show that the sample-corrected σ_X can be expressed using the sample variance S^2 . Let x^k denote the k -th entry of the vector x .

$$\begin{aligned}
& \sum_{i=1}^N \sum_{j=1}^N \|x_i - x_j\|^2 \\
&= \sum_{k=1}^D \sum_{i=1}^N \sum_{j=1}^N (x_i^k - x_j^k)^2 \\
&= \sum_{k=1}^D \sum_{i=1}^N \sum_{j=1}^N ((x_i^k - \mu_k) - (x_j^k - \mu_k))^2 \\
&= \sum_{k=1}^D \sum_{i=1}^N \sum_{j=1}^N ((x_i^k - \mu_k)^2 - 2(x_i^k - \mu_k)(x_j^k - \mu_k) + (x_j^k - \mu_k)^2) \\
&= \sum_{k=1}^D \sum_{i=1}^N \left[N(x_i^k - \mu_k)^2 - 2(x_i^k - \mu_k) \sum_{j=1}^N (x_j^k - \mu_k) + \sum_{j=1}^N (x_j^k - \mu_k)^2 \right] \\
&= \sum_{k=1}^D \sum_{i=1}^N \left[N(x_i^k - \mu_k)^2 - 2(x_i^k - \mu_k)(N\mu_k - N\mu_k) + (N-1)s_k^2 \right] \\
&= \sum_{k=1}^D [N(N-1)s_k^2 - 2(N\mu_k - N\mu_k)(N\mu_k - N\mu_k) + N(N-1)s_k^2] \\
&= \sum_{k=1}^D 2N(N-1)s_k^2
\end{aligned}$$

where we made use of $\sum_{i=1}^N x_i^k = N\mu_k$ and $\sum_{i=1}^N (x_i^k - \mu_k)^2 = (N-1)s_k^2$ which follow directly from their respective definitions.

Ultimately, we have:

$$\sigma_X = \sum_{k=1}^D 2S_k^2 \quad (\text{A.5})$$

Then, since the sample variance S_k^2 is an unbiased estimator for the population variance (i.e. $E[S_k^2] = \sigma_k^2$), the sample-corrected σ_X in Equation A.4 is an unbiased estimator for the dispersion of the multivariate point cloud given by x_1, \dots, x_N . Likewise, the the original definition of σ_X in Equation A.2 is a biased estimator. For δ_{XY} we observe robustness to the number of cells empirically to begin with, so we did not investigate bias of this term.

Despite the theoretical stability of this estimator, when applied to real single-cell data after PCA, we still observe that the E-distance increases for small numbers of cells (see Figure 5 in the main text). This low-n behavior is due to the nature of PCA; the assumptions underlying PCA's ability to fully represent the data break down in the low sample regime⁹⁹. To confirm, we simulated 30 dimensions of data using standard Gaussian, then applied PCA before calculating E-distance. As expected, we observe strong deviations in both δ and σ attributed to PCA (Fig A.2). The estimators converge at 100 samples. As such, at least 200 cells per condition is likely sufficient for a stable distance measure.

Unless otherwise noted, all E-distance calculations reported in the main paper and in this supplement use the squared Euclidean metric and the bias correction factor.

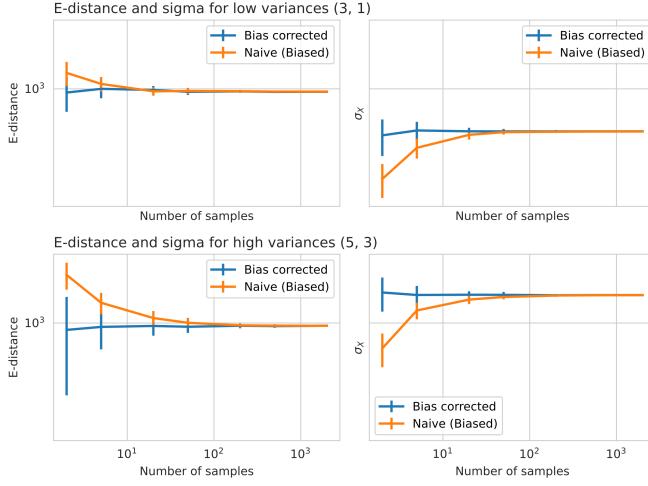


Figure A.1: Sample correcting the calculation for σ removes count bias of the E-distance with respect to sample size. The naive estimator uses Equation A.2, and the bias corrected uses Equation A.4. Vertical lines show standard deviation across 50 simulation runs. The mean of σ and E-distance remain confident as fewer data points are sampled. Each of the 30 dimensions used in this example is an independent Gaussian with the same mean and variance.

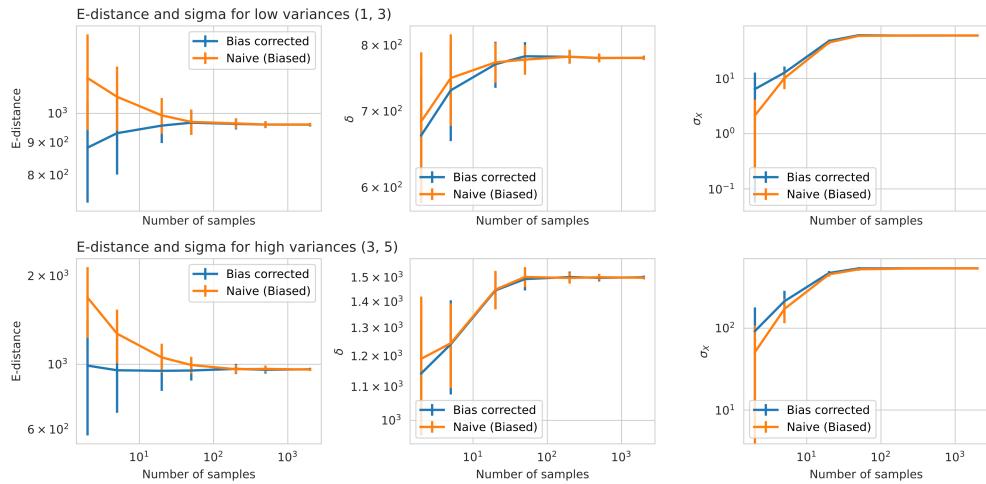


Figure A.2: PCA prior to E-distance biases results at low n . Both sigma and delta change due to PCA. Convergence at around 100 cells. Dimensions were independent.

B

Robustness Analysis of E-statistics

Most CRISPR perturbation studies sequence many more unperturbed cells than perturbed ones.

Moreover, although we equalized perturbation cell counts for benchmarking, in a real-world dataset this may not be feasible or may result in discarding too many cells (as in the example analysis of⁷³).

In order to explore the effect of asymmetric sample sizes, we varied the number of control cells while keeping the number of cells per perturbation fixed at 200 cells. For each perturbation in¹⁵³, we calculated the E-distance to control and evaluated E-test significance (Figure B.1). Subsampling and

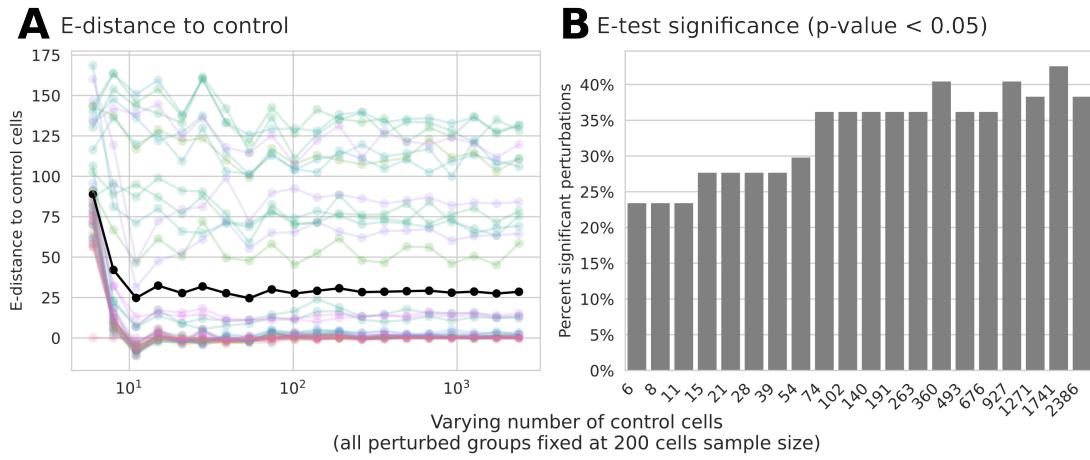


Figure B.1: Asymmetric sample sizes: as long as both conditions have at least 200 cells, E-distances and E-tests are well-behaved. For varying number of cells in the control group and fixed ($n=200$) number of cells per perturbed group we recorded for each perturbation the (A) E-distance to control cells and (B) the number of perturbations with a significant E-test ($p\text{-value} < 0.05$). Black line corresponds to average over all perturbations, all colored lines to different perturbations. RNA data from ¹⁵³.

PCA was performed independently for each control cell count. As previously described for symmetric sample sizes (Figure A.2), the E-distance increased in the ultra-low-cell regime (Figure B.1A). Increasing the number of control cells above 200 did not consistently increase the number of perturbations flagged as significant (Figure B.1B).

Thus, E-statistics are empirically robust to asymmetric sample sizes as long as each sample consists of at least 200 cells. Since the control population is used as reference for every single statistical test for each perturbation respectively, it is crucial to include at least this many control cells when designing an experiment.

We evaluated the impact of feature selection and feature counts on E-distance by changing the which genes are used to compute PCA (Figure B.2A). highly variable genes (HVGs) were computed using the 'seurat_v3' flavor in `scipy.tl.highly_variable`. We investigated this comparison using two drug datasets with varying numbers of perturbations ^{153,223}. In both cases, distances are largely stable above 2000 HVGs (Figure B.2A,B). Moreover, for most datasets the number of HVGs used

to compute the principle components (PCs) had almost no effect on E-testing above 500 HVGs (Figure B.4B). The SchraivogelSteinmetz2020¹⁷⁶ is TAP-seq, so has fewer genes measured than all other datasets. The faster decrease in significance observed in this dataset indicates stronger sensitivity on the number of PCs with fewer features available. Based on this analysis, we use 2000 highly variable genes prior to PCA throughout.

Notably, the algorithm used to identify highly variable genes does not take into account perturbation or cell labels. In a prior work applying E-distance to single cell RNA-seq data, highly variable genes were enriched by including additional differentially expressed genes for each perturbation¹⁶⁶. Particularly in cases where the number of perturbations is large, or if the effect of each perturbation is small, this selective enrichment could be necessary to evaluate distinguishability of more modest perturbations. To test this, we compared E-statistics outcomes across multiple ways of selecting variable genes on two datasets^{153,166}. As can be seen in Figure B.2C and Figure B.2D, augmented gene lists have little impact on the calculated E-distances after PCA.

In the “hybrid” approach, the set of differentially expressed genes (DEGs) was used to augment the list of HVGs, as described in¹⁶⁶. Although the inclusion of DEGs is conceptually attractive, DEG identification is not a settled question; there is poor overlap of DEGs across methods⁹¹ and single-cell DEGs are not a consistent measure of effect size¹⁸⁰. Depending on the dataset and the heterogeneity of perturbations in the dataset the number of unique genes in the union of 50 top DEGs varies substantially (Figure B.3). Due to this inconsistency, the challenge of properly performing DE testing, and its minimal impact on observed distances, we recommend using the 2000 HVG E-distance.

We also examined how choices made in computing the PCs used in distance calculations affects E-statistics. The number of PCs used from PCA to compute the E-distance had a moderate effect on E-test results, mildly decreasing the number of significant perturbations (Figure B.4A). Interestingly, E-test significance was lost most rapidly in a TAP-seq (targeted Perturb-seq) experiment¹⁷⁶.

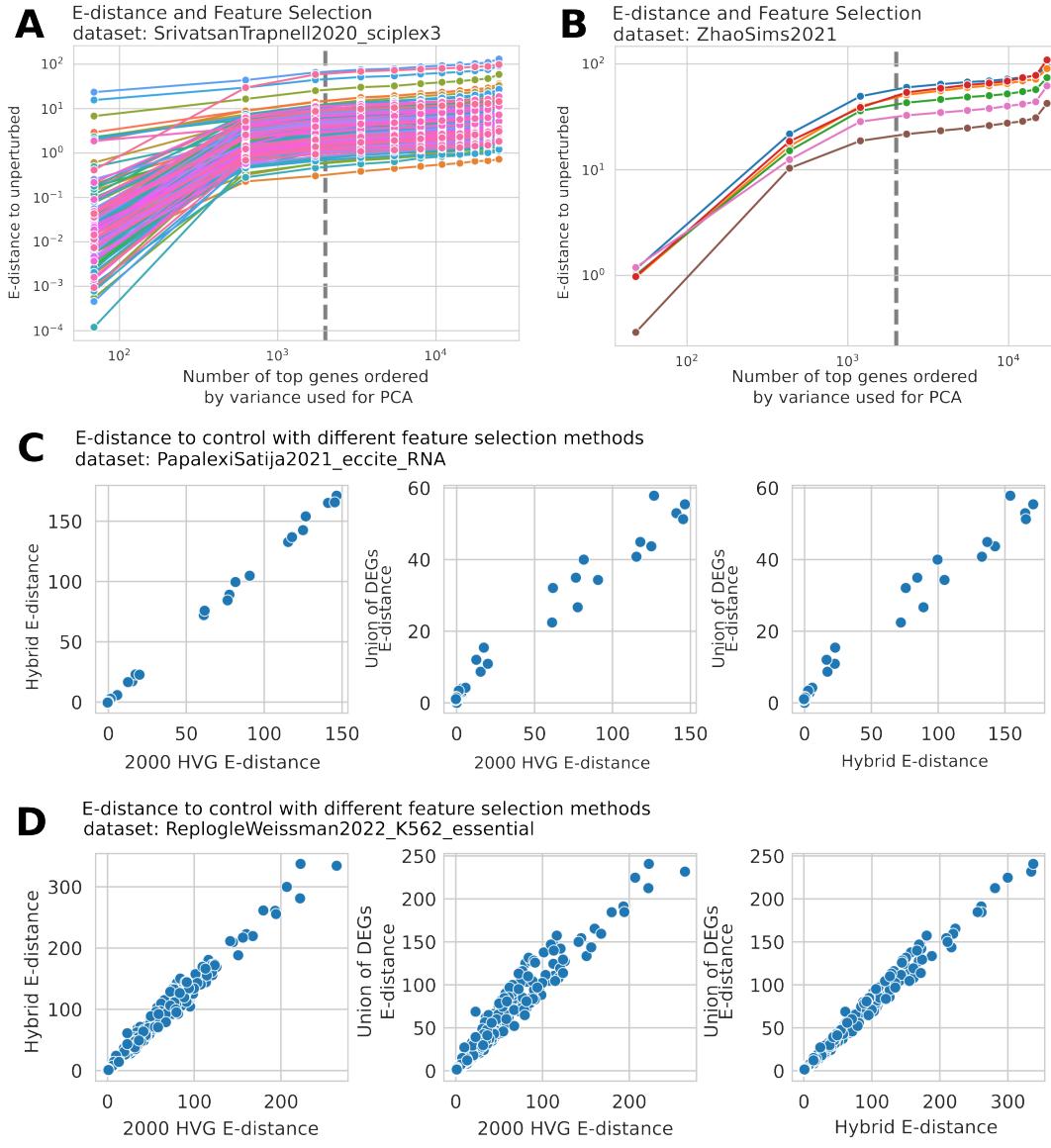


Figure B.2: (A),(B) E-distance is largely stable when at least 2000 (gray line) genes are used to compute PCA. The dotted line is at 2000 highly variable genes. (C),(D) Modification of feature selection to specifically use genes which are differentially expressed under perturbations minimally effects the E-distance. HVG: highly variable genes; DEG: differentially expressed genes; hybrid: the union of HVGs and DEGs.

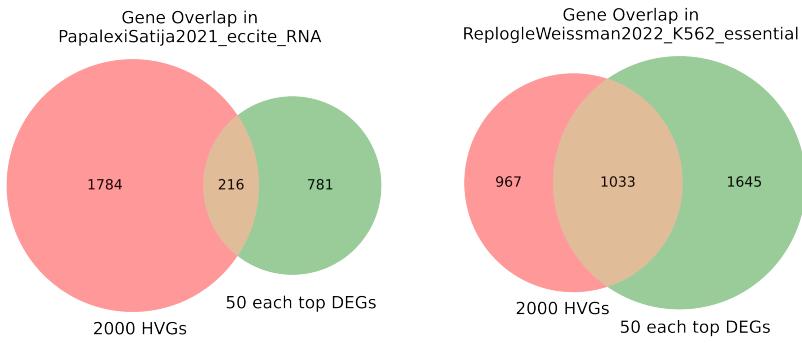


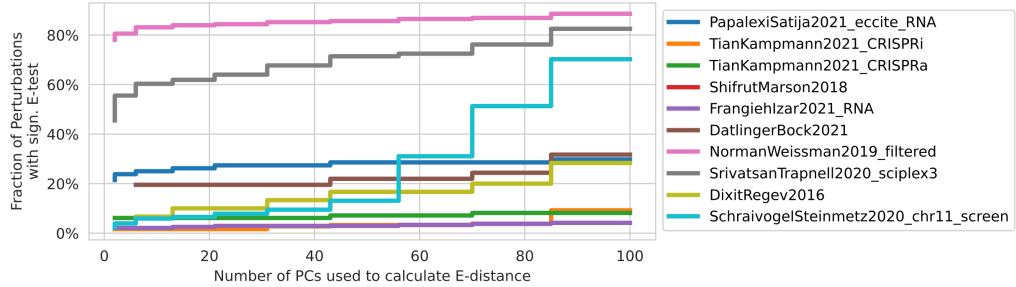
Figure B.3: Overlap of genes from different feature selection methods in the two datasets considered.

TAP-seq only measures approximately 3000 pre-defined genes of interest, and thus has far fewer starting features than other datasets. This leads to reduced correlation between genes in the resulting expression matrix, and thus fewer PCs are needed to sufficiently describe the data. Computing PCs separately for each perturbation rather than jointly across all perturbations in a given dataset similarly had minimal impact on the resulting E-distances (Figure B.4C). Taken together, this analysis indicates that E-statistics can be calculated as part of an existing computational workflow, which already includes calculating PCs across the full dataset.

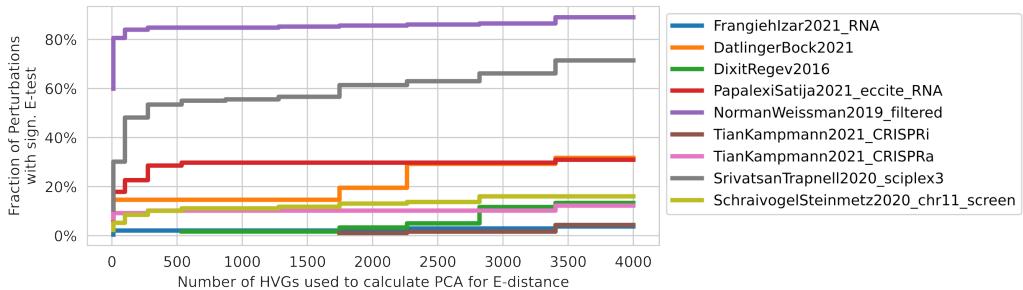
To examine the interplay between the possible confounding factors, we jointly varied the number of cells per perturbation, the average number of counts per cell, the number of HVGs used for PCA, and the number of PCs used to compute the E-distance for representative CRISPR and drug perturbation datasets (NormanWeissman2019_filtered¹⁴⁹; ZhaoSims2021²²³). We recorded the average E-distance to unperturbed, the average p-value in the E-test to unperturbed, and the number of perturbations with a significant (p-value < 0.05) E-test (Figure B.5).

Interestingly, when reducing the number of counts the E-test significance results in NormanWeissman2019_filtered did not converge for 200 cells; 500 cells were required to stabilize findings. In ZhaoSims2021, however, significance of the perturbation was stable even for low cell or count numbers. This supports our observations that drug perturbations induce stronger shifts in tran-

A Impact of reducing the number of PCs on significance



B Impact of reducing the number of HVGs on significance



C Comparison between joint PCA or separate PCA for E-distances

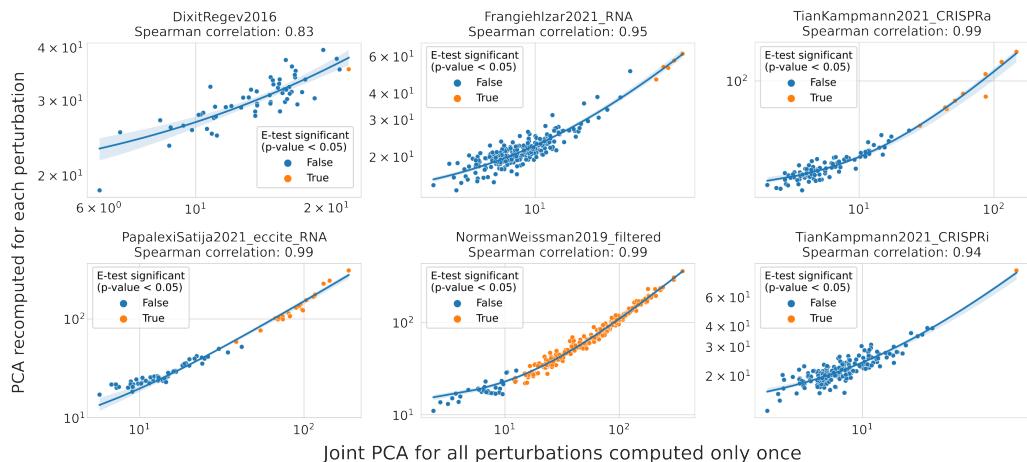


Figure B.4: Tests on robustness of E-statistics to dataset properties and parameters. (A) Changing the number of principal components from PCA has a small effect on the E-test for most datasets. (B) For most datasets, E-test results are stable between 500 and 4000 HVGs. (C) E-distance computed in a single, joint PCA is highly correlated with E-distance computed in a separate PCA per perturbed-unperturbed combination across three exemplary datasets. Consistently high Pearson correlations indicate strong equivalence between both approaches across datasets. Perturbations were subset to 200 cells prior to other calculations; perturbations with fewer than 200 cells were removed.

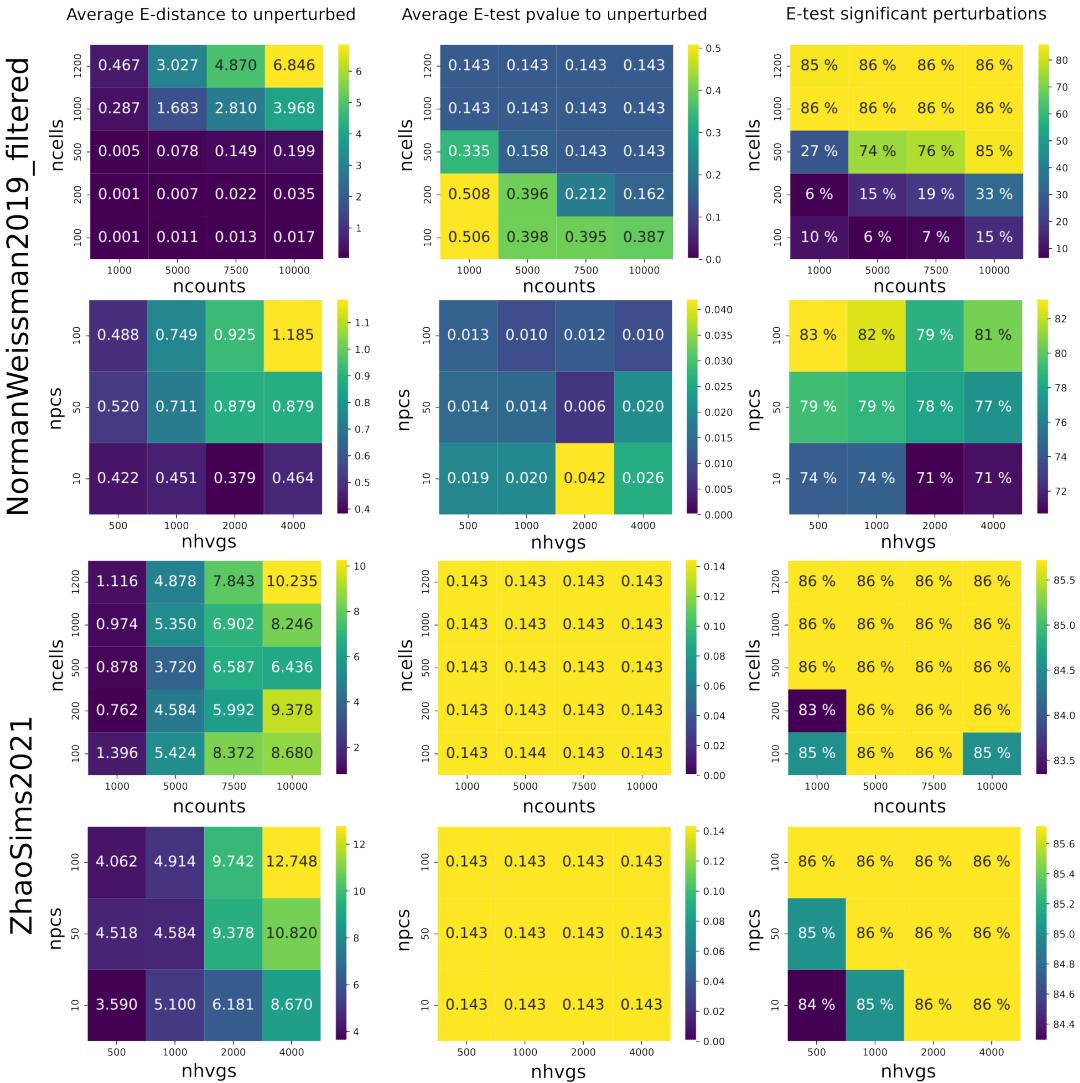


Figure B.5: Impact of jointly varying multiple parameters on empirical E-statistics results. For two datasets (top: `NormanWeissman2019_filtered`¹⁴⁹; bottom: `ZhaoSims2021`²²³) the following were jointly varied: cells per perturbation, average counts per cell, HVGs used for PCA, and PCs used to compute the E-distance.

scriptome profiles relative to CRISPR perturbations. Additionally, while the E-distance seems to quickly converge with number of cells, where convergence speed strongly depends on the dataset, higher numbers of counts increase the signal-to-noise ratio and therefore the E-distance would likely continue to increase at counts per cell beyond the maximum of 10000 counts per cell considered here.

We also investigated the impact of varying both the number of HVGs and the number of PCs, as each represents a type of feature selection or weighting method prior to E-distance calculation. In almost all cases 50 PCs and 2000 HVGs give stable E-test results, with minimally higher E-distances for higher numbers of PCs and HVGs.

Note that additional confounding factors such as the number of cells actually affected by a perturbation (which in the case of CRISPR-Cas9 is lower than 100%) and batch effects can also influence E-distance, and analysis should be applied with care in those settings. We recommend performing E-testing per batch and aggregating the results, or applying a batch correction method of choice prior to PCA. Future work on statistical methods should work towards accounting for batches.

C

Computational complexity of the E-test

We perform E-testing as a Monte Carlo permutation test (MCPT) using the E-distance based on pairwise distances in PCA space as test statistic between groups of cells. Therefore, the computational complexity of the E-test is composed of

$$\text{Complexity}_{\text{E-test}} \sim \text{Complexity}_{\text{PCA}} + \text{Complexity}_{\text{Pairwise distance}} + \text{Complexity}_{\text{MCPT}} \quad (\text{C.1})$$

We will examine each term separately. First, let us define some parameters (typical values in parentheses):

- n : Number of cells per perturbation, assuming same sizes for simplicity (100-2000)
- m : Number of control cells (usually a bit larger than n)
- k : Number of perturbations in the dataset (2-30000, usually depending on the perturbation method)
- s : Number of permutations for the MCPT (we recommend at least 1000-10000)
- d : Dimension of the data for PCA (2000 highly variable genes)
- p : Number of PCs used from PCA (50, usually between 10 and 100)

PCA essentially consist of two steps: first calculating the covariance matrix of the data $O(d^2(kn + m))$, then performing eigendecomposition of that matrix $O(d^3)$. This complexity can be reduced by using less features d or by using chunked PCA, as implemented by `scipy.pp.pca`²¹².

Pairwise distance computation requires to calculate distances only across each perturbation and the control cells in the lower-dimensional PCA space with p dimensions. Hence, it scales with $O(pk(nm)^2)$. Note that this scales linearly with the number of perturbations k , but quadratically with the number of cells per perturbation or control. We expect that future datasets will be focused on increasing k instead of n or m .

The computational complexity of the MCPT is given by $O(skp(n^2 + m^2 + nm))$, scaling linearly in the number of permutations s and the number of perturbations k respectively. Crucially, this holds because we compute the pairwise distances once at the beginning instead of recomputing them at every permutation. Thus, a permutation only requires summations over different sets of vectors. This reduces the complexity of one permutation to $O(p(n^2 + m^2 + nm))$, which is given

by the calculation of averages from pre-computed distances in the p -dimensional PCA space for the E-distance terms σ_n , σ_m , and δ_{nm} respectively. In addition, we parallelized the permutations of the MCPT across multiple threads, yielding a further increase in computation speed by a factor of 16 on our machines.

Taken together, the whole E-test procedure scales linearly with the number of permutations k , while scaling quadratically in the number of cells. We think that down-sampling cells is a feasible option—in case of too large datasets—as our investigations indicate that the E-distance converges quickly with the number of cells provided.

For the average dataset sub-sampled to 200 cells per perturbation, the E-test with 10,000 permutations runs within a few minutes on a laptop with 16 GB of memory. For the largest dataset in the scPerturb database, a genome-wide perturbation screen from ¹⁶⁶, we also sub-sampled to 200 cells for each of approximately 5000 perturbations. In this case, running the E-test with 10,000 permutations on 16 threads took 9 hours.

Bibliography

- [1] Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., & Weissman, J. S. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7), 1867–1882.e21.
- [2] Adlung, L. & Amit, I. (2018). From the Human Cell Atlas to dynamic immune maps in human disease. *Nature Reviews Immunology*, 18(10), 597–598.
- [3] Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., Gajda, A. M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., Feldman, L. E., Green, S. J., Lopez-Bigas, N., Frolov, M. V., & Benevolenskaya, E. V. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature Communications*, 12(1), 1628.
- [4] Amirifar, P., Yazdani, R., Azizi, G., Ranjouri, M. R., Durandy, A., Plebani, A., Lougaris, V., Hammarstrom, L., Aghamohammadi, A., & Abolhassani, H. (2021). Known and potential molecules associated with altered B cell development leading to predominantly antibody deficiencies. *Pediatric Allergy and Immunology*, 32(8), 1601–1615.
- [5] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106.
- [6] Armingol, E., Officer, A., Harismendy, O., & Lewis, N. E. (2021). Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2), 71–88.
- [7] Artis, D. & Spits, H. (2015). The biology of innate lymphoid cells. *Nature*, 517(7534), 293–301.
- [8] Atsaves, V., Leventaki, V., Rassidakis, G. Z., & Claret, F. X. (2019). AP-1 Transcription Factors as Regulators of Immune Responses in Cancer. *Cancers*, 11(7), 1037.

- [9] Baccin, C., Al-Sabah, J., Velten, L., Helbling, P. M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L. M., Trumpp, A., & Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature Cell Biology*, 22(1), 38–48.
- [10] Bairoch, A. (2018). The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of Biomolecular Techniques*, 29(2), 25–38.
- [11] Ballester Lopez, C., Conlon, T., Ertüz, Z., Eickelberg, O., & Yildirim, A. (2018). The novel notch ligand DNER modulates IFNG levels in recruited macrophages by enhancing non-canonical notch during COPD progression. In *Asthma and COPD: the best of respiratory structure and function*, American Thoracic Society International Conference Abstracts (pp. A7408–A7408). American Thoracic Society.
- [12] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippe, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1), D991–D995.
- [13] Bayik, D. & Lathia, J. D. (2021). Cancer stem cell-immune cell crosstalk in tumour progression. *Nature Reviews Cancer*, 21(8), 526–536.
- [14] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1), 38–44.
- [15] Bechtel, T. J., Reyes-Robles, T., Fadeyi, O. O., & Oslund, R. C. (2021). Strategies for monitoring cell–cell interactions. *Nature Chemical Biology*, 17(6), 641–652.
- [16] Bertin, P., Rector-Brooks, J., Sharma, D., Gaudelet, T., Anighoro, A., Gross, T., Martinez-Pena, F., Tang, E. L., S, S. M., Regep, C., Hayter, J., Korablyov, M., Valiante, N., van der Sloot, A., Tyers, M., Roberts, C., Bronstein, M. M., Lairson, L. L., Taylor-King, J. P., & Ben-gio, Y. (2022). RECOVER: sequential model optimization platform for combination drug repurposing identifies novel synergistic compounds in vitro.
- [17] Bhat, S., Rotti, H., Prasad, K., Kabekkodu, S. P., Saadi, A. V., Shenoy, S. P., Joshi, K. S., Nesari, T. M., Shengule, S. A., Dedge, A. P., Gadgil, M. S., Dhumal, V. R., Salvi, S., & Satyamoorthy, K. (2023). Genome-wide DNA methylation profiling after Ayurveda intervention to bronchial asthmatics identifies differential methylation in several transcription factors with immune process related function. *Journal of Ayurveda and Integrative Medicine*, 14(2), 100692.
- [18] Bhosale, P. B., Kim, H. H., Abusaliya, A., Vetrivel, P., Ha, S. E., Park, M. Y., Lee, H. J., & Kim, G. S. (2022). Structural and functional properties of activator protein-1 in cancer and

inflammation. *Evidence-based Complementary and Alternative Medicine : eCAM*, 2022, 9797929.

- [19] Bobolea, I., del Pozo, V., Sanz, V., Cabañas, R., Fiandor, A., Alfonso-Carrillo, C., Salcedo, M. A., Heredia Revuelto, R., & Quirce, S. (2018). Aspirin desensitization in aspirin-exacerbated respiratory disease: New insights into the molecular mechanisms. *Respiratory Medicine*, 143, 39–41.
- [20] Bonner, K., Kariyawasam, H. H., Ali, F. R., Clark, P., & Kay, A. B. (2010). Expression of functional receptor activity modifying protein 1 by airway epithelial cells with dysregulation in asthma. *Journal of Allergy and Clinical Immunology*, 126(6), 1277–1283.e3.
- [21] Bredikhin, D., Kats, I., & Stegle, O. (2022). MUON: multimodal omics analysis framework. *Genome Biology*, 23(1), 42.
- [22] Bridge, K. S. & Sharp, T. V. (2012). Regulators of the hypoxic response: a growing family. *Future Oncology*, 8(5), 491–493.
- [23] Broad Institute (2022). Single Cell Portal.
- [24] Browaeys, R., Saelens, W., & Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2), 159–162.
- [25] Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., & John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *GigaScience*, 11.
- [26] Buccitelli, C. & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), 630–644.
- [27] Buchheit, K. M., Dwyer, D. F., Ordovas-Montanes, J., Katz, H. R., Lewis, E., Vukovic, M., Lai, J., Bankova, L. G., Bhattacharyya, N., Shalek, A. K., Barrett, N. A., Boyce, J. A., & Laidlaw, T. M. (2020). IL-5R α marks nasal polyp IgG4- and IgE-expressing cells in aspirin-exacerbated respiratory disease. *Journal of Allergy and Clinical Immunology*, 145(6), 1574–1584.
- [28] Buchheit, K. M. & Hulse, K. E. (2021). Local immunoglobulin production in nasal tissues: A key to pathogenesis in chronic rhinosinusitis with nasal polyps and aspirin-exacerbated respiratory disease. *Annals of Allergy, Asthma & Immunology*, 126(2), 127–134.
- [29] Buchheit, K. M., Lewis, E., Gakpo, D., Hacker, J., Sohail, A., Taliaferro, F., Giron, E. B., Asare, C., Vukovic, M., Bensko, J. C., Dwyer, D. F., Shalek, A. K., Ordovas-Montanes, J., & Laidlaw, T. M. (2021). Mepolizumab targets multiple immune cells in aspirin-exacerbated respiratory disease. *Journal of Allergy and Clinical Immunology*, 148(2), 574–584.

- [30] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561), 486–490.
- [31] Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2), 155–160.
- [32] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β -VAE.
- [33] Burkhardt, D. B., Stanley, J. S., Tong, A., Perdigoto, A. L., Gigante, S. A., Herold, K. C., Wolf, G., Giraldez, A. J., van Dijk, D., & Krishnaswamy, S. (2021). Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39(5), 619–629.
- [34] Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., Mills, G. B., Lau, C. C., & Brown, P. H. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7), 1688–1698.
- [35] Butler, A., Hoffman, P., Smibert, P., Papalex, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5), 411–420.
- [36] Cahill, K. N., Bensko, J. C., Boyce, J. A., & Laidlaw, T. M. (2015). Prostaglandin D₂: A dominant mediator of aspirin-exacerbated respiratory disease. *Journal of Allergy and Clinical Immunology*, 135(1), 245–252.
- [37] Cahill, K. N. & Boyce, J. A. (2017). Aspirin-exacerbated respiratory disease: Mediators and mechanisms of a clinical disease. *Journal of Allergy and Clinical Immunology*, 139(3), 764–766.
- [38] Chan Zuckerberg Initiative (2022). CELLxGENE Discover.
- [39] Chang, M. T., Shanahan, F., Nguyen, T. T. T., Staben, S. T., Gazzard, L., Yamazoe, S., Wertz, I. E., Piskol, R., Yang, Y. A., Modrusan, Z., Haley, B., Evangelista, M., Malek, S., Foster, S. A., & Ye, X. (2022). Identifying transcriptional programs underlying cancer drug response with TraCe-seq. *Nature Biotechnology*, 40(1), 86–93.
- [40] Chari, T. & Pachter, L. (2022). The specious art of single-cell genomics.
- [41] Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, 20(1), 241.

- [42] Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., & Han, J.-D. J. (2023). Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1), 223.
- [43] Chen, W. S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., & Krishnaswamy, S. (2020). Uncovering axes of variation among single-cell cancer specimens. *Nature Methods*, 17(3), 302–310.
- [44] Comhair, S. A. A., Bochenek, G., Baicker-McKee, S., Wang, Z., Stachura, T., Sanak, M., Hammel, J. P., Hazen, S. L., Erzurum, S. C., & Nizankowska-Mogilnicka, E. (2018). The utility of biomarkers in diagnosis of aspirin exacerbated respiratory disease. *Respiratory Research*, 19(1), 210.
- [45] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., & Wang, B. (2023). scGPT: towards building a foundation model for single-cell multi-omics using generative AI.
- [46] Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237), 910–914.
- [47] Dagogo-Jack, I. & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2), 81–94.
- [48] Dahlin, A. & Weiss, S. T. (2016). Genetic and epigenetic components of aspirin-exacerbated respiratory disease. *Immunology and allergy clinics of North America*, 36(4), 765–789.
- [49] Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., & Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2), 245–253.
- [50] Date, K., Yamazaki, T., Toyoda, Y., Hoshi, K., & Ogawa, H. (2020). Alpha-Amylase expressed in human small intestinal epithelial cells is essential for cell proliferation and differentiation. *Journal of Cellular Biochemistry*, 121(2), 1238–1249.
- [51] Datlinger, P., Rendeiro, A. F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., & Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods*, 18(6), 635–642.
- [52] Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3), 297–301.
- [53] Deng, J., Thennavan, A., Shah, S., Bagdatlioglu, E., Klar, N., Heguy, A., Marier, C., Meyn, P., Zhang, Y., Labbe, K., Almonte, C., Krosgaard, M., Perou, C. M., Wong, K.-K., & Adams, S. (2021). Serial single-cell profiling analysis of metastatic TNBC during Nab-paclitaxel and pembrolizumab treatment. *Breast Cancer Research and Treatment*, 185(1), 85–94.

- [54] Dhapola, P., Rodhe, J., Olofzon, R., Bonald, T., Erlandsson, E., Soneji, S., & Karlsson, G. (2022). Scarf enables a highly memory-efficient analysis of large-scale single-cell genomics data. *Nature Communications*, 13(1), 4616.
- [55] Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C. E., Vasilevsky, N. A., Haendel, M. A., Blake, J. A., & Mungall, C. J. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1), 44.
- [56] Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., Ramirez Flores, R. O., Kim, H., Szalai, B., Costa, I. G., Valdeolivas, A., Dugourd, A., & Saez-Rodriguez, J. (2022). Comparison of methods and resources for cell-cell communication inference from single-cell RNA-seq data. *Nature Communications*, 13(1), 3224.
- [57] Dixit, A., Parnas, O., Li, B., & Chen, J. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7).
- [58] Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., Zhang, Z., Wang, P., Sun, S., & Liu, Q. (2019). Model-based understanding of single-cell CRISPR screening. *Nature Communications*, 10, 2233.
- [59] Dudakov, J. A., Hanash, A. M., & van den Brink, M. R. (2015). Interleukin-22: immunobiology and pathology. *Annual review of immunology*, 33, 747–785.
- [60] Edqvist, P.-H. D., Fagerberg, L., Hallström, B. M., Danielsson, A., Edlund, K., Uhlén, M., & Pontén, F. (2015). Expression of human skin-specific genes defined by transcriptomics and antibody-based profiling. *Journal of Histochemistry & Cytochemistry*, 63(2), 129–141.
- [61] Eferl, R. & Wagner, E. F. (2003). AP-1: a double-edged sword in tumorigenesis. *Nature Reviews Cancer*, 3(11), 859–868.
- [62] Efremova, M. & Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. *Nature Methods*, 17.
- [63] Feng, X., Zhang, M., Wang, B., Zhou, C., Mu, Y., Li, J., Liu, X., Wang, Y., Song, Z., & Liu, P. (2019). CRABP2 regulates invasion and metastasis of breast cancer through hippo pathway dependent on ER status. *Journal of Experimental & Clinical Cancer Research*, 38(1), 361.
- [64] Fischer, D. S., Dony, L., König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H., & Theis, F. J. (2021). Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology*, 22(1), 248.
- [65] Forcato, M., Romano, O., & Bicciato, S. (2021). Computational methods for the integrative analysis of single-cell data. *Briefings in Bioinformatics*, 22(3).

- [66] Forno, E., Zhang, R., Jiang, Y., Kim, S., Yan, Q., Ren, Z., Han, Y.-Y., Boutaoui, N., Rosser, F., Weeks, D. E., Acosta-Pérez, E., Colón-Semidey, A., Alvarez, M., Canino, G., Chen, W., & Celedón, J. C. (2020). Transcriptome-wide and differential expression network analyses of childhood asthma in nasal epithelium. *Journal of Allergy and Clinical Immunology*, 146(3), 671–675.
- [67] Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M. S., Zhao, M., Ager, C. R., Rogava, M., Hovey, L., Rotem, A., Bernatchez, C., Wucherpfennig, K. W., Johnson, B. E., Rozenblatt-Rosen, O., Schadendorf, D., Regev, A., & Izar, B. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3), 332–341.
- [68] Franz, A., Coscia, F., Shen, C., Charaoui, L., Mann, M., & Sander, C. (2021). Molecular response to PARP1 inhibition in ovarian cancer cells as determined by mass spectrometry based proteomics. *Journal of Ovarian Research*, 14(1), 140.
- [69] Gasperini, M., Hill, A. J., McFadine-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1), 377–390.e19.
- [70] Gatto, L., Aebersold, R., Cox, J., Demichev, V., Derk, J., Emmott, E., Franks, A. M., Ivanov, A. R., Kelly, R. T., Khouri, L., Leduc, A., MacCoss, M. J., Nemes, P., Perlman, D. H., Petelski, A. A., Rose, C. M., Schoof, E. M., Van Eyk, J., Vanderaa, C., Yates III, J. R., & Slavov, N. (2022). Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments.
- [71] Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., Svensson, V., da Veiga Beltrame, E., Kleshchevnikov, V., Talavera-López, C., Pachter, L., Theis, F. J., Streets, A., Jordan, M. I., Regier, J., & Yosef, N. (2022). A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2), 163–166.
- [72] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3), 272–282.
- [73] Gehring, J., Hwee Park, J., Chen, S., Thomson, M., & Pachter, L. (2020). Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nature Biotechnology*, 38(1), 35–38.

- [74] Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M., & Weissman, J. S. (2014). Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, 159(3), 647–661.
- [75] Goyal, Y., Busch, G. T., Pillai, M., Li, J., Boe, R. H., Grody, E. I., Chelvanambi, M., Dardani, I. P., Emert, B., Bodkin, N., Braun, J., Fingerman, D., Kaur, A., Jain, N., Ravindran, P. T., Mellis, I. A., Kiani, K., Alicea, G. M., Fane, M. E., Ahmed, S. S., Li, H., Chen, Y., Chai, C., Kaster, J., Witt, R. G., Lazcano, R., Ingram, D. R., Johnson, S. B., Wani, K., Dunagin, M. C., Lazar, A. J., Weeraratna, A. T., Wargo, J. A., Herlyn, M., & Raj, A. (2023). Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. *Nature*, (pp. 1–9).
- [76] Grandclaudon, M., Perrot-Dockès, M., Trichot, C., Karpf, L., Abouzid, O., Chauvin, C., Sirven, P., Abou-Jaoudé, W., Berger, F., Hupé, P., Thieffry, D., Sansonnet, L., Chiquet, J., Lévy-Leduc, C., & Soumelis, V. (2019). A quantitative multivariate model of human dendritic cell-T helper cell communication. *Cell*, 179(2), 432–447.e21.
- [77] Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3), 403–411.
- [78] Gronbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., & Winther, O. (2020). scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics*.
- [79] Gross, T. & Blüthgen, N. (2020). Identifiability and experimental design in perturbation studies. *Bioinformatics*, 36(Supplement 1), i482–i489.
- [80] Gross, T., Wongchenko, M. J., Yan, Y., & Blüthgen, N. (2019). Robust network inference using response logic. *Bioinformatics*, 35(14), i634–i642.
- [81] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573–3587.e29.
- [82] Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), 75.
- [83] He, J. & Wang, H. (2019). HspA1B is a prognostic biomarker and correlated with immune infiltrates in different subtypes of breast cancers. *bioRxiv*.

- [84] He, R., Wu, S., Gao, R., Chen, J., Peng, Q., Hu, H., Zhu, L., Du, Y., Sun, W., Ma, X., Zhang, H., Cui, Z., Wang, H., Martin, B. N., Wang, Y., Zhang, C.-j., & Wang, C. (2021). Identification of a long noncoding RNA TRAF3IP2-AS1 as key regulator of IL-17 signaling through the SRSF10–IRF1–Act1 axis in autoimmune diseases. *The Journal of Immunology*, 206(10), 2353–2365.
- [85] Hernandez-Pacheco, N., Vijverberg, S. J., Herrera-Luis, E., Li, J., Sio, Y. Y., Granell, R., Corrales, A., Maroteau, C., Lethem, R., Perez-Garcia, J., Farzan, N., Repnik, K., Gorenjak, M., Soares, P., Karimi, L., Schieck, M., Pérez-Méndez, L., Berce, V., Tavendale, R., Eng, C., Sardon, O., Kull, I., Mukhopadhyay, S., Pirmohamed, M., Verhamme, K. M., Burchard, E. G., Kabesch, M., Hawcutt, D. B., Melén, E., Potočnik, U., Chew, F. T., Tantisira, K. G., Turner, S., Palmer, C. N., Flores, C., Pino-Yanes, M., & Maitland-van Der Zee, A. H. (2021). Genome-wide association study of asthma exacerbations despite inhaled corticosteroid use. *European Respiratory Journal*, 57(5), 2003388.
- [86] Heumos, L., Zhang, X., Ji, Y., Peidli, S., & Green, T. (2023). pertpy (GitHub).
- [87] Hu, Y., Peng, T., Gao, L., & Tan, K. (2021). CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances*, 7(16).
- [88] Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8), 1–14.
- [89] Hysi, P., Kabesch, M., Moffatt, M. F., Schedel, M., Carr, D., Zhang, Y., Boardman, B., von Mutius, E., Weiland, S. K., Leupold, W., Fritzsch, C., Klopp, N., Musk, A. W., James, A., Nunez, G., Inohara, N., & Cookson, W. O. (2005). NOD1 variation, immunoglobulin E and asthma. *Human Molecular Genetics*, 14(7), 935–941.
- [90] Innes, B. T. & Bader, G. D. (2021). Transcriptional signatures of cell-cell interactions are dependent on cellular context. *bioRxiv*.
- [91] Jaakkola, M. K., Seyednasrollah, F., Mehmood, A., & Elo, L. L. (2017). Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics*, 18(5), 735–743.
- [92] Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., Oudenaarden, A. v., & Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167(7), 1883–1896.e15.
- [93] Jerby-Arnon, L. & Regev, A. (2022). DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nature Biotechnology*, 40(10), 1467–1477.
- [94] Ji, Y., Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2021). Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6), 522–537.

- [95] Jiang, Y.-Z., Ma, D., Suo, C., Shi, J., Xue, M., Hu, X., Xiao, Y., Yu, K.-D., Liu, Y.-R., Yu, Y., Zheng, Y., Li, X., Zhang, C., Hu, P., Zhang, J., Hua, Q., Zhang, J., Hou, W., Ren, L., Bao, D., Li, B., Yang, J., Yao, L., Zuo, W.-J., Zhao, S., Gong, Y., Ren, Y.-X., Zhao, Y.-X., Yang, Y.-S., Niu, Z., Cao, Z.-G., Stover, D. G., Verschraegen, C., Kaklamani, V., Daemen, A., Benson, J. R., Takabe, K., Bai, F., Li, D.-Q., Wang, P., Shi, L., Huang, W., & Shao, Z.-M. (2019). Genomic and transcriptomic landscape of triple-negative breast cancers: Subtypes and treatment strategies. *Cancer Cell*, 35(3), 428–440.e5.
- [96] Jin, K., Schnell, D., Li, G., Salomonis, N., Prasath, V. B. S., Szczesniak, R., & Aronow, B. J. (2022). CellDrift: Inferring perturbation responses in temporally-sampled single cell data.
- [97] Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3), 531–547.
- [98] Joshi, R. S., Kanugula, S. S., Sudhir, S., Pereira, M. P., Jain, S., & Aghi, M. K. (2021). The role of cancer-associated fibroblasts in tumor progression. *Cancers*, 13(6), 1399.
- [99] Jung, S. & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B).
- [100] Kabat, A. M., Harrison, O. J., Riffelmacher, T., Moghaddam, A. E., Pearson, C. F., Laing, A., Abeler-Dörner, L., Forman, S. P., Grencis, R. K., Sattentau, Q., Simon, A. K., Pott, J., & Maloy, K. J. (2016). The autophagy gene Atg16l1 differentially regulates Treg and TH₂ cells to control intestinal inflammation. *eLife*, 5, e12444.
- [101] Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., Edfors, F., Oksvold, P., von Feilitzen, K., Zwahlen, M., Arif, M., Altay, O., Li, X., Ozcan, M., Mardinoglu, A., Fagerberg, L., Mulder, J., Luo, Y., Ponten, F., Uhlén, M., & Lindskog, C. (2021). A single-cell type transcriptomics map of human tissues. *Science Advances*, 7(31).
- [102] Katzenelenbogen, Y., Sheban, F., Yalin, A., Yofe, I., Svetlichnyy, D., Jaitin, D. A., Bornstein, C., Moshe, A., Keren-Shaul, H., Cohen, M., Wang, S.-Y., Li, B., David, E., Salame, T.-M., Weiner, A., & Amit, I. (2020). Coupled scRNA-seq and intracellular protein activity reveal an immunosuppressive role of TREM₂ in cancer. *Cell*, 182(4), 872–885.e19.
- [103] Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7), 723–732.
- [104] Kim, S., Forno, E., Zhang, R., Park, H. J., Xu, Z., Yan, Q., Boutaoui, N., Acosta-Pérez, E., Canino, G., Chen, W., & Celedón, J. C. (2020). Expression quantitative trait methylation analysis reveals methylomic associations with gene expression in childhood asthma. *Chest*, 158(5), 1841–1856.

- [105] Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D., & Kirschner, M. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5), 1187–1201.
- [106] Kobak, D. & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1), 5416.
- [107] Koppensteiner, L., Mathieson, L., O'Connor, R. A., & Akram, A. R. (2022). Cancer associated fibroblasts: An impediment to effective anti-cancer T cell immunity. *Frontiers in Immunology*, 13, 887380.
- [108] Kubo, T., Yagi, T., & Kamiya, R. (2012). Tubulin polyglutamylation regulates flagellar motility by controlling a specific inner-arm dynein that interacts with the dynein regulatory complex. *Cytoskeleton*, 69(12), 1059–1068.
- [109] Kumar, V., Cheng, S.-C., Johnson, M. D., Smeekens, S. P., Wojtowicz, A., Giamarellos-Bourboulis, E., Karjalainen, J., Franke, L., Withoff, S., Plantinga, T. S., van de Veerdonk, F. L., van der Meer, J. W. M., Joosten, L. A. B., Sokol, H., Bauer, H., Herrmann, B. G., Bochud, P.-Y., Marchetti, O., Perfect, J. R., Xavier, R. J., Kullberg, B. J., Wijmenga, C., & Netea, M. G. (2014). Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. *Nature Communications*, 5(1), 4675.
- [110] Kumari, C., Gupta, R., Sharma, M., Jacob, J., Narayan, R. K., Sahni, D., & Kumar, A. (2023). Morpho-functional characterization of the submucosal glands at the nasopharyngeal end of the auditory tube in humans. *Journal of Anatomy*, 242(5), 771–780.
- [111] Laidlaw, T. M. (2018). Pathogenesis of NSAID-induced reactions in aspirin-exacerbated respiratory disease. *World Journal of Otorhinolaryngology - Head and Neck Surgery*, 4(3), 162–168.
- [112] Laidlaw, T. M. & Boyce, J. A. (2016). Aspirin-exacerbated respiratory disease—new prime suspects. *New England Journal of Medicine*, 374(5), 484–488.
- [113] Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., Liu, Q., Russikh, N., Ryazantsev, G., Ohler, U., Participants, N. . M. d. i. c., Pisco, A. O., Bloom, J., Krishnaswamy, S., & Theis, F. J. (2022). Multimodal single cell data integration challenge: results and lessons learned.
- [114] Larsson, I., Dalmo, E., Elgendy, R., Niklasson, M., Doroszko, M., Segerman, A., Jörnsten, R., Westermark, B., & Nelander, S. (2021). Modeling glioblastoma heterogeneity as a dynamic network of cell states. *Molecular Systems Biology*, 17(9), e10105.
- [115] Lee, H. & Han, B. (2022). FastRNA: An efficient solution for PCA of single-cell RNA-sequencing data based on a batch-accounting count model. *The American Journal of Human Genetics*, 109(11), 1974–1985.

- [116] Lee, R. U. & Stevenson, D. D. (2011). Aspirin-Exacerbated Respiratory Disease: Evaluation and Management. *Allergy, Asthma and Immunology Research*, 3(1), 3.
- [117] Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., Rosen, Y., Slyper, M., Kowalczyk, M. S., Villani, A.-C., Tickle, T., Hacohen, N., Rozenblatt-Rosen, O., & Regev, A. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nature Methods*, 17(8), 793–798.
- [118] Lin, J., Zhu, Z., Xiao, H., Wakefield, M. R., Ding, V. A., Bai, Q., & Fang, Y. (2017). The role of IL-7 in immunity and cancer. *Anticancer Research*, 37(3), 963–967.
- [119] Liscovitch-Brauer, N., Montalbano, A., Deng, J., Méndez-Mancilla, A., Wessels, H.-H., Moss, N. G., Kung, C.-Y., Sookdeo, A., Guo, X., Geller, E., Jaini, S., Smibert, P., & Sanjana, N. E. (2021). Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nature Biotechnology*, 39(10), 1270–1277.
- [120] Liu, C., Li, N., Liu, G., & Feng, X. (2021a). Annexin A3 and cancer. *Oncology Letters*, 22(6), 834.
- [121] Liu, C., Tu, C., Wang, L., Wu, H., Houston, B. J., Mastrorosa, F. K., Zhang, W., Shen, Y., Wang, J., Tian, S., Meng, L., Cong, J., Yang, S., Jiang, Y., Tang, S., Zeng, Y., Lv, M., Lin, G., Li, J., Saiyin, H., He, X., Jin, L., Touré, A., Ray, P. F., Veltman, J. A., Shi, Q., O'Bryan, M. K., Cao, Y., Tan, Y.-Q., & Zhang, F. (2021b). Deleterious variants in X-linked CFAP47 induce asthenoteratozoospermia and primary male infertility. *The American Journal of Human Genetics*, 108(2), 309–323.
- [122] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053–1058.
- [123] Lorenzo, G., Jarrett, A. M., Meyer, C. T., Quaranta, V., Tyson, D. R., & Yankeelov, T. E. (2022). Identifying mechanisms driving the early response of triple negative breast cancer patients to neoadjuvant chemotherapy using a mechanistic model integrating in vitro and in vivo imaging data.
- [124] Lotfollahi, M., Dony, L., Agarwala, H., & Theis, F. (2021). Out-of-distribution prediction with disentangled representations for single-cell rna sequencing data. *bioRxiv*.
- [125] Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., Shendure, J., McFaline-Figueroa, J. L., Boyeau, P., Wolf, F. A., Yakubova, N., Günemann, S., Trapnell, C., Lopez-Paz, D., & Theis, F. J. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), e11517.

- [126] Lotfollahi, M., Naghipourfar, M., Theis, F. J., & Wolf, F. A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. *Bioinformatics*, 36(Supplement_2), i610–i617.
- [127] Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8), 715–721.
- [128] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
- [129] Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), 41–50.
- [130] Luecken, M. D. & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), e8746.
- [131] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerewinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., & Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1), 31.
- [132] Ma, A., McDermaid, A., Xu, J., Chang, Y., & Ma, Q. (2020). Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*, 38(9), 1007–1022.
- [133] Maaten, L. v. d. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.
- [134] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202–1214.
- [135] Maman, S. & Witz, I. P. (2018). A history of exploring cancer in context. *Nature Reviews Cancer*, 18(6), 359–376.
- [136] Masternak, M. M. & Bartke, A. (2012). Growth hormone, inflammation and aging. *Pathobiology of Aging & Age-related Diseases*, 2(1), 17293.

- [137] McFarland, J. M., Paolella, B. R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W. N., Jones, A., Chambers, E., Dionne, D., Bender, S., Wolpin, B. M., Ghandi, M., Tirosh, I., Rozenblatt-Rosen, O., Roth, J. A., Golub, T. R., Regev, A., Aguirre, A. J., Vazquez, F., & Tsherniak, A. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11(1), 4296.
- [138] McGeachie, M. J., Sordillo, J. E., Dahlin, A., Wang, A. L., Lutz, S. M., Tantisira, K. G., Pan-ganiban, R., Lu, Q., Sajuthi, S., Urbanek, C., Kelly, R., Saef, B., Eng, C., Oh, S. S., Kho, A. T., Croteau-Chonka, D. C., Weiss, S. T., Raby, B. A., Mak, A. C. Y., Rodriguez-Santana, J. R., Burchard, E. G., Seibold, M. A., & Wu, A. C. (2020). Expression of SMARCD1 interacts with age in association with asthma control on inhaled corticosteroid therapy. *Respiratory Research*, 21(1), 31.
- [139] McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.
- [140] Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalex, E., Ouyang, Z., Satija, R., Sanjana, N. E., Koralov, S. B., & Smibert, P. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5), 409–412.
- [141] Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B. Z., Papalex, E., Thakore, P. I., Kibayashi, T., Wing, J. B., Hata, M., Satija, R., Nazor, K. L., Sakaguchi, S., Ludwig, L. S., Sankaran, V. G., Regev, A., & Smibert, P. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*, 39(10), 1246–1258.
- [142] Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y. Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Magalhães, J. P. d., & Larbi, A. (2019). RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*, 26(6), 1627–1640.e7.
- [143] Mongia, A., Saunders, D. C., Wang, Y. J., Brissova, M., Powers, A. C., Kaestner, K. H., Vahedi, G., Naji, A., Schwartz, G. W., & Faryabi, R. B. (2023). AnnoSpat annotates cell types and quantifies cellular arrangements from spatial proteomics.
- [144] Morita, H., Kubo, T., Rückert, B., Ravindran, A., Soyka, M. B., Rinaldi, A. O., Sugita, K., Wawrzyniak, M., Wawrzyniak, P., Motomura, K., Tamari, M., Orimo, K., Okada, N., Arae, K., Saito, K., Altunbulakli, C., Castro-Giner, F., Tan, G., Neumann, A., Sudo, K., O’Mahony, L., Honda, K., Nakae, S., Saito, H., Mjösberg, J., Nilsson, G., Matsumoto, K., Akdis, M., & Akdis, C. A. (2019). Induction of human regulatory innate lymphoid cells from group 2 innate lymphoid cells by retinoic acid. *Journal of Allergy and Clinical Immunology*, 143(6), 2190–2201.e9.

- [145] Moses, L., Einarsson, P. H., Jackson, K., Luebbert, L., Booeshaghi, A. S., Antonsson, S., Bray, N., Melsted, P., & Pachter, L. (2023). Voyager: exploratory single-cell genomics data analysis with geospatial statistics. *bioRxiv*.
- [146] Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). *Sustainable data analysis with Snakemake*. Technical Report 10:33, F1000Research.
- [147] Navarro, R., Compte, M., Álvarez Vallina, L., & Sanz, L. (2016). Immune regulation by pericytes: modulating innate and adaptive immunity. *Frontiers in Immunology*, 7, 480.
- [148] Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7), 773–782.
- [149] Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., & Weissman, J. S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455), 786–793.
- [150] Ordovas-Montanes, J., Dwyer, D. F., Nyquist, S. K., Buchheit, K. M., Vukovic, M., Deb, C., Wadsworth, M. H., Hughes, T. K., Kazer, S. W., Yoshimoto, E., Cahill, K. N., Bhattacharyya, N., Katz, H. R., Berger, B., Laidlaw, T. M., Boyce, J. A., Barrett, N. A., & Shalek, A. K. (2018). Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature*, 560(7720), 649–654.
- [151] Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., Bryant, V. L., Penington, J. S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G. B., kConFab, Papenfuss, A. T., Lindeman, G. J., Smyth, G. K., & Visvader, J. E. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO Journal*, 40(11), e107333.
- [152] Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., Lotfollahi, M., Richter, S., & Theis, F. J. (2022). Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2), 171–178.
- [153] Papalexis, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., & Satija, R. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53(3), 322–331.
- [154] Park, S.-M., Park, J. S., Park, H.-S., & Park, C.-S. (2013). Unraveling the genetic basis of aspirin hypersensitivity in aspirin beyond arachidonate pathways. *Allergy, Asthma & Immunology Research*, 5(5), 258.

- [155] Pavlidis, P., Tsakmaki, A., Pantazi, E., Li, K., Cozzetto, D., Digby-Bell, J., Yang, F., Lo, J. W., Alberts, E., Sa, A. C. C., Niazi, U., Friedman, J., Long, A. K., Ding, Y., Carey, C. D., Lamb, C., Saqi, M., Madgwick, M., Gul, L., Treveil, A., Korcsmaros, T., Macdonald, T. T., Lord, G. M., Bewick, G., & Powell, N. (2022). Interleukin-22 regulates neutrophil recruitment in ulcerative colitis and is associated with resistance to ustekinumab therapy. *Nature Communications*, 13(1), 5820.
- [156] Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., Mc-Clanahan, T. K., Sadekova, S., & Klappenebach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10), 936–939.
- [157] Pierce, S. E., Granja, J. M., & Greenleaf, W. J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications*, 12(1), 2969.
- [158] Pierson, E. & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1), 241.
- [159] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2), 147–154.
- [160] Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34(9), 1538–1546.
- [161] Przybyla, L. & Gilbert, L. A. (2021). A new era in functional genomics screens. *Nature Reviews Genetics*.
- [162] Qi, C., Vonk, J. M., van der Plaat, D. A., Nieuwenhuis, M. A. E., Dijk, F. N., Aïssi, D., Siroux, V., Boezen, H. M., Xu, C.-j., Koppelman, G. H., & BIOS Consortium (2020). Epigenome-wide association study identifies DNA methylation markers for asthma remission in whole blood and nasal epithelium. *Clinical and Translational Allergy*, 10(1), 60.
- [163] Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., Wauters, E., Pomella, V., Verbandt, S., Busschaert, P., Bassez, A., Franken, A., Bempt, M. V., Xiong, J., Weynand, B., Van Herck, Y., Antoranz, A., Bosisio, F. M., Thienpont, B., Floris, G., Vergote, I., Smeets, A., Tejpar, S., & Lambrechts, D. (2020). A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research*, 30(9), 745–762.
- [164] Qiao, W., Wang, W., Laurenti, E., Turinsky, A. L., Wodak, S. J., Bader, G. D., Dick, J. E., & Zandstra, P. W. (2014). Intercellular network structure and regulatory motifs in the human hematopoietic system. *Molecular Systems Biology*, 10(7), 741.

- [165] Ren, L., Li, J., Wang, C., Lou, Z., Gao, S., Zhao, L., Wang, S., Chaulagain, A., Zhang, M., Li, X., & Tang, J. (2021). Single cell RNA sequencing for breast cancer: present and future. *Cell Death Discovery*, 7(1), 104.
- [166] Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mas-cibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., & Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14), 2559–2575.e28.
- [167] Rieckmann, J. C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S. S., Lanzavecchia, A., Mann, M., & Meissner, F. (2017). Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature Immunology*, 18(5), 583–593.
- [168] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., & Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1), 284.
- [169] Rizzo, M. L. & Székely, G. J. (2016). Energy distance. *WIREs Computational Statistics*, 8(1), 27–38.
- [170] Roy, D. G., Kaymak, I., Williams, K. S., Ma, E. H., & Jones, R. G. (2021). Immunometabolism in the tumor microenvironment. *Annual Review of Cancer Biology*, 5(1), 137–159.
- [171] Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., Mumbach, M. R., Ji, A. L., Kim, D. S., Cho, S. W., Zarnegar, B. J., Greenleaf, W. J., Chang, H. Y., & Khavari, P. A. (2019). Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, 176(1), 361–376.e17.
- [172] Sarin, K. Y., Lin, Y., Daneshjou, R., Ziyatdinov, A., Thorleifsson, G., Rubin, A., Pardo, L. M., Wu, W., Khavari, P. A., Uitterlinden, A., Nijsten, T., Toland, A. E., Olafsson, J. H., Sigurgeirsson, B., Thorisdottir, K., Jorgensen, E., Whittemore, A. S., Kraft, P., Stacey, S. N., Stefansson, K., Asgari, M. M., & Han, J. (2020). Genome-wide meta-analysis identifies eight new susceptibility loci for cutaneous squamous cell carcinoma. *Nature Communications*, 11(1), 820.
- [173] Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A. L. S., Zheng, G. X. Y., Greenleaf, W. J., & Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8), 925–936.

- [174] Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, 14(10), 975–978.
- [175] Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., & Lander, E. S. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4), 928–943.e22.
- [176] Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., & Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, 17(6), 629–635.
- [177] Schroder, K., Hertzog, P. J., Ravasi, T., & Hume, D. A. (2004). Interferon- γ : an overview of signals, mechanisms and functions. *Journal of Leukocyte Biology*, 75(2), 163–189.
- [178] Seshadri, S., Lu, X., Purkey, M. R., Homma, T., Choi, A. W., Carter, R., Suh, L., Norton, J., Harris, K. E., Conley, D. B., Kato, A., Avila, P. C., Czarnocka, B., Kopp, P. A., Peters, A. T., Grammer, L. C., Chandra, R. K., Tan, B. K., Liu, Z., Kern, R. C., & Schleimer, R. P. (2015). Increased expression of the epithelial anion transporter pendrin/SLC26A4 in nasal polyps of patients with chronic rhinosinusitis. *Journal of Allergy and Clinical Immunology*, 136(6), 1548–1558.e7.
- [179] Shifrut, E., Carnevale, J., Tobin, V., Roth, T. L., Woo, J. M., Bui, C. T., Li, P. J., Diolaiti, M. E., Ashworth, A., & Marson, A. (2018). Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell*, 175(7), 1958–1971.e15.
- [180] Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinner, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 5692.
- [181] Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., & Trapnell, C. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473), 45–51.
- [182] Stacey, S. N., Helgason, H., Gudjonsson, S. A., Thorleifsson, G., Zink, F., Sigurdsson, A., Kehr, B., Gudmundsson, J., Sulem, P., Sigurgeirsson, B., Benediktsdottir, K. R., Thorisdottir, K., Ragnarsson, R., Fuentelsaz, V., Corredera, C., Gilaberte, Y., Grasa, M., Planelles, D., Sanmartin, O., Rudnai, P., Gurzau, E., Koppova, K., Nexø, B. A., Tjønneland, A., Overvad, K., Jonasson, J. G., Tryggvadottir, L., Johannsdottir, H., Kristinsdottir, A. M., Stefansson,

- H., Masson, G., Magnusson, O. T., Halldorsson, B. V., Kong, A., Rafnar, T., Thorsteinsdotir, U., Vogel, U., Kumar, R., Nagore, E., Mayordomo, J. I., Gudbjartsson, D. F., Olafsson, J. H., & Stefansson, K. (2015). New basal cell carcinoma susceptibility loci. *Nature Communications*, 6(1), 6825.
- [183] Stathias, V., Turner, J., Koleti, A., Vidovic, D., Cooper, D., Fazel-Najafabadi, M., Pilarczyk, M., Terryn, R., Chung, C., Umeano, A., Clarke, D. J. B., Lachmann, A., Evangelista, J. E., Ma'ayan, A., Medvedovic, M., & Schürer, S. C. (2020). LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Research*, 48(D1), D431–D439.
- [184] Steinke, J. W., Negri, J., Liu, L., Payne, S. C., & Borish, L. (2014). Aspirin activation of eosinophils and mast cells: implications in the pathogenesis of aspirin-exacerbated respiratory disease. *The Journal of Immunology*, 193(1), 41–47.
- [185] Stevenson, D. D., Hankammer, M. A., Mathison, D. A., Christiansen, S. C., & Simon, R. A. (1996). Aspirin desensitization treatment of aspirin-sensitive patients with rhinosinusitis-asthma: long-term outcomes. *The Journal of Allergy and Clinical Immunology*, 98(4), 751–758.
- [186] Stevenson, D. D., Simon, R. A., & Mathison, D. A. (1980). Aspirin-sensitive asthma: tolerance to aspirin after positive oral aspirin challenges. *The Journal of Allergy and Clinical Immunology*, 66(1), 82–88.
- [187] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9), 865–868.
- [188] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), 1888–1902.e21.
- [189] Su, Y., Xu, C., Sun, Z., Liang, Y., Li, G., Tong, T., & Chen, J. (2019). S100A13 promotes senescence-associated secretory phenotype and cellular senescence via modulation of non-classical secretion of IL-1 α . *Aging (Albany NY)*, 11(2), 549–572.
- [190] Svensson, V., da Veiga Beltrame, E., & Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. *Database*, 2020.
- [191] Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., & Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4), 381–387.
- [192] Szczeklik, A., Niżankowska, E., Dupлага, M., & the Aiane Investigators, o. b. o. (2000). Natural history of aspirin-induced asthma. *European Respiratory Journal*, 16(3), 432.

- [193] Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8), 1249–1272.
- [194] Tachibana, M., Morioka, H., Tanimura, F., MacHino, M., & Mizukoshi, O. (1986). Amylase secretion by nasal glands: An immunocytochemical study. *Annals of Otology, Rhinology & Laryngology*, 95(3), 284–287.
- [195] Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616–624.
- [196] Tian, R., Abarrientos, A., Hong, J., Hashemi, S. H., Yan, R., Dräger, N., Leng, K., Nalls, M. A., Singleton, A. B., Xu, K., Faghri, F., & Kampmann, M. (2021). Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24(7), 1020–1034.
- [197] Tian, R., Gachechiladze, M. A., Ludwig, C. H., Laurie, M. T., Hong, J. Y., Nathaniel, D., Prabhu, A. V., Fernandopulle, M. S., Patel, R., Abshari, M., Ward, M. E., & Kampmann, M. (2019). CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. *Neuron*, 104(2), 239–255.e12.
- [198] Toney, N. J., Opdenaker, L. M., Cicek, K., Frerichs, L., Kennington, C. R., Oberly, S., Archinal, H., Somasundaram, R., & Sims-Mourtada, J. (2022). Tumor-B-cell interactions promote isotype switching to an immunosuppressive IgG4 antibody response through upregulation of IL-10 in triple negative breast cancers. *Journal of Translational Medicine*, 20(1), 112.
- [199] Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, 25(10), 1491–1498.
- [200] Tsai, M.-H. & Lee, C.-K. (2018). STAT3 cooperates with phospholipid scramblase 2 to suppress type I interferon response. *Frontiers in Immunology*, 9, 1886.
- [201] Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., Ghandi, M., Garraway, L. A., Root, D. E., Golub, T. R., Boehm, J. S., & Hahn, W. C. (2017). Defining a cancer dependency map. *Cell*, 170(3), 564–576.e16.
- [202] Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J. T., Haniffa, M., Moffett, A., & Teichmann, S. A. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731), 347–353.

- [203] Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, 583(7818), 729–736.
- [204] Vieth, B., Ziegenhain, C., Parekh, S., Enard, W., & Hellmann, I. (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21), 3486–3488.
- [205] Vivier, E., Artis, D., Colonna, M., Diefenbach, A., Santo, J. P. D., Eberl, G., Koyasu, S., Locksley, R. M., McKenzie, A. N. J., Mebius, R. E., Powrie, F., & Spits, H. (2018). Innate lymphoid cells: 10 years on. *Cell*, 174(5), 1054–1066.
- [206] Wang, Y., Sun, X., & Zhao, H. (2022). Benchmarking automated cell type annotation tools for single-cell ATAC-seq data. *Frontiers in Genetics*, 13.
- [207] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479).
- [208] Wessels, H.-H., Méndez-Mancilla, A., Papalexopoulou, A., Mauck, W. M., Lu, L., Morris, J. A., Mimou, E., Smibert, P., Sanjana, N. E., & Satija, R. (2022). Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq.
- [209] White, A. A. & Stevenson, D. D. (2018). Aspirin-Exacerbated Respiratory Disease. *New England Journal of Medicine*, 379(11), 1060–1070.
- [210] Wilk, A. J., Shalek, A. K., Holmes, S., & Blish, C. A. (2023). Comparative analysis of cell–cell communication at single-cell resolution. *Nature Biotechnology*, (pp. 1–14).
- [211] Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515–534.
- [212] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15.
- [213] Wu, K. E., Yost, K. E., Chang, H. Y., & Zou, J. (2021). BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15), e2023070118.
- [214] Xie, S., Duan, J., Li, B., Zhou, P., & Hon, G. C. (2017). Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Molecular Cell*, 66(2), 285–299.e5.
- [215] Yamaguchi, N., Suzuki, A., Yoshida, A., Tanaka, T., Aoyama, K., Oishi, H., Hara, Y., Ogi, T., Amano, I., Kameo, S., Koibuchi, N., Shibata, Y., Ugawa, S., Mizuno, H., & Saitoh, S. (2022). The iodide transporter Slc26a7 impacts thyroid function more strongly than Slc26a4 in mice. *Scientific Reports*, 12(1), 11259.

- [216] Yin, L., Duan, J.-J., Bian, X.-W., & Yu, S.-c. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research*, 22(1), 61.
- [217] Young, M. D. & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12), giaa151.
- [218] Yuan, X., Wang, J., Huang, Y., Shangguan, D., & Zhang, P. (2021). Single-cell profiling to explore immunological heterogeneity of tumor microenvironment in breast cancer. *Frontiers in Immunology*, 12, 643692.
- [219] Zhang, X., Biagini Myers, J. M., Burleson, J., Ulm, A., Bryan, K. S., Chen, X., Weirauch, M. T., Baker, T. A., Butsch Kovacic, M. S., & Ji, H. (2018). Nasal DNA methylation is associated with childhood asthma. *Epigenomics*, 10(5), 629–641.
- [220] Zhang, Y., Chen, H., Mo, H., Hu, X., Gao, R., Zhao, Y., Liu, B., Niu, L., Sun, X., Yu, X., Wang, Y., Chang, Q., Gong, T., Guan, X., Hu, T., Qian, T., Xu, B., Ma, F., Zhang, Z., & Liu, Z. (2021). Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. *Cancer Cell*, 39(12), 1578–1593.e8.
- [221] Zhang, Y.-N., Song, J., Zhai, G.-T., Wang, H., Luo, R.-Z., Li, J.-X., Liao, B., Ma, J., Wang, H., Lu, X., Liu, D.-B., & Liu, Z. (2020). Evidence for the presence of long-lived plasma cells in nasal polyps. *Allergy, Asthma & Immunology Research*, 12(2), 274–291.
- [222] Zhao, C., Wang, W., Yao, H., & Wang, X. (2018). SOCS3 is upregulated and targeted by miR30a-5p in allergic rhinitis. *International Archives of Allergy and Immunology*, 175(4), 209–219.
- [223] Zhao, W., Dovas, A., Spinazzi, E. F., Levitin, H. M., Banu, M. A., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M. L., Sisti, M. B., Bruce, J. N., Canoll, P., & Sims, P. A. (2021). Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Medicine*, 13(1), 82.
- [224] Zhou, B., Lawrence, T., & Liang, Y. (2021). The role of plasmacytoid dendritic cells in cancers. *Frontiers in Immunology*, 12.
- [225] Zhou, Z., Ye, C., Wang, J., & Zhang, N. R. (2020). Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11(1), 651.