Title
Quantifying Reproducibility and Bias in Single-Cell Sequencing Analyses of Human Immunity

Author
Cole, Michael Blake

Publication Date
2018

Peer reviewed|Thesis/dissertation

Quantifying Reproducibility and Bias in Single-Cell Sequencing Analyses of Human Immunity

by

Michael Blake Cole

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy
in
Physics
and the Designated Emphasis
in
Computational and Genomic Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Nir Yosef, Co-Chair
Professor Michael DeWeese, Co-Chair
Professor Lisa Barcellos
Professor Hernan Garcia-Melan

Summer 2018

ABSTRACT

Quantifying Reproducibility and Bias in Single-Cell Sequencing Analyses of Human Immunity

by

Michael Blake Cole

Doctor of Philosophy in Physics
Designated Emphasis in Computational and Genomic Biology
University of California, Berkeley
Professor Nir Yosef, Co-Chair
Professor Michael DeWeese, Co-Chair

Single-cell RNA sequencing technologies have evolved rapidly over the past few years, but the newest protocols and platforms are still limited by bias and noise. These technical issues can present serious challenges to downstream data analyses when samples are collected from multiple human donors recruited from multiple sites. My dissertation outlines methods for assessing bias and quantifying reproducibility in single-cell RNA sequencing studies. These tools are applicable to a new class of single-cell studies of human disease that move beyond tissue-level case–control comparisons. I have implemented these methods in two software packages, scone and scRAD, both developed to improve the quality of biological insights derived from single-cell RNA sequencing data.

To Mom and Dad

ACKNOWLEDGMENTS

Thank you, Nir, for your strong commitment to advising - you have been my advocate every step of the way. Thanks to Lisa for your contagious enthusiasm, your encouragement and guidance. Thanks to Alex for your tireless support, your perspectives and advice. Thank you to my co-authors – Davide, Kellie, and Enrique – for your expertise, your perseverance, and your professionalism. Thanks to everyone in the Yosef Lab – working with you has been a privilege. Thanks to the National Institute of Dental and Craniofacial Research, the National Institutes of Health, and the taxpayers who fund them.

Thanks to my friends from Stanford for your warmth from near and afar. Thanks to my friends from Berkeley for your laughter and for challenging me to grow.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Ab	antibody
ACR	American College of Rheumatology
`AME`	Analysis of Motif Enrichment
ANOVA	analysis of variance
ARI	adjusted Rand index
ASW	average silhouette width
AUC	area under the curve
auto-Ab	autoantibody
BH	Benjamini-Hochberg
BMDC	bone marrow-derived dendritic cell
bp	base pair
BY	Benjamini-Yekutieli
cART	combination antiretroviral therapy
CBMC	cord blood mononuclear cells
CD	cluster of differentiation
CDF	cumulative distribution function
cDNA	complementary DNA
CFSE	carboxyfluorescein succinimidyl ester
CIS	clinically isolated syndrome
CITE-seq	cellular indexing of transcriptome and epitopes by sequencing

CSEA	cell set enrichment analysis
CP	chronic progressor
CpG	5'—C—phosphate—G—3'
CpH	5'—C—phosphate—H—3'
CPU	central processing unit
CTL	cytotoxic T cell
CSF	cerebrospinal fluid
DC	dendritic cell
DE	differential expression
DMP	differentially methylated position
DNA	deoxyribonucleic acid
ds	double-stranded
EC	elite controller
ES	enrichment score
EDA	exploratory data analysis
EM	expectation–maximization
ERCC	External RNA Controls Consortium
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FET	Fisher's exact test
FLS	focal lymphocytic sialadenitis
FNR	false negative rate
FQ	full quantile
FWER	family-wise error rate

GEO	Gene Expression Omnibus
GFP	green fluorescent protein
GLM	generalized linear model
GO	Gene Ontology
GSEA	gene set enrichment analysis
GWAS	genome-wide association study
HD	healthy donor
HIV-1	human immunodeficiency virus type 1
HLA	human leukocyte antigen
IDR	irreproducible discovery rate
IFN	interferon
IIH	idiopathic intracranial hypertension
IL	interleukin
`IPA`	Ingenuity pathway analysis
iPSC	Induced Pluripotent Stem Cells
ISG	interferon-stimulated gene
IQR	interquartile range
KCS	keratoconjunctivitis sicca
KS	Kolmogorov–Smirnov
KW	Kruskal–Wallis
k-NN	k-nearest neighbors
LCMV	lymphocytic choriomeningitis virus
lfc	log fold change
LPS	lipopolysaccharide

LSG	labial salivary gland
LTR	long terminal repeat
MAD	median absolute deviation
mDC	myeloid dendritic cell
MHC	major histocompatibility complex
miRNA	microRNA
MLE	maximum likelihood estimation
MOI	multiplicity of infection
mRNA	messenger RNA
MS	multiple sclerosis
MSigDB	Molecular Signatures Database
MWW	Mann–Whitney–Wilcoxon
NK	natural killer
NPC	neural progenitor cell
OSS	ocular staining score
PAM	Partitioning Around Medoids
Pam	Pam3CSK4
PAMP	pathogen associated molecular pattern
PBMC	peripheral blood mononuclear cell
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
poly(A)	polyadenylation
PRR	pattern recognition receptor

QC	quality control
qPCR	quantitative polymerase chain reaction
Q–Q	quantile-quantile
RA	rheumatoid arthritis
RLE	relative log expression
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC	receiver operating characteristic
RT	reverse transcriptase
RUV	remove unwanted variation
`scone`	Single-Cell Overview of Normalized Expression data
`scRAD`	single-cell Reproducibility Across Donors
scRNA-seq	single-cell RNA-seq
SCVI	scVIsingle-cell Variational Inference
SD	standard deviation
shRNA	short hairpin RNA
SICCA	Sjögren's International Collaborative Clinical Alliance
SNP	single nucleotide polymorphism
SNR	signal-to-noise ratio
r_s	Spearman's rank-order correlation
SPRI	Solid Phase Reversible Immobilization
SRA	Sequence Read Archive
SS	Sjögren's syndrome
ss	single-stranded

SSA	Sjögren's syndrome-related antigen A
SSB	Sjögren's syndrome-related antigen B
SVA	surrogate variable analysis
SVD	singular value decomposition
SW	silhouette width
TC	total count
TF	transcription factor
TFBM	transcription factor binding motif
T_{FH}	follicular helper T
T_{FR}	follicular helper regulatory T
T_H	helper T
T_{Reg}	regulatory T
TLR	toll-like receptors
TMM	trimmed mean of M values
TPM	transcripts per million
tSNE	t-Distributed Stochastic Neighbor Embedding
UMI	unique molecular identifier
UQ	upper quartile
UV	unwanted variation
VL	viral load
VSV-G	vesicular stomatitis virus G glycoprotein
WTA	whole transcriptome amplification
WV	wanted variation
ZI	zero inflation

`ZINB-WaVE`	Zero-inflated Negative Binomial based Wanted Variation Extraction

1

INTRODUCTION

Human disease is fundamentally a system-level phenomenon: symptoms affecting quality and extent of life emerge from the complex interactions of many players – self and non-self – at many scales. To understand disease is to understand the relevant players and their interactions, and in turn, how these shape the system-level disease environment affecting a patient. Even when causal players can be identified (e. g. mutations, pathogens) - any methodology aimed at treatment demands an understanding of systems-level interactions.

One of the most exciting subsystems of the human body is the immune system, responsible for identifying harmful non-self entities and coordinating their destruction or removal; effective immune responses are founded upon the orchestrated dynamics of complex cellular ensembles. Breakdowns in human immunity can have morbid and even deadly effects. Autoimmune diseases such as Sjögren's syndrome (SS) or multiple sclerosis (MS) are believed to be caused by aberrant immune responses to self-tissue, resulting in autoinflammation. Pathogen and host immune system are naturally locked into an adversarial posture: e. g. human immunodeficiency virus type 1 (HIV-1) significantly and detrimentally alters the host's immune response.

Part I of my dissertation motivates the use of observational studies to characterize these disease states and discusses the prospects of applying single-cell technologies to these and other similar studies. Chapter 2 illustrates core *systems immunology* analysis principles using my published work on epigenetic *case–control* differences in tissue collected from SS donors. This chapter demonstrates the importance of *quality control (QC)* and *data normalization*, touching upon issues of *reproducibility* in human subjects studies. The results of this analysis highlight issues associated with bulk tissue experiments and motivate to use of single-cell measurements. Chapter 3 reviews RNA sequencing (RNA-seq) transcriptome analysis and its application to single-cell contexts via new single-cell RNA-seq (scRNA-seq) technologies. This chapter

will describe some of the challenges for scRNA-seq analysis (e. g. modeling missing data) and emphasize how *normalization* techniques may be deployed to address measurement *bias*.

Part II discusses the software tools I have developed to tackle the challenges described in Chapter 3. Chapter 4 outlines the Single-Cell Overview of Normalized Expression data (`scone`) computational workflow for scRNA-seq QC and normalization performance assessment. `scone`'s normalization performance metrics are validated in a series of public data sets and simulations. Chapter 5 outlines the general single-cell Reproducibility Across Donors (`scRAD`) framework for quantifying reproducibility of signals measured over multiple replicate samples. Main software modules are discussed in detail, including an overview of important modifications made to the reproducibility model of Li et al. [1].

Part III describes examples of studies in which I have used the tools in Part II to probe human disease with the help of scRNA-seq. Chapter 6 follows a published work in which I have applied these tools to study reproducible dendritic cell (DC) induction profiles in cells collected from HIV-1 elite controllers (ECs). With the help of my experimental collaborators, I have identify an antiviral DC response phenotype that is characteristic of ECs and inducible in cells from healthy donors (HDs). Chapter 7 describes an unpublished single-cell case-control analysis of cerebrospinal fluid (CSF) cells collected from MS donors. This analysis uncovers important compositional differences in the CSF and leverages a new method – cell set enrichment analysis (CSEA) – to describe subtle changes in T cell expression states.

PART I

BACKGROUND

2

DISEASE SIGNATURES

2.1 INTRODUCTION

A powerful observational approach to understanding immune system processes in human disease is the genetic case–control study: contrasting the genetic characteristics of individuals exhibiting symptoms against those who do not [2]. For example, many studies of immunological control of human immunodeficiency virus type 1 (HIV-1) infection have focused on persons resistant to HIV-1, including elite controllers (ECs)—a rare (~0.5%) subset of HIV-1 infected individuals who naturally suppress viral replication without combination antiretroviral therapy (cART) [3, 4]. Studies contrasting these populations against control populations have uncovered the protective effects of specific *CCR5* and *HLA-B* genetic variants in HIV-1 infection [5–7].

Although genetic association studies like these are well postured to identify candidate loci with causal effects on disease outcomes, findings in HIV-1 have proven insufficient to explain the frequency control in the general population. In particular, these studies have not suggested clinically actionable targets for eliciting an EC-like phenotype in other HIV-1-infected individuals, promoting interest in other cellular components or interactions that could be implicated in coordinating effective host defense.

The study of autoimmune disease suffers from a similar problem of *missing heritability* – individual genetic variants are insufficient to explain disease risk. Missing heritability is observed in Sjögren's syndrome (SS)—a chronic autoimmune disease characterized by progressive destruction of the exocrine glands, with subsequent mucosal and conjunctival dryness [8, 9]. Although the precise cause of SS remains unknown, it is understood to be a complex genetic disease, with multiple weakly associated genetic risk factors [10, 11]. Elucidation of how other, non-genetic factors correlate with disease should significantly improve understanding of this complex disorder.

It is important to note that there is widespread clinical heterogeneity in SS, reflecting differences in underlying disease mechanisms. Current approaches to SS research and research for other diseases are compromised by such phenotypic heterogeneity, motivating careful and methodical phenotyping. Paired with the fact that human tissue samples are precious resources that must be ethically obtained, well designed human subjects recruitment and minimally invasive biopsy procedures are both critically important to guarantee power and generalizability of findings. Studies of circulating blood cells are well suited to reveal novel mechanisms in disease etiology due to ease of sample collection and access to naive cell populations. However, disease-associated changes observed in these cells likely reflect systemic aspects of the disease, rather than tissue-specific disease states driven by local inflammation.

Modern technologies such as microarrays and high-throughput sequencing have made it possible to measure thousands of biological features (e. g. genomic, epigenomic, transcriptomic) in parallel for a single tissue sample. Case–control studies based on these technologies can be very helpful in highlighting the tissues, cell types, and molecular pathways most perturbed by the disease state. While causality can not be demonstrated without experimental intervention, systems-level inferences can be built on many measurements of disease-affected tissue, powered by an ever-growing knowledge base relating measured features to known biological processes.

Even the best efforts to recruit representative sample populations and process samples with uniform quality are susceptible to various forms of unwanted *bias* and *noise*. Furthermore the high dimensionality of high-throughput data sets introduces additional challenges and hurdles, such as the burden of *multiple comparisons*. All of these complications may imperil *reproducibility* and necessitate methods development. In this chapter[1], I will discuss these concepts within the context of a case–control analysis I have performed in SS based on microarray-based epigenetic measurements.

2.2 SJÖGREN'S SYNDROME AND DNA METHYLATION

A growing body of evidence has implicated epigenetic factors, in particular, altered patterns of CpG dinucleotide methylation across nuclear DNA (or

1 This chapter is adapted from a published paper in Arthritis & Rheumatology: "Epigenetic Signatures of Salivary Gland Inflammation in Sjögren's Syndrome." [12] © The Authors and Arthritis & Rheumatology 2016, reproduced with permission.

"DNA methylation"), in models of autoimmune disease [13, 14]. CpG methylation is a reversible chemical modification to DNA and it is actively modulated by chromatin regulators, participating in various chromatin regulatory feedback loops [15]. With only simple chemical modification the methylation state of a CpG site (methylated v. unmethylated) can be measured using the same technologies developed for single nucleotide polymorphism (SNP) profiling [16].

Furthermore, studies characterizing the DNA methylation profiles of naive CD4$^+$ T cells, B cells, and salivary gland epithelial cells provide evidence for aberrant DNA methylation profiles in SS donors [17–20]. While it is unknown which differences, if any, reflect causal determinants of risk, it is likely that many of these patterns reflect subtle differences in the cell type composition of the tissue [21]. Furthermore, not every tissue comparison is the same: e. g. methylation patterns in synoviocytes targeted by rheumatoid arthritis (RA) differ from disease-associated patterns in peripheral blood compartments, with interesting exceptions [22]. One informative compartments for analyzing immunoregulatory heterogeneity in SS is labial salivary gland (LSG) tissue, an accessible target of disease-specific processes [10, 23]. I will discuss below how I applied statistical hypothesis testing to identify thousands of disease-associated DNA methylation differences marking LSG-specific immune processes in SS, implicating both immune-related and cell lineage-specific pathways in disease pathogenesis.

2.3 STUDY SAMPLE

My study analyzed samples of LSG tissue biopsied from 28 female participants in the Sjögren's International Collaborative Clinical Alliance (SICCA) Registry (Table 1). All study subjects were participants in the SICCA Registry, and all were women, the group predominantly affected by SS [24]. The Institutional Review Boards at the University of California, San Francisco and the University of California, Berkeley approved the study protocol.

2.3.1 STUDY SUBJECTS

As part of their enrollment into the SICCA Registry, all subjects were evaluated for clinical criteria of SS at one or two time points; LSG tissue was biopsied at least once during these visits, frozen and subsequently stored. Case–

	Cases (n = 13)	Controls (n = 13)	Test P
Focus score	3.4 ± 2	0.07 ± 0.13	9.1×10^{-6}
Two-eye mean OSS	6.1 ± 2.8	1.2 ± 0.7	1.5×10^{-5}
I (SSA$^+$)	0.92	0	2.7×10^{-6}
I (SSB$^+$)	0.54	0	5.2×10^{-3}
Age in years	55 ± 13	53 ± 7.9	0.84
PC1 of ancestry	0.005 ± 0.003	−0.014 ± 0.026	0.035

Table 1: *Covariates across study groups.* Values are the mean±SD of covariates, except for SSA/SSB seropositivity, represented as indicator variables (TRUE/-FALSE): only the mean values (proportions) are reported for seropositivity phenotypes. Two-sided P-values were computed using MWW tests for focus score, OSS, and age. FET P-values were reported for SSA/SSB seropositivity.

control status was determined according to the 2012 American College of Rheumatology (ACR) criteria for SS [23]; as a result, this analysis targeted "cases" with severe SS, requiring they meet all three of the following criteria:

- autoantibody (auto-Ab) seropositivity: positive anti-Sjögren's syndrome-related antigen A (SSA) and/or anti-Sjögren's syndrome-related antigen B (SSB) auto-Ab serology;
- keratoconjunctivitis sicca (KCS): ocular staining score (OSS) of ≥ 3 in at least one eye;
- focal lymphocytic sialadenitis (FLS): LSG biopsy section with focus score ≥ 1 focus /4 mm^2,

"Control" subjects must not satisfy any of these criteria. Samples were designated as case or control based on the clinical evaluation at the time of biopsy. Two of the study subjects met only the high OSS criterion at time of sample collection, referred to here as "KCS-only" subjects. Importantly, neither cases nor controls were disqualified based on an additional systemic autoimmune disease diagnosis (e. g. RA or Hashimoto's disease). Based on these criteria, I classified 13 SS cases, 13 controls, and 2 subjects with KCS-only phenotypes. It is possible that SS case subgroups (e. g. cases with specific extraglandular manifestations) exhibit unique DNA methylation profiles; the study was not large enough to test this hypothesis.

Self-reported medication data was collected for all participants. Medications may be used to treat symptoms of disease, and they can have a substantive effect on tissue environments of interest (e. g. inflammation): I conducted Fisher's exact tests (FETs) to compare medication usage between cases and controls in order to identify medications that could confound case–control analysis. The four drugs shown in Table 2 exhibited the smallest P-values of all 53 drugs reported (data not shown), providing no significant evidence for rejecting the null hypothesis of independence. This result boosted confidence in the absence of a confounding effect of medication.

Medication	Cases (n = 13)	Controls (n = 13)	FET P
Levothyroxine	8	3	0.11
Folic acid	0	3	0.22
Calcium	5	9	0.24
Vitamin D	2	5	0.38

Table 2: *Self-reported medication by case–control status.* Counts of subjects self-reporting medication from case and control groups. P-values were computed by Fisher's exact test for independence.

Prior to this study, the 28 subjects had been genotyped using the HumanOmni2.5-Quad BeadChip (Illumina), as part of a genome-wide association study (GWAS) [25]. In addition to sample verification (Subsection 2.3.2), these data were used to evaluate the genetic ancestry of the study subjects. Collaborators had applied `EigenStrat` [26] to genotypes from the full GWAS data set in order to derive principal components (PCs) reflecting the primary axes of genetic variation. The 28 study subjects fell within 2 standard deviations (SDs) of the mean of the first 2 PCs of self-identified Europeans from the original study; all GWAS subjects within this range represent a relatively homogenous genetic background.

One of the strengths of the study is its restriction to women with similar ancestry: both genetic ancestry and sex have been shown to influence DNA methylation profiles [27, 28]. While this design minimizes the potential for confounding by genetic ancestry or sex it also limits the generalizability of findings extended to non-European or male populations. Importantly, there are known to be many important immunologic differences between the sexes [29]. As a result, epigenetic studies comparing male cases and controls might yield a different set of SS-associated patterns.

Because the DNA methylation signal of interest is located on genetic material, it is important to consider whether biased ancestry sampling - even at the intra-European-level - could confound the measurement. In order to represent intra-European ancestry for our study subjects, collaborators applied `EigenStrat` analysis to genotypes from subjects with European ancestry, as defined above. The first 4 PCs were considered in this study of DNA methylation. Non-parametric Mann–Whitney–Wilcoxon (MWW) tests provided no significant evidence of case–control difference in the first ancestry PC. Similarly, age - another potential confounder - showed no significant association with case–control status (Table 1).

2.3.2 TISSUE AND MEASUREMENT PLATFORM

Whole LSG DNA methylation data were obtained for each sample using the Illumina 450K Infinium Methylation BeadChip (or "450K chip") platform for bulk sample methylation profiling [16]. The 450K chip allows for high-throughput interrogation of >450,000 highly informative CpG sites spanning ~ 22,000 genes across the genome. The primary measure of DNA methylation at each CpG site is the β-value: defined as the ratio of the intensities of fluorescent signals from methylated and unmethylated alleles. Sample identity was verified by comparing GWAS genotypes to the genotypes measured by 35 SNP probes on the 450K chip. Samples were prepared on three separate plates or "batches" on different dates. Three of the LSG DNA samples were divided into two within-batch technical replicates, contributing to a total of 31 samples for subsequent DNA methylation analysis (Table 3).

Plate (batch)	Case (n = 15)	Control (n = 14)	KCS-only (n = 2)
Batch 1	8 + 1	8	2
Batch 2	5 + 1	1	0
Batch 3	0	4 + 1	0

Table 3: *Distribution of sample types per batch.* Counts of case, control, and KCS-only samples (including replicates) for each plate (batch). Replicate samples share the same batch: batches 1 and 2 each have one case replicate pair, and batch 3 contains a control replicate pair. Without double-counting replicated samples, $P = 0.094$, computed by FET for independence; $P = 0.044$ if KCS-only subjects are excluded from the test.

2.4 DATA FILTERING

Subsection 2.3.1 described some problematic subject-level covariates that could bias or add noise to my analysis. There are plenty of reasons that subsets of measurements could be problematic as well. This study considers far more more measurements per sample than samples, giving me license to throw out the most problematic probe channels. One important reason to remove a measurement from consideration is irrelevance; the 450K chip includes 3,091 CpH probes and 65 SNP probes; all were removed from downstream analysis due to my interest in CpG methylation. Another important reason to remove a features is a lack of interpretability; Chen et al. [30] had identified a large number of CpG probes that were "cross-reactive": hybridizing with multiple off-target sequences across the genome. Because the signal observed in these probes are not easily localized to any part of the genome, I removed all 16,177 from the analysis.

Although I did not see significant genetic ancestry differences between cases and controls, I can not rule out genetic heterogeneity at the single base pair (bp)-level. Genetic heterogeneity at or around these sites can introduce noise and obscure interpretable signals. In order to avoid the direct effect of genotype variation on CpG sites, I removed 1,213 CpG probes targeting SNPs known to be variable from the matched GWAS data. I also considered the larger set of SNPs from the 1000 Genomes project lying within 450k chip probe-hybridizing sequence as tabulated by Chen et al. [30]. Overlapping the probe list with the SNP138 track in the UCSC Genome Browser [31, 32], I identified and removed 62,220 CpG probes neighboring known SNPs.

It may be useful to remove features which are too noisy for technical reasons; an additional 3,392 CpG probes were removed from the analysis due to high detection P-values ($P > 0.05$) in one or more samples, as computed by Illumina's `GenomeStudio` software. The signal in these probes could not be distinguished from background. After filtering, I considered a total of 404,353 CpG probes for downstream analysis.

2.5 DATA NORMALIZATION

The 450k chip generates fluorescence-based readouts, and these signals must be normalized computationally so that they correlate with the biological signal of interests rather than technical factors. The DNA methylation normal-

ization pipeline used in this study was implemented entirely in R [33] and leveraged the methylumi data representation in Bioconductor [34, 35].

2.5.1 LOW-LEVEL NORMALIZATION

Before the probe filtering described in Section 2.4, I had previously applied the normal-exponential convolution method on out-of-band probe intensities ("noob") to correct each sample for technical variation in background fluorescence [36]. The red and green intensity channels on the 450K chip were normalized so as to be comparable, using the all-sample mean normalization method—a natural extension of the Illumina `GenomeStudio` color-channel normalization protocol [37]. After probe filtering, I corrected each sample for within-sample probe design bias using the beta-mixture quantile normalization method [38].

2.5.2 BATCH CORRECTION

One approach for understanding high-dimensional data is to consider the primary axes of variation in that data, using *dimensionality reduction* methods such as principal component analysis (PCA), followed by *post hoc* correlational analyses to annotate these axes in terms of known sample-level covariates, including technical covariates such as batch. Despite my first-pass normalization using standard methods, PCA clearly separated samples according to batch (Figure 1). Downstream analysis, particularly unsupervised analysis could be improved by batch correction.

However, direct adjustment for batch effects using a standard batch correction method, such as ComBat [39], may removes case–control differences along with the batch artifacts. This is because the cases and control samples are not evenly distributed across the three plates (Table 3). Although the distribution of samples is not inconsistent with random uniform sampling (FET $P > 0.01$), the noise nevertheless presents challenges for batch correction. For example, batch 3 contains only control samples; any adjustment of the data that guarantees methylation profiles from batch 3 closely resemble profiles in the other batches could easily misrepresent true biological signals. Another - more fundamental - problem with a standard batch correction strategy is that it relies on a categorical batch-level covariate as a proxy for differences in individual sample qualities. Sample quality varies within batches as well as between them and qualities may overlap between batches. I therefore

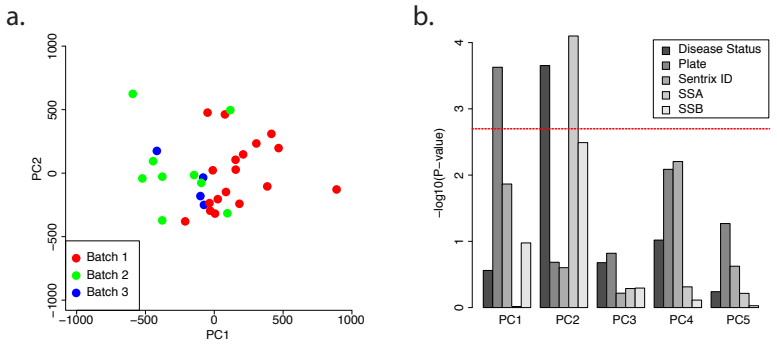


Figure 1: *PCA of genome-wide gland DNA methylation, before batch correction.* **(a)** Before batch correction, the first PC separates the three batches. **(b)** PCA on genome-wide gland DNA methylation, before batch correction. KW test P-values for Fisher-transformed r_s, showing significance of correlation between PCs 1-5 and categorical covariates: disease status (three-level), plate (batch), sentrix ID (chip), SSA and SSB seropositivity. Red dashed line indicates Bonferroni-corrected significance threshold for 25 tests, controlling the FWER.

adjusted the data against quality control (QC)-based proxies of known batch effects, rather than batch identifiers; this approach is similar in spirit to the method of Fortin et al. [40].

The 450K chip includes 850 QC probes (1,696 color channels) that measure different aspects of sample quality, invariant to biological context. I use these control probe channels to adjust for subtle technical variation across batches, using PCA to derive a sample-level quality factor, or "qPC," representing the primary axis of technical variation in the data. In selecting the input to this PCA, I score all control probe channel intensities according to their correlation with the PCs of the CpG *M-value* matrix. M-values are logit[2] transformed β-values for which PCs can be thought of as logistic modulations in relative methylated CpG ratios, rather than additive modulations. The correlation score, s_i, of a control probe channel i is computed as the weighted mean of that channel intensity's squared Spearman's rank-order correlation (r_s) with all PCs of the CpG M-value matrix, weighted by the variance of each PC:

$$s_i = \frac{\sum_n \mathrm{Var}(PC_n^{(\text{M-value})}) r_s(p_i, PC_n^{(\text{M-value})})^2}{\sum_n \mathrm{Var}(PC_n^{(\text{M-value})})} \tag{1}$$

2 the logit function is defined as $\mathrm{logit}(x) \equiv \log(x/(1-x))$ for $x \in [0,1]$

Control probe channels are selected as candidate confounding QC channels if their correlation score exceeds one median absolute deviation (MAD) from the median score across all control probe channels. PCA is performed on log-transformed confounding QC channel intensities and the first PC is used as the qPC.

The effect of qPC can be removed from the data in many different ways (e. g. beta regression). I implemented a bin-based centering procedure, placing each sample j into one of four (# of batches +1) equally-sized bins according to the value of their quality parameter: $b_j = f(qPC)$. These bins can be thought of as "pseudo-batches," grouping samples with similar sample quality. For each CpG, I computed the mean β-value across all samples ("global mean") as well as four bin-specific means.

$$\begin{aligned} \overline{\beta}_i &= \frac{1}{J}\sum_j \beta_{ij} \\ \forall k \in \{1,2,3,4\}, \tilde{\beta}_{ik} &= \frac{1}{\sum_j \mathrm{I}\left(b_j = k\right)} \sum_j \beta_{ij}\,\mathrm{I}\left(b_j = k\right) \end{aligned} \tag{2}$$

I defined an adjusted M-value by taking the original M-value, subtracting the logit transformed bin-specific mean β-value, and adding the logit transformed global mean β-value. The adjusted β-value is calculated by performing the inverse logit (logistic or "expit") transformation on the adjusted M-value. This last step naturally limits the adjusted β-value to the original range from zero to one.

$$\mathrm{logit}\left(\beta_{ij}^{(adjusted)}\right) = \mathrm{logit}\left(\beta_{ij}\right) - \mathrm{logit}\left(\tilde{\beta}_{ib_j}\right) + \mathrm{logit}\left(\overline{\beta}_i\right) \tag{3}$$

After adjusting for these technical effects, none of the top 5 DNA methylation PCs (60% of variance) showed significant association with the sample batch (Figure 2).

Averaging DNA methylation PC values for replicates, I tested the top 5 PCs for association with age and genetic ancestry PCs, applying two-tailed Z-tests to Fisher-transformed r_s (Table 4). There didn't appear to be any significant associations between age, ancestry, and PCs of methylation, and thus no need to adjust for ancestry or age as we have for batch. Nevertheless, I will screen any case-associated DNA methylation differences for marginal effects of ancestry PC1 and age.

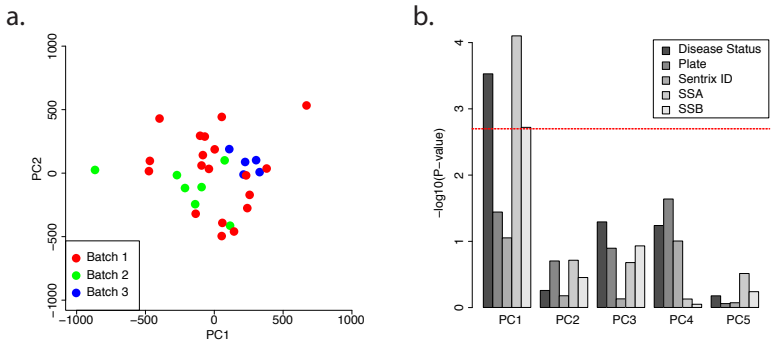


Figure 2: *PCA of genome-wide gland DNA methylation, after batch correction.* **(a)** Following batch correction, batches are mixed. **(b)** PCA on genome-wide gland DNA methylation, after batch correction. KW test P-values showing significance of correlation between PCs 1-5 and categorical covariates: disease status (three-level), plate (batch), sentrix ID (chip), SSA and SSB seropositivity. Significance of correlation with batch is low across first five PCs, while preserving associations with phenotype. Red dashed line indicates Bonferroni-corrected significance threshold for 25 tests.

2.6 ANALYZING DISEASE SIGNATURES

2.6.1 UNSUPERVISED ANALYSIS

Above we have describes an unsupervised dimensionality reduction / correlative approach for exploring correlated signals in high-throughput data - applying it to understanding technical bias. This approach may also be useful when multiple biological conditions or factors are present in the data, because it facilitates prioritization of specific hypothesis testing regimes. I tested the top 5 PCs for association with focus score and mean OSS, as I had tested age and ancestry above. The first PC was strongly associated with the focus score ($Q = 2.1 \times 10^{-5}$) and the mean OSS ($Q = 5.3 \times 10^{-4}$) (Table 4), suggesting that this axis captures disease-associated processes in the gland.

As shown in Figure 2, Kruskal–Wallis (KW) one-way analysis of variance (ANOVA) tests provide evidence for PC1's association with disease status and seropositivity (Bonferroni-adjusted $P < 0.05$). MWW testing on replicate-averaged PC values suggests that the first PC of DNA methylation in LSG tissue is associated with case–control status ($P = 1.3 \times 10^{-5}$). Plots of the first 2 PCs place the two KCS-only individuals between the cases and the con-

	r_s (PC1, PC2)	Z-test Q (PC1, PC2)
Focus score	−0.77, 0.20	2.1×10^{-5}, 0.070
Two-eye mean OSS	−0.70, 0.21	5.3×10^{-4}, 0.70
Age in years	−0.15, 0.22	0.76, 0.70
PC1 of ancestry	−0.41, −0.17	0.32, 0.70

Table 4: *Spearman's rank-order correlation between first and second PCs of DNA methylation, and continuous covariates.* BH Q-values for Z-tests on Fisher-transformed r_s of first five PCs against continuous covariates.

trols, consistent with an intermediate phenotype (Figure 3). Perhaps it should come as no surprise, given the study design, that the primary axis is correlated with case–control status. However, recall that this was not true before batch correction.

2.6.2 SUPERVISED ANALYSIS

MWW testing was used to test each CpG's β-value for association with case–control status, followed by the Benjamini-Yekutieli (BY) adjustment for controlling the false discovery rate (FDR) under multiple comparisons. The BY adjustment, as implemented in `p.adjust` [33], is a more conservative version of the Benjamini-Hochberg (BH) FDR procedure, which may be preferable when test statistics are correlated [41]. Given large correlations between CpG methylation levels in this data, I chose to use this more conservative FDR procedure. I set no constraints on the magnitude of significant differences in methylation. I refer to disease-associated CpGs ($Q < 0.01$) as differentially methylated positions (DMPs). The β-values for replicate samples were averaged prior to single CpG-site association tests.

This association study identified 7,820 DMPs associated with SS case status. The median absolute β-difference between cases and controls ($\Delta\beta$) was 0.10 for DMPs, demonstrating that most SS-associated DMPs identified in the current study showed modest-to-large differences in DNA methylation.

Of the 7,820 DMPs tested, 5,699 (73%) were hypomethylated in cases. The set of DMPs contained far more hypomethylated CpGs than was expected by the distribution of non-DMPs ($P < 2.2 \times 10^{-16}$ by FET) (Table 5), suggesting that CpGs are generally more hypomethylated in whole LSG tissue from SS

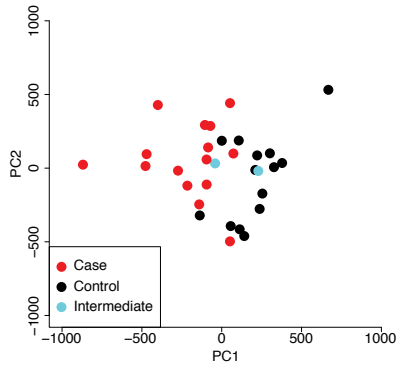


Figure 3: *PCA of genome-wide DNA methylation in all LSG tissue samples, including replicates.* PC1 separates SS cases from controls, with samples from the two subjects with KCS-only phenotype (OSS ≥ 3 in at least one eye) between those of the cases and the controls. PC2 represents a spread of sample DNA methylation profiles orthogonal to the primary case–control contrast. This axis may represent biological between-donor heterogeneity.

cases. Thabet et al. [17] previously reported whole genome hypomethylation in cultured LSG epithelial cells from SS donors. Despite the hypomethylation enrichment cited above, no significant differences in mean genome DNA methylation were observed across all CpGs (1.1 fold hypermethylation in SS cases; $P = 0.26$ by MWW). Of course, the tissue samples analyzed in my study are not cultured and are composed of LSG epithelial cells as well as other cell types; the discrepancy here could be due to differences in sample biology. Alternatively, the 450k chip targets may poorly represent the distribution of CpG sites on the genome-level.

I used the `lm` function in R [33] to fit a linear model relating each of the 7,820 DMP DNA methylation levels (logit transformed) to the first PC of genetic ancestry or age at biopsy. No DMP was significantly associated with either factor by Student's t-test at a BH controlled FDR of 0.05. These two factors may affect DNA methylation levels of SS-associated DMPs, but their average effects are too small to resolve in my study.

Prior to this analysis, Imgenberg-Kreuz et al. [20] reported results from their study of DNA methylation in minor salivary gland biopsies from 15 primary SS cases and 13 controls in which they used the same 450K chip. In addition to a parametric analysis approach, the authors used a conservative Bonferroni-adjusted P-value reporting criterion for DMPs. While one top

	$Q < 0.01$	$Q > 0.01$
hypermethylated	2,121	227,666
hypomethylated	5,699	168,867

Table 5: *Global DNA methylation differences between LSGs from the SS cases and the controls.* Global DNA methylation differences between LSGs from SS cases and controls. Proportions of hyper- and hypomethylated CpGs with Q-values < 0.01 or > 0.01 by MWW test. The latter represent a control set of CpGs, or non-DMPs. Direction of methylation is determined by the sign of the difference in the mean β-values between cases and controls. DMPs are significantly enriched for hypomethylated sites as compared to non-DMPs ($P < 2.2 \times 10^{-16}$ by FET).

"hit" from that study – cg20870559 in *OAS2* – was successfully replicated in the current study, only two of the remaining 44 DMP hits reported by that study were replicated here: cg12560128 and cg16596716. Both study populations were small, and differences in phenotype or age may have contributed to the lack of replication.

Previous studies have defined a gene as being differentially methylated if it contains a number of DMPs exceeding a given threshold [42]. One problem with this approach is that it is biased toward reporting genes with higher CpG coverage. Assuming that false-positive results would be randomly distributed across the 450K chip, a gene with better coverage will have more false-positive results. Coverage is also problematically associated with biologic function [43], but enrichment tests, such as the hypergeometric test, will take this coverage into account. Enrichment analyses and more comprehensive analyses of extended patterns of DNA methylation may be better approaches to characterizing profiles associated with case status than single CpG-site testing.

2.7 ENRICHMENT ANALYSIS

Once differences between cases and controls are identified, it can be asked what these changes mean within the greater systems context. CpG methylation is known to be associated with chromatin silencing [15]; therefore, it is reasonable to hypothesize that CpG sites neighboring up-regulated genes are hypomethylated in the disease state. The microarray study of Hjelmervik

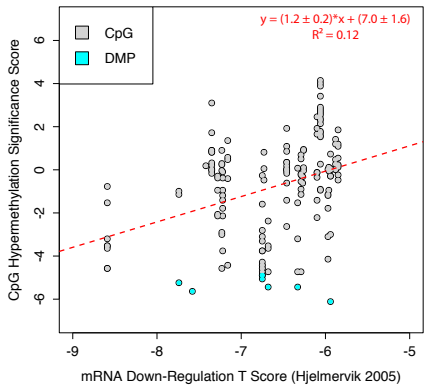


Figure 4: *Correlation between DNA methylation and expression disease-associations from two studies.* The x-axis shows extent of mRNA down-regulation from Hjelmervik et al. [44] and the y-axis shows the significance of DNA hypermethylation - from the present study - for CpGs in the promoter of the corresponding gene. DMPs from the present study are highlighted in cyan.

et al. [44] reported 50 ribonucleic acids (RNAs) that are highly differentially expressed in SS cases compared to controls, allowing us to resolve the relationship between methylation and expression at a population-level.

CpGs were mapped to promoters using the `BEDTools` suite [45]. For each RefSeq entry in the UCSC RefGene track [32], I defined as "promoter" the genomic interval spanning 2,500 bp upstream and 500 bp downstream of the annotated transcription start site, similar to the definition described by Whitaker et al. [42]. RefSeq identifiers were mapped to gene symbols using the `org.Hs.eg.db` package in Bioconductor [46]; all unmapped RefSeq entries were excluded from the analysis. For the 42 gene symbols overlapping between my study and Hjelmervik et al. [44], I assigned hypomethylation (decrease in methylation) significance scores to each CpG falling in their promoters:

$$\text{score} = \text{sign}\left(-\Delta\beta\right) \times -\log P \tag{4}$$

Linear modeling suggests that the average hypomethylation score across a promoter is positively associated with the extent of messenger RNA up-regulation reported in SS-affected tissue (Figure 4).

2.7.1 DIFFERENTIALLY METHYLATED PROMOTERS

The predictive power of differential expression suggests that many DNA methylation differences in LSGs from SS cases are associated with the same upstream biologic factors driving differential transcription in SS. I searched genome-wide for differentially methylated promoters using hypergeometric hypothesis testing for DMP enrichment, as described by Nakano et al. [47]. Enrichment P-values were controlled for multiple testing using a BH Q-value threshold of 0.05. To avoid promoter-specific bias, I excluded all CpGs that did not fall within promoters; enrichment tests were performed solely on promoter CpGs. Furthermore, to protect against biases associated with double-counting CpGs sitting in the intersection of multiple loci, I excluded any CpGs mapping to two or more promoters.

Differentially methylated promoter analysis identified 57 genes (Table 6). This list includes a large number of genes encoding transcription factors (TFs) (e. g. *RUNX3* and *SPI1*, latter not shown in table: 22-fold enrichment and $Q = 0.018$) and known cell-differentiation markers (e. g. *TNFRSF13B*, *CCR6*, *BST2*, *BTLA*, and *CXCR5*). In addition to protein-coding genes, the list contains a number of RNA genes, including several antisense RNA genes (e. g. *PSMB8-AS1*) and microRNA (miRNA) genes (e. g. *MIR339*). These results could reflect differential regulation of neighboring coding genes or primary transcripts. Interestingly, three of the differentially methylated promoters are located within one interval of the major histocompatibility complex (MHC) genomic region: *PSMB8*, *PSMB8-AS1*, and *TAP1* (Figure 5).

The most significant promoter-level DMP enrichment was for *PSMB8-AS1*, a long noncoding RNA gene neighboring the *PSMB8* locus (a.k.a. "*PSMB5i*" or "*LMP7*") in the MHC region. This gene encodes an antisense RNA, and is in a head-to-head configuration with *PSMB8* gene (Figure 5). *PSMB8*, whose promoter is also hypomethylated in SS cases, encodes a subunit of the immunoproteasome that has been reported to be up-regulated in the salivary glands of donors with SS [48]. I have also presented evidence here for promoter hypomethylation of *TAP1* ($Q = 0.0016$), neighboring both *PSMB8* and *PSMB9*. Rare variants of *TAP1* and extended HLA haplotypes are thought to confer disease risk upon some SS donors [49]. Given their specific roles in antigen presentation, most DMPs observed across these three neighboring loci are likely to be directly associated with an increased proportion of immune cells in the tissue.

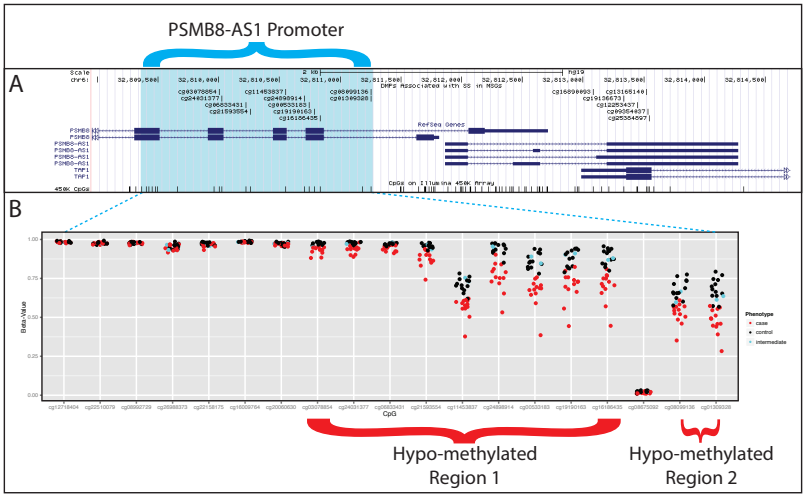


Figure 5: *Extended differential methylation in PSMB8-AS1 promoter.* **(a)** Highlighted region shows region designated as *PSMB8-AS1* promoter, sitting within the gene body of *PSMB8*. All SS-associated DMPs (promoter and non-promoter) are annotated in the top track, UCSC Genome Browser RefGene annotations in the middle, and all 450K chip CpG sites at the bottom. **(b)** Evidence of an SS-associated differentially hypomethylated region within the promoter of *PSMB8-AS1*.

Thabet et al. [17] report that disease-associated gland up-regulation of ICAM1/CD54 [10], a gene critically involved in the processes of intercellular adhesion and trans-endothelial migration, was associated with global hypomethylation of salivary gland epithelial cell genomes. The investigators hypothesized that global hypomethylation could be a regulatory mechanism upstream of increased expression [17]. I found no evidence of differential methylation in or around the *ICAM1* promoter, an observation consistent with other mechanisms being more directly responsible. However, due to the heterogeneous nature of gland tissue used in the current study, both direct and indirect effects may be masked by cell proportion differences in tissue.

Promoter enrichment analysis highlighted a miRNA (*miR-339*)—a potential posttranscriptional regulator of *ICAM1* [50]. Although this *ICAM1* regulatory mechanism is an attractive explanation of the data, there exists little evidence to support it within the context of SS, beyond down-regulation of *miR-339* reported in a microarray study of SS-affected glands [51]. Any mechanistic interpretation is further complicated by the hypomethylation observed in the promoter, which would support up-regulation of this gene product based

on a simple model of DNA methylation-associated epigenetic regulation. Despite the unknown biologic role of the striking hypomethylation I identified at this miRNA locus, the proposed regulatory potential of *miR-339* makes it an intriguing candidate for functional followup.

2.7.2 GENE SET ENRICHMENT ANALYSIS

After identifying the set of genes with significantly differentially methylated promoters, I considered whether the gene set is enriched for categories of biologic function or genomic position. Hypergeometric gene set enrichment analysis was used to test 2,666 gene sets from Molecular Signatures Database (MSigDB) [52, 53] for enrichment of differentially methylated promoters, including "hallmark" gene sets, positional gene sets, motif gene sets, and gene ontology gene sets, using a BH Q-value cutoff of 0.05.

I further tested two candidate gene sets for enrichment of genes possessing differentially methylated promoters:

1. genes encoding the 50 RNA transcripts showing the greatest fold change in LSG expression between SS cases and controls in Hjelmervik et al. [44],

2. genes highlighted in recent SS GWAS: *GTF2I*, *TNFAIP3*, *IRF5*, *STAT4*, *IL12A*, *BLK*, *CXCR5*, *TNIP1*, *HLA-DRA*, *HLA-DQB1*, *HLA-DRB1*, *HLA-DPB1*, and *COL11A2* [54, 55].

Although promoter-level gene set enrichment analysis is a valuable tool for understanding the distribution of differentially methylated promoters, the DMPs on which this analysis is based are called at single-bp resolution. This discrepancy can lead to biased reporting due to the variation in promoter coverage across the 450K chip platform; some promoters contain far more probed CpGs than others resulting in greater power to resolve extended differences in those regions. Some of this bias of differential power can be avoided by considering CpG sets rather than gene sets.

For each of the differentially methylated gene sets identified in the hypergeometric gene set enrichment analysis, as well as the two candidate gene sets, a CpG set was defined containing all CpGs mapping to promoters of the corresponding gene set. DMP enrichment was then performed using hypergeometric tests, as before, although CpGs mapping to multiple sets were included in this analysis. The CpG set enrichment analysis was adjusted for multiple

testing using Bonferroni-adjusted P-values, accounting for the 2,668 gene set enrichment tests used to select CpG sets. CpG sets with a adjusted P-value less than 0.01 were considered enriched for DMPs.

CpG set enrichment results emphasized both the inflammation and tissue specificity of the observed DNA methylation differences. DMPs were found to be enriched for several gene ontology terms involving immune response and signal transduction. Only a small number of these genes have been highlighted by SS GWAS (*CXCR5* and *BLK*) [54] or are known to be differentially expressed at the transcription-level in SS-affected LSG tissue (*ARHGAP25*) [44]; however, the promoter CpG sets corresponding to both of these candidate gene sets were significantly enriched for DMPs (Bonferroni-adjusted $P = 9.2 \times 10^{-7}$ and 6.0×10^{-4}, respectively). I also observed evidence of enrichment for promoters containing transcription factor binding motifs (TFBMs) for PU.1 or Ets2 (in mouse, by orthology; Table 7), likely representing differences in cell composition and activity resulting from SS pathogenesis. This tissue heterogeneity interpretation is further supported by the abundance of differentially methylated cell differentiation markers noted in my DMP enrichment analyses; this enrichment could indicate that many-to-most of the extended DNA methylation differences observed in this study are consequences of varying cell proportions in the gland tissue. As a deeper understanding of cell type-specific DNA methylation motifs in immune- and tissue-specific cells becomes available, the patterns observed in target tissue may serve as clues to which cell types are driving recurring inflammation in SS donors.

2.7.3 TFBM ANALYSIS

Given the intimate relationship between transcription-factor binding and chromatin state, I also considered whether disease-associated DNA methylation changes colocalize with specific TFBMs, using the Analysis of Motif Enrichment (`AME`) tool [56] to identify enriched TFBMs in the sequence surrounding disease-associated DMPs (Figure 6). For each DMP, I extracted a window of the UCSC hg19 reference genome within 150 bp of the annotated CpG position—a half-width similar to the length of a nucleosome. Overlapping intervals were merged, producing a set of DMP-associated sequences. A control set of CpG-neighboring sequences was generated using the same procedure applied to all non-DMPs.

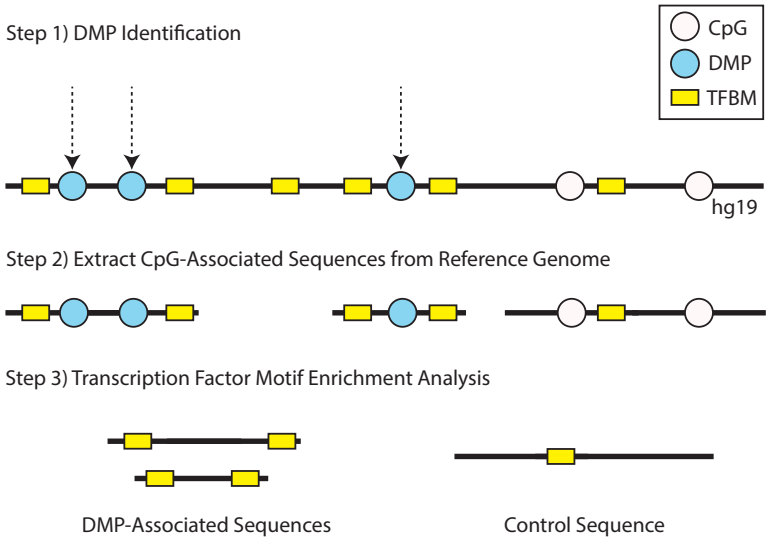


Figure 6: *JASPAR motifs enriched in the neighborhood of SS-associated DMPs in labial salivary gland tissue.* Schematic of TFBM enrichment analysis.

Using the AME, I tested DMP-associated sequences for enrichment of 205 TFBMs from the JASPAR CORE 2014 vertebrates set [57], adjusting for sequence length and using the control set as a sequence control. AME was performed using three motif affinity options that use different scoring methods to evaluate the number and strength of motif matches: total number of matches above a threshold ("totalhits"), sum of motif scores ("sum"), and average motif score ("avg"). Default thresholds were used for all choices of motif affinity function, and observed enrichment was evaluated for statistical significance using FETs. Motifs were considered enriched if the corresponding Bonferroni-adjusted P-value is below 0.01 for any of the three affinity functions (controlling the family-wise error rate (FWER) for $3 \times 205 = 615$ tests).

The AME tool identified three enriched motifs in the immediate neighborhood of DMPs (Table 8). The most significant motif was annotated for TCF11/MafG [58], an antioxidant response element binding complex that is reported to play a role in proteasome regulation and stability [59]. A second enriched motif was annotated for the STAT1/STAT2 heterodimer, targeting interferon (IFN)-stimulated response elements [60]. The final motif is the conserved binding motif of PU-box-binding TF PU.1 [61].

The greater proteasome regulatory network was previously implicated by the enrichments reported above around *PSMB8*. While these differences in

DNA methylation may be functionally related, there is no clear evidence of *immuno*-proteasome regulation by the TCF11/MAFG complex [59].

The TF PU.1 was also highlighted multiple times in the current study. Not only was extended hypomethylation observed in the promoter region of this gene, but there also appeared to be a spatial association between differential methylation patterns and PU.1 binding motifs, both at the promoter-level (CpG set enrichment analysis) and at the nucleosome-level (TFBM enrichment analysis). PU.1 is a known factor involved in B cell and macrophage differentiation, binding to the enhancers of many lineage-specific genes [62], and it may directly recruit DNA methylation machinery to repress target genes [63]. As such, differential proportions of immune cell types (i. e. B lymphoid versus myeloid lineage) may drive PU.1 target enrichments in inflamed tissue. In particular, the abundance of hypomethylated B cell and lymphoid markers, including *CD19* ($Q = 0.046$), *CD79B* ($Q = 0.046$), *PTPRCAP* ($Q = 1.2 \times 10^{-7}$), and *TNFRSF13B* ($Q = 1.8 \times 10^{-5}$), further supports this interpretation.

2.8 CONCLUSIONS

Through whole-genome DNA methylation profiling of a clinically well-characterized sample of European women, I identified a strong signature of disease-associated immune processes in LSG tissue. I observed evidence of targeted hypomethylation at the tissue-level in SS cases as compared to controls. Further, my findings showed that epigenetic states of inflammatory genes and immune-cell markers are major contributors to DNA methylation differences that distinguish SS cases. While results from this observational study cannot establish a causal role for the observed DNA methylation patterns in the risk of SS, my DMP-based CpG set, and TFBM enrichment analyses all demonstrated that DNA methylation profiling in SS cases and controls provides unique insights into tissue-specific differences involved in disease [64].

Labial salivary gland biopsy is a minimally invasive procedure that provides investigators access to tissue targets of SS and may help to illuminate processes specific to a disease in progress. Furthermore, as a target tissue, these samples may prove more useful in characterizing disease phenotypes in donors with early evidence of SS symptoms. Insights from this study and larger studies may soon yield new epigenetic biomarkers for this complex and heterogeneous disease and may help to inform the development of novel treatment strategies in the future.

Promoter	DMPs	Fold enrichment	Q-value	Direction
PSMB8-AS1	11	38.3	1.4×10^{-11}	⇓ hypo-
CTSZ	10	26.5	1.9×10^{-8}	⇓ hypo-
PTPRCAP	8	35.3	1.2×10^{-7}	⇓ hypo-
LTA	7	38.6	7.6×10^{-7}	⇓ hypo-
MIR339	7	30.9	4.8×10^{-6}	⇓ hypo-
TNFRSF13B	5	55.1	1.8×10^{-5}	⇓ hypo-
PSMB8	7	22	5.7×10^{-5}	⇓ hypo-
MTNR1A	5	33.1	0.00053	⇑ hyper-
MPEG1	4	52.9	0.00066	⇓ hypo-
CCR6	5	27.6	0.0013	⇓ hypo-
TAP1	4	44.1	0.0016	⇓ hypo-
SSH3	5	23.6	0.0027	⇑ hyper-
BST2	4	37.8	0.0029	⇓ hypo-
PPFIA4	4	37.8	0.0029	⇓ hypo-
AIM2	3	66.1	0.0036	⇓ hypo-
BTLA	3	66.1	0.0036	⇓ hypo-
CXCR5	5	20.7	0.0036	⇓ hypo-
FCRL3	4	33.1	0.0036	⇓ hypo-
KCNQ1DN	7	10.8	0.0036	⇑ hyper-
LINC00926	3	66.1	0.0036	⇓ hypo-
MIR3186	4	33.1	0.0036	⇓ hypo-
MIR4269	3	66.1	0.0036	⇓ hypo-
WDFY4	5	20.7	0.0036	⇓ hypo-
RUNX3	7	10.1	0.0055	⇓ hypo-
FERMT3	4	26.5	0.0093	⇓ hypo-

Table 6: *Top promoter DMPs enrichments in LSGs from SS donors.* Promoter enrichment results are shown for the most-significant promoters ($Q < 0.01$). Both the total number of DMPs and the fold enrichment for DMPs in the region are reported. Q-values are reported for one-tailed hypergeometric tests for enrichment. "Direction" column notes whether all DMPs in the promoter were hypomethylated (⇓) or all were hypermethylated (⇑).

MSigDB gene set	Differentially methylated promoters	P (Adjusted)
Immune response GO:0006955	*CCR6, BST2, AIM2, LCP2, CD79B, MADCAM1*	2.9×10^{-8}
Intrinsic to plasma membrane GO:0031226	*TNFRSF13B, MTNR1A, CCR6, BST2, CXCR5, NCKAP1L, CD160, CD19, CD79B, IL12RB1*	1.7×10^{-7}
Genes with promoters containing Ets2 motif RYTTCCTG M14654	*PTPRCAP, TNFRSF13B, KCNQ1DN, RUNX3, FERMT3, LCP1, SPI1, SLAMF1, CD19, ERG, PIK3CG*	3.6×10^{-7}
Immune system process GO:0002376	*CCR6, BST2, AIM2, SPI1, LCP2, CD79B, MADCAM1*	1.0×10^{-6}
Cell surface receptor–linked signal transduction GO:0007166	*TNFRSF13B, MTNR1A, CXCR5, CD160, GNB3, LCP2, CD19, IL12RB1, PIK3CG*	2.9×10^{-6}
Genes with promoters containing PU.1 motif WGAGGAAG M14376	*PTPRCAP, LTA, NCKAP1L, LCP2, NR1H3, PIK3CG*	4.7×10^{-5}
Signal transduction GO:0007165	*LTA, TNFRSF13B, MTNR1A, CCR6, BST2, CXCR5, CD160, GNB3, BLK, LCP2, CD19, ERG, IL12RB1, KALRN, MADCAM1, PIK3CG*	2.8×10^{-4}

Table 7: *Differentially methylated CpG sets in LSGs from SS donors.* These gene sets from the MSigDB were selected as candidates for CpG enrichment because they contained a significantly high fraction of differentially methylated promoters, listed here. Bonferroni-adjusted P-values are reported for hypergeometric CpG set enrichment tests.

JASPAR ID	Annotated TF complex	Targets	FET P (Adjusted)
MA0089.1	TCF11/MAFG heterodimer	Antioxidant response elements	5.2×10^{-5}
MA0517.1	STAT2/STAT1 heterodimer	IFN-stimulated response elements	7.5×10^{-4}
MA0080.3	PU.1	PU box	5.9×10^{-3}

Table 8: *DMP-associated motifs identified by AME.* P-values were computed from FET, with Bonferroni-adjustment for multiple testing.

3

SINGLE-CELL TRANSCRIPTOMICS

3.1 INTRODUCTION

Substantial work has been done to catalog the cell types, states, and interactions that inform immune behaviors [65–71]. More recent studies, empowered by state-of-the-art technologies, show that seemingly identical cell populations can exhibit functionally relevant heterogeneities [65, 72–76]. As illustrated in Chapter 2, tissue-level disease signatures derived from bulk measurements are often understood in terms of shifts in cell type composition (e. g. B cells infiltrating LSGs); these effects obscure cell type specific phenomena and impede efforts to answer critical biological questions surrounding the immune system. The degree of cellular diversity involved in immune system processes demands higher resolution measurements.

Single-cell technologies provide new opportunities for characterizing systems-level immune responses. In this chapter I will discuss a single-cell technology popularized in recent years: single-cell RNA-seq (scRNA-seq). I will begin by reviewing the RNA sequencing (RNA-seq) foundation of the technology and critical modifications made to enable single-cell transcriptome profiling, followed by an illustration of scRNA-seq analysis challenges and discussion of methods for addressing them (e. g. *data normalization*).

3.2 RNA-SEQ

Unlike hybridization-based transcriptome quantification approaches (e. g. microarray), RNA-seq technologies measure the underlying sequence of cellular RNA. This approach provides sequence-level sampling of transcripts with relatively low levels of bias [77]. Full-length transcript sequencing methods can provide important information about splicing, supporting the discovery of novel isoforms.

This chapter focuses on technologies that target long messenger RNA (mRNA) species, but RNA sequencing technologies can measure many other types of RNA species. The first step of these protocols involves RNA isolation and purification of the polyadenylation (poly(A))-positive fraction to avoid highly-abundant ribosomal RNAs. These RNAs are typically reverse transcribed into complementary DNAs (cDNAs) using a modified reverse transcriptase (RT) enzyme, followed by enzymatic or mechanical fragmentation. During this process, the fragmented cDNAs are attached to polymerase chain reaction (PCR) adaptors, facilitating their amplification and subsequent high throughput sequencing.

Short sequences measured on the sequencer are computationally aligned to a reference genome or transcriptome (e. g. RSEM [78]), supporting many kinds of downstream analyses (e. g. variant calling, isoform analysis, or expression quantification). Expression quantification typically relies on gene-level summaries reflecting the number of reads in a library aligning to a gene. The number of aligned reads provides a quantitative measurement proportional to the number of gene transcripts in the original sample. Comparison of read counts between samples can provide insights into subtle transcription differences across thousands of molecular species. Importantly, the unbiased nature of these platforms guarantees greater generalizability: disease signatures quantified by these methods can be compared to future unbiased transcriptome measurements across very different biological conditions. Cross-study comparisons like these increase impact and discovery potential.

Bulk RNA-seq measurements provide only an imperfect picture of the RNA content of a tissue because they do not distinguish between constituent cells. Notably, a bulk profile can not be thought of as simple average of cell transcriptomes: cells with greater RNA content contribute more to the pool than smaller cells.

3.3 SINGLE-CELL TECHNOLOGIES

More recently, these technologies have been modified to measure RNA at the cell-level. scRNA-seq [66, 73, 74, 79, 80] technologies tie transcriptomic observations to individual cellular players, facilitating important modeling of heterogeneity and cell-cell interactions. As a result, scRNA-seq affords a direct means of identifying and comprehensively characterizing functionally important subsets of cells and their complex underlying biology.

3.3.1 SINGLE-CELL ISOLATION

Some of the earliest methods for scRNA-seq were manual and low-throughput, limited to 96 cells per batch (e. g. Martin-Gayo et al. [81]). Sufficient sampling of cell types may require a large number of these small batches, especially when investigators wish to accommodate rare cell types without sacrificing unbiased sampling. Lack of automation and intra-batch heterogeneity introduces many varied forms of technical noise and bias. In some cases the low throughput demands a significant tissue cost, requiring large amounts of tissue for a relatively small numbers of cells.

Microfluidic technology serves an important role in automating critical scRNA-seq protocol steps, giving investigators a means to uniformly lyse, process and amplify single-cell transcriptomes in preparation for subsequent sequencing [82, 83]. Early approaches integrating these technologies exhibited improved reproducibility but do not address limits on the cell number.

Early on in the development of these technologies it became clear that low sequencing coverage is sufficient for resolving biologically meaningful cell types [82, 84]. This revelation motivated technological approaches that could significantly increase the number of cells at the cost of sequencing coverage. Cellular barcoding techniques allowed technologies such as the 10x Chromium platform [85] to perform pooled reaction steps, greatly increasing the number of cells that could be extracted and profiled from one sample.

3.3.2 LEVERAGING UMIS

There are several biases associated with read-based expression summaries that complicate their interpretation as expression levels.

First, the number of reads aligning to a transcript can depend on the length of the gene, particularly for full-length RNA-seq protocols. This may be redressed by methods that normalize read summaries for gene length – an example of *within-sample* normalization – but length normalization approaches are limited by transcriptome reference accuracy. Furthermore, protocol-specific coverage bias along the length of transcripts (e. g. preference for 3' or 5' ends) can change the effective length of a transcript.

Second, PCR amplification is known to be biased by sequence GC-content. This bias worsens with increasing number of PCR cycles, and other sources of noise may be amplified by PCR as well. These biases can be worse in scRNA-seq due to the small amounts of starting material.

Recently, unique molecular identifier (UMI) techniques have been developed to tag cDNAs with unique molecular sequences shared by all amplified reads associated with the original molecule. These identifiers allow read data to be collapsed down to UMI count summaries, counting the number of times a unique UMI is paired with any read aligning to a specific gene. UMI-based quantification removes most of the PCR read biases described above, and it has been implemented on several platforms, including Fluidigm C1 [86] and 10x Chromium [85].

3.3.3 MULTI-OMICS

The success of scRNA-seq technology has led to an exciting new area of development involving joint measurement of genome, epigenome, proteome or other omic data with transcriptome measurement [87]. These approaches are motivated by the same desire to measure biological covariances at the cell-level. Joint measurements will significantly improve our understanding of fundamental biology, including gene regulation.

3.4 CHALLENGES

As scRNA-seq evolves into a mainstream technology, many of the fundamental issues complicating interpretation of the first scRNA-seq data sets continue to affect new protocols and platforms. In this section I will highlight two of these general scRNA-seq challenges: i) missing data issues and ii) cell quality heterogeneity.

3.4.1 MISSING DATA IN RNA-SEQ

Due to small starting amounts of RNA, scRNA-seq is burdened by a large number of *dropout* events, in which an whole class of RNA (e. g. gene product) is not detected after sequencing despite being expressed in the cell [88–90]. In many cases, dropped transcripts have been skipped by RT, degraded non-specifically, or simply missed by sequencing: dropouts of this kind are natural consequences of low mean sampling. When degradation or transcript capture differs across transcripts, these biases fall into a category similar to the length and GC-content PCR biases described above, modulating the effective sampling mean.

In some single-cell data sets (e. g. SMART-seq read counts) the number of zeros far exceeds predictions from standard low-mean count distributions [90, 91]. This missing data problem is referred to as zero inflation (ZI) because missing data cannot be easily distinguished from "legitimate" zero values. ZI can occur for sampling reasons: a single cDNAs in the amplified component can produce many reads while a transcript in the non-captured component will surely have no reads. ZI may also reflect important biological heterogeneity: i. e. the sample mixture includes a subpopulation of non-expressing cells. Distinguishing these two contributions and adjusting data to account for dropouts falls under the topic of *data imputation.* Data imputation in scRNA-seq is still in its infancy; only a handful of methods exist [92, 93] and it is unclear whether their promised advantages outweigh their limitations.

Beyond data imputation, there are other approaches that can account for ZI. In many statistical analyses there is a natural place for *data weights*, in this case reflecting confidence that a zero measurement comes from an amplified component. These weights can be computed within the context of ZI mixture models (e. g. Zero-inflated Negative Binomial based Wanted Variation Extraction (`ZINB-WaVE`) [94] and scVI (SCVI) [95]) and may be used in the context of various dimensionality reduction techniques.

In this subsection I will develop a simple model for partitioning zeros, describe inference, and define data weights based on model outputs.

FNR MODELING WITH HOUSEKEEPING GENES. The probability of dropout is mainly a function of transcript abundance at cell lysis; empirical observations suggest that dropout rates and log abundance typically follow a logistic relationship [88, 90]. When determining whether zeros are technical, one may employ *housekeeping genes* which are expected to be expressed both highly and uniformly across cells. Identification of housekeeping genes is not always easy, especially in the single-cell study contexts where single-cell expression heterogeneity is often the primary research motivation. However difficult they may be to define, these genes can serve as powerful negative controls for aspects of model estimation: model parameters fit using housekeeping genes may be extended to non-housekeeping genes in which assumptions of homogeneity do not hold.

Let y_{ij} denote the `log1p` transformed read count of gene $i \in \{1, \ldots, m\}$ in cell $j \in \{1, \ldots, n\}$ and let $\mathcal{I} \subseteq \{1, \ldots, m\}$ denote a set of m_0 housekeeping genes. The housekeeping assumption for these genes allow us to treat

zeros as technical dropouts rather than biological ZI, hence the interpretation of $\Pr(y_{ij} \leq \epsilon | \bar{y}_{i\cdot})$ as a false negative rate (FNR). Here, $\epsilon \geq 0$ is a detection threshold commonly set to 0. For each cell j, we may fit the following *logistic regression* model to housekeeping genes only:

$$\operatorname{logit} \mathrm{E}[\tilde{y}_{ij} | \bar{y}_{i\cdot}] = \beta_{j0} + \beta_{j1} \bar{y}_{i\cdot}, \qquad i \in \mathcal{I}_0, \tag{5}$$

where $\tilde{y}_{ij} \equiv \mathrm{I}(y_{ij} > \epsilon)$ is an indicator variable equal to 1 if gene i is detected and zero otherwise, $\bar{y}_{i\cdot}$ is the median log-read count of gene i across all non zero cells, and $\beta_j = (\beta_{j0}, \beta_{j1}) \in \mathbb{R}^2$ are cell-specific regression parameters. This model relates the detection rate $\Pr(y_{ij} > \epsilon | \bar{y}_{i\cdot})$ of gene i in cell j to the baseline expression measure $\bar{y}_{i\cdot}$ of the gene in a cell-specific manner (i. e. cell-specific regression parameters β_j).

Equation (5) yields FNR curves for each cell, where each point on the curve corresponds to a gene's dropout rate $\Pr(y_{ij} \leq \epsilon | \bar{y}_{i\cdot})$ at a baseline expression $\bar{y}_{i\cdot}$:

$$\mathrm{FNR}_j(\bar{y}_{i\cdot}) = 1 - \operatorname{expit}\left(\beta_{j0} + \beta_{j1} \bar{y}_{i\cdot}\right). \tag{6}$$

For the purpose of QC, it may be informative to examine these characteristics as they might reveal problematic cells. For instance, cells with an unusually high proportion of zero counts for highly-expressed genes might be uniquely affected by technical bias. Generally, the higher the FNR_j curve – the lower the quality of the cell. Cells may be compared based on the area under the curve (AUC) of their respective FNR curves (FNR AUC).

FULL BIOLOGICAL ZI MODEL. In the interest of modeling zeros across all cells, we consider Z, a hidden binary matrix of the same dimension as Y encoding the binary expression state of a gene in a cell. θ_i is a modeled expression rate (i. e. % cells that express the gene i),

$$\mathrm{E}\left[\mathrm{I}\left(z_{ij} = 1\right)\right] = \theta_i \tag{7}$$

EM INFERENCE. The generalized linear model (GLM) fitting described above yielded FNR_j in Equation 6. These fitted functions may be used to describe the conditional detection rate in non-housekeeping genes:

$$\begin{aligned} &\Pr(\tilde{y}_{ij} | z_{ij} = 1) = 1 - \mathrm{FNR}_j\left(\bar{y}_{i\cdot}\right), \\ &\Pr(\tilde{y}_{ij} | z_{ij} = 0) = 0. \end{aligned} \tag{8}$$

Using these equations we can estimate Z and θ_i with the expectation–maximization (EM) algorithm.

As all observations are conditionally independent, the likelihood can be written as a product:

$$\mathcal{L}(\theta|\tilde{Y}, Z) = \prod_{ij} \Pr(\tilde{y}_{ij}|z_{ij}) \Pr(z_{ij}|\theta_i) \tag{9}$$

The conditional distribution of the missing value matrix Z, given the observed detection matrix $\tilde{Y}$ and expression rates θ_i, is separable due to conditional independence of Z elements given gene-level expression rates θ_i.

$$\Pr(Z|\tilde{Y}, \theta) = \prod_{ij} \Pr\left(z_{ij}|\tilde{y}_{ij}, \theta_i\right) \tag{10}$$

Each term in this product can be evaluated using Bayes' theorem:

$$\Pr(z_{ij} = 1|\tilde{y}_{ij}, \theta_i) = \tilde{y}_{ij} + (1 - \tilde{y}_{ij}) \operatorname{expit}\left(\log\left(\frac{\mathrm{FNR}_j(\bar{Y}_{i\cdot})}{1/\theta_i - 1}\right)\right). \tag{11}$$

We can see that as FNR_j increases, the probability of an undetected gene's hidden expression rises.

EXPECTATION STEP. This step computes the expected value of the log likelihood function with respect to the current estimate of parameters θ_i. By the definition of probability, the expected value can be expressed as a simple sum, with each pair of indices i, j contributing one term:

$$Q(\theta', \theta) = \sum_{ij} \sum_{z_{ij}} \log \Pr(\tilde{y}_{ij}, z_{ij}|\theta_i') \Pr\left(z_{ij}|\tilde{y}_{ij}, \theta_i\right). \tag{12}$$

It's helpful to adopt notation for these individual terms Q_{ij}:

$$\begin{aligned} Q_{ij}(\theta_i', \theta_i) = {} & \log\left(\Pr\left(\tilde{y}_{ij}|z_{ij} = 1\right)\theta_i'\right) \Pr\left(z_{ij} = 1|\tilde{y}_{ij}, \theta_i\right) \\ & + \log\left(\Pr\left(\tilde{y}_{ij}|z_{ij} = 0\right)\left(1 - \theta_i'\right)\right) \Pr\left(z_{ij} = 0|\tilde{y}_{ij}, \theta_i\right). \end{aligned} \tag{13}$$

MAXIMIZATION STEP. This step involves finding parameter values that maximize the sum of Q_{ij}, by computing the first derivatives with respect to each θ_i':

$$\partial_{\theta_i'} Q(\theta', \theta) = \sum_j \frac{\Pr\left(z_{ij} = 1|\tilde{y}_{ij}, \theta_i\right)}{\theta_i'} - \frac{\Pr\left(z_{ij} = 0|\tilde{y}_{ij}, \theta_i\right)}{1 - \theta_i'}. \tag{14}$$

Equating each of these derivatives to zero yields updated estimates for θ_i:

$$\hat{\theta}_i' = \frac{1}{n} \sum_j \Pr\left(z_{ij} = 1 | \tilde{y}_{ij}, \theta_i\right). \tag{15}$$

When all cells have the same characteristic FNR_j, this expression can be solved ($\hat{\theta}_i' = \theta_i$) without iteration.

DATA WEIGHTS The resulting posterior Z probabilities (the expected z_{ij} matrix) may be used as a weight matrix w_{ij}, capturing the posterior probability that gene i did not dropout in cell j, but was, rather, unexpressed.

3.4.2 CELL QUALITY HETEROGENEITY

Single-cell RNA-seq data can exhibit strong, transcriptome-wide nuisance effects (e. g. batch), comparable in magnitude to the biological effects of interest [96]. Uneven sample quality, e. g. in terms of alignment rates and nucleotide composition [97] can also induce significant within-batch technical heterogeneity, obscuring biological signals.

In this section I illustrate these difficulties using a published SMART-seq data set, processed on Fluidigm C1 [83] (Appendix A). I will consider the subset of 337 mouse $\text{T}_\text{h}17$ T cells harvested after *in vitro* differentiation of CD4^+ naive T cells under 48 hours of pathogenic (IL-1β+IL-6+IL-23) or non-pathogenic (TGF-β1+IL-6) conditioning. Prior to conditioning, these cells had been extracted from two strains of mice: wild-type B6 and transgenic B6 mice with an IL-17A green fluorescent protein (GFP) reporter.

I aligned the publicly available reads to the mouse genome, counting over RefSeq gene intervals to generate a count matrix. For assessing scRNA-seq quality at the cell-level, I rely on over a dozen QC metrics evaluated by software packages such as `FastQC` [98], `Picard` [99], and `Cell Ranger` [100]. These QC measures summarize various aspects of genome alignment and nucleotide composition (Table 9).

Mouse-specific effects are prominently featured in a PCA of the `log1p` transformed data matrix (m =7,590 genes; Figure 7a). In the space defined by the first two PCs, the distances between cells from mouse 7 and mouse 8 are larger than the distances between pathogenic and non-pathogenic cells collected from the same mouse. Due to the partially confounded design, these mouse effects may result from multiple sources, including i) true biological differences between mice or ii) mouse-specific technical biases. The study

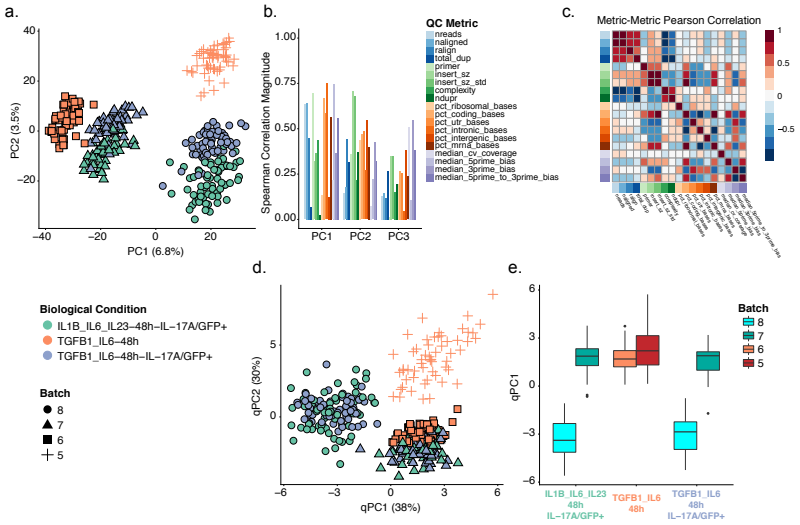


Figure 7: *Exploratory data analysis of mouse T_H17 data set [83].* **(a)** PCA of the log-transformed, TC-normalized read count data. Cells are color-coded by biological condition; shape represents the donor mouse (batch). For two of the three conditions, samples were extracted from only one mouse (IL-1β_IL-6_IL-23-48h-IL-17A/GFP$^+$ and TGF-β1_IL-6-48h-IL-17A/GFP$^+$ from mice 7 and 8, respectively), while samples from the third condition (TGF-β1_IL-6-48h) came from two distinct mice (mice 5 and 6). Cells cluster by both biological condition and batch, the latter representing unwanted variation. **(b)** Absolute r_s coefficient between the first three PCs of the expression measures (as computed in (a)) and a set of QC measures (Table 9). **(c)** Heatmap of pairwise Pearson correlation coefficients between QC measures. **(d)** PCA of the QC measures for all cells in (a). PCs of QC measures are labeled "qPCs" to distinguish them from expression PCs. Single-cell QC profiles cluster by batch, representing important aspects of batch covariation. **(e)** Boxplot of the first qPC, stratified by both biological condition and batch. Note that there are different numbers of cells in each stratum.

Name	Description	Source
NREADS	Total number of sequenced reads	Picard
NALIGNED	Total number of aligned reads	Picard
RALIGN	Percentage of mapped reads	Picard
TOTAL_DUP	Number of duplicate reads	FastQC
PRIMER	Percentage of primer sequence reads	FastQC
INSERT_SZ	Average insert size	Picard
INSERT_SZ_STD	Insert size variance	Picard
COMPLEXITY	Sequence Complexity	Picard
NDUPR	Percentage of unique reads	Picard
PCT_RIBOSOMAL_BASES	Percentage of ribosomal bases	Picard
PCT_CODING_BASES	Percentage of coding bases	Picard
PCT_UTR_BASES	Percentage of UTR bases	Picard
PCT_INTRONIC_BASES	Percentage of intronic bases	Picard
PCT_INTERGENIC_BASES	Percentage of intergenic bases	Picard
PCT_MRNA_BASES	Percentage of mRNA bases	Picard
MEDIAN_CV_COVERAGE	Median coefficient of variation of coverage	Picard
MEDIAN_5PRIME_BIAS	Mean 5' coverage bias	Picard
MEDIAN_3PRIME_BIAS	Mean 3' coverage bias	Picard
MEDIAN_5PRIME_TO_3PRIME_BIAS	Mean 5' to 3' coverage bias	Picard

Table 9: *Cell-level QC measures for full-length scRNA-seq protocols.*

design prevents us from teasing apart these two effects, but we can account for technical contributions to the read counts by examining the association of the expression PCs with RNA-seq library QC metrics (Fig. 7b).

The first three expression PCs exhibit large correlations with measures of genomic alignment rate, primer contamination, intronic alignment rate, and 5' bias. The correlation structure between these QC measures reflects constraints on library quality in the study (Fig. 7c). While some of these pairwise associations represent natural dependencies between similar QC measures (e. g. total number of reads and total number of aligned reads are positively correlated), others reflect mouse-specific technical biases in the study. Applying PCA to the matrix of QC measures, we can see how these metrics provide a candidate basis for representing batch (i. e. mouse) effects (Fig. 7d). Inter-batch QC differences are relatively large as in Figure 7a, while intra-batch differences between pathogenic and non-pathogenic cells are noticeably smaller (Fig. 7e). Mouse 6 libraries are technically similar to mouse 7 libraries, while cells from mice 5 and 8 are technically distinct; these relationships are similar to those observed in the PCA of gene expression measures, suggesting

that the corresponding structure in the expression data is artifactual. It is surely possible that some of the observed associations between read counts and QC measures result from biological confounding rather than direct technical bias, i. e. a cell's biological state may impact transcriptome integrity and sequencing viability [101, 102]. However, unlike factors such as mouse-of-origin, there exist simple interpretations for correlations between quantified expression measures and library alignment statistics.

3.5 DATA NORMALIZATION

Normalization is a common preprocessing step in the analysis of omic data, such as high-throughput transcriptome microarray and sequencing data. The goal of normalization is to account for observed differences in measurements between observations (e. g. cells) and/or features (e. g. genes) resulting from technical artifacts or unwanted biological effects (e. g. batch effects) rather than biological effects of interest. Accordingly, two types of normalization are often considered: within-sample normalization [103], which adjusts for gene-specific (and possibly sample-specific) effects, e. g. related to gene length and GC-content, and between-sample normalization, which adjusts for effects related to distributional differences in read counts between samples, e. g. sequencing depth, C1 run, library preparation. This section focuses on the former. In order to derive gene expression measures from scRNA-seq data and subsequently compare these measures between cells, analysts must normalize read counts (or other expression measures) to adjust for obvious differences in sequencing depths. When there are other significant biases in expression quantification, it may be necessary to further adjust expression measures for more complex unwanted technical factors related to sample and library preparation.

As previously discussed [101, 104], normalization of scRNA-seq data is often accomplished via methods developed for bulk RNA-seq or even microarray data. These methods tend to neglect prominent features of scRNA-seq data. In particular, widely-used global-scaling methods, such as reads per million (RPM) [105], trimmed mean of M values (TMM) [106], and relative log expression (RLE) normalization in `DESeq` [107], are not well suited to handle large or complex batch effects and may be biased by low counts and zero inflation [101]. Other more flexible methods, such as remove unwanted variation (RUV) [108, 109] and surrogate variable analysis (SVA) [110, 111],

depend on tuning parameters (e. g. the number of unknown factors of unwanted variation).

A handful of normalization methods specifically designed for scRNA-seq data have also been proposed. These include scaling methods [112, 113], regression-based methods for known nuisance factors [114, 115], and methods that rely on spike-in sequences from the External RNA Controls Consortium (ERCC)[116, 117]. While these methods address some of the problems affecting bulk normalization methods, each suffers from limitations with respect to their applicability across diverse study designs and experimental protocols. Global-scaling methods define a single normalization factor per cell and thus are unable to account for complex batch effects. Explicit regression on known nuisance factors (e. g. batch, number of reads in a library) may miss unknown, yet unwanted variation, which may still confound the data [109]. Unsupervised normalization methods that regress gene expression measures on unknown unwanted factors may perform poorly with default parameters (e. g. number of factors adjusted for) and require tuning, while ERCC control-based methods suffer from differences between endogenous and spiked-in transcripts [101, 109]. Protocols using UMI still require normalization; while UMIs remove amplification biases, they are often sensitive to sequencing depth and differences in capture efficiency before reverse transcription [101].

3.5.1 NORMALIZATION PROCEDURES

Most between-sample normalization methods proposed to date are adaptations of methods for bulk RNA-seq and microarrays and range, as described below, from simple global scaling to regression on gene- and cell-level covariates.

3.5.2 GLOBAL-SCALING NORMALIZATION

Only a fraction of a cell's RNA content can be captured and sequenced by scRNA-seq technologies. Meaningful comparisons are usually made by inspecting compositional differences in RNA content between cells, rather than absolute molecule counts; the total RNA content or number of molecules captured in a cell are often treated as an uninformative *cell size* factors. For technologies without UMIs, an additional *library size* factor – the total number of reads sequenced in a cell – adds technical variance to the measurement.

For full-length methods, one may consider the model of Robinson and Oshlack [106]:

$$\begin{aligned} \mathrm{E}[r_{ij}] &= \frac{L_i \mu_{ij}}{S_j} N_j \\ S_j &= \sum_i L_i \mu_{ij} \end{aligned} \tag{16}$$

Where r_{ij} is the read count matrix, μ_{ij} is the true gene expression of the cell, L_i is the length of gene i, and N_j is a size factor quantifying the total number of reads in cell j. S_j is an unknown *cell size* factor describing the true RNA content of the cell.

Depending on the application, researchers may be interested in estimating either μ_{ij} or $\alpha_{ij} = \frac{\mu_{ij}}{\sum_i \mu_{ij}}$ from the data. Linear *global-scaling normalization* procedures are commonly used for this purpose, scaling gene-level read counts by a single factor per cell.

TOTAL COUNT (TC). The scaling factor is the sum of the read counts across all genes, as in the widely-used reads per million (RPM), counts per million (CPM), and reads per kilobase of exon model per million mapped reads (RPKM) [105].

UPPER QUARTILE (UQ). The scaling factor is the upper-quartile (upper quartile (UQ)) of the gene-level count distribution, [118].

TRIMMED MEAN OF M VALUES (TMM). The scaling factor is based on a robust estimate of the overall expression fold change between the sample and a reference sample [106]. TMM is implemented in the Bioconductor R package `edgeR` [119]. The default behavior of this implementation is to select a reference sample that has an upper quartile closest to the mean upper quartile of all samples.

RELATIVE LOG-EXPRESSION (`DESEQ`). The scaling factor for a given sample is defined as the median fold change between that sample and a synthetic reference sample whose counts are defined as the geometric means of the counts across samples [107]. The method is implemented in the Bioconductor R packages `DESeq` and `edgeR` (as "RLE") [119, 120]. Note that the method discards any gene having zero count in at least one sample; as zeros are common in single-cell data, the scaling factors are often based on only a handful of genes.

POOL-BASED SCALING (SCRAN). To reduce the effect of single-cell noise on normalization, the scaling factors are computed on pooled expression measures and then deconvolved to obtain cell-specific factors [112]. The method is implemented in the Bioconductor R package `scran` [121]. Optionally, cells can be clustered prior to normalization to relax the assumption that the majority of genes are not differentially expressed across groups of cells.

3.5.3 NON-LINEAR SCALING NORMALIZATION

In some cases, a single scaling factor per sample may not be sufficient to capture the non-linear effects that affect gene expression measures. Hence, some authors have proposed non-linear normalization methods. Although, these are not technically "scaling" methods, they are aimed at making the between-sample distributions of expression measures more similar, rather than explicitly correcting for batch or other confounding factors.

FULL QUANTILE (FQ). All quantiles of the read count distributions are matched between samples [118]. Specifically, for each sample, the distribution of sorted read counts is matched to a reference distribution defined in terms of a function of the sorted counts (e. g. median) across samples. This approach, inspired by the microarray literature [122], is implemented in the Bioconductor R package `EDASeq` [123].

QUANTILE REGRESSION (SCNORM). Bacher et al. [115] noted that a single scaling factor per sample is not enough to account for the systematic variation in the relationship between gene-specific expression measures and sequencing depth. To address this problem, they use quantile regression to estimate the dependence of gene expression measures on sequencing depth, group genes with similar dependence, and use a second quantile regression to estimate scaling factors within each group. In this way, gene expression measures are normalized differently across the range of expression (i. e. highly-expressed genes are scaled differently than lowly-expressed genes). The method is implemented in the Bioconductor R package `SCnorm` [124]).

3.5.4 REGRESSION-BASED NORMALIZATION

Consider the following GLM which allows adjustment for known and unknown factors of *unwanted variation*:

$$g(\mathrm{E}[R|X,U,W]) = \beta X + \gamma U + \alpha W, \tag{17}$$

where R is the $m \times n$ matrix of gene-level read counts, X is an $M \times n$ design matrix corresponding to the M covariates of interest/factors of *wanted variation* (e. g. treatment) and β its associated $m \times M$ matrix of parameters of interest, U is an $H \times n$ matrix corresponding to known factors of unwanted variation (e. g. batch, sample QC measures) and γ its associated $m \times H$ matrix of nuisance parameters, W is an $K \times n$ matrix corresponding to unknown factors of unwanted variation and α its associated $m \times K$ matrix of nuisance parameters, and g is a link function, such as the logarithm in Poisson/log-linear regression.

The γU and αW terms correspond, respectively, to supervised and unsupervised removal of unwanted variation. A fully supervised version of Equation (17), without W, reduces to a simple GLM fit. The RUV model of Risso et al. [109] arises as a special case of Equation (17), when one omits the known unwanted factors U. In many cases, the data-driven unsupervised version of Equation (17), without U, captures effects related to U; for instance, W is often associated with QC measures. However, in many cases, U should still include known batches (e. g. set of samples processed at the same time), as W could capture effects related to sample quality within batches that are important to remove in addition to the batch effects captured by U. In practice, the computationally simpler approach of fitting a linear model to log-transformed counts, Y, yields good results, relying on a linear model version of Equation (17) with identity link function ($g(x) = x$).

As detailed in Risso et al. [109] and implemented in the Bioconductor R package `RUVSeq` [125], the unknown, unwanted factors W can be estimated by singular value decomposition (SVD) using several main approaches, e. g. RUVg estimates the factors of unwanted variation based on *negative control genes*, assumed to have constant expression across all samples ($\beta = 0$).

3.5.5 ADJUSTMENT FOR NESTED EXPERIMENTAL DESIGNS

The model of Equation (17) should be applied with caution and only after a careful examination of the experimental design. In particular, a common

limitation of scRNA-seq data sets is the *nesting* of unwanted technical effects within the biological effects of interest. For instance, the Induced Pluripotent Stem Cells (iPSC) data set of Tung et al. [86] contains samples derived from three individuals, each processed in three batches. Regressing read counts on the batch covariate in U without adjusting for the covariate of interest in X would then remove the effect of interest (effect of individual donor). Additionally, to avoid collinearity issues between columns of U and X due to nesting, one could either specify suitable contrasts or use a mixed effect model where technical effects are viewed as random.

Specifically, to account for nested batch effects, consider the following model for each gene. For illustration purposes, I will not include the additional known or unknown covariates allowed in Equation (17). Let y_{ij}^{cb} denote the transformed expression measure of sample j in condition c and batch b, with $c = 1, \ldots, a$ conditions, $b = 1, \ldots, b_c$ batches for condition c, and $j = 1, \ldots, n_{cb}$ samples for batch b of condition c. Consider a fit for the following model:

$$\mathrm{E}[y_{ij}^{cb}|X, U] = \alpha_i + \beta_i^c + \gamma_i^{cb}, \tag{18}$$

where α_i are gene-specific intercepts, β_i^c are biological effects of interest, and γ_i^{cb} are nested batch effects. Given the $a + 1$ constraints for each i:

$$\begin{aligned} \sum_{c=1}^{a} \beta_i^c &= 0, \\ \sum_{b=1}^{b_c} \gamma_i^{cb} &= 0, \qquad c = 1, \ldots, a, \end{aligned} \tag{19}$$

the model is identifiable and can be fit using standard R functions such as `lm`. The batch-corrected gene expression measures are given by the residuals $y_{ij}^{cb} - \hat{\gamma}_i^{cb}$. When additional factors of unwanted variation are included in the model, their effects are similarly subtracted from the original matrix to produce the normalized matrix.

Note that no adjustment method is able to remove batch effects while preserving biological effects of interest if the experimental design is completely confounded. For instance, if only one batch had been processed per individual in the iPSC data set of Tung et al. [86], it would have been impossible to determine if the differences between cells were due to batch effects or biology [96].

3.6 CONCLUSIONS

In this chapter I have introduced scRNA-seq technologies and highlighted some of the unique and inherited challenges of the RNA-seq based measurement, including dropout events, ZI, sample heterogeneity and batch effects. I have described some methods that can be used to mitigate these difficulties, including ZI modeling and data normalization. In particular, I discussed many data normalization solutions without including a detailed discussion of the trade-offs between different approaches. The problem of normalization assessment will be discussed in the next chapter.

PART II

COMPUTATIONAL TOOLS

4

SCONE

4.1 INTRODUCTION

Chapter 3 discusses how systematic measurement biases make data normalization an essential preprocessing step in scRNA-seq analysis. As illustrated there in Figure 7, simple global scaling alone is insufficient for normalizing scRNA-seq data and more flexible and aggressive procedures aimed at removing unwanted variation (UV) (e. g. batch effects) may be generally beneficial. This example was drawn from one study [83], but the observations are not unique to this data set and are indeed general features of scRNA-seq. To highlight this, I have performed a similar exploratory data analysis (EDA) on a set of developing human cortical neurons assayed using a 2014 Fluidigm protocol [82] and a set of peripheral blood mononuclear cells (PBMCs) assayed using the 10x Chromium platform [85] (Figures 8 and 9). Basic processing details for these public data sets and others analyzed in this chapter can be found in Appendix A.

From these examples it appears that there is considerable room for normalization to improve data quality and downstream results – e. g. clustering and differential expression (DE). On the other hand, depending on the study design there may be varying, competing considerations behind the assessment of normalization performance. Due to the prevalence of confounding in single-cell experiments, the lack of a uniformly optimal normalization across data sets, and the ambiguity in tuning parameter guidelines for commonly-used normalization methods, it makes sense to try many method combinations, using data-driven metrics to guide the selection of suitable approaches.

Collaborators and I have developed the Single-Cell Overview of Normalized Expression data (`scone`)[1] framework for implementing and assessing

1 This chapter is based on a preprint paper hosted on bioRxiv: "Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-seq." [126] © The Authors 2018, reproduced with permission.

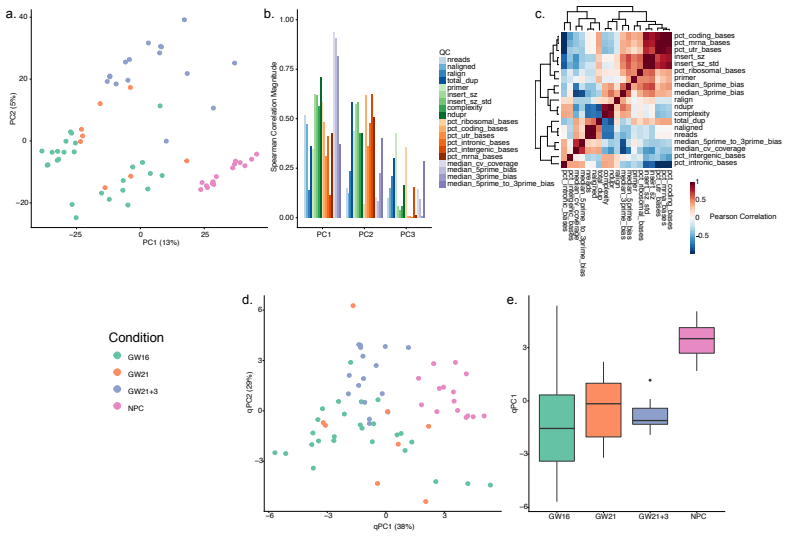


Figure 8: *Exploratory data analysis of human cortex cells from Pollen et al. [82].* **(a)** PCA of the log-transformed, TC-normalized read count data using all genes passing quality filtering (Subsection 4.2.2). Cells are color-coded by biological condition. Cells cluster partially by biological condition, with significant intra-condition heterogeneity. The design of this study is fully confounded (one batch per biological condition): batch adjustment is not advisable, as it would remove the biological effects of interest. **(b)** r_S coefficient magnitude between the first three PCs of the expression data (as computed in (a)) and a set of QC measures (Table 9). **(c)** Heatmap of pairwise Pearson correlation coefficients between QC measures. **(d)** PCA of the QC measures for all cells in (a). Single-cell QC profiles cluster by biological condition, suggestive of technical confounding. **(e)** Boxplot of the first qPC, stratified by biological condition. QC measures differ significantly between NPCs and other biological conditions / batches.

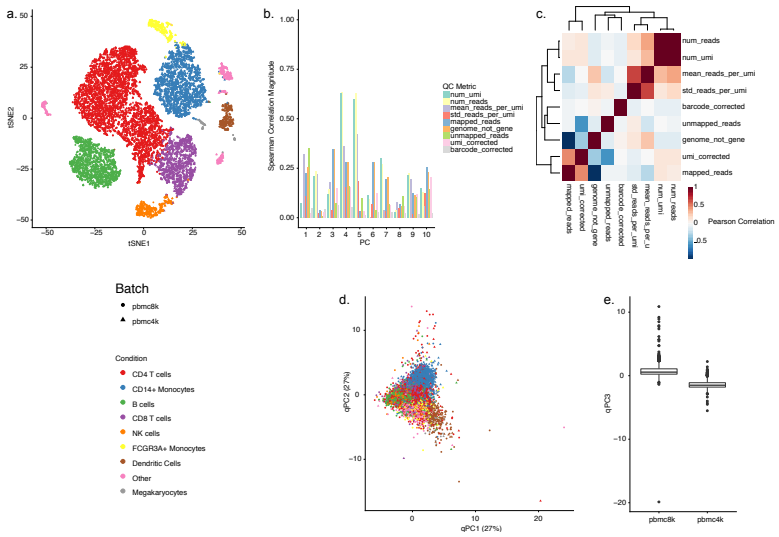


Figure 9: *Exploratory data analysis of PBMCs sequenced on the 10x Chromium platform [85].* **(a)** tSNE of the first 10 PCs of the log-transformed, TC-normalized UMI count data for all genes and cells passing quality filtering (Subsection 4.2.2). Cells are color-coded by a `Seurat`-based manual annotation of major PBMC subtypes (Appendix A); shape represents the 10x batch. cells from both batches ("pbmc4k" and a larger "pbmc8k") originated from the same healthy human donor. Cells clearly cluster by data-derived biological condition, one consequence of being clustered jointly in `Seurat`. **(b)** Absolute r_s coefficient magnitude between the first ten PCs of the expression data (as computed in (a)) and a set of QC measures (Table 10). **(c)** Heatmap of pairwise Pearson correlation coefficients between QC measures. **(d)** PCA of the QC measures for all cells in (a). Single-cell QC profiles partially cluster by data-derived biology (especially $CD14^+$ monocytes), with no clear clustering by batch. **(e)** boxplot of the third qPC, stratified by batch. The third qPC is the qPC with the highest correlation with batch.

the performance of a range of *normalization procedures*, each consisting of defined normalization steps, such as scaling and supervised or unsupervised regression-based adjustments. scone evaluates the performance of each procedure and ranks them by aggregating over a panel of performance metrics that consider different aspects of a desired normalization outcome, including removal of UV and preservation of wanted variation (WV). Through graphical summaries and quantitative reports, scone summarizes performance trade-offs and ranks large numbers of normalization methods by aggregate panel performance. The modularity of the open-source R [33] software package scone allows researchers to tune and compare a set of default normalizations as well as to include user-defined methods, providing a useful framework for both practitioners and method developers.

Below I demonstrate that the scone methodology is generally applicable to different scRNA-seq protocols and study designs and show that top-performing normalization methods lead to better agreement with independent validation data.

4.2 scone: AN EXPLORATORY FRAMEWORK FOR SCRNA-SEQ NORMALIZATION

The scone workflow consists of five steps: i) Quantifying QC; ii) data filtering; iii) normalization procedures; iv) normalization performance assessment; v) exploratory analysis of normalized data. In this section I will provide specific details for each of these steps.

4.2.1 QUANTIFYING QC

I have already touched on the utility of QC metrics in Chapter 3, using library-level metrics (Table 9) to probe confounding in scRNA-seq data. My results in Chapter 2 demonstrate how low-level QC metrics can also be used to adjust data and remove UV. scone uses QC metrics for both purposes, and thus an important first step of the framework is the extraction or definition of cell-level measures. For the analysis of 10x in this chapter I have considered an alternative set of metrics (Table 10) tailored to the 10x platform, quantifying aspects of UMI and barcode processing.

In some cases it may be difficult to obtain low-level (e. g. alignment-based) QC measures. This can occur if sequence data are filtered or omitted when

Name	Description	source
num_umi	Number of unique UMI sequences	Cell Ranger
num_reads	Total number of reads (regardless of mapping)	Cell Ranger
mean_reads_per_umi	The average number of reads supporting each UMI	Cell Ranger
std_reads_per_umi	Standard deviation of the number of reads supporting each UMI	Cell Ranger
mapped_reads	Proportion of reads which confidently mapped to a gene	Cell Ranger
genome_not_gene	Proportion of reads mapping to the genome, but not to a gene	Cell Ranger
unmapped_reads	Proportion of reads which did not align	Cell Ranger
umi_corrected	Proportion of reads whose UMI sequence was corrected by Cell Ranger	Cell Ranger
barcode_corrected	Proportion of reads whose barcode sequence was corrected by Cell Ranger	Cell Ranger

Table 10: *Cell-level QC measures for 10x Chromium.*

uploading data sets to a public-facing repository. QC measures may be derived from count matrices instead of alignments, using summary tools such as those provided in the `scater` package [127].

4.2.2 DATA FILTERING

The goal of data filtering is to remove problematic or noisy observations from downstream analysis. This can simplify many aspects of the analyses, including normalization. `scone` uses gene and cell filtering to combat dropouts and poor data quality. The canonical `scone` filtering step has three substeps, the latter two reducing the size of the data set:

1. Define *common genes* based on read counts: Genes with more than n_r reads in at least f_s of cells, where n_r is the upper-quantile of the non-zero elements of the count matrix and f_s is a user-specified percentage, with default value 25%.

2. Filter cells based on QC measures: Remove cells with low numbers of reads, low proportions of mapped reads, low numbers of detected common genes, or high the FNR AUC as defined in Chapter 3.

   Thresholds calculated by `scone::metric_sample_filter` are defined data-adaptively and for each metric: A cell may fail any criterion if the associated metric under-performs by z_{cut} standard deviations from the mean metric value or by z_{cut} median absolute deviations from the median metric value. For all cell-filtered data sets described in Appendix A I have set $z_{cut} = 2$. This substep was not applied to the Pollen et al. [82] data set due to the small number of cells.

3. Filter genes based on read counts: Remove genes with fewer than n_r reads in at least n_s cells, where n_r is the UQ of non-zero count matrix elements for the submatrix of cells passing the filtering described in the previous step. I have set a default of $n_s = 5$ to accommodate markers of rare populations.

   This substep ensures that included genes are detected in a sufficient number of cells after cell filtering. Due to the small number of negative control genes in the data set from Gaublomme et al. [83] (Appendix A), I had forced inclusion of all detected negative controls.

4.2.3 NORMALIZATION PROCEDURES

The two scone steps prior to normalization result in the selection of high-quality cell gene expression profiles. Following these initial steps, scone uses a two-part normalization template to define an ensemble of normalization procedures:

1. Scaling: scaling of the counts to account for between-library differences in sequencing depth and other parameters of the read count distributions: e. g. total count (TC) scaling or more robust scaling procedures designed to reduce the effect of outliers, like TMM [106] or RLE [107].

2. Adustment: regression-based adjustment for known unwanted factors, such as processing batches, and unknown unwanted factors. Confounding factors can be adjusted for by regressing scaled gene expression measures on quantities known to influence them (e. g. batches or "qPCs" like those shown in Figure 7). Alternatively, unsupervised procedures can estimate hidden unwanted factors and regress them out of the data (e. g. RUVg [109]).

4.2.4 NORMALIZATION PERFORMANCE ASSESSMENT

Different normalization procedures can lead to vastly different distributions of gene expression measures. In the context of bulk RNA-seq, the choice of normalization procedure had a greater impact on differential expression results than the choice of DE test statistic [118]. A natural and essential question is therefore whether normalization is beneficial and, if so, which method is

most appropriate for a given data set. In order to address this question, scone calculates a set of eight performance metrics, aimed at capturing different aspects of successfully normalized data. These metrics fall into three broad categories:

1. Clustering properties: clustering of cells according to factors of wanted and UV;
2. Association with control genes and QC metrics: association of expression PCs with factors of WV and UV;
3. Global distributional properties: between-cell distributional properties of the expression measures;

CLUSTERING PROPERTIES. The following three metrics evaluate normalization procedures based on how well the cells are grouped according to factors of WV and UV: Clustering by wanted factors is desirable, while clustering by unwanted factors is undesirable.

I use silhouette widths (SWs) [128] as clustering quality measures. For any clustering of n cells, the SW of cell i is defined as

$$sil(i) \equiv \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1], \tag{20}$$

where $a(i)$ denotes the average distance between the ith cell and all other cells in the cluster to which i was assigned and $b(i)$ denotes the minimum average distance between the ith cell and cells in other clusters. Intuitively, the larger the SWs, the better the clustering. Thus, the average silhouette width (ASW) across all n cells provides an overall quality measure for a given clustering. Here I considered Euclidean distance over first three PCs of expression measures.

- *BIO_SIL*: Group the n cells according to the value of a categorical covariate of interest (e. g. known cell type, genotype) and compute the ASW for the resulting clustering.
- *BATCH_SIL*: Group the n cells according to the value of a nuisance categorical covariate (e. g. batch) and compute the ASW for the resulting clustering.

- *PAM_SIL*: Cluster the n cells using Partitioning Around Medoids (PAM) for a range of user-supplied numbers of clusters and compute the maximum ASW for these clusterings. `scone` provides an option for stratified scoring, computing the PAM_SIL metric in all distinct strata defined jointly by biological and batch classification. The reported PAM_SIL metric is a weighted average across all strata, weighing by the total number of cells in each. This option is useful when prior biological classifications are poor proxies for cell types, i. e. when additional heterogeneity is expected. I used this option for all data sets listed in Appendix A but the 10x Genomics PBMC and cellular indexing of transcriptome and epitopes by sequencing (CITE-seq) cord blood mononuclear cells (CBMC) data sets for which the biological classification were data-derived clusters (Appendix A).

Large values of BIO_SIL and PAM_SIL and low values of BATCH_SIL are desirable.

ASSOCIATION WITH CONTROL GENES AND QC METRICS. The next three metrics concern the association of PCs of normalized log-counts with "evaluation" PCs of wanted or UV.

- *EXP_QC_COR*: The weighted *coefficient of determination* $\bar{R}^2$, defined below, for the regression of expression PCs on all PCs of scaled QC measures (qPCs).
- *EXP_UV_COR*: The weighted coefficient of determination $\bar{R}^2$ for the regression of expression PCs on factors of UV derived from negative control genes, preferably different from those used in RUV. The submatrix of log-transformed unnormalized counts for negative control genes is row-centered and scaled (i. e. for each row/gene, expression measures are transformed to have mean zero and variance one across columns/cells) and factors of UV are defined as the right-singular vectors as computed by the `rARPACK::svds` function.
- *EXP_WV_COR*: The weighted coefficient of determination $\bar{R}^2$ for the regression of log-count PCs on factors of WV derived from positive control genes. The WV factors are computed in the same way as the UV factors above, but with positive instead of negative control genes.

Large values of EXP_WV_COR and low values of EXP_QC_COR and EXP_UV_COR are desirable.

The weighted coefficients of determination are computed as follows. For each type of evaluation criterion (i. e. QC, UV, or WV), regress each expression PC on all supplied evaluation PCs (here, three). Let SST_k, SSR_k, and SSE_k denote, respectively, the total sum of squares, the regression sum of squares, and the residual sum of squares for the regression for the kth expression PC. The coefficient of determination is defined as usual as

$$R_k^2 \equiv \frac{SSR_k}{SST_k} = 1 - \frac{SSE_k}{SST_k},$$

and the weighted average coefficient of determination as

$$\bar{R}^2 \equiv \frac{\sum_k SST_k R_k^2}{\sum_k SST_k} = \frac{\sum_k SSR_k}{\sum_k SST_k} = 1 - \frac{\sum_k SSE_k}{\sum_k SST_k}. \tag{21}$$

GLOBAL DISTRIBUTIONAL PROPERTIES. When comparing distributions of expression measures between cells, gene-level RLE measures, defined as log-ratios of read counts to median read counts across cells, are more informative than log-counts [129]:

$$RLE_{ij} \equiv \log \frac{r_{ij}}{\text{Median}_j\, r_{ij}} = y_{ij} - \text{Median}_j\, y_{ij}, \tag{22}$$

for gene i in cell j. For similar distributions, the RLE distribution should be centered around zero and have have similar spread across cells.

- *RLE_MED*: Mean squared median RLE:

$$\frac{1}{n} \sum_j \left(\text{Median}_i\, RLE_{ij}\right)^2, \tag{23}$$

- *RLE_IQR*: Variance of interquartile range (IQR) of RLE:

$$\frac{1}{n} \sum_j \left(\text{IQR}_i RLE_{ij} - \frac{1}{n} \sum_{j'} \text{IQR}_{i'} RLE_{i'j'} \right)^2, \tag{24}$$

Low values of RLE_MED and RLE_IQR are desirable.

RANKING AND SELECTING NORMALIZATION PROCEDURES. Within the scone framework, expression measures are normalized according many method combinations and the eight metrics above are computed for each normalized

data set[2]. The performance assessment results can be visualized using biplots [130] and the normalization procedures ranked based on a function of the performance metrics. In particular, I define a *performance score* by orienting the metrics (multiplying by ±1) so that large values correspond to good performance, ranking procedures by each metric, and averaging the ranks across metrics.

Note that a careful, global interpretation of the metrics is recommended, as some metrics tend to favor certain methods over others, e. g. EXP_UV_COR naturally favors RUVg, especially when the same set of negative control genes are used for normalization and evaluation. I have used non-overlapping sets of control genes for all of the analyses discussed below.

Overall, these metrics capture the trade-offs between the ability of a normalization procedure to remove UV, preserve biological variation of interest, and maintain minimum global technical expression variability. These trade-offs are rooted in the confounding commonly encountered in single-cell assays; scone provides a reproducible basis for managing those trade-offs via normalization.

4.2.5 EXPLORATORY ANALYSIS OF NORMALIZED DATA

The multidimensional aspect of normalization performance is lost in a simple ranking of normalization procedures, even the two-dimensional biplot representation. Furthermore, the metrics do not necessarily capture all effects of normalization on the data. In order to address this I have developed the scone report browser (Figure 10) for inspecting effects on normalization performance.

4.3 scone REMOVES UV AND PRESERVES WV

Below I draw on evidence from multiple single-cell data sets (Appendix A), generated from various technological platforms, to show how no single normalization method is uniformly optimal: insights from the scone framework will highlight how performance depends on the design of the experiment and other characteristics of the data.

2 In order to avoid an evaluation dominated by zero value handling, I force to zero all values that are initially zeros as well as any negative values produced by a normalization procedure.

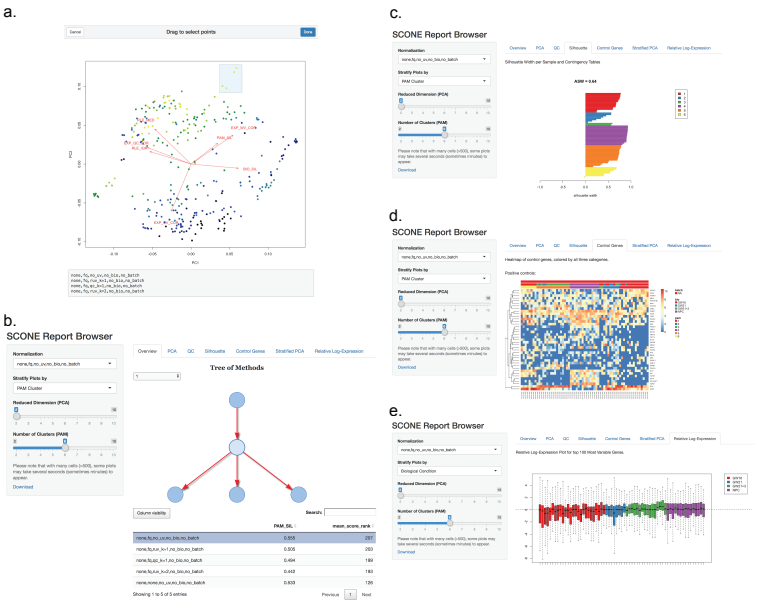


Figure 10: *Report Browser Shiny interface.* **(a)** Selecting normalization procedures of interest using the interactive biplot function `biplot_interactive` and its drag-and-drop window selection tool. This tool is useful for exploring performance clusters and selecting procedures that perform similarly across the eight performance metrics. **(b)** Browsing normalized products. The `scone` Report Browser presents an interactive tree representation (top-right panel) of selected procedures. Procedures may be further selected via a sortable performance table (bottom-right panel) or a drop-down menu (side panel). The report will then produce plots corresponding to various analyses of the normalized data. **(c)** Report Browser "Silhouette" tab: For the selected procedure, the SW of each normalized cell is computed, grouping cells by biological condition, batch, or PAM clustering. The drop-down menu in the left bar allows the user to switch between the three categorical labels; the slider in the left panel allows the user to select the number of clusters for PAM, recomputed for each normalization procedure. **(d)** Report Browser "Control Gene" tab: If the user provides positive and negative control genes, the gene-level expression measures for these genes are visualized using silhouette-sorted heatmaps, including annotations for biological condition, batch, and PAM clustering. **(e)** Report Browser "Relative log-Expression" tab: A boxplot of RLE measures is shown for the selected normalization procedure. Boxes (per-cell) are color-coded by biological condition, batch, or PAM clustering (drop-down selection in the left panel). If the majority of genes are not expected to be differentially expressed, the RLE distributions of the cells should be similar and centered around zero.

A useful representation of the normalization performance landscape is the *biplot* [130], mentioned above, in which each point corresponds to a normalization procedure and the dimensions of variation, represented by red arrows, correspond to scone performance metrics (Figure 11a-c). The scone biplot naturally represents trade-offs between these metrics, as illustrated in Figure 11 for three data sets. For the cortical neuron data set of Pollen et al. [82] (Figure 11a), there are two major bundles of red arrows, representing i) preservation of biological heterogeneity and ii) distributional uniformity irrespective of library quality. The existence of this trade-off suggests that WV is confounded by measurement artifacts. The top-ranked procedure according to scone involves full quantile (FQ) normalization followed by adjustment for 6 qPCs. Tracing a path from the performance coordinates of unnormalized data to those of the top-ranked normalization, notice that FQ without adjustment also performs very well according to scone, occupying a middle position between these two trade-offs. Both procedures may reasonably be carried to downstream analysis, but the scone biplot highlights an important tension in this decision.

The scone biplot for the SMART-seq data set of Gaublomme et al. [83] demonstrates a more complex "fan" of trade-offs between batch effect removal and preservation of WV (Figure 11b). Compared to no normalization, global-scaling and FQ normalization primarily improve distributional properties of the data, reducing the amount of global expression variability between cells (captured by the RLE metrics). Regression-based normalization, including batch regression and RUV, remove UV at the expense of WV; the biplot can help identifying those normalization that balance the trade-off between removing too much biological variation and too little technical variation. Unlike the biplot for Pollen et al. [82], the arrows corresponding to associations with QC metrics or negative control genes are closely aligned. However, the factors of UV computed from negative control genes (RUVg) and the QC measures are not always correlated (Figure 12), suggesting that regression normalizations based on these factors are complementary approaches.

I observe similar trade-offs between removal of technical variation and preservation of biological variation in the 10x Genomics data set (Figure 11c). Unique to this case is the relatively good performance of no normalization: most normalization procedures scored worse than doing nothing. Nevertheless, scone identifies a procedure that balances the trade-offs between the different metrics, involving the RLE scale factor and regression-based adjustment for all 8 PCs of the QC matrix.

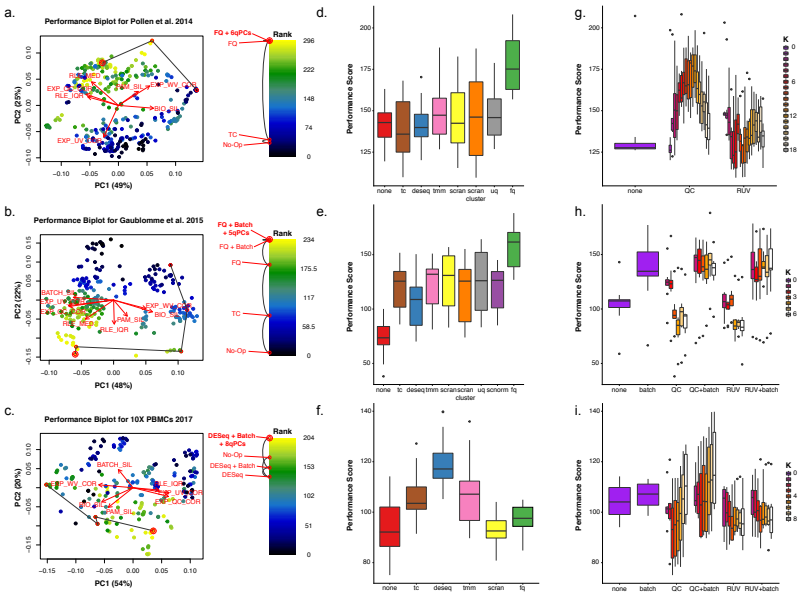


Figure 11: *Normalization performance assessment for three scRNA-seq data sets [82, 83, 85].* **(a-c)** Biplot [130] showing the first two PCs of eight rank-transformed scone performance metrics, or fewer if some are undefined or invariant: Preservation of biological clustering ("BIO_SIL"), batch effect removal ("BATCH_SIL"), cluster heterogeneity ("PAM_SIL"), preservation of association with positive control genes ("EXP_WV_COR"), removal of unwanted associations (negative control genes, "EXP_UV_COR", or cell-level QC measures, "EXP_WC_COR"), and global distributional uniformity ("RLE_MED" and "RLE_IQR"). Each point corresponds to a normalization procedure and is color-coded by the rank of the scone performance score (mean of eight scone performance metric ranks). The red arrows correspond to the PCA loadings for the eight performance metric ranks. The direction and length of a red arrow can be interpreted as a measure of how much each metric contributes to the first two PCs. Red circles mark the best normalization (w/ double circle), no normalization, and other normalization procedures relating the two. Key: "No-Op" = No normalization, "DESeq" = RLE scaling [107], "Batch" = Regression-based batch normalization, "*k*qPCs" = Regression-based adjustment for first *k* qPCs. **(d-f)** Boxplot of scone performance score, stratified by scaling normalization method, for the three scRNA-seq data sets presented in the same order as in (a-c). **(g-i)** Boxplot of scone performance score, stratified by regression-based normalization method (batch, QC, and RUV), for the three scRNA-seq data sets presented in the same order as in (a-c).

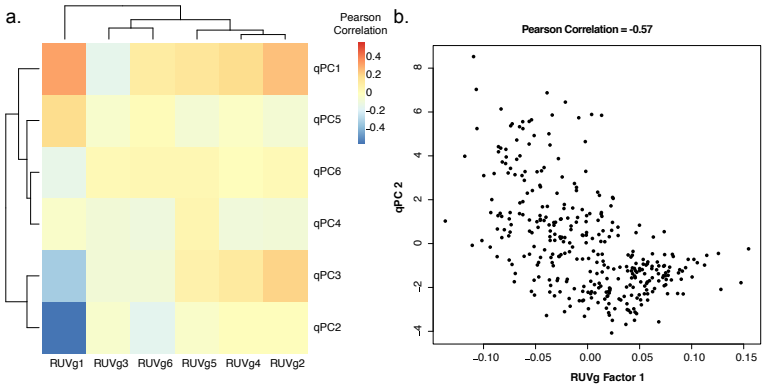


Figure 12: *Factors of UV in Gaublomme et al. [83].* **(a)** Heatmap of Pearson correlation coefficients between RUVg-derived factors of UV [109] and qPCs. Row and column clustering is generated from the R `hclust` function with default parameters. **(b)** Scatter plot of one anticorrelated pair of RUVg factor and qPCs, selected based on their high correlation magnitude displayed in (a).

4.4 DATA-ADAPTIVE PERFORMANCE RANKING

Global-scaling normalization methods are ranked similarly for both C1 data sets (Figure 11d-e), under-performing the more aggressive FQ normalization. Importantly, for the data set of Pollen et al., globally-scaled data do not show improved performance when compared to unscaled data (Figure 11b). Conversely, scaling by RLE size factors outperforms other scaling normalizations and FQ normalization for the 10x Genomics PBMC data set (Figure 11f). For all of these data sets, single-cell-specific methods, such as those implemented in the R packages `scran` and `SCnorm`, do not outperform methods developed for bulk RNA-seq.

The inclusion of a batch regression step in the normalization strategy is desirable for the SMART-seq data set; procedures including QC or RUV factors without batch normalization perform poorly (Figure 11h). This result indicates that, for this study, preexisting batch classifications are better proxies of inter-batch effects than QC or RUVg factors, despite their problematic associations with biological condition. In contrast, QC-based regression normalization outperforms RUVg for the Pollen et al. data set (Figure 11g), as well as in the 10x data set when paired with batch adjustment (Figure 11i). Taken together, these observations suggest that there is no single normaliza-

tion method that uniformly outperforms the others and that scone is able to identify appropriate normalization procedures in a data-dependent fashion.

4.5 SUBSAMPLED PERFORMANCE RANKING

One potential drawback of scone is its computational complexity, implementing and ranking hundreds of normalization procedures per data set. This can be especially problematic when applying scone to large data sets. In such cases, an efficient strategy is to use a random subset of the cells for the purpose of ranking normalizations, applying only the best performing normalization procedure to the full data set. For the 10x PBMC data set, this subsampling strategy leads to a ranking that is highly consistent with the ranking based on the full data, as illustrated in Figure 13. Importantly, as little as 10% of the cells is enough to yield more than 80% correlation with the full ranking (Figure 13d).

4.6 EXTERNAL MEASURES OF DE

I validate scone's performance assessment by relating normalized expression measures to controls derived from external DE studies. For Pollen et al. [82], I consider a set of positive and negative control genes for DE between CP+SP (cortical plate and subplate) and SZ+VZ (subventricular zone and ventricular zone) tissues an independent bulk microarray data set Miller et al. [131], available from the BrainSpan atlas (http://brainspan.org/static/download.html). I select the 1,000 most significant DE genes from that study as positive controls, ranking by limma P-values [132] (all 1,000 had P-value < 0.01). As negative controls, I took the 1,000 least significantly DE genes. I assess my ability to discriminate these two sets of genes in a comparison of GW16 (gestational week 16) and GW21+3 (gestational week 21, cultured for 3 weeks) cells based on normalized Pollen et al. [82] expression. DE analysis was performed using limma with voom weights [133], generating receiver operating characteristic (ROC) curves parametrized by P-value threshold. My approach identifies two clusters of procedures, including one cluster with low ROC AUC and low-to-moderate scone performance and another cluster with high ROC AUC and moderate-to-high scone performance (Figure 14a). The latter includes all FQ procedures as well as a subset of UQ procedures paired with QC adjustment. This example may suggest an

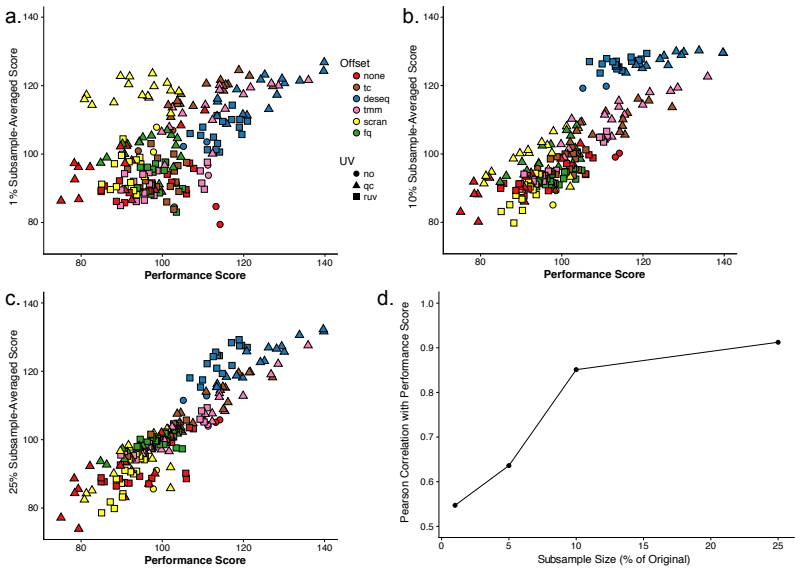


Figure 13: *scone analyses for subsamples of 10x PBMC data set [85].* **(a-c)** Average subsample performance score v. full-sample performance score. I randomly extracted 10 subsamples from the full data set corresponding to a fixed percentage of the original sample size, applied scone independently for each subsample, and averaged the 10 performance scores to obtain a final performance score per procedure. Plots are shown for subsamples comprising (a) 1% (b) 10%, and (c) 25% of the original sample. **(d)** Pearson correlation coefficient between average subsample performance score and full-sample performance score for different subsample percentages. When sampling at least 10% of the cells, I observed correlations greater than 0.8 with scores for the full data.

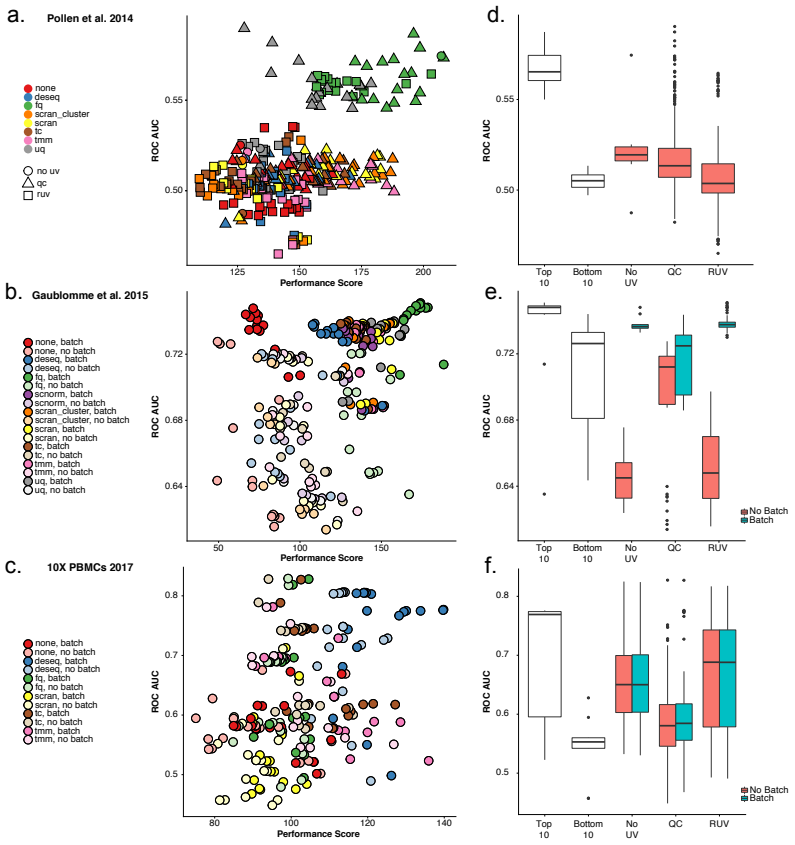


Figure 14: *Relationship between* `scone` *performance scores and external differential expression validation in three scRNA-seq data sets [82, 83, 85].* **(a-c)** ROC AUC v. `scone` performance score. Normalization procedures in the top-right corner are deemed best both by `scone` and by independent differential expression DE validation. **(a)** Comparing GW16 (gestational week 16) and GW21+3 (gestational week 21, cultured for 3 weeks) cells in [82], highlighting performance differences between scaling methods and the type of regression-based adjustment. **(b)** Comparing pathogenic and non-pathogenic cells in [83], performance differs between scaling methods and regression-based batch adjustment. **(c)** Comparing B cells and dendritic cells in 10x data set [85]; performance differs between scaling methods but not by batch adjustment. **(d-f)** Boxplots of ROC AUC for the bottom 10 (bot10) and top 10 (top10) procedures as ranked by `scone` and for procedures with RUV, QC adjustment, and neither ("No_UV"). Boxplots are further stratified by batch adjustment, when appropriate. Data sets are presented in the same order as in (a-c).

advantage in considering method classes rather than individual procedures, as in Figure 11d: while there is a spread in `scone` performance scores for any one scaling method, FQ performs well on average and this performance is validated by external comparisons.

For the SMART-seq data set of Gaublomme et al. [83], I utilize a separate bulk study to define two more control sets of genes in DE and no-DE between pathogenic and non-pathogenic T_h17 cells [134], available on Gene Expression Omnibus (GEO) with accession GSE39820. For each scRNA-seq normalization procedure, I tested for differences in expression between T_h17-positive pathogenic cells and unsorted non-pathogenic cells and generated ROC curves for the control set classification. I observe a relatively low correlation between the ROC AUC and the `scone` performance scores (Figure 14b; $r_s = 0.4$). However, the improved performances of the FQ method (Figure 11e) and batch adjustment method (Figure 11h) are validated by external DE data.

For the PBMC data set [85], I process an independent bulk microarray data set of Nakaya et al. [135], available on GEO with accession GSE29618. I computed sets of positive and negative control genes by comparing baseline B cell and baseline dendritic cell dendritic cell (DC) microarray samples. For each normalization procedure, I use these sets to evaluate DE between the single-cell clusters of B cells and dendritic cells, as defined by Seurat's clustering procedure [136] (Appendix A). RLE scaling performs well on average, as suggested by the `scone` performance score (Figure 11f).

While the `scone` ranking is not necessarily correlated with the AUC-based ranking across the whole performance range, I find an overall high level of agreement between the two rankings at the level of method classes, with the exception of RUV and QC methods. When considering many normalization procedures, users may also rely on top-ranking procedures to provide a basis for further exploration and downstream analysis. I found that the top ten normalizations as ranked by `scone` consistently performed well in terms of ROC AUC and better than procedures that consisted of scaling only (Figure 14d-f). Taken together, these results indicate that the `scone` performance ranking is a good way of identifying suitable normalization procedures for a given data set.

4.7 IMPROVED REPRESENTATION OF CELL-CELL SIMILARITY

My validation of the `scone` performance scoring in the previous section assumes that there were different cell populations to compare. In many scRNA-seq studies, however, the goal is to identify novel cell subpopulations via clustering analysis. Here, I aim to assess the ability of `scone` to identify normalization procedures or classes thereof that will lead to the best clustering of a given data set, using some notion of ground truth for cell clusters.

I simulate 10 independent data sets (20,000 genes in 1,000 cells) using the Bioconductor R package `splatter` [137] (Figure 15a). Simulation parameters are inferred from a subset of 100 cells from the 10x Genomics [85] "pbmc4k" data set, setting the DE probability to 0.3 and adding five cell populations (or "groups") of different sizes: one population comprising 50% of the cells, one comprising 20%, and the remaining three populations comprising 10% of cells each. I include dropouts in the simulation and add a batch effect (two batches of 500 cells) to make normalization more challenging.

I run `scone` on each simulated UMI data set, using a cell filtering scheme similar to the one above with the requirements of at least 1,000 UMIs, greater than 80% of common genes detected, below 0.65 AUC, and using a z_{cut} of 3 for greater data-adaptive leniency. Negative control genes (200 for evaluation) are *ideal* and extracted from the simulation. 200 positive control genes are selected based on maximum absolute average log fold change (lfc) as reported by the simulation.

The `scater` package [127] is used to compute the QC measures per simulated library, including: $\log_{10}$ total UMIs, $\log_{10}$ total UMI features, and percent UMIs in top 50, 100, 200, and 500 features. Batch information – not group information – was extracted from the simulation. `scone` assesses procedures with zero to three factors of QC or RUVg. PCA is used to decompose the log-normalized count matrix following each normalization, followed by k-means clustering on the space of the first 10 PCs. The correct number of clusters ($k = 5$) are used for every k-means clustering. I compute the adjusted Rand index (ARI) between the true simulated clusters and the clusters inferred by k-means clustering, reporting the average ARI across the 10 simulated data sets. The `scone` performance score is highly correlated with the ARI calculated between the simulated clusters and the clusters identified by k-means on the normalized data (Figure 15b-c).

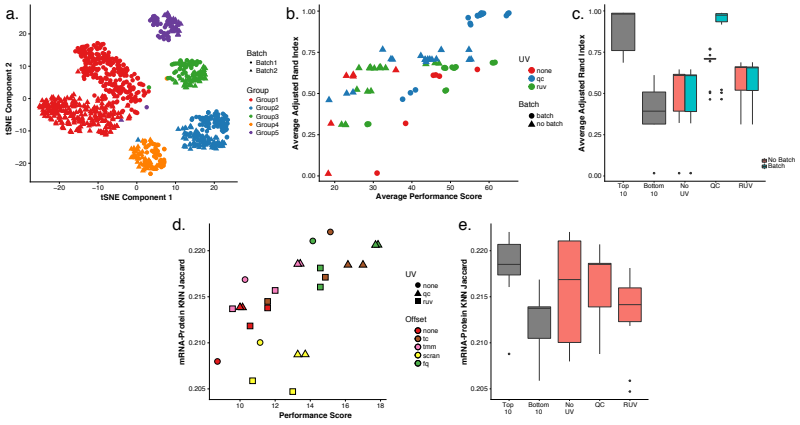


Figure 15: *Validating* `scone` *performance with simulated data and external cell-level data.* **(a)** tSNE of the first 10 PCs of the log-transformed, TC-normalized UMI counts for a data set simulated using `splatter`, with parameters inferred from the 10x PBMC data set [85]. **(b)** Average ARI between the true simulated clusters and k-means clusters ($k = 5$) for normalized data v. `scone` performance score (without BIO_SIL score), across 10 `splatter` simulations. A Pearson correlation of 0.73 between the two metrics highlights the ability of `scone` to select procedures that optimize aspects of clustering that are not explicitly accounted for in the performance panel. The top-performing procedure was FQ with adjustment for batch and 1 qPC. **(c)** Boxplot of average ARI for the bottom 10 (bot10) and top 10 (top10) procedures as ranked by `scone` and for procedures with RUV, QC adjustment, and neither ("No_UV"). The boxplot is stratified by batch adjustment for the latter 3 categories. **(d)** Jaccard score between k-NN graph of protein abundance measures and k-NN graph of normalized expression measures ($k = 792$, 10% of cells) v. `scone` performance score. A Pearson correlation of 0.60 between these metrics demonstrates how `scone` selects procedures that improve local representations of cell-cell similarity. **(e)** Boxplot of Jaccard score for the bottom 10 (bot10) and top 10 (top10) procedures as ranked by `scone`, procedures with no non-batch UV normalization ("No_UV"), and procedures with RUV or QC adjustment.

I also apply scone to the recent CITE-seq data set of Stoeckius et al. [87], in which gene expression and antibody levels for 13 cell-surface proteins had been jointly measured for the same cells. Specifically, I use scone to normalize the transcriptome measures and then examine manner in which their consistency with protein measures varies with normalization. Antibody-associated CITE-seq UMI counts are extracted from GEO entry GSE100866, corresponding to a sample of human CBMCs and mouse cells [87]. Cell UMI profiles are transformed using the centered log-ratio. Means and SDs for each of the 13 antibody measures are computed across all mouse cells (< 0.1 human RNA UMI fraction; Appendix A) and the mean plus SD is subtracted from all abundances, as described in Stoeckius et al. [87]. For human cells (mouse cells were not utilized beyond preprocessing; Appendix A), I constructed a k-nearest neighbors (k-NN) graph using the Euclidean metric in 13-dimensional antibody space (k = 792 or 10% of all human cells). For each normalization procedure, I apply PCA to log-transformed scRNA-seq data, selecting the top 10 PCs, and using them to construct a k-NN graph with the same choice of k. The ranking of normalization procedures by scone is compared to the ranking by the Jaccard similarity score of the RNA and protein k-NN adjacency matrices. I have considered other values for k (e. g. k = 8 or 1% of cells; data not shown), but found that the mean and range of Jaccard similarity scores decreased considerably for smaller neighborhood sizes. Computing k-NN graphs (k = 792, 10% of cells) for the two spaces, namely, protein abundance and transcript abundance (10 PCs), I observe an increase in overlap between the two graphs as the scone performance score increased (Figure 15d-e), reflecting how procedures ranked highly by scone are better at representing surface marker expression similarity.

4.8 USING CONTRASTS TO ADJUST FOR BATCH

The scone analysis of the T_h17 data set (Figure 11) demonstrates the importance of correcting for batch effects in single-cell RNA-seq data. However, depending on the experimental design, simply including a batch variable in the model is not always a viable option. As an extreme case, imagine a completely *confounded design*, in which each biological condition is assayed in a distinct batch. In such a case, regressing out the batch indicator from the expression measures will result in the removal of biological effects; conversely, not accounting for batch will make it impossible to attribute the observed

differences in expression measures to biological differences between conditions or technical differences between batches. Note that this is not just a thought experiment; several examples of data sets with suboptimal designs are discussed in Hicks et al. [96].

On the opposite end of the spectrum are experiments designed in such a way that each batch contains cells from each biological condition. Such *factorial designs* are the optimal choice, when possible. Hicks et al. [96] and Tung et al. [86] discuss practical aspects of designing factorial experiments in the context of scRNA-seq.

Although optimal, factorial experiments are not always possible or practical. An alternative strategy is to collect multiple batches of cells from each biological condition of interest – a nested design, as exemplified by the iPSC data set from Tung et al. [86] (Chapter 3; Figure 16). After scaling normalization, the cells clearly cluster by individual, but the cells for each individual are further clustered by batch (Figure 16a). Blindly removing these batch effects with a standard batch correction method, such as ComBat [39], removes the biological effects of interest along with the batch effects. Moreover, the QC measures collected as part of the `scone` pipeline are not able to completely capture the batch effects (Figure 16b), as the space of the first two PCs of the QC measures is dominated by the difference between a subset of low-quality cells and the rest of the cells. Explicitly accounting for the nested nature of the design while adjusting for batch effects is the only strategy that effectively removes the unwanted technical variation and preserves the biological signal of interest (Figure 16c).

The `scone` package is able to detect nested designs by examining the cross-tabulation of the biological and batch factors. Given a nested design, the nested batch effect adjustment based on Equations (18) and (19) is automatically applied as one of the different normalization strategies to be compared. Nested designs are common in single-cell studies due to various practical constraints (e. g. processing material from different tissues separately). The `scone` performance scores (Figure 16d) show how only procedures that remove the nested batch effects rank high in the evaluation step.

4.9 USER INTERFACE

As part of the Bioconductor R package `scone`, I have developed a Shiny app [138] that allows users to interactively explore the data at various stages of the

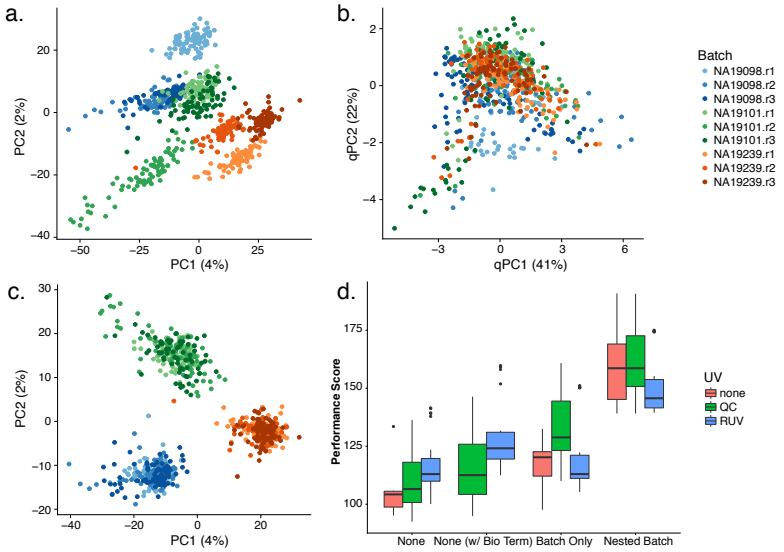


Figure 16: *scone results for human iPSC data set with nested study design [86].* **(a)** PCA of the log-transformed, TC-normalized UMI counts for all genes and cells passing quality filtering, with points coded by donor (color) and batch (shade). The cells cluster by batch, indicating substantial batch effects. **(b)** PCA of QC measures, with points coded by donor and batch. The QC measures do not appear to capture batch effects, but rather intra-batch technical variation. **(c)** PCA of log-transformed expression measures after FQ normalization followed by normalization for nested batch effects (top-performing procedure in `scone`), with points coded by donor and batch. As desired, cells cluster by donor, but not by batch. **(d)** Boxplot of `scone` performance score, stratified by regression-based normalization. Normalization procedures including a nested batch correction performed better than those without that step.

`scone` workflow. In Figure 10, I use the cortical neurons data set of Pollen et al. [82] to illustrate the app's functionality.

The `scone` package provides a function to display an interactive version of the biplot, allowing the user to select a group of normalizations for further exploration (Figure 10a).

The app also provides a hierarchical overview of all the compared normalization strategies (Figure 10b). The hierarchy is based on the series of algorithmic choices that constitute a given normalization strategy: In the example, the first level of the hierarchy represents scaling (e. g. FQ, TMM), while the second represents regression-based methods (QC or RUVg). In general, additional levels are present for the optional imputation and batch correction steps. Alternatively, the user can select a normalization strategy using the drop-down menu in the left panel of the app or using the interactive table at the bottom of the screen; this table can be sorted by the `scone` performance score or by any individual performance metric, making it easy to select, for instance, the procedure that maximizes the preservation of WV (as measured by the EXP_WV_COR metric).

Once a normalization approach has been selected for inspection, the Shiny app provides six exploratory tabs for an extended view of the normalized data, that should guide the selection of the final procedure. Here, I focus on three example tabs. The "Silhouette" tab (Figure 10c) shows the SW for each cell, for clustering based on PAM, clustering by batch, or clustering by biological condition (if available). If the user provides a set of positive and negative control genes, these are visualized in the "Control Genes" panel (Figure 10d) in a heatmap that includes batch, biology, and PAM clustering information. Similarly, the "Relative log-Expression" tab displays boxplots of the RLE measures for the normalized data (Figure 10e).

4.10 DISCUSSION

Many different normalization schemes are available, either specifically designed for scRNA-seq or borrowed from the bulk RNA-seq and microarray literatures. Here, I have demonstrated that simple global-scaling normalization is not always sufficient to correctly normalize the data and that more sophisticated strategies may be needed. However, different normalization strategies may perform differently across data sets, depending on the experimental design, protocol, and platform.

The main idea behind scone is to use a data-driven approach to select an appropriate normalization strategy for the data at hand. Although it may be infeasible to select a "best" normalization, as this would depend on a somewhat subjective definition of optimality, scone provides a set of performance metrics (and clustering via the biplot) that can be used to reduce the number of normalization procedures to further explore in the selection of a suitable strategy. One advantage of a panel-based normalization selection framework is that it can be communicated and reproduced by other investigators. I have shown using real data spanning different labs and technologies that scone is able to reliably rank normalizations by summarizing multiple dimensions of data quality.

Although scone has demonstrated its usefulness in published studies (e. g. Fletcher et al. [139], Afik et al. [140], Gadye et al. [141], and Martin-Gayo et al. [142]) and normalization can help remove UV from the data, a careful experimental design is the most important aspect of a successful use of scRNA-seq. In fact, if the biological effects of interest are completely confounded by unwanted technical effects, no statistical method will be able to extract meaningful signal from the data [96].

My discussion here surrounds normalization, but the scone framework is more general, facilitating the comparison of imputation methods [92, 93], dimensionality reduction techniques [94, 143, 144], and additional preprocessing steps such as gene and cell filtering. An alternative approach to imputation is to model the zeros as part of dimensionality reduction. An example of such a method, ZINB-WaVE [94], has the ability to include additional covariates to produce a low-dimensional representation of the data that is not influenced by UV. The covariates (e. g. QC or RUV factors) selected by scone as important for normalization can be included in ZINB-WaVE to provide a more robust projection of the data. Performance analyses of normalization as a preprocessing step can also aid such methods, for instance, by informing them on which covariates to include to adjust for unwanted technical effects.

This methodology and software are general and applicable to other types of non-scRNA-seq assays, including microarray, bulk RNA-seq, adductomic, and metabolomic assays. In particular, the user could extend the package by adding different metrics for scRNA-seq, as well as metrics specific to other assays. The scone package implementation leverages core Bioconductor packages for efficient parallel computation and on-disk data representation, both essential when analyzing large data sets [145–147].

The computational complexity of scone is directly related to the complexity of the normalization methods included in the comparisons. In particular, all scaling methods, RUV, and regression-based methods are very efficient, leading to reasonable computing time (e. g. 60 hours with 1 central processing unit (CPU) for the 10x Genomics PBMC data set of 12,039 cells). With parallelization, this computation can be sped up considerably (e. g. 11 hours with 10 processors). For very large data sets, subsampling can be used to decrease computation: One or more random subsets of the cells can be used to evaluate a set of normalization procedures using the scone metrics and only the selected normalization can be subsequently applied to the full data set.

4.11 CONCLUSIONS

Overall, scone provides a flexible and modular framework for the preprocessing of scRNA-seq data that can be used by practitioners to evaluate the impact of the statistical design of a given study and select an appropriate normalization, as well as by method developers to systematically compare a proposed strategy to state-of-the art approaches. Normalization can improve data representation and statistical analyses (i. e. regression analysis), but the performance metrics discussed in this chapter are articulated in terms of prior biological or technical assumptions (e. g. controls). Another way to monitor the quality of scRNA-seq analysis is to assess the replicability of findings across replicate single-cell experiments. Modeling the reproducibility of scRNA-seq signals is the subject of the next chapter.

5

SINGLE-CELL REPRODUCIBILITY ACROSS DONORS (scRAD)

5.1 INTRODUCTION

As single-cell technologies mature and push the boundaries of study resolution and scale, they promise to enrich our understanding of human disease, facilitating the development of new and exciting biological signals. Single-cell studies of a single donor's tissue may already prove very fruitful in the characterization and cataloging of diverse cell types; in some cases these single-donor observations may lead to bold hypotheses. Of course these approaches do not guarantee results which *generalize* to other individuals. When considering samples from human donors, differences in genetics, behavior, exposure, environment, and/or collection point, can produce uncontrolled axes of experimental variation and confound computational analyses that seek to identify and prioritize putative shared features. If the goal of an analysis is to predict phenomena linked to a specific phenotype or disease state, it is natural to consider reproducibility of signals across multiple sources. In this chapter I motivate and describe single-cell Reproducibility Across Donors (scRAD) – a generally applicable framework based on irreproducible discovery rate (IDR) analysis [1] – for charactering reproducibility of within-donor and between-donor signals.

In Section 5.2, I discuss coarse-grained structures in the single-cell expression data and the problem of clustering cells across multiple donors. In Section 5.3, I describe how scRAD leverages donor information in order to identify reproducible gene modules: axes of expression variation that are common to all (or most) individuals. Section 5.4 introduces quantitative tools built from the Li et al. [1] framework for assessing cross-donor reproducibility of within-donor signals. Section 5.5 summarizes additional tools for IDR-based and IDR-free meta-analysis applicable to multi-donor studies. Finally,

Section 5.6 describes lower-level utilities for monitoring reproducibility across donors.

5.2 CLUSTERING ACROSS DONORS

Quantifying the heterogeneity of gene expression in individual cells is often the initial focus of single-cell analyses, but understanding describing the data in terms of cell type distributions and modeling the response of that distribution to perturbation is an important downstream analysis priority. When performing single-cell analyses of multiple donors, it is important to compare cells between donors – i. e. to quantify the extent to which samples from different donors contain cells that are representative of the same cell state. There may be biologically meaningful differences in gene expression between similar cells collected from different subjects, but modeling and estimating those differences is a challenging task, especially when cells are viewed on a high-dimensional expression manifold. As discussed in Chapter 3, there may be subtle, transcriptome-wide batch effects due to technical differences in extraction and/or sample processing. Efforts to remove or adjust for these effects (Chapter 4) may be burdened by the fact that some cell types are represented in greater or lesser numbers across donors and others are unique to specific donors.

If either i) cell profiles are properly normalized and jointly clustered across biological conditions or ii) clusterings are harmonized using sophisticated computational techniques [148], the compositional differences between samples can be decoupled from other effects of interest. Furthermore, donor differences in cell cluster memberships can provide important insights into the differences between biological conditions (e. g. disease-affected tissue v. control tissue). Therefore, clustering analysis is a critical step in many multi-donor analyses of single-cell data, accounting for important between-donor variability. Below, I discuss a tool for clustering normalized scRNA-seq-based expression estimates. In addition to clustering, this tool selects an intrinsic dimensionality for the clustering, facilitating visualization.

5.2.1 PAM CLUSTERING AND DIMENSION SELECTION

Once expression data from multiple batches has been properly normalized and variance has been appropriately stabilized, I may attempt to relate cell

expression profiles to each other based on their pairwise Euclidean distance. In high-dimensional scRNA-seq data, many axes of variation are random and noisy: they may misinform us of the underlying cell type. Dimensionality reduction techniques, such as PCA, allow us to represent the data using a smaller number of features with high signal-to-noise ratios (SNRs). Distances computed on these features may be used to partition or cluster cells by cell type.

PAM is a clustering algorithm that searches for a set of k medoids spanning the data set, associating each cell with the closest cluster medoid. Because the number of clusters, k, is unspecified at the beginning of an analysis, it is common to choose a value of k that maximizes the ASW – that is, the average over all cell SWs, quantifying the extent to which cells cluster with their own cluster over others. This procedure is implemented by `fpc:pamk`.

The `scRAD::pamkd` function, introduced in the `scRAD` package[1], performs PAM clustering over a range of clusters, similar to `fpc:pamk`, but it examines a range of PCs, varying the dimension over which distances are computed. After clustering over a range of PC dimensions, this function selects a dimension for the data that maximizes the cluster number.

Let $\bar{s}(k,d)$ represent the ASW of a PAM k-clustering on d dimensions. I define $k(d)$ as the unique choice of k that maximizes $\bar{s}(k,d)$ for any choice of d. I selected d so as to maximize cluster number and tightness:

$$\begin{aligned} k(d) &\geq k(d') \forall \{d' | d' \neq d\} \\ \bar{s}(k(d),d) &\geq \bar{s}(k(d'),d') \forall \{d' | k(d') = k(d)\} \end{aligned} \tag{25}$$

The maximal cluster criterion biases the analysis to a representation with the largest possible number of clusters constrained by the conditional `fpc:pamk` objective; `scRAD::pamkd` fails in the absence of tight clusters, spuriously selecting large or maximal values of k. Therefore, the results of this method should only be applied when its selections are robust with respect to varying the maximum k or maximum d.

One advantage of the simultaneous k and d selection above is that I are left with a small number of PCs that are well-suited for downstream analyses such as non-linear dimensionality reduction (e. g. t-Distributed Stochastic Neighbor Embedding (tSNE)).

1 `scRAD` is available for download at `https://github.com/YosefLab/scRAD`

5.3 REPRODUCIBLE GENE MODULES

Given access to multiple samples of homologous single-cell populations, it may be important to identify and prioritize common axes of variation. Clustering analysis accounts for the distribution of cell types across donors, but the covariance structures these clusters reflect are not necessarily shared across donors. One way to prioritize these covariances is to ignore clusters that are biased toward any one donor. An alternative and complementary unsupervised approach aims to identify transcripts serving as reliable proxies of reproducible gene expression patterns. Reproducibly correlated gene modules may inform us of novel biological mechanisms that drive these systems. These genes should form genes modules that are consistently co-regulated across donor samples. If cell populations are heterogeneous, this kind of analysis may naturally follow a clustering step, uncovering reproducible modes of intra-cluster variation. Alternatively, reproducible correlation analysis may be done separately from clustering analysis altogether; reproducible modules can then serve as a basis for reproducible inter-cluster comparisons.

The `scRAD` package facilitates a three-step reproducible gene module analysis. The first step involves identifying gene-gene pairs that are reproducibly correlated, generating a gene-gene adjacency matrix. The second step is hub-identification in the reproducible gene-gene graph. The goal of this step is to identify genes that play critical roles in the reproducible covariance structure. The third step involves clustering hub genes into reproducible modules.

5.3.1 DEFINING A REPRODUCIBLE GENE-GENE ADJACENCY MATRIX

Normalized log-expression data is pooled for each donor and for each I separately compute the gene-gene Pearson correlation matrix. Correlation values are Fisher-transformed and subsequently centered and scaled resulting in robust Z-values with zero median and unit MAD over the upper triangle of the correlation matrix. A pair is called "reproducible" if its Z-value corresponds to a two-tailed P-value below a threshold (e. g. $P < 0.01$) in all donors. I build a gene-gene adjacency matrix by drawing undirected edges between any two genes in such a pair. This step is implemented by the `scRAD::get.repro.thresh.adjacency` function.

5.3.2 HUB IDENTIFICATION

For each gene, I tally the number of reproducible gene pairs to which it belongs. I considered whether I could find genes with significantly more pairs than would be expected by a Poisson model of vertex degree; these genes could serve as reliable proxies of reproducible correlations. The distribution of pair counts was modeled as a zero-inflated Poisson process, using an unconnected zero-component to model noisy genes with correlations consistently below threshold Under this null model, I computed upper-tail P-values using the `scRAD::pzipdegree` function, As these genes are connected to a large number of reproducible gene pairs, I called these proxy genes "reproducible module genes."

5.3.3 HUB CLUSTERING AND REPRODUCIBLE MODULE ANNOTATION

In the last step of our reproducible module analysis, I aim to cluster hub genes into reproducible modules. `scRAD` provides no tools for this step, but I may rely on other common libraries as there are many ways to cluster genes based on their correlation submatrix. Several `igraph` functions could then be used with the resulting graph in order to identify communities of hubs.

5.4 REPRODUCIBILITY-BASED DIFFERENTIAL EXPRESSION AND SIGNATURE ANALYSES

In this section I discuss new tools for supervised tasks such as gene-level DE, quantifying the extent to which transcripts are up- or down-regulated in one cell type vs all other cells.

Any DE analysis downstream of *de novo* clustering analysis demands careful consideration. Traditional DE analysis aims at identifying transcripts that vary markedly by cell class; a common goal is to rank the relative importance of transcripts in characterizing underlying expression states. Within the single-cell context, cell class is frequently defined based on low-dimensional representations of expression data. Therefore, the assumption that most genes are not differentially expressed between classes may not hold. Null models based on this assumption are ill-suited to the data, and will naturally yield uncalibrated probabilistic-based scores, e. g. artificially deflated P-value distributions.

In addition to biological factors, library intrinsic technical factors and batch-level features can drive broad expression covariance in scRNA-seq data. While some of these effects are random, others can confound DE analyses by systematically distorting transcriptome-wide differences between biologically distinct cell types. As discussed in Chapter 4, without sufficient modeling efforts batch specific biases can skew cluster classifications and reorder the ranks of DE genes.

A natural way to calibrate DE scores and avoid batch-specific effects is to apply *meta-analysis* to replicate experiments. Unfortunately, there is no natural analog for biological replicates in the single-cell context; I do not yet wield the necessary experimental controls to reproduce a specific sample of a transcriptional state. At the very least, I can map clusters from replicate experiments so that cluster contrasts are made comparable (Section 5.2).

5.4.1 IDR

The meta-analytical DE approach implemented in `scRAD` relies on a reproducibility metric known as IDR [1]. Rather than using hypothesis testing to identify genes showing "significant" evidence of DE, the IDR framework supports DE ranking based on *reproducible* across donors.

This metric evaluates a matched set of signals measured in two or more replicate experiments. The `scRAD::kruskalIDRm` tool performs DE analysis within each replicate sample using KW tests; for each single-cell comparison and for each donor sample, this yielded a list of lfcs and a list of P-values. The two-component IDR mixture model is used to fit the joint distribution of $-\log(P)$-values obtained from these tests. For each gene, I estimate a probability that the gene is a member of an *irreproducible component* for which P-values are high and uncorrelated rather than a member of an *reproducible component* for which P-values are low and correlated. Sorting genes by increasing probability of irreproducibility, one can compute the cumulative probability of membership for all genes of same or lower rank, defining the IDR. Genes can then be reported as differentially expressed after placing a threshold on cumulative irreproducible component membership probability

5.4.2 IDR WITH MANY REPLICATES

The `scRAD` package modifies the EM algorithm from the `idr` package to handle three or more replicates, as prescribed in the authors' original manuscript.

The many-replicate `scRAD::est.IDRm` function is adapted from the two-replicate `idr::est.IDR`, including some new error messages.

COPULA MIXTURE MODEL. In Li et al. [1], the authors describe a model generating M matched lists of I signals: x_{im}, where $i \in \{1, \cdots, I\}$ and $m \in \{1, \cdots, M\}$. Each signal i is drawn from a two-component mixture model. Membership to the reproducible component is indicated by a random variable $K_i \sim \text{Bernoulli}(\pi_1)$. Conditioned on membership K_i, M replicate latent variables, z_{im} are drawn from a multivariate normal: $\vec{z}_i | K_i = k \sim N(\vec{\mu}_k, \mathbf{\Sigma}_k)$. For each component, $k = 0, 1$, the m-dimensional mean vector and covariance matrix can be written:

$$\begin{aligned} \vec{\mu}_k &= \mu_k \vec{1} \\ \mathbf{\Sigma}_k &= \sigma_k^2 \left[(1 - \rho_k) I + \rho_k \mathbb{1} \right] \end{aligned} \tag{26}$$

where $\mu_0 = 0$, $\mu_1 > 0$, $\sigma_0^2 = 1$, $\rho_0 = 0$, $0 < \rho_1 < 1$.

Consider the marginal cumulative distribution function (CDF) of each signal under this model:

$$G(z) = \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z) \tag{27}$$

The true observations are generated in terms of continuous and unknown replicate-specific marginal CDFs F_m:

$$x_{im} = F_m^{-1}(G(z_{im})) \tag{28}$$

ESTIMATION. The estimation algorithm for this mixture model involves two major iterative steps that repeat until convergence [1]. The first step computes *pseudo-data* from observations using the empirical CDFs $\hat{F}_m$ and current parameter estimates $\theta = (\pi_1, \mu_1, \sigma_1, \rho_1)$:

$$z_{im} = G^{-1}\left(\hat{F}_m(x_{im}) | \theta \right) \tag{29}$$

The second step is application of an EM algorithm for the M-dimensional normal mixture model, maximizing likelihood of the pseudo-data and the hidden K variables. Derivations relevant to the M-dimensional extension can be found below.

COVARIANCE MATRIX DETERMINANT AND INVERSE. The matrix determinant of $\mathbf{\Sigma}_k$ can be computed via Sylvester's determinant theorem:

$$|\mathbf{\Sigma}_k| = \sigma_k^{2M} (1 - \rho_k)^{M-1} (1 + (M - 1) \rho_k) \tag{30}$$

The matrix inverse can be computed via the Sherman-Morrison formula:

$$\mathbf{\Sigma}_k^{-1} = \frac{1}{\sigma_k^2 (1-\rho_k)} \left(I - \frac{\rho_k}{1 + (M-1)\rho_k} \mathbb{1} \right) \tag{31}$$

MAXIMUM LOG-LIKELIHOOD OF PSEUDO-DATA. Considering the general model above, I can extend Equation (1.5) from Section 1 of the Supplementary Materials for Li et al. [1], "Estimation algorithm for the copula mixture model." This equation represents the second term of the expected log-likelihood $Q(\theta, \theta^{(t)})$, the only term that varies by μ_1, σ_1, or ρ_1:

$$\begin{aligned} \mathrm{E}\left[l_z | \theta^{(t)} \right] = \sum_{i=1}^{n} \mathrm{E}\, K_i \Bigg\{ & \log \left(\frac{1}{\sqrt{(2\pi)^M \sigma_1^{2M} (1-\rho_1)^{M-1} (1 + (M-1)\rho_1)}} \right) \\ & - \frac{1}{2\sigma_1^2 (1-\rho_1)} \sum_{p,q}^{M} \left\{ \left(z_{i,p} - \mu_1 \right) \left(z_{i,q} - \mu_1 \right) \left(I_{pq} - \frac{\rho_1}{1 + (M-1)\rho_1} \mathbb{1}_{pq} \right) \right\} \Bigg\} \end{aligned} \tag{32}$$

I can obtain estimates for model parameters (μ_1, σ_1, and ρ_1) by maximizing the expected likelihood.

μ_1 MLE. Taking a derivative with respect to the mean parameter, μ_1, setting it to zero, and solving for μ_1 (assuming $\sigma_1 > 0$, $\rho_1 < 1$, and $M > 0$), I can express the maximum likelihood estimation (MLE) estimate in terms of $\pi_i^{(t+1)} = \mathrm{E}\left[K_i | \vec{z}_i, \theta^{(t)} \right]$ as a weighted mean of replicate means $z_{i\cdot}$:

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} \pi_i^{(t+1)} z_{i\cdot}}{\sum_{i=1}^{n} \pi_i^{(t+1)}} \tag{33}$$

MLE FOR σ_1 AND ρ_1. Taking derivatives with respect to σ_1 and ρ_1 yields a system of equations for these parameters that can be solved and expressed in terms of the following quantities:

$$\begin{aligned} C_{pq}^{(t+1)} &= \sum_{i=1}^{n} \pi_i^{(t+1)} \left(z_{i,p} - \mu_1^{(t+1)} \right) \left(z_{i,q} - \mu_1^{(t+1)} \right), \\ W^{(t+1)} &= \sum_{i=1}^{n} \pi_i^{(t+1)}. \end{aligned} \tag{34}$$

The trace and sum of $\boldsymbol{C}^{(t+1)}$ are important data summaries:

$$\begin{aligned} T^{(t+1)} &= \sum_{p,q}^{M} C_{pq}^{(t+1)} I_{pq}, \\ S^{(t+1)} &= \sum_{p,q}^{M} C_{pq}^{(t+1)} \mathbb{1}_{pq}. \end{aligned} \tag{35}$$

The solutions for our variance and correlation estimates are:

$$\begin{aligned} \rho_1^{(t+1)} &= \frac{S^{(t+1)} - T^{(t+1)}}{T^{(t+1)}\,(M-1)}, \\ \sigma_1^{2(t+1)} &= \frac{T^{(t+1)}}{W^{(t+1)} M}. \end{aligned} \tag{36}$$

5.4.3 APPLICATIONS BEYOND REPRODUCTION

The IDR model was built to quantify reproducibility, but the mixture model has other applications. Consider three populations of cells, A, B, and C. If I assume the difference between A and B is small compared to their common differences with cluster C, I may claim that greater DE in A v. C will correspond to greater chance of reproducible DE in B v. C. By this assumption, IDR analysis can be applied to multiple lists of IDR values from similar experiments in order to identify genes for which signals obtained from both comparisons are correlated. Genes passing this threshold and having common sign of DE across multiple comparisons may be called "shared" genes.

5.5 BIOMARKER AND SIGNATURE ANALYSES

Before diving into low-level IDR analysis methods in the last section, I discuss additional high-level analyses that can be implemented using `scRAD` tools, including new examples of general IDR-based analyses and meta-analysis.

5.5.1 MARKER PREDICTION

One may consider multiple criteria when selecting candidate markers based on DE analysis, especially when a search is targeting surface marker proteins. The `scRAD::getMarkers` tool nominates potential biomarkers for subpop-

ulations by synthesizing multiple analyses. The tool reports the intersection of three gene sets:

1. genes reproducibly differentially expressed between two conditions (e. g. target cluster v. all). This criterion is based on the IDR-based DE analysis output of `scRAD::kruskalIDRm`. This differential expression condition requires specificity.

2. reproducible module hub genes identified using the procedure described above (Section 5.3). The hub condition requires that selected transcripts show robust and reproducible co-expression with many other genes.

3. external list of known candidate gene products (e. g. predicted membrane molecules from the Human Protein Atlas: `http://www.proteinatlas.org`)

5.5.2 OTHER IDR APPLICATIONS: IDENTIFICATION OF UPSTREAM REGULATORS

In Section 5.4, I applied the IDR framework to differential expression. In this subsection we explore how these methods can be applied to other kinds of signals, such as gene signatures defined from related data sets.

E. g. in the short hairpin RNA (shRNA)-knockdown study of Chevrier et al. [149] investigators considered the effect of knocking down signaling regulators, transcriptional regulators, and phosphoproteins on the mRNA expression of mouse bone marrow-derived dendritic cells (BMDCs). Correlating single-cell gene expression profiles of similar cells (i. e. DCs) with bulk expression profiles of these knockdowns, one may identify potential upstream regulators mediating measured single-cell responses.

As discussed in Chapter 3, observations of zeros in scRNA-seq are both plentiful and unreliable. I can therefore use a weighted correlation of cells in terms of the FNR weight matrix $w_g j$:

$$\begin{aligned}
\mu_j^{(W)} &= \frac{\sum_{g=1}^{G} w_{gj} y_{gj}}{\sum_{g=1}^{G} w_{gj}} \\
\mathrm{Cov}_{jj'}^{(W)} &= \frac{\sum_{g=1}^{G} w_{gj} w_{gj'} \left(y_{gj} - \mu_j^{(W)}\right)\left(y_{gj'} - \mu_j^{(W)}\right)}{\sum_{g=1}^{G} w_{gj} w_{gj'}} \\
\mathrm{Cor}_{jj'}^{(W)} &= \frac{\mathrm{Cov}_{jj'}^{(W)}}{\sqrt{\mathrm{Cov}_{jj}^{(W)}\,\mathrm{Cov}_{j'j'}^{(W)}}}
\end{aligned} \tag{37}$$

where weights of population data are set to unity. In this case, the opposite of the correlation can be referred to as an "upstream regulatory score," as it measures the extent to which a single-cell response is anticorrelated with an expression profile in which specific regulators have been inhibited. These scores may not be too meaningful in absolute terms, but differences in these signatures could hint at specific regulatory activity differences between cell populations. Testing for differences in many signatures between cell population is conceptually similar to differential expression: therefore `scRAD::kruskalIDRm` supports this type of differential signature analysis.

5.5.3 NON-IDR SIGNATURE META-ANALYSIS UTILITIES

IDR mixture model estimation may not feasible in some cases, including situations in which the number of genes and/or signatures is small. For example, a fit may quickly converge to a point where all tests are called irreproducible with 100% probability. In these cases it is more straightforward to consider an alternative meta-analysis approach, such as Stouffer's z-method for P-value aggregation. The `scRAD::kruskalMeta` function implements a routine similar to `scRAD::kruskalIDRm` except that it applies P-value aggregation rather than estimating IDR values. While this method does not quantify reproducibility, it is not fit based, requiring no choices of initial parameters.

5.5.4 IDR-BASED REPRODUCIBLE MODULE ANALYSIS

In this subsection I introduce one more IDR-based analysis in `scRAD`. In the previous section, I have discussed multiple steps for reproducible module analysis. In the first step of this analysis, I define a gene-gene graph by identifying pairs with reproducibly high levels of correlation. The `scRAD::get.repro.thresh.adjacency` function identifies these pairs by calling correlation below a modeled P-value threshold. I offer a second function, `scRAD::get.repro.idr.adjacency`, that selects these pairs by running IDR analysis on the same matrix of P-values, calling pairs below a modeled IDR threshold. Unfortunately, inference is slow enough to prohibit its application to full scRNA-seq expression matrices, but it may be applied to much smaller sets of genes, or even signature scores. The latter, "reproducible signature module" analysis may be useful in predicting relationships between the regulators within the context of any single-cell study.

5.6 OTHER EXTENSIONS TO IDR ANALYSIS

In addition to high-level functions such as `scRAD::kruskalIDRm`, `scRAD` implements new tools for probing lower-level IDR analysis. It was mentioned above that `scRAD::est.IDRm` is the underlying function behind IDR analysis in `scRAD`. The arguments to this function include a matrix of signals (e. g. log-transformed P-values from gene-level KW tests) and initialized parameter values for the underlying fit.

5.6.1 SUBSAMPLING TESTS

`scRAD` offers another low-level function, `scRAD::est.IDRm.sample`, which runs IDR analysis over various subsamples of tests. By running the analysis many times, I can evaluate the stability of our outputs and compare the range of fit parameters values after I vary our initial parameter estimates. These fits should robustly converges on the same final values. Another element of the output allows us to monitor the relationship between the mean probability of membership of a test to the irreproducible component and its relation to the standard deviation of those estimate across samples. Extreme membership probabilities should be quite robust, while intermediate estimates may be noisy and unreliable.

5.6.2 PAIRWISE CORRELATION METRIC

At times, it may be useful to consider pairwise signal correlations between M replicates signal vectors. `scRAD` offers a `scRAD::corIDR` function that runs pairwise IDR analysis between all replicate pairs, returning the correlation parameter from the fit. This parameter translates to the correlation of signals in the reproducible component of the signals (ρ_1). These results can be useful in discriminating outlier replicate pools: samples that exhibit lower signal correlation (e. g. < 0.5) with most replicates are reproducing signals poorly and may be candidates for removal.

5.7 CONCLUSIONS

In this chapter I introduced a number of tools in `scRAD` that can be used to monitor the reproducibility of signals across replicate scRNA-seq experiments. In the next section I will describe applications of methods and concepts from both `scRAD` and `scone`, illustrating their usefulness in the context of human studies.

PART III

APPLICATIONS TO HUMAN IMMUNOLOGY

6

DONORS WITH COMMON PHENOTYPE

6.1 INTRODUCTION

Human immunity relies on the coordinated responses of many cellular subsets and functional states. Inter-individual variation in cellular composition and communication could thus potentially alter host protection. Work on ECs has demonstrated enhanced cytotoxic CD8$^+$ T cell responses [6, 7] and improved crosstalk between the innate and adaptive immune systems [150–152]. Collaborators recently reported that enhanced cell-intrinsic responses to HIV-1 in primary myeloid dendritic cells (mDCs) from ECs lead to effective priming of HIV-1-specific CD8$^+$ T cell responses *in vitro* [151]. Nevertheless, the master regulators driving this mDC functional state, the fraction of EC mDCs that assume it, its biomarkers, and how to potentially enrich for it are unknown. Here[1] I analyze scRNA-seq data to examine viral responses among the DCs of three ECs of HIV-1 infection.

I used the `scRAD` toolset to overcome the confounding effects of donor variability and identify reproducible patterns in gene expression across donors who share the EC classification. My analysis highlights a functional antiviral DC state in ECs. Integrating existing genomic databases into my reproducibility modeling framework, I identify immunomodulators that increase the fractional abundance of this state in primary PBMCs from healthy individuals *in vitro*.

My results demonstrate how single-cell approaches can reveal previously unappreciated, yet important, immune behaviors and empower rational frameworks for modulating systems-level immune responses that may prove therapeutically and prophylactically useful.

1 This chapter is based on a published paper in BMC Genome Biology: "A Reproducibility-Based Computational Framework Identifies an Inducible, Enhanced Antiviral State in Dendritic Cells from HIV-1 Elite Controllers." [81] © The Authors 2018, reproduced with permission.

6.2 STUDY OVERVIEW

My collaborators and I have applied scRNA-seq to evaluate heterogeneity of transcriptional responses of mDCs (CD14^{-}, CD11c^{Hi}, HLA-DR^{+}) from three EC individuals after *in vitro* exposure to a vesicular stomatitis virus G glycoprotein (VSV-G) pseudotyped HIV-1 virus or media control. We developed a broadly applicable strategy combining reproducibility-based computational analyses with targeted experimentation in order to resolve, characterize, and modulate common response states across multiple donors (Figure 17). I utilized tools developed by my group for single-cell data analysis, including `FastProject` [153] and my own `scone` [126] (Chapter 4) and `scRAD` (Chapter 5) to identify reproducible response states, pathways, and biomarkers across multiple donors who share the EC classification.

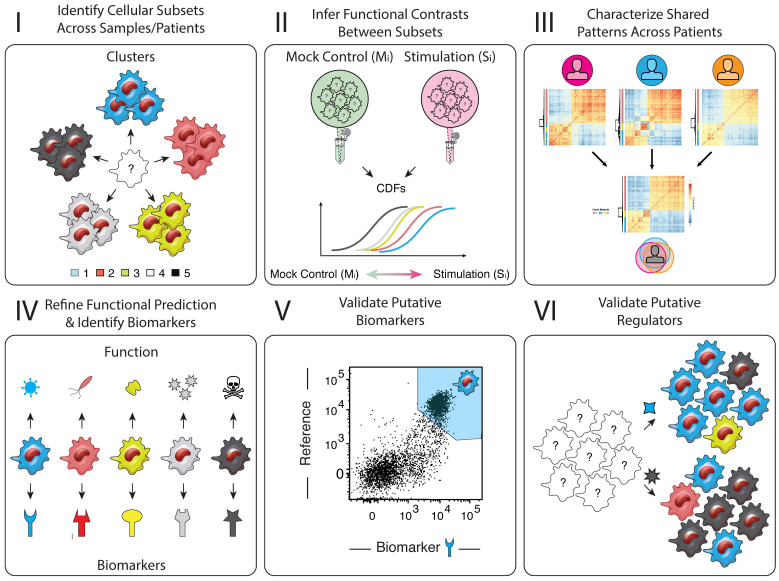


Figure 17: A generally applicable framework used to resolve, characterize and then modulate response states across multiple donor sources. **(I)** Resolve the individual mDC subtypes and states that comprise the system under study. **(II)** Define putative functions for each and identify biologically meaningful contrasts using existing databases. **(III)** Characterize patterns of differential expression that are common across donors. **(IV)** Nominate potential biomarkers and relevant cellular circuitry based on accumulated knowledge. **(V)** Isolate and characterize interesting subsets. **(VI)** Validate inferred regulators.

6.3 SHARED SUBSETS

In order to identify features of mDC innate immune responses to HIV-1 shared across ECs, my collaborators performed scRNA-seq on PBMCs from three ECs ("p1", "p2", and "p3") exposed *in vitro* to either a VSV-G pseudotyped HIV-1 virus or a media control for 48 hours (Figure 18). Stimulating PBMCs

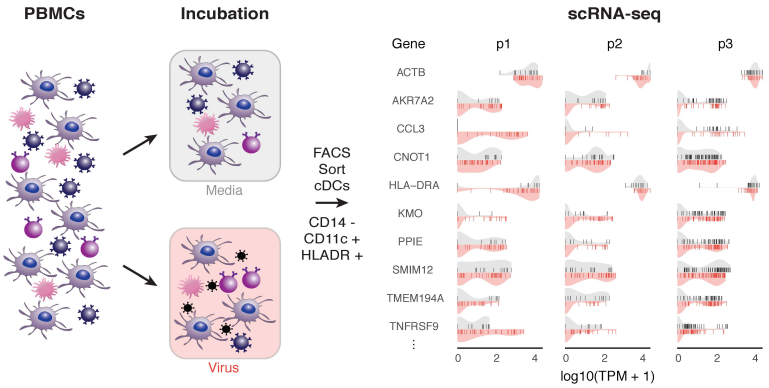


Figure 18: *EC-only scRNA-seq study design.* **Left:** Schematic representation of experimental system. After incubation with virus or a media control for 48 hours, mDCs were isolated from PBMCs by FACS and profiled by scRNA-seq. **Right:** Violin plots of single-cell expression levels for ten select genes for each EC donor (p1, p2, p3). Vertical lines represent individual cellular values; the upper (gray) half of the violin shows the distribution of values for the media control and the bottom (red) shows the same for virus-exposed cells.

mimics some of the critical physiological interactions that occur between mDCs and other immune cell types, while the use of a VSV-G pseudotyped HIV-1 particles enhances mDC infection efficiency [154]. Given the potential bias of viability sorting, which may discard dying DC undergoing viral stress responses, my collaborators sequenced two sets of plates, sorted and unsorted based on LIVE/DEAD blue viability dye (Invitrogen, Carlsbad, CA, USA). Following incubation, collaborators sorted single mDCs and performed a SMART-seq2 based scRNA-seq protocol [155].

6.3.1 SINGLE-CELL EXPRESSION QUANTIFICATION

RNA-seq reads were aligned to the RefSeq hg38 transcriptome (GRCh38.2) using `Bowtie2` [156]. The resulting transcriptomic alignments were processed by `RSEM` to estimate the abundance (expected counts and transcripts per million (TPM)) of RefSeq transcripts [78]. Several genes were quantified multiple times due to alternative isoforms unrelated by RefSeq annotation. Before expression data normalization, these TPM estimates were summed to produce a single TPM estimate per RefSeq gene symbol.

After estimating gene expression levels, I applied the `scone` [126] pipeline to filter out single-cell libraries with poor alignment characteristics ("*in silico* cell filtering") and normalize the remaining data to minimize the impacts of these characteristics on expression quantification.

6.3.2 DATA FILTERING

For each single-cell library, I computed transcriptome alignment and QC metrics analogous to the ones listed in Table 9. I used the `scone::metric_sample_filter` function to flag libraries with

1. low numbers of aligned reads (< 28,840; Figure 19a)
2. low percentages of aligned reads (< 15%; Figure 19b)
3. low percentages of detected transcripts (< 33.4% of Ensembl GRCh38.80 protein-coding genes expressed at > 100 TPM in at least 10% of cells, or "common genes"; Figure 19c),

I further identified 99 genes of candidate constitutive expression by fitting a population-wide Fano factor as a linear function of mean TPM, selecting the 99 common genes with minimal fit residual. These genes covered a range of 50.0-35,000 TPM. For each cell, we modeled a FNR curve, and used the AUC to distinguish cells with poor detection properties (Figure 19d). Viability-sorted mDC data exhibited only a two- to three-fold enrichment in high-quality cells compared to unsorted cells, suggesting that incubated primary cells from HIV-1 infected donors represent a fragile source material (Figure 20). Out of 2489 initial cells, only 393 (318 at 48 hours and 75 at 24 hours) cells passed this primary filter. Following cell filtering, genes were retained for downstream analysis if they were annotated as protein-coding and expressed at levels > 100 TPM in at least five high-quality cells.

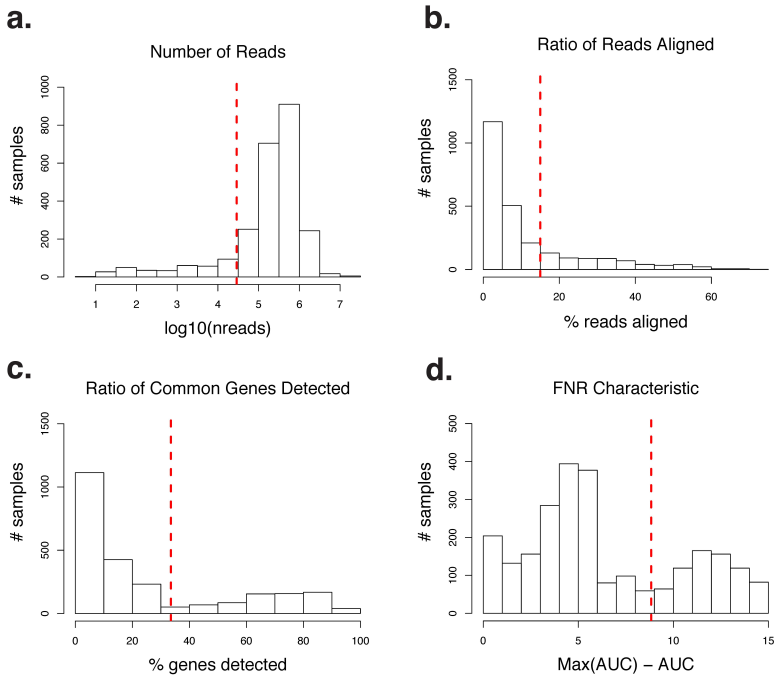


Figure 19: *Distributions of single-cell sample (24 and 48 h) filtering metrics.* Red lines represent adaptive threshold below which all cells (n = 2489) were removed from further analysis. **(a)** Distribution of number of paired-end reads per library. **(b)** Distribution of transcriptome read alignment ratio per library. **(c)** Distribution of the fraction of common genes detected per library. **(d)** Distribution of fit FNR AUC per library.

6.3.3 DATA NORMALIZATION

In order to normalize TPM data between cells, I applied the FQ normalization method, restoring original zero values to zero following normalization. This restoration step was necessary due to widespread zero-ties. I used normalization metrics of the `scone` [126] package to assess performance of this strategy.

The first three scores measure the correlation between the first three PCs of the TPM matrix and the first three PCs of: i) the matrix of library-level QC metrics, ii) the unnormalized matrix of TPM estimates for the negative control, MSigDB "HSIAO-HOUSEKEEPING-GENES" gene set, and iii) the un-normalized matrix of TPM estimates for the positive control, MSigDB "REACTOME-INNATE-IMMUNE-SYSTEM" gene set. Following normal-

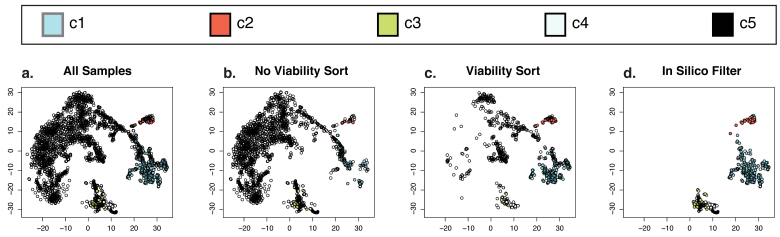


Figure 20: *In silico cell filtering.* tSNE plots of (un-normalized) $\log(\text{TPM}+1)$ expression, including all cells from 24 hours and 48 hours, HIV-1 and media exposures, with or without viability gating. Points are colored according to a 48 hour cells' membership to clusters c1-c5. Various subsets are plotted independently, including **(a)** All single-cell samples. **(b)** Cells that were not sorted on viability. **(c)** Cells passing viability sorting. **(d)** Cells passing *in silico* cell filter. Viability sorting tends to exclude cells from low-quality clusters, enriching the fraction of cells passing the quality filter.

ization, the first two scores decreased while the third increased slightly, suggesting that technical structure has been removed from the data while retaining structure associated with the biological processes at hand (Figure 21a).

The three ASW scores were defined for i) biological class = donor ID × exposure × time point × viability, ii) batch class = well plate batch, and iii) stratified PAM clustering. Following normalization, the first two scores decrease, suggesting that confounding by biological and batch factors could not be addressed by this normalization. However, the rise of the third score suggests greater intra-stratum clusterability following normalization (Figure 21b).

The last two scores i) the median absolute RLE and ii) the variance of the RLE IQR both decreased, implying reduced global DE following normalization (Figure 21c).

6.3.4 CLUSTERING ANALYSIS AND VISUALIZATION

PCA was applied to all filtered and normalized single-cell log-TPM data collected at the 48 hour time point; consequent analysis was limited to the first 50 PC values explaining 32% of expression variance. Unsupervised k-medoids clustering with the `scRAD::pamkd` function revealed five distinct transcriptional response states (clusters c1-c5; Figure 22a). Due to the high-dimensionality of the underlying expression space, clustering was visualized using a two-dimensional tSNE projection applied to the $d = 7$ Euclidean distance metric.

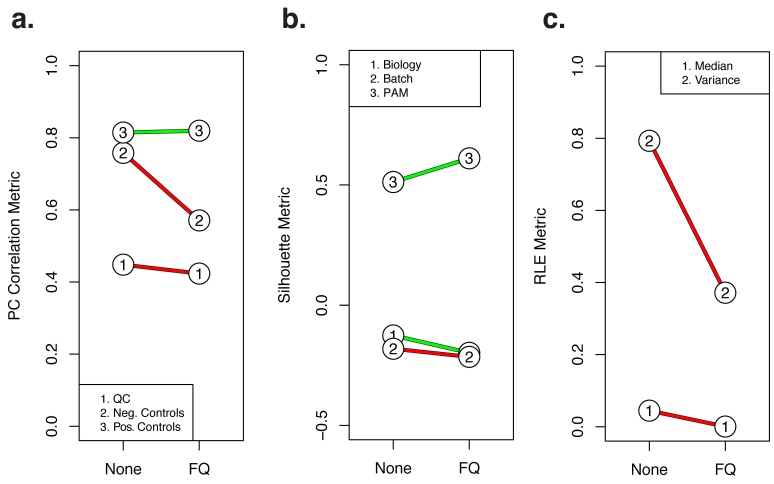


Figure 21: *Differences in* `scone` *metrics before and after FQ normalization.* **(a)** Correlation between the first three expression PCs and the first three PCs computed across negative controls (Alignment QC metrics and housekeeping genes) tend to decrease while correlations with the first three PCs across positive controls (innate immune system genes) tends to increase. **(b)** The ASW of biological condition (donor x exposure x time point x viability sort) and the ASW of batch both decrease. However, the ASW of *de novo* PAM clustering tends to increase. **(c)** The mean of cell-median RLEs decreases, as does the variance of the cell-IQR RLE decrease: both global differential expression and differential expression variability is reduced.

Low-dimensional representation of normalized expression estimates with tSNE illustrates how cells from each of the three EC donors span a common expression state-space: cells from different donors often share similar expression profiles, forming mixed clusters; all but one state (c5) is observed in all three donors.

After clustering, I applied linear regression to model each gene i's log-expression in cell j as a function of donor, exposure, and cell type:

$$\mathrm{E}\left[\log y_{ij}\right] = \alpha_i + \beta_i^p \times \mathrm{donor}_j + \beta_i^e \times \mathrm{Exposure}_j + \beta_i^c \times \mathrm{Cluster}_j \quad (38)$$

donor features were coded p1 v. p3 and p1 v. p2, exposure coded hiv v. media, and cluster coded c1 v. c2-c5. Two-sided t-tests identified 131 and 14 genes that were significantly associated with donor and exposure, respectively (Bonferroni-adjusted P-value < 0.01), while 1170 genes were significantly associated with cluster contrasts. These numbers suggest that cluster

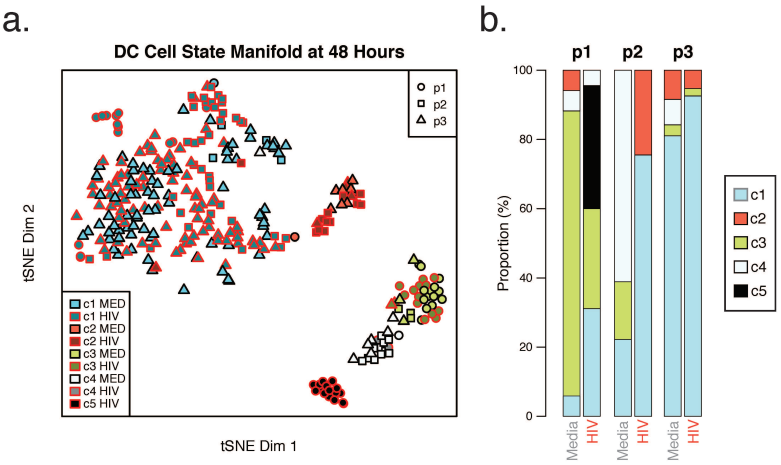


Figure 22: *scRNA-seq identifies five response clusters among EC mDCs.* **(a)** tSNE of all FACS sorted mDCs across three EC subjects passing quality filters (p1: circles, p2: triangles, p3: squares). Virus exposed cells are outlined in red; media exposed cells have no outline. Cells separate into five distinct clusters (c1-5). **(b)** Stacked bar plot depicting the percentage of total mDCs in each cluster for each donor under media and viral exposure conditions

identity is far more determinant of global gene expression than donor or exposure.

Cluster proportions are themselves associated with donor and exposure condition: for c1-c4, I modeled the relative abundance of cluster k as a logistic model of donor and exposure:

$$\text{logit}\left(\text{E}\left[\text{I}\left(c_j = k\right)\right]\right) = a_k + b^p \times \text{donor}_j + b^e \times \text{Exposure}_j \tag{39}$$

The fractional abundance of c1-c4 varied significantly ($P < 0.05$) across the three donors and two exposure conditions. Among these, the c1 response state was consistently enriched among virally exposed mDCs (P-value=8.5×10^{-6}, logistic regression, Wald test) while c3 and c4 were more common among media-exposed cells (P-value= 1.3×10^{-4} and 1.1×10^{-5}, respectively, logistic regression, Wald test) (Figure 22b).

Donor p1 cells at 24 hours were assigned partial cluster identities by projecting their profiles into the first seven PCs of the 48 hours data. Following projection, the 30 nearest 48 hours neighbors by Euclidean distance were identified and used to assign partial memberships proportional to the memberships of the neighbors. Similar, though less pronounced, fractional abundance shifts were observed in these cells (Figure 23).

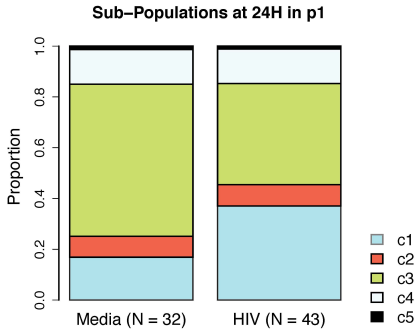


Figure 23: *Single-cell distribution at 24 hours in donor p1.* Stacked bar plot depicting expected percentage of total mDCs in each cluster for p1 at 24 hours under media and viral exposure conditions.

6.3.5 QUANTIFYING VIRAL ABUNDANCE

Within the virus-exposed p1 mDCs, I detected viral product primed from adenine-rich regions in the pseudotyped HIV-1, allowing me to consider associations between cell intrinsic responses and viral sequences (Figure 24). For each cell, viral abundance was quantified as a mean of RSEM TPM estimates for Gag and Pol gene segments[2], given the even coverage observed across those segments. I applied FETs to compare HIV-1 detection across virally exposed subpopulations (excluding p1-specific cluster 5), but found no significant trends: viral product was observed at comparable frequencies across the four universal clusters c1-c4. Similarly, KW tests comparing gene expression in HIV-1 positive and negative groups (all exposed cells) found no significant intra-cluster variation. These findings as well as those in the last subsection suggest that average virus-induced expression changes in the DC compartment are well explained by shifts in the frequencies of invariant cell types.

2 GenBank accession AF324493.

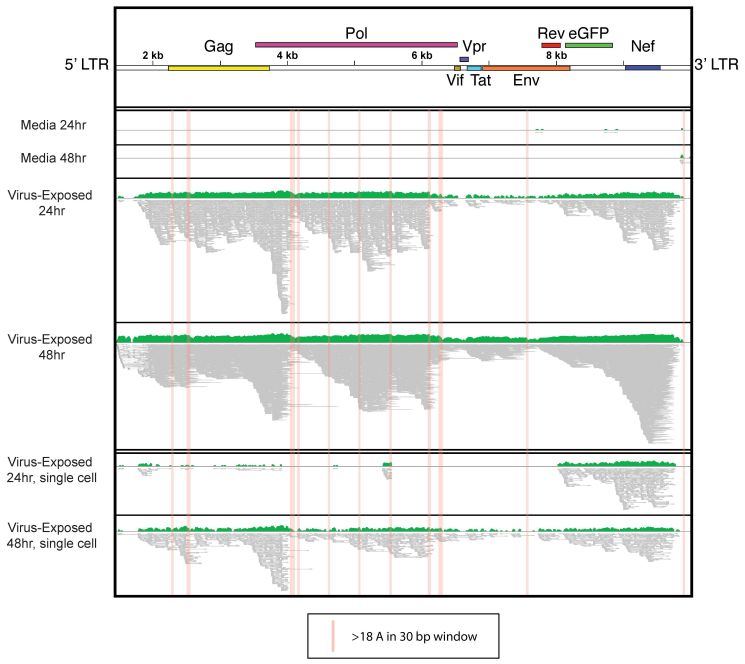


Figure 24: *Intracellular viral products can be captured with scRNA-seq.* Alignment of reads (reads in grey; histogram of reads in green) from pooled media or virus-exposed p1 cells at 24 and 48 hours (top) to the viral sequence between the 5' and 3' LTRs of the pseudotyped viral plasmid (bar at top, colored by gene). Representative single cells are shown at bottom. Vertical bars mark positions in the plasmid sequence where there are at least 18 adenines in a 30-base pair window.

6.4 REPRODUCIBILITY-BASED FUNCTIONAL ANALYSIS

To further examine these five EC mDC response states, I utilized `FastProject` [153]: a software package for visualization and interpretation of scRNA-seq data with reference to prior biological knowledge (Figure 25). I searched GEO (https://www.ncbi.nlm.nih.gov/geo/) for all study entries matching the query 11, utilizing the results to identify relevant expression signatures from the MSigDB C7 collection. Signature inputs include the selected MSigDB signatures, a curated signature of 28 IFN-response genes [151], and a precomputed cluster signature. I selected a few of the top signatures from my `FastProject` analysis, considering the cumulative distribution of signatures

[Organism]	AND [Data Set Type]	AND [Sample Source]
"homo sapiens"	"expression profiling by array"	"dendritic cell"
NOT "mus musculus"	OR "expression profiling by' high throughput sequencing"	OR "dendritic cells"

Table 11: *GEO query parameters.*

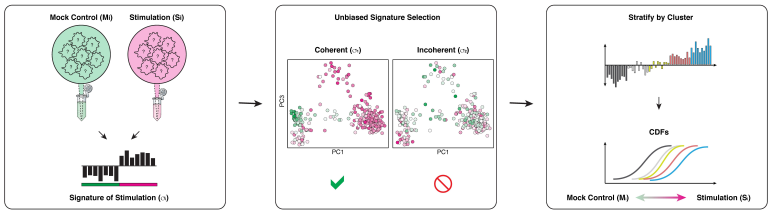


Figure 25: *Characterization of transcriptional single-cell response groups.* **Left:** Schematic of signature database. The expression of a bulk sample of simulated DCs (S_i) is compared to the expression of a mock control (M_i). Highly ranked up-regulated and down-regulated genes comprise the signature σ_i. **Middle:** σ_i is applied to all cells in the study and `FastProject` identifies pairs of expression data projections and σ_i for which σ_i varies coherently across the projection. **Right:** Coherent σ_i values are binned by cluster to nominate specific cluster contrasts as biologically meaningful.

across each of the five clusters. Two-sided Kolmogorov–Smirnov (KS) tests were performed between the signature distributions of clusters in order to monitor the extent to which these signatures discriminate the populations. Coherently varying gene expression signatures identified by `FastProject` frequently implicated c1 and c2 – but not c3-c5 – as responses associated with elevated DC activation (Figure 26). The transcriptional behavior of c1 mDCs appeared more consistent with elevated innate antiviral activity, displaying greater signature values for DCs exposed to viruses such as HIV-1 or Newcastle virus. c2 was well distinguished by signatures of DCs stimulated through alternative pathogen associated molecular patterns (PAMPs), such as lipopolysaccharide (LPS) and R848, or by specific bacteria or parasites (Figure 26).

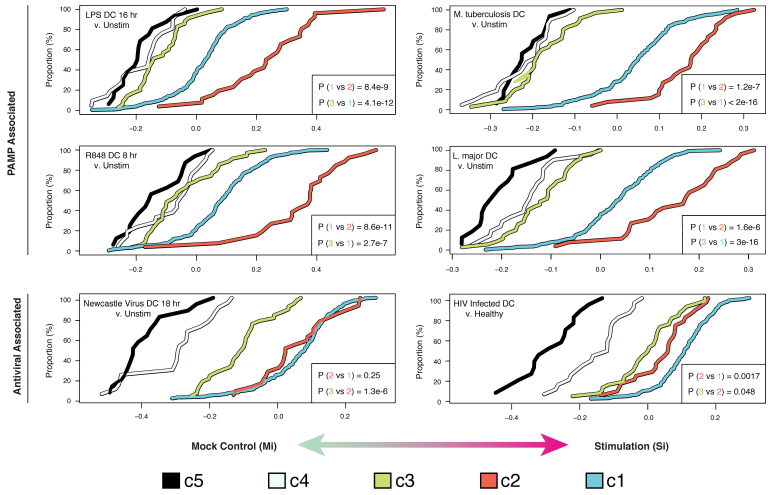


Figure 26: CDF comparisons for single cells from each cluster identified in Figure 17 with `FastProject` gene signatures derived from MSigDB records of GSE360 [157], GSE14000 [158], GSE22589 [154], GSE18791 [159], and GSE2706 [160]. The single-cell signature value quantifies the extent to which each cell is polarized toward a stimulated instead of unstimulated expression state. Clusters with gene expression signatures more closely mapping to the stimulated condition shift right, while clusters characteristic of unstimulated shift left. Two-sided KS test P-values highlight significant differences in these signatures between the first three clusters (c1, n = 220; c2, n = 26; c3, n = 35).

6.4.1 DE ANALYSIS

Motivated by the biological relevance of signatures contrasting c1 and c2 against the remaining clusters, I tested each gene for DE between these two populations and the pool of c3, c4, and c5 cells ("c3-5"). As in most experiments involving non-model organisms, inter-subject biological and technical variability poses a substantial confounding risk by systematically distorting or exaggerating transcriptome-wide differences between groups. To address this, I applied the KW-based DE module of `scRAD` (Section 5.4). I pooled cells from donors p1 and p2 together because they had the fewest high-quality cells: pooling them together increased average stratum power. IDR tools implemented in the CRAN IDR package are designed to analyze only two replicates, a limitation addressed by `scRAD`.

I compared this approach to DE effects estimated according to a more standard linear batch adjustment of log-expression in gene i in cell j:

$$\mathrm{E}\left[\log y_{ij}\right] = \alpha_i + \beta_i^p \times \mathrm{donor}_j + \beta_i^t \times \mathrm{CellType}_j \tag{40}$$

For each gene, I can estimate a separate offset, cluster effect (c1 v. c3-5), and donor effects (p1 or p2 v. p3) to model the expression of that gene across all cells in the study occupying the extreme clusters. 29% of the genes that are called as significantly differentially expressed (adjusted P-value < 0.01) are not reproducibly so (IDR < 0.01; Figure 27b). These genes are differentially expressed, but they do not vary consistently across the two cell pools – i. e. a gene may be differentially expressed in one but not the other. A gene could also be differentially expressed in both pools, but the rank of the difference varies substantially between pools.

On the other hand, only 8% of genes that exhibit reproducible P-value rank fall below my significance threshold. The presence of more tests like this (i.e., a higher fraction) could suggest that the significance threshold is too stringent or that there is an issue with the underlying null model used for computing P-values. Beyond the small number of insignificant reproducible tests, the IDR criterion appears to be stricter than the batch-adjusted significance criterion, selecting a smaller set of tests with uniform results across replicate pools. The IDR approach also appears to better emphasize aspects of clustering that are reproduced over multiple donors (Figure 27b). In order to partition differentially expressed genes (IDR < 0.01) into a "common" set from both clusters (c1 and c2) and two cluster-specific sets, I used `scRAD` again, this time performing meta-analysis to aggregate the DE results obtained independently

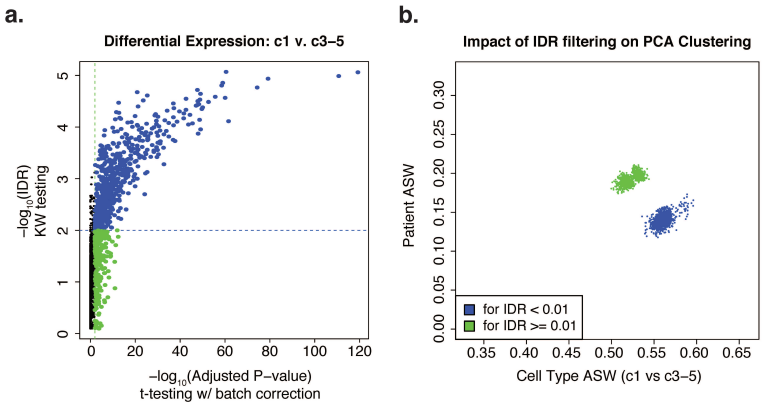


Figure 27: *Batch effects in IDR analysis.* **(a)** For each gene, comparison of IDR KW differential expression criterion to linear regression t-test criterion, the latter adjusting for batch (donor) effect. Blue genes meet both criteria, while green genes meet only the traditional criterion; IDR selection is generally more conservative than the alternative. **(b)** Each point corresponds to a subsampled PCA analysis. Low IDR (blue) and high IDR (green) genes from (a) are subsampled 1000 times to maintain comparable expression means across sets. An Euclidean cell distance metric is computed over each set, filtering expression data to the top third of PC variance. ASWs are computed for donor condition and cell type cluster condition; while donor effects decline upon IDR selection, cell type differences improve.

c1 v. c3-5 and c2 v. c3-5. In line with known pathway elements shared between the DC antiviral and bacterial and parasitic response pathways [149, 161], I uncovered 121 genes that were commonly up-regulated when comparing either c1 or c2 compared to c3-5 (Figure 28). Some of the remaining differentially expressed genes from these two comparisons were partitioned into three additional groups: i) c1-specific, for which a gene is called differentially expressed in both c1 v. c3-5 AND c1 v. c2 comparisons, but not c2 v. c3-5; ii) c2-specific, which is analogously defined; and (3) "discordant" for which genes are called differentially expressed in all three comparisons. I identified 103 genes that were uniquely called as up- or down-regulated in c1 or c2 relative to the remaining clusters (Figure 28). Genes preferentially expressed by c1 include the interferon-stimulated gene (ISG) IFIT3, whereas genes preferentially expressed by c2 encode molecules associated with endocytosis and antigen presentation (e. g. LAMP3 [162], Figure 28), suggesting different levels of activation or polarization between c1 and c2.

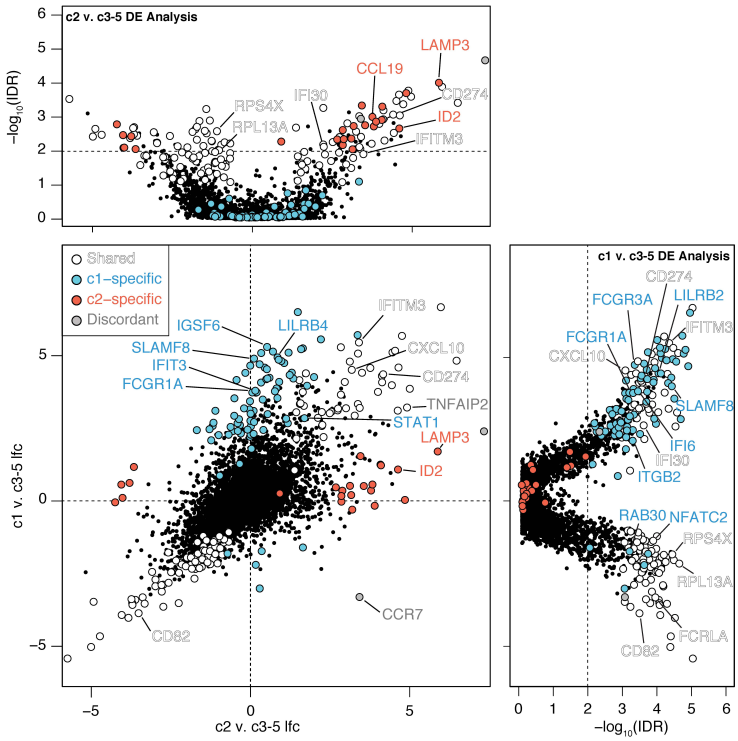


Figure 28: Potential genes specific for c1 (cyan), c2 (orange), shared between c1 and c2 (white) or inconsistent across individuals (gray). Individual volcano plots of negative $-\log_{10}$ (IDR) v. mean lfc between clusters c1 and c3-5 (right) and c2 v. c3-5 (left).

A targeted analysis of the expression of 28 ISGs regulated by HIV-1 [151, 163] suggested that c1 displayed the most potent and coherent interferon-induced transcriptional signatures (P-value=2.5×10^{-7}, two-sided KS test c1 v. c2; c1, n=220; c2, n=26; Figure 29). Given the large number of c1 cells

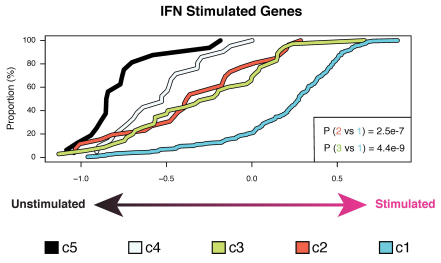


Figure 29: CDF plot for an unsigned `FastProject` signature of (n = 28) ISGs. As in Figure 26, clusters with stronger IFN stimulated gene signatures are shifted right. KS tests show c1 has a significantly higher IFN signature than c2 or c3.

at 48 hours, I additionally considered the expression modulating effects of viral exposure on cells from that cluster. Due to the small number of cells tested I imposed an additional reporting criterion of two-fold difference to call genes in DE. Several canonical antiviral response genes were differentially expressed between virus- and media-exposed c1 cells, highlighting that stimulation-induced changes also contribute modestly to measured transcriptional variation (Figure 30).

6.4.2 IPA

For each of the main three DE comparisons, I applied Ingenuity pathway analysis (IPA)[3] [164] to the list of lfc (pool mean) and IDR, setting a less restrictive cutoff of IDR < 0.05. The data set was used as the reference background for P-value calculation and all experimentally verified mammalian associations were included in the analysis. IPA reported BH Q-values for canonical pathways enrichments and I performed my own Bonferroni P-value adjustment for all reported upstream analysis P-values. The analysis revealed

3 QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis

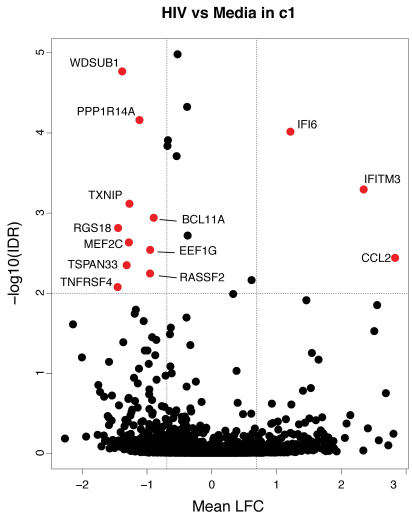


Figure 30: Impact of HIV-1 condition on c1 cells. Volcano plot representing genes enriched in VSV-G pseudotyped HIV-1 v. media exposure conditions across cells from c1: $-\log_{10}$ (IDR) is plotted against mean lfc across the donor pools. Genes differentially up-regulated in HIV-1 (right) or media control condition (left) are highlighted in red and labeled.

that the gene set reproducibly differentiating c1 from c3-5 is enriched for pathways related to DC maturation (BH Q-value=4×10^{-6}), innate recognition of microbes by pattern recognition receptor (PRR) (Q=8×10^{-5}), interferon (Q=3×10^{-3}) and toll-like receptors (TLR) signaling (Q=0.03, Figure 31). These pathway enrichments do not reach significance for c2. Several molecules were associated with antiviral responses with enhanced activity in c1 (IFNG, IFNA, STAT1). Significant TLR activation (TLR3, TLR4) enrichments were observed for c1 but not c2 (Figure 32). Overall, these observations suggest that c1 represents a subset of mDCs in an activated viral response state that could potentially inform the effective innate antiviral immune responses observed in bulk mDC from ECs [151].

6.5 REPRODUCIBLE BIOMARKER ANALYSIS

To further study the c1 response state, I sought to identify putative markers for prospectively isolating c1 cells after exposure to HIV-1 across ECs. I used

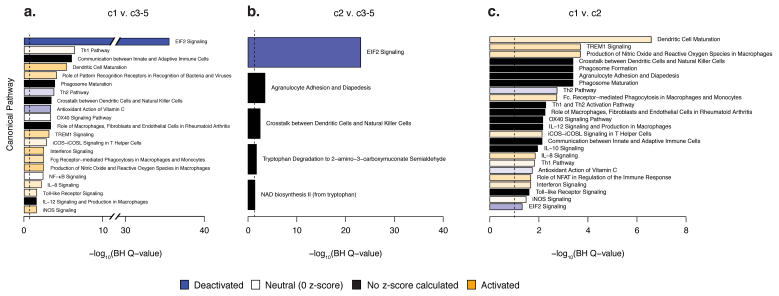


Figure 31: *IPA Canonical Pathways analysis.* Selected results for canonical pathways significantly (BH Q-value < 0.01) deactivated (blue), neutral (white: with IPA z score; black: without z score), or activated (orange) in **(a)** c1 versus c3-5. **(b)** c2 versus c3-5. **(c)** c1 versus c2.

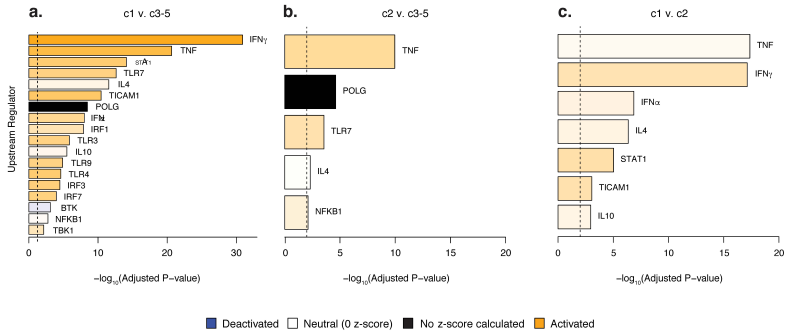


Figure 32: *IPA Upstream Regulators analysis.* Selected results for upstream regulators significantly (Bonferroni-adjusted P-value < 0.05) deactivated (blue), neutral (white: with z score; black: without z score), or activated (orange) in **(a)** c1 versus c3-5. **(b)** c2 versus c3-5. **(c)** c1 versus c2.

the two reproducibility-based criteria for surface marker candidacy implemented in the biomarker selection module of scRAD:

1. The surface marker must be encoded by a transcript that is reproducibly up-regulated in c1 v. c3-5 (IDR < 0.01).

2. The transcript encoding the surface marker should be correlated with sufficiently many genes, in a reproducible manner, across all donors (Figure 33)

I generated a list of 74 candidate c1 mDC markers using this procedure (Figure 34). Based on antibody availability, my collaborators selected five

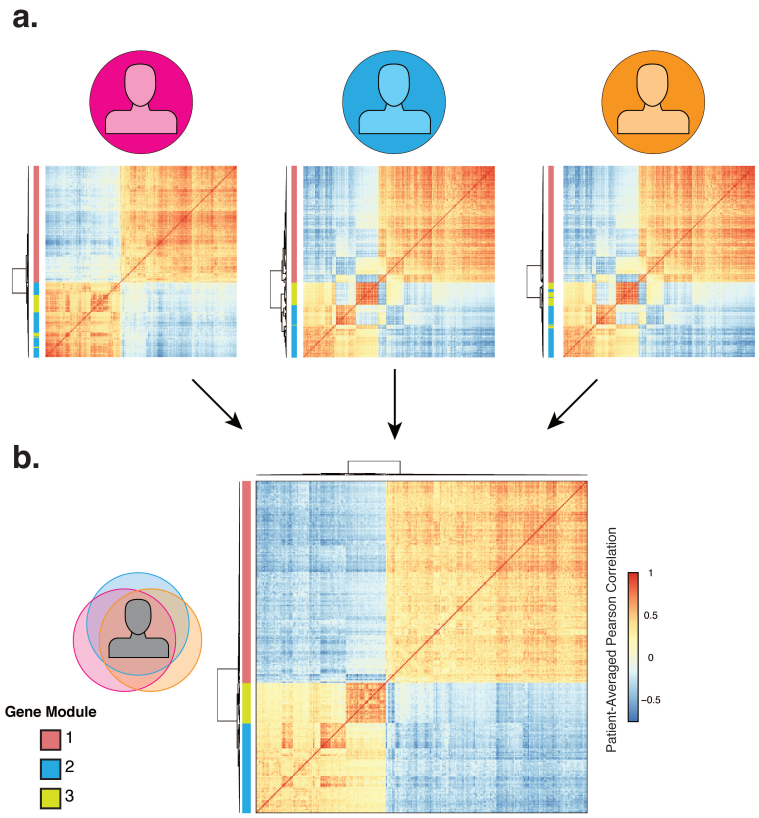


Figure 33: *Reproducible gene modules across three ECs.* Correlations were scaled to Z-values with 0 median and MAD equal to 0.67. Only gene pairs with $|Z| > 2.4$ in all three donor matrices were considered reproducible. 263 reproducible hub genes were called at P-values < 0.01 following Bonferroni adjustment. **(a)** Hierarchical clusterings of the gene-gene correlation matrix for hub genes across all three EC donors. Genes are clustered by complete-linkage clustering on correlation distance. **(b)** Hierarchical clustering of the median gene-gene correlation matrix. Reproducible hub genes may be clustered into three modules (m1-m3).

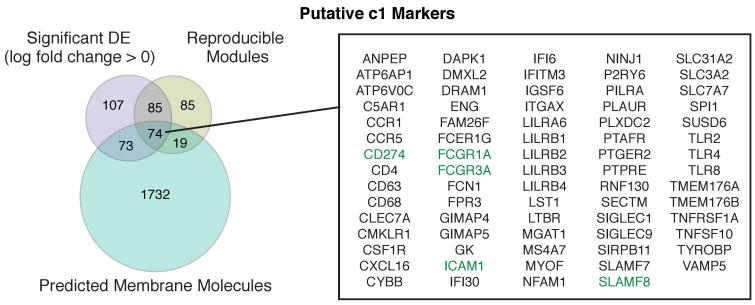


Figure 34: *Marker selection for c1-like cells.* 74 genes (listed in box) were: i) differentially expressed between c1 and c3-5, ii) reproducibly correlated with other c1 genes across all three ECs profiled, and iii) predicted membrane proteins. Candidate markers shown in green were selected for validation by FACS.

proteins (FCGR3, FCGR1, CD274, ICAM1, SLAMF8) to profile in mDCs by flow cytometry 24 hours after infection with pseudotyped HIV-1. Among these, both CD64 (FCGR1A) and PD-L1 (CD274) exhibited the most dramatic and consistent virus-induced up-regulation among mDCs isolated from the PBMCs of the three ECs characterized by scRNA-seq, as well as those from five additional EC donors (Figure 35a; P-value=7.8×10^{-3}; two-tailed Wilcoxon signed-rank test; n=8).

CD64 is an Fc-receptor for IgG [165], while PD-L1 has been implicated in mediating the balance between T cell activation and immunopathology, as well as immediate effector differentiation and long-term memory formation in T cells [166]. Importantly, high expression of PD-L1 has also been found on tolerogenic murine mDCs in chronic lymphocytic choriomeningitis virus (LCMV) infection [167] and in inflammatory lymph node-resident mDCs from HIV-1 infected individuals [168]. Nevertheless, high expression of IFN and inflammatory cytokines identified in my pathway analysis of c1 (Figure 31) and high CD86 expression levels on CD64^{Hi} and PD-L1^{Hi} cells indicates that these cells are highly activated inflammatory DCs.

When collaborators analyzed mDCs based on surface expression levels of CD64 and PD-L1 following viral stimulation, they observed two dominant mDC populations: one CD64^{Hi},PD-L1^{Hi} and the other CD64^{Lo},PD-L1^{Lo} (Figure 35b).

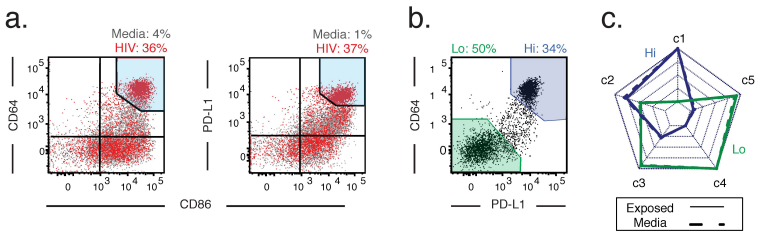


Figure 35: *CD64 and PD-L1 enriched for highly functional c1-like mDCs.* **(a)** Flow cytometry analysis of either CD64 (y-axis, left panel) or PD-L1 (y-axis, right panel) v. CD86 (x-axis) expression in mDCs from EC donor 1 (p1). Numbers above represent the percentage of CD64^{Hi}/PD-L1^{Hi} cells (top right gate; light blue) at 24 hours in media (gray) and VSV-G pseudotyped HIV-1 virus exposure (red) conditions. **(b)** Flow cytometry plots showing analysis of CD64 v. PD-L1 expression on mDCs exposed to VSV-G pseudotyped HIV-1 for 24 hours, defining two populations: CD64^{Hi},PD-L1^{Hi} (Hi; blue) and CD64^{Lo},PD-L1^{Lo} (Lo; green). Percentage in each gate is listed above. **(c)** Radar plots representing relative similarities of each subset (c1-c5) to population-level RNA-seq data from cells in the Hi and Lo PD-L1, CD64 gates 48 hours after viral (solid line) or media exposure (dashed line).

6.5.1 VALIDATION OF C1 POPULATION

We next applied population-level transcriptional profiling to mDCs sorted on CD64^{Hi},PD-L1^{Hi} or CD64^{Lo},PD-L1^{Lo} at both 24 and 48 hours post-viral stimulation. As with the scRNA-seq data, I applied RSEM alignment and sample-filtering procedures to population RNA-seq samples. Expression values for 6557 genes were normalized using RLE scaling normalization [169], followed by a one factor qPC adjustment. A total of 576 of the DE gene symbols from the c1 v. c3-5 comparison, were detected in population experiments. Correlations were computed between sorted populations and FNR-weighted means after `log1p`-transforming both data sets. Radar plot cycles representing these correlations are presented on a min-max scale per bulk condition (Figure 35c), revealing CD64^{Hi},PD-L1^{Hi} gene expression profiles dominated by the signature of the c1 and, to a lesser extent, c2 response states. mDCs sorted on CD64^{Lo},PD-L1^{Lo} matched a mixture of c3-5, supporting our conclusion that CD64 and PD-L1 co-expression enriches for c1 cells. Importantly, while these two markers are predominantly associated with c1 re-

sponses, they are not necessarily causally involved in inducing either phenotype.

6.6 FUNCTIONAL CHARACTERIZATION

Given the ties between strong antiviral activation and immune control of HIV-1, my collaborators and I wondered whether the CD64^{Hi},PD-L1^{Hi} mDC phenotype, common to ECs, was uniquely enriched within these individuals and might be linked to common features of immune control against HIV-1. While this phenotype was consistently and efficiently induced in HIV-1 exposed mDCs from ECs, markedly lower proportions of it were observed in HIV-1 exposed mDCs from chronic progressors (CPs) and healthy donors (HDs) (Figure 36a). Correlating the fractional abundance of CD64^{Hi},PD-L1^{Hi}

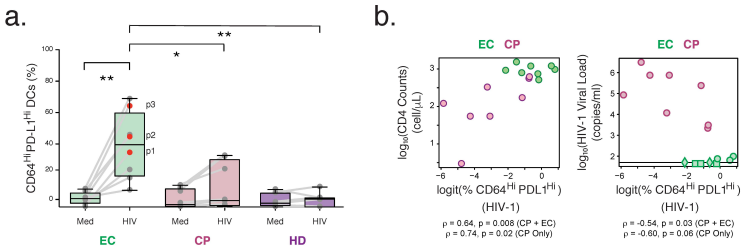


Figure 36: **(a)** Proportions of CD64^{Hi},PD-L1^{Hi} mDCs induced from multiple ECs (n = 8), untreated CPs (n = 8), and HDs (n = 7) after 24 hours of culture in media or VSV-G pseudotyped HIV-1 (*, P < 0.05; **, P < 0.01; two-tailed Wilcoxon signed-rank test). **(b) Left:** Correlation between the proportions of CD64^{Hi},PD-L1^{Hi} mDCs induced and clinical CD4$^+$ T cell count. ECs (n=8) and untreated CPs (n=8) were pooled together (P-value=8×10^{-3}, two-sided permutation-based P-value on Spearman correlation). CPs were also considered separately (P-value=2×10^{-2} (one-sided)). **(b) Right:** Correlation between the proportions of CD64^{Hi},PD-L1^{Hi} mDCs induced and HIV-1 viral load P=3×10^{-2} (two-sided) for ECs and untreated CPs. P=6×10^{-2} (one-sided) for just CPs. Diamond and square points represent indeterminate viral loads of < 20 and < 50 copies/mL, respectively.

mDCs after HIV-1 exposure against clinical phenotypes, I observed a significant positive association with CD4$^+$ T cell count across both CPs (one-sided) and ECs+CPs (two-sided; P-value=2×10^{-2} and 8×10^{-3}, respectively; two-

sided permutation-based P-value on Spearman correlation). Plasma HIV-1 viral loads, meanwhile, were negatively associated with percentages of CD64^{Hi},PD-L1^{Hi} mDCs across all donors (P-value=3×10^{-2}, Spearman correlation two-sided permutation P-value), with insignificant association in CPs alone (P-value=6×10^{-2}, one-sided P-value, Figure 36b). These associations show that a donor's CD64^{Hi},PD-L1^{Hi} mDC fraction after viral stimulation tracks traditional biomarkers along a spectrum of HIV-1 control, suggesting that the ability to induce c1-like cells might be a useful biomarker of enhanced protective immune responses against HIV-1.

Our collaborators sought to directly probe the association between the induction of c1 responses and the enhanced functionality observed in bulk mDCs from EC. They first examined the putative enhanced antigen presentation and T cell activation abilities of the c1-like subset of mDCs by performing mixed leukocyte reactions to compare my CD64,PD-L1 high and low mDC subpopulations. In these experiments, the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC population demonstrated superior ability to stimulate CD4^{+} and CD8^{+} T cell proliferation relative to CD64^{Lo},PD-L1^{Lo} mDCs across multiple ECs (Figure 37a, P-value=1.6×10^{-2} and P-value=3.1×10^{-2}, respectively; two-tailed Wilcoxon signed-rank test; n=6). Similar results were observed in assays conducted with T cells from ECs, where CD64^{Hi},PD-L1^{Hi} mDCs were capable of efficiently stimulating the production of IFNγ^{+} in a significantly higher proportion of autologous CD8^{+} T cells as compared to CD64^{Lo},PD-L1^{Lo} mDCs (Figure 37b; P-value=3×10^{-2}; two-tailed Wilcoxon signed-rank test; n=5). Further, IFNγ^{+} CD8^{+} T cells primed in the presence of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs expressed significantly higher levels of both the degranulation markers CD107a and TNFα (Figure 37c; P-value=1.5×10^{-2}; two-tailed Wilcoxon signed-rank test; n=7), mirroring the polyfunctional cytotoxic T cell (CTL) responses observed in ECs [6, 7].

6.7 ADJUVANT SIGNATURE META-ANALYSIS

Given the possible therapeutic and prophylactic potential of c1-like DCs for studies in non-EC populations with less efficient responses to *in vitro* viral stimulation (Figure 36a), I sought to predict the common signaling pathways involved in the acquisition of the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC phenotype. IPA results for c1 had highlighted several signatures of human DC stimulation, including multiple components of several TLR signaling path-

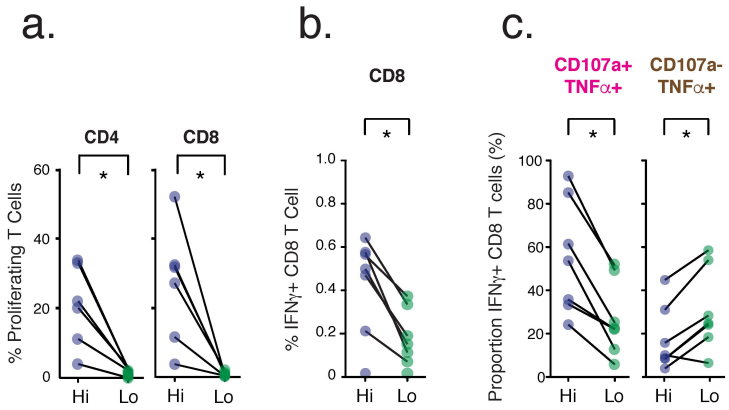


Figure 37: **(a)** Proportion of proliferating $CD4^+$ (left) and $CD8^+$ (right) T cells co-cultured with the Hi and Lo sorted virus-exposed mDCs populations (n = 6 donors). **(b)** Proportion of total $IFN\gamma^+$ $CD8^+$ T cells cultured with the Hi and Lo sorted virus-exposed mDCs populations (n = 7 donors). **(c)** Scatter plots of proportions of $CD107a^+$, TNF^+ (left) and $CD107a^+$, TNF^- (right) $CD8^+$ T cells cultured with Hi and Lo mDCs (n = 7 donors). Statistical significance was evaluated using a two-tailed Wilcoxon signed-rank test (*, $P < 0.05$).

ways (Figures 31 and 32); thus, I aimed to compare my single-cell expression profiles to perturbed bulk expression data in order to determine which TLR pathways were most compatible with the c1 signature v. c3-5.

RSEM and sample-filtering procedures were applied to population RNA-seq data collected from DCs incubated for 48 hours with or without various TLR ligands (no TLR, TLR2/3/4/8). Expression values for 18,482 genes were normalized using RLE scaling normalization [169]. I used weighted correlation to define, for every cell and every TLR ligand I tested, a stimulation score which reflects the similarity between the a cell's transcriptional profile and the one induced by the ligand. I then scored each ligand by the extent to which its respective stimulation scores in c1 cells are higher than in clusters c3-5 (using a KW test). Finally, using the differential signature analysis module in `scRAD`, I combine the resulting P-values across donors. Notably, for this analysis I used the Stouffer-Z P-value combination method since the number of hypotheses (i. e. TLR ligands) is small, leading to instabilities in the IDR inference. My meta-analysis showed that c1 cells correlated most positively with TLR3 stimulation via Poly I:C compared to the c3-5 (FDR < 0.01; Figure 38a), generating the actionable hypothesis that triggering the endosomal double-

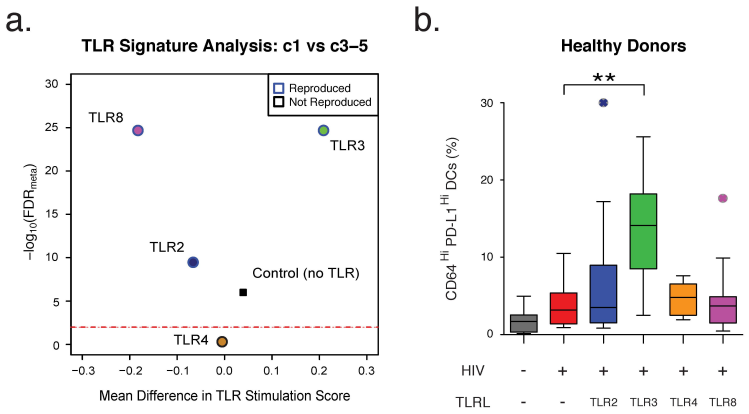


Figure 38: **(a)** Volcano plot of meta-analysis $-\log(\text{FDR})$ v. mean difference in TLR stimulation score between c1 and c3-5. Scores are computed from weighted correlations between single-cell profiles and transcriptional patterns from human DCs after 48 hours of stimulation with media control (black) or agonists for either TLR2 (Pam, dark blue), TLR3 (Poly I:C, green), TLR4 (LPS, orange), TLR7/8 (Gard, purple), or TLR9 (CpG, light blue). Tests reproduced with FDR < 0.01 in both stratified analyses are highlighted in blue. **(b)** Proportion of CD64^{Hi},PDL1^{Hi} cells among mDCs from PBMCs isolated from HIV-1-negative individuals cultured in the absence or the presence of VSV-G pseudotyped HIV-1, alone or in combination with TLR ligands (TLRL: TLR2L, PGNA, n = 11; TLR3L, Poly I:C, n = 11; TLR4L, LPS, n=8; TLR8L, CL097, n = 11). Statistical significance was calculated using KW and Dunn's tests (**, $P < 0.01$).

stranded (ds)RNA sensor TLR3 might selectively activate downstream pathways that synergize with innate viral sensing mechanisms to increase the fraction of mDCs maturing towards a c1-enriched/CD64^{Hi},PD-L1^{Hi} phenotype.

To directly test this hypothesis, collaborators incubated PBMCs from several HDs (n=7) – which do not spontaneously generate significant numbers of c1-enriched/CD64^{Hi},PD-L1^{Hi} cells *in vitro* in the presence of VSV-G pseudotyped HIV-1 (Figure 36a) – with virus and different TLR agonists for 24 hours. In contrast to the other TLR ligands tested, I observed that co-incubation of mDCs with virus and Poly I:C led to a significant increase in the proportion of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs in PBMCs from healthy individuals (TLR3L: P-value=0.0091, n=11; KW and post-hoc Dunn's test; TLR2L, TLR4L, and TLR8L, not significant; n=11, 8, 11, respectively) (Figure 38b).

To explore the generality and therapeutic applicability of this adjuvant strategy, we next examined whether we could couple the same TLR3 activation with direct DNA-based targeting of the cytosolic innate immune recognition machinery that senses viral DNA products [170] rather than use the virus itself. To address this, my collaborators incubated PBMCs from HDs or ECs simultaneously with a TLR3 agonist (Poly I:C) and single-stranded (ss)- or ds HIV-1 Gag DNA (ssDNA or dsDNA, respectively) encapsulated in polymeric nanoparticles. A similar delivery vehicle has previously been shown to selectively activate cGAS- and STING-dependent immune recognition pathways, which are involved in innate immune sensing of HIV-1 during natural infection [171]. When we analyzed the fraction of mDCs differentiating into c1-enriched/CD64^{Hi},PD-L1^{Hi} cells, we found that activation with either ss/dsDNA or Poly I:C (TLR3 agonist) alone in PBMCs from HDs was less efficient at inducing c1-enriched responses (P-value=7×10^{-2}, nano v. Poly I:C alone; P-value=5×10^{-2}, nano v. ssDNA; P-value=1×10^{-2}, nano v. dsDNA; two-tailed Wilcoxon signed-rank test; n=8; Figure 39, comparisons not highlighted).

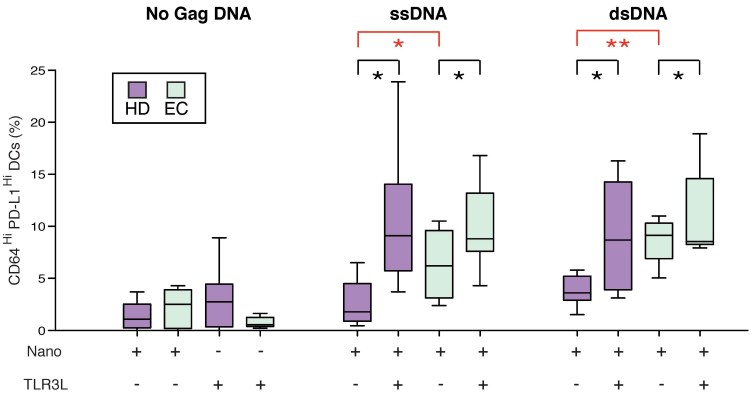


Figure 39: Proportions of CD64^{Hi}, PD-L1^{Hi} cells among mDCs from healthy individuals (indigo) and elite controllers (olive) cultured in the absence or the presence of Poly I:C and polymer nanoparticles loaded with ss or ds 100 nucleotide HIV-1 DNA (n = 8, HIV-1 negative individuals; n = 7, ECs). Statistical significance was calculated using either two-tailed Wilcoxon signed-rank test (black) or two-tailed MWW test (red) to compare differences within or between donor groups, respectively (**, $P < 0.01$; *, $P < 0.05$).

Combining both stimuli, however, significantly increased the proportion of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs in PBMCs isolated from HDs (P-value=1.6×10^{-2} and P-value=3.1×10^{-2} for ss- and dsDNA, respectively; two-tailed Wilcoxon signed-rank test; n=8; Figure 39). Similar results were obtained with cells from ECs (P-value=0.0469 for both ss- and dsDNA; two-tailed Wilcoxon signed-rank test; n=7; Figure 39), with the notable exception that, in ECs, exposure to dsDNA alone led to significantly higher levels of c1-like/CD64^{Hi},PD-L1^{Hi} mDCs relative to cells cultured only in media (P-value=3×10^{-2}; Wilcoxon signed-rank test; n=7; Figure 39, comparison not highlighted), suggesting a heightened baseline predisposition of EC to respond to intracellular DNA.

In mixed leukocyte reactions, the CD64^{Hi},PD-L1^{Hi} mDCs generated from HDs incubated with TLRL3 and nanoparticles containing gag dsDNA stimulated greater proliferation in CD4$^+$ and CD8$^+$ T cells compared to the CD64^{Lo},PD-L1^{Lo} mDCs from the same assay (P-value=3.5×10^{-2} and P-value=3.1×10^{-2}, respectively; two-tailed Wilcoxon signed-rank test; n=6), suggesting that adjuvant induced CD64^{Hi},PD-L1^{Hi} mDCs in HDs are highly functional antigen presenting cells like their EC counterparts (Figure 40).

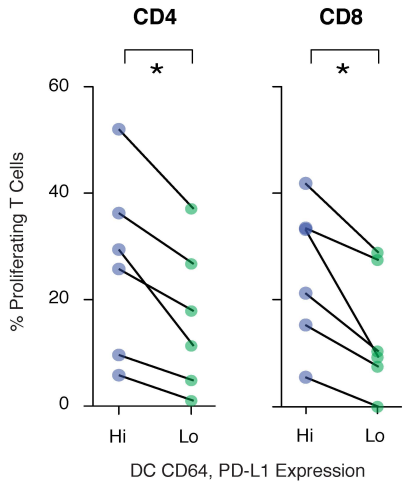


Figure 40: Proportion of proliferating CD4$^+$ or CD8$^+$ T cells after culture with Hi or Lo mDC from a HD stimulated with TLRL3 and nanoparticles containing gag ssDNA (*, $P < 0.05$; two-tailed Wilcoxon signed-rank test, $n = 6$).

6.8 REPRODUCIBLE DIFFERENTIAL SIGNATURE ANALYSIS

To identify additional nodes for rationally modulating the acquisition of the c1 functional state, as well as to examine the general applicability of the IDR-framework for uncovering putative regulators of c1's (or any other state's) induction, I again applied the differential signature module of `scRAD`; in this instance, due to limited public availability of human perturbation data, I turned to a published data set of the transcriptional effect of ~200 transcription factor and signaling molecule perturbations in LPS-stimulated mouse DCs – a pathway that is highly conserved with humans [149, 161].

Publicly available and normalized nCounter population data were mapped to unique human homologs, log-scaled and centered. Correlation-based signature scores were computed as in the TLR analysis above for each shRNA experiment. I applied the `scRAD::kruskalIDRm` analysis as in the DE analysis, defining IDR < 0.05 as my threshold for calling differential signatures. I ranked the perturbations by the degree to which they reproducibly favored the generation of one or more (here, c1) responses over others (here, c3-5). The resulting meta-analysis nominated several putative regulators for modulating the fractional abundance of c1 mDCs in response to a virus or virus-like stimulation (Figure 41a).

Among my top positive regulators of c1 was TBK1, a recognized signal mediator that is activated downstream of multiple innate immune sensing pathways at the convergence of the organelle-associated adaptors MAVS, TRIF (downstream effector of TLR3, TLR4), and STING (effector of the intracellular DNA sensor cGAS) [172–174], some of which were previously detected in my `IPA` Upstream Analysis (Figure 32). Notably, the cGAS-STING pathway is known to play a key role in the recognition of cytoplasmic HIV-1 DNA in myeloid cells, including those from ECs [151, 170], and cGAS itself (MB21D1) was up-regulated in c1 cells (lfc=1.9, IDR < 0.05). To evaluate whether signaling through TBK1 significantly contributes to the maturation of mDCs into the c1-enriched/CD64^{Hi},PD-L1^{Hi} subset in ECs, my collaborator added BX795, a TBK-1 antagonist, to PBMCs from ECs at the time of viral addition and examined the impact on mDC responses. As shown in Figure 41b, inhibition of TBK1 during viral exposure led to a dramatic and significant abrogation of the induction of the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC population in ECs (P-value=2.0×10^{-3}; two-tailed Wilcoxon signed-rank test;

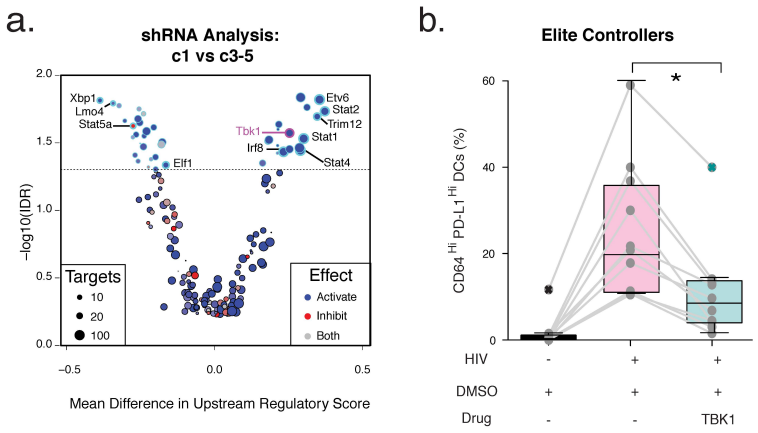


Figure 41: **(a)** Volcano plot of $-\log(\text{IDR})$ in upstream regulatory score between c1 and c3-5 based on single-cell correlations with shRNA-perturbation profiles from mouse DCs stimulated with LPS for 6 hours (adapted from Chevrier et al. [149]). The net effect (activate, inhibit, both) of each perturbation is denoted by color (red, blue, gray, respectively), as is its breadth (size). **(b)** Proportions of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ cells among EC mDCs cultured in the presence or absence of virus and DMSO (control, magenta) or BX795 TBK1 inhibitor (cyan; $n = 10$). Statistical significance was calculated using a two-tailed Wilcoxon signed-rank test (*, $P < 0.05$)

n=10), suggesting that TBK1 is a key driver of the acquisition of the c1 phenotype in mDCs and validating the promise of my computational framework.

6.9 CONCLUSIONS

In summary, by studying elite immune control of HIV-1 infection across donors that are linked by a common enhanced immunity, I identified an mDC response state that displays gene expression features consistent with profound functional activation and heightened antiviral activity. This subset of mDCs, enriched among cells expressing the surface molecules PD-L1 and CD64, is:

1. induced more efficiently in ECs than in HIV-1 CPs or HDs after *in vitro* viral exposure

2. associated with both higher $CD4^+$ T cell counts and lower HIV-1 viral loads

3. more effective at stimulating T cell proliferation *in vitro*

4. more efficient at inducing HIV-1-specific polyfunctional cytotoxic $CD8^+$ T cells

All of these are canonical correlates of antiviral immunity in EC [175]. By leveraging `scRAD` to re-examine publicly available transcriptome data sets, I identified key regulatory molecules and adjuvants for modulating the acquisition of this functional mDC response state in the general population, with potential therapeutic and prophylactic implications. This example illustrates how `scRAD` tools can be applied to a wide variety of common scRNA-seq analyses and derive robustness from a reliance on multiple donors.

The heterogeneity of mDC responses identified in my study should invoke recent work by Villani et al. [71] that describes at least four subsets of circulating mDC in HDs. Interestingly, my c1-mDC response state shares important characteristics with the DC4 ($CD11c^+MHC\text{-}II^+CD1C^-CD141^-CD16^+$) subset described in that work, exhibiting its characteristic antiviral signature as well as reproducible up-regulation of all five published marker genes [71]. Given the dissimilarities between cohorts and experimental conditions, future studies will be required to fully elucidate the functional and transcriptional relationships between these mDC groups and their ontogeny.

7

DONORS WITH HETEROGENOUS PHENOTYPES

7.1 INTRODUCTION

To conclude my dissertation I will summarize findings from an ongoing analysis in partnership with collaborators[1]. My collaborators have performed scRNA-seq on total unsorted cerebrospinal fluid (CSF) cells from multiple sclerosis (MS) donors and matched controls, as part of a single-cell case–control analysis. Outside of the field of cancer [176], there are only a handful of studies that utilize scRNA-seq technology to compare tissue samples from disease-affected donors against those of separate control donors in a clinically relevant setting [177, 178]. Therefore, our study illustrates a rare and exciting mode of scRNA-seq analysis, and presents new analytical challenges and opportunities for discovery. The goals of this analysis are similar to goals of the EC DC study in Chapter 6, including

1. characterization of cell states,
2. analysis of composition differences,
3. differential expression analysis,

where differences are defined between groups of donors rather than in terms of a within-donor response phenotype (i. e. viral stimulation).

I have employed a scRNA-seq analysis pipeline that takes advantage of my own computational modules `scone` and `scRAD`, among others (Figure 42). My pipeline uses `scone` to select a normalization procedure for removing donor-specific effects from downstream clustering and signature analyses. DE modeling incorporates factors of unwanted variation prioritized by `scone`

1 This chapter is based on ongoing work in collaboration with David Schafflick, Maike Hartlehnert, Tobias Lautwein, Konrad Buscher, Jolien Wolbert, Sven G. Meuth, Mark Stettner, Christoph Kleinschnitz, Tanja Kuhlmann, Catharina Gross, Heinz Wiendl, Nir Yosef, and Gerd Meyer zu Horste.

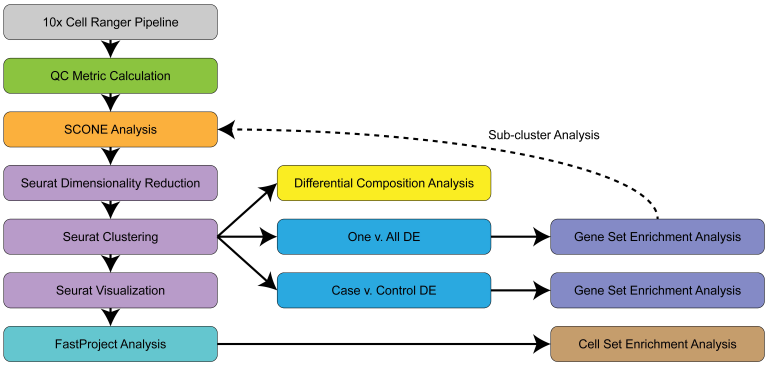


Figure 42: *scRNA-seq case–control analysis pipeline.* Scheme depicting the scRNA-seq analysis workflow utilized in this study. Analysis begins with 10x `Cell Ranger` processing and cell-level QC metric evaluation, followed by `scone` data filtering and normalization, `Seurat` dimensionality reduction, clustering and visualization. Results from these analyses are input into *FastProject* `VISION` for signature calculation and consistency testing. These signatures may be used for CSEA testing. Differential abundance analysis is performed based on the `Seurat` clustering, and various forms of differential expression testing, including one v. all, "marker" analysis and cluster-specific case v. control analysis are performed using a meta-analysis approach that supports IDR modeling with `scRAD` tools. GSEA testing is used to ascribe biologic meaning to differential expression results, motivating further subclustering analysis, in which a cluster is analyzed using an identical analytical procedure.

but respects the original count data context. I have adopted a DE meta-analysis approach to accommodate reproducibility assessments with `scRAD` tools. Below I will step through an outline of the pipeline workflow, applied in two stages: I will first describe a computational analysis of all CSF cells, followed by a focused analysis on a cell subset defined by a shared cell-level phenotype, i. e. $CD4^+$ cells. The cells in this subcluster analysis exhibit a continuum phenotype that isn't well characterized by traditional clustering analyses. For this reason I have developed a new single-cell analysis method, cell set enrichment analysis (CSEA), based on the popular gene set enrichment analysis (GSEA) test of Subramanian et al. [179]: CSEA provides a statistical test for enrichments of cell groups (e. g. MS-derived cells) in high or low signature tails.

7.2 STUDY SAMPLE

MS is a phenotypically heterogenous disease; our interpretation of case–control differences hinges on the definition of "cases." This study considers six treatment-naive MS donors with either i) clinically isolated syndrome (CIS) indicative of MS or ii) a first diagnosis of relapsing-remitting MS. For simplicity, I refer to all of these subjects as MS donors. Lumbar punctures are necessary for CSF sample collection but cannot be performed in healthy volunteers. My collaborators recruited six non-MS donors with idiopathic intracranial hypertension (IIH) for which normal CSF samples could be obtained in compliance with ethical standards. All donors gave written informed consent. The study was performed in accordance with the declaration of Helsinki and approved by local ethics committees. All shared data was anonymized to protect donor identities.

The barcoded single-cell mRNA libraries were constructed and cells were called using the 10x Chromium platform. Two samples (donors) were disqualified at an early stage due to evidence of substantial contamination. The remaining five MS and five IIH samples were aggregated via `Cell Ranger` read downsampling, guaranteeing that the average number of reads per cell barcode were uniform across samples (Table 12).

Subject ID	Total barcodes	Mean UMIs	Mean species
MS1	4,098	3,752.788	1,079.8538
MS2	417	3,416.008	1,059.5766
MS3	2,426	2,033.123	810.7696
MS4	3,005	3,165.410	981.8682
MS5	920	3,908.745	1,061.7293
IIH1	3,498	3,591.566	1,008.1838
IIH2	1,106	3,848.090	973.9295
IIH3	8,129	3,630.284	1,032.6788
IIH4	1,218	3,896.800	1,033.3760
IIH5	1,099	3,464.908	902.2602

Table 12: *Summary of aggregated 10x `Cell Ranger` filtered output* The raw UMI matrix contains 33,694 gene-level features for 26,916 barcodes. Mean reads were standardized by aggregation at 20,750 ± 2.

7.3 QC METRIC CALCULATION

In the interest of applying the `scone` workflow, I extracted all QC metrics listed in Table 10 from the `Cell Ranger` filtered molecule info files. All of these metrics are summarized below in Table 13. Barcodes are associated

QC metric	Mean±SD
num_umi	$6 \pm 4 \times 10^3$
num_reads	$4 \pm 3 \times 10^4$
mean_reads_per_umi	7 ± 1.5
std_reads_per_umi	5 ± 1
mapped_reads	0.5 ± 0.1
genome_not_gene	0.34 ± 0.06
unmapped_reads	0.16 ± 0.055
umi_corrected	0.009 ± 0.005
barcode_corrected	0.03 ± 0.02

Table 13: *Summary of QC metrics based on 10x `Cell Ranger` unfiltered output*

with over a thousand UMIs on average, at ~7 reads per UMI.

7.4 DATA FILTERING

I filtered genes and cells using a scheme similar to the one described in Chapter 4, involving a new third step.

1. Define *common genes* based on UMI counts. Genes with n_u or more UMIs in at least 25% of barcodes, where n_u is the UQ of the non-zero elements of the UMIs matrix.

2. Filter cells based on QC metric. Remove cells with low numbers of reads ("num_reads" metric), low proportions of mapped reads ("mapped_reads" metric), or low numbers of detected common genes (Figure 43). The threshold for each measure is defined data-adaptively: A cell may fail any criterion if the associated metric under-performs by z_{cut} standard deviations from the mean metric value or by z_{cut} median absolute deviations from the median metric value. Here I have used $z_{cut} = 2$.

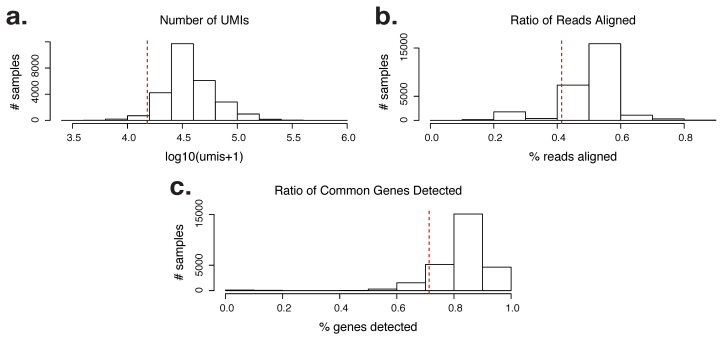


Figure 43: *Distributions of single-cell filtering metrics.* Red lines represent adaptive threshold below which all cells were removed from further analysis. **(a)** Distribution of number of reads per barcoded library. **(b)** Distribution of transcriptome read alignment ratio per library. **(c)** Distribution of the fraction of common genes detected per library.

3. Remove barcodes from donors with fewer than 100 barcodes following cell filtering. These donors have contributed too few high-quality cells to reliably estimate donor-specific effects. Only seven cells were removed in this step.

4. Filter genes based on UMI counts: Genes with n_u or more UMIs in at least n_s barcodes, where n_u is the UQ of the non-zero elements of the cell-filtered UMIs matrix. I have set $n_s = 5$ to accommodate markers of rare populations. This substep ensures that included genes are detected in a sufficient number of cells after cell filtering.

A summary of the filtered data can be found in Table 14. After filtering, only 4 MS samples and 4 IIH samples remain.

7.5 NORMALIZATION

I utilized the `scone` package [126] (Chapter 4) to select an appropriate normalization based on a standardized panel of performance criteria. Clustering and correlation evaluations were PCA-based, using ten PCs.

SCALING NORMALIZATION. I considered a number of scaling methods with the `scone` package, including: no normalization, TC normalization, TMM

Subject ID	Total barcodes	Mean UMIs	Mean species
MS1	4,000	3,743.200	1,057.1047
MS2	1,278	3,558.925	1,073.4570
MS3	1,739	2,323.858	873.3197
MS4	2,635	3,393.682	1,016.5905
MS5	0	NA	NA
IIH1	2,931	3,968.076	1,069.8857
IIH2	977	4,102.382	1,029.3603
IIH3	7,843	3,666.721	1,018.2217
IIH4	0	NA	NA
IIH5	954	3,735.062	953.2537

Table 14: *Summary of aggregated 10x `Cell Ranger` filtered output, following `scone` cell and gene filtering* The raw UMI matrix contains 10,267 gene-level features for 22,357 barcodes. Spread of mean reads per barcode increased considerably after filtering at 22,000 ± 1,000.

normalization, UQ normalization, FQ normalization, and RLE normalization.

CATEGORICAL COVARIATES. I also considered normalization procedures that include a linear regression-based batch adjustment for log-transformed expression data. Donor ID was treated as a batch covariate. Normalized UMI matrices were scored for batch mixing using the scone batch ASW score. I also monitored the silhouette score of case–control status, although I never explicitly included this categorical biological covariate as part of the adjustment model. The stratified PAM argument was applied to the evaluation of *de novo* PAM clusters, considering a range of K from 2 to 8.

CONTROL GENES. Positive controls were selected from the top 500 most common gene symbols referenced in the MSigDB C7 collection of immunological signatures. Negative controls were selected from Eisenberg and Levanon [180]. In order to match sets for mean expression, genes were binned according to the rounded mean $\log_2$-expression (adding 1 to each observation). Genes for the positive control set, and two negative control gene sets (adjustment and evaluation) were drawn in equal numbers (max) from each expression bin, for a total of 207 genes each.

UNWANTED VARIATION. qPC-based adjustment were based on the matrix of QC metrics discussed above. Both RUVg and qPC adjustments were performed over a range of 0 to 8 factors.

SELECTED NORMALIZATIONS. The top performing normalization as ranked by `scone` involve RLE scaling, qPC-based adjustment, and batch adjustment. This normalization included all eight qPCs.

7.6 SEURAT ANALYSIS

After cell filtering, I loaded the normalized log-transformed UMI matrix into the popular `Seurat` analysis pipeline [136]. Following data scaling and PCA, I clustered the cells in the first ten PCs using the `Seurat::FindClusters` function. Clustering resolution was set to 0.6. `Seurat` identified over 10 clusters, but these were manually collapsed during annotation (Figure 44). tSNE data representations were computed using that fast option in `Seurat::RunTSNE`.

7.7 VISION ANALYSIS

I passed raw and normalized UMI data to the `VISION` R workflow[2], an updated implementation of *FastProject* tools. Before running `VISION`, I computed the mean expression per gene symbol per cell in order to make the gene features relatable to signatures based on gene symbols. The goal of *FastProject* analysis is to uncover biologically meaningful gene signatures that vary coherently across single-cell neighborhoods [153]. These signatures can help assign meaning to the dominant expression differences between clusters. In addition to raw data, I passed QC, donor, status, and `Seurat` cluster covariates for exploratory analysis and visualization. I also included the `Seurat` tSNE as a precomputed projection.

My signature set includes:

- Human cell cycle genes from Macosko et al. [80], representing sets of genes marking G1/S, S, G2/M, M, and M/G1 phases.
- The MSigDB C7 immunological signature collection.
- helper T (T_H) signatures compiled in Lönnberg et al. [181].

2 Available for download at `https://github.com/YosefLab/VISION`

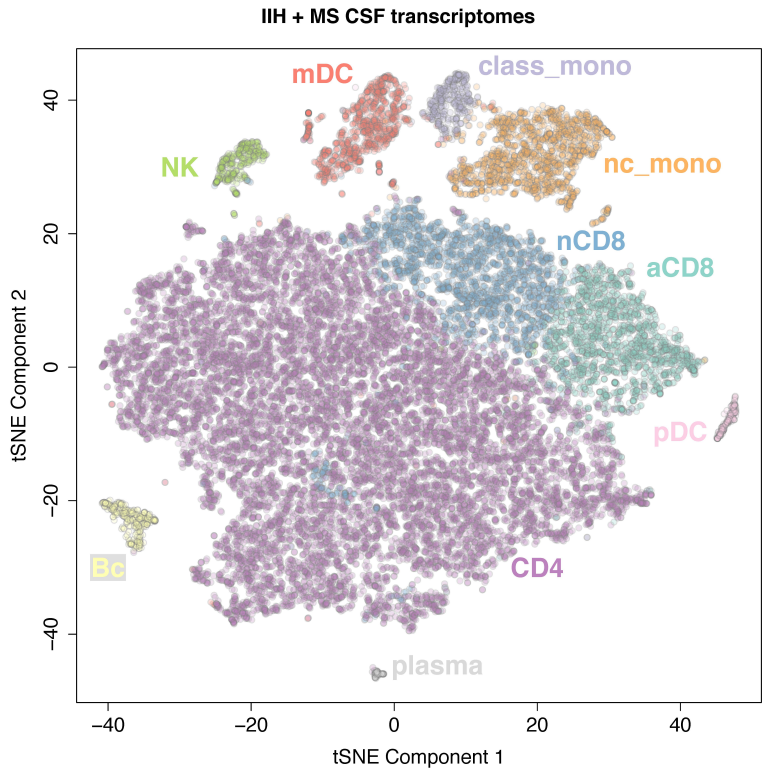


Figure 44: *`Seurat` clustering analysis of CSF cells.* tSNE plot of 10 cell clusters identified by scRNA-seq after quality control filtering and normalization in 22,357 total merged IIH- (n = 4) and MS-derived (n = 4) CSF cells. Cluster identity was manually assigned based on marker gene expression.

- NetPath database signatures [182].
- Curated T cell signatures used in Gaublomme et al. [83].
- follicular helper T (T_{FH}) marker genes derived from Crotty [183] and Liu et al. [184].

Housekeeping genes were referenced from the same source as the `scone` negative controls above [180].

In addition to data exploration, signature values calculated per cell may be used for downstream analyses such as CSEA, below.

7.8 ONE V. ALL DE

In order annotate clusters, I performed one v. all comparisons following each clustering analysis. One v. all DE tests P-values were used to rank genes by the extent they are up-regulated in one cluster over all others. Tests were performed separately for each donor sample with at least 10 cells in the target cluster. qPC factors used for normalization above were incorporated into a linear predictor for DE testing with `limma` `voom`. Results for each donor sample were combined in multiple ways, calculating median lfcs, meta-analysis P-values for one-sided tests using Stouffer's method, and IDRs for two-sided tests using the `scRAD::est.IDRm` [81] (Chapter 5) for all genes and comparisons.

These comparisons highlighted a number of genes that mark known cell types (Figure 45), including monocytes, DCs, natural killer (NK) cells, B cells, and T cells.

7.9 DIFFERENTIAL COMPOSITION ANALYSIS

I used the `limma` package to test for differential `Seurat` cluster log-abundance between MS and IIH. I observed significant excesses in natural killer (NK) cells, B cells, and plasma cells, reflecting a unique immune signature of MS in the CSF.

7.10 CLUSTER-SPECIFIC DISEASE SIGNATURES

I also considered cluster-specific case–control differences, similar to HIV-1 v. media comparison for c1 DCs (Chapter 6). Donors were only included in a comparison if 10 or more cells from the target cluster were detected in the donor's sample. All pairings of MS donors with control donors were considered (up to 16). For each valid case–control pair, DE analysis was performed using `limma` with `voom`, as in the marker analysis, but comparing case cells against control cells. lfc was summarized by the median of lfcs estimated across the donor pairs. Meta-analysis was performed on all possible pairings of cases and controls (up to $4! = 24$); the median meta-analysis P-value was reported. IDR modeling was applied at the pair level, modeling the reproducibility of up to 16 replicate significance signals. Some genes are very

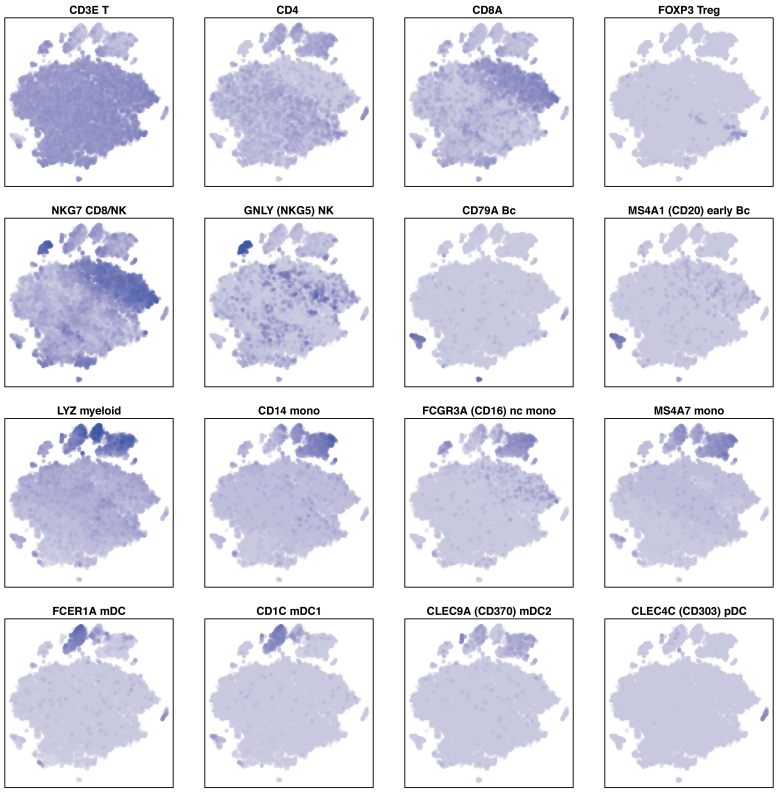


Figure 45: *CSF marker expression.* Feature plots for 16 marker genes highlighted in one v. all DE tests. Plots are labels by gene symbol and marked cell type. Darker blue indicates higher expression.

lowly expressed across individual clusters, resulting in unstable statistical estimation for those genes. Genes were filtered before DE if they had mean un-normalized UMI counts below 0.05.

7.11 GSEA

After deriving important lists of genes (e. g. marker genes or disease signature genes), I may seek to uncover enrichment for particular gene sets in order to capture important biological differences between cells. But gene sets compiled in the literature can often be large and over-inclusive.

GSEA is a hypothesis testing method for simultaneously uncovering enrichments and identifying subsets of gene sets of importance [179]. In this section I will describe this method in specific detail so that I can build on it in

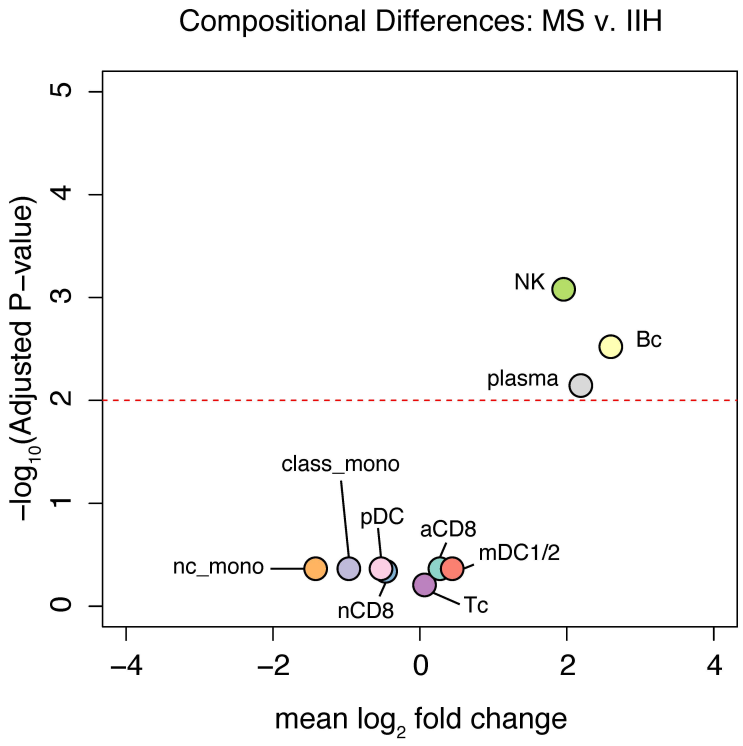


Figure 46: *scRNA-seq differential composition analysis* Volcano plot representing differential abundance of cell types in MS v. IIH. Bonferroni adjustment controls the FWER of P-values reported by `limma`.

the discussion of CSEA below. The input to GSEA is a list of N genes, rank-ordered by some input signal (e. g. log fold change, log transformed P-value). Using a similar notation as Subramanian et al. [179], I will let σ_j denote the gene j's signal; indices have been sorted so that $\sigma_j > \sigma_{j+1}$ (alternatively in decreasing order: $\sigma_j < \sigma_{j+1}$). The test involves considering all genes up to a specific position, i. A "hit" score is defined as the cumulative sum of signal magnitudes (optionally exponentiated by parameter p: $|\sigma_j|^p$) for members of gene set S, divided by the sum over all set members in the list.

$$\begin{aligned} P_{\text{hit}}(S, i) &= \sum_{j}^{j \leq i} \frac{\mathrm{I}\left(j \in S\right)|\sigma_j|^p}{N_R} \\ N_R &= \sum_{j} \mathrm{I}\left(j \in S\right)|\sigma_j|^p \end{aligned} \tag{41}$$

A "miss" score is similarly calculated for non-members of S, but without weighing by signal magnitudes.

$$\begin{aligned} P_{\text{miss}}(S,i) &= \sum_{j}^{j \leq i} \frac{\mathrm{I}(j \notin S)}{N - N_H} \\ N_H &= \sum_{j} \mathrm{I}(j \in S) \end{aligned} \tag{42}$$

The GSEA enrichment score (ES) is defined as the maximum of $P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i)$, with respect to index i. When $p = 0$, the ES reduces to a one-sided KS test statistic. Subramanian et al. [179] recommend simulating a null distribution for ES at the signal-level (e. g. recomputing lfcs for shuffled cell labels): For a random S, ES should be small, but if the list is concentrated at the top of the list, ES will be close to 1. Unfortunately, this particular permutation approach is often impractical when calculating the underlying score is costly. In my analyses I generated null distributions of ESs by shuffling S memberships, assigning empirical one-sided P-values based on simulation [185].

For $p \neq 0$, GSEA cannot be seen as a simple rank-based enrichment test: GSEA tests for enrichment of a gene set at the high tail (or low tail) of the signal distribution, but additionally weighs the set elements according to their signature value. This reduces the effects of low-magnitude genes in S, whereas all genes not in S are treated the same no matter the magnitude of their signal. GSEA tests if high magnitude (positive or negative) genes are enriched at a specific tail, applying permutation tests to account for the additional variability induced by the magnitude weights. The set of indices up to where $P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i)$ reaches its maximum also holds significance - referred to in Subramanian et al. [179] as the *leading-edge* of the enrichment test. The intersection of the set S and the leading-edge is the *leading-edge subset*, representing an important core subset of genes driving the enrichment.

I apply GSEA tests to all DE tests above – using signed significance scores based on meta-analysis P-values as gene signals – applying the Bonferroni adjustment to control FWER for each category of hypotheses.

While enrichments for one v. all comparisons were surely helpful in annotating and merging `Seurat` clusters, enrichments for disease signature genes were far more interesting, reflecting ongoing immune cell activation in the CSF of MS donors. For example, disease signature genes up-regulated in CD4$^+$ cells and monocytes are enriched for several immunity related Gene Ontology (GO) terms, including immune response and immune defense. The monocyte disease signature set also showed enrichment for antigen process-

ing, antigen presentation, and endocytosis. Monocytes and CD8$^+$ cells were enriched in housekeeping genes, consistent with activation. Finally, NK cells exhibited specific enrichments in genes associated with oxidative phosphorylation.

7.12 SUBCLUSTER ANALYSIS

CD4$^+$ T cells represent a critical piece of the adaptive immune system, and they also make up most of the scRNA-seq data set. In the interest of centering my analysis on these important cells, I subset all cells in the CD4$^+$ T cell cluster and re-analyzed them. This involved running `scone` a second time: the same normalization procedure was selected, except only four qPCs were recommended – a less intrusive normalization. I subsequently ran a `Seurat` analysis on the subset, defining CD4$^+$ T cell subclusters and identifying computational contaminants (Figure 47).

`VISION` analysis highlights several biologically meaningful signatures, corresponding to known differences between T cell subsets (e. g. naive v. memory). Nevertheless, the clusters were not easily annotated based on these and other conventional observations (e. g. DE results), with the exception of one cluster of regulatory T (T_{Reg}) cells. Furthermore, only one, unannotated cluster exhibited significant excesses in MS (Figure 48). My collaborators and I speculated that this conventional approach would be insensitive to gene signatures or cell states that are poorly represented by tight clustering. I therefore developed a new technique – CSEA – applying the GSEA testing procedure to ranked cell lists.

7.13 CSEA

For each `VISION` signature, I treated the computed signature scores as cell signals σ_j in a transposed GSEA analysis where j now indexes over cells rather than genes. The sets under consideration in this new CSEA were the mutually exclusive sets of MS and control cells. The goal of CSEA is to identify core sets of cells that drive each biological condition's enrichment for high or low signature values (Figure 49). Contaminating subpopulations in the CD4$^+$ cluster were removed prior to CSEA.

There are generally two kinds of signatures highlighted by CSEA: the first captures coherent signatures (i. e. low `VISION` consistency P-values), for which

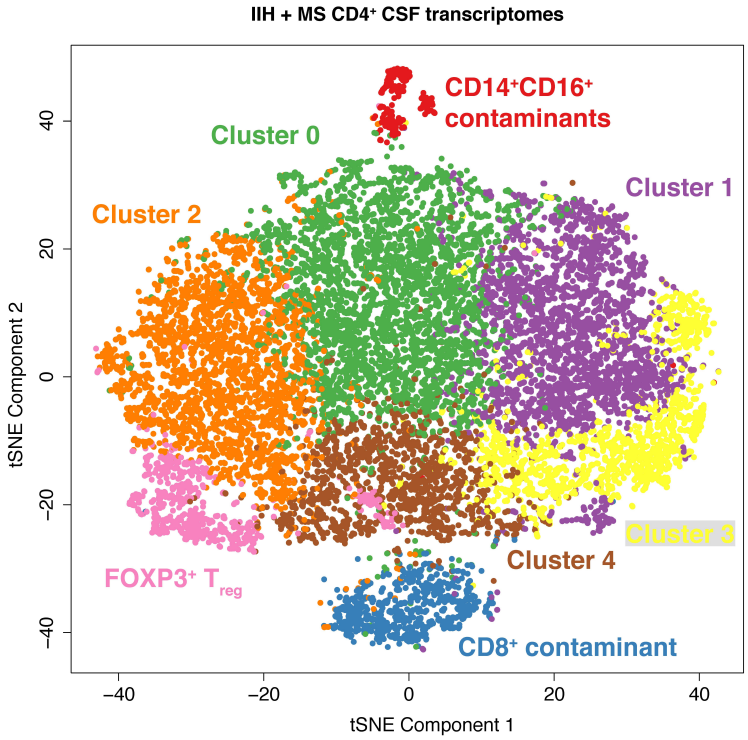


Figure 47: *`Seurat` clustering analysis of CD4⁺ CSF cells.* tSNE representing subclustering of single T cell expression profiles for $CD4^+$ CSF T cells pooled from 8 human donors, including 4 MS and 4 IIH controls.

one tail or the other is enriched for a biological condition (e. g. T_H1-like signature enriched for MS donor cells; Figure 50a). The second type of signature is incoherent with respect to the overall cell profile clustering, representative of a gene module rather than a cell type (e. g. M-phase cell-cycle signature enriched for MS donor cells; Figure 50b). This novel analytical approach decouples clustering of cells from disease-state signature enrichment, providing a new framework for interpreting complex scRNA-seq data sets.

7.14 CONCLUSIONS

The study above is just a snapshot of what is possible now that the field's understanding of scRNA-seq data has matured. Additional donor-level covariates (e. g. ancestry, age, treatment, disease phenotypes) will need to be inte-

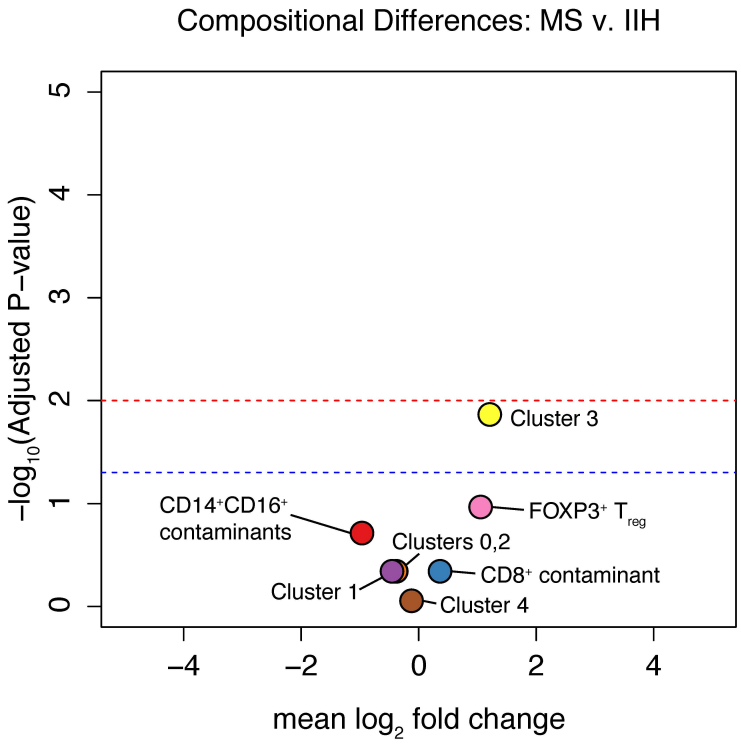
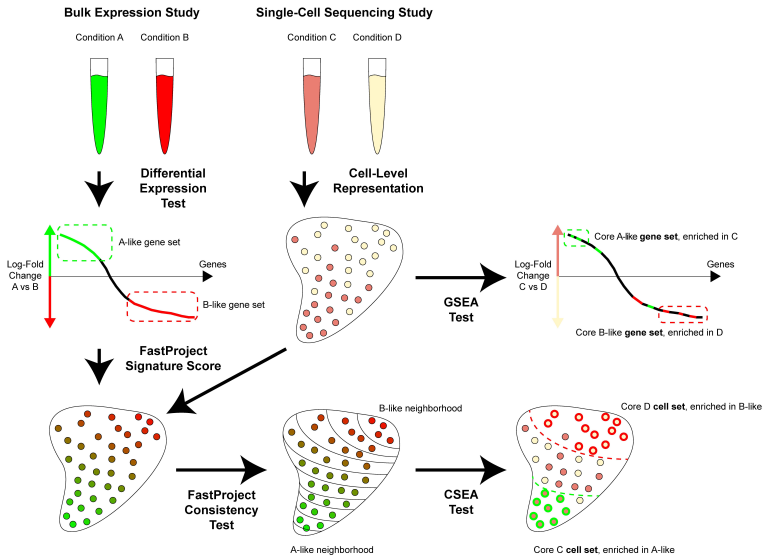


Figure 48: *scRNA-seq differential composition analysis for T cell subclusters.* Volcano plot representing differential abundance of CD4$^+$ T cell subtypes in MS v. IIH. Bonferroni adjustment controls the FWER of P-values reported by `limma`.

grated into these analyses as the numbers of donors increase. Unwanted variation factors will need to be monitored for confounding and removed from the data via data normalization methods. Donor-level biological covariates may be correlated against single-cell measurements to generate interesting new disease hypotheses, and multi-omic technologies promise to challenge are understanding of condition-specific cell-level associations. Single-cell data may also be used to derive quantitative phenotypes of interest, for use as covariates in analyses of other parts of the data: e. g. how does the distribution of T cell substates correlate with the total abundance of B cells measured by scRNA-seq? One thing is for sure - scRNA-seq and other single-cell technologies will continue to provoke new and interesting questions – biological and theoretical – and motivate innovative methods development for years to come.



Publicly available bulk microarray or RNA-seq data are used to identify gene signature sets characterizing immune cell populations. These gene sets are used for either i) GSEA of our scRNA-seq differential expression results or ii) single-cell `VISION` signature scores, input to both `VISION` Consistency testing and CSEA testing.

Figure 49: *Scheme of GSEA/`VISION`/CSEA analysis.*

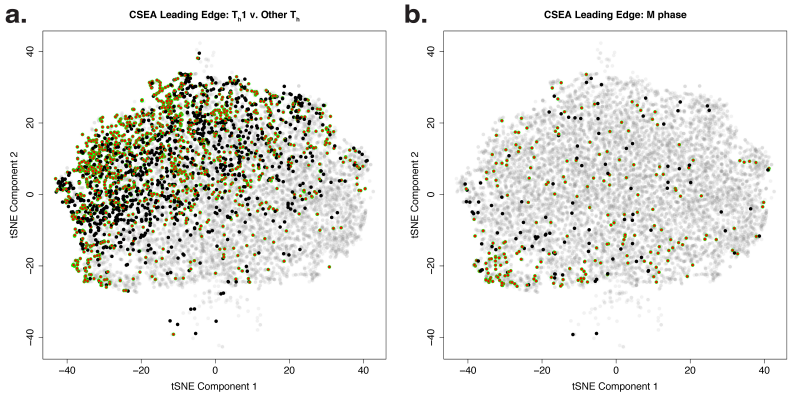


Figure 50: *CSEA examples.* Two very different examples of significant CSEA tests. In both cases, MS cells are enriched in the upper tail of the signature distribution. **(a)** This signature measures the extent a cell profile resembles T_H1 v. other T_H cell types. **(b)** This signature measures the extent to which M-phase specific genes are expressed in a cell. Red points with green outline are the core MS set, black cells are IIH members of the leading edge cell set.

PART IV

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. "MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS". *Annals of Applied Statistics* **5**:3 (2011), 1752–1779.

[2] Jean-Laurent Casanova. "Human genetic basis of interindividual variability in the course of infection". *Proceedings of the National Academy of Sciences of the United States of America* **112**:51 (2015), E7118–E7127.

[3] Joel N. Blankson. "Effector mechanisms in HIV-1 infected elite controllers: Highly active immune responses?" *Antiviral Research* **85**:1 (2010), 295–302.

[4] Asier Saez-Cirion, Christine Lacabaratz, Olivier Lambotte, Pierre Versmisse, Alejandra Urrutia, Faroudy Boufassa, Francoise Barre-Sinoussi, Jean-Francois Delfraissy, Martine Sinet, Gianfranco Pancino, Alain Venet, and Sida Agence Nationale Recherches. "HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype". *Proceedings of the National Academy of Sciences of the United States of America* **104**:16 (2007), 6776–6781.

[5] R. Liu, W. A. Paxton, S. Choe, D. Ceradini, S. R. Martin, R. Horuk, M. E. MacDonald, H. Stuhlmann, R. A. Koup, and N. R. Landau. "Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection". *Cell* **86**:3 (1996), 367–377.

[6] P. Kiepiela, A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, S. Chetty, P. Rathnavalu, C. Moore, K. J. Pfafferott, L. Hilton, P. Zimbwa, S. Moore, T. Allen, C. Brander, M. M. Addo, M. Altfeld, I. James, S. Mallal, M. Bunce, L. D. Barber, J. Szinger, C. Day, P. Klenerman, J. Mullins, B. Korber, H. M. Coovadia, B. D. Walker, and P. J. R. Goulder. "Dominant influence of HLA-B in mediating the potential coevolution of HIV and HLA". *Nature* **432**:7018 (2004), 769–774.

[7] X. J. Gao, A. Bashirova, A. K. N. Iversen, J. Phair, J. J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, M. Altfeld, S. J. O'Brien, and M. Carrington. "AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis". *Nature Medicine* **11**:12 (2005), 1290–1292.

[8] E. Theander, R. Manthorpe, and L. T. H. Jacobsson. "Mortality and causes of death in primary Sjogren's syndrome - A prospective cohort study". *Arthritis and Rheumatism* **50**:4 (2004), 1262–1269.

[9] Manuel Ramos-Casals, Pilar Brito-Zeron, Antoni Siso-Almirall, and Xavier Bosch. "PRACTICE POINTER Primary Sjogren syndrome". *British Medical Journal* **344**: (2012).

[10] Athanasios G. Tzioufas, Efstathia K. Kapsogeorgou, and Haralampos M. Moutsopoulos. "Pathogenesis of Sjogren's syndrome: What we know and what we should learn". *Journal of Autoimmunity* **39**:1-2 (2012), 4–8.

[11] John A. Ice, He Li, Indra Adrianto, Paul Chee Lin, Jennifer A. Kelly, Courtney G. Montgomery, Christopher J. Lessard, and Kathy L. Moser. "Genetics of Sjogren's syndrome in the genome-wide association era". *Journal of Autoimmunity* **39**:1-2 (2012), 57–63.

[12] M. B. Cole, H. Quach, D. Quach, A. Baker, K. E. Taylor, L. F. Barcellos, and L. A. Criswell. "Epigenetic Signatures of Salivary Gland Inflammation in Sjogren's Syndrome". *Arthritis & Rheumatology* **68**:12 (2016), 2936–2944.

[13] Anura Hewagama and Bruce Richardson. "The genetics and epigenetics of autoimmune diseases". *Journal of Autoimmunity* **33**:1 (2009), 3–11.

[14] Carlo Selmi, Patrick S. C. Leung, David H. Sherr, Marilyn Diaz, Jennifer F. Nyland, Marc Monestier, Noel R. Rose, and M. Eric Gershwin. "Mechanisms of environmental influence on human autoimmunity: A national institute of environmental health sciences expert panel workshop". *Journal of Autoimmunity* **39**:4 (2012), 272–284.

[15] K. D. Robertson. "DNA methylation and human disease". *Nature Reviews Genetics* **6**:8 (2005), 597–610.

[16] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks. "A comprehensive overview of Infinium HumanMethylation450 data processing". *Briefings in Bioinformatics* **15**:6 (2014), 929–941.

[17] Yosra Thabet, Christelle Le Dantec, Ibtissem Ghedira, Valerie Devauchelle, Divi Cornec, Jacques-Olivier Pers, and Yves Renaudineau. "Epigenetic dysregulation in salivary glands from patients with primary Sjogren's syndrome may be ascribed to infiltrating B cells". *Journal of Autoimmunity* **41**: (2013), 175–181.

[18] Nezam Altorok, Patrick Coit, Travis Hughes, Kristi A. Koelsch, Donald U. Stone, Astrid Rasmussen, Lida Radfar, R. Hal Scofield, Kathy L. Sivils, A. Darise Farris, and Amr H. Sawalha. "Genome-Wide DNA Methylation Patterns in Naive CD4+T Cells From Patients With Primary Sjogren's Syndrome". *Arthritis & Rheumatology* **66**:3 (2014), 731–739.

[19] Corinne Miceli-Richard, Shu-Fang Wang-Renault, Saida Boudaoud, Florence Busato, Celine Lallemand, Kevin Bethune, Rakiba Belkhir, Gaetane Nocturne, Xavier Mariette, and Joerg Tost. "Overlap between differentially methylated DNA regions in blood B lymphocytes and genetic at-risk loci in primary Sjogren's syndrome". *Annals of the Rheumatic Diseases* **75**:5 (2016), 933–940.

[20] J. Imgenberg-Kreuz, J. K. Sandling, J. C. Almlof, J. Nordlund, L. Signer, K. B. Norheim, R. Omdal, L. Ronnblom, M. L. Eloranta, A. C. Syvanen, and G. Nordmark. "Genome-wide DNA methylation analysis in multiple tissues in primary Sjogren's syndrome reveals regulatory effects at interferon-induced genes". *Annals of the Rheumatic Diseases* **75**:11 (2016), 2029–2036.

[21] Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. "DNA methylation arrays as surrogate measures of cell mixture distribution". *Bmc Bioinformatics* **13**: (2012).

[22] B. Rhead, C. Holingue, M. Cole, X. R. Shao, H. L. Quach, D. Quach, K. Shah, E. Sinclair, J. Graf, T. Link, R. Harrison, E. Rahmani, E. Halperin, W. Wang, G. S. Firestein, L. F. Barcellos, and L. A. Criswell. "Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With Synoviocytes". *Arthritis & Rheumatology* **69**:3 (2017), 550–559.

[23] S. C. Shiboski, C. H. Shiboski, L. A. Criswell, A. N. Baer, S. Challacombe, H. Lanfranchi, M. Schiodt, H. Umehara, F. Vivino, Y. Zhao, Y. Dong, D. Greenspan, A. M. Heidenreich, P. Helin, B. Kirkham, K. Kitagawa, G. Larkin, M. Li, T. Lietman, J. Lindegaard, N. McNamara,

K. Sack, P. Shirlaw, S. Sugai, C. Vollenweider, J. Whitcher, A. Wu, S. Zhang, W. Zhang, J. S. Greenspan, T. E. Daniels, and Sicca Res Grp. "American College of Rheumatology classification criteria for Sjogren's syndrome: A data-driven, expert consensus approach in the Sjogren's International Collaborative Clinical Alliance Cohort". *Arthritis Care & Research* **64**:4 (2012), 475–487.

[24] Robert I Fox. "Sjögren's syndrome". *The Lancet* **366**:9482 (2005), 321–331.

[25] Lindsey A. Criswell, Kimberly E. Taylor, Quenna Wong, David M. Levine, Caitlin McHugh, Cathy Laurie, Kimberly Doheny, Mi Y. Lam, Alan N. Baer, Stephen Challacombe, Yi Dong, Hector Lanfranchi, Morten Schiodt, M. Srinivasan, Susumu Sugai, Hisanori Umehara, Frederick B. Vivino, Zhao Yan, Stephen Shiboski, Troy Daniels, John S. Greenspan, Caroline H. Shiboski, and Collaborative Sjogren's Syndrome. "The Genetic Basis of Sjogren's Syndrome (SS) Clinical Manifestations from Genome-Wide Association Analysis of Subphenotype Extremes in an International Cohort". *Arthritis & Rheumatology* **66**: (2014), S228–S229.

[26] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. "Principal components analysis corrects for stratification in genome-wide association studies". *Nature Genetics* **38**:8 (2006), 904–909.

[27] Richard T. Barfield, Lynn M. Almli, Varun Kilaru, Alicia K. Smith, Kristina B. Mercer, Richard Duncan, Torsten Klengel, Divya Mehta, Elisabeth B. Binder, Michael P. Epstein, Kerry J. Ressler, and Karen N. Conneely. "Accounting for Population Stratification in DNA Methylation Studies". *Genetic Epidemiology* **38**:3 (2014), 231–241.

[28] Paul Yousefi, Karen Huen, Veronica Dave, Lisa Barcellos, Brenda Eskenazi, and Nina Holland. "Sex differences in DNA methylation assessed by 450 K BeadChip in newborns". *Bmc Genomics* **16**: (2015).

[29] C. C. Whitacre, S. C. Reingold, P. A. O'Looney, and Autoimmuni Task Force Gender Multiple Sclerosis. "Biomedicine - A gender cap in autoimmunity". *Science* **283**:5406 (1999), 1277–1278.

[30] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. "Discovery of cross-reactive probes and poly-

morphic CpGs in the Illumina Infinium HumanMethylation450 microarray". *Epigenetics* **8**:2 (2013), 203–209.

[31] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. "The UCSC Table Browser data retrieval tool". *Nucleic Acids Research* **32**: (2004), D493–D496.

[32] Kate R. Rosenbloom, Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Glenn Hickey, Angie S. Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T. Lee, Chin H. Li, Karen H. Miga, Ngan Nguyen, Benedict Paten, Brian J. Raney, Arian F. A. Smit, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. "The UCSC Genome Browser database: 2015 update". *Nucleic Acids Research* **43**:D1 (2015), D670–D681.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org.

[34] Sean Davis, Pan Du, Sven Bilke, Jr. Triche Timothy J., and Moiz Bootwalla. *methylumi: Handle Illumina methylation data*. R package version 2.14.0. 2015. URL: https://bioconductor.org/packages/methylumi/.

[35] Wolfgang Huber, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D. Hansen, Rafael A. Irizarry, Michael Lawrence, Michael I. Love, James MacDonald, Valerie Obenchain, Andrzej K. Oles, Herve Pages, Alejandro Reyes, Paul Shannon, Gordon K. Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. "Orchestrating high-throughput genomic analysis with Bioconductor". *Nature Methods* **12**:2 (2015), 115–121.

[36] Jr. Triche Timothy J., Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. "Low-level processing of Illumina Infinium DNA Methylation BeadArrays". *Nucleic Acids Research* **41**:7 (2013).

[37] Paul Yousefi, Karen Huen, Raul Aguilar Schall, Anna Decker, Emon Elboudwarej, Hong Quach, Lisa Barcellos, and Nina Holland. "Considerations for normalization of DNA methylation data by Illumina

450K BeadChip assay in population studies". *Epigenetics* **8**:11 (2013), 1141–1152.

[38] Andrew E. Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data". *Bioinformatics* **29**:2 (2013), 189–196.

[39] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, John D. Storey, Yuqing Zhang, and Leonardo Collado Torres. *sva: Surrogate Variable Analysis*. R package version 3.28.0. 2018. URL: https://bioconductor.org/packages/sva/.

[40] Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia MT Greenwood, and Kasper D. Hansen. "Functional normalization of 450k methylation array data improves replication in large cancer studies". *Genome Biology* **15**:11 (2014), 503.

[41] Y. Benjamini and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics* **29**:4 (2001), 1165–1188.

[42] John W. Whitaker, Robert Shoemaker, David L. Boyle, Josh Hillman, David Anderson, Wei Wang, and Gary S. Firestein. "An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype". *Genome Medicine* **5**: (2013).

[43] Kristin N. Harper, Brandilyn A. Peters, and Mary V. Gamble. "Batch Effects and Pathway Analysis: Two Potential Perils in Cancer Studies Involving DNA Methylation Array Analysis". *Cancer Epidemiology Biomarkers & Prevention* **22**:6 (2013), 1052–1060.

[44] T. O. R. Hjelmervik, K. Petersen, I. Jonassen, R. Jonsson, and A. I. Bolstad. "Gene expression profiling of minor salivary glands clearly distinguishes primary Sjogren's syndrome patients from healthy control subjects". *Arthritis and Rheumatism* **52**:5 (2005), 1534–1544.

[45] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". *Bioinformatics* **26**:6 (2010), 841–842.

[46] Marc Carlson. *Genome wide annotation for Human*. R package version 3.1.2. 2015. URL: http://bioconductor.org/packages/org.Hs.eg.db/.

[47] Kazuhisa Nakano, John W. Whitaker, David L. Boyle, Wei Wang, and Gary S. Firestein. "DNA methylome signature in rheumatoid arthritis". *Annals of the Rheumatic Diseases* **72**:1 (2013), 110–117.

[48] Thomas Egerer, Lorena Martinez-Gamboa, Anja Dankof, Bruno Stuhlmueller, Thomas Doerner, Veit Krenn, Karl Egerer, Paul E. Rudolph, Gerd-R. Burmester, and Eugen Feist. "Tissue-specific up-regulation of the proteasome subunit beta 5i (LMP7) in Sjogren's syndrome". *Arthritis and Rheumatism* **54**:5 (2006), 1501–1508.

[49] R. I. Fox, J. Tornwall, and P. Michelson. "Current issues in the diagnosis and treatment of Sjogren's syndrome". *Current opinion in rheumatology* **11**:5 (1999), 364–71.

[50] Ryo Ueda, Gary Kohanbash, Kotaro Sasaki, Mitsugu Fujita, Xinmei Zhu, Edward R. Kastenhuber, Heather A. McDonald, Douglas M. Potter, Ronald L. Hamilton, Michael T. Lotze, Saleem A. Khan, Robert W. Sobol, and Hideho Okada. "Dicer-regulated microRNAs 222 and 339 promote resistance of cancer cells to cytotoxic T-lymphocytes by down-regulation of ICAM-1". *Proceedings of the National Academy of Sciences of the United States of America* **106**:26 (2009), 10746–10751.

[51] Ilias Alevizos, Stefanie Alexander, R. James Turner, and Gabor G. Illei. "MicroRNA Expression Profiles as Biomarkers of Minor Salivary Gland Inflammation and Dysfunction in Sjogren's Syndrome". *Arthritis and Rheumatism* **63**:2 (2011), 535–544.

[52] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdottir, Pablo Tamayo, and Jill P. Mesirov. "Molecular signatures database (MSigDB) 3.0". *Bioinformatics* **27**:12 (2011), 1739–1740.

[53] Arthur Liberzon, Chet Birger, Helga Thorvaldsdottir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. "The Molecular Signatures Database Hallmark Gene Set Collection". *Cell Systems* **1**:6 (2015), 417–425.

[54] Christopher J. Lessard, He Li, Indra Adrianto, John A. Ice, Astrid Rasmussen, Kiely M. Grundahl, Jennifer A. Kelly, Mikhail G. Dozmorov, Corinne Miceli-Richard, Simon Bowman, Sue Lester, Per Eriksson, Maija-Leena Eloranta, Johan G. Brun, Lasse G. Goransson, Erna Harboe, Joel M. Guthridge, Kenneth M. Kaufman, Marika Kvarnstrom,

Helmi Jazebi, Deborah S. Cunninghame Graham, Martha E. Grandits, Abu N. M. Nazmul-Hossain, Ketan Patel, Adam J. Adler, Jacen S. Maier-Moore, A. Darise Farris, Michael T. Brennan, James A. Lessard, James Chodosh, Rajaram Gopalakrishnan, Kimberly S. Hefner, Glen D. Houston, Andrew J. W. Huang, Pamela J. Hughes, David M. Lewis, Lida Radfar, Michael D. Rohrer, Donald U. Stone, Jonathan D. Wren, Timothy J. Vyse, Patrick M. Gaffney, Judith A. James, Roald Omdal, Marie Wahren-Herlenius, Gabor G. Illei, Torsten Witte, Roland Jonsson, Maureen Rischmueller, Lars Ronnblom, Gunnel Nordmark, Ng Wan-Fai, Xavier Mariette, Juan-Manuel Anaya, Nelson L. Rhodus, Barbara M. Segal, R. Hal Scofield, Courtney G. Montgomery, John B. Harley, Kathy L. Sivils, and U. K. Primary Sjogren's Syndrome Regi. "Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjogren's syndrome". *Nature Genetics* **45**:11 (2013), 1284–+.

[55] Yongzhe Li, Kunlin Zhang, Hua Chen, Fei Sun, Juanjuan Xu, Ziyan Wu, Ping Li, Liuyan Zhang, Yang Du, Haixia Luan, Xi Li, Lijun Wu, Hongbin Li, Huaxiang Wu, Xiangpei Li, Xiaomei Li, Xiao Zhang, Lu Gong, Lie Dai, Lingyun Sun, Xiaoxia Zuo, Jianhua Xu, Huiping Gong, Zhijun Li, Shengquan Tong, Min Wu, Xiaofeng Li, Weiguo Xiao, Guochun Wang, Ping Zhu, Min Shen, Shengyun Liu, Dongbao Zhao, Wei Liu, Yi Wang, Cibo Huang, Quan Jiang, Guijian Liu, Bin Liu, Shaoxian Hu, Wen Zhang, Zhuoli Zhang, Xin You, Mengtao Li, Weixin Hao, Cheng Zhao, Xiaomei Leng, Liqi Bi, Yongfu Wang, Fengxiao Zhang, Qun Shi, Wencheng Qi, Xuewu Zhang, Yuan Jia, Jinmei Su, Qin Li, Yong Hou, Qingjun Wu, Dong Xu, Wenjie Zheng, Miaojia Zhang, Qian Wang, Yunyun Fei, Xuan Zhang, Jing Li, Ying Jiang, Xinping Tian, Lidan Zhao, Li Wang, Bin Zhou, Yang Li, Yan Zhao, Xiaofeng Zeng, Jurg Ott, Jing Wang, and Fengchun Zhang. "A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjogren's syndrome at 7q11.23". *Nature Genetics* **45**:11 (2013), 1361–+.

[56] Robert C. McLeay and Timothy L. Bailey. "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data". *Bmc Bioinformatics* **11**: (2010).

[57] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, Francois Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge

Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles". *Nucleic Acids Research* **42**:D1 (2014), D142–D147.

[58] O. Johnsen, P. Murphy, H. Prydz, and A. B. Kolsto. "Interaction of the CNC-bZIP factor TCF11/LCR-F1/Nrf1 with MafG: binding-site selection and regulation of transcription". *Nucleic Acids Research* **26**:2 (1998), 512–520.

[59] Janos Steffen, Michael Seeger, Annett Koch, and Elke Krueger. "Proteasomal Degradation Is Transcriptionally Controlled by TCF11 via an ERAD-Dependent Feedback Loop". *Molecular Cell* **40**:1 (2010), 147–158.

[60] S. E. Hartman, P. Bertone, A. K. Nath, T. E. Royce, M. Gerstein, S. Weissman, and M. Snyder. "Global changes in STAT target selection and transcription regulation upon interferon treatments". *Genes & Development* **19**:24 (2005), 2953–2968.

[61] Elodie Portales-Casamar, Stefan Kirov, Jonathan Lim, Stuart Lithwick, Magdalena I. Swanson, Amy Ticoll, Jay Snoddy, and Wyeth W. Wasserman. "PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation". *Genome Biology* **8**:10 (2007).

[62] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". *Molecular Cell* **38**:4 (2010), 576–589.

[63] M. Suzuki, T. Yamada, F. Kihara-Negishi, T. Sakurai, E. Hara, D. G. Tenen, N. Hozumi, and T. Oikawa. "Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b". *Oncogene* **25**:17 (2006), 2477–2488.

[64] Y. Renaudineau and E. Ballestar. "DNA methylation signatures in Sjogren syndrome". *Nature Reviews Rheumatology* **12**:10 (2016), 565–+.

[65] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. "Single-Cell

Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum". *Science* **332**:6030 (2011), 687–696.

[66] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors (vol 163, pg 1663, 2015)". *Cell* **164**:1-2 (2016), 325–325.

[67] David Allman and Shiv Pillai. "Peripheral B cell subsets". *Current Opinion in Immunology* **20**:2 (2008), 149–157.

[68] Akiko Iwasaki and Ruslan Medzhitov. "Control of adaptive immunity by the innate immune system". *Nature Immunology* **16**:4 (2015), 343–353.

[69] Yuka Kanno, Golnaz Vahedi, Kiyoshi Hirahara, Kentner Singleton, and John J. O'Shea. "Transcriptional and Epigenetic Control of T Helper Cell Specification: Molecular Mechanisms Underlying Commitment and Plasticity". *Annual Review of Immunology, Vol 30*. Ed. by W. E. Paul. Vol. 30. Annual Review of Immunology. 2012, 707–731.

[70] Miriam Merad, Priyanka Sathe, Julie Helft, Jennifer Miller, and Arthur Mortha. "The Dendritic Cell Lineage: Ontogeny and Function of Dendritic Cells and Their Subsets in the Steady State and the Inflamed Setting". *Annual Review of Immunology, Vol 31*. Ed. by D. R. Littman and W. M. Yokoyama. Vol. 31. Annual Review of Immunology. 2013, 563–604.

[71] Alexandra-Chloe Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L. De Jager, Orit Rozenblatt-Rosen, Andrew A. Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors". *Science* **356**:6335 (2017).

[72] A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon. "Dynamic Proteomics of Individual Cancer Cells in Response to a Drug". *Science* **322**:5907 (2008), 1511–1516.

[73] Alex K. Shalek, Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J. Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z. Levin, Hongkun Park, and Aviv Regev. "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells". *Nature* **498**:7453 (2013), 236–240.

[74] Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation". *Nature* **510**:7505 (2014), 363–+.

[75] Nir Yosef, Alex K. Shalek, Jellert T. Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, David Gennert, Rahul Satija, Arvind Shakya, Diana Y. Lu, John J. Trombetta, Meenu R. Pillai, Peter J. Ratcliffe, Mathew L. Coleman, Mark Bix, Dean Tantin, Hongkun Park, Vijay K. Kuchroo, and Aviv Regev. "Dynamic regulatory network controlling T(H)17 cell differentiation". *Nature* **496**:7446 (2013), 461–+.

[76] Ofer Feinerman, Garrit Jentsch, Karen E. Tkach, Jesse W. Coward, Matthew M. Hathorn, Michael W. Sneddon, Thierry Emonet, Kendall A. Smith, and Gregoire Altan-Bonnet. "Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response". *Molecular Systems Biology* **6**: (2010).

[77] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews Genetics* **10**:1 (2009), 57–63.

[78] Bo Li and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". *Bmc Bioinformatics* **12**: (2011).

[79] Dominic Grun, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. "Single-cell messenger RNA sequencing reveals rare intestinal cell types". *Nature* **525**:7568 (2015), 251–+.

[80] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell* **161**:5 (2015), 1202–1214.

[81] E. Martin-Gayo, M. B. Cole, K. E. Kolb, Z. Y. Ouyang, J. Cronin, S. W. Kazer, J. Ordovas-Montanes, M. Lichterfeld, B. D. Walker, N. Yosef, A. K. Shalek, and X. G. Yu. "A Reproducibility-Based Computational Framework Identifies an Inducible, Enhanced Antiviral State in Dendritic Cells from HIV-1 Elite Controllers". *Genome Biology* **19**: (2018), 21.

[82] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, Naveen Ramalingam, Gang Sun, Myo Thu, Michael Norris, Ronald Lebofsky, Dominique Toppani, Darnell W Kemp, Michael Wong, Barry Clerkson, Brittnee N Jones, Shiquan Wu, Lawrence Knutsson, Beatriz Alvarado, Jing Wang, Lesley S Weaver, Andrew P May, Robert C Jones, Marc A Unger, Arnold R Kriegstein, and Jay A A West. "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex". *Nature Biotechnology* **32**:10 (2014), 1053–8. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4191988{\&}tool=pmcentrez{\&}rendertype=abstract.

[83] Jellert T. Gaublomme, Nir Yosef, Youjin Lee, Rona S. Gertner, Li V. Yang, Chuan Wu, Pier Paolo Pandolfi, Tak Mak, Rahul Satija, Alex K. Shalek, Vijay K. Kuchroo, Hongkun Park, and Aviv Regev. "Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity". *Cell* **163**:6 (2015), 1400–1412.

[84] Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. "Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing". *Cell Systems* **2**:4 (2016), 239–250.

[85] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. "Massively parallel digital transcriptional profiling of single cells". *Nature Communications* **8**: (2017), 14049.

[86] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. "Batch effects and the effective design of single-cell gene expression studies". *Scientific Reports* **7**: (2017), 39921.

[87] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. "Simultaneous epitope and transcriptome measurement in single cells". *Nature Methods* **14**: (2017). URL: http://dx.doi.org/10.1038/nmeth.4380.

[88] Daniel Ramskold, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R. Faridani, Gregory A. Daniels, Irina Khrebtukova, Jeanne F. Loring, Louise C. Laurent, Gary P. Schroth, and Rickard Sandberg. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". *Nature Biotechnology* **30**:8 (2012), 777–782.

[89] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A. Teichmann, John C. Marioni, and Marcus G. Heisler. "Accounting for technical noise in single-cell RNA-seq experiments". *Nature Methods* **10**:11 (2013), 1093–1095.

[90] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. "Bayesian approach to single-cell differential expression analysis". *Nature Methods* **11**:7 (2014), 740–U184.

[91] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, and others. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data". *Genome Biology* **16**:1 (2015), 1.

[92] David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. "MAGIC: A diffusion-based imputation method

reveals gene-gene interactions in single-cell RNA-sequencing data". *bioRxiv* (2017), 111591.

[93] Wei Vivian Li and Jingyi Jessica Li. "scImpute: accurate and robust imputation for single cell RNA-seq data". *bioRxiv* (2017), 141598.

[94] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. "ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data". *bioRxiv* (2017), 125112.

[95] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael Jordan, and Nir Yosef. "Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing". *bioRxiv* (2018). eprint: https://www.biorxiv.org/content/early/2018/03/30/292037.full.pdf. URL: https://www.biorxiv.org/content/early/2018/03/30/292037.

[96] Stephanie C Hicks, F W Townes, Mingxiang Teng, and Rafael A Irizarry. "Missing data and technical variability in single-cell RNA-sequencing experiments". *Biostatistics* (2017), kxx053.

[97] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. "Classification of low quality cells from single-cell RNA-seq data". *Genome Biology* **17**:1 (2016), 1. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0888-1.

[98] Simon Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2010. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[99] *picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*. URL: http://broadinstitute.github.io/picard.

[100] 10x Genomics. *Cell Ranger: Single Cell Analysis Pipelines*. URL: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger.

[101] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. "Challenges in the normalization of single-cell RNA sequencing datasets". *Nature Methods* **14**: (2017), 565–571.

[102] Allon Wagner, Aviv Regev, and Nir Yosef. "Revealing the vectors of cellular identity with single-cell genomics". *Nature Biotechnology* **8 Nov**: (2016). URL: http://dx.doi.org/10.1038/nbt.3711.

[103] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. "GC-Content Normalization for RNA-Seq Data". *BMC Bioinformatics* **12**: (2011), Article 480. URL: http://www.biomedcentral.com/1471-2105/12/480/abstract.

[104] Rhonda Bacher and Christina Kendziorski. "Design and computational analysis of single-cell RNA-sequencing experiments". *Genome Biology* **17**:1 (2016), 1.

[105] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* **5**:7 (2008), 621–628.

[106] M. D. Robinson and A. Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". *Genome Biology* **11**:3 (2010), R25.

[107] S. Anders and W. Huber. "Differential expression analysis for sequence count data". *Genome Biology* **11**:10 (2010), R106.

[108] Johann A Gagnon-Bartsch and Terence P Speed. "Using control genes to correct for unwanted variation in microarray data". *Biostatistics* **13**:3 (2012), 539–552.

[109] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. "Normalization of RNA-seq data using factor analysis of control genes or samples". *Nature Biotechnology* **32**:9 (2014), 896–902. URL: http://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.2931.html.

[110] Jeffrey T Leek and John D Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". *PLoS genetics* **3**:9 (2007), e161.

[111] Jeffrey T Leek. "svaseq: removing batch effects and other unwanted noise from sequencing data". *Nucleic Acids Research* **42**:21 (2014), e161–e161.

[112] Aaron TL Lun, Karsten Bach, and John C Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". *Genome Biology* **17**:1 (2016), 1.

[113] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. "Single-cell mRNA quantification and differential analysis with Census". *Nature Methods* **14**:3 (2017), 309–315.

[114] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". *Nature Biotechnology* **33**:2 (2015), 155–160.

[115] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. "SCnorm: robust normalization of single-cell RNA-seq data". *Nature Methods* **14**:6 (2017), 584–586.

[116] Bo Ding, Lina Zheng, Yun Zhu, Nan Li, Haiyang Jia, Rizi Ai, Andre Wildberg, and Wei Wang. "Normalization and noise reduction for single cell RNA-seq experiments". *Bioinformatics* (2015), btv122.

[117] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. "BASiCS: Bayesian analysis of single-cell sequencing data". *PLoS Computational Biology* **11**:6 (2015), e1004333.

[118] J. H. Bullard, E. A. Purdom, K. D. Hansen, and S. Dudoit. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". *BMC Bioinformatics* **11**: (2010), Article 94. URL: http://www.biomedcentral.com/1471-2105/11/94/abstract.

[119] Yunshun Chen, Aaron Lun, Davis McCarthy, Xiaobei Zhou, Mark Robinson, and Gordon Smyth. *edgeR: Empirical Analysis of Digital Gene Expression Data in R*. 2010. URL: https://bioconductor.org/packages/edgeR/.

[120] Simon Anders. *DESeq: Differential gene expression analysis based on the negative binomial distribution*. 2010. URL: https://bioconductor.org/packages/DESeq/.

[121] Aaron Lun, Karsten Bach, Jong Kyoung Kim, Antonio Scialdone, and Laleh Haghverdi. *scran: Methods for Single-Cell RNA-Seq Data Analysis*. 2016. URL: https://bioconductor.org/packages/scran/.

[122] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data". *Biostatistics* **4**:2 (2003), 249–264.

[123] Davide Risso, Sandrine Dudoit, and Ludwig Geistlinger. *EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq*. 2010. URL: https://bioconductor.org/packages/EDASeq/.

[124] Rhonda Bacher. *scnorm: Normalization of single cell RNA-seq data*. 2017. URL: https://bioconductor.org/packages/SCnorm/.

[125] Davide Risso, Sandrine Dudoit, Lorena Pantano, and Kamil Slowikowski. *RUVSeq: Remove Unwanted Variation from RNA-Seq Data*. 2014. URL: https://bioconductor.org/packages/RUVSeq.

[126] Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. "Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-seq". *bioRxiv* (2018). eprint: https://www.biorxiv.org/content/early/2018/05/18/235382.full.pdf. URL: https://www.biorxiv.org/content/early/2018/05/18/235382.

[127] Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. "Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R". *Bioinformatics* **14 Jan**: (2017). URL: http://dx.doi.org/10.1093/bioinformatics/btw777.

[128] P. J. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* **20**: (1987), 53–65.

[129] Luke C Gandolfo and Terence P Speed. "RLE Plots: Visualising Unwanted Variation in High Dimensional Data". *arXiv preprint arXiv:1704.03590* (2017).

[130] Karl Ruben Gabriel. "The biplot graphic display of matrices with application to principal component analysis". *Biometrika* **58**:3 (1971), 453–467.

[131] Jeremy A Miller, Song-Lin Ding, Susan M Sunkin, Kimberly A Smith, Lydia Ng, Aaron Szafer, Amanda Ebbert, Zackery L Riley, Joshua J Royall, Kaylynn Aiona, et al. "Transcriptional landscape of the prenatal human brain". *Nature* **508**:7495 (2014), 199–206.

[132] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* **43**:7 (2015), e47.

[133] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biology* **15**:2 (2014), R29.

[134] Youjin Lee, Amit Awasthi, Nir Yosef, Francisco J Quintana, Sheng Xiao, Anneli Peters, Chuan Wu, Markus Kleinewietfeld, Sharon Kunder, David A Hafler, et al. "Induction and molecular signature of pathogenic TH17 cells". *Nature Immunology* **13**:10 (2012), 991–999.

[135] Helder I Nakaya, Jens Wrammert, Eva K Lee, Luigi Racioppi, Stephanie Marie-Kunze, W Nicholas Haining, Anthony R Means, Sudhir P Kasturi, Nooruddin Khan, Gui-Mei Li, et al. "Systems biology of vaccination for seasonal influenza in humans". *Nature Immunology* **12**:8 (2011), 786–795.

[136] Rahul Satija, Andrew Butler, and Paul Hoffman. *Seurat: Tools for Single Cell Genomics*. R package version 2.1.0. 2017. URL: https://CRAN.R-project.org/package=Seurat.

[137] Luke Zappia. *splatter: Simple Simulation of Single-cell RNA Sequencing Data*. 2017. URL: https://bioconductor.org/packages/splatter.

[138] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*. R package version 1.0.5.9000. URL: http://shiny.rstudio.com.

[139] R. B. Fletcher, D. Das, L. Gadye, K. N. Street, A. Baudhuin, A. Wagner, M. B. Cole, Q. Flores, Y. G. Choi, N. Yosef, E. Purdom, S. Dudoit, D. Risso, and J. Ngai. "Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution". *Cell Stem Cell* **20**: (2017), 817–830.

[140] Shaked Afik, Kathleen B Yates, Kevin Bi, Samuel Darko, Jernej Godec, Ulrike Gerdemann, Leo Swadling, Daniel C Douek, Paul Klenerman, Eleanor J Barnes, et al. "Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state". *Nucleic Acids Research* (2017).

[141] Levi Gadye, Diya Das, Michael A. Sanchez11, Kelly Street, Ariane Baudhuin, Allon Wagner, Michael B. Cole, Yoon Gi Choi, Nir Yosef, Elizabeth Purdom, Sandrine Dudoit, Davide Risso, John Ngai, and Russell B. Fletcher. "Injury Activates Transient Olfactory Stem Cell States with Diverse Lineage Capacities". *Cell Stem Cell* **21**:6 (2017), 775 –790.

[142] Enrique Martin-Gayo, Michael B. Cole, Kellie E. Kolb, Zhengyu Ouyang, Jacqueline Cronin, Samuel W. Kazer, Jose Ordovas-Montanes, Mathias Lichterfeld, Bruce D. Walker, Nir Yosef, Alex K. Shalek, and Xu G. Yu. "A Reproducibility-Based Computational Framework Identifies an Inducible, Enhanced Antiviral State in Dendritic Cells from HIV-1 Elite Controllers". *Genome Biology* **19**:10 (2018).

[143] Emma Pierson and Christopher Yau. "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". *Genome Biology* **16**:1 (2015), 241.

[144] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. "Varying-Censoring Aware Matrix Factorization for Single Cell RNA-Sequencing". *bioRxiv* (2017), 166736.

[145] Robert C Gentleman, Vincent J Carey, Douglas M Bates, and others. "Bioconductor: Open software development for computational biology and bioinformatics". *Genome Biology* **5**: (2004), R80. URL: http://genomebiology.com/2004/5/10/R80.

[146] Martin Morgan, Valerie Obenchain, Michel Lang, and Ryan Thompson. *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.11.11. 2017. URL: https://github.com/Bioconductor/BiocParallel.

[147] Bernd Fischer, Gregoire Pau, and Mike Smith. *rhdf5: HDF5 interface to R*. R package version 2.21.6. 2017.

[148] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. "scmap: projection of single-cell RNA-seq data across data sets". *Nature Methods* **15**: (Apr. 2018), 359 EP –. URL: http://dx.doi.org/10.1038/nmeth.4644.

[149] Nicolas Chevrier, Philipp Mertins, Maxim N. Artyomov, Alex K. Shalek, Matteo Iannacone, Mark F. Ciaccio, Irit Gat-Viks, Elena Tonti, Marciela M. DeGrace, Karl R. Clauser, Manuel Garber, Thomas M. Eisenhaure, Nir Yosef, Jacob Robinson, Amy Sutton, Mette S. Andersen, David E. Root, Ulrich von Andrian, Richard B. Jones, Hongkun Park, Steven A. Carr, Aviv Regev, Ido Amit, and Nir Hacohen. "Systematic Discovery of TLR Signaling Components Delineates Viral-Sensing Circuits". *Cell* **147**:4 (2011), 853–867.

[150] Jinghe Huang, Patrick S. Burke, Thai Duong Hong Cung, Florencia Pereyra, Ildiko Toth, Bruce D. Walker, Luis Borges, Mathias Lichterfeld, and Xu G. Yu. "Leukocyte Immunoglobulin-Like Receptors Maintain Unique Antigen-Presenting Properties of Circulating Myeloid Dendritic Cells in HIV-1-Infected Elite Controllers". *Journal of Virology* **84**:18 (2010), 9463–9471.

[151] Enrique Martin-Gayo, Maria Jose Buzon, Zhengyu Ouyang, Taylor Hickman, Jacqueline Cronin, Dina Pimenova, Bruce D. Walker, Mathias Lichterfeld, and Xu G. Yu. "Potent Cell-Intrinsic Immune Responses in Dendritic Cells Facilitate HIV-1-Specific T Cell Immunity in HIV-1 Elite Controllers". *Plos Pathogens* **11**:6 (2015).

[152] Galit Alter, David Heckerman, Arne Schneidewind, Lena Fadda, Carl M. Kadie, Jonathan M. Carlson, Cesar Oniangue-Ndza, Maureen Martin, Bin Li, Salim I. Khakoo, Mary Carrington, Todd M. Allen, and Marcus Altfeld. "HIV-1 adaptation to NK-cell-mediated immune pressure". *Nature* **476**:7358 (2011), 96–+.

[153] David DeTomaso and Nir Yosef. "FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data". *Bmc Bioinformatics* **17**: (2016).

[154] Nicolas Manel, Brandon Hogstad, Yaming Wang, David E. Levy, Derya Unutmaz, and Dan R. Littman. "A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells". *Nature* **467**:7312 (2010), 214–U104.

[155] Simone Picelli, Asa K. Bjorklund, Omid R. Faridani, Sven Sagasser, Gosta Winberg, and Rickard Sandberg. "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". *Nature Methods* **10**:11 (2013), 1096–1098.

[156] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2". *Nature Methods* **9**:4 (2012), 357–U54.

[157] Damien Chaussabel, Roshanak Tolouei Semnani, Mary Ann McDowell, David Sacks, Alan Sher, and Thomas B. Nutman. "Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites". *Blood* **102**:2 (2003), 672–681. eprint: http://www.bloodjournal.org/content/102/2/672.full.pdf.

[158] Maurizio Ceppi, Giovanna Clavarino, Evelina Gatti, Enrico K. Schmidt, Aude de Gassart, Derek Blankenship, Gerald Ogola, Jacques Banchereau, Damien Chaussabel, and Philippe Pierre. "Ribosomal protein mRNAs are translationally-regulated during human dendritic cells activation by LPS". *Immunome research* **5**: (2009), 5–5.

[159] Elena Zaslavsky, Uri Hershberg, Jeremy Seto, Alissa M. Pham, Susanna Marquez, Jamie L. Duke, James G. Wetmur, Benjamin R. tenOever, Stuart C. Sealfon, and Steven H. Kleinstein. "Antiviral Response Dictated by Choreographed Cascade of Transcription Factors". *Journal of Immunology* **184**:6 (2010), 2908–2917.

[160] G. Napolitani, A. Rinaldi, F. Bertoni, F. Sallusto, and A. Lanzavecchia. "Selected Toll-like receptor agonist combinations synergistically trigger a T helper type 1-polarizing program in dendritic cells". *Nature Immunology* **6**:8 (2005), 769–776.

[161] Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K. Grenier, Weibo Li, Or Zuk, Lisa A. Schubert, Brian Birditt, Tal Shay, Alon Goren, Xiaolan Zhang, Zachary Smith, Raquel Deering, Rebecca C. McDonald, Moran Cabili, Bradley E. Bernstein, John L. Rinn, Alex Meissner, David E. Root, Nir Hacohen, and Aviv Regev. "Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses". *Science* **326**:5950 (2009), 257–263.

[162] B. de Saint-Vis, J. Vincent, S. Vandenabeele, B. Vanbervliet, J. J. Pin, S. Ait-Yahia, S. Patel, M. G. Mattei, J. Banchereau, S. Zurawski, J. Davoust, C. Caux, and S. Lebecque. "A novel lysosome-associated membrane glycoprotein, DC-LAMP, induced upon DC maturation, is transiently expressed in MHC class II compartment". *Immunity* **9**:3 (1998), 325–336.

[163] Andrew N. Harman, Joey Lai, Stuart Turville, Shamith Samarajiwa, Lachlan Gray, Valerie Marsden, Sarah Mercier, Kate Jones, Najla Nasr, Arjun Rustagi, Helen Cumming, Heather Donaghy, Johnson Mak, Jr. Gale Michael, Melissa Churchill, Paul Hertzog, and Anthony L. Cunningham. "HIV infection of dendritic cells subverts the IFN induction pathway via IRF-1 and inhibits type 1 IFN production". *Blood* **118**:2 (2011), 298–308.

[164] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. "Causal analysis approaches in Ingenuity Pathway Analysis". *Bioinformatics* **30**:4 (2014), 523–530.

[165] Cees E. van der Poel, Robbert M. Spaapen, Jan G. J. van de Winkel, and Jeanette H. W. Leusen. "Functional Characteristics of the High Affinity IgG Receptor, Fc gamma RI". *Journal of Immunology* **186**:5 (2011), 2699–2704.

[166] Pamela M. Odorizzi, Kristen E. Pauken, Michael A. Paley, Arlene Sharpe, and E. John Wherry. "Genetic absence of PD-1 promotes accumulation of terminally differentiated exhausted CD8(+) T cells". *Journal of Experimental Medicine* **212**:7 (2015), 1125–1137.

[167] Cameron R. Cunningham, Ameya Champhekar, Michael V. Tullius, Barbara Jane Dillon, Anjie Zhen, Justin Rafael de la Fuente, Jonathan Herskovitz, Heidi Elsaesser, Laura M. Snell, Elizabeth B. Wilson, Juan Carlos de la Torre, Scott G. Kitchen, Marcus A. Horwitz, Steven J. Bensinger, Stephen T. Smale, and David G. Brooks. "Type I and Type II Interferon Coordinately Regulate Suppressive Dendritic Cell Fate and Function during Viral Persistence". *Plos Pathogens* **12**:1 (2016).

[168] Paloma Carranza, Perla M. Del Rio Estrada, Dafne Diaz Rivera, Yuria Ablanedo-Terrazas, and Gustavo Reyes-Teran. "Lymph nodes from HIV-infected individuals harbor mature dendritic cells and increased numbers of PD-L1+conventional dendritic cells". *Human Immunology* 77:7 (2016), 584–593.

[169] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology* **15**:12 (2014), 550. URL: https://doi.org/10.1186/s13059-014-0550-8.

[170] Daxing Gao, Jiaxi Wu, You-Tong Wu, Fenghe Du, Chukwuemika Aroh, Nan Yan, Lijun Sun, and Zhijian J. Chen. "Cyclic GMP-AMP Synthase Is an Innate Immune Sensor of HIV and Other Retroviruses". *Science* **341**:6148 (2013), 903–906.

[171] Nan Yan, Ashton D. Regalado-Magdos, Bart Stiggelbout, Min Ae Lee-Kirsch, and Judy Lieberman. "The cytosolic exonuclease TREX1 inhibits the innate immune response to human immunodeficiency virus type 1". *Nature Immunology* **11**:11 (2010), 1005–U53.

[172] Matthias Habjan and Andreas Pichlmair. "Cytoplasmic sensing of viral nucleic acids". *Current Opinion in Virology* **11**: (2015), 31–37.

[173] Yueh-Ming Loo and Jr. Gale Michael. "Immune Signaling by RIG-I-like Receptors". *Immunity* **34**:5 (2011), 680–692.

[174] Zhe Ma and Blossom Damania. "The cGAS-STING Defense Pathway and Its Counteraction by Viruses". *Cell Host & Microbe* **19**:2 (2016), 150–158.

[175] Alejandra Peris-Pertusa, Mariola Lopez, Norma I. Rallon, Clara Restrepo, Vincent Soriano, and Jose M. Benito. "Evolution of the Functional Profile of HIV-Specific CD8(+) T Cells in Patients With Different Progression of HIV Infection Over 4 Years". *Jaids-Journal of Acquired Immune Deficiency Syndromes* **55**:1 (2010), 29–38.

[176] M. G. Filbin, I. Tirosh, V. Hovestadt, M. L. Shaw, L. E. Escalante, N. D. Mathewson, C. Neftel, N. Frank, K. Pelton, C. Hebert, C. Haberler, K. Yizhak, J. Gojo, K. Egervari, C. Mount, P. van Galen, D. M. Bonal, Q. D. Nguyen, A. Beck, C. Sinai, T. Czech, C. Dorfer, L. Goumnerova, C. Lavarino, A. M. Carcaboso, J. Mora, R. Mylvaganam, C. C. Luo, A. Peyrl, M. Popovic, A. Azizi, T. T. Batchelor, M. P. Frosch, M. Martinez-Lage, M. W. Kieran, P. Bandopadhayay, R. Beroukhim, G. Fritsch, G. Getz, O. Rozenblatt-Rosen, K. W. Wucherpfennig, D. N. Louis, M. Monje, I. Slavc, K. L. Ligon, T. R. Golub, A. Regev, B. E. Bernstein, and M. L. Suva. "Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq". *Science* **360**:6386 (2018), 331–335.

[177] H. A. Zhang, C. A. A. Lee, Z. L. Li, J. R. Garbe, C. R. Eide, R. Petegrosso, R. Kuang, and J. Tolar. "A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa". *Plos Computational Biology* **14**:4 (2018).

[178] E. Der, S. Ranabothu, H. Suryawanshi, K. M. Akat, R. Clancy, P. Morozov, M. Kustagi, M. Czuppa, P. Izmirly, H. M. Belmont, T. Wang, N. Jordan, N. Bornkamp, J. Nwaukoni, J. Martinez, B. Goilav, J. P. Buyon, T. Tuschl, and C. Putterman. "Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis". *Jci Insight* **2**:9 (2017).

[179] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences of the United States of America* **102**:43 (2005), 15545–15550.

[180] E. Eisenberg and E. Y. Levanon. "Human housekeeping genes are compact". *Trends in Genetics* **19**:7 (2003), 362–365.

[181] Tapio Lönnberg, Valentine Svensson, Kylie R. James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan S. F. Soon, Lily G. Fogg, Arya Sheela Nair, Urijah N. Liligeto, Michael J. T. Stubbington, Lam-Ha Ly, Frederik Otzen Bagger, Max Zwiessele, Neil D. Lawrence, Fernando Souza-Fonseca-Guimaraes, Patrick T. Bunn, Christian R. Engwerda, William R. Heath, Oliver Billker, Oliver Stegle, Ashraful Haque, and Sarah A. Teichmann. "Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria". *Science Immunology* **2**:9 (2017). eprint: http://immunology.sciencemag.org/content/2/9/eaal2192.full.pdf.

[182] Kumaran Kandasamy, S. Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S. Sameer Kumar, Abhilash K. Venugopal, Deepthi Telikicherla, J. Daniel Navarro, Suresh Mathivanan, Christian Pecquet, Sashi Kanth Gollapudi, Sudhir Gopal Tattikota, Shyam Mohan, Hariprasad Padhukasahasram, Yashwanth Subbannayya, Renu Goel, Harrys KC Jacob, Jun Zhong, Raja Sekhar, Vishalakshi Nanjappa, Lavanya Balakrishnan, Roopashree Subbaiah, YL Ramachandra, B. Abdul Rahiman, TS Keshava Prasad, Jian-Xin Lin, Jon CD Houtman, Stephen Desiderio, Jean-Christophe Renauld, Stefan N. Constantinescu, Osamu Ohara, Toshio Hirano, Masato Kubo, Sujay Singh, Purvesh Khatri, Sorin Draghici, Gary D. Bader, Chris Sander, Warren J. Leonard, and Akhilesh Pandey. "NetPath: a public resource of curated signal transduction pathways". *Genome Biology* **11**:1 (2010), R3.

[183] S. Crotty. "Follicular Helper CD4 T Cells (T-FH)". *Annual Review of Immunology, Vol 29* **29**: (2011), 621–663.

[184] X. D. Liu, X. Chen, B. Zhong, A. B. Wang, X. H. Wang, F. L. Chu, R. I. Nurieva, X. W. Yan, P. Chen, L. G. van der Flier, H. Nakatsukasa, S. S. Neelapu, W. J. Chen, H. Clevers, Q. Tian, H. Qi, L. Wei, and C. Dong. "Transcription factor achaete-scute homologue 2 initiates follicular T-helper-cell development". *Nature* **507**:7493 (2014), 513–+.

[185] Daniel Greene. *gsEasy: Gene Set Enrichment Analysis in R*. R package version 1.3. 2018. URL: https://CRAN.R-project.org/package=gsEasy.

[186] C Trapnell, L Pachter, and S L Salzberg. "TopHat: discovering splice junctions with RNA-Seq". *Bioinformatics* **25**:9 (2009), 1105–1111.

[187] Yang Liao, Gordon K. Smyth, and Wei Shi. "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features". *Bioinformatics* **30**:7 (2014), 923–930.

[188] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, David Gennert, Rahul Satija, Arvind Shakya, Diana Y Lu, John J Trombetta, Meenu R Pillai, Peter J Ratcliffe, Mathew L Coleman, Mark Bix, Dean Tantin, Hongkun Park, Vijay K Kuchroo, and Aviv Regev. "Dynamic regulatory network controlling TH17 cell differentiation". *Nature* **25 April**: (2013). URL: http://dx.doi.org/10.1038/nature11981.

PART V

APPENDIX

A

PUBLIC DATA SETS

A.1 DATA PROCESSING

For the T_H17 [83] and cortex data sets [82], Sequence Read Archive (SRA)-format files were downloaded from the SRA and transformed to FASTQ format using the SRA toolkit. Reads were aligned with `TopHat` (v. 2.0.11) [186] to the appropriate reference genome (GRCh38 for human cells, GRCm38 for mouse). RefSeq mouse gene annotation (GCF_000001635.23_GRCm38.p3) was downloaded from NCBI on Dec. 28, 2014. RefSeq human gene annotation (GCF_000001405.28) was downloaded from NCBI on Jun. 22, 2015. `featureCounts` (v. 1.4.6-p3) [187] was used to compute gene-level read counts.

SMART-SEQ_C1 T_H17 DATA SET. Cells were harvested from two C57BL/6J and three IL-17A – GFP^+ mice [83]. Unsorted non-pathogenic cells were collected from the first two mice and both IL-17A-sorted pathogenic and non-pathogenic cells were collected from the three remaining mice. Cells were sorted and a Fluidigm C1-based SMART-seq protocol was used for single-cell RNA extraction and sequencing. Following cell filtering, 337 cells were retained from four donor mice – one mouse was filtered out due to the small number of acceptable cells. Filtered expression data included 7,590 gene features over these 337 cells. For `scone`, we provided negative and positive control genes based on Supplementary Table S6 from Yosef et al. [188].

FLUIDIGM C1 CORTEX DATA SET. 65 cells from the developing cortex were assayed using the Fluidigm C1 microfluidics system [82]. Each cell was sequenced at both high and low depths; I focus on the high-coverage data. The data are available as part of the Bioconductor R package `scRNAseq` (`https://bioconductor.org/packages/scRNAseq`). No cell filtering was applied to this data set and 4,706 genes were retained following gene filtering. For

scone, we provided default negative control genes from the "housekeeping" list and positive control genes related to neurogenesis as annotated in MSigDB: JEPSEN_SMRT_TARGETS and GO_NEURAL_PRECURSOR_CELL_PROLIFERATION; http://software.broadinstitute.org/gsea/msigdb/cards/.

FLUIDIGM C1 IPSC DATA SET. Three batches of 96 libraries from each of three YRI iPSC lines were sequenced using the Fluidigm C1 microfluidics system [86]. The full data set, including UMI counts, read counts, and quality metrics, was obtained from https://github.com/jdblischak/singleCellSeq. Library-level QC measures included:

1. Proportion of reads aligning to ERCC spike-ins (matching the pattern "^ERCC");
2. number of unique molecules;
3. well number as reported in online metadata ("well");
4. concentration as reported in online metadata ("concentration");
5. number of detected molecule classes (genes with more than zero UMI).

Following gene and cell filtering, we retained 6,818 genes and 731 libraries; retained cells had more than 24,546 reads, more than 80% of common genes detected, and FNR AUC below 0.65. For scone, we provided default negative control genes from the "housekeeping" list, as well as positive control genes from the "cellcycle_genes" default list. ERCC genes were used as negative controls for RUVg normalization. Donor was used as a proxy for biological condition, while batch was defined as an individual C1 run.

10X GENOMICS PBMC DATA SET. I considered scRNA-seq data from two batches of PBMCs from a healthy donor (4k PBMCs and 8k PBMCs). The data were downloaded from the 10x Genomics website (https://www.10xgenomics.com/single-cell/) using the `cellrangerRkit` R package (v. 1.1.0). After filtering, 12,039 cells and 10,310 genes were retained. For scone, I provided default negative control genes from the "housekeeping" list and positive control genes as the top 513 most common genes annotated in the MSigDB C7 immunological signature collection (http://software.broadinstitute.org/gsea/msigdb/collections.jsp). `Seurat`-derived clusters [136] were used as a biological condition (see below), while batch was defined as an individual 10x run.

CITE-SEQ DATA SET. The CITE-seq data set was extracted from GEO entry GSE100866 (CBMC_8K_13AB_10X), for a collection of human CBMCs and mouse cells [87]. 8,005 cells were called as human based on greater than 90% human-mapped UMI fraction. After filtering, 7,978 cells and 7,231 genes were retained. For `scone`, I provided default negative control genes from the "housekeeping" list and positive control genes as the top 513 most common genes annotated in the MSigDB C7 immunological signature collection (`http://software.broadinstitute.org/gsea/msigdb/collections.jsp`). `Seurat`-derived clusters [136] were used as a biological condition (see below). QC features were limited to the fraction of human, mouse, and ERCC UMIs (three features).

A.2 SEURAT CLUSTERING ANALYSES

It is common to see clear biological clustering at early stages of a single-cell analysis (e. g. major blood cell types). However, an important asset of single-cell approaches is their ability to resolve deeper and more subtle biological heterogeneities. Thus, one might wish to maintain the large-scale clustering evident in loosely normalized data, by passing this clustering to `scone` as a biological classification to be preserved after normalization.

I took this approach for the two largest data sets, namely, the 10x PBMC and CITE-seq CBMC data sets. After cell filtering, I loaded the UMI matrices for these data sets into the widely-used `Seurat` analysis pipeline [136]. Following TC normalization, log-transformation, scaling, and PCA, I clustered the cells in the first 10 PC at a "resolution" of 0.6. The resulting clusters were treated as biological conditions for evaluating the biological cluster tightness.

For the PBMC data set, I manually collapsed clusters based on expression of the following marker genes: IL7R (CD4$^+$ T cells), CD14 and LYZ (CD14$^+$ monocytes), MS4A1 (B cells), CD8A (CD8$^+$ T cells), FCGR3A and MS4A7 (FCGR3A$^+$ monocytes), GNLY and NKG7 (NK cells), FCER1A and CST3 (DCs), and PPBP (megakaryocytes). These RNA markers were discussed as part of the official `Seurat` vignette.