Discovering Gene Expression Programs Underlying the Host Response to Ebola Virus Infection With Single-Cell RNA-Seq

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365527

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Discovering gene expression programs underlying the host response to Ebola virus infection with single-cell RNA-Seq

ABSTRACT

The regulation of gene expression determines the molecular tools a cell has at its disposal to perform its basic functions, respond to its environment, and replicate itself and its genetic material. With the development of single-cell transcriptomics, we have an unprecedented ability to observe the gene expression of individual cells and how it varies across cell types and environmental conditions. However, a great challenge of biology remains to understand the complex biochemical circuitry that underlies expression and how it gets corrupted in disease.

In this thesis, I describe two projects utilizing single-cell transcriptomics to characterize and interpret gene expression. In the first project, we develop a tool for exploring single-cell data, consensus non-negative matrix factorization (cNMF), which identifies signatures of co-varying genes and their relative contributions to each cell. Unlike hard clustering, cNMF models each cell as a weighted mixture of expression signatures, allowing it to account for multiple independent axes of variation within cell. We show with simulated data and in several real datasets that the signatures inferred by cNMF can frequently be interpreted as identity programs that are characteristic of discrete cell types, or “activity” programs that are induced in cells carrying out specific activities (e.g. cell replication, adaptation to a hypoxic environment, or responding to signals from neighboring cells). We identify unexpected ac-
tivity programs that reproduce across multiple datasets, highlighting cNMF’s ability to draw out the latent sources of variation within cells in a tissue.

In the second project, we characterize changes in the transcriptomes and in selected protein markers of circulating immune cells in a lethal challenge model of Ebola virus in Rhesus Macaque. We describe the dynamics of the host response in each circulating immune cell type, and relate these to the clinical phenomena in Ebola virus disease. For example, T-cells acquire an apoptosis expression signature consistent with observed lymphopenia, and monocytes lose expression of MHC class II proteins, which could underlie the lack of an adaptive immune response observed in severe clinical cases. Moreover, as Ebola virus has an RNA genome and makes poly-adenylated mRNA, we detect infected cells and identify host expression markers associated with tropism. We find that monocytes expressing a macrophage-differentiation program and co-expressing CD14 and CD16 are the predominant target of EBOV amongst circulating cells, in vivo. Finally, while virtually all uninfected cells express the antiviral interferon response program, we show that EBOV suppresses this program in infected cells while up-regulating putative pro-viral genes such as DYNLL1 and HSPA5. We thus demonstrate that Ebola virus shapes the gene expression of infected cells to evade the antiviral response and enhance its own replication.

In summary, this thesis describes two projects centered around the analysis of single-cell data. The first develops a general approach for discerning gene expression programs underlying cell type and cellular activities. The second employs this method as part of a general analysis of the dynamics of host response to Ebola virus infection. This body of work constitutes distinct advances in our ability to interpret transcriptomic data and our specific understanding of the immune response to Ebola virus.
Table of contents

Abstract iii
List of figures ix
Dedication xi
Acknowledgments xii
Collaborator contributions xiv

1 Introduction 1
  1.1 The programmatic view of biology 2
  1.2 Inference of programs from gene expression data 8

2 Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq 13
  2.1 Abstract 13
  2.2 Introduction 14
  2.3 Results 20
    2.3.1 Evaluation of matrix factorization for GEP inference in simulated data 20
    2.3.2 cNMF deconvolutes hypoxia and cell-cycle activity GEPs from identity GEPs in brain organoid data 25
    2.3.3 cNMF identifies depolarization induced and novel activity programs in scRNA-Seq of mouse visual cortex neurons 30
  2.4 Discussion 35
  2.5 Materials and methods 40
    2.5.1 Simulations 40
    2.5.2 Data preprocessing 42
    2.5.3 Consensus Non-negative Matrix Factorization (cNMF) 43
    2.5.4 Identification of marker genes 49
    2.5.5 Choosing the number of components 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.6</td>
<td>Comparison of cNMF with other methods</td>
<td>53</td>
</tr>
<tr>
<td>2.5.7</td>
<td>Testing enrichment of gene sets in programs</td>
<td>54</td>
</tr>
<tr>
<td>2.5.8</td>
<td>Data availability</td>
<td>55</td>
</tr>
<tr>
<td>2.5.9</td>
<td>Code availability</td>
<td>55</td>
</tr>
<tr>
<td>2.6</td>
<td>Acknowledgments</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>Single-cell profiling defines dynamics of host response and viral</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>replication in Ebola virus infected rhesus macaques</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Abstract</td>
<td>56</td>
</tr>
<tr>
<td>3.2</td>
<td>Introduction</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>Results</td>
<td>61</td>
</tr>
<tr>
<td>3.3.1</td>
<td>A single-cell atlas of RNA and protein expression in circulating</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>immune cells of Ebola virus infected rhesus macaques</td>
<td></td>
</tr>
<tr>
<td>3.3.2</td>
<td>Interferon response drives a gene expression program shared across</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>multiple cell types</td>
<td></td>
</tr>
<tr>
<td>3.3.3</td>
<td>Cell-type and/or temporally-specific modules help explain cell states</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>and clinical phenomena in EVD</td>
<td></td>
</tr>
<tr>
<td>3.3.4</td>
<td>Monocytes lose expression of MHC class II mRNAs and proteins</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>independent of infection status</td>
<td></td>
</tr>
<tr>
<td>3.3.5</td>
<td>Interferon stimulated genes are down-regulated in infected monocytes</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>relative to bystanders</td>
<td></td>
</tr>
<tr>
<td>3.3.6</td>
<td>Emergence of CD14loCD16lo and CD14hiCD16hi monocyte subsets and</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>enrichment of EBOV infection in the CD14hiCD16hi compartment</td>
<td></td>
</tr>
<tr>
<td>3.3.7</td>
<td>Monocytes expressing markers of macrophage differentiation are</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>enriched for EBOV infected cells</td>
<td></td>
</tr>
<tr>
<td>3.3.8</td>
<td>scRNA-Seq of PBMCs infected ex vivo provides high resolution view of</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Ebola infection dynamics</td>
<td></td>
</tr>
<tr>
<td>3.3.9</td>
<td>Ebola infection down-regulates host antiviral genes and up-regulates</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>putative pro-viral genes</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>92</td>
</tr>
<tr>
<td>3.5</td>
<td>Materials and methods</td>
<td>99</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Natural history study</td>
<td>99</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Clinical observations and scoring</td>
<td>99</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Whole blood collection</td>
<td>100</td>
</tr>
<tr>
<td>3.5.4</td>
<td>Hematology and complete blood counts</td>
<td>100</td>
</tr>
<tr>
<td>3.5.5</td>
<td>EBOV viral load by RT-qPCR</td>
<td>101</td>
</tr>
<tr>
<td>3.5.6</td>
<td>PBMC purification</td>
<td>101</td>
</tr>
<tr>
<td>3.5.7</td>
<td>Seq-Well</td>
<td>102</td>
</tr>
<tr>
<td>3.5.8</td>
<td>DASH of select Seq-Well libraries</td>
<td>103</td>
</tr>
<tr>
<td>3.5.9</td>
<td>CyTOF</td>
<td>103</td>
</tr>
<tr>
<td>3.5.10</td>
<td>De-multiplexing the raw scRNA-Seq data</td>
<td>104</td>
</tr>
</tbody>
</table>
4 General Discussion 111
  4.1 Overview of the previous chapters 111
  4.2 Interpreting the function of sets of co-expressed genes 114
  4.3 Identifying sub-modules that co-vary in most natural contexts 120

APPENDIX A Supplementary Note and Figures for Chapter 2 124
  A.1 Supplementary Note - Analysis of between-sample variability 124
  A.2 Supplementary Figures and Tables 127

APPENDIX B Supplementary Figures for Chapter 3 145
  B.1 Supplementary Figures and Tables 145

References 180
### List of figures

2.1 Schematics of matrix factorization for single-cell RNA-Seq analysis and the consensus matrix factorization pipeline ........................................... 17
2.2 cNMF infers identity and activity expression programs in simulated data ................................................................. 22
2.3 Deconvolution of activity programs from cell identity in brain organoid data ....................................................... 27
2.4 Identification of activity programs in neurons of the visual cortex ................................................................. 32
2.5 Robustness of cNMF to the number of components (K) ................................................................. 37

3.1 Studying the natural history of Ebola virus disease in rhesus macaques using single-cell profiling of RNA and protein expression in circulating immune cells ................................................... 62
3.2 Single-cell measurements reveal changing cell-type abundance, proliferation, and infection status throughout EVD ................................................... 64
3.3 Patterns of differential expression across disease stages and cell types in EVD ................................................... 69
3.4 Monocytes reduce expression of MHC class II proteins independent of infection status ................................................... 73
3.5 ISG suppression, co-expression of CD14 and CD16, and expression of macrophage-linked genes, are associated with monocyte infectivity ................................................... 77
3.6 Viral transcriptional dynamics of infected monocytes in vivo and ex vivo ................................................... 86
3.7 Ebola virus infection down-regulates host antiviral genes and up-regulates putative pro-viral genes ................................................... 89

A.1 Robustness of matrix factorization methods ................................................... 128
A.2 Deconvolution accuracy of matrix factorization methods ................................................... 129
A.3 Diagnostic plot for cNMF on an example simulated dataset ................................................... 130
A.4 GEP usage inference ................................................... 131
A.5 Accuracy of identifying genes in each GEP ................................................... 132
A.6 Comparison of run-times for different matrix factorization algorithms ................................................... 133
A.7 cNMF demonstration on simulated dataset with 50% doublets ................................................... 134
A.8 Impact of variable cell-type proportions on GEP inference ................................................... 135
A.9 Diagnostic plot for cNMF on the Quadrato et al., 2017 brain organoid dataset ................................................... 136
A.10 Correlation between GEP spectra pairs and fraction of cells that use both programs ................................................... 137
A.11  t-SNE plots of identity and activity GEPs in the Quadrato et al., 2017 brain organoid dataset ................................. 138
A.12  Comparison of cNMF usages with the cell-type clusters from Quadrato et al., 2017 .................................................. 139
A.13  Diagnostic plot for cNMF on the Hrvatini et al., 2018 visual cortex dataset ......................................................... 140
A.14  Comparison of cNMF usages with the cell-type clusters from Hrvatini et al., 2018 ................................................. 141
A.15  Comparison of GEPs identified in the Hrvatini et al., 2018 and Tasic et al., 2016 visual cortex datasets ...................... 142
A.16  Characterization of GEP usage across biological replicates .......................................................... 143
A.17  Comparison of cNMF usages with the published cell-type clusters from Baron et al., 2016 ................................. 144

B.1  Blood sampling overview and clinical time course per animal replicate ................................. 147
B.2  Cell type markers for Seq-well and CyTOF clusters ................................................................. 148
B.3  UMAP embeddings of in vivo data colored by Seq-well samples and CyTOF batches ................................. 149
B.4  Estimates of cell type abundance over the time course ................................................................. 150
B.5  Estimates of cell type replication rate over the time course ................................................................. 151
B.6  Receiver operating characteristic for identifying Ebola virus infected cells ................................................................. 152
B.7  Quantification of cytokine expression and enrichment of response signatures ................................................................. 154
B.8  Extended characterization of gene expression signals associated with EBOV infection status in monocytes ................................................................. 155
B.9  Relationship between monocyte subsets and differentiation score ................................................................. 156
B.10  Overview of the ex vivo Ebola virus infection dataset ......................................................................... 157
B.11  Ebola virus infection dynamics in the ex vivo dataset ......................................................................... 159
FOR SABA AND SAVTA AND GRANDMA AND POP
Acknowledgments

First of all thank you to my parents, my grand parents, my aunts and uncles and cousins, and my brother Ian. I really couldn’t have done this without you;

Thank you to my advisor Pardis Sabeti. You’re an incredible mentor and I am constantly in awe of your resilience, your ability to find the win-wins in a situation, your insightfulness, and your ability inspire greatness in the people around you. I am also immensely grateful to everyone who has overlapped with me in the lab, you are truly exceptional people and it has been a privilege to know each one of you;

Thank you to my undergraduate mentor Dana Pe’er for believing in me and giving me the opportunity to start down this scientific path. You, Oren Litvin, Uri-david Akavia, and Felix Sanchez-Garcia were my first scientific role models. Thank you for sharing your wisdom about how to analyze data and your grand visions about how much we can learn with a good model and sound statistics;

Thank you Kayla Barnes, Aaron Lin, Adrian Veres, Eran Hodis, Shervin Tabrizi, Ryan Tewhey, Jacob Lemieux, and Gordon Adams–my direct scientific partners in one or more projects and my good friends. Science is truly a team sport and it has been an incredibly rewarding experience to play
it with you all;

Thank you to the CGTA discussion group—Adrian Veres, Eran Hodis, Yakir Reshef, and Aurel Nagy, for lifting me up, for being my friends and my scientific role models. Thank you to my MD/PhD and HST classmates, I am honored to be included among your ranks;

Thank you Stephen Blacklow, Thomas Michel, Loren Walensky, Amy Cohen, Richard Mitchell, Patty Cunningham, and all of the others who keep the MD/PhD and HST programs running. Thank you for inviting me into your community, and for creating such a rich environment for exploring biology and medicine. Thank you Allon Klein, Yonatan Grad, Joel Hirschorn, Sean Whelan, Jon Bloom, Sam Wolock, Eric Lander, Aviv Regev, Gad Getz, and so many others for providing excellent feedback on my projects at decisive moments: your advice helped steer me in the right direction;

Finally, thank you to all of my friends who have kept me afloat over the past decade. There are too many to name but I’ll specifically list Vincent Troyani, Micah Urbach, Nick Vanarthos, Rachel Leslie, Dan Schwartz, Tripp Odom, Shai Chester, Lee Cohen, Audrey Baker, Bridget Chak, Andres Colubri, Amy Chen, Nicole Torres, Alina Xu, and so many others, thank you for your love and support over the years. It has really meant the world to me.
Collaborator contributions

1. Chapter 1 was written and researched by Dylan Kotliar (DK)

2. Chapter 2 derived from the hard work of the following individuals: Dylan Kotliar (DK), Adrian Veres (AV), M Aurel Nagy (MAN), Shervin Tabrizi (ST), Eran Hodis (EH), Douglas A Melton (DAM), Pardis C Sabeti (PCS). DAK and AV lead and conceived of the project, and carried out the data analysis and interpretation. The project was supervised by PCS and DAM. MAN provided and helped with interpretation of the visual cortex data. EH and ST helped with data analysis during an early version of the project that impacted the final method. DK, AV, PCS, and DAM wrote the manuscript.

3. Chapter 3 derived from the hard work of the following individuals: Dylan Kotliar (DK), Aaron E. Lin (AEL), James Logue (JL), Kayla G. Barnes (KGB), Travis Hughes (TH), Siddharth S. Raju (SSR), Nadine Khoury (NK), Marc H. Wadsworth II (MHW), Garry P. Nolan (GN), Richard S. Bennett (RSB), David W. McIlwain (DWM), Alex K. Shalek (AKS), Lisa E. Hensley (LEH), Pardis C. Sabeti (PCS). DAK and AEL lead and conceived of the project, designed the ex vivo experiment, and were supervised by PCS. DAK carried out most of the data analysis and DAK and AEL interpreted the data, and wrote the manuscript. SSR contributed analysis of cytokine expression levels. LEH supervised and JL and RSB conducted the non-human primate work including Seq-well loading and cell staining for CyTOF. AEL lead the experimental work for sequencing of single cell libraries with support from KGB and NK. TH, MHW, and AKS provided Seq-well reagents, helped with troubleshooting, and providing training for running Seq-well. DWM and GN planned the CyTOF data collection and DWM oversaw the CyTOF run.

4. Chapter 4 was written and researched by DAK
Introduction

If your question is what is a mouse and how does it come about to be a mouse, it may be that the only way to give the answer to this is to specify an algorithm for how you could build a mouse in the way the mouse builds itself... I believe that in biology, programmatic explanations will be algorithmic explanations. You will have to say, next switch on gene group number fifty-eight. And then one has that whole lot of molecular biology—what is gene group fifty-eight and what does it do. And one takes for granted that gene group fifty-eight performs its computation, and then the
simultaneous steps, the alternative steps, the sequential steps in great detail... Molecular biology has to go in this direction--to explore the high level logical computations, the programs, the algorithms of development in molecular terms... to move between the molecular hardware and the logical software of how it’s all organized, without feeling they are different sciences. –Sidney Brenner, 1976 (Judson, 1979)

1.1 The programmatic view of biology

The human genome is 3 billion nucleotides long, and includes 20,000 protein coding and 40,000 non-coding genes, that get spliced into over 200,000 distinct mRNA transcripts (Frankish et al., 2019) and translated into 100,000 proteoforms (Aebersold et al., 2018). However, these fixed numbers don’t adequately reflect the immense biological complexity of a human, from the myriad cell-types to the unparalleled computing power of the central nervous system. Biological complexity is not a direct consequence of an organism’s genome size or number of genes. In fact, Caenorhabditis elegans, the modest nematode worm has more protein coding genes than humans, and the common onion, Allium cepa, has a genome that is more than 4x longer than that of humans (Palazzo and Gregory, 2014).

The complexity of mammalian life is better reflected by the enormous amount of information that gets encoded in a genome of fixed size. All of the cells in a human–bacteria-killing macrophages and signal-processing neurons alike–operate based on the same set of instructions in the genome, but they read different parts of these instructions, and interpret them in distinct ways, endowing them with
tremendously different behaviors and appearances.

The ability of cells to perform different functions based on the same set of genetic material is a consequence of gene expression regulation, the process of deciding which genes get expressed as RNA and protein, and at what levels. There are believed to be hundreds of thousands of DNA regulatory elements that act as “switches” and “knobs”, increasing or decreasing the rates of RNA synthesis for specific nearby genes depending on whether or not they are bound by their matching transcription factors (TFs) (ENCODE Project Consortium, 2012). Transcription of a gene thus relies on the necessary TFs being themselves present in the cell (and hence the necessary conditions for their transcription being met). Moreover, TFs frequently must be activated through post translational modifications or binding of a co-factor before they can move to the nucleus and transcribe their targets. And finally, the target DNA element must be accessible in order to be bound by the TF—i.e. not blocked by nucleosomes, contained within heterochromatin, or already bound by competing proteins. Whether or not the necessary criteria for transcription are met can depend on specific signals from the environment, for example whether or not a receptor has bound its ligand which transduces a signal to phosphorylate the TF. In addition, the ability to transcribe a gene frequently depends on the specific developmental lineage that gave rise to the cell, which shapes the regions of the genome that are potentially accessible to TF binding.

The activation of TFs and their binding to DNA are just a small set of the vast number of physical interactions between biological molecules in a cell that cumulatively determine which genes get transcribed and translated, and how the gene products will function together to carry out the cell’s activities. The set of all physical relationships between biological molecules that can be assigned a role in
maintaining the cell’s functioning is frequently referred to as the cellular circuitry. This is an analogy with computer hardware, acknowledging that the physical interactions between biomolecules encodes the logic and implements the algorithms that determine the cell’s behavior.

The task of enumerating all of the components of the cellular circuitry is daunting. Whereas a relatively small set of individual transistors can be combined in an electrical circuit to generate hardware capable of complex computations, there are hundreds of thousands of distinct components in a biological system, each with its own specific biochemistry. Furthermore, there are many levels in a cell’s circuitry such as epigenetic modifications to DNA, RNA and protein levels, and post-translational modifications to proteins, and there are dense levels of physical interactions within and between all layers of the network (Mitra et al., 2013; Kolch et al., 2015). For example, many proteins physically interact with other proteins in complexes that can impact their activity in crucial ways (Barabási and Oltvai, 2004). Furthermore, the transcription of most genes is regulated by combinatorial binding of multiple transcription factors and cofactors at a DNA locus, and those interactions can have complex effects on transcription (Reiter et al., 2017). The number of regulatory relationships playing a role in the cell’s circuitry is currently unknown, but presumably it could be extremely large, surpassing what can be comprehensively perturbed and measured with our current tools, in a reasonable timeframe.

However, as the quote from Sidney Brenner at the opening of this chapter suggests, there is an alternative conception of the cell that abstracts the detailed biochemistry of each individual molecule and instead considers a set of “programs” that are executed to carry-out specific functions in the cell. This has been described as a “modular” view of the cell: it conceives of distinct programs that get composed and sequentially ordered in different combinations to generate the diversity of cellular be-
behaviors observed in an organism. Cellular programs and the biological molecules that constitute them, are assumed to be relatively insulated from one another, either through spatial isolation in the cell, as is the case with multi-protein complexes such as the ribosome, or through chemical specificity, such as the specific interactions that govern a signal transduction cascade (Hartwell et al., 1999). This insulation allows programs to be combined in new combinations without disrupting their individual functions. The conception of a modular cell makes an analogy with software design: complex computer programs are composed by combining simpler reusable sub-programs, that can be counted on to do their job in any context, without interfering with the functioning of other sub-programs. As with software, more complicated biological programs could be constructed by combining multiple sub-programs. For example, a DNA replication program could involve calls to sub-programs such as nucleotide metabolism, DNA helicase activity, DNA repair, and DNA ligation (Ashburner et al., 2000).

Modularity in biology could help explain the evolvability of complex structures in the same way that it explains how software can become increasingly complex while remaining maintainable. Software engineers can tinker with a single program while generally assuming that other programs will continue to work as expected. This is because discrete programs are only connected through a few well-demarcated input and output variables. Modularity in biology would expect most genes to function primarily within a specific program which allows their role to be enhanced by evolution without negative consequences for other programs. A relatively smaller number of genes would connect programs to couple their regulation or function in various ways. New links between programs could be formed or links could be removed by specifically mutating those linker genes, without disrupting the
core program functionality. Genes that are highly pleiotropic would be the most difficult to evolve to enhance one function, as that could likely have a negative impact on other functions (Hartwell et al., 1999).

If a set of gene products must cooperate to carry-out a function, as is the case with enzymes in a biosynthetic pathway or components of a multi-protein complex, one would expect that they should be coordinately regulated at the transcriptional level to ensure the appropriate stoichiometric ratios are maintained. This would reduce the chance that one or more components of the program would be missing which could render the entire program ineffective, resulting in wasted energy, or production of toxic byproducts.

The expectation of transcriptional co-regulation of co-functional genes was born out with the advent of high-throughput genomics technologies that allowed for simultaneous quantification of expression levels of thousands of genes. In early expression profiling experiments, it was observed that genes that function as part of a complex such as the ribosome, or as part of a common process such as cell-division, were coordinately regulated in response to various stimuli. A specific stimulus applied to cells would result in changes to many genes, some of which shared known functions and other of which did not. However, looking across a large number of stimuli, the signals could be dissected into several distinct programs based on patterns of correlated gene expression profiles. The term “gene expression program” was thus coined to describe sets of genes that are coordinately regulated transcriptionally in order for their gene products to act cooperatively to carry out a distinct function.

The concept of the gene expression program is exciting because it suggests that we can begin to understand the high level organization of cells, the algorithm of how a “mouse builds itself” as Brenner
puts it, without first understanding all of the complicated biochemistry of every component in the program.

However, even if we could hypothetically predict the consequence of any cellular perturbation on gene expression with a black-box predictor, it is likely that we would still prefer a higher-level explanation in terms of the sequential activation and connections between expression programs. Such an explanation is easier for us to interpret, reason about, and potentially manipulate. Analogously, knowing about the states of each transistor in a circuit is much less interpretable (at least for humans) than reading the computer programs. Reading computer code teaches us about their underlying logic, allowing us to adapt programs for new purposes, or even fix bugs. Perhaps the same can be said of biological programs.

Beyond its interest as a fundamental topic in biology, understanding the structure and organization of gene expression programs could provide insights for pressing translational questions and applications. In studies of complex trait genetics, it has become clear that heritability is generally mediated by the cumulative impact of large numbers of non-coding variants of small effect. While it has proven challenging to interpret the impact of these variants in isolation, it might be possible to investigate their cumulative effect on a gene expression program (Smillie et al., 2019). In addition, understanding the choreographed activation of gene expression programs that occur during development could be very useful for cellular medicine, which seeks to program cells into specific cell types. We typically seek to recapitulate these programs through exogenous administration of sets of transcription factors (Takahashi and Yamanaka, 2006) and knowing the relevant programs from normal differentiation could guide the development of improved protocols (Veres et al., 2019). Finally, certain gene
expression programs might become excessively activated in specific disease states, such as immune exhausition of T-cells in cancer or excessive activation states in autoimmune diseases. The ability to read out which gene expression programs are being activated could provide a deeper understanding of the disease processes that could guide therapeutic development (Rao et al., 2017).

In the next section of the introduction, I provide a brief review of the history of gene expression program inference, beginning with the advent of genomic technologies through to the current era of single-cell transcriptomics. This sets the stage for the data chapters of the thesis which describe a general approach for inferring gene expression programs from single-cell RNA-Seq data, and a specific analysis of the gene expression programs that get induced in circulating immune cells during Ebola virus disease.

1.2 Inference of programs from gene expression data

Many of the earliest and most successful strategies for systematically analyzing gene expression programs were developed in studies of the baker’s yeast, *Saccharomyces cerevisiae* in the late 1990s and early 2000s. *S. cerevisiae* was the first eukaryote to have its complete genome sequenced in 1996 (Goffeau et al., 1996) which enabled the first transcriptome-wide expression profiling or any organism in 1997 (Velculescu et al., 1997). By 2000, transcriptomic profiles had been generated for hundreds of *S. cerevisiae* cultures at multiple time-points following environmental perturbations such as amino acid starvation, temperature shocks, and transfer to glucose-poor media (Gasch et al., 2000). This dataset lay the groundwork for the development of several computational approaches that are still used to
analyze transcriptomic data to this day.

In an early exploration of this data, Eisen et al. used unsupervised clustering to find groups of genes that had correlated expression profiles when viewed across samples in different environmental conditions (Eisen et al., 1998a). This identifies genes that are coordinately up-regulated or down-regulated in response to the same external signals. For example, they found 126 genes including ribosomal proteins, tRNA synthetases, and other components of the protein synthesis machinery were down-regulated in response to the same stress conditions, suggesting that stress leads to the broad down-regulation of a translation program. Other co-expressed genes included proteasome components, histone genes, and components of the tricarboxylic acid cycle. Concurrent work looking at yeast cultures in synchronized cell-cycle stages, found that genes known to be involved in each cell cycle phase were co-expressed and could be clustered into distinct modules (Spellman et al., 1998). For example, histone genes were co-expressed along with genes involved in DNA synthesis (S phase) while proteins involved in cytokinesis such as CTSi were grouped into a distinct co-expression cluster.

Insight into the regulatory mechanism underlying co-expression was found through analyzing the upstream promoter sequences of genes in the same cluster. Spellman et al. found that specific DNA sequence motifs were enriched in the promoters of genes in the same co-expression clusters, many of which corresponded to known transcription factors. For example, 56% of genes in a large cluster associated with G1 phase contained the MluI cell cycle box motif, ACGCGT, which binds the MBF transcription factor complex—far more than would be expected by chance. Systematic discovery of motifs from co-expressed gene modules identified dozens of motifs that could in turn, predict patterns of expression in held out data with impressive accuracy (Beer and Tavazoie, 2004).
The observation of common DNA regulatory motifs and co-expression of genes known to act together to carry-out distinct biological functions demonstrated the empirical reality of the gene expression programs concept in yeast. Since then, similar analyses have become wide-spread in the study of mammalian cells. At the time of this writing, there are 3,457,530 transcriptomic profiles posted on the Gene Expression Omnibus database corresponding to 125,783 distinct experiments. Of these 54% of transcriptomes derive from human and 25% derive from mouse samples. Most of the transcriptomic analyses utilize the differential expression study design, which compares average expression levels of cells in a treatment group to those in a control group. This produces an expression signature defining which genes show statistically significant changes in response to the treatment. Compendia of these differential expression signatures have been compiled in databases such as the Molecular signatures database (MSigDB) (Liberzon et al., 2011) allowing researchers to compare their own differential expression analyses against those of previous experiments.

While the general concept of co-regulated gene expression programs of co-functional genes appears to be consistent in mammalian transcriptomes, discerning and interpreting those functions has proved far more challenging. First of all, the yeast genome only encodes 6000 genes, and because of its powerful tractability for forward and reverse genetics, more than half of the genes had been characterized by genetic or biochemical assays by as early as 2001 (Kumar and Snyder, 2001), and complete gene knock-out libraries had been generated by 2002 (Giaever and Nislow, 2004). By contrast, in 2016, over 35% of human protein-coding genes still lacked even a basic functional annotation (Hu et al., 2016). Thus, when differentially expressed genes are identified in mammalian gene expression datasets, it can be more challenging to interpret the lists of genes to discern whether or not there is a shared function.
In addition, the multi-cellularity of mammals has posed a practical challenge for understanding gene expression programs. Traditional bulk gene expression profiles quantify average expression levels across thousands of cells. While this isn’t a problem for cultures of yeast and other unicellular organisms where cells and relatively homogenous, it can mask considerable heterogeneity within mammalian tissues which are composed of diverse cell-types. For example, many co-expression modules have been identified through transcriptomic profiling of bulk tumor profiles (Hoadley et al., 2014; Cancer Genome Atlas Research Network, 2012, 2011; Berger et al., 2018; Segal et al., 2004). However, these coordinated gene expression changes can reflect variable cell type compositions within the tumor microenvironment (e.g. differential levels of immune cell infiltration) as well as changing states of individual cancer or stromal cells types, and is likely a mixture of the two. This is a significant limitation to the interpretability of a gene expression signature since it is unclear whether genes are actually being transcriptionally regulated or rather reflect the presence of a new cell type in the tissue.

The challenge of disentangling gene expression changes resulting from changing cell type composition and changing cell state has largely been overcome with the invention and wide-spread adoption of single-cell transcriptomic technologies such as single-cell RNA-Seq (scRNA-Seq) technologies. scRNA-Seq approaches rely on the isolation of individual cells, either within microfluidic droplets (Macosko et al., 2015; Klein et al., 2015a) or in microwells (Picelli et al., 2013; Shalek et al., 2013; Gierahn et al., 2017; Hashimshony et al., 2012), amplifying and barcoding their transcriptomes and then reading out the barcoded profiles with traditional RNA sequencing. By directly reading out transcriptome profiles from individual cells, it is possible to broadly characterize the heterogeneity of cells in the system, including both the cell-type composition and the characteristic expression states of each
cell type.

Systematic efforts are now under way to catalogue the defining molecular features associated with each cell type in the mouse (Consortium et al., 2018) and human (Regev et al., 2017), with a particular emphasis on determining the characteristic gene expression profile based on scRNA-Seq. The characteristic transcriptomic profile of a cell type is a high-level program encompassing the set of genes that the cell type uses to maintain homeostasis while performing its particular functions within the organism. It also defines a reference state that can be compared against cells in various disease states such as cancer (Suvà and Tiros, 2019), autoimmunity (Zhang et al., 2019), and viral infection (Steuerman et al., 2018a). By comparing the gene expression profiles of each cell type in disease states to healthy ones, these studies can begin to discern the gene expression programs that are perturbed in disease processes.

In the next two chapters, I describe a general tool for inferring gene expression programs from single-cell transcriptome data, and a specific analysis of the gene expression programs involved in the immune response to Ebola virus infection. This work is very much in the same spirit as the early gene expression clustering analyses in yeast that began to resolve a high-level structure to the regulation of gene expression. By bringing this analysis into the mammalian context, it represents a key step toward the vision reflected in the quote from Sidney Brenner at the beginning of this chapter, “to move between the molecular hardware and the logical software of how it’s all organized, without feeling they are different sciences.”
Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq

2.1 Abstract

Identifying gene expression programs underlying both cell-type identity and cellular activities (e.g. life-cycle processes, responses to environmental cues) is crucial for understanding the organization of cells and tissues. Although single-cell RNA-Seq (scRNA-Seq) can quantify transcripts in individual
cells, each cell’s expression profile may be a mixture of both types of programs, making them difficult to disentangle. Here we benchmark and enhance the use of matrix factorization to solve this problem. We show with simulations that a method we call consensus non-negative matrix factorization (cNMF) accurately infers identity and activity programs, including their relative contributions in each cell. To illustrate the insights this approach enables, we apply it to published brain organoid and visual cortex scRNA-Seq datasets; cNMF refines cell types and identifies both expected (e.g. cell cycle and hypoxia) and novel activity programs, including programs that may underlie a neurosecretory phenotype and synaptogenesis.

2.2 Introduction

Genes act in concert to maintain a cell’s identity as a specific cell type, to respond to external signals, and to carry out complex cellular activities such as replication and metabolism. Coordinating the necessary genes for these functions is frequently achieved through transcriptional co-regulation, where genes are induced together as a gene expression program (GEP) in response to the appropriate internal or external signal (Eisen et al., 1998b; Segal et al., 2003a). By enabling unbiased measurement of the whole transcriptome, profiling technologies such as RNA-Seq are paving the way for systematically discovering GEPs and shedding light on the biological mechanisms that they govern (Liberzon et al., 2015a).

Single-cell RNA-Seq (scRNA-Seq) has greatly enhanced our potential to resolve GEPs by making

*The material in this chapter is adapted from a publication with the same title in eLIFE (Kotliar et al., 2019).
it possible to observe variation in gene expression over many individual cells. Even so, inferring GEPs remains challenging as scRNA-Seq data is noisy and high-dimensional, requiring computational approaches to uncover the underlying patterns. In addition, technical artifacts such as doublets (where two or more distinct cells are mistakenly collapsed into one) can confound analysis. Methodological advances in dimensionality reduction, clustering, lineage trajectory tracing, and differential expression analysis have helped overcome some of these issues (Amir et al., 2013; Kharchenko et al., 2014; Satija et al., 2015; Trapnell et al., 2014).

Here, we focus on a key challenge of inferring expression programs from scRNA-Seq data: the fact that individual cells may express multiple GEPs but we only detect cellular expression profiles that reflect their combination, rather than the GEPs themselves. A cell’s gene expression is shaped by many factors including its cell type, its state in time-dependent processes such as the cell cycle, and its response to varied environmental stimuli (Wagner et al., 2016). We group these into two broad classes of expression programs that can be detectable in scRNA-Seq data: (1) GEPs that correspond to the identity of a specific cell type such as hepatocyte or melanocyte (identity programs) and (2) GEPs that are expressed independently of cell type, in any cell that is carrying out a specific activity such as cell division or immune cell activation (activity programs). In this formulation, identity programs are expressed uniquely in cells of a specific cell type, while activity programs may vary dynamically in cells of one or multiple types and may be continuous or discrete.

Thus far, the vast majority of scRNA-Seq studies have focused on systematically identifying and characterizing the expression programs of cell types composing a given tissue, i.e. identity GEPs. Substantially less progress has been made in identifying activity GEPs, primarily through direct manipula-
tion of cells in controlled experiments, for example comparing stimulated and unstimulated neurons (Hrvatin et al., 2018a) or cells pre- and post-viral infection (Steuerman et al., 2018b).

If a subset of cells profiled by scRNA-Seq expresses a given activity GEP, there is a potential to directly infer the program from the data without the need for controlled experiments. However, this can be significantly more challenging than ascertaining identity GEPs; while some cells may have expression profiles that are predominantly the output of an identity program, activity programs will always be expressed alongside the identity programs of one or frequently many cell types. Thus, while finding the average expression of clusters of similar cells may often be sufficient for finding reasonably accurate identity GEPs, it will often fail for activity GEPs.

We hypothesized that we could infer activity GEPs directly from variation in single cell expression profiles using matrix factorization. In this context, matrix factorization would model the gene expression data matrix as the product of two lower rank matrices, one encoding the relative contribution of each gene to each program, and a second specifying the proportions in which the programs are combined for each cell. We refer to the second matrix as a ‘usage’ matrix as it specifies how much each GEP is ‘used’ by each cell in the dataset (Stein-O’Brien et al., 2018) (Figure 1a). Unlike hard clustering, which reduces all cells in a cluster to a single shared GEP, matrix factorization allows cells to express multiple GEPs. Thus, this computational approach would allow cells to express one or more activity GEPs in addition to their expected cell-type GEP, and could correctly model doublets as a combination of the identity GEPs for the combined cell types. To the best of our knowledge, no previously reported studies have benchmarked the ability of matrix factorization methods to accurately learn identity and activity GEPs from scRNA-Seq profiles.
Figure 2.1: Schematics of matrix factorization for single-cell RNA-Seq analysis and the consensus matrix factorization pipeline

(a) Schematic representation of how cellular gene expression can be modeled with matrix factorization. The gene expression profiles of individual cells (left bar charts) are weighted mixtures of a set of global gene expression programs (right bar charts) with distinct weights reflecting the usage of each program (middle pie chart). Varying levels of usage of the activity programs is represented by the number of blue circles and orange squares in each cell.

(b) Schematic of the consensus matrix factorization pipeline.
We see three primary motivations for jointly inferring identity and activity GEPs in scRNA-Seq data. First, systematic discovery of GEPs could reveal unexpected or novel activity programs reflecting important biological processes (e.g. immune activation or hypoxia) in the context of the native biological tissue. Second, it could enable characterization of the prevalence of each activity GEP across cell types in the tissue. Finally, accounting for activity programs could improve inference of identity programs by avoiding spurious inclusion of activity program genes in the latter. GEPs corresponding to different phases of the cell cycle are examples of widespread activity programs and are well-known to confound identity (cell type) program inference in scRNA-Seq data (Chen and Zhou, 2017; Scialdone et al., 2015). However, cell-cycle is just one instance of the broader problem of confounding of identity and activity programs.

While matrix factorization is widely used as a preprocessing step in scRNA-Seq analysis, a priori it is unclear which, if any, factorization approaches would be most appropriate for inferring biologically meaningful GEPs. In particular, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999) have been used for dimensionality reduction of data prior to downstream analysis or as an approach to cell clustering. However, while PCA (Shalek et al., 2014; Steuerman et al., 2018b), NMF (Puram et al., 2017) and ICA (Saunders et al., 2018) components have been interpreted as activity programs, the dimensions inferred by these or other matrix factorization algorithms may not necessarily align with biologically meaningful gene expression programs and are frequently ignored in practice. This is because each method makes different simplifying assumptions that are potentially inappropriate for gene expression data. For example, NMF and LDA are non-
negative and so cannot directly model repression. ICA components are statistically independent, PCA components are mutually orthogonal, and both allow gene expression to be negative. Furthermore, none of these methods, except LDA, explicitly accounts for the count distribution of expression data in their error models.

In this study, we motivate, validate, and enhance the use of matrix factorization for GEP inference. Using simulations, we show that despite their simplifying assumptions, ICA, LDA, and NMF—but not PCA—can accurately discover both activity and identity GEPs. However, due to inherent randomness in their algorithms, they give substantially varying results when repeated multiple times, which hinders their interpretability. We therefore implemented a meta-analysis approach (Figure 1b), which demonstrably increased robustness and accuracy. Overall, the meta-analysis of NMF, which we call Consensus NMF (cNMF), gave the best performance in these simulations.

Applied to three real datasets generated by 3 different scRNA-Seq platforms, cNMF inferred expected activity programs (cell-cycle programs in a brain organoid dataset and depolarization induced programs in visual cortex neurons), an unanticipated hypoxia program, and intriguing novel activity programs. It also enhanced cell type characterization and enabled estimation of rates of activity across cell types. These findings on real datasets further validate our approach as a useful analysis tool to understand complex signals within scRNA-Seq data.
2.3 Results

2.3.1 Evaluation of matrix factorization for GEP inference in simulated data

We sought to establish whether components inferred by simple matrix factorizations would align with GEPs in scRNA-seq data. We evaluated this in simulated data of 15,000 cells composed of 13 cell types, one cellular activity program that is active to varying extents in a subset of cells of four cell types, and a 6% doublet rate (Figure 2a). We generated 20 replicates of this simulation, each at three different ‘signal to noise’ ratios, in order to determine how matrix factorization accuracy varies with noise level (Materials and methods).

We first analyzed the performance of ICA, LDA, and NMF and noticed that they yielded different solutions when run several times on the same input simulated data. We ran each method 200 times and assigned the components in each run to their most correlated ground-truth program. We saw that there was significant variability among the components assigned to the same program – particularly for NMF and LDA (Figure A.1). Unlike PCA, which has an exact solution, these factorizations use stochastic optimization algorithms to obtain approximate solutions in a solution space including many local optima. We observed that such local optima frequently corresponded to solutions where a simulated GEP was split into multiple inferred components and/or multiple GEPs were merged into a single component (Figure A.2a). This variability reduces the interpretability of the solutions and may decrease the accuracy as well.
To overcome the issue of variability of solutions, we employed a meta-analysis approach, which we call consensus matrix factorization, that averages over multiple replicates to increase the robustness of the solution. The method which is adapted from a similar procedure in mutational signature discovery (Alexandrov et al., 2013) proceeds as follows: we run the factorization multiple times, filter outlier components (which tend to represent noise or merges/splits of GEPs), cluster the components over all replicates combined, and take the cluster medians as our consensus estimates. With these estimates fixed, we are able to compute a final usage matrix specifying the contribution of each GEP in each cell and to transform our GEP estimates from normalized units to biologically meaningful ones such as transcripts per million (TPM). This approach also provides us with a guide for determining K, the number of components to use, by selecting a value that provides a reasonable trade-off between error and stability (Figure A.3a, see Methods for details). We refer to this approach as consensus matrix factorization based on its analogy with consensus clustering (Monti et al., 2003) and to its application to LDA, NMF, and ICA, as cLDA, cNMF, and cICA respectively. While consensus clustering has been previously applied to bulk gene expression analysis using hard-clustering derived by binarizing NMF factors (Brunet et al., 2004), our approach does not require any hard cluster assignments.

Consensus matrix factorization inferred components underlying the GEPs as well as which cells expressed each GEP (Figure 2b-c, A.4a). By contrast, principal components were linear combinations of the true GEPs. Beyond increasing the robustness of the solution, the consensus approach also increased the ability of factorization to deconvolute the true GEPs - most dramatically for LDA and NMF which had the most stochastic variability. cNMF successfully deconvoluted the activity and identity GEPs more frequently than the other matrix factorizations considered (Figure 2d, A.2b-c).
Figure 2.2: cNMF infers identity and activity expression programs in simulated data
(a) t-distributed stochastic neighbor embedding (tSNE) plot of an example simulation showing different cell types with marker colors, doublets as gray Xs, and cells expressing the activity gene expression program (GEP) with a black edge.
(b) Pearson correlation between the true GEPs and the GEPs inferred by cNMF for the simulation in (a).
(c) Same tSNE plot as (a) but colored by the simulated or the cNMF inferred usage of an example identity program (left) or the activity program (right).
(d) Percentage of 20 simulation replicates where an inferred GEP had Pearson correlation greater than 0.80 with the true activity program for each signal to noise ratio (parameterized by the mean log2 fold-change for a differentially expressed gene).
(e) Receiver Operator Characteristic (except with false discovery rate rather than false positive rate) showing prediction accuracy of genes associated with the activity GEP.
(f) Scatter plot comparing the simulated activity GEP usage and the usage inferred by cNMF for the simulation in (a). For cells with a simulated usage of 0, the inferred usage is shown as a box and whisker plot with the box corresponding to interquartile range and the whiskers corresponding to 5th and 95th percentiles.
(g) Contour plot of the true GEP usage on the Y-axis and the second true GEP usage for doublets or the second highest GEP usage inferred by cNMF for singletons for the simulation in (a). 1000 randomly selected cells are overlayed as a scatter plot for each group.
We next sought to benchmark the sensitivity and specificity of each matrix factorization method for inferring which genes are associated with each GEP. We also evaluated the performance of hard clustering for this task because clustering is the most common way GEPs are identified in practice. We evaluated the commonly used Louvain community detection clustering algorithm (Blondel et al., 2008a; Levine et al., 2015) but also considered an upper bound on how well any discrete clustering could perform by using ground-truth to assign cells to a cluster of its cell type or to an activity cluster if it had \(\geq 40\%\) simulated contribution from the activity GEP (Figure A.4b). We evaluated the association between genes and GEPs using linear regression and measured accuracy using a receiver operator characteristic (Materials and methods).

We found that cNMF was most accurate at inferring genes in the activity program, with a sensitivity of 61\% at a false discovery rate (FDR) of 5\% (Figure 2e). cICA and the ground-truth clustering were the next most accurate with 57\% and 56\% sensitivity at a 5\% FDR respectively. cNMF also performed the best at inferring identity GEPs of the 4 cell types that expressed the activity (A.5). As expected, the clustering approaches performed worse as they inappropriately assigned activity GEP genes to these identity programs, resulting in an elevated FDR. This illustrates how matrix factorization can outperform clustering for inference of the genes associated with activity and identity GEPs.

We decided to proceed with cNMF to analyze the real datasets due its accuracy, processing speed, and interpretability. First, it yielded the most accurate inferences in our simulated data. Second, NMF was the fastest of the basic factorization algorithms considered, which is especially useful given the need to run multiple replicates and given the growing sizes of scRNA-Seq datasets (Figure A.6). Third, the non-negativity assumption of NMF naturally results in usage and component matrices that can be
normalized and interpreted as probability distributions—i.e. where the usage matrix reflects the probability of each GEP being used in each cell, and the component matrix reflects the probability of a specific transcript expressed in a GEP being a specific gene. The other high performing factorization method, cICA, produced negative values in the components and usages which precludes this interpretation.

Beyond identifying the activity program itself, we found that cNMF could also accurately infer which cells expressed the activity program and what proportion of their expression was derived from the activity program (Figure 2f). With an expression usage threshold of 10%, cNMF accurately classified 91% of cells expressing the activity program and 94% of cells that did not express the program. Moreover, we observed a high Pearson correlation between the inferred and simulated usages in cells that expressed the program (R=0.74 for all simulations combined, R=0.68 for the example simulation in Figure 2a). Thus, cNMF can be used both to infer which cells express the activity program, as well as what proportion of their transcripts derive from that program.

We further demonstrated that cNMF was robust to the presence of doublets—instances where two cells are mistakenly labeled as a single cell. Due to limitations in the current tissue dissociation and single-cell sequencing technologies, some number of “cells” in an scRNA-Seq dataset will actually correspond to doublets. Several computational methods have been developed to identify cells that correspond to doublets but this is still an important artifact in scRNA-Seq data (McGinnis et al., 2018; Wolock et al., 2018). We found that cNMF correctly modeled doublets as a combination of the GEPs for the two combined cell types (Figure 2g). Moreover, we found that cNMF could accurately infer the GEPs even in a simulated dataset composed of 50% doublets (Figure A.7). This illustrates another
benefit of representing cells in scRNA-Seq data as a mixture of GEPs rather than classifying them into discrete clusters.

In all of the simulations described above, the 13 cell-types occurred at uniform frequencies. This allowed us to treat all identity programs as replicates of each other for evaluating inference accuracy, rather than having to separately consider rare GEPs which should, all else equal, be harder to infer than common ones. However, this is an approximation of reality where cell-type proportions can vary over multiple orders of magnitude. We therefore also performed simulations containing biologically plausible cell-type proportions derived from the published clustering of a dataset analyzed later in this manuscript (Hrvatin et al., 2018b) (Materials and methods). When we kept all of the other simulation parameters identical to those of the initial simulations, some identity GEPs from rare cell-types were missed by cNMF, cICA, and Louvain clustering (Figure A.8a). However, when we increased the distinctness of the identity GEPs of the cell types, they could still be inferred by both cICA and cNMF with similar relative performances to what we saw in the primary benchmarking analysis (Figure A.8b). This suggests that the simplification of uniform cell-type frequencies does not significantly impact our conclusions.

2.3.2 cNMF deconvolutes hypoxia and cell-cycle activity GEPs from identity GEPs in brain organoid data

Having demonstrated its performance and utility on simulated data, we then used cNMF to re-analyze a published scRNA-Seq dataset of 52,600 single cells isolated from human brain organoids (Quadrato et al., 2017). The initial report of this data confirmed that organoids contain excitatory cell types ho-
mologous to those in the cerebral cortex and retina as well as unexpected cells of mesodermal lineage, but further resolution can be gained on the precise cell types and how they differentiate over time. As organoids contain many proliferating cell types, we sought to use this data to confirm that cNMF could detect activity programs – in this case, cell cycles programs – in real data, and to explore what biological insights could be gained from their identification.

We identified 31 distinct programs in this dataset that could be further parsed into identity and activity programs (Figure A.9). We distinguished between identity and activity programs by using the fact that activity programs can occur in multiple diverse cell types while identity programs represent a single cell-type. Most cells had high usage of just a single GEP, which is consistent with expressing just an identity program (Figure 3a). When cells expressed multiple GEPs, those typically had correlated expression profiles, suggesting that they correspond to identity programs of closely related cell types or cells transitioning between two developmental states, rather than activity programs (Figure A.10).

By contrast, 3 GEPs were co-expressed with many distinct and uncorrelated programs, suggesting that they represent activity programs that occur across diverse cell types (Figure 3a-b). Consistent with this, the 28 suspected identity programs were well separated by the cell-type clusters reported in Quadrato et al., 2017 while the three suspected activity programs were expressed by cells across multiple clusters (Figure A.11-12). Except for a few specific cases discussed below, we used these published cluster labels to annotate our identity GEPs.

Our 28 identity programs further refined the 10 primary cell-type clusters originally reported for this dataset. For example, we noticed that cells previously annotated as mesodermal predominantly expressed one of three GEPs that were significantly enriched for genes in the ‘Muscle Contraction’ Gene
Figure 2.3: Deconvolution of activity programs from cell identity in brain organoid data
(a) Heatmap showing percent usage of all GEPs (rows) in all cells (columns). Identity GEPs are shown on top and activity GEPs are shown below. Cells are grouped by their maximum identity GEP and fit into columns of a fixed width for each identity GEP.
(b) tSNE plot of the brain organoid dataset with cells colored by their maximally used identity GEP, and with a black edge for cells with >10% usage of the G1/S or G2/M activity GEP or a maroon edge for cells with >10% usage of the hypoxia GEP.
(c) Table of P-values for the top six Gene Ontology gene set enrichments for the three activity GEPs.
(d) Heatmap of z-scores of top genes associated with three mesodermal programs, in those programs (top), and in all other programs (bottom).
(e) Heatmap of z-scores of top genes associated with three activity GEPs, in those programs (top), and in all other programs (bottom).
(f) Proportion of cells assigned to each identity GEP that express the G1/S or G2/M program with a percent usage greater than 10%.
(g) Proportion of cells assigned to each identity GEP that express the hypoxia program with a percent usage greater than 10%.
Ontology (GO) set (P<1x10-10 vs. P>.19 for all other GEPs). They therefore likely represent muscle cells. Inspecting the genes associated with these 3 GEPs, we noticed that they include genes characteristic of different classes of skeletal muscle: (1) immature skeletal muscle (e.g. MYOG, TNNT2, NES), (2) fast-twitch muscle (e.g. TNNT3, TNNC2, MYOZ1), and (3) slow-twitch muscle (e.g. TNNT1, TNNC3, TPM3) (Figure 3d, Supplementary Table 1). This unexpected finding suggests that distinct populations of skeletal muscle cells – excitatory cell types with many similarities to neurons – are differentiating in these brain organoids.

Of the three activity programs identified, we found that two were strongly enriched for cell cycle Gene Ontology (GO) sets, suggesting that they correspond to separate phases of the cell cycle (Figure 3c). One showed stronger enrichment for genesets involved in DNA replication (e.g. DNA Replication P=3x10-52 compared to P=2x10-3) while the other showed stronger enrichment for genesets involved in mitosis (e.g. Mitotic Nuclear Division, P=4x10-61 compared to P=2x10-46). These enrichments and inspection of the genes most associated with these programs implied that one represents a G1/S checkpoint program and the other represents a G2/M checkpoint program (Figure 3e). Thus, cNMF discovered two activity programs corresponding to separate phases of the cell cycle directly from the data.

The third activity program is characterized by high levels of hypoxia related genes (e.g. VEGFA, PGK1, CA9, P4HA1, HILPDA) suggesting it represents a hypoxia program (Figure 3e). This is consistent with the lack of vasculature in organoids which makes hypoxia an important growth constraint (Kelava and Lancaster, 2016). This GEP is most significantly enriched for genesets related to protein localization to the endoplasmic reticulum and nonsense mediated decay (P=3x10-37, P=5x10-31) (Figure
3c), consistent with reports that hypoxia post-transcriptionally increases expression of genes that are translated in the ER (Staudacher et al., 2015) and modulates nonsense mediated decay activity (Gardner, 2008). In the initial report of this data, staining for a single hypoxia gene, \(HIF1A\), failed to detect significant levels of hypoxia. Indeed, \(HIF1A\) is not strongly associated with this GEP, at least not at the transcriptional level. This highlights the ability of our unbiased approach to detect unanticipated activity programs in scRNA-Seq data.

Having identified proliferation and hypoxia activity programs, we sought to quantify their relative rates across cell types in the data. We found that 3079 cells (5.9%) expressed the G1/S program and 2043 cells (3.9%) expressed the G2/M program (with usage>=10%). Classifying cells into cell types according to their most used identity program, we found that many distinct populations were replicating. For example, cNMF detected a rare population, included with the forebrain cluster in the original report, that we label as “stem-like” based on high expression of pluripotency markers (e.g. \(LIN28A\), \(LiTD1\), \(MIR302B\), \(DNMT3B\)) (Supplementary Table 1). These cells showed the highest rates of proliferation with over 38% of them expressing a cell-cycle program in addition to the “stem-like” identity GEP (Figure 3f).

cNMF was further able to refine cell types by disentangling the contributions of identity and activity programs to the gene expression of cells. For example, we found that a cell cluster labeled in Quadrato et al., 2017 as “proliferative precursors”, based on high expression of cell-cycle genes, is composed of multiple cell types including immature muscle and dopaminergic neurons (Figure A.12). The predominant identity GEP of cells in this cluster is most strongly associated with the gene \(PAX7\), a marker of self-renewing muscle stem cells (Pawlikowski et al., 2009) (Supplementary Table 1). Indeed,
this GEP has high (>10%) usage in 41% of cells who’s most used GEP is the immature muscle program, suggesting it may be a precursor of muscle cells. This relationship was not readily identifiable by clustering because the majority of genes associated with the cluster were cell cycle related.

We also saw a wide range of cell types expressing the hypoxia program, with the highest rates in C6-1, neuroepithelial-1, type 2 muscle, and dopaminergic-2 cell types. The lowest levels of hypoxia program usage occurred in forebrain, astroglial, retinal, and type 1 muscle cell types (Figure 3g). The hypoxia response program is widespread in this dataset with 5,788 cells (11%) of all cells expressing it (usage > 10%). This illustrates how inferring activity programs in scRNA-Seq data using cNMF makes it possible to compare the rates of cellular activities across cell types.

2.3.3 cNMF identifies depolarization induced and novel activity programs in scRNA-Seq of mouse visual cortex neurons

Next we turned to another published dataset to further validate cNMF and to illustrate how it can be combined with scRNA-Seq of experimentally manipulated cells to uncover more subtle activity programs. We re-analyzed scRNA-Seq data from 15,011 excitatory pyramidal neurons or inhibitory interneurons from the visual cortex of dark-reared mice that were suddenly exposed to 0 hours, 1 hours, or 4 hours of light (Hrvatin et al., 2018b). This allowed the authors to identify transcriptional changes induced by repeated depolarization, a phenomenon believed to be critical for proper cortical function. To increase our resolution to identify neuronal activity programs, we used the published clustering to exclude any non-neuronal cells that were also captured by the experiment from our dataset. We sought to determine whether cNMF would identify the relatively modest activity programs (60 genes with
fold-change $\geq 2$ and FDR $< 0.05$) elicited by the experiment without knowledge of the experimental design labels. Furthermore, since the authors identified heterogeneity in stimulus-responsive genes between neuronal subtypes, we wondered if cNMF would identify a common activity program and whether it could tease out patterns in what is shared or divergent across neuron subtypes.

We ran cNMF on neurons combined from all three exposure conditions and identified 20 GEPs, interpreting 14 as identity and 6 as activity programs (Figure A.13). As we saw in the organoid data, the activity programs were co-expressed with many distinct and uncorrelated GEPs while the identity programs only overlapped in related cell types (Figure 4a-b). In addition, the identity programs were well separated by the published clusters while the activity programs were spread across multiple clusters (Figure A.14). We thus used the published cluster annotations to label the identity GEPs.

Three activity programs were correlated with the stimulus, which indicates that they are induced by repeated depolarization (Figure 4c). One of these was induced at 1H and thus corresponds to an early response program (ERP). The others were primarily induced at 4H and thus correspond to late response programs (LRPs). These programs overlapped significantly with the sets of differentially expressed genes reported in Hrvatin et al., 2017 ($P=8 \times 10^{-34}$ for the ERP and genes induced at 1H; $P=4 \times 10^{-22}$, $P=4 \times 10^{-14}$ for the LRPs and genes induced at 4Hs, one-sided Mann Whitney test). Intriguingly, one LRP was more induced in superficial cortical layers, while the other was more induced in deeper layers. This supports a recently proposed model where the ERP is predominantly shared across excitatory neurons, while LRPs vary more substantially across neuron subtypes (Hrvatin et al., 2018b). It also illustrates cNMF’s sensitivity: in the initial report, only 64 and 53 genes were identified as differentially expressed in at least one excitatory cell type at 1H and 4Hs (fold-change $\geq 2$, FDR $< 0.05$).
Figure 2.4: Identification of activity programs in neurons of the visual cortex
(a) Heatmap showing percent usage of all GEPs (rows) in all cells (columns). Identity GEPs are shown on top and activity GEPs are shown below. Cells are grouped by their maximum identity GEP and fit into columns of a fixed width for each identity GEP.
(b) t-SNE plots of cells colored by maximum identity GEP usage (left) or by absolute usage of each activity GEP (right).
(c) Box and whisker plot showing the percent usage of activity programs (rows) in cells classified according to their maximum identity GEP (columns) and stratified by the stimulus condition of the cells (hue). The central line represents the median, boxes represent the interquartile range, and whiskers represent the 5th and 95th quantile.
(d) Scatter plot of z-scores of the superficial late response GEP in the primary dataset against the corresponding GEP in the Tasic et al. (2016) dataset.
(e) Same as (d) but for the neurosecretory activity program.
Nevertheless, cNMF was able to find this program in an unsupervised manner, without knowledge of the experimental design.

cNMF was also able to identify a depolarization induced program in visual cortex neurons that were not experimentally manipulated to elicit them. We analyzed an additional scRNA-Seq dataset of 1,573 neurons from the visual cortex of adult mice that, unlike in the primary dataset, were not reared in darkness or treated with a specific light stimulus (Tasic et al., 2016). In this dataset, cNMF identified a matching GEP for all visual cortex cell types found in the primary dataset except for a single subtype of excitatory layer 5 (Figure A.15a). Moreover, it identified a GEP that showed striking concordance with the superficial LRP found in the primary dataset (Fisher Exact Test of genes with association Z-score>.0015, OR=127, P=1x10^{-118}, Pearson Correlation=.645) (Figure 4d). This program was predominantly expressed in excitatory cells of the more superficial layers of the cortex as would be expected based on the results in the primary dataset. For example, over 40% of the excitatory layer 2 (Exc. L2) type neurons expressed this activity program (Figure A.15b). This demonstrates that cNMF could also find the depolarization induced activity program in scRNA-Seq of cells that had not been experimentally manipulated to elicit it.

Finally, cNMF identified three additional activity programs in the primary visual cortex dataset that were not well correlated with the light stimulus but were expressed broadly across multiple neuronal cell types (Figure 4b-c). We labeled one of these, that was specific to excitatory neurons, as Neurosecretory (NS) because it is characterized by high expression of several secreted neuropeptides including Vgf, Adeyap1, Seg2, Cck, Seg3, and Dkk3, and has high expression of genes that facilitate protein secretion such as Cpe, Cadps2, and Scamps (Supplementary Table 2). The top expressed gene—Vgf (VGF nerve
growth factor inducible) is induced by nerve growth factor (Salton et al., 1991), suggesting that this program may be regulated by external growth factor signals. Notably, we found a matching program in the Tasic et al. dataset (Fisher Exact Test of genes with association Z-score > 0.0015, OR = 53.8, P = 8x10-21, Pearson Correlation = 0.356) (Figure 4e). Thus, this neurosecretory activity program is reproducible across multiple single-cell datasets.

An additional activity program which we labeled Synaptogenesis (Syn) was characterized by expression of genes that play a crucial role in synapse formation, including the transcriptional regulator Mef2c (Barbosa et al., 2008; Flavell et al., 2006; Harrington et al., 2016), synaptic adhesion molecules Ncam1 (Hata et al., 2018) and Cadms (Biederer et al., 2002; Robbins et al., 2010), membrane vesicle traffickers Syt1 and Syt4, Actb which constitutes the predominant synapse cytoskeletal protein, and others with a strong connection to synapse biology such as Ywhaz (Foote et al., 2015; Ramser et al., 2010; Xu et al., 2015), Biedt (Aguirre-Chen et al., 2011; Li et al., 2010). It was also significantly enriched for relevant Gene Ontology sets including postsynapse, glutamatergic synapse, postsynaptic density, and dendrite morphogenesis (P < 3.25x10-6, Supplementary Table 3) which further suggests its interpretation as a program involved in the formation or regulation of synapses. The last activity program (labeled Other) was characterized by high expression of the maternally expressed long non-coding RNA Meg3 (Yan et al., 2016) and other genes that are associated with cerebral ischemic injury (e.g. Glgt (Zhang et al., 1996), Rtn1 (Gong et al., 2017)). Our functional interpretations of the novel activity programs are speculative, but they highlight the ability of cNMF to identify intriguing GEPs in an unbiased fashion.
2.4 Discussion

In this study, we distinguished between cell type (identity) and cell type independent (activity) gene expression programs (GEPs) to motivate our use of matrix factorization, which represents cells as linear combinations of multiple GEPs. However, we note that some biological programs are not nearly classified as either identity or activity GEPs. For example, cell states reflecting oncogenic transformation, or a cell’s position along a morphological gradient blur the distinction between identity and activity. In addition, stochastic fluctuations in individual transcription factors could result in coordinated gene expression changes (Thattai and van Oudenaarden, 2001) and might be better described as a third program category, rather than as an identity or activity GEP. While the identity/activity distinction might not be appropriate in every case, matrix factorization should, in principle, be appropriate for representing all gene expression states that can be reasonably approximated as a linear mixture of programs.

Furthermore, in this study, we have provided an empirical foundation for the use of matrix factorization to simultaneously infer identity and activity programs from scRNA-Seq data. We first show with simulations that despite their simplifying assumptions, ICA, LDA, and NMF (but not PCA) can infer components that align well with GEPs. However, due to the stochastic nature of these algorithms, the interpretability and accuracy of individual solutions can be low. This led us to develop a consensus approach that empirically increased the accuracy and robustness of the solutions. cNMF inferred the most accurate identity and activity programs of all the methods we tested. Moreover, it yielded results in interpretable units of gene expression (transcripts per million) and could accurately infer the percentage of each cell’s expression that was derived from each GEP. These properties made
it the most promising approach for GEP inference on real datasets.

We then explored the utility of cNMF on real data, recapitulating known GEPs, identifying novel ones, and further characterizing their usage. We first validated cNMF with several expected activity programs serving as positive controls. We then identified several unexpected but highly plausible programs, a hypoxia program in brain organoids and a depolarization induced activity program in untreated neurons. Finally, we identified three novel programs in visual cortex neurons that we speculate may correspond to a neurosecretary phenotype, new synapse formation, and a stress response program. Beyond simply discovering activity programs, cNMF clarified the underlying cell types in these datasets by disentangling activity and identity programs from the mixed single-cell profiles. For example, we found that a brain organoid subpopulation that was initially annotated as proliferative precursors actually includes replicating cells of several cell types such as an immature skeletal muscle cell that is differentiating into slow-twitch and fast twitch muscle populations. Furthermore, joint analysis of identity and activity GEPs allowed us to quantify the relative prevalence of activities across cell types. For example, we found in the visual cortex data that one depolarization-induced late response program was predominantly expressed in neurons of superficial cortical layers, while the other was mainly expressed in deeper layers. This suggests that an anatomical or developmental factor may underlie variability in the response. While commonly used approaches based on clustering or pseudotemporal ordering of cells are poorly suited to achieve such insights from single-cell data, these findings emerge naturally from our matrix factorization approach.

We have made our tools and analyses readily accessible so that researchers can readily use cNMF and further develop on the approach. We have deposited all of the cNMF code on Github https://github.com/dylkot/cNMF/
and have made available all of the analysis scripts for figures contained in this manuscript on Code Ocean (https://doi.org/10.24433/CO.9044782e-cb96-4733-8244-072c21399e6) for easy exploration and re-execution.

Figure 2.5: Robustness of cNMF to the number of components (K)
Line plots of the maximum Pearson correlation between each of the cNMF components presented in the main analysis, and the cNMF components that result from multiple choices of K. For the simulated data, for which we have access to ground truth, we plot the correlation between the inferred components for each choice of K and the ground truth 14 components. We highlight components corresponding to activity GEPs with distinct colors and denote the number of identity GEPs contained on the same plot in parenthesis in the legend. A dashed line indicates the K choice that was presented in the main analysis. Pearson correlations are computed considering only the 2000 most over-dispersed genes and on vectors normalized by the computed sample standard deviation of each gene.

As others apply this approach, one key consideration will be the choice of the 3 input parameters required by cNMF: the number of components to be found (K), the percentage of replicates to use as nearest neighbors for outlier-detection, and a distance threshold for defining outliers. While the choice
of $K$ must ultimately reflect the resolution desired by the analyst, we propose two simple decision aids based on (1) considering the trade-off between factorization stability and reconstruction error and (2) looking at the proportion of variance explained by $K$ principal components to estimate the dimensionality of the data (Figure A.3, A.9, A.13). In addition, we noticed that choosing consecutive values of $K$ primarily influenced individual components at the margin, suggesting that cNMF may be robust to this choice within a reasonable range of options (Figure 5 and “Choosing the number of components” section of the Materials and methods).

The additional 2 parameters allow users to optionally identify outlier replicates to filter before averaging across replicates. This improves overall accuracy by removing infrequent solutions that often represent merges or splits of the true GEPs. Using 30% of the number of replicates as nearest neighbors worked well for all datasets we analyzed, and an appropriate outlier distance threshold was clear in our applications based on the long tail in the distance distribution (Figure A.3, A.9, A.13).

Our approach is an initial step towards disentangling identity and activity GEPs, and will benefit from subsequent development. For example, cNMF does not specifically address the count nature of gene expression. Recently developed statistical frameworks that address these aspects of scRNA-Seq data such as Hierarchical Poisson Factorization (Levitin et al., 2018) may therefore increase the accuracy of GEP inference. In addition, NMF often yields low but non-zero usages for many GEPs even though we expect most cells to express a small number of identity and activity GEPs. This lack of sparsity is likely due to over-fitting and could be addressed by adding regularization to the model (Taslaman and Nilsson, 2012). Such refinements and any new matrix factorization that relies on stochastic optimization can be readily combined with our consensus approach to potentially improve accuracy and
interpretability.

A more fundamental limitation of matrix factorizations, including cNMF, is the built-in assumption that cells can be modeled as linear combinations of GEPs. Notably, this precludes modeling of transcriptional repression, where one or more genes that would be induced by one GEP are significantly reduced in expression when a second repressing GEP is active in the same cell. To our knowledge such relationships have not been represented in a matrix factorization framework, but they may be easier to incorporate in new classes of latent variable models such as variational autoencoders (VAEs) (Ding et al., 2018; Grønbech et al., 2018). VAEs represent cells in a highly flexible latent space that can capture non-linearities and interactions between latent variables. However, while the latent variables are designed to facilitate accurate reconstruction of the input gene expression data, it remains to be shown whether they can be directly or indirectly interpreted as distinct GEPs and GEP usages. For the foreseeable future, there may be a trade-off between the flexibility of these models and the difficulty in training them and interpreting their output.

With ongoing technological progress in RNA capture efficiency and throughput, scRNA-Seq data is likely to become richer and more expansive. This will make it possible to detect increasingly subtle GEPs, reflecting biological variability in cell types, cell states, and activities. Here, we have demonstrated a computational framework that can be used to infer such GEPs directly from the scRNA-Seq data without the need for experimental manipulations, providing key insights into the behavior of cells and tissues.
2.5 Materials and methods

2.5.1 Simulations

Our simulation framework is based on Splatter (Zappia et al., 2017) but is re-implemented in Python and adapted to allow simulation of doublets and activity programs. 14 gene-expression programs - $Z \in \mathbb{N}^M$ (13 identity programs $Z_1, \ldots, Z_{13}$ and one activity program $Z_a$ each a vector of $M$ genes) - were simulated as in Splatter. Cells were then randomly assigned to an identity program with equal probability for each class. 30% of cells of 4 cell types were randomly selected to express the activity program at a usage $\varphi_i$ uniformly distributed between 10% and 70%. In the Splatter notation, the pre-trended mean gene-expression profile $\lambda'_i$ for each cell $i=1, \ldots, 15,000$ was computed as the weighted sum of the identity and the activity program:

$$\lambda'_i = L_i (\varphi_i Z_a + (1 - \varphi_i) Z_{I(i)})$$

where $L_i$ is the simulated library size for a cell, $I(i)$ is the identity program assignment for cell $i$, and $\varphi_i = 0$ for cells that do not express the activity GEP and $\varphi_i \sim \text{Uniform}(0.1, 0.7)$ for cells that do.

Doublets were constructed by randomly sampling pairs of cells, summing their simulated count vectors, and randomly down-sampling the counts until the total number of counts equaled the maximum of the library sizes of the original two cells. We simulated 25,000 genes, 1,000 of which were associated with the activity program. The probability of a gene being differentially expressed in a cell identity program was set to 2.5%. The differential expression scale parameter was 1.0 for all simulations.
and the location parameter was either 1.0, 0.75, or 0.5 to simulate different signal to noise levels. Other splatter parameters were:

- $\text{lib.loc} = 7.64$
- $\text{libscale} = 0.78$
- $\text{mean\_rate} = 7.68$
- $\text{mean\_shape} = 0.34$
- $\text{expoutprob} = 0.00286$
- $\text{expoutloc} = 6.15$
- $\text{expoutscale} = 0.49$
- $\text{diffexpdownprob} = 0$
- $\text{bcv\_dispersion} = 0.448$
- $\text{bcv\_dof} = 22.087$

These values were inferred from 8000 randomly sampled cells of the Quadrato et al., 2017 organoid dataset using Splatter. A differential expression location parameter of 1.0 were used for the 50% doublet simulation and all other parameters were kept the same (Figure A.7). Differential expression location parameters of 1.0 and 2.0 were used for the two variable cell-type proportion simulations, and cell-type assignments were sampled from a multinomial with parameters (0.215, 0.210, 0.195, 0.130, 0.092, 0.0328, 0.028, 0.0269, 0.0149, 0.0119, 0.0114, 0.009, 0.009, 0.008, 0.006) based on the proportions of neuronal cell-type clusters in Hrvatin et al., 2017 (Figure A.8). All other parameters were as described above.
2.5.2 Data preprocessing

For each dataset, we removed cells with fewer than 1000 unique molecular identifiers (UMIs) detected. We also filtered out genes that were not detected in at least 1 out of 500 cells. We denote this filtered count matrix as $C_{ij}$ ($i = 1 \ldots N$ cells, $j = 1 \ldots M$ genes). We denote matrices in bold.

We then selected the $H$ most over-dispersed genes as determined by the v-score (Klein et al., 2015b) for input to cNMF. $H$ was set to 2000 for all datasets analyzed in this manuscript. It is essential to select a subset of over-dispersed genes prior to normalization because afterwards, variation in lower-variance genes due to noise will be on the same scale as biologically meaningful variation. 2000 reflects a trade-off between including enough genes to detect subtle biological signals, and speeding up computation by not including too many extraneous genes.

Each of the $H$ over-dispersed genes was then scaled to unit variance before running cNMF resulting in a normalized matrix:

$$\tilde{C} = \left[ \begin{array}{ccc} \vdots & \frac{c_j}{s(c_j)} & \vdots \end{array} \right] \text{ for genes } j \in \{ \text{Overdispersed genes} \}$$

where $C_j$ denotes the $j$th column of $C$, $s(C_j)$ denotes the sample standard deviation of $C_j$, and \{H overdispersed genes\} denotes the previously selected set of over-dispersed genes. This variance-scaling is similar to the log transformation that is commonly applied to scRNA-Seq data in that it ensures genes on different expression scales contribute comparable amounts of information to the GEP inference. However, we prefer variance scaling because log transformation requires addition of
an arbitrary pseudo-count which can substantially impact downstream analysis (William Townes et al., 2019). In addition log-transformation renders fold-changes of 0.1 to 1, and 100 to 1000 equivalent in absolute terms. This is undesirable given the current resolution of scRNA-Seq data because the former change can frequently be attributable to noise while we have much greater confidence in the biological significance of the latter. Variance scaling the data avoids the need for any modulation to the shape of a gene’s distribution and avoids the need for addition of an arbitrary pseudocount. We do not mean center the genes so as to preserve the non-negativity of the expression data which is a requirement for NMF.

Note that we do not perform any cell count normalization (i.e. normalization of the rows of $C$ prior to cNMF). This is because cells with more counts can contribute more information to the model. Technical variation in transcript abundances across cells are captured in the usage matrix rather than the component matrix. However, for the Tasic Et. Al. dataset, which is based on full-transcript sequencing rather than digital UMI counting, we variance-normalized high-variance genes from the TPM matrix directly rather than from the raw count matrix as in the other datasets.

As a final step in cNMF, the consensus programs can be re-fit in physically meaningful (non-normalized) biological units of the user’s choice, such as transcripts per million and including all genes (not just the H over-dispersed ones). See below for details.

2.5.3 **Consensus Non-negative Matrix Factorization (cNMF)**

We use non-negative matrix factorization implemented in scikit-learn (version 0.20.0) with the default parameters except for random initialization, tolerance for the stopping condition of $10^{-4}$, and a max-
imum number of iterations of 400.

R replicates of NMF are run on the same normalized dataset with the same number of components K but with different randomly selected seeds, resulting in R instances of usage matrices $U^{(r)}$ (N cells x K programs) and program matrices $G^{(r)}$ (K programs x H genes):

$$U^{(r)} \times G^{(r)} \approx C \text{ for } r = 1 \ldots R$$

For each replicate r, the rows of $G^{(r)}$ are normalized to have L2 norm of 1:

$$\tilde{G}^{(r)} = \begin{bmatrix} \frac{G_1^{(r)}}{||G_1^{(r)}||_2} \\ \vdots \\ \frac{G_K^{(r)}}{||G_K^{(r)}||_2} \end{bmatrix}$$

where $G_k^{(r)}$ is the kth row of the programs matrix for the rth NMF replicate $G^{(r)}$ and $|| \cdot ||_2$ denotes the L2 norm.

The component matrices from each replicate are then concatenated vertically into a single RK x H dimensional matrix, $G$, where each row is a component from one replicate:
Components with high mean Euclidean distance from their L nearest neighbors are then filtered out as below:

\[
L = \varepsilon R
\]

\[
D(G_i, L) = \frac{1}{L} \sum_{G_n \in N_L(G_i)} \| G_i - G_n \|_2
\]

\[
\mathbf{G}^{(l)} = \begin{bmatrix}
\cdots \\
G_i \\
\cdots
\end{bmatrix}
\text{for } l = 1\ldots Rk \text{ if } D(G_i, L) < \tau
\]

where \(G_i\) is the \(l\)th row of \(\mathbf{G}\), \(N_L(G_i)\) is the set of \(L\) nearest neighbors of \(G_i\) and \(\mathbf{G}^{(l)}\) is the matrix of rows that passed the \(L\) nearest neighbors distance threshold filter.

Two user-specified parameters, \(\varepsilon\) and \(\tau\) determine which replicate components are filtered out and which are kept. \(\varepsilon\) denotes the fraction of NMF replicates to be used as nearest neighbors. Intuitively, \(\varepsilon\) can be thought of as the fraction of replicates that must yield a component approximately matching a
program in order for that program to be kept by cNMF. \( \tau \) is a distance threshold that determines how close a component must be to its nearest neighbors in Euclidean space to be considered “approximately matching”.

Our choice of \( \xi \) and \( \tau \) were guided by inspection of the clustergram and histogram output by cNMF with the goal of filtering out outlier components and yielding clean correlation blocks on the clustergrams (Figures A.3, A.9, A.13). We set \( \xi = 0.3 \) for all datasets analyzed in this manuscript which reflects a tolerance to identify components that occur approximately in 30% or more replicates. We find this to be an appropriate default setting. We set \( \tau = 0.03, 0.10, 0.08, 0.10, \) and \( 0.04 \) for the simulation, organoid, main visual cortex (Hrvatin et al., 2018a), secondary visual cortex (Tasic et al., 2016), and the Pancreatic islets datasets (discussed in the appendix A.1 - supplementary note on between-sample variability) based on truncating the long tail of the distance to KNN histogram.

Next, the rows of \( G^{(l)} \) are clustered using KMeans with the Euclidean distance metric and the same number of clusters (\( K \)) as the number of components for the NMF runs. This defines sets \( A_k = \) rows \( l \) assigned to cluster \( k \) containing the indices of the rows of \( G^{(l)} \) that are assigned to the \( k \)th cluster.

Each cluster of replicate components is then collapsed down to a single consensus vector by taking the median value for each gene across components in a cluster:

\[
G_{kj} = \text{median}(\{C_{lj}^{(l)} \text{ for } l \in A_k\})
\]

where \( l \) indexes over rows of \( G^{(l)} \) and \( j \) indexes over columns (genes), and with the median taken separately for each gene \( j \). This defines a KxH consensus programs matrix \( G^{(c)} \) where the \( (c) \) super-
script denotes consensus. The merged GEP components are then L1 normalized:

\[
\tilde{G}^{(c)} = \left[ \begin{array}{c}
\frac{G_1^{(c)}}{|G_1^{(c)}|_1} \\
\vdots \\
\frac{G_K^{(c)}}{|G_K^{(c)}|_1} 
\end{array} \right]
\]

where \(G_k^{(c)}\) is the kth row of \(G^{(c)}\) and \(| \cdot |_1\) denotes the L1 norm. A consensus usage matrix is then fit by running one last iteration of NMF with the component matrix fixed to \(\tilde{G}^{(c)}\). This amounts to fitting non-negative least squares regressions of each cell’s normalized expression profile \(\tilde{C}_i\) against \(\tilde{G}^{(c)}\) by solving the following optimization:

\[
\min_{U_{i1} \ldots U_{iK} \geq 0} |\tilde{G}_i - \sum_{k=1}^{K} U_{ik} \tilde{G}_k^{(c)}|_2
\]

where \(\tilde{G}_k^{(c)}\) is the H-dimensional normalized consensus program vector for the kth GEP, i indexes over cells, and we are maximizing with respect to GEP usage values \(U_{i1} \ldots U_{iK}\) which are constrained to be non-negative. We concatenate all of these coefficients into a consensus usage matrix \(U^{(c)}\):

\[
U^{(c)} = \left[ \begin{array}{c}
U_{i1} \ldots U_{iK} \\
\vdots \\
U_{Ni} \ldots U_{NK}
\end{array} \right]
\]
and normalize it so that the usage values for each cell sum to 1:

\[
\tilde{\mathbf{U}}(\epsilon) = \begin{bmatrix}
\frac{u_1^{(\epsilon)}}{||u_1^{(\epsilon)}||_1} \\
\vdots \\
\frac{u_k^{(\epsilon)}}{||u_k^{(\epsilon)}||_1} \\
\vdots \\
\frac{u_m^{(\epsilon)}}{||u_m^{(\epsilon)}||_1}
\end{bmatrix}
\]

where \(u_i^{(\epsilon)}\) is the \(i\)th row of the consensus usage matrix, corresponding to the \(i\)th cell. With this normalized, consensus usage matrix fixed, final program estimates can be computed in desired units, and for all genes—including genes that were not initially included among the over-dispersed set. This is done by running a last iteration of NMF with the usage matrix fixed as \(\tilde{\mathbf{U}}(\epsilon)\) and the input data reflecting the desired final units. To convert the estimated programs to TPM units and to obtain program vectors spanning the full set of input genes, we refit against the matrix of TPM values, \(T\):

\[
T = \begin{bmatrix}
\frac{10^6 c_1}{||c||_1} \\
\vdots \\
\frac{10^6 c_l}{||c||_1} \\
\vdots \\
\frac{10^6 c_N}{||c||_N}
\end{bmatrix}
\]

\[
\min_{G_j^{(TPM)} \ldots G_j^{(TPM)}} \sum_{k=1}^K \left( ||T_j - \sum_{j=1}^K \tilde{U}_j^{(\epsilon)} G_{jk}^{(TPM)}||_1 \right)
\]
where $T_j$ is the N-dimensional TPM profile for the jth gene, $\bar{U}^{(c)}_k$ is the N-dimensional normalized consensus profile for the kth GEP over all cells, and $G_{jk}^{\text{(TPM)}}$ is an estimated coefficient reflecting how much the TPM of gene j is expected to increase per a unit increase in usage of GEP k, if all other usages were held constant. Note that the TPM matrix $T$ is calculated using a raw count matrix $C$ that includes all genes, even those that were filtered out for falling below the count threshold. We repeat this for all M genes in the filtered count matrix and combine all of these coefficients into a consensus program matrix:

$$
G^{\text{(TPM)}} = 
\begin{bmatrix}
G_{n1}^{\text{(TPM)}} & \cdots & G_{nK}^{\text{(TPM)}} \\
\vdots & \ddots & \vdots \\
G_{Ki}^{\text{(TPM)}} & \cdots & G_{Kr}^{\text{(TPM)}}
\end{bmatrix}
$$

2.5.4 Identification of marker genes

We identify marker genes (genes that are statistically associated with each GEP) using multiple least squares regression of normalized (Z-scored) gene expression against the consensus GEP usage matrix. This amounts to finding the genes that have higher than average expression for cells that use a specific GEP. We compute the Z-score of the TPM profile like so:

$$Z_j = \frac{T_j - \mu_j}{\sigma_j}$$

where $T_j$ is the TPM profile of the jth gene, $\mu_j$ is the sample mean, and $\sigma_j$ is the sample standard
deviation of $T_j$. Then we fit coefficients, $B_{ij} \ldots B_{kj}$, reflecting the association between GEP $k$ and gene $j$ using ordinary least squares regression by solving the minimization:

$$\min_{B_{ij} \ldots B_{kj}} ||Z_j - \sum_{k=1}^{K} U_{k}^{(c)} B_{kj}||_2$$

where $U_{k}^{(c)}$ is the $k$th column of the un-normalized consensus usage matrix. The regression coefficients $B_{kj}$ can then be interpreted as by how many standard deviations the expression of gene $j$ should increase for an additional count of usage being attributed to GEP $k$. We regress against Z-scored expression values rather than the un-normalized expression values so that the coefficients will be comparable between genes expressed on different scales. For discrete clustering methods, the usage matrix is a binary indicator matrix containing a 1 for the cluster (column) each cell is assigned to, and a 0 for all other columns. In the discrete clustering context, $B_{kj}$ can be interpreted as the average expression of gene $j$ in cells assigned to cluster $k$. Identifying marker genes through multivariate regression in this fashion, rather than through separate tests for each GEP, reduces the risk of confounding that can occur when GEPs tend to be expressed in the same cells. For example, if an activity GEP is predominantly expressed in cells of a specific cell-type, it avoids misattributing activity genes to the identity program of that cell-type and vice versa.

We note that because gene-expression data is not normally distributed, the residuals of the regression will not be normal, which violates an assumption of OLS regression. However, the coefficient estimates will still be unbiased even if normality is violated. In practice, we do not use the P-values of the regressions at any point in our analysis as those can be inaccurate due to non-normality. We
recommend testing for gene-set enrichment on regression coefficients directly (as we discuss below) rather than setting thresholds on regression P-values.

2.5.5 Choosing the number of components

Determining the number of components (K) to use for cNMF is an important but challenging step without a simple approach that can work for all datasets and applications. We use two diagnostic plots to help guide this decision. The first plot shows the stability of the solution (as captured by the silhouette score) and the Frobenius reconstruction error as a function of K as described previously in (Alexandrov et al., 2013). However, unlike in (Alexandrov et al., 2013), we run NMF on normalized data matrices rather than count matrices and therefore do not resample counts but simply repeat NMF with different randomly selected seeds. We compute the Frobenius error using the consensus NMF solution but without any outlier filtering. We also use the Euclidean distance on L2 normalized components as the metric for the silhouette score rather than Cosine distance. Silhouette score is calculated using the Scikit-learn version 20.0 silhouette_score function. We parallelized the individual factorization steps over cores on a multi-core virtual machine using GNU Parallel (Tange and Others, 2011).

As another approach to confirm the appropriateness of our choice of K, we use scree plots which depict the proportion of variance explained per principal component (Figures A.3, A.9, A.13). This is motivated by the fact that choosing the optimal number of principal components and choosing the number of NMF components can both be framed as estimating the rank for a low-dimensional representation of the input data. As a consequence of the Eckart Young Mirsky theorem, PCA necessarily
provides the matrix factorization with minimum Frobenius reconstruction error for any choice of K
(Eckart and Young, 1936) and we also use the Frobenius error in our NMF model. Because principal
components are orthogonal to each other, and loadings of NMF components can never be negative,
K principal components will always span a larger sub-space than K NMF components. This suggests
that the optimal number of NMF components will likely not be smaller than the optimal number of
PCs. The scree plot is a commonly used tool to estimate the number of principal components and we
use it to help guide the number of NMF components as well.

We note that these 2 plots merely provide a general aid for the choice of K and we considered the
biological interpretability of factors found from several choices of K before proceeding. We do not
recommend necessarily using the maximum stability solution of the error vs. stability plot as this
can frequently miss true biological signal and, indeed would have led to the incorrect choice for the
simulated data (Figure A.3).

Given the uncertainty of the choice of K, we confirmed that the conclusions of this manuscript
are robust to this decision. When we varied K within a range of ± four around the choice used in
the manuscript, we found approximately the same core set of GEPs with a single new GEP being
discerned with each consecutive step in K. For each step below the selected K, approximately a single
GEP was lost, but for choices above the selected K, components approximately matching the original
K programs (i.e. with Pearson correlation >0.7) were found (Figure 5). This suggests that cNMF
yields relatively stable solutions for a moderate range of K values.
2.5.6 Comparison of cNMF with other methods

We compared cNMF with consensus and standard versions of LDA and ICA as well as with PCA, Louvain clustering and a hard clustering based on assignment of cells to their ground-truth labels. We used the implementations of LDA, ICA, and PCA in scikit-learn and the implementation of Louvain clustering in scanpy (Wolf et al., 2018a). For ICA, we used the FastICA implementation with default options for all the parameters. For LDA, we used the batch algorithm and all other parameters as defaults. We defined the consensus estimates across 200 replicates in the same way as for cNMF but with a slight modification for ICA. Because ICA is under-determined with respect to the signs of the solutions, some iterations will yield a given component pointed in one direction while others produce approximately the same component but pointed in the opposite direction (multiplied by -1). Therefore, we aligned the orientation of components from across replicates by identifying any components whose median usage across all cells was positive and scaled those and the corresponding usages by -1.

For Louvain clustering, we used 14 principal components to compute distances between cells and used 200 nearest neighbors to define the KNN graph. We chose 14 principal components based on the fact that the data was simulated based on a 14-dimensional basis and, therefore, the biological variation in the data can be captured by 14 PCs and subsequent components correspond to noise. This choice is also justified by choosing the elbow on scree plot in Figure A.3. We used 200 nearest neighbors for the clustering as this is a relatively large number to minimize variance but it is still smaller than the smallest discrete population (0.3 x 15000 x (1/13) = 346 cells from a specific cell-type that expresses the activity program).
For ground-truth assignment clustering, we assigned each cell to a cluster defined by its true identity program, except for cells which had greater than 40% usage of the activity program, which we assigned to an activity program cluster. Then we determined a GEP corresponding to each cluster as the mean TPM value for each gene over cells in the cluster.

To evaluate the accuracy of these various methods, we first calculated the Z-score coefficient for associating each gene with each program as described above. We then calculated sensitivity and false discovery rate (FDR) for each threshold on those coefficients and plotted those as an ROC-curve, except with FDR on the X-axis instead of false positive rate. For this evaluation, we considered a gene as truly associated with a GEP if it had a ground-truth fold-change of \( \geq 2 \) and truly unassociated with a GEP if the ground-truth fold-change was 1. Genes with a fold-change between 1 and 2 were ignored for this evaluation.

2.5.7 Testing enrichment of genesets in programs

We used the Z-score regression coefficients identified as above as input for a one-sided Mann Whitney U Test (with tie correction) comparing the median of genes in each geneset to those of genes not in the geneset. We first floored all negative coefficients to equal zero prior to the test. Coefficients less than 0 indicate genes that are expressed at higher levels in cells that do not use the GEP (all other things equal) than in cells that do. We floor these values so that variation in genes that are not directly part of a GEP (which can make up the majority of genes) do not substantially impact the Mann-Whitney statistic for that GEP.
2.5.8 Data availability

The organoid data described in the manuscript is accessible at NCBI GEO accession number GSE86153. However, we obtained the clustering and unnormalized data by request from the authors. The visual cortex datasets used for Figure 3 are accessible at NCBI GEO, accession numbers GSE102827 and GSE71585.

2.5.9 Code availability

Code for running cNMF is available on at Github https://github.com/dylkot/cNMF, as is code for simulating data with doublets and activity programs at https://github.com/dylkot/scsim.

All datasets and analysis used in this manuscript are available for download, exploration and re-execution on Code Ocean: https://doi.org/10.24433/CO.9044782e-cb96-4733-8a4f-bf42c21399e6

2.6 Acknowledgments

We thank all members of the Greenberg lab for providing critical feedback during the course of the project as well as the preparation of the manuscript. In particular, we thank Thomas Vierbuchen, Emi Ling, Lisa Boxer, Eric Griffith, Eric Greer, and David Ginty for their feedback and critical reading of the manuscript.
Single-cell profiling defines dynamics of host response and viral replication in Ebola virus infected rhesus macaques

3.1 Abstract

Recent and ongoing outbreaks of Ebola virus disease (EVD) have caused >30,000 human infections with case fatality rates exceeding 40%, yet little is known about innate and adaptive immunity, cell tropism, and Ebola virus (EBOV) replication dynamics, in vivo. Working with a biosafety level 4
laboratory, we adapted two orthogonal single-cell technologies – Seq-Well, a portable single-cell RNA-Seq platform, and CyTOF, mass cytometry – to characterize peripheral immune cells during EVD in rhesus macaques. We obtained mRNA profiles from 100,000 peripheral immune cells and 42 protein profiles from >15 million cells, and characterize changes in frequency and phenotype of all major cell types. We identify global and cell type-specific gene expression changes that may explain disease; for example, monocytes lose antigen presentation capacity, a likely cause of failed adaptive immunity during EVD. By comparing infected versus uninfected monocytes, we identify key factors influencing EBOV tropism: changes in canonical monocyte markers CD14 and CD16, and monocyte-to-macrophage differentiation. By studying heterogeneity among infected monocytes, we find that intracellular EBOV RNA concentration spans multiple orders of magnitude. Moreover, EBOV infection down-regulates STAT1 and interferon signaling at the transcriptional level, and increases expression of putative pro-viral genes such as DYNLL1 and HSPA5, hinting at key cellular pathways the virus manipulates for its replication. Our dataset provides a high resolution map of the dynamics of viral replication and host response in infected and bystander cells which can serve as a resource for future investigations.

3.2 Introduction

Ebola virus (EBOV) is among the world’s most deadly pathogens, with an estimated case fatality rate of 66% in the ongoing epidemic in the Democratic Republic of Congo (Ilunga Kalenga et al., 2019; World Health Organization, 2019) and 40% in the 2013–2016 epidemic in west Africa (Lo et al., 2017). EBOV infection causes Ebola virus disease (EVD), characterized by fever, malaise, muscle aches, and
gastrointestinal distress, rapidly progressing to coagulopathy, shock, and multi-organ failure (Malvy et al., 2019). While recently developed vaccines (Kennedy et al., 2017) and monoclonal antibody therapeutics (Mulangu et al., 2019) have shown great promise for treating EVD, case fatality rates in treated patients still exceed 30%, highlighting the crucial need for further research into disease pathogenesis.

Experimental studies of EVD pathogenesis, while paramount, face numerous logistical challenges and have thus been limited in scope relative to studies of other pathogens. Experiments involving live EBOV require maximum containment (e.g., biosafety level 4 [BSL-4]) and therefore are restricted to a small number of highly specialized research facilities. In vivo studies are especially challenging: thoroughly studying human EVD is difficult to do in the context of deadly outbreaks in resource limited countries, necessitating animal models of EVD. However, commonly used laboratory mouse lines are immune to naturally occurring EBOV isolates, limiting their utility for research (Bray, 2001; Rasmussen et al., 2014). Moreover, rodents and other small animal models such as ferrets lack the primate-specific NPC1 genotype (Diehl et al., 2016), the key cellular receptor for EBOV entry (Carette et al., 2011; Côté et al., 2011), and do not always result in human EVD-like disease. EVD in non-human primates (NHPs) most closely resembles human EVD, but NHP studies are often limited to small sample sizes (Bennett et al., 2017; St Claire et al., 2017).

The two predominant approaches to studying EVD – analyzing infected cells in culture or infected animals in vivo – have revealed important, if somewhat contradictory aspects of how EBOV impacts the host immune system. In cell culture, EBOV potently inhibits both production of type 1 interferon (Basler et al., 2003; Gupta et al., 2001; Harcourt et al., 1998) and signal transduction downstream of interferon receptors (Harcourt et al., 1999; Kash et al., 2006; Leung et al., 2006; Reid et al., 2008). Under-
activation of this key innate antiviral response hinders the ability of antigen-presenting cells to activate the adaptive immune system to combat infection (Bosio et al., 2003; Lubaki et al., 2013; Mahanty et al., 2003). Similarly, in human EVD cases, failure to mount T-cell and humoral immune responses is a key determinant of fatal outcomes (Baize et al., 1999) and could be due to reduced presentation of viral proteins by antigen-presenting cells (Lüdtke et al., 2016). However, in contrast to the culture-based findings, EVD in vivo is characterized by high fever, dramatic up-regulation of hundreds of interferon stimulated genes (Caballero et al., 2016; Liu et al., 2017; Reynard et al., 2019; Speranza et al., 2018) and dozens of inflammatory cytokines (Caballero et al., 2016; Reynard et al., 2019; Wauquier et al., 2010), suggesting that an aberrant over-activation of innate immunity underlies much of EVD pathology, rather than solely virus-mediated cytotoxicity (Geisbert et al., 2003b,c).

High-throughput single-cell technologies such as single-cell RNA-Seq (scRNA-Seq) and CyTOF have made it possible to dissect host-virus interactions at high resolution: comprehensively cataloging changes in cell types and states during disease, identifying host factors that influence cell tropism, and determining how viral replication alters cellular pathways to engineer a pro-viral environment. By generating high resolution mRNA profiles for thousands to millions of individual cells in a sample, scRNA-Seq can quantify the cell type composition and gene expression programs of individual cells, signals that get convolved in bulk RNA-Seq. By sequencing viral RNA within cells, scRNA-Seq allows comparison of gene expression between infected and uninfected bystander cells, yielding insights into both host and viral transcriptional dynamics of infection (Hein and Weissman, 2019; Russell et al., 2018; Steuerman et al., 2018a; Zanini et al., 2018b). The ability to separately characterize gene expression of infected and uninfected cells is especially exciting as a means to disentangle the direct effects
of EBOV infection within a cell from the effect of immunopathology driven by the inflammatory cytokine milieu. However, many scRNA-Seq technologies require droplet generator devices that can be challenging to establish in a maximum containment facility; as a result, such approaches have yet to be applied to a risk group 4 (RG-4) pathogen such as EBOV.

Here we describe the first investigation of a RG-4 agent under maximum containment with scRNA-Seq and high-throughput multi-parameter proteomics. We apply Seq-Well, a portable single-cell RNA-seq platform (Gierahn et al., 2017; Hughes et al., 2019), and CyTOF to a combined total of 64 samples of peripheral blood mononuclear cells (PBMCs) collected from 27 rhesus macaques at multiple timepoints before and after lethal EBOV challenge. These data provide unprecedented resolution to characterize changes in cell type composition and gene expression of the major immune cell types over the course of EVD. Moreover, as EBOV harbors an RNA genome and transcribes polyadenylated mRNAs, we detect viral RNA within individual cells, allowing us to define Ebola tropism with high resolution and to identify EBOV-associated transcriptional changes in putative pro- and antiviral genes.

We find that EVD leads to wide-spread changes in the circulating monocyte populations, with loss of the conventional classical (CD14+) and non-classical (CD16+) subsets, which are in turn replaced by a double positive (CD14hi CD16hi) subset that is enriched for EBOV infected cells, and a highly proliferative double negative subset (CD14lo CD16lo). Furthermore, we resolve the apparent contradiction between in vivo and in vitro studies of Ebola by noting that bystander cells of all major immune cell types express an interferon response program but that this response is suppressed specifically within infected myeloid cells, consistent with previous studies in culture.
To gain further insight into host and viral dynamics in infected cells, we performed additional scRNA-Seq of PBMCs infected ex vivo which allowed us to replicate our major in vivo findings with a larger number of infected cells. We find that EBOV infection down-regulates anti-viral interferon response genes while up-regulating candidate pro-viral genes such as DYNLL1 and HSPA5, within infected cells.

Taken together, this dataset constitutes a unique resource for the study of Ebola, enabling the study of host immune response in infected and bystander cells across cell types, and in natural (in vivo) and experimentally controlled (ex vivo) contexts.

3.3 Results

3.3.1 A single-cell atlas of RNA and protein expression in circulating immune cells of Ebola virus infected rhesus macaques

To comprehensively profile EBOV induced immune dysfunction in vivo, we collected circulating immune cells from rhesus macaques prior to infection, or at multiple days post-infection (DPI), corresponding to distinct phases of EVD (Figure 3.1). Cohorts of ≥3 non-human primates (NHPs) were sacrificed as baseline uninfected controls (B) or during early disease (E, DPI 3), middle disease (M, DPI 4–5), or late disease up to a pre-determined humane euthanasia criteria (L, DPI 6–8) (Figure B.1a). Viral load first became detectable in all animals on DPI 3, preceding detectable symptom onset by 1–2 days (Figure 3.2a). EVD symptoms progressed until humane euthanasia criteria was uniformly reached between DPI 5–8 (Figures B.1b). Cells were collected serially for CyTOF or prior to infection.
and again at sacrifice for Seq-Well.

**Figure 3.1: Studying the natural history of Ebola virus disease in rhesus macaques using single-cell profiling of RNA and protein expression in circulating immune cells**

27 rhesus macaques were challenged with 1000 plaque-forming units of Ebola virus (EBOV) Kikwit variant via intramuscular injection. Under biosafety level 4 (BSL-4) containment, we measured clinical parameters and drew whole blood for single-cell and cytokine profiling throughout disease. We enriched peripheral blood mononuclear cells (PBMCs) by density gradient centrifugation, and performed Seq-Well to quantify host (black) and EBOV (red) RNA expression and CyTOF for 41 protein markers. See also Figure S1A.

After quality control filters, we obtained single cell transcriptomes from 58,000 PBMCs and 42-protein CyTOF profiles from 15,000,000 PBMCs. We visualized this data with UMAP non-linear dimensionality reduction (Figures 3.2a, 3.2c, 3.2e, 3.2f, and 3.2h). Unsupervised clustering of the transcriptomes and a down-sampled set of 335,670 CyTOF profiles yielded clusters that could be readily identified as the major circulating immune cell types using well-known RNA and protein markers (Figures 3.2b, 3.2c, and B.2) (Materials and methods). After batch correction of the CyTOF data and
integration of the transcriptomes to adjust for technical sources of variation (Materials and methods), samples were well distributed across the cell-type clusters (Figures B.3a and B.3b) but separated by DPI (Figures 3.2e and 3.2f), consistent with changing cell states over the course of disease progression.

In addition to the major PBMC cell types, we noticed the emergence of immature neutrophils during EVD, a cluster of cells marked by high gene expression of CD177 and SOD2, and protein expression of CD66 and CD11b. Though neutrophils are typically removed during density-based PBMC isolations, immature neutrophils – which are less dense than polynuclear mature neutrophils – can be released from the bone marrow and co-isolate with PBMCs in infectious and autoimmune inflammatory conditions (Carmona-Rivera and Kaplan, 2013; Darcy et al., 2014; Deng et al., 2016), including during EVD (Eisfeld et al., 2017). The neutrophils in our data were almost entirely absent from baseline samples but rose to high frequency by late EVD (scRNA-Seq: 0.08% of baseline cells compared to 15.6% of late EVD cells by scRNA-Seq; CyTOF: 5% of baseline cells compared to 45% of late EVD cells), which supports the hypothesis that immature band cells are released into the periphery from the bone marrow in response to inflammatory cytokines elicited during EVD.

Next, we quantified absolute abundance of each cell type over the course of EVD by combining CyTOF data with complete blood counts (CBC) (Materials and methods). In agreement with previous NHP studies, there was a dramatic >5-fold increase in neutrophil abundance by DPI 4 relative to baseline, before levels returned to baseline in the final phase of illness (Fisher-Hoch et al., 1985; Ebihara et al., 2011) (Figures 3.2d and B.4a). Also consistent with previous studies, we observed a dramatic decrease in lymphocyte abundance with NK cells declining 1 day before the other cell types. Interestingly, all lymphocyte populations rebounded to baseline levels in 4 animals that survived until DPI
Figure 3.2: Single-cell measurements reveal changing cell-type abundance, proliferation, and infection status throughout EVD

(a) Time course of viral load (red, left y-axis, log10 scale) and clinical score (blue, right y-axis). Markers denote the mean ± 1 standard deviation (SD). See also Figure B.1b.

(b-c) Uniform Manifold Approximation and Projection (UMAP) embedding of Seq-Well (b) and CyTOF (c) data, colored by annotated cluster assignment. See also Figures B.2 and B.3.

(d) Fold change (log2 scale) in the absolute abundance (cells/µL of whole blood) of each cell type relative to baseline. Markers denote the mean ± 1 SD. See also Figure B.4.

(e-f) UMAP embedding of Seq-Well (e) and CyTOF (f) data, colored by the day post-infection on which each cell was sampled.

(g) Percentage of Ki67-positive cells (Ki67 intensity > 1.8) of each cell type. Markers denote the mean ± 1 SD. See also Figure B.5.

(h) UMAP embedding of Seq-Well data, colored by the percentage of cellular transcripts mapping to EBOV.

(i) Percentage of infected cells by cell type. Dashed line denotes the 1% false discovery rate threshold used for calling infected cells. See also Figure S6.
Monocyte abundance initially increased more than 2-fold before declining precipitously between DPI 4 and 5, which, to our knowledge, has not previously been described. We relied on CyTOF data for calculating cell-type percentages because we had serial CyTOF measurements for each animal, and because there was significant variability in library size between Seq-Well arrays that could introduce bias in cell-type percentages. Nevertheless, there was generally good agreement between cell-type percentage estimates based on CyTOF and Seq-Well (Figure B.4b).

Changes in circulating cell-type abundance could reflect cell proliferation and death, as well as movement of cells to and from bone marrow, lymph, and tissues. While we were unable to directly quantify rates of death or movement between different compartments, we estimated the fraction of actively dividing cells using the proliferation marker Ki67 in the CyTOF data and cell-cycle genes in our scRNA-Seq data. The fraction of Ki67+ monocytes increased dramatically over the course of infection, suggesting an increase in proliferation (Figures 3.2g and B.5a). By contrast, neutrophil proliferation remained roughly constant (Figures 3.2g and B.5a) despite the dramatic changes in abundance (Figures 3.2d and B.4a), further evidence that immature neutrophils were released from the bone marrow during disease (Summers et al., 2010). Intriguingly, the fraction of dividing T cells and NK cells stayed relatively constant, but it increased dramatically in the 2 NHPs (out of 12 total) that survived until DPI 8 (Figure B.5a). There was general agreement between CyTOF and scRNA-Seq estimates of proliferation rates (Figure B.5b), further supporting the hypothesis that lymphocyte recovery may delay disease progression and even lead to recovery.

Not all cell types support EBOV replication; here, we were able to identify which cells were infected using scRNA-Seq because EBOV has an RNA genome and produces poly-adenylated mRNA tran-
scripts (Figure 3.2h). However, uninfected cells may nevertheless contain EBOV reads due to ambient RNAs that contaminate the single-cell profiles (Young and Behjati, 2018; Fleming et al., 2019). We therefore developed a statistical approach to identify infected cells as those that contain more EBOV transcripts than would be expected by chance. Our approach defines an adaptive threshold based on 1) the number of EBOV-mapped transcripts (marked by unique molecular indices [UMIs]) in a cell, 2) the cell’s library size, 3) the proportion of ambient RNA in a cell, and 4) the proportion of EBOV reads in the ambient RNA of a sample (Materials and methods). This allows us to control the false positive rate (FPR) at a pre-specified level while maximizing power to call infected cells. At a FPR of 1%, we were well-powered to identify an infected cell when ≥1% of its transcripts mapped to EBOV, and estimated a 51% sensitivity when ≥0.1% of its transcripts derived from EBOV (Figure B.6).

As an experimental negative control to test our statistical model, we spiked Madin-Darby canine kidney (MDCK) cells, constituting 5% of the total sample, into a subset of PBMC samples from EVD NHPs immediately before scRNA-Seq. As MDCKs were not exposed to EBOV, viral reads in these transcriptomes should be due to ambient RNA contamination (Russell et al., 2018).

Monocytes comprised the main infected cell population in vivo, first detectable at DPI 4, with an increasing fraction of infected monocytes thereafter (Figure 3.2i). Consistent with previous studies, T cells, B cells, and neutrophils were not called as infected more often than would be expected by chance at a 1% FPR threshold, or more than what was observed in the MDCK control cells. We did not observe any infected plasmacytoid (pDC) or conventional dendritic cells (cDC) in circulation, though DCs have been observed to be infected in culture and in lymph nodes (Geisbert et al., 2003a).
3.3.2 Interferon response drives a gene expression program shared across multiple cell types

We next sought to comprehensively catalog changes in the gene expression profiles of each circulating immune cell type during EVD. To increase statistical power to detect differentially expressed genes, we grouped cells into the baseline, early, middle, and late EVD phases (Figure 3.1). We then compared transcriptomes of cells from each EVD phase to baseline for each cell type individually (Materials and methods).

1,169 genes were differentially expressed in at least one cell type and EVD phase with an FDR corrected q-value < 0.1 and a fold-change of greater than 35%. To identify cell-type and temporal patterns of gene expression, we performed unsupervised clustering of the differential expression signatures and identified 14 modules (Figure 3.3a, Materials and methods). We excluded neutrophils, platelets, pDCs, cDCs, and plasmablasts from this clustering analysis as they had too few cells in one or more EVD phases to accurately estimate fold-change. Two modules which we term "Global" were up or down-regulated across cell types, and the remaining modules were cell-type and/or temporally specific.

The “Global up” module contains 117 genes, consisting of regulators and targets of the interferon (IFN) alpha (α) and gamma (γ) signal transduction cascade such as STAT1, IRF7, MX1, and ISG15. Gene sets labeled “response to interferon alpha”, and “response to interferon gamma” were significantly enriched in this module (Fisher’s exact test, IFNα: OR=89.6, q=1x10-39; IFNγ: OR=49.6, q=3x10-41, Table S1). The emergence of an IFN response is consistent with the fact that we observe up-regulation of type 1 (α/β) and type 2 (γ) IFNs in monocytes and T cells respectively, along with a
large number of other cytokines (Figure B.7a, Table S1). However, as IFN mRNAs are unstable and expressed transiently (Lin et al., 2011), we observed them at extremely low levels, and IFNα was not detected as statistically significant at any EVD phase compared to baseline. To further characterize the dynamics of the IFN response, we scored cells for the expression of literature-annotated interferon stimulated genes (ISGs) that overlap with the "Global up" module (Materials and methods). The average expression of ISGs increased monotonically throughout EVD across each cell type (Figure 3.3b).

As there is substantial overlap between the genes stimulated by IFNα and IFNγ, we sought to determine if one cytokine predominated, or if both acted independently. We therefore identified genes that were annotated as regulated by IFNα but not IFNγ (i.e., uniquely IFNα-regulated) and vice versa. Both uniquely IFNα- and uniquely IFNγ-regulated genes were still significantly enriched in the “Global up” module (FDR corrected q-value < 0.001), with a larger fold-change for the uniquely IFNα-regulated genes than the uniquely IFNγ-regulated genes (OR = 28.6 vs. OR = 14.6). This pattern held true for each cell type and EVD phase separately (Figure B.7b). These results suggest that both IFNα and IFNγ substantially and independently influence the gene expression profiles of circulating cells during EVD.

The "Global down” module reflects core gene expression programs that are down-regulated during EVD. It contains numerous regulators of translation initiation and elongation (e.g. EEF2, EIF3D, and PABPC1) which is consistent with a core function of IFN signaling being the down-regulation of translation (Li et al., 2015). Consistent with this interpretation, the REACTOME_TRANSLATION gene set is significantly enriched (q=2.3×10⁻⁷, Table S1). This module was most significantly down-regulated in late EVD, with most genes showing monotonically decreasing expression over time.
Figure 3.3: Patterns of differential expression across disease stages and cell types in EVD
(a) Fold changes (log2 scale) of 1,169 differentially expressed genes (rows) in each cell type at early (E), middle (M), and late (L) stages of disease, relative to baseline levels. Genes were grouped into modules through unsupervised clustering.
(b) Distribution of interferon stimulated gene (ISG) scores for each cell type at baseline (blue), early (yellow), mid (orange), or late (red) EVD. Markers denote the median and the interquartile range. See also Figure B.7.
(c) Volcano plot of differentially expressed genes between monocytes in late EVD compared to baseline. Genes are colored by membership in the "Global down" (blue), "Global up" (red), or any of the 4 monocyte specific modules (green), or if it is annotated as an MHC class II gene (purple).
3.3.3 Cell-type and/or temporally-specific modules help explain cell states and clinical phenomena in EVD

The “NK early up” module is highly specific to NK cells and is most enriched for genes in the CDC42/RAC pathway (OR=219.4, q=1.4x10⁻⁶). This pathway is essential for the polarization of cytolytic granules in NK cells which is a requirement for effective cytotoxicity (Sinai et al., 2010; Tybulewicz and Henderson, 2009) suggesting that NK cells became activated with increased cytotoxicity early, which was lost in mid and late EVD.

The 3 modules “B/T early up”, “B/T late up”, and “B/T late up, mono down” reflect changing gene expression states of lymphocytes at different phases of EVD. “B/T early up” is characterized by many lymphoid activation genes including the canonical marker CD69 (Testi et al., 1994), CD48 (McArdel et al., 2016), and the transcription factor FOS (Foletta et al., 1998). This module is unlikely to represent antigen-dependent activation via the BCR/TCR as it occurs in most lymphocytes and does not coincide with proliferation. Indeed, several of the top up-regulated genes, such as GADD45B and ZFP36L2, are associated with growth arrest. This module therefore is more likely to represent cytokine-mediated activation of lymphocytes.

Later during EVD, the “B/T late up, mono down” module arises. The top associated gene sets implicate DNA repair of UV damage and apoptosis via TRAIL (DACOSTA_UV_RESPONSE_VIA_ERCC3_DN q=3.5x10⁻¹³ and HAMAI_APOPTOSIS_VIA_TRAIL_UP q=3.5x10⁻¹³). This latter gene set is potentially consistent with previous reports of T-cell apoptosis in EVD (Geisbert et al., 2000; Wauquier et al., 2010; Iampietro et al., 2017) and with the lymphopenia in our dataset (Figure 3.2D).
Monocytes were of particular interest; in addition to being the preferred target of EBOV, they had far more gene expression changes during EVD than the other cell types. 834 genes (8.03% of total genes detected) were differentially expressed in monocytes at one or more EVD phases versus baseline, compared to 364 genes (4.03%) for CD4+ T cells, the next highest cell type. These 834 genes contributed to 4 modules that were predominantly associated with monocytes ("Mono early up", "Mono late up", "Mono down", "Mono late down") (Figures 3.3a and 3.3c). We therefore focused our attention on characterizing monocytes in detail.

3.3.4 Monocytes lose expression of MHC class II mRNAs and proteins independent of infection status

Numerous MHC class II (MHC-II) genes, essential intermediaries for engaging the adaptive immune system, were strikingly down-regulated in monocytes by mid and late EVD (Figures 3.3a and 3.3c). Professional antigen-presenting cells such as monocytes, DCs, and B cells internalize, load, and display peptides derived from extracellular proteins on MHC-II proteins at the cell surface to stimulate B and T cells. Genetic deficiency of MHC-II causes severe immunodeficiency (Klein et al., 1993), highlighting the import of this system in defending against pathogens. During a typical infection, IFNγ activates antigen-presenting cells and up-regulates the production of MHC-II (Steimle et al., 1994); during EVD, we observed production of IFNγ (Figure B.7a) and a significant IFNγ response signature (Figures 3.3a and B.7b, Table S1), so it is unexpected that MHC-II would be down-regulated. Previous reports have described loss of HLA-DR during EBOV infection of monocytes ex vivo (Hensley et al., 2002), in experimentally-infected NHPs (Menicucci et al., 2017), and in human EVD cases (Lüdtke
et al., 2016), similar to observations of reduced HLA-DR on monocytes in patients experiencing septic shock (Wolk et al., 2000). However, the dynamics, cell-type specificity, impacted MHC genes, and relationship with EBOV infection status have not been previously described.

We found that monocytes dramatically reduced expression of all MHC-II genes by the early phase of disease and had effectively lost expression of many genes by the middle and late phases (Figure 3.4a). By contrast, B-cells mainly showed reduced expression of MAMU-DRB1 and plasmacytoid and conventional dendritic cells showed modest or no reduction of any of the MHCII genes. This pattern was consistent at the protein level as well for HLA-DR proteins measured by CyTOF which supports the conclusion that these changes in gene expression levels are reflected in protein expression as well (Figure 3.4b). Indeed, even as monocytes upregulated CD8, an activation marker (Amici et al., 2018), they showed dramatic down-regulation of HLA-DR in the protein level (Figures 3.4c and 3.4d). We also noted decreased expression of MHC class I A genes in T-cells, B-cells, and to a lesser extent monocytes.

MHCII levels were almost entirely absent from monocytes by the middle phase when only a small percentage of cells were infected, suggesting that these changes could not be specific to infected cells relative to bystanders. We confirmed that MHC class II levels were comparable if not in fact modestly higher in cells called as infected compared to cells with no detectable EBOV RNA (Figure 4E). Thus we conclude that the MHCII decrease in monocytes observed in EVD is independent of viral infection and is likely to be regulated by the cytokine milieu.

To gain further insight into the regulatory program underlying the decrease in MHC class II, we looked for other non MHC genes that had correlated expression with MHCII in monocytes. We
Figure 3.4: Monocytes reduce expression of MHC class II proteins independent of infection status
(a) Dot-plot showing expression levels of major histocompatibility (MHC) or MHC-associated genes (rows) in several cell types at baseline (B), early (E), middle (M), or late (L) stages of disease (columns). Marker size represents the percentage of cells in that group in which the gene was detected, and color intensity denotes the average expression level of that gene in the group, in log2 transcripts per 10,000 (TP10K).
(b) CyTOF intensity values for the HLA-DR protein in antigen presenting cells at different stages of EVD. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles.
(c-d) Fold change (log2 scale) in CD38 (c) and HLA-DR (d) CyTOF intensity on monocytes at each day post infection (DPI) relative to baseline. Colored lines connect samples from the same animal.
(e) Average gene expression (log2 TP10K) for four MHC class II genes in monocytes at different DPI, stratified by cell infection status. Error bars denote 95% bootstrap confidence intervals.
found that many of the genes most correlated with MHCII expression were functionally involved in the antigen presentation pathway such as CD74 (Spearman $\rho=0.36$), which chaperones MHCII to the endosome and prevents premature binding of antigen (Schröder, 2016), LGMN (Spearman $\rho=0.37$), a protease that cleaves peptides to facilitate presentation on MHCII (Dall and Brandstetter, 2016), and B2M (Spearman $\rho=0.33$), a component of the MHC class I complex (Figure 3.44). In addition, one of the most associated genes was ZFP36 (Spearman $\rho=0.38$), a protein that directly regulates mRNA stability and turnover of MHC class II RNAs (Pisapia et al., 2019). These findings suggest that down-regulation of MHCII in monocytes in EVD is a component of a broader transcriptional module including genes involved in antigen presentation.

3.3.5 **Interferon stimulated genes are down-regulated in infected monocytes relative to bystanders**

Next we characterized genes that are differentially expressed between infected and bystander monocytes as these could represent host restriction factors or genes that change in response to EBOV replication within a cell. For this and all subsequent differential expression analyses, Ebola genes were excluded from the denominator when normalizing cells by library size to avoid a downward bias in the estimated expression levels of host genes in infected cells. We identified 682 genes that were differentially expressed between infected and bystander monocytes (FDR corrected q-value < 0.1) of which 234 changed by more than 35% (Figure 3.5a). 163 (19%) of the 849 genes that were differentially expressed in monocytes at one or more phases of EVD relative to baseline were also differentially expressed in infected monocytes relative to bystanders.
First we noticed that several key ISGs such as MX1, MX2, LAP3, and MNDA were among the genes expressed at lower levels in infected cells relative to bystanders. To further characterize this observation, we compared the average ISG score in infected and bystander cells (Figure 3.5b). ISG levels were higher in both bystander and infected monocytes of late time points than baseline monocytes. However, ISG levels were significantly lower in infected cells relative to bystanders (Ranksum test $P=7.21 \times 10^{-3}$). This effect was even more dramatic for specific ISGs such as MX1 (Ranksum test $P=1.44 \times 10^{-8}$, Figure B.8a). We also analyzed the relationship between ISG score and the percentage of cellular transcripts attributed to EBOV. There was a striking negative correlation between viral load within a cell and the expression of ISGs (Spearman $\rho=-0.60$ $P=2.39 \times 10^{-17}$ - Figure 3.5c). In conjunction with previously published data showing that EBOV inhibits IFN signaling in cell culture, our data suggest that EBOV inhibits this signaling in vivo as well.

3.3.6 Emergence of CD14loCD16lo and CD14hiCD16hi monocyte subsets and enrichment of EBOV infection in the CD14hiCD16hi compartment

Among the genes most associated with EBOV infected cells were CD14 – the lipopolysaccharide co-receptor – and FCGR3 (CD16) – the Fc receptor of IgG antibodies, two genes that respectively mark classical and non-classical monocytes, the two dominant sub-populations in the blood of healthy individuals (Kapellos et al., 2019). To better understand the impact of such heterogeneity in monocyte populations in EVD, we visualized the monocytes with 2 dimensional UMAP embeddings (Figure 3.5d). The monocytes separated by DPI and by EBOV infection status, which is consistent with changing cell states across infection time points and between infected cells and bystanders. In addition, the baseline
Figure 3.5: ISG suppression, co-expression of CD14 and CD16, and expression of macrophage-linked genes, are associated with monocyte infectivity

(a) Volcano plot of fold change (log2 scale) in gene expression between infected and bystander monocytes from DPI 5–8. Genes are colored by membership in gene sets of interferon stimulated genes (ISG), and genes up-regulated (Mac. Up) or down-regulated (Mac. Down) when monocytes differentiate into macrophages in culture.

(b) ISG scores of monocytes at baseline, and uninfected bystanders or infected cells in late infection. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles. See also Figure B.8a.

(c) Scatter plot of ISG score (y-axis) versus percentage of cellular transcripts mapping to EBOV (x-axis) for infected monocytes in late infection.

(d) Uniform Manifold Approximation and Projection (UMAP) embedding of monocyte gene expression data, colored by day post infection, percentage of cellular transcripts derived from virus, CD14 expression values, and CD16 expression values (log2 transcripts per 10,000).

(e) Scatter plot of smoothed expression values of CD14 and CD16 for monocytes in baseline, early, mid, and late disease stages. Cells are colored by infection status. Boxes define the CD14+, CD16+, CD16int, DN, and DP subsets described in the text and numbers denote the percentage of cells in the disease stage falling into each subset. See also Figures B.8B–B.8D

(f) Same as e but showing intensity values for CD14 and CD16 protein levels as measured by CyTOF. The distribution is displayed as a bivariate kernel density plot with 200 randomly sampled cells over-layered as a scatter plot in black.

(g) Average gene expression (log2 TP10K) for four MHC class II genes in monocytes at different DPI, stratified by cell infection status. Error bars denote 95% bootstrap confidence intervals.

(h) Percentage of infected monocytes in each subset in late infection. Height of the bar denotes the mean, and error bars denote 95% bootstrap confidence intervals for the mean.

(i) Association between differentiation score (X-axis), infection probability (left axis, red), and expression of NR1H3, a marker of macrophage differentiation (right axis, blue). Monocytes from late disease are ordered by differentiation score and averages are computed within a sliding window of 400 cells. See also Figures B.9b–B.9d
monocytes separated into 2 distinct clusters, marked by the expression of CD14 (classical monocytes) or CD16 (non-classical monocytes), consistent with the conventional subtyping.

However, the separation between CD14 and CD16 monocyte subsets became less distinct in cells during EVD, with the appearance of cells with reduced expression of both markers, and cells with higher expression of both markers. To better visualize dynamics in the expression of CD14 and CD16, we made biaxial plots of smoothed gene expression values for the different EVD phases (Materials and methods, Figure 3.3e). While 86.6% of cells fell into single CD14hi or CD16hi bins at baseline, this number fell to 22.2%, 30.3%, and 11.8% in the early, middle, and late phases respectively. Instead, a population of CD14lo CD16lo cells (double negatives or DN)s and CD14hi CD16hi double-positive cells (DPs) rose to make up 70.4% and 12.1% of the monocytes respectively by the late phase. We confirmed these trends with CyTOF data where we observed that the loss of CD14hi and CD16hi subsets and the emergence of DP and DN subsets occurred at the protein level as well as the RNA level (Figure 3.5f). While a similar increase in DP monocytes has previously been observed in sepsis (Fingerle et al., 1993; Nockher and Scherberich, 1997), it has not, to our knowledge, been described in any viral infections. Moreover, an expansion of DN monocytes like what we observed has not been described in any pathological context, to our knowledge.

We also observed a set of monocytes with low CD14 and intermediate CD16 mRNA levels (CD16ints) that rose from 3.8% to 12.7% of monocytes at the early time point (DPI 3) before returning to baseline levels (Figure 3.5e). At the protein level, this population had high levels of both CD14 and CD16 positive (thus appearing like DPs). We note that DPI 3 also saw a nearly 2-fold increase in monocyte abundance that was sustained until DPI 4 before levels declined (Figure 3.2d). This population of
early, non-proliferating (Figure B.8b), monocytes therefore likely represents immature cells that are released from the bone marrow along with the neutrophils band cells.

Next, we characterized the rates of EBOV infection and cell cycle replication within the monocyte subsets. Consistent with the fact that CD14 and CD16 were both independently higher in EBOV infected monocytes than bystanders, we found that the DP population harbored a disproportionately high percentage of EBOV infected cells (Figure 3.5g). At the late infection timepoints, 24.8% of DPs were infected compared to 16.7% of CD16ints, 14.5% of CD16hi’s, 5.2% of CD14hi’s and 2.3% of DN’s. Thus, the dominant infected circulating cell population in EVD is a specific double positive subset that emerges in the middle and late phases of infection.

We also characterized the replication rates of the different populations using the Ki67 marker and cell cycle score in protein and RNA levels respectively (Figures B.8b, B.8d). This revealed that the DN population is highly proliferative, with over 9% of cells expressing cell cycle markers by the late time point. Thus, the increase in monocyte replication rates we observed in Figure 3.2g is largely a consequence of the expansion of DN cells.

3.3.7 Monocytes expressing markers of macrophage differentiation are enriched for EBOV infected cells

We observed that a large number of genes that are associated with monocyte to macrophage differentiation were expressed at higher levels in infected monocytes compared to bystanders (e.g. NR1H3, ADAMDEC1, and several Cathepsins such as CTSH and CTSC) (Figure 3.5A). It was previously observed that freshly isolated monocytes are largely refractory to EBOV infection in cell culture but
that EBOV entry factors are up-regulated during macrophage differentiation, allowing increased infection (Martinez et al., 2013). Indeed, the known entry restriction factors CTSL, CTSB, and GNPTAB (Carette et al., 2011; Gnirß et al., 2012) were all expressed at significantly higher levels in cells that were infected with EBOV than bystanders (FDR-corrected q-value $2.1 \times 10^{-7}$, $3.7 \times 10^{-4}$, and $1.3 \times 10^{-2}$, respectively). By contrast, the entry restriction factor NPC1 was not significantly differentially expressed, suggesting that natural variation in the expression of NPC1 do not influence EBOV infectivity within circulating monocytes in rhesus macaques. We suspected that up-regulation of the entry restriction factors CTSL, CTSB, and GNPTAB might be occurring as part of a general differentiation program.

To test the hypothesis that macrophage-like monocytes are more likely to be infected with EBOV than less macrophage-like cells, we sought to determine whether genes associated with macrophage differentiation were up-regulated in infected cells relative to bystanders. We obtained sets of genes that significantly increased or decreased during in vitro differentiation of monocytes into macrophages from previously published RNA-Seq data (Dong et al., 2013). We found that genes that are up-regulated during differentiation were significantly enriched in infected cells (Fisher exact test OR=$3.665$, $P=1.7 \times 10^{-19}$) and genes that were down-regulated during differentiation were significantly enriched in uninfected bystanders (Fisher exact test OR=$4.1$, $P=4.1 \times 10^{-14}$), combined Chi-square goodness of fit test $P=5.8 \times 10^{-37}$). We confirmed these findings by observing similar enrichments of macrophage or monocyte associated genes based on 2 other RNA-Seq studies of in vitro monocyte differentiation (Chi-square goodness of fit test $P=2.3 \times 10^{-13}$ (Saeed et al., 2014), Fisher-exact test $P=3.0 \times 10^{-14}$ (Italiani et al., 2014)). Genes associated with M2 polarized macrophages were more enriched among EBOV-infection associated genes than those of M1 polarized macrophages (OR=$8.9$, $P=3.8 \times 10^{-14}$)
compared to OR=2.8, P=4.5 x 10^{-3} (Italiani et al., 2014)).

To further characterize the role of macrophage differentiation status in EBOV tropism, we assessed how the proportion of infected cells changed along a continuous dimension reflecting the trajectory of macrophage differentiation. To do so, we computed a macrophage differentiation score for each monocyte as a sum of genes previously identified as differentially expressed between monocytes and differentiated macrophages, weighted by the magnitude of the expression change that occurs during differentiation (Materials and methods). Ordering cells from lowest to highest differentiation score, we observe that the infected cell proportion rises nearly three-fold from 4% to nearly 15%. Thus, our data strongly suggests that of all circulating cells, EBOV predominantly infects monocytes that are further along the trajectory of differentiation into macrophages.

We next sought to understand the relationship between the CD14 and CD16 defined monocyte subsets we described previously and the macrophage-differentiation status of monocytes, to determine if these are independent predictors of EBOV tropism. First, we looked at genes that are differentially expressed between DN and DP populations and found that a number of macrophage-differentiation associated genes were enriched in DPs relative to DNs (Figure B.9a). Comparing the overall differentiation score of the different CD14/CD16-marked sub-population further confirmed that DPs had the highest differentiation scores while DNs and CD14hi cells had the lowest (Figure B.9b). Thus, some of the enrichment of infected cells among the DP subset can be attributed to their higher differentiation status, or vice versa.

However, there is substantial heterogeneity in differentiation scores within the CD14/CD16-marked subsets so we sought to determine if differentiation and subset status were independent determinants
of infectivity. We grouped cells into low and high differentiation score bins based on whether or not they were lower or higher than the median score across all monocytes. We then looked at the fraction of infected cells within each CD14/CD16-marked subset, stratified by differentiation score bin (Figure B.9c). This showed that double negatives in the differentiation high bin were less likely to be infected than double positives in the differentiation low bin, even though all cells in the former had a higher differentiation score than all cells in the latter (Fisher exact test \( P=0.037 \)). The same was true when comparing DNs in the differentiation high bin to CD16-int or CD16+ cells in the differentiation low bin (Fisher exact test \( P=6.3 \times 10^{-6} \), \( P=0.027 \)) respectively). However, within the CD14+ and DP subsets, EBOV infectivity was significantly higher in the differentiation-high bin than the differentiation-low bin (Fisher exact test \( P=0.029 \) and \( P=0.038 \)) respectively). As an alternative way to see this, we observed that within the DP and CD16+ subsets, infected cells had higher differentiation scores than bystanders (Figure B.9D). As a final confirmation of the independent effects of CD14/CD16 subset and differentiation score, we fit a logistic regression predicting the binary infection status of each cell using the subset-membership as dummy variables, differentiation score as a continuous variable, and an intercept. All subset-membership dummy variables and the differentiation score were significant predictors of EBOV infection (\( P<0.005 \) for all predictors except DN-status which was \( P=0.071 \)). These findings demonstrate that monocyte CD14/CD16 subset, and differentiation status, independently impact the probability of a cell being infected with EBOV, in vivo.
3.3.8 scRNA-Seq of PBMCs infected ex vivo provides high resolution view of Ebola infection dynamics

We sought to obtain data for a large number of cells at various stages post infection with EBOV in order to better probe how viral and host gene expression changes with viral load in a cell. To do so, we isolated PBMCs from 2 healthy macaques and treated those cells ex vivo, with either live EBOV, EBOV inactivated by irradiation, or media only as a control, and then performed single-cell RNA-Seq on them using Seq-well at 4 hours or 24 hours post treatment (Figure B.10a). In this setting, we could use a multiplicity of infection (MOI) of 0.1 to ensure that a large proportion of cells would be infected. With the 4 hour and 24 hour timepoint, we can identify gene expression changes that occur within a timescale of hours following infection, which we cannot do from infected cells in vivo.

We obtained single cell transcriptomes from 50,646 PBMCs infected, ex vivo, and observed similar cell type representation, clustering by treatment condition, and distribution of EBOV infected cells to what we observed previously with the in vivo collections, with a few notable exceptions (Figures B.10b–d). First, we observed that cells from the 2 animal donors (identified as NHP1 and NHP2) separated in UMAP embeddings of the data (Figure B.10e). This was associated with higher expression of ISGs such as MX1 in cells derived from NHP1 compared to NHP2, suggesting that the difference reflects the fact that one animal mounted an IFN response while the other didn’t (Figures B.10f–g). The ISG signal was specific to cells from the 24H time point that were treated with either irradiated or live virus (Figure B.10h). The fact that the ISG response happened for both the irradiated and live virus conditions, specifically in one animal replicate and not in the other, suggests that it might re-
fect genuine biological differences between the two animals. However, more technical and biological replicates are necessary to confirm this. For the following analyses, we analyze cells from each animal separately as well as together to ensure that any observed signals are not an artifact of confounding with differences between the 2 animal replicates.

Over 60% of total monocytes treated with live virus were called as infected by 24H in both animal replicates (Figures B.11a and B.11b). Surprisingly, 10.4% of the monocytes treated in the irradiated virus condition also contained a statistically significant number of viral reads by 24H. Cells treated with irradiated virus tended to have a significantly lower percent of EBOV reads than those treated with live virus (Figure B.11c). Moreover, viral reads from the cells treated with irradiated virus were substantially more likely to derive from the viruses negative strand, consistent with being reads from genomic RNA rather than mRNA transcripts (Figure B.11d). We estimated that 57% and 61% of EBOV transcripts from cells treated with irradiated virus came from genomic mRNA at 4H and 24H, compared to only 25% and 9% in cells treated with live virus at the same time point. Reads mapping to the positive strand, which are consistent with mRNA transcripts, were enriched for the most 3’ gene NP, with substantially less detection of more 3’ genes than in cells treated with live virus. This suggests that our method can detect RNA from irradiated virus that has entered cells but, as expected, these infections are non-productive and do not generate large quantities of poly-adenylated mRNAs.

Another difference is that while we did not observe infected dendritic cells in vivo, we found that 19.8% of dendritic cells were infected ex vivo (Figure B.11a). This difference could be due to a number of factors such as the increased density of cells in culture, the higher effective MOI we used in the infection, or even changes to the expression states of DCs associated with culture conditions.
Next, we exploited the increased resolution of the ex vivo dataset to analyze patterns of variation in viral transcript abundances across cells. We observed variability over several orders of magnitude in the percentage of virus-derived transcripts in a cell, both in vivo and ex vivo (Figures 3.6a and 3.6b). We refer to the percentage of a cell’s transcripts that are virus-derived as the viral load. While most cells harbored viral loads below 0.1%, a substantial minority had loads of >10%, with maximum detected loads of 52% and 57% for cells in vivo and ex vivo respectively. The observed heterogeneity in viral load was not due to different numbers of transcripts detected per cell, as low and high viral load cells could have a wide range of total transcripts (unique molecular identifiers or UMIs) detected (Figures B.11e and B.11f). A larger percentage of high (>8%) viral load cells were detected ex vivo than in vivo: 33.3% vs 15.1%. We suspect that this is due to two possibilities. First, 18–48 hours is approximately the time necessary for 1 EBOV replication cycle (Nanbo et al., 2013) and so we are potentially more likely to detect cells at late stages of infection than we would be in vivo, where cells may have been infected over a wider range of time periods. Second, we used an MOI of 0.1 which – since monocytes are approximately 10% of PBMCs – amounts to an effective MOI of 1 if virions are only entering monocytes. This is likely much higher than an in vivo MOI where the number of infectible monocytes is much higher and constantly changing due to release from the bone marrow and extravasation into tissues.

We next analyzed the dynamics of gene expression of Ebola’s 7 genes over the course of infection. Transcription of EBOV mRNAs by the viral RNA-directed RNA polymerase (RdRp) L follows the canonical stop-start mechanism described for filoviruses and other non-segmented negative-strand RNA viruses (Brauburger et al., 2014, 2016). L initiates transcription de novo (Deflubé et al., 2019)
Figure 3.6: Viral transcriptional dynamics of infected monocytes in vivo and ex vivo

(a-b) Histogram of the percentage of cellular reads derived from virus (viral load) in monocytes treated with live virus ex vivo (a) or in vivo (b), including only cells containing at least one viral read. See also Figures B.10a–B.10d.

(c) Schematic of the start-stop model of filovirus transcription.

(d-e) Changes in the relative proportion of EBOV genes with viral load, ex vivo (d) or in vivo (e). Infected monocytes treated with live virus (d) or from mid-late stage of disease (e) were ordered by viral load and the average proportion of viral genes was computed within a sliding window of 50 cells. The thick line denotes the window mean and the surrounding band denotes 1 standard error around the mean. The X-axis is plotted on a log10 scale.
at the 3’ end of the genome and processes from 5’ to 3’; at each gene transcription termination signal, L pauses and either falls off the genomic RNA template or reinitiates transcription of a new mRNA for the subsequent gene (Figure 3.6c - reviewed in (Mühlberger, 2007)). As a consequence, NP is the first gene to be transcribed and is transcribed at the highest level, proceeding down the genome to the polymerase gene L being transcribed last and at the lowest level.

To understand how viral gene expression evolves over the course of infection, we plotted the relative expression levels of EBOV genes as a function of viral load (Figures 3.6d-e). At low viral loads, both in vivo and ex vivo, the gene expression distribution roughly matched the expected pattern based on the stop-start mechanism, with most of the transcripts derived from 5’ genes and in particular NP, and the fewest transcripts derived from the most 3’ genes VP30, VP24, and L. However, as viral load increased, we found that GP became the most highly expressed viral transcript. GP is processed into 3 potential transcripts based on “RNA editing”, a process where viral polymerase upon reaching a 7-nucleotide poly-U tract in the middle of the GP gene can incorporate either 7 or more As (Shabman et al., 2014), leading to 3 different reading frames for the 3’ end of the gene which results in 3 different gene products with different functions. The dominant 7A reading frame encodes the secreted glycoprotein (sGP) that has been shown to be an important antibody decoy while the 8A ORF encodes the full-length viral glycoprotein used for viral entry (de La Vega et al., 2015). This observation could be explained by differences in the RNA stability of GP relative to other transcripts, the presence of an additional transcription initiation sites upstream of GP, or even the presence of defective interfering, subgenomic RNAs (Calain et al., 1999) that predominantly transcribe GP. Subsequent experiments are needed to test these different hypotheses.
3.3.9 Ebola infection down-regulates host antiviral genes and up-regulates putative pro-viral genes

We next identified host genes that change in abundance as EBOV levels increase over the course of infection. Instead of testing for differential expression between infected and bystander cells as we did previously to define tropism-associated genes, we looked for continuous association between viral abundance and host transcript levels specifically in infected monocytes. We identified 316 genes that varied inversely and 16 genes that varied directly with EBOV levels in infected cells *ex vivo* with FDR corrected q-value < 0.1 and with more than a 15% change in expression per 10-fold change in EBOV viral load (Figure 3.7a).

Consistent with our previous observation that ISG score decreased with viral load within a cell, we observed that the genes that were most negatively associated with EBOV levels were ISGs, both in vivo and *ex vivo* (Figures 3.7a and 3.7b). In the *ex vivo* data, the most negatively associated gene was *STAT1*—the master transcriptional regulator of the IFN response. EBOV protein VP24 inhibits STAT1 activity by blocking its translocation to the nucleus (Leung et al. 2006) and also through direct binding (Zhang et al. 2012). However, this is the first observation that *STAT1* mRNA levels decrease over the course of EBOV infection.

We visualized the dynamics in the mRNA levels of *STAT1* and several other negatively regulated ISGs (Figure 3.7c). These genes appeared to remain relatively constant as EBOV levels rose to 1% of cellular transcripts, before declining precipitously. This suggests that there is a lag time before EBOV can down-regulate host antiviral genes, perhaps during which time it is transcribing and translates proteins.
Figure 3.7: Ebola virus infection down-regulates host antiviral genes and up-regulates putative pro-viral genes (a-b) Volcano plot of association between host genes expression and viral load, within infected cells treated with live virus ex vivo (a) or in vivo (b).
(c-d) Association between gene expression and viral load for selected host genes in monocytes treated with live virus, ex vivo, at 24 hours post infection. In the right sub-plots, infected cells are ordered by viral load and the average expression of each gene is computed within a sliding window of 100 cells. The X-axis is plotted on a log10 scale. In the left sub-plots, box plots show the distributions of gene expression in uninfected bystander cells, with boxes denoting the interquartile range and whiskers denoting the 2.5th and 97.5th percentiles. Genes that are significantly associated with viral load are shown in color while unassociated negative control genes are shown in gray. Spearman correlation coefficients for the association between viral load and gene expression are listed in the legend. See also Figures B.10e–B.10h.
such as VP24 that effect the inhibition. We also visualized the trajectories of these genes separately for cells from each donor animal to see if the trends would differ given that NHP1 had mounted a broad IFN response while NHP2 hadn’t (Figure B.11g). While ISGs $MX_1$ and $MX_2$ started from a higher expression level in bystanders and low viral load cells of NHP1 than NHP2, they declined following similar trajectories and were independently significant considering cells of each animal separately. This suggests that EBOV is able to productively infect cells that are mounting an ISG response.

We also identified host genes that increased in expression level alongside viral expression levels. The most dramatically up-regulated gene was $DYNLL_1$, which was significantly up-regulated in vivo and ex vivo. $DYNLL_1$ is a multi-functional protein that acts as a “dimerization engine” linking numerous proteins to the dynein motor, the cytoskeletal network, or to other proteins (Barbar, 2008). Intriguingly $DYNLL_1$ was previously shown to interact with the EBOV protein VP35 (Kubota et al., 2009) and to increase EBOV replication in a mini-genome reporter assay (Luthra et al., 2015), and increases replication of rabies virus (Tan et al., 2007), a NNS virus with a similar genome organization. Our data show that it is also up-regulated transcriptionally, starting when EBOV RNA constitutes between 0.1% and 1% of the transcripts in a cell (Figure 3.7d), in a consistent pattern between the two donor animals (Figure B.11h). Combined with the previous observations that experimental over-expression of $DYNLL_1$ increases viral replication, our findings suggest a mechanism whereby this pro-viral gene is, in fact, transcriptionally up-regulated in a natural viral infection context, leading to enhanced viral replication.

Several other genes that we identified as up-regulated alongside viral replication have known or speculative pro-viral functions. For example, HSPA5 is a chaperone protein that is involved in en-
try and maturation of numerous viruses and was previously shown to be an essential host factor for Ebola (Reid et al., 2014). However, this is the first observation that \( \text{HSPA}_5 \) mRNA is up-regulated in EBOV infection. *IARS* - isoleucyl-tRNA synthetase - was also significantly up-regulated, which is interesting, as the availability of tRNA molecules can be a fundamental limitation on the rate of viral translation and replication, which leads several viral families to encode their own tRNAs (Albers and Czech, 2016). In addition, several gene sets associated with amino acid depletion were significantly enriched among genes that were up-regulated by virus. The 2nd and 4th most enriched genes sets were KRIDGE\_RESPONSE\_TO\_TOSEDOSTAT\_24HR\_UP (Tesedostat is an aminopeptidase inhibitor, FDR-corrected \( q = 0.0061 \)) and KRIDGE\_AMINO\_ACID\_DEPRIVATION, (FDR-corrected \( q = 0.049 \)). Our observation that cells containing replicating virus begin to up-regulate genes reflecting depletion of amino acids is consistent with a previous that fatal EVD cases are associated with loss of amino acids in plasma (Eisfeld et al., 2017).

Other up-regulated genes included *NFE2L1* (also known as *NRF1*) and *DDIT3*, two genes that sense endoplasmic reticulum stress and oxidative stress (Kim et al., 2016). The *NFE2L1* binding motif was previously noted to be enriched in differentially expressed genes following infection of a cell line with Marburg virus (Hölzer et al., 2016). Our data now suggests that the transcription factor is up-regulated at the transcriptional level as viral load increases in infected cells. *DDIT3* was previously shown to be up-regulated in monocytes in a non-human primate model of EBOV (Menicucci et al., 2017) and our data now shows that it is up-regulated in response to viral load in infected cells, rather than by the cytokine milieu.
3.4 Discussion

This work represents the first ever study of a risk group 4 pathogen with scRNA-Seq and CyTOF, orthogonal and complementary single-cell methods to quantify RNA and protein as measures of cell type and state. We show that changes in cell type abundance and phenotype occur at distinct phases of EVD, corresponding to the appearance of viremia (early) or clinical signs (mid and late EVD) in a non-human primate model of disease. These data provide an unparalleled opportunity to separately study EBOV-infected and uninfected bystander cells. In agreement with cell culture studies, we find that EBOV down-regulates interferon signaling in infected cells in vivo, but uninfected bystander cells propagate an interferon response, which explains the ISG signatures that have been observed previously in human and animal studies of EVD. In total, our data shed new light on global and cell typespecific gene expression changes throughout EVD, define which cells are preferred targets of EBOV in vivo, and identify genes regulated either by the cytokine milieu or by direct EBOV infection.

Our data demonstrates that monocytes comprise the vast majority of infected circulating immune cells in vivo. Though we detect sporadic EBOV reads in other cell types, these reads could be due to ambient RNA or template-switching that swaps cell barcodes. We develop a statistical model for determining infected cells and show that no other cell types had frequencies of EBOV transcripts exceeding what would be expected by chance, nor did they have higher levels of EBOV transcripts than uninfected cultured MDCK cells which were spiked into PBMC samples as a negative control prior to scRNA-Seq. This supports existing literature which has repeatedly demonstrated that myeloid cells are major targets of EBOV in vivo (Geisbert et al., 2003a,c). We note a couple of discrepancies between
our study and others. First, a recent report shows that human T cells can be infected with EBOV ex vivo, resulting in viral transcription but ultimately abortive infection (Younan et al., 2019); we did not observe significant levels of viral RNA in T cells, either in vivo nor ex vivo, which could be differences due to species (human versus rhesus macaque), MOI (1 versus 0.1 pfu/cell) or experimental setup (T cells versus PBMCs). Second, DCs are also commonly believed to be a major target of EBOV textitex vivo and in lymph nodes textitin vivo (Geisbert et al., 2003a). We observed infected DCs textitex vivo but not in our circulating immune cells textitin vivo, which might reflect an interesting biological difference. Some factors shared between textitex vivo conditions and a lymph node but lacking from circulating DCs, such as high cell density, cell-to-cell contact, or high MOI, might be required for infection of DCs. Further studies of immune cells in various tissues by scRNA-Seq or CyTOF would allow us to better define how cell-to-cell contact and similar factors influence cell susceptibility and permissivity to infection.

We identified global and cell type-specific changes in gene expression, focusing on monocytes in depth because they were the major infected cell population in our study. Monocytes are key innate immune cells that release cytokines and can present antigens on MHC-II molecules to stimulate lymphocytes. Normally these functions are positively correlated; interferon signaling triggers monocyte activation as well as increased MHC-II expression (Steimle et al., 1994). However, throughout EVD, both infected and uninfected monocytes lost expression of MHC-II at the RNA and protein level, despite being highly activated. We observed additional interesting patterns of cell-type and temporally specific regulation of gene expression, modules we label – ”NK early up”, ”B/T early up”, ”B/T late up”, and ”B/T late up, mono down”. We provide insights into these programs, and have made our
data publicly accessible as a resource for the filovirus community. It may be possible to utilize these cell type gene expression profiles to estimate the frequency of each cell type (Avila Cobos et al., 2018), for example, during immune infiltration of tissues, for which bulk RNA-seq data exists (Caballero et al., 2016).

By comparing infected and uninfected bystander monocytes, we identified genes and expression programs likely involved in susceptibility and permissivity to EBOV infection, including CD14hi CD16hi intermediate monocytes and a monocyte-to-macrophage differentiation program. We observed surprising changes in the expression levels of CD14 and CD16 markers, at both the mRNA and protein levels over the course of EVD. Baseline PBMCs contained predominantly CD14hi CD16lo classical monocytes and CD16hi CD14lo non-classical monocytes. Yet throughout EVD, the CD16hi CD14lo population disappeared as previously reported (Lüdtke et al., 2016), along with the CD14hi CD16lo population at the RNA and protein level. These populations were replaced with CD14hi CD16hi intermediate monocytes, which are known to increase in other diseases such as sepsis (Fingerle et al., 1993; Nockher and Scherberich, 1998), as well as an unusual CD14lo CD16lo population, that, to our knowledge, have not been previously described in viral infections. EBOV-infected cells were significantly enriched among the CD14hi CD16hi intermediate monocytes, which correlated with increased abundance of differentiation genes, including known entry factors like Cathepsin B (Martinez et al., 2013; Chandran et al., 2005; Schornberg et al., 2006), in agreement with cell culture studies of EBOV infection (Martinez et al., 2013). On the other hand, the CD14lo CD16lo monocytes were highly proliferative and appeared to be immature monocyte precursors such as monoblasts (Lambert et al., 2017), perhaps the result of emergency myelopoiesis (Chiba et al., 2018; Sayed et al., 2019). This finding in
particular highlights the power of high-parameter methods such as scRNA-Seq and CyTOF; despite little-to-no detection of CD14 or CD16, there were enough other RNA and protein markers to able to reliably detect these CD14lo CD16lo cells as monocytes. Previous work using FACS or IF/IHC relying solely on CD14 and CD16 markers would have failed to detect monocyte precursors, which is perhaps why emergency myelopoiesis has not previously been described for EVD, to the best of our knowledge.

Even among infected monocytes, there was a wide heterogeneity of viral RNA levels, allowing us to identify genes that were likely down- or up-regulated by infection itself, as opposed to being regulated by the cytokine milieu. We observed only a handful of host genes that positively correlated with EBOV RNA, but they appeared to be pro-viral genes that are strong candidates for follow-up functional characterization studies. DYNLL1 was the top hit both in vivo and ex vivo. DYNLL1 interacts with VP35 (Kubota et al. 2009), and this binding increases EBOV RNA synthesis in an minigenome assay (Luthra et al. 2015). DYNLL1 is also pro-viral for live rabies virus (Tan et al. 2007) and may be for other viruses as well. Here, we show that EBOV infection actually up-regulates DYNLL1 expression in vivo, suggesting that EBOV is manipulating cellular pathways to set up a pro-viral cellular environment. Nuclear DYNLL1 typically represses its own transcription factor ATMIN (Jurado et al. 2012); if cytoplasmic VP35 were to sequester DYNLL1 in the cytoplasm as well, ATMIN would be free to drive transcription of DYNLL1. This is one plausible mechanism that would explain how infection up-regulates DYNLL1. HSPA5, NFE2L1, DDIT3, GTF2A1, and IARS are additional intriguing hits, though their effect sizes are much larger ex vivo than in vivo. HSPA5 in particular is essential for EBOV replication (Reid et al., 2014). Thus, our analysis nominates several putative pro-viral genes for
further study.

On the other hand, ISG expression strongly negatively correlates with viral load, strongly suggesting that viral infection down-regulates ISG expression in vivo. While we cannot prove that virus is specifically down-regulating ISGs, rather than preferentially infecting cells with low ISG levels, there are several reasons to favor this interpretation. First, there are multiple well-established mechanisms by which EBOV down-regulates transcription of ISGs, such as by preventing the master transcription factor, STAT1, from translocating to the nucleus (Kash et al., 2006; Harcourt et al., 1999; Leung et al., 2006; Reid et al., 2008). Second, the percentage of EBOV infected cells increases gradually from DPI 4 onward (Figure 3.2i) despite the fact that monocytes are mounting a strong ISG response by then. This suggests that EBOV is able to overcome the inhibitory activities of ISGs and continue replicating. Third, when EBOV infects a cell, viral levels must start at a low level before increasing with replication and transcription. However, we don’t observe any cells with low (0.01%-0.1%) viral titers and low ISG levels (Figures 3.4c and 3.6c). If the causal direction was reversed and low ISG levels supported EBOV progression and high ISG levels inhibited EBOV progression, we would expect to see some cells with low ISG and low EBOV. Thus our data suggests that EBOV is able to infect monocytes that are mounting a full interferon response, overcome the inhibitory activities of ISGs, and transcribe its mRNA levels to high levels.

Finally, we observed extreme heterogeneity of viral RNA within infected monocytes, including increased abundance of GP RNA, above expectation based on the start-stop model of NNS virus transcription. We were able to confidently identify infected cells with 0.1% EBOV RNA out of total RNA, but some cells had up to 60% EBOV RNA, a range of 2.5 orders of magnitude. Similar het-
egeneity has been observed in other infected cells analyzed by scRNA-Seq, including the segmented -ssRNA virus influenza (Russell et al., 2018), the +ssRNA virus dengue (Zanini et al., 2018a,b), and the dsDNA virus HCMV (Hein and Weissman, 2019), but this is the first demonstration of such heterogeneity for an NNS virus. Moreover, viral gene expression varied depending on the total viral RNA in the cell. At low levels of viral RNA, likely reflecting early stages of the viral life cycle, RNA abundance mirrored the genome organization of EBOV, as expected for a non-segmented negative strand (NNS) virus (Brauburger et al., 2014, 2016), with the highest expression of NP and the lowest expression of L. But at high levels of viral RNA, late during the viral life cycle, GP RNA was more abundant than NP RNA. It is possible that there is a stoichiometric advantage to the virus to produce more GP during later phases of the infection when it is preparing for virus release, compared to at the early phase of infection when its focus is more on genome replication. Temporal regulation of viral gene expression is common to virtually all viral families; transcription of genes involved in genome replication and evasion of host detection are typically transcribed at higher levels early in infection and those needed for viral particle production are up-regulated later on. However, such dynamics of viral gene expression have not been previously observed for NNS viruses. Our data suggests that transcription of Ebola virus genes matches the common temporal pattern by increasing abundance of the viral glycoprotein later into infection. The mechanism by which GP would have higher than NP expression is not clear from our data but potentially could be explained by differences in stability of GP mRNA.

Our scRNA-Seq and CyTOF data, identified several molecular phenomena in the natural history of EVD that appear to be shared with immunosuppressive septic shock (Bray and Mahanty, 2003). Loss of MHC-II expression in monocytes (Monneret and Venet, 2014), increased CD14hi CD16hi
intermediate monocytes (Fingerle et al., 1993), and we hypothesize emergency myelopoiesis (Cuenca et al., 2015; Bomans et al., 2018) – are shared with immunosuppressive sepsis (Bray and Mahanty, 2003). Soluble mediators, including cytokines and glucocorticoids, could be key drivers of both phenomena. TNFα signaling has been extensively described as a key inflammatory molecule in EVD that lead to systemic loss of vascular resistance and shock. Glucocorticoids have been less well studied in EVD, but have been shown to decrease MHC-II ex vivo (Hawrylowicz et al., 1994) and in vivo during sepsis (Tulzo et al., 2004), and can reduce CD14 expression (Nockher and Scherberich, 1997) while increasing the abundance of intermediate monocytes (Liu et al., 2015). Indeed, the connection between EVD and sepsis may be direct: studies have found bacterial invasion during EVD in NHPs (Reisler et al., 2018) and in humans (Carroll et al., 2017), with immune signatures that resemble sepsis (Eisfeld et al., 2017). The loss of MHC-II and alteration of monocyte frequency and function in particular will be important considerations for vaccine design, which depend on monocyte differentiation and antigen presentation.

In summary, we have deeply characterized the transcriptomes of circulating immune cells in a natural history model of EVD. With an unbiased single-cell approach, we identified novel monocyte subpopulations that play a key role in EVD pathogenesis and discerned the molecular correlates of EBOV tropism in vivo. Furthermore, we analyzed patterns associated with the progression of viral infection and demonstrated previously unappreciated temporal dynamics in EBOV expression. This dataset can serve as a valuable resource for future studies of the the host response and viral transcriptional dynamics in Ebola virus disease.
3.5 Materials and methods

3.5.1 Natural history study

This study utilized the the Ebola virus/H. sapiens-tc/COD/1995/Kikwit-9510621 (EBOV/Kikwit; Gen-Bank accession; Filoviridae: Zaire ebolavirus) isolate for the in vivo and ex vivo challenges as it is the standard challenge stock defined by the filovirus animal non-clinical group [FANG] for testing product efficacy for FDA approval and is well characterized.

This study included 27 outbred rhesus macaques of Chinese origin. Two baseline blood samples were collected from all study animals around 30 and 14 days prior to infection. Animals were exposed to the EBOV/Kikwit isolate diluted to a target concentration of 1,000 PFU in a volume of 1 mL/dose. Injections were delivered intramuscularly into the left lateral triceps muscle within a 5 month period.

3.5.2 Clinical observations and scoring

Beginning on day post-infection (DPI) 0, animals were observed once daily in the morning and given a clinical score based on five criteria: overall clinical appearance and signs of hemorrhage; respiratory rate, mucous membrane color, and dyspnea; recumbency; non-responsiveness; and core temperature. Each criterium was assigned a score between 1 and 10 (based on Appendix C) and all scores were added together. When at least one animal in a cohort progressed to a clinical score of three, a second officially documented observation was performed in the afternoon. When at least one study animal achieved a clinical score of five, a third observation was added and officially documented. This observation oc-
curred at any point following the second observation but was ideally conducted as late in the daylight cycle as possible. Once an animal reached a combined score of at least ten, the animal was humanely euthanized.

3.5.3 Whole blood collection

Blood was drawn from anesthetized animals into BD vacutainer plastic serum separator tubes (SST) and in BD vacutainer plastic blood collection tubes with K$_3$EDTA (Becton Dickinson, Franklin Lakes, NJ, USA) and were incubated at room temperature for at least 30 minutes prior to being centrifuged at room temperature for 10 minutes at 1800 x g. The separated serum was then transferred into a clean 2 mL Sarstedt tube. The sample condition was recorded for any presence of hemolysis, icterus, and lipemia. K$_3$EDTA tubes were again mixed by gentle inversion prior to hematology and PBMC isolation.

3.5.4 Hematology and complete blood counts

250 μL of each whole blood sample was transferred to a clean 0.5 mL Sarstedt tube and analyzed on a Sysmex 2000i XT (Sysmex Corporation, Kobe, Hyogo Prefecture, Japan) following an initial quality check on the instrument. Parameters analyzed by this instrument were: counts of basophils, eosinophils, lymphocytes, monocytes, neutrophils, white blood cell count; percentages of each cell type; and mean platelet volume.

To estimate the abundance of lymphocyte cell types, we multiplied the CBC lymphocyte count by the proportion of lymphocytes of each cell type (CD8 T cells, CD4 T cells, NK cells, and B cells) which
was calculated from the unsupervised clustering of the CyTOF data (see below).

3.5.5 EBOV viral load by RT-qPCR

Harvested serum was stored in a -80 °C freezer for no longer than 90 days before viral titers were determined by RT-qPCR. 70 μL of sample inactivated by Trizol LS was added to 280 μL of buffer AVL (Qiagen, Germantown, MD) with carrier RNA. Samples were then extracted using the QIAamp Viral RNA Mini Kit (Qiagen) in accordance with the manufacturer’s instructions, eluted in 70 μL of buffer AVE (Qiagen), aliquoted, and frozen. Viral RNA titer was determined using an experimental BEI Resources Critical Reagents Program EZt RT-qPCR kit assay in accordance with the manufacturer’s instructions. Sample titers were reported as RNA genome equivalents (GE) per mL of sample.

3.5.6 PBMC purification

Following centrifugation at 1800 x g for 10 min at RT and removal of EDTA plasma from the K3EDTA blood tube, phosphate buffered saline (PBS, Gibco, ThermoFisher, Waltham MA) was added to the pelleted cells to double the original whole blood volume collected in each tube. The PBS-blood cell mixture was then gently poured into an Accuspin tube (Sigma, St. Louis MO) containing RT Histopaque over the supplied filter. Accuspin tubes were then centrifuged at 1000 x g for 10 min at RT with the brake set to 1. Following centrifugation, the top, clear supernatant layer was removed to within 0.5 cm of the cloudy white layer containing PBMCs. The cloudy PBMC layer was then transferred to a clean 15 mL conical tube and brought up to 10 mL in PBS + 2% fetal bovine serum and mixed by inversion. The sample was then centrifuged at 300 x g for 10 min at 4 °C with the brake set to 1. After centrifuga-
tion, the supernatant was removed and the cell pellet was resuspended in PBS + 2% HI FBS to a final volume 10 mL and mixed using gentle raking to wash the cells. The wash step was repeated 2 more times with the centrifuge set to 200 x g for 10 min at 4 °C with the brake set to 1. Samples were then resuspended in 9.5mL PBS + 2% HI FBS for counting. PBMC counts were then determined using the Countess Cell Counting system (Thermo Fisher Scientific).

Once cell counts were determined, 0.5 mL was transferred to another tube for Seq-Well, while the remaining 9 mL volume was fixed for CyTOF.

3.5.7 Seq-Well

We performed Seq-Well as described previously (Gierahn et al., 2017) with the S3 protocol (Hughes et al., 2019) and some modifications to adhere to the BSL-4 environment.

After loading beads and cells and sealing Seq-Well arrays, we placed them in a -80 °C freezer until further processing – this step was required due to time constraints in the BSL-4. Later, sealed Seq-Well arrays were removed from the -80 °C freezer, placed in 4-well dishes, and allowed to come up to room temperature for at least 30 min. Arrays were then covered in 5 mL Lysis Buffer per protocol.

We performed RNA hybridization and RT as specified in the protocol (Hughes et al., 2019). After RT, we collected beads by centrifugation at 1000 x g for 1 min at room temperature. We resuspended beads with GeneXpert lysis buffer for inactivation, which was required prior to removal from the BSL-4 laboratory according to standard operating procedures. After removal, we washed beads thrice with TE buffer containing 0.01% Tween 20 and shipped at 4 °C for further library construction and sequencing, which was performed with the S3 protocol (Hughes et al. 2019).
3.5.8 DASH of select Seq-Well libraries

We performed DASH according to reaction conditions laid out for S. aureus Cas9 (SauCas9) (Yourik et al., 2019). We incubated 10 pmol gRNA with 5 pmol SauCas9 at 25 °C for 10 min, and then added up to 5 fmol DNA (2000:1000:1 RNA:protein:DNA ratio) and NEBuffer 3.1 (NEB) to 1X. We incubated this reaction at 37 °C for 2 h, and then quenched by adding EDTA to 50 μM, SDS to 1%, and 4 total U of Proteinase K (NEB) at room temp for 10 min. We removed degraded fragments with two consecutive 0.8X SPRI purifications using Ampure XP beads, eluted with water, and performed 6–9 cycles of PCR with the NEBNext Ultra II Q5 Master Mix (NEB) using Illumina P7 and the Seq-Well P5-TSO hybrid primer.

3.5.9 CyTOF

1 mL of 16% paraformaldehyde (PFA) was added to the remaining 9 mL PBMCs to fix the cells. Samples were incubated at RT for 10 min followed by a final centrifugation at 600 x g for 5 min at 4 °C with the brake set to 9. The supernatant was then removed and 1 mL of PBS + 5% HI FBS was added for every 3 x 106 cells in each sample (i.e. 2 mL for a sample containing 6 x 106 cells), aliquoted into 1 mL aliquots in cryovials, and stored in a -80 °C freezer to await further processing.

Isolated PBMC samples fixed in 1.6% PFA were allowed to come to room temperature before being transferred to a 96-well plate. After being transferred, barcoding, surface cell staining, and intracellular staining were performed as instructed by protocols provided by the Nolan laboratory.

Following CyTOF staining, samples were inactivated according by adding 250 μL of 16% PFA to 750
μL of each sample, for a final concentration of 4.0% PFA and incubated at 4 °C overnight. Following this incubation, samples were centrifuged at 600 x g for 5 minutes at 4 °C and the supernatant removed to 100 μL. Samples were then resuspended in 1 mL 4% PFA in PBS and transferred to a clean 2 mL cryovial. Samples were then removed from the BSL-4 using a dunk tank and frozen at -80 °C within 30 min of PFA addition. Samples were processed on a CyTOF instrument within a BSL-2 lab.

3.5.10 Demultiplexing the raw scRNA-Seq data

We demultiplexed the raw sequencing image files using bcl2fastq with the parameters: --mask-short-adapter-reads 10 --minimum-trimmed-read-length 10 --no-lane-splitting --barcode-mismatches 1

3.5.11 Obtaining digital gene expression count matrices


We aligned to a hybrid genome of Mmul_8.01 and EBOV Kikwit accession KU182905.1 and Ensembl gene build 92.

3.5.12 Filtering, preprocessing, and clustering the scRNA-Seq data

We utilized the Scanpy analysis toolkit to preprocess, cluster, and visualize our scRNA-Seq data (Wolf et al., 2018b) first removed cells with fewer than 300 genes detected, and genes detected in fewer than 10 cells. Next we filtered cells with >10% of their UMIs derived from mitochondrial genes and cells where
more than 95% of the UMIs were not assigned to the sense strand of an exon in our gene annotation. We then filtered genes that were no longer detected in at least 10 cells.

We transformed expression values to \( \log_2(\text{TP}10\text{K}+1) \), i.e. normalized each cell profile to sum to 10,000, added 1 to every expression value, and then took the natural log of every expression value. We subsetted the gene expression data to overdispersed genes using the default parameters for the Scanpy highly_variable_genes function (which resulted in 1,816 genes being kept).

We then performed an initial round of clustering to identify and filter spiked-in MDCKs. We used the Louvain community detection algorithm (Blondel et al., 2008b) with 200 nearest neighbors, 20 Principal Components, and resolution=1.0. In this and every clustering iteration, we determined the number of Principal Components by choosing an elbow on the log Eigen-value plot. We identified MDCKs as a group of cells that separated visually by UMAP, was characterized by expression of epithelial cell markers (e.g. COL3A2 and EMP1), and was split into 3 clusters by the Louvain clustering.

To ensure that no cells that were doublets of MDCKs were retained, we determined the average expression profile of each cluster (in \( \log(\text{TP}10\text{K}+1) \)) and regressed this against the gene expression profile matrix using Non-negative least squares. Cells that clustered into the 3 MDCK clusters were marked as MDCKs and cells that had a cumulative usage of the 3 MDCK clusters in the regression profile were filtered as possible doublets with MDCKs.

Next we identified genes that were correlated with whether the cells were loaded fresh, or underwent a freeze-thaw cycle prior to Seqowell array loading. To do so, we first classified cells into T cells, B cells, monocytes, and neutrophils based on the initial Louvain clustering. We then looked for differential expression between fresh and freeze-thawed cells of each cell-type in baseline, early, middle, or late
periods using a Ranksum test. We filtered genes that changed by more than 2-fold and had P-value < 0.01 in 3 or more of the 16 individual tests which excluded 396 genes.

We then used Canonical Correlation Analysis (CCA) as implemented in the SEURAT package (Stuart et al., 2019) to remove variation due to whether the cells derived from fresh or freeze-thawed PBMCs. We ran CCA with the following parameters: selection.method = “vst”, nfeatures = 2000, dims = 1:21, anchor.features=2000.

After CCA, we ran Louvain clustering with 200 nearest neighbors, 21 PCs, and resolution = 1.0 and the clusters separated on UMAP into B-cell T-cell/NK-cell, Mono/DC, Neutrophil, and a cluster that seemed to contain doublets and/or high levels of ambient RNA because it expressed markers of several cell types.

We next used iterative Louvain clustering to further sub-stratify the broad clusters into more finely resolved cell type clusters while removing suspected doublets at each iteration. At each iteration, we subsetted only cells from a broad cluster (Mono/DC, B-cell T-cell/NK-cell, Neutrophil). We included cells from the ambiguous cluster with Mono/DCs. We then repeated the normalization and high-variance gene selection procedure and ran Louvain clustering with a resolution between 1.0 and 2.0. We identified clusters that co-expressed markers of two canonical cell-types and filtered those out prior to each clustering iteration. 3 clustering iterations per cell type was sufficient to remove apparent doublets and achieve the cell-type resolution shown in Figure 2B-C.
3.5.13 Differential expression testing

We performed differential expression testing using MAST (Finak et al., 2015) using log2(TP10K + 1) transformed values as recommended. For comparisons of cell types at each disease stage compared to baseline, we adjusted for the following covariates: (1) the percentage of mitochondrial reads in a cell (2) the number of genes detected in a cell and (3) a binary variable reflecting if the cell came from a fresh sample or one that was freeze-thawed. For the in vivo comparisons of infected vs. bystander monocytes, we adjusted for those same 3 variables as well. However, we also restricted to only looking in monocytes in DPI 5-8 and calculated TP10K values after excluding all viral genes. There was no fresh vs. freeze-thawed covariate for the ex vivo data. However we also used a covariate corresponding to which animal replicate the cell derived from. We only considered cells treated with live virus at the 24H timepoint for this comparison.

For differential expression tests correlating viral load with gene expression, we used log10 viral transcripts per 10,000 as the regressor with the same covariates as the bystander vs. infected comparisons. We only included cells with 1 or more reads for these comparisons.

3.5.14 Identifying differential gene expression modules

We first identified genes that showed statistically significant differential expression (q<0.01) and a fold-change of >35% in at least one cell type and disease stage relative to baseline. We clustered the log fold-changes profiles of genes the remaining genes. As log fold-changes can be high for non-significant genes reflecting noise due to low expression levels, we set the log fold-change of genes that had a P-
value of cell-type/stage above 0.2 to 0. We then clustered the genes using KMeans with the default parameters in scikit-learn 0.21.3 (Pedregosa et al., 2011). We tried several values of K above and below 12 and observed that above 12 the global module was being split into multiple sub-modules with similar profiles. Below 12, highly consistent programs were being obtained but the NK-cell specific module was being merged with the B/T up module.

3.5.15 Detecting EBOV infected cells

Our first step in detecting infected cells is estimating the percentage of each cell’s transcripts that are due to ambient RNA. To do so, we first obtain a representation of each cell in gene expression program usage space using cNMF (Kotliar et al., 2019). We run cNMF using the default parameters for K=6 to K=24 and 100 iterations. We selected K=21 based on the PCA Skree plot, noting that since we are merely obtaining a representation for the cells, it is better to over-estimate the dimensionality rather than under-estimate.

cNMF returned a spectra of 21 GEPs that capture cell-types and activities in the data. However, it does not learn a GEP corresponding to ambient RNA contamination. We therefore estimate the percentage contribution due to ambient RNA, treating the 21 GEPs as covariates. First we normalize the TPM GEP spectra returned by cNMF to each sum to 1. We obtain an estimate of the ambient RNA profile of each sample by considering all cell barcodes that had fewer than 50 UMIs and obtaining the proportion of reads assigned to each gene. We append this ambient profile to the normalized GEP spectra and fit non-negative least squares to re-infer the usage for each GEP while accounting for the ambient RNA profile. We repeat this separately for each sample. Finally we normalize the
returned usage profile to sum to 1. The coefficient for the ambient RNA profile thus corresponds to the proportion of transcripts predicted to be due to ambient RNA.

We then calculate for each cell a threshold number of reads for calling the cell EBOV positive such that it will only be positive by chance 1% or less of the time. We assume the number of Ebola reads due to ambient RNA is binomially distributed with \( \mathcal{N}_i \) corresponding to the library size of the cell and \( p_i \) corresponding to the probability of a read in cell i mapping to Ebola due to ambient RNA. \( p_i = x_i \cdot p \) where \( x_i \) is the estimated proportion of ambient RNA in the cell and \( p \) is the proportion of EBOV RNA in the ambient RNA profile of the cell. We determine the threshold RNA by using the inverse survival distribution function of the Binomial distribution with the \( p_i \) and \( \mathcal{N}_i \) parameters and the 1% threshold.

3.5.16 Scoring cells for cell cycle and interferon stimulated gene levels

We scored cells for cell cycle using the Scanpy score_genes function with TOP2A and MKI67 as the markers and default parameters.

For interferon stimulated genes, we determined marker genes by intersecting the 117 genes in the global up module with the union of the following literature-based genesets:

HECKER_IFNB1_TARGETS, BROWNE_INTERFERON_RESPONSIVE_GENES, MOSERLE_IFNA_RESPON HALLMARK_INTERFERON_ALPHA_RESPONSE HALLMARK_INTERFERON_GAMMA_RESPONSE
downloaded from msigdb (Liberzon et al., 2015b). The result was a set of 48 genes. Again, we used the Scanpy score_genes function with these 48 genes as the markers and default parameters.
3.5.17 Gene expression smoothing for visualization

We smoothed CD14 and FCGR3 (CD16) expression values for visualization of scatter plots in figure 5e. We used MAGIC (Moon et al., 2017) as implemented in Scanpy for this purpose using log2(TPtoK) values for all genes computed just on monocytes and with the following parameters: knn=5, t=3, solver="exact".

3.5.18 CyTOF clustering analysis

We first down-sampled cells by randomly sampling a fixed number of cells from each disease stage to yield 100,000 cells from each period. We then normalized the data to intensity values by dividing by 5 and using the hyperbolic arcsin transformation. We then filtered doublets by plotting biaxial plots of CD3 vs. CD19, CD19 vs. CD19, CD3 vs. CD19, and CD4 vs. CD8 and drawing thresholds to filter the double positives. We also set a ceiling on the maximum value for each gene at the 99.999th percentile to reduce the contribution of outliers. Finally we clustered the data with 11 PCs (determined by inspecting the Skree plot), n_neighbors=200, and resolution=1.0.
General Discussion

4.1 Overview of the previous chapters

Single-cell transcriptomics is enabling the systematic dissection of gene expression in mammalian tissues. Exciting efforts are underway to catalog the characteristic expression profiles of every cell type in humans and mice (Regev et al., 2017; Consortium et al., 2018), to trace the lineage trajectories that give rise to each cell type, to localize gene expression in spatial context within tissues, and to define patterns of dysregulated expression that correspond to disease states. This rapid progress has
necessitated the development of new conceptual models and computational tools to make sense of the large-scale high-dimensional datasets that are being generated.

In the introduction to this thesis, I reviewed a framework for understanding gene expression as modular: genes involved in the same biological processes are co-regulated at the transcriptional level so that they can be co-expressed in response to the appropriate biological context. Such co-regulated, co-functional genes constitute a gene expression program (GEP) – an analogy to computer programs which are called to accomplish discrete tasks and can be combined by a software engineer to build ever more complicated programs. This framework is simple and intuitive, can help explain the evolvability of biological complexity, and is consistent with the patterns of co-expression that have been observed since the first transcriptome-scale gene expression profiling experiments in yeast.

This introduction set the stage for the first data chapter where I described a new approach to discovering GEPs from single-cell data, Consensus non-negative matrix factorization (cNMF). The main advantage of cNMF over hard-clustering based approaches is that it models every cell as a mixture of multiple GEPs, rather than implicitly assuming that all cells express a single GEP. This allows it to appropriately model instances where multiple discrete programs are contributing to the expression profile of a cell. In practice, we show that with existing single-cell RNA-Seq (scRNA-Seq) datasets, it frequently discerns identity GEPs that capture the characteristic expression profile of a cell type, and activity GEPs that correspond to the characteristic set of genes induced by any cells carrying out a corresponding cellular process. Application to simulated and real scRNA-Seq datasets showed that cNMF could discover activity GEPs in an unsupervised fashion. Many of the GEPs we found could easily be assigned a biological function based on interpreting their genes (e.g. hypoxia, cell-cycle, rewiring
of neurons following repeat depolarization). Other signatures replicated in multiple distinct datasets suggesting that they reflect an important axis source of variation in gene expression but we could only speculate about the corresponding biological function (e.g. a putative neurosecretory program).

In the third chapter, I analyzed scRNA-Seq and CyTOF data from a non-human primate model of Ebola virus disease to characterize the gene expression changes that occur in immune cells during Ebola virus infection. While cNMF proved to be a useful tool for representing the variation in this data and correcting for technical artifacts, many analyses benefited from well-defined cell phenotype labels (e.g. infected cell vs. uninfected bystander) that allowed us to analyze the data in a supervised fashion with differential expression techniques.

Nevertheless, the GEP framework played a key role in our analysis. We found that an interferon response program was up-regulated widely across immune cell types, but specifically suppressed by virus infection. We identified several GEPs with interesting links to observed clinical phenomena in Ebola virus disease (EVD): for example, T-cells up-regulated genes mediating apoptosis through the TRAIL pathway potentially explaining the observed lymphopenia in our data; Monocytes down-regulated MHC genes and accessory proteins, consistent with failed induction of an adaptive immune response. An unexpected finding in the data was that monocytes varied based on their degree of induction of a macrophage differentiation program and EBOV infection was highly enriched among the most differentiated cells. This could be explained by the fact that several known EBOV permissivity factors such as CTSB, CTSL, and GPNMB are up-regulated as part of this program.

Taken together, the GEP concept helped motivate and shape the analyses of the two projects described in this thesis, first in the development of a tool applicable to diverse datasets, then in a specific
analysis of Ebola virus disease. While this concept was useful, there were several stumbling blocks and limitations I repeatedly encountered when I attempted to interpret co-expressed genes as a functional program. First, methods to ascribe function to co-expressed genes based on gene-set enrichment can often be unsatisfactory for all but the most prominent GEPs (e.g. interferon response and cell cycle). Second, it is generally not possible to infer the regulator of a GEP without subsequent experiments that are mostly bespoke to specific disease contexts. Finally, while single-cell analysis can frequently infer broad identity and activity GEPs that vary independently in a dataset, such GEPs are typically made up of multiple sub-programs that generally cannot be teased apart from one another. In the next sections I briefly overview these challenges and limitations and suggest some possible directions forward to tackle them.

4.2 Interpreting the function of sets of co-expressed genes

Gene set enrichment analysis (GSEA*) is a bedrock tool for the exploration and interpretation of transcriptomic data. I performed GSEA at some point for essentially every dataset analyzed in this thesis. However, as a method, it is frequently viewed with a mixture of scorn and frustration by bioinformaticians. Even when it is helpful, its results must typically be taken with a grain of salt. Why is this?

GSEA is the process of determining whether a user-generated geneset shows statistically significant overlap with a second gene-set that is typically derived from the literature. There are different formulations of this task for discrete gene-sets vs continuous vectors of gene scores (such as the GEP signatures

---

* I am using the abbreviation GSEA to describe the general process of looking for statistical enrichment of a specific query gene-set within a list of genes or a differential expression profile, not referring to the specific statistical test that is also called GSEA (Subramanian et al., 2005).
that are output by cNMF). For simplicity, I describe only the case of testing one list of genes for enrichment against a second list of genes, but analogous considerations exist for continuous gene profiles as well. The typical context for this is when the data analyst generates a list of genes from their data from a clustering analysis to identify groups of co-expressed genes, or by setting a threshold for statistical significance on a differential expression analysis. By recognizing some of the genes in the list, and reviewing their literature, one can begin to speculate about what biology processes underlying the data. However, we also need a systematic way to understand the results and to ensure that interpretations are statistically sound. This is where GSEA comes in.

Typically the user-generated query set of genes is tested for statistical enrichment against all of the literature-derived genesets in a database. Common databases for this purpose include Gene Ontology (Harris et al., 2004) which contains manually curated gene sets corresponding to “molecular functions”, “cellular components”, and “biological processes”, as well as the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) which includes gene sets derived from previous differential expression experiments involving chemical or genetic perturbations to cells. If the query geneset shows significant enrichment for a Gene Ontology gene-set annotated as a biological process such as “cell cycle” or a MSIGDB gene-set such as “genes up-regulated by interferon alpha,” then they have evidence that cell cycle or interferon response is contributing to the gene expression signal in their data. Ideally, one would like to conclude from this analysis that a cell cycle GEP or an interferon alpha response GEP is being induced or repressed.

In my experience, the frustration elicited by such analyses derives from the fact that it can output impressively significant enrichments for gene-sets which are totally unrelated to the biological system...
that generated the query. This can occur due to statistical artifacts in the testing procedure. For examples, gene set databases are biased for highly-expressed and well-studied genes, larger gene sets tend to yield inflated significance values, and variables such as gene length can confound the tests (Newton and Wang, 2015; Goeman and Bühlmann, 2007; Simillion et al., 2017).

However, it can also be due to the fact that most programs are made up of many sub-programs and any overlap in the sub-programs can result in a statistically significant enrichment that may not correspond to the most interpretable enrichment. For example, in my enrichment testing for genes that were 122 down-regulated across multiple cell-types in Ebola virus disease, the two most significantly enriched Gene-set related were “Housekeeping genes”, and genes down-regulated in PBMCs of patients with Sickle-cell relative to controls. The third most significant geneset was a set of regulators of Translation, which could make sense in our data as inhibition of translation is a potent anti-viral response regulated by interferon (Li et al., 2015). And the genes in set overlapped significantly with the “Housekeeping genes” and “Sickle-cell associated genes” which potentially explains why those were also associated.

I see two fundamental aspects of the limited interpretability of GSEA. First, mere statistical enrichment often does not capture what we are looking for: just because a gene-set overlaps more than expected by chance doesn’t mean that there is some coordinated activity of the genes taking place. Second, gene-sets are typically noisy derived from myriad different laboratories introducing different biases and confounders into the gene-sets that can reduce interpretability of the enrichment.

The first challenge might be partially addressable by new statistical testing procedures that move beyond simply testing marginal association of each gene-set in a database. There is an analogy between
this and the problem of fine-mapping genetic variants that impact gene expression in cis (expression quantitative trait loci or eQTL studies). Traditionally, those studies had tested each variant within a window for marginal association with gene expression and did not seek to discern which specific variants might be causally driving gene expression. Because variants are typically highly correlated with each other because of linkage dis-equilibrium, many variants would be statistically significant and there was no guarantee in marginal association testing that the causal variant would tend to have the highest association (Broekema et al., 2020). New procedures for jointly testing sets of variants are able to at least return multiple “credible sets” of variants that are equally likely to be causally impacting gene expression, while distinguishing distinct credible sets that are acting independently (Wang et al., 2019; Benner et al., 2016).

The analogy for gene-set enrichment testing would be to jointly consider related gene-sets in a database which typically include large numbers of overlapping genes, and try to estimate which individual genes are likely driving an enrichment of multiple genesets and to consider some model, which may be specific for different types of queries, to determine which gene-sets are most relevant to the signal. For example, Gene Ontology gene-sets are explicitly structured into hierarchies where all genes in child-nodes in the hierarchy are contained in their parent nodes as well. This could lend itself to hierarchical models for determining which level of the hierarchy is the best stopping place for defining the enrichment. A few approaches to jointly testing gene-set enrichments of multiple sets are starting to be developed with this kind of idea in mind (e.g. (Simillion et al., 2017; Jiang and Gentleman, 2007; Lu et al., 2008; Falcon and Gentleman, 2007). However, these methods are rarely used in practice and there is likely room for development in this space.
The two challenges to GSEA—simple statistical enrichment being under-specified for many biological questions, and limitation in the availability of comprehensive gene-sets—are actually potentially related. The availability of a comprehensive catalog of query states can directly enable certain types of enrichment analyses. For example, consider the problem of assigning cell-type labels to cells in a single-cell transcriptome dataset. This can be formulated as a specific case of GSEA: one would like to know which if the genes that are expressed in a cell statistically “match” some previously defined cell-type profiles. Methods that directly tackle this problem have been developed (reviewed here (Abdelaal et al., 2019)). Such methods can take advantage of the systematically defined catalogs of cell-type profiles that are being generated by projects such as the Human Cell Atlas and Tabular Muris (Regev et al., 2017; Consortium et al., 2018).

Perhaps the approach described in the previous paragraph—generating a reference catalog of cell-type profiles to enable inference of cell-types in new data—could be applied as a more general approach to structured GSEA. Given a standardized catalog of transcriptional responses, e.g. to cytokines treatments or transcription factor over-expression, could we determine which individual or combination of factors are driving observed differential expression patterns in a new dataset? Identifying the transcriptional regulator(s) of gene expression programs

At various points in this thesis, we identified a gene expression signature that we hypothesized constituted a GEP, performed GSEA with more or less success, and were left wondering what signal transduction cascade and transcription factors could be regulating the gene. In Ebola virus disease, many of the expression changes are believed to be mediated by inflammatory cytokines released into systemic circulation. Canonically, the most dominant expression change we observed, the interferon response,
is regulated by extracellular interferon. However, a systematic approach to ascertaining the regulators of expression programs would greatly enhance our ability to interpret gene expression and to perform experiments to understand the role of specific gene expression programs in the relevant context such as a specific disease.

Seeking to understand the regulators of the differential expression modules in my Ebola data, or the signatures in the cNMF analysis, I had 3 main approaches at my disposal: (i) look for correlations between the expression levels of putative regulator genes (e.g. TFs) and the co-expressed genes (2) GSEA enrichment analysis with gene-sets annotated for known regulators or using known transcription factor motif elements.

The first approach has been utilized in a number of cases to generate hypothetical links between putative regulators (e.g. transcription factors or genes up-regulated in cancer through copy-number aberration) and gene-expression programs (e.g. (Segal et al., 2003b; Akavia et al., 2010)). In the context of the EBOV data, we have begun to look if cytokine protein levels measured in the serum of blood sample correlate with the levels of induction of differential expression modules in cells from the same sample. However, as regulators may themselves be transcriptionally regulated, it is generally not possible to prove a causal link from such approaches without additional data. For example, interferon alpha levels in the blood are less correlated with our interferon response signature than lLirRN which are likely induced as interferon-stimulated genes. Nevertheless, this approach can perhaps prioritize candidate regulators that correlate with the expression levels of a gene expression program for subsequent experimental testing.

The second approach I undertook is to do GSEA with gene-sets associated with the response to
known regulators. A particular case of this is the use of gene-sets consisting of genes that have a specific motifs in their promoter that respond to a known transcription factor. The MEMe suite constitutes a databases of DNA motif elements and tools for searching for them in the promoters of a set of genes that can be used for this purpose (Bailey et al., 2009). This approach is mostly hindered by our limited awareness of transcription factor regulatory logic for most transcription factors. The ENCODE project is advancing this area by systematically cataloguing transcription factor binding sites with CHiP-Seq. In addition, novel computational methods are increasingly available for turning those catalogues into predictive models (e.g. (Kelley et al., 2016)). So it may soon become possible to integrate predictive models of TF binding to interpret patterns of co-expressed genes.

4.3 Identifying sub-modules that co-vary in most natural contexts

Many gene expression sub-programs will tend to be tightly co-regulated in many natural conditions, even if there are subtle differences in their gene expression regulation. For example, in the type 1 interferon response, most interferon stimulated genes are induced by the ISGF3 transcription factor complex which binds to the canonical Interferon-Stimulated Response Element (IRSE) motif, but a substantial minority are induced primarily by non-cannonical transcriptional regulators in response to the same interferon cytokine such as IRF1 (Schneider et al., 2014; Mostafavi et al., 2016; Schoggins et al., 2011). The modular gene expression program concept leads to the hypothesis that there is a reason why these subsets of genes have distinct patterns of co-regulation, namely they perform distinct sub-functions that frequently get called together by the interferon response.
Another example of this concept is in the context of cell-type specification. In the course of its normal functioning, a given cell-type must express a number of distinct transcriptional programs. However, the relative levels of those transcription programs do not necessarily vary independently of one another in a given single-cell transcriptomic dataset. In datasets consisting of multiple cell-types all of the distinct sub-programs used by a cell-type will all get grouped together into an “identity program” as was the case with cNMF Kotliar2019-on. Even in datasets consisting of a single cell-type, there might not be sufficient natural variation in the relative activation levels of these sub-programs to disentangle them.

A recently developed technology that is very well suited to uncovering modular sub-structure in gene expression programs is Perturb-seq (Dixit et al., 2016; Adamson et al., 2016). Perturb-seq uses pooled CRISPR screening to test the effect of knocking out or inhibiting 100s of targets and reads out the effect on individual cells by single-cell RNA-Seq. Depending on the experimental conditions, Perturb-seq can be used to study the effect of perturbing individual genes on individual transcriptomes as well as combinations of genes.

In a pioneering proof of principal, Dixit et al. used CRISPR to knock out 24 transcription factors in bone-marrow derived dendritic cells before treating those cells with Lipopolysaccharide, a component of the outer-membrane of gram negative bacteria, that triggers a gene expression program of over 2000 genes (Dixit et al., 2016). While those 2000 genes would potentially be tightly linked in single-cell transcriptomic data in a natural bacterial infection context, the perturbation clearly separated them into 5 distinct sub-programs corresponding to anti-viral response; antigen presentation, cytoskeleton, and ribosomal proteins; mitochondrial function and biogenesis; an interferon gamma response; and an in-
flammatory TNF response. Each program had a distinct pattern of transcription factor perturbations that impacted it.

In a concurrent pioneering study, Adamson et al. sought to further resolve the unfolded protein response (UPR), a transcriptional response that occurs following the detection of an accumulation of unfolded proteins in the endoplasmic reticulum (ER) by dedicated sensor proteins (Adamson et al., 2016). They then inhibiting each of the 3 UPR sensors independently or in various combinations, prior to induction of UPR with several drugs and read out the various transcriptional consequences by scRNA-Seq. This decoupled the UPR into 3 sub-programs with overlapping but distinct transcriptional responses and able to function independently of one another. Thus, a high-level UPR GEP can be dissected into the 3 sub-programs that are typically coordinately induced to generate the overall program.

These analyses provide a proof of principal as to how we can dissect the broad gene expression programs that are observed in observational scRNA-Seq experiments—programs that reflect the genes evolution has shaped a cell to use in a given environment—into the specific sequence of sub-programs that give rise to the overall program. This allows us to take a step into the programs of nature, to understand the various sub-routines that are being called. As a practical matter, knowledge of the sub-programs could be useful for therapeutic interventions because inhibiting the regulator of a sub-program might have less side-effects than inhibiting the master regulator of the complete program.

Concluding thoughts

Over hundreds of millions of years, evolution shapes the genomes of living organisms in a greedy and random process, optimizing the propagation of genomes in their immediate context, without re-
gard for what might be in store for those genomes millions of years down the road. This has yielded a genetic blueprint that is fundamentally unlike what a human engineer would construct, full of dense connections and interdependencies, redundancies, and feedback loops that befuddle our intuitions. With the ever-expanding arsenal of high-throughput technologies for quantifying and perturbing the biological molecules in the cell, we are increasingly able to describe the physical phenomena that constitute life. Simultaneously, it is an ongoing challenge to organize the massive amounts of data into conceptual frameworks and algorithms that allow us to understand and predict the behavior of cells. The convergence of these two arms of biology is driving us toward the vision conveyed in Sidney Brenner’s quote at the opening of this thesis: “to explore the high level logical computations, the programs, the algorithms of development in molecular terms... to move between the molecular hardware and the logical software of how it’s all organized” (Judson, 1979).
Appendix A: Supplementary Note and Figures for Chapter 2

A.1 Supplementary Note - Analysis of between-sample variability

We sought to understand how variability between sample replicates and batches would impact the results of cNMF. We therefore considered how GEP usage varies across replicates in the primary datasets analyzed in this manuscript, as well as in a previously published scRNA-Seq dataset of human pancreatic islets with noted batch-effect (Baron et al., 2016).

First, we analyzed the aggregate GEP usage of cells in organoid replicates in the (Quadrato et al., 2017) data, and mouse replicates in the (Hrvatin et al., 2018a) visual cortex data. For this purpose, we
defined the aggregate GEP profile of a replicate as the sum of the GEP usage of all cells derived from that replicate. The visual cortex data showed relative uniformity of GEP usage across mouse replicates, with the only clear pattern being the expected association between depolarization-induced GEPs and mice treated with the stimulus (Figure A.16a - left). By contrast, there was significant variability between organoids in the (Quadrato et al., 2017) data that was primarily associated with the bioreactors in which the organoids were grown (Figure A.16b - left). This variability was discussed in the original manuscript and validated using immunohistochemistry, and thus represents true biological signal that we would hope for cNMF to discern.

We also considered whether any GEPs could be attributed to just one or a small number of replicates which could suggest that they are not reproducible within the experiment. We therefore looked at what percentage of the aggregate usage of a GEP derived from cells in each replicate. We found that each GEP contributed to cells from multiple independent replicates in both datasets (Figure A.16, right panels). No GEP derived more than 15% of its usage from a single replicate in the visual cortex data or more than 45% of its usage from a single replicate in the organoid data. Furthermore, each organoid GEP was the maximum contributing GEP for a cell in at least 6 distinct organoid replicates, and each visual cortex GEP was the maximum contributor for a cell in at least 10 distinct mouse replicates. This supports our conclusion that the inferred GEPs represent reproducible signals within the primary organoid and visual cortex datasets.

We also analyzed a human pancreatic islet scRNA-Seq dataset where variability between 4 donors resulted in more substantial batch-effects to see how that would impact the behavior of cNMF (Baron et al., 2016). Applied to this dataset of 10,939 cells, cNMF identified 16 GEPs that corresponded well
with the cell-type clusters described in the initial publication (Figure A.17). Our application of cNMF failed to identify GEPs corresponding to a few cell-types described in (Baron et al., 2016) (e.g. cells distinguished as delta and gamma cell-types were assigned the same GEP). However, many of the cell-types that were missed by cNMF were only distinguished through iterative sub-clustering in the initial publication, which we did not attempt.

Notably, we identified multiple GEPs for many cell-type clusters that corresponded to “donor of origin.” For example, we identified separate GEPs corresponding to acinar cells derived from donor 1 and 3, and acinar cells derived from donors 2 and 4, and similarly for alpha, ductal, and stellate cells. One potential contributor to the batch-effect could be that donors 1 and 3 were male and donors 2 and 4 were female. Consistent with this, we noticed that among the genes that were most differentially expressed between donors 1 and 3 compared to donors 2 and 4 in alpha, beta, and acinar cells were XIST on the X chromosome and RPSY1 on the Y chromosome (linear regression F-test P-values < 5×10^{-243} for XIST and P-values < 4×10^{-145} for RPSY1 for all 3 cell-types tested). But in general, the fact that cNMF is discerning multiple GEPs for the same cell-types suggests that technical sources of variation such as batch-effect can confound the identification of identity and activity GEPs.

In this instance, cNMF did not learn a single GEP for each donor (i.e. batch) but rather identified multiple hybrid identity-donor GEPs corresponding to individual cell-types derived from distinct sets of donors. This is likely due to the fact that the batch effect modulated the expression of different sets of genes in different cell-types, and therefore, no single shared “batch-effect” GEP could capture the impact on each cell-type. To avoid incorporating variation between batches into the inferred GEPs for datasets containing significant batch-effect, batch-effect correction can be performed prior to running
A.2 Supplementary Figures and Tables
Figure A.1: Robustness of matrix factorization methods

(a) Pairwise correlation coefficients for all components combined across 200 replicates for multiple simulations (rows) and multiple matrix factorization algorithms (columns). All 14 components for each replicate were assigned in a one-to-one fashion to a ground truth GEP in order from most correlated to least correlated. Components are ordered in each heatmap by the assigned ground truth GEP which is denoted by the outer color bars. The three simulations had different average signal strengths as parameterized by the mean log2 fold-change for differentially expressed genes in a GEP (Online Methods) and are ordered from least signal (top row) to most signal (bottom row).

(b) Components from the 200 replicates were assigned to a ground truth GEP, as in (a), and were correlated with the median of their assigned group. Then, for each factorization replicate, the 14 components were ranked in order from most to least correlated. The plots show the % of replicates where the kth most correlated component had a Pearson correlation >0.9 with the median of the components assigned to the same ground truth GEP. We plot the mean and standard deviation of this percentage across the 20 replicates as bars around the mean, and plot individual dots for simulation replicates that exceeded one standard deviation of the mean.
Figure A.2: Deconvolution accuracy of matrix factorization methods
(a) Pearson correlation between ground truth GEP means (rows) and GEPs inferred by different iterations of NMF (left), cNMF, or PCA (right) for an example simulation. All correlations are computed considering only the 2000 most over-dispersed genes and on vectors normalized by the computed sample standard deviation of each gene. Arrows annotate cases where GEPs were merged into a single component or a GEP was split into two components.
(b) Pearson correlation between inferred GEPs and true simulated GEPs for several matrix factorization methods across all 20 simulation replicates for each average signal level.
(c) Percentage of simulation replicates for which GEPs of each type had a pearson correlation of $R > 0.8$ with the true simulated GEP, as a function of the average signal level.
Figure A.3: Diagnostic plot for cNMF on an example simulated dataset
(a) Number of cNMF components (K) against solution stability (blue, left axis) measured by the euclidean distance silhouette score of the clustering, and Frobenius error of the consensus solution (red, right axis), with the K selection used in the manuscript indicated with a line.
(b) Plot of the Log10 proportion of variance explained per principal component, with the K selection used in the manuscript indicated with a dotted line.
(c) Clustergram showing the clustered NMF components for K = 14, combined across 200 replicates, before (left) and after filtering (right). In between is a histogram of the average distance of each component to its 60 nearest neighbors with a dashed line where we set the threshold for filtering outliers.
a

**Simulation overview**  

13 cell identity programs

**Ground Truth Cluster Assignment**

b

**GEP usage inference**  

(a) Comparison of GEP usage inference by cNMF, cLDA, and cICA for an example identity GEP (top row) and the activity GEP (bottom row). Each cell is represented as a point and its usage is represented by the marker color.

(b) Comparison of the results of ground truth cluster assignment and Louvain clustering represented on a t-SNE plot. (Left) Reproduction of Figure 1b which shows cells colored based on their true identity program, doublets marked with an X, and cells that express the activity program with a black border. (Middle) Cells are colored based on Louvain clustering. (Right) Cells with activity GEP usage of greater than 40% are assigned to an activity cluster, and all other cells were assigned to their identity cluster. This shows how an optimal hard clustering might behave in the context of mixed membership.

Figure A.4: GEP usage inference
Figure A.5: Accuracy of identifying genes in each GEP
(a) Receiver operating characteristics (ROCs) but showing false discovery rate (FDR) on the X-axis against sensitivity on the Y-axis for detection of genes in GEPs. Separate curves are shown for cNMF, cICA, cLDA, NMF, ICA, LDA, ground truth cluster assignment, Louvain clustering, or PCA. These show combined results for all 20 simulations with mean log2 fold-change = 1.00. Sensitivity is calculated considering genes with a differential expression fold-change of \( \geq 2 \) and FDR is calculated considering genes with no differential expression (fold change = 1).
(b) Same as (a) but only showing the results for the activity GEP and plotted separately for each of the 20 simulations at the mean differential expression log2 fold-change of 1.00.
Figure A.6: Comparison of run-times for different matrix factorization algorithms
Run-time in seconds for NMF, ICA, and LDA for a simulated scRNA-Seq dataset down-sampled to 6250, 12,500, 25,000, or 50,000 cells, run either using eight CPUs or four CPUs. Estimates are the average of three independent replicates with different seeds.
Figure A.7: cNMF demonstration on simulated dataset with 50% doublets
(a) tSNE plot for a simulated dataset with 50% doublets with marker color and edge color representing the simulated cell types.
(b) K selection diagnostic plot showing solution stability (measured by the silhouette distance) in blue and Frobenius error of the consensus solution in red.
(c) Pairwise Pearson correlation between ground truth GEP means (rows) and GEPs inferred by cNMF (columns) for the 50% doublet simulation dataset.
Figure A.8: Impact of variable cell-type proportions on GEP inference

(a) Visualization and inference of simulated scRNA-Seq data with the same parameters presented in Figure 2a but with 15 cell-types at variable frequencies corresponding to the proportions of cell-type clusters in the Hrvatin Et. Al visual cortex dataset.

(b) The same as (a) but using a differential expression location parameter of 2.0 instead of 1.0, corresponding to GEPs that are more divergent from each other. The leftmost panels show the t-distributed stochastic neighbor embedding (tSNE) plot for the simulations, representing cell types with distinct marker colors, doublets as gray Xs, and cells expressing the activity gene expression program (GEP) with a black edge. In the heatmaps to the right, we display the Pearson correlation of the ground truth GEPs with the programs inferred by cNMF, cICA, and Louvain clustering. All correlations are computed considering only the 2000 most over-dispersed genes and on vectors normalized by the computed sample standard deviation of each gene. The GEPs are labeled by the type of GEP (activity, I for identity only, and I + A for cell-types that express the activity GEP) and with the frequency of the cell-type in the data.
Figure A.9: Diagnostic plot for cNMF on the Quadrato et al., 2017 brain organoid dataset
(a) Number of cNMF components (K) against solution stability (blue, left axis) measured by the euclidean distance silhouette score of the clustering, and Frobenius error of the consensus solution (red, right axis), with the K selection used in the manuscript indicated with a line.
(b) Plot of the Log10 proportion of variance explained per principal component, with the K selection used in the manuscript indicated with a dotted line.
(c) Clustergram showing the clustered NMF components for K = 31, combined across 500 replicates, before (left) and after filtering (right). In between is a histogram of the average distance of each component to its 150 nearest neighbors with a dashed line where we set the threshold for filtering outliers.
**Figure A.10: Correlation between GEP spectra pairs and fraction of cells that use both programs**

Scatter plot of the Pearson correlation between pairs of GEPs (X axis) and the fraction of cells that co-use the GEP pair (Y axis). Co-usage is defined as the number of cells with usage $>0.1$ for both programs divided by the number of cells that use the less common of the programs with usage $>0.1$. Dots are colored by whether or not the GEP pair is made up of identity or any of the three activity programs.
Organoid cell-types and activities

• Astro-1  • FB-3
• Astro-2  • Dop-1
• Astro-3  • Dop-2
• Astro-4  • NE-1
• Astro-5  • NE-2
• Astro-6  • Stem-like
• Ret-1    • PP
• Ret-2    • Musc-T1
• Ret-3    • Musc-Im
• Ret-4    • Musc-T2
• Ret-5    • C6-1
• Ret-6    • C6-2
• FB-1     • C7
• FB-2     • C8

Figure A.11: t-SNE plots of identity and activity GEPs in the Quadrato et al., 2017 brain organoid dataset. t-SNE plots of cells colored by maximum identity GEP usage (left) or by absolute usage of each activity GEP (right).
Figure A.12: Comparison of cNMF usages with the cell-type clusters from Quadrato et al., 2017
Box and whisker plot of the usage of each GEP (column) in cells of the clusters from Quadrato et al., 2017 (rows). Boxes represent interquartile range, whiskers represent 5th and 95th percentiles.
Figure A.13: Diagnostic plot for cNMF on the Hrvatin et al., 2018 visual cortex dataset
(a) Number of cNMF components (K) against solution stability (blue, left axis) measured by the euclidean distance silhouette score of the clustering, and Frobenius error of the consensus solution (red, right axis), with the K selection used in the manuscript indicated with a line.
(b) Plot of the Log10 proportion of variance explained per principal component, with the K selection used in the manuscript indicated with a dotted line.
(c) Clustergram showing the clustered NMF components for K = 20, combined across 500 replicates, before (left) and after filtering (right). In between is a histogram of the average distance of each component to its 150 nearest neighbors with a dashed line where we set the threshold for filtering outliers.
Figure A.14: Comparison of cNMF usages with the cell-type clusters from Hrvatin et al., 2018
Box and whisker plot of the usage of each GEP (column) in cells of the clusters from Hrvatin et al., 2018 (rows). Boxes represent interquartile range, whiskers represent 5th and 95th percentiles.
Figure A.15: Comparison of GEPs identified in the Hrvatin et al., 2018 and Tasic et al., 2016 visual cortex datasets
(a) Heatmap showing the odds ratio for the intersection of top associated genes in each inferred GEP in the Hrvatin et al., 2018 and Tasic et al., 2016 datasets. Top associated genes were defined as those with an association score > 0.0015. Odds ratios above 100 were set to 100 for better visualization of pairs in the lower range. GEPs from the Tasic et al. dataset are labeled as ABA for Allen Brain Atlas.
(b) Proportion of cells of each cell type that express the superficial LRP with greater than 10% usage in the Tasic et al. dataset. Cells were assigned to a cell type based on their most used identity GEP.
Figure A.16: Characterization of GEP usage across biological replicates

(a) Heatmaps of aggregated GEP profiles for each biological replicate of the Hrvatin et al., 2018 data, derived by summing the GEP usage vectors for all cells from a replicate. We plot the GEP profiles normalized to sum to one for each replicate (row) in the left panel or to sum to one for each GEP (column) in the right panel. The left panel contrasts the composition of the replicates while the right panel visualizes which replicates contributed most to each GEP. We use yellow arrows to highlight the usage of the depolarization-induced GEPs (ERP, LRP-S, and LRP-D) in replicates of mice treated with the stimulus condition (replicate names ending in _1hr or _4hr denoting the 1 hr or 4 hr treatments, respectively). For all heatmaps, rows are ordered by hierarchical clustering of the row-normalized matrix using the cosine metric and average linkage method.

(b) The same as (a) but for the Quadrato et al., 2017 organoid data. We use yellow arrows to highlight the variability corresponding to the bioreactor from which the organoids derived (indicated in the beginning of the name).
Figure A.17: Comparison of cNMF usages with the published cell-type clusters from Baron et al., 2016

Box and whisker plot of the usage of each GEP (column) in cells of each cluster from Baron et al. (2016) (rows) stratified by the donor of origin of each cell (hue). Boxes represent interquartile range, whiskers represent 5th and 95th percentiles.
Appendix B: Supplementary Figures for Chapter 3

B.1 Supplementary Figures and Tables
Days Post-Infection (DPI) | Normal | Infection | Necropsy
--- | --- | --- | ---
Baseline | ✔ | ✔ | ✔
-30 | | |
-14 | | |
0 | | |
1 | | |
2 | | |
3 | | |
4 | | |
5 | | |
6 | | |
7-9 | | |
N=3

Uninfected controls
Scheduled D3 necropsy
Scheduled D4 necropsy
Scheduled D5 necropsy
Scheduled D6 necropsy
Terminal - No manip.
Terminal - Routine manip.
N=6
Figure B.1: Blood sampling overview and clinical time course per animal replicate

(a) Overview of study cohorts and blood draw timelines. Animals were grouped into cohorts with pre-scheduled necropsy times (at baseline, or day post infection 3, 4, 5, 6 - N=3 each), or allowed to progress until reaching terminal disease, defined by a clinical score exceeding 10 (terminal groups). There were two terminal groups of 6 animals, one with multiple intermediate blood draws (Routine manipulations) and one without intermediate blood draws (No manipulations). Dots indicate scheduled blood draws for each cohort with red denoting a normal (non-necropsy) draw, and gray indicating a draw that coincided with euthanasia and necropsy. Necropsy and baseline normal draws were used for Seq-well and CyTOF while intermediate post-infection draws were available only for CyTOF.

(b) Each panel represents the time course of log viral load (left axis, red) and clinical score (right axis, blue) for a specific animal replicate. Panels are organized into rows based on the cohorts.
Figure B.2: Cell type markers for Seq-well and CyTOF clusters
(a) Dot-plot showing expression levels of cell type marker genes (columns) for cell type clusters (rows) based on the in vivo Seq-well data. Marker size represents the percentage of cells in that group in which the gene was detected, and color intensity denotes the average expression level of that gene in the group, in units of log2 transcripts per 10,000.
(b) Heatmap of average intensity values of cell type marker genes (columns), for cell type clusters (rows), based on the CyTOF data.
Figure B.3: UMAP embeddings of in vivo data colored by Seq-well samples and CyTOF batches
(a) Uniform Manifold Approximation and Projection (UMAP) embedding of Seq-well data colored by the sample source (animal replicate plus day post infection). Samples with more than 500 cells were first down-sampled to 500 cells so they wouldn’t dominate the plot.
(b) UMAP embedding of the CyTOF data with each cell colored by the multiplex batch in which it was pooled, and analyzed by CyTOF.
Figure B.4: Estimates of cell type abundance over the time course
(a) Estimates of the abundance of each cell type (rows) for each animal replicate (individual markers) in units of 1000 cells per uL or whole blood. The average value at a day post infection is shown as a black line.
(b) Scatter plot of the percentage of cells of each cell type in a sample, inferred from Seq-well (x-axis), or CyTOF (y-axis), for several cell types (panels). Each dot represents a sample, and is colored by the day post infection on which the sample was taken.
Figure B.5: Estimates of cell type replication rate over the time course
(a) Estimates of the percentage of Ki67 positive cells (CyTOF intensity > 1.8) of each cell type (rows) for each animal replicate (markers). The average value at a day post infection is shown as a black line.
(b) Scatter plots of the percentage of Ki67 positive cells in a sample inferred from Seq-well (x-axis) or CyTOF (y-axis) for several cell types (panels). Each dot represents a sample and is colored by the day post infection on which the sample was taken. Ki67 values for Seq-well are first smoothed using MAGIC with t=1 and cells with smoothed expression above 0.1 were called Ki67 positive.
EBOV detection accuracy

Figure B.6: Receiver operating characteristic for identifying Ebola virus infected cells

Estimates of sensitivity to detect an infected cell at various false positive rate thresholds in vivo (left) and ex vivo (right). Curves are estimated separately for a hypothetical viral load of 0.1% (blue line) and 1% (orange line). A theoretical sensitivity is computed separately for a random sample of 2000 cells from the relevant condition (DPI 3-8 in vivo, live virus treatment at 4H or 24H for ex vivo), and we plot the average sensitivity across cells.
Figure B.7: Quantification of cytokine expression and enrichment of response signatures
(a) Average expression values (MAGIC smoothed log2 transcripts per 10,000) of literature annotated cytokines (columns) across cell-types and stages of EVD (rows). Values are plotted as a ratio relative to the maximum value considering all cell types and stages. Genes that had a statistically significant difference from baseline in a given cell type and stage are indicated with a blue star.
(b) Heatmap of scores reflecting the difference in median log fold-changes for genes in a specific gene set (rows) compared to genes not in the geneset. The log fold changes were defined from differential expression profiles of cell-types at each EVD phase (columns) relative to baseline. Scores are the normal approximation Z-score to the Mann Whitney U statistic. Five genesets were tested - three from the Hallmark database (IFN ALPHA, IFN GAMMA, and TNF ALPHA VIA NFKB) (Liberzon et al., 2015a) and 2 constructed from the hallmark sets, as genes in "IFN ALPHA" but not "IFN GAMMA" ("IFN ALPHA - GAMMA"), and vice versa ("IFN GAMMA - ALPHA")
Figure B.8: Extended characterization of gene expression signals associated with EBOV infection status in monocytes
(a) MX1 expression (log2 transcripts per 10,000) in monocytes at baseline, and uninfected bystanders or infected cells in late infection. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles.
(b) Scatter plot of MAGIC smoothed expression values of CD14 and CD16 for monocytes in baseline, early, mid, and late disease stages. Cells are colored by their cell-cycle score. Boxes indicate the CD14+, CD16+, CD16int, DN, and DP subsets described in the text and the numbers denote the percentage of cells falling into each subset.
(c) Percentage of replicating monocytes in each EVD phase stratified by monocyte subset. Replicating cells are defined as those with a cell cycle score in the Seq-well data of greater than 3. Height of the bar denotes the mean and error bars denote 95% bootstrap confidence intervals for the mean.
(d) Ki67 marker intensity in the CyTOF data for monocytes in each EVD phase stratified by monocyte subset. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles.
a) Volcano plot of differentially expressed genes between double positive and double negative monocyte subsets from day post infection 5-8. Genes are colored by membership in cell cycle, macrophage up-regulated (Mac. Up), and macrophage down-regulated (Mac. Down) genesets, or by whether they are an Ebola virus gene.

(b) Differentiation scores for monocytes in the late disease phase for each subset. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles.

(c) Percentage of infected monocytes in each subset in late disease, stratified by low or high differentiation score. Height of the bar denotes the mean, and error bars denote 95% bootstrap confidence intervals for the mean.

(d) Differentiation scores for monocytes in the late EVD phase, for each subset, stratified by monocyte infection status. Boxes denote the median and interquartile range, and whiskers denote the 2.5th and 97.5th percentiles.
Figure B.10: Overview of the ex vivo Ebola virus infection dataset
(a) Schematic of ex vivo study design.
(b-g) Uniform Manifold Approximation and Projection (UMAP) embedding of Seq-well data colored by annotated cluster assignment, treatment condition, viral load, animal donor, MX1 gene expression (log2 transcripts per 10,000), and interferon stimulated gene (ISG) score.
(h) Distributions of ISG scores across monocytes from each treatment condition, stratified by animal donor replicate. The central white marker denotes the median and the black bar denotes the interquartile range.
In vivo

Marker: P=5.49e-52
MX1: P=6.54e-31
MX2: P=2.45e-35

Positive correlation with viral load

Marker: P=2.54e-27
HSPA5: P=7.48e-24
IARS: P=1.11e-07

Inverse correlation with viral load

Marker: P=2.35e-31
HSPA5: P=2.05e-27
IARS: P=2.11e-07
Figure B.11: Ebola virus infection dynamics in the ex vivo dataset

(a) Estimated percentage of infected cells of each cell type in the ex vivo dataset. The dashed line denotes the 1% false positive rate threshold used for calling infected cells.

(b) Percentage of Ebola virus (EBOV) positive cells in monocytes from each ex vivo treatment condition, stratified by animal donor replicate. Height of the bar denotes the mean, and error bars denote 95% bootstrap confidence intervals for the mean.

(c) Violin plot showing distributions of viral loads across monocytes from different treatment conditions. The central white marker denotes the median and the black bar represents the interquartile range.

(d) Estimated percentage of EBOV transcripts derived from the EBOV genome or each EBOV gene, stratified by treatment conditions. Prior to averaging, the counts of EBOV genes for each cell was normalized to sum to one, so each cell contributes uniformly to the estimate, regardless of its total number of EBOV reads. Height of the bar denotes the mean, and error bars denote 95% bootstrap confidence intervals for the mean.

(e-f) Scatter plot of total transcripts (unique molecular identifiers) detected in a cell (X-axis) against percentage of cellular transcripts derived from virus (Y-axis) for cells with one or more viral reads ex vivo (e) or in vivo (f). Cells called as infected are colored in red and otherwise colored in blue. The X and Y axes are plotted on a log10 scale.

(g-h) Association between gene expression and viral load for selected host genes in monocytes, 24 hours following treatment with live virus, ex vivo. In the right sub-plots, infected cells are ordered by viral load and the average expression of each gene is computed within a sliding window of 100 cells. The X-axis is plotted on a log10 scale. In the left sub-plots, distributions of gene expression in uninfected bystander cells are shown as a boxplot with boxes denoting the interquartile range, and whiskers denoting the 2.5th and 97.5th percentiles. Curves are box-plots shown separately for the 2 donor animals. P-values for the Spearman correlation between viral load and gene expression for each animal separately are listed in the legend.
References


173


178


