Title
Single-cell mapping of progressive fetal-to-adult transition in human naïve T cells

Author
Bunis, Daniel

Publication Date
2020

Peer reviewed|Thesis/dissertation

Single-cell mapping of progressive fetal-to-adult transition in human naïve T cells

by
Daniel Bunis

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Mark Anderson
657D61B3C699406...

Mark Anderson
Chair

DocuSigned by:
Alexander Marson
[illegible]4A8...

Alexander Marson

DocuSigned by:
Trevor D. Burt
[illegible]44D...

Trevor D. Burt

DocuSigned by:
Marina Sirota
925B61AB9C41499...

Marina Sirota

Committee Members

To my parents, Gayle and Larry, for their unconditional love and support even when I wanted to move all the way to California.

To my husband, Phil, for coming with me, and being my rock, on this journey.

Acknowledgements

Looking back at the windy path I took during graduate school, I am reminded of all of the growth I see in myself, both in my approach to science and also to life in general, compared to the naïve student I was when I first arrived at UCSF. And I have to say that I am incredibly grateful to have landed at UCSF for these past 5 years. For the self-confidence, in so many areas, that I have gained from my many interactions here, both with my direct supervisors and lab mates but also with collaborators far and wide. For the Pride in myself fostered not only by UCSF, but also the vibrant city of San Francisco. For the classmates who helped me through the harder times – no PhD is a walk in the park. For the scientific advisors who helped ensure that the path I took was one that would lead to success. There are many individuals in particular whom I need to thank.

Firstly, to Trevor, my original mentor who encouraged me to think freely, to trust my instincts, and who gave me a vast appreciation for the complexity of the system yet also for my own ability to make sense of its individual pieces. Looking back, I realize that your consistent confidence in my ability to pick up, and make good use of, new skills was the best encouragement that I could have gotten early in my graduate career. The near absolute independence to drive the project forward which I felt I'd had, and to learn from failures and press on, was incredibly beneficial for the self-confidence as a scientist that I have today. Another thing I will be eternally grateful for is that your attention to detail when editing writing has helped make me a better writer and reviewer myself.

To Marina, my original thesis committee chair turned co-mentor, the structure and direction that you gave in the last year and a half after I joined your lab helped foster a flurry of additional projects to come to fruition that would never have happened otherwise. I am amazed at my productivity during the final stretch of my PhD, and a huge contributor to that productivity were our weekly meetings. Your many nudges and reassurances that our goals were not crazy,

as well as your insightful ideas when I was momentarily stumped, were inherent to that productivity and have helped shape me into the confident, forward-looking scientist that I am today. On top of that, your openness to discussion of many tough topics helped me grow both as a scientist and as a person. While working in your lab, I felt truly empowered to accomplish all of my goals. Moving forward, I take that feeling with me, largely, because of the key bits of advice and all of the encouragement that I received from you and from others within the lab.

To Mark Anderson, my eventual thesis committee chair, and Alex Marson, the final member of my thesis committee, thank you both for helping to guide my journey and ensuring that the paths taken were ones that could lead to success. You were both part of my qualifications exam committee as well, and the direct, intuitive, suggestions that I received from each of you, along the entire way, were truly instrumental in shaping my project into a what it is today. Among others certainly, special thank you for the strong suggestion that I chase down more than just contributions from a single transcription factor, Bcl11a.

To my lab families, Melissa, Elze, Joanna, Ventura, and Norm, as well as, eventually, the entire Sirota lab, you all helped make my time at UCSF so much more fun and rewarding. From random discussions while working side-by-side in the tissue culture room, to practice talks, lab meetings, and our one-on-one discussions, I was constantly learning from everyone in the Burt lab. And I know that will continue in the future as we will be life-long friends. Same goes for my Sirota lab family. The settings of conversations may have been different in our computational lab space, but the outcomes were similar. I learned so much from all of you and built many relationships that I hope will last forever. I truly would not be at the end of this path I took for my PhD, right now, if it were not for all of your support throughout these years.

To all of my collaborators and co-authors, both at UCSF and elsewhere, science does not operate in a vacuum and I’m so proud of what we were able to conduct together, and I’m so grateful for the efforts and support from so many people along the way.

Special thanks, as well, to my classmates, especially Dror, Rebecca, Jessica, Casey, Ariane, Carlos, Simone, Emily, Nate, and Natanya for emotional support, camaraderie, venting, and motivation. Our “wine downs” and hang outs, both in person and over zoom, were quite necessary, and also formed memories that I will cherish forever. In particular, Dror, our countless coffee trips were integral to both my mental health and my experimental planning. Rebecca, our many study and coding sessions helped make learning so many new things easier and I will forever be grateful for all the code we shared and the troubleshooting practice that we gave each other.

To my family as well, I would certainly not be at this point today without the love and support from all of you. To my parents specifically who only ever want the best for me, who allowed me to spread my wings and fly to the West Coast, and who instilled from a young age that trying your best matters more than doing the best; I would not have been able to grow in so many different directions during graduate school had you not taught me, long ago, that even if initial attempts at something may be naïve or ineffective, pride can still be taken in making a concerted effort. And much of my recent growth has come from making continuous, concerted efforts on fronts that I knew very little about before graduate school. And finally, to my husband, Philip, thank you all the food, for all fun and the breaks from science that I got to take with you, and for all of your encouragement. But most of all, Phil, thank you for your patience. I truly cannot imagine what this journey could have been like without you in my life.

Contributions

The work presented in this dissertation was performed under the direct supervision and guidance of Dr. Trevor D. Burt, M.D., and Dr. Marina Sirota, Ph.D. Additional guidance and insight were provided by thesis committee members Dr. Mark Anderson, M.D., Ph.D, and Dr. Alexander Marson, M.D., Ph.D.

Chapter 1 parts 1-2 of this work are largely adapted from a published review article:
Silvia Pineda*, **Daniel G. Bunis***, Idit Kosti, Marina Sirota. “Data Integration for Immunology,” Annual Review of Biomedical Data Science, 2020.
Conceptualization: S.P., M.S. Writing – original draft preparation: S.P., D.G.B., I.K; S.P. lead the data integration section, D.G.B. lead the molecular technologies section. Writing – review and editing: all authors. Supervision: M.S. *S.P. and D.G.B. contributed equally to this manuscript.

Chapter 1 parts 3-4, Chapter 2, Chapter 4, and the Materials & Methods of this work are largely adapted from a manuscript in-press at *Cell Reports*:
Daniel G. Bunis, Yelena Bronevetsky, Elisabeth Krow-Lucal, Nirav R. Bhakta, Charles C. Kim, Srilaxmi Nerella, Norman Jones, Ventura F. Mendoza, Yvonne J. Bryson, James E. Gern, Rachel L. Rutishauser, Chun Jimmie Ye, Marina Sirota, Joseph M. McCune, Trevor D. Burt. "Single-cell mapping of progressive fetal-to-adult transition in human naive T cells," in-press with Cell Reports, 2020.
D.B., Y.Bro., E.K.L., M.S, T.D.B., and J.M.M. designed the study and analyzed the data. D.B., Y.Bro, E.K.L, V.M, and N.J. performed the experiments. M.S. advised on the computational analysis, whereas T.D.B. and J.M.M. advised on experiments. N.R.B. and S.N. provided critical bioinformatic assistance. C.C.K. assisted with microarray experimental workflow and data analysis. J.Y. contributed to Demuxlet analysis. R.L.R. contributed to data analysis of URECA

cohort samples. J.G. and Y.Bry. provided technical counseling and samples. D.B., Y.Bro., E.K.L., M.S., T.D.B., and J.M.M. wrote the paper. All authors reviewed and commented on the manuscript.

Chapter 3 of this work is largely adapted from a manuscript under review at *Bioinformatics*:

Daniel G. Bunis, Jared Andrews, Gabriela K. Fragiadakis, Trevor D. Burt, and Marina Sirota. "dittoSeq: Universal User-Friendly Single-Cell and Bulk RNA Sequencing Visualization Toolkit," manuscript under review, 2020.

Conceptualization: D.G.B. Software: D.G.B., J.A. Supervision: G.K.F., T.D.B., M.S. Funding acquisition: T.D.B., M.S. Writing – original draft preparation: D.G.B. Writing – review and editing: all authors.

Single-cell mapping of progressive fetal-to-adult transition in human naïve T cells

Daniel Bunis

Abstract

Whereas the human fetal immune system is poised to generate immune tolerance and suppress inflammation in utero, an adult-like immune system emerges to orchestrate anti-pathogen immune responses in post-natal life. Compared to their adult counterparts, fetal naïve T cells more readily differentiate into tolerance-prone regulatory T cells (T_{regs}) upon stimulation, and fetal monocytes exhibit distinct responses to cytokine stimulation and impaired antigen presentation capacity. It is believed that tolerogenic responses across various immune cells are adaptations that allow the fetal immune system to suppress responses to non-inherited maternal antigens that might otherwise lead to pregnancy complications. After birth, the balance between tolerance versus protection must shift. It has been posited that cells of the adult immune system arise as a discrete ontological “layer” of hematopoietic stem-progenitor cells (HSPCs) and their progeny; evidence supporting this model in humans has, however, been inconclusive. Although it has been shown that various immune cell transitions during early gestation may be modeled by the layering hypothesis, it is unknown whether the transition from fetal-to-adult follows a similar mechanism. Here, we combine bulk and single-cell transcriptional profiling of lymphoid, myeloid, and HSPCs from fetal, perinatal, and adult developmental stages to demonstrate that the fetal-to-adult transition occurs progressively along a continuum of maturity—with a substantial degree of interindividual variation at the time of birth—rather than via a transition between discrete waves. We find that (1) newborn immune cells are relatively homogenous, each with an intermediate transitional phenotype, rather than having either a fetal or adult

phenotype, (2) UCB HSPCs are not fully adult-transitioned, (3) the progression of transition at birth shows a high degree of inter-individual variability, and (4) pathways known to affect T cell polarization are among those expressed more highly in newborn than adult naïve CD4 T cells. These findings have important implications in the design of strategies for prophylaxis against infection in the newborn, and for the use of umbilical cord blood (UCB) in the setting of transplantation. I also present dittoSeq, a universal bulk and single-cell visualization toolkit that powered the analysis of these data. dittoSeq visualizations are color blindness-friendly by default, robustly documented to power ease-of-use by both novice and experienced coders, and allow highly customizable generation of both daily-use and publication-quality figures.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

Material for this chapter was modified from a published review article & a separate manuscript currently in-press.

Parts 1-2 are adapted from:

Silvia Pineda*, **Daniel G. Bunis***, Idit Kosti, Marina Sirota. “Data Integration for Immunology,” Annual Review of Biomedical Data Science, 2020.

*These authors contributed equally to this article

Parts 3-4 are adapted from:

Daniel G. Bunis, Yelena Bronevetsky, Elisabeth Krow-Lucal, Nirav R. Bhakta, Charles C. Kim, Srilaxmi Nerella, Norman Jones, Ventura F. Mendoza, Yvonne J. Bryson, James E. Gern, Rachel L. Rutishauser, Chun Jimmie Ye, Marina Sirota, Joseph M. McCune, Trevor D. Burt. "Single-cell mapping of progressive fetal-to-adult transition in human naive T cells," in-press with Cell Reports, 2020.

Part 1 – Introduction to the Immune System

The immune system is probably the most complex system in the body. Over the last several decades, we have learned that the immune system is not just responsible for protection against diseases and host defense but also plays a role in tissue maintenance and repair. As such, the system is spread throughout not only the blood and lymphatic systems but also most tissues (**Fig. 1.1**). We know that the immune system is made up of many distinct cell types borne from hematopoietic stem cells, including innate immune cells (e.g., granulocytes, macrophages, dendritic cells, etc.) and adaptive immune cells (e.g., B cells, T cells, plasma cells, etc.), each with its own distribution pattern throughout the body, as well as antigens, cytokines, and chemokines that these cells use to communicate (**Fig. 1.1** and **Fig. 1.2**). Each of the distinct cell types responds to chemokines and cytokines in a different way, and many non-immune cells display robust responses to cytokines or can release certain cytokines themselves in response to infection or tissue damage. Furthermore, there are thousands of genes involved in immune regulation within individual cells. With all of these components, the individual study of each cell, each stimulus, each cytokine, or each gene cannot always describe the interconnecting pathways that control immune system responses. Advances in technology and methodology have transformed the landscape of immunology research (Fernandez et al., 2019; Thorsson et al., 2018) and have made it feasible to study systems immunology as a whole, but a clear understanding of much of this complexity remains elusive.

The main effectors of the immune system are specialized cells that communicate with each other and tissues via cell-surface receptors, cytokines, and chemokines (Mark M. Davis et al., 2017). The many cells of the immune system play distinct roles. Some play a continuous role, such as breaking down potential antigens and presenting those to other immune cells for evaluation. Others lie dormant until being presented with a specific antigen. Upon stimulation, certain immune cells release cytotoxic agents that directly kill infected cells, whereas others release cytokines and chemokines that modify the responses of other immune cells.

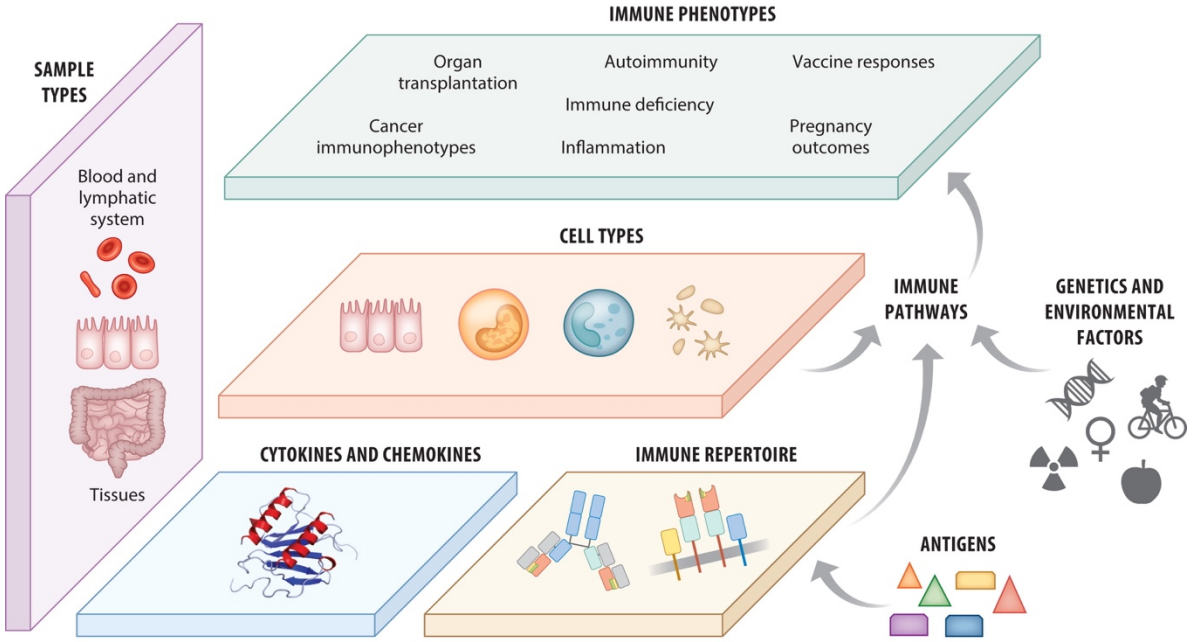
On top of this, it is known that this complex system is heavily influenced by genetic variation and the environment. Smoking, as well as demographic factors such as age and sex, together with allelic variability at genetic loci have been shown to affect human immune phenotypes (Liston & Goris, 2018; Patin et al., 2018).

There are several ways in which the immune system can fail. In contrast to conventional immune-mediated elimination of foreign antigens, immune tolerance is a state of unresponsiveness of the immune system to substances or tissues that have the capacity to elicit an immune response. While central tolerance—whereby adaptive immune cells specific for self-antigens are negatively selected during their development and are either eliminated altogether or pushed towards anergic or anti-inflammatory fates—is the main way the immune system learns to discriminate self from non-self, peripheral tolerance—whereby presence, or absence, of certain signals, during a post-maturation activation, push adaptive immune cells towards anergic or anti-inflammatory fates—is key to preventing over-reactivity of the immune system to various environmental entities. Deficits in central or peripheral tolerance can lead to autoimmune disease; on the flip side, an overly tolerant immune system can result in unresponsiveness to infection.

Many recent findings have been powered by technological advances that have enabled deeper profiling of immune cells and tissues. Advances in technology and methodology improve our ability to associate differences in gene expression, genetics, epigenetics, and microbiome and/or immune repertoire composition with various immune phenotypes and clinical outcomes, and are thus transforming the landscape of immunology research. By combining measurements taken from multiple technologies, and from either the resolution of whole-tissues versus individual cells, the immune system can now be examined on a more holistic level, and new layers of complex interactions can be elucidated (Thorsson et al., 2018).

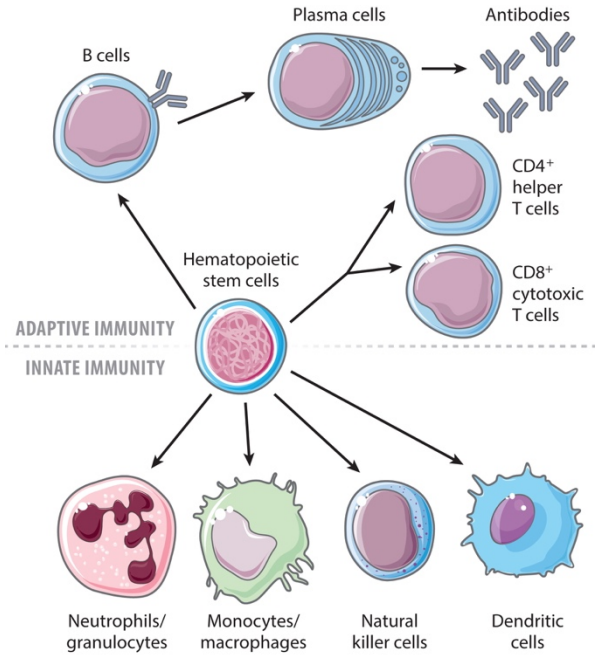
In the next section, Part 2, I discuss some of the most important data types and computational approaches used in immunology (**Fig. 1.3**) with a discussion of applications

where such approaches have been used. Then, in Parts 3 & 4, I discuss a particular application of a few of these technologies to study early human immune development.



Pineda S, et al. 2020.
Annu. Rev. Biomed. Data Sci. 3:113–36

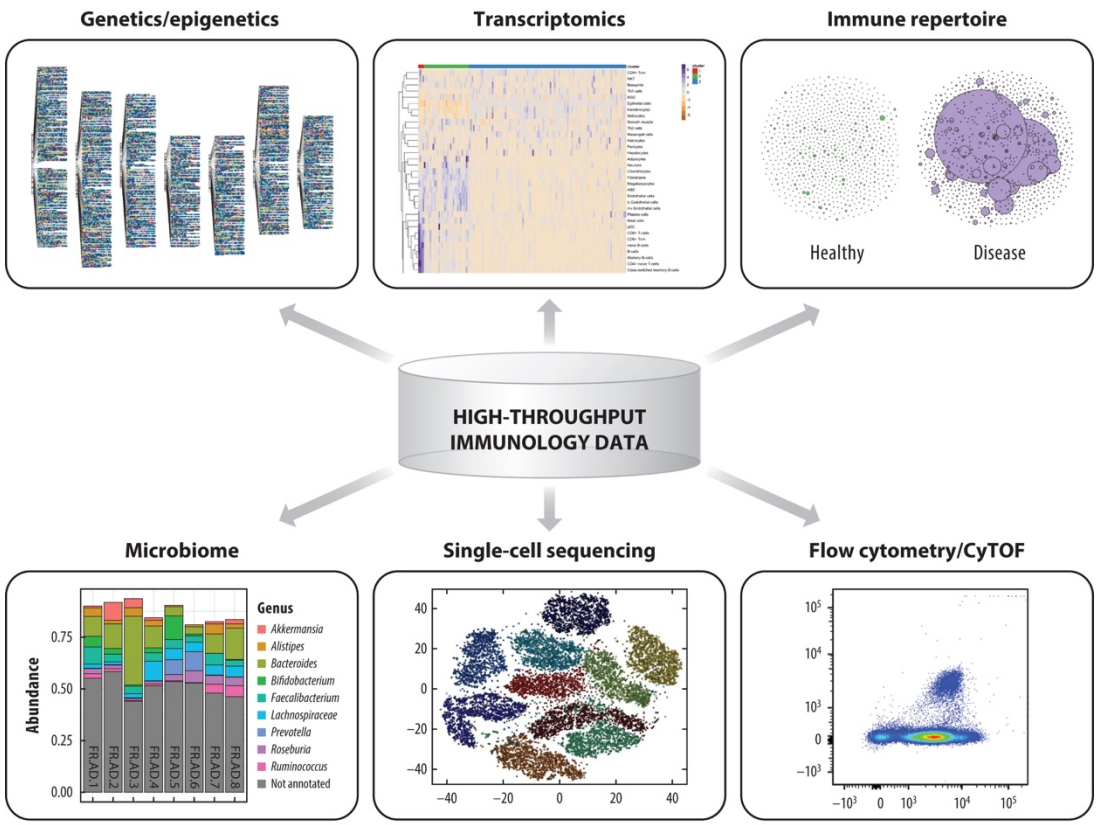
Figure 1.1: A systems immunology overview
The immune system is made up of many cell types spread throughout the body that respond to antigens or communicate through cytokines and chemokines in distinct ways. An individual's antibodies and B cell and T cell immune repertoire also play a major role in shaping responses of the immune system. These factors, together with genetic and environmental factors, affect immune pathways and thereby the development and presentation of immune-related diseases and phenotypes.



Pineda S, et al. 2020.
Annu. Rev. Biomed. Data Sci. 3:113–36

Figure 1.2: Cells of the immune system

Hematopoietic stem cells give rise to all cells of the immune system. Innate immunity (bottom) refers to nonspecific defense mechanisms that come into play soon after a pathogen or other stimulus enters the body. Major cell types responsible for the innate response include neutrophils, granulocytes, monocytes, macrophages, natural killer cells, and dendritic cells. Adaptive immunity (top) refers to antigen-specific immune responses. Once an antigen has been processed and recognized by B or T cells of the adaptive immune system, these cells proliferate and differentiate into an army of specialized effector cells specifically designed for protection from distinct types of antigens. Adaptive immunity also has a memory component that makes future responses against the returning antigens more efficient by reducing the time it takes for the army of antigen-specific adaptive immune cells to respond and regrow. Images created using Servier Medical Art (CC BY 3.0).



Pineda S, et al. 2020.
Annu. Rev. Biomed. Data Sci. 3:113–36

Figure 1.3: An overview of the diverse high-throughput data types publicly available for studying immunology

Genetics/epigenetics, transcriptomics, immune repertoire, microbiome, single-cell sequencing, and flow cytometry/cytometry by time-of-flight (CyTOF). Genetics/epigenetics image adapted with permission from the NHGRI-EBI (National Human Genome Research Institute–European Bioinformatics Institute) GWAS (Genome-Wide Association Study) Catalog diagram (https://www.ebi.ac.uk/gwas/docs/diagram-downloads); all other images were generated from publicly available data.

Part 2 – Leveraging High-Throughput Data to Profile the Immune System, From Bulk Tissue to Single-Cell

Genetics

Genome-wide association studies (GWAS) utilize genotyping information from hundreds to hundreds of thousands of individuals in order to identify specific genetic mutations that are significantly associated with disease states. GWAS studies have elucidated genetic variation within both coding and noncoding regions that show associations with many immune-related disorders, including the revelation of genetic susceptibility loci for multiple sclerosis (Baranzini et al., 2013), lupus (Martin et al., 2013), rheumatoid arthritis (Stahl et al., 2010), and many other immune-related diseases.

Exactly how such variation might relate to particular disorders is not always known. One method of examining these connections is to compare mutations versus wild type in an animal model. For example, variants in noncoding enhancer regions near the CD25 gene (*IL2RA*), a gene encoding a subunit of the high-affinity receptor for T cell proliferation cytokines, have been linked to alterations of CD25 upregulation kinetics after activation and to a subsequent effect of skewing $CD4^+$ helper T cell differentiation toward the proinflammatory T helper 17 fate but away from the induced regulatory T cell (T_{reg}) fate (Simeonov et al., 2017). Another widely used method for linking particular genetic variants to associated mechanisms of effect, and one which directly involves integration of multiple data types, is through expression quantitative trait loci (eQTL) analysis. We describe this technique in the next section, alongside the transcriptomic profiling approaches upon which eQTL studies often rely.

One particular region in which GWAS studies repeatedly identify linkages to immune-related diseases is the human leukocyte antigen (HLA) region (Kennedy et al., 2017; Trowsdale & Knight, 2013). This region contains the genes that encode for major histocompatibility complexes (MHCs) and is thought to be the most highly polymorphic region of the human

genome (Horton et al., 2008). MHCs are major effector molecules of the immune system that are utilized by antigen-presenting cells (APCs) to present peptide antigens to T cells. Polymorphisms in this region can affect the set of peptides that an individual's APCs can present to T cells (Unanue et al., 2016). Thus, this particular region has an outsized impact on responses by the immune system and often deserves particular attention within systems immunology data integration approaches.

Meta-analysis—the re-analysis of combined data from multiple, original studies—of multiple GWAS studies can increase the power for identifying genetic variation associated with individual diseases (Baranzini et al., 2013; Martin et al., 2013; Stahl et al., 2010) and allows additional insights to be drawn about how genetic risk varies between diseases. We have previously carried out a meta-analysis of GWAS studies by performing pairwise comparisons of genetic relationships of six autoimmune diseases and five nonautoimmune diseases, which revealed broad clusters of single-nucleotide polymorphisms (SNPs) that increase susceptibility to one disease while providing protection from another (Sirota et al., 2009). For powering such meta-analyses, in addition to integrative approaches that include other data types described below, there are multiple databases containing genomic data from many studies (**Table 1.1**). The database of Genotypes and Phenotypes (Mailman et al., 2007) (**https://www.ncbi.nlm.nih.gov/gap**) contains genomic data from 1,203 studies; the GWAS catalog (Buniello et al., 2019) (**https://www.ebi.ac.uk/gwas/**) contains data from 7,796 publications; and the UK Biobank (Bycroft et al., 2018) (**https://www.ukbiobank.ac.uk/**) includes around 500,000 genotyped participants along with deidentified health information covering a range of serious and life-threatening illnesses.

Table 1.1: Data Repositories for High-throughput Immunology Data
[a]Sample and study numbers were obtained, when available, from repository websites in December 2019.
[b]These metrics include both single-cell and bulk transcriptomics datasets.
[c]These metrics come from searching the Sequence Read Archive for "TCR OR BCR OR repertoire."
[d]These metrics come from searching the Sequence Read Archive for "microbiome."
Abbreviations: CyTOF, cytometry by time-of-flight; NA, not available.

Repository Name	Data Types	Size[a]	
		Studies	Samples
dbGaP: Database of Genotypes and Phenotypes, https://www.ncbi.nlm.nih.gov/gap	Genetics Transcriptomics Epigenetics	NA	2,843,764 159,097 20,784
GWAS catalog, https://www.ebi.ac.uk/gwas/	Genetics	4,298 publications	161525 associations
UK Biobank, https://www.ukbiobank.ac.uk/	Genetics	NA	500,000
GEO: Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/	Transcriptomics Single-cell Transcriptomics	4,348[b]	3,323,951[b]
Immunological Genome Project (ImmGen), ImmGen.org	Transcriptomics	20	NA
ENCODE: Encyclopedia of DNA Elements, https://www.encodeproject.org/	Genetics: Genotyping DNA sequencing Transcriptomics Epigenetics: DNA binding DNA accessibility DNA methylation RNA binding 3D chromatin structure Other: Replication timing Proteomics RNA structure	NA	 168 29 3797 9562 1200 874 744 135 182 14 6
SRA: Short Reads Archive, https://www.ncbi.nlm.nih.gov/sra	Immune Repertoire Microbiome	1,141[c] 6,621[d]	8,268[c] 262,613[d]
ImmPort, https://www.immport.org/home	Immune Repertoire Flow Cytometry CyTOF	NA 140 27	55,607 (total)
Pan Immune Repertoire Database, https://db.cngb.org/pird/	Immune Repertoire	12	3600
Human Microbiome Project & Integrative Human Microbiome Project, https://hmpdacc.org/	Microbiome	18	31,596
JingleBells, http://jinglebells.bgu.ac.il/	Single-cell Transcriptomics	120	500,175 cells

Repository Name	Data Types	Size[a]	
		Studies	Samples
PanglaoDB, https://panglaodb.se/	Single-cell Transcriptomics		1368 samples 5,586,348 cells
Single Cell Expression Atlas, https://www.ebi.ac.uk/gxa/sc/home	Single-cell Transcriptomics	132	1,028,590 cells
FlowRepository, https://flowrepository.org/	Flow Cytometry	37	NA
Cytobank, https://www.cytobank.org/data-sets.html	CyTOF	19	NA

Transcriptomics

Whole-genome transcriptional profiling allows unbiased comparison of gene expression across samples, tissues, and cells and has become standard in the last decade. Massive amounts of data have been generated via two distinct approaches: whole-transcriptome shotgun sequencing using next-generation sequencing platforms, typically referred to as RNA sequencing (RNA-seq), and fluorescence-based quantification of transcript abundance via imaging after hybridization to specialized chips (typically referred to as microarrays) that contain thousands of probes, where each probe represents an individual gene or transcript. Transcriptional profiling by these methods has been utilized to compare expression levels in healthy versus disease settings, in steady state versus activation or other perturbation settings, and across cell types in a range of settings; these have been integral parts of many major advancements in immunology. For example, our understanding of the process of T cell development was extended immensely through the characterization of the expression differences between successive cell types throughout the thymic T cell differentiation process (Mingueneau et al., 2013). In general, transcriptomic profiling studies have identified genes, and pathways of multiple genes, that play specific roles in various immune cells.

In recent years, useful methods for querying RNA-seq data outside of differential gene expression have been developed as well. Cell deconvolution and enrichment methods like CIBERSORT (Newman et al., 2015) and xCell (Aran et al., 2017) can be used to infer cell types,

and cell type percentages, that likely exist within bulk tissue samples. These methods rely on construction of mixture models of previously analyzed RNA-seq data from pure cell types, and have been used extensively, especially in the characterization of immune cell tumor infiltrates (Kim et al., 2019; Stewart et al., 2019). Pipelines for extracting T cell receptor (TCR) and B cell receptor (BCR) sequences from bulk RNA-seq data have also been developed and can be used to extract information about immune repertoires (described in more detail below in the section titled Immune Repertoire) directly from nontargeted RNA-seq datasets. Additionally, pipelines have been developed for studying gene expression changes and genetic variation together in order to identify specific genetic variations (eQTLs) associated with changes in expression levels of specific genes. For example, 417 eQTLs associated with varying effects between stimulation conditions were identified in a recent study in which monocytes from 134 genotyped individuals were activated in vitro (Kim-Hellmuth et al., 2017). Such studies help create mechanistic links between genetic variations and immunological phenotype differences.

Meta-analysis of multiple transcriptomics studies can add increased power to allow further insights to be drawn about how genes and pathways relate to different diseases. However such combined re-analyses require careful consideration of potential batch effects. Batch effects—systemic variation in the data that may be unrelated to target biological effects—may come from various technical sources including, but not limited to, differences in sample collection and tissue processing, in methods for RNA preparation, in the sequencing or microarray technologies utilized, in overall study design, and in overall quality of distinct datasets. Given proper metadata for all samples, common methods such as ComBat (W. E. Johnson et al., 2007) or SVA (Leek & Storey, 2007) can be used to reduce contributions from potential sources of batch effects, but care must be taken to ensure that all potential sources are accounted for properly. Improper modeling of batch effect sources can result in the appearance of artificial biological signals, whereas, alternatively, overcorrection for batch effects can remove signal of true biological effects. When sufficient care is given to accounting for

batch effects, meta-analyses of transcriptomics studies can lead to novel insights. Recently, for example, a meta-analysis of three individual small studies relating to spontaneous preterm birth (sPTB) identified a set of 210 differentially expressed genes between sPTB and full-term maternal whole-blood samples after inter-study batch effects were reduced using ComBat. Pathway analysis of the gene expression differences revealed an upregulation of innate immunity and downregulation of adaptive immunity in women who delivered preterm (Vora et al., 2018). Data for such meta-analyses, as well as integrative approaches aiming to analyze expression data alongside of other data types, can be found in the Gene Expression Omnibus (GEO) (Barrett et al., 2013) (**https://www.ncbi.nlm.nih.gov/geo/**) (**Table 1.1**). This is a database containing most published transcriptomics datasets, including more than 21,000 related to the search term "immune." GEO contains metadata associated with each dataset, as well as tools for downloading and interpreting the data. An additional repository of murine immunological transcriptomic data is the Immunological Genome Project Consortium (ImmGen) (Heng et al., 2008) (**http://ImmGen.org**), which includes gene expression profiles of various, purified mouse immune cells in carefully standardized conditions.

Epigenomics

Epigenomic characterizations allow users to study chromatin features known to affect gene expression regulation, such as DNA accessibility, DNA methylation, histone modifications, and transcription factor binding. The development of techniques for studying such elements was, in large part, motivated by the fact that most GWAS-identified variants reside in noncoding regions, but the data has proven quite informative outside the realm of GWAS interpretation (Rotival, 2019). Chromatin accessibility; cytosine methylation; presence of activation- and repression-associated histone modifications; binding by activation-, activatability-, and repression-associated transcription factors; and proximity to other genomic regions associated with such markers 1) can all be studied with current epigenomic characterization techniques,

and 2) have all been associated with regulation of gene expression. Chromatin accessibility can be assessed with the assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) (J. Buenrostro et al., 2015). DNA methylation at cytosines can be studied through bisulfite sequencing (Y. Li & Tollefsbol, 2011). Regions containing particular histone modifications and regions that are bound by particular transcription factors can be identified with chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Barski et al., 2007; D. S. Johnson et al., 2007). Chromatin conformation capture techniques, reviewed recently (Rotival, 2019), can be used to identify regions of the genome within close proximity of each other, thus helping to link faraway enhancers and transcription factor binding sites to the genes they might regulate.

Information for histone modifications that are associated with activation, repression, or poising of expression can be used to identify genes and regions of the genome that are repressed, activated, or poised for activation in immune cells at different stages of the immune response. Accessibility and histone modifications have been characterized in a large-scale, systematic way for many different immune cell types and activation settings by various different groups (Calderon et al., 2019; ENCODE Project Consortium, 2012). Identification of genes regulated by individual transcription factors within specific immune cell types before and after stimulation have also been explored by multiple groups, and links have been found between, for example, specific transcription factors and differentiation, or aging, of CD8+ cytotoxic T cells (Moskowitz et al., 2017).

Combined with transcriptional profiling or genetic variation, epigenetic information provides mechanistic insights into how immune cells differentiate after stimulation (B. Yu et al., 2017) and how noncoding region genetic variations might specifically affect immune responses. Taken further, these types of analyses have been used to implicate particular immune cells that play roles in specific diseases caused by immune dysfunction (Farh et al., 2015). For powering meta-analyses and integrative analyses utilizing epigenetic data, the ENCODE (Encyclopedia of

DNA Elements) project (C. A. Davis et al., 2018; ENCODE Project Consortium, 2012) (**https://www.encodeproject.org/**) is an ongoing international consortium of research groups that has put together a vast epigenetic data repository that spans the various different epigenetic characterization techniques and that contains more than 15,000 individual samples (**Table 1.1**).

Immune Repertoire

Recent technological advances have made it possible to apply high-throughput sequencing to characterize BCR and TCR repertoires (Georgiou et al., 2014; Heather et al., 2018). For T cells, these receptors represent the other side of the MHC–TCR interaction described above in the section titled Genetics. For B cells, these receptors are involved in direct binding to target epitopes, but they also represent a proxy for an individual's antibodies because BCR gene regions are themselves converted into secreted antibodies. Thus, BCR and TCR sequencing allows for broad examination of B cell, T cell, and antibody responses by following changes in clonal and population dynamics, as well as in the functional cellular markers associated with each clone.

A key feature of the repertoire is the enormous diversity each individual is capable of producing (M. M. Davis & Bjorkman, 1988; Elhanati et al., 2018; Schroeder, 2006), which enables the recognition of a vast array of antigenic epitopes. This huge diversity occurs mainly within a particular variable domain of the BCR and TCR loci, known as the complementarity-determining region 3 (CDR3), by recombination of a set of variable (V), diversity (D), and joining (J) gene segments that form the germline-encoded B and T cell immune repertoire. With the advances of next-generation sequencing and robust computational approaches, we can study the VDJ region in fine detail (Bradley & Thomas, 2019; López-Santibáñez-Jácome et al., 2019; Shugay et al., 2015; Yaari & Kleinstein, 2015). In B and T cell immune repertoire sequencing, the approximately 100 base pairs composing the VDJ region are studied. The CDR3 region is

the main region that defines the specificity of the receptor or antibody and is often used to group B and T cells into clonotypes based on CDR3 length and sequence similarity. A common analysis is to explore immune repertoire diversity using a measure of richness of clonal diversity. Measures of diversity such as Shannon entropy and Simpson index are used to study not only richness but also the expansion of the clones. Immune repertoire data can be naturally represented as a network based on sequence diversity (Bashford-Rogers et al., 2013) as we have shown in our kidney transplant study, which demonstrated that individuals with a more active immune system prior to transplant are more likely to reject the organ (Pineda et al., 2019). In this work, we proposed a pipeline for analysis of immune repertoire data that uses both a network approach and a sequence-specific analysis at the clone and the V gene levels, which can be extended to other diseases of interest. There are numerous other examples of how immune repertoire data have been used to study the role of antibodies in disease such as multiple sclerosis (Palanichamy et al., 2014; Pineda et al., 2019; von Büdingen et al., 2012), influenza vaccine responses (Strauli & Hernandez, 2016), and primary immunodeficiency (Roskin et al., 2015).

One important development and future application in the BCR/TCR repertoire field is the prediction of antigen epitopes that receptors recognize. This is an advancement that would allow an unprecedented level of precision in clinical diagnosis, antibody drug discovery, and vaccine development. Some work has been done for TCRs of αß T cells, but such prediction is a complex problem that relies on many facets. Namely, each TCR binds peptides that are presented by certain MHCs and each MHC molecule binds and presents only a restricted set of potential peptide antigens (collectively termed MHC-restriction). Additionally, TCRs are heterodimeric and their peptide specificity can be determined by both the α and ß chains, yet these are not always sequenced in a linked fashion. A database of structures of TCRs bound to peptide-MHC complexes has been assembled which can serve as a training set for antigen-prediction methods (Gowthaman & Pierce, 2019), and work is underway to learn from known

structures to build predictive methods (Dash et al., 2017; Glanville et al., 2017). However, on top of being able to account for both MHC-restriction and missing-chain uncertainty, an ideal TCR antigen prediction method for clinical diagnostics concerning emerging infectious diseases would additionally be able to predict specificity towards novel peptide antigens that have not been studied previously.

Recently there has also been some integrative analysis incorporating genomic profiles and immune repertoire across 11 tumor types that has shown that high expression of T and B cell signatures predicts improved overall survival across breast, lung, and melanoma cancer types (Iglesia et al., 2016). Many TCR and BCR sequencing studies have deposited their data within the Sequence Read Archive (SRA) (Leinonen et al., 2011) (**https://www.ncbi.nlm.nih.gov/sra**) (**Table 1.1**). However, the development of new computational and analytical strategies to analyze and integrate this type of data are still in progress and the publicly available data need to be standardized. For that reason the Adaptive Immune Receptor Repertoire community is working on a consistent strategy for annotation, storage, and availability of this type of data (Rubelt et al., 2017), and several specialized repositories are available for this particular data type in order to perform that standardization, such as ImmPort (Bhattacharya et al., 2018) (**https://www.immport.org/home**) and the Pan Immune Repertoire Database (Zhang et al., n.d.) (**https://db.cngb.org/pird/**).

Microbiome

The study of microbial composition and its relation to health and disease has been on the rise in the past decade alongside the findings that particular microbial communities can both protect humans from and predispose them to many diseases. There are two popular approaches for sequencing microbiota. The first method is sequencing the conserved prokaryotic 16S ribosomal RNA gene. This approach involves amplification of a particular region of the gene followed by next-generation sequencing and comparison of the sequences to known

taxonomy. A caveat of this approach is that it leaves ambiguity in distinguishing between highly similar microbiota strains. The second method is using whole-genome shotgun sequencing, which usually has enhanced detection of bacterial species, increased detection of diversity, and increased prediction of genes, but this method is more expensive and it generates significantly larger quantities of data that require additional processing steps. Qiime, UPARSE, and other pipelines have been developed for the analysis of these data types (Caporaso et al., 2010; Edgar, 2013).

Study of the microbiome has elucidated many ways that this complex collection of organisms can interact with the immune system, and has highlighted various associations between particular microbial components with potential protection/predisposition of individuals from/to diseases. For example, studies have linked differences in gut microbial products, or particular microbiota, to the prevention of *Clostridium difficile* infection (Theriot & Young, 2015), of antimicrobial resistance (Relman & Lipsitch, 2018), and of tuberculosis infection (Dumas et al., 2018). On the mechanistic side, some strains of bacteria have been directly linked to suppression of other harmful pathogenic strains (Chiu et al., 2017), and particular metabolites of *Lactobacillus* species have been found to modulate T cell responses by promoting generation of tolerogenic T_{regs} (Ding et al., 2017).

It is worth noting that studies of the microbiome often highlight associations between immune perturbation phenotypes with presence or absence of particular microbial components, but the direction of causality—whether such immune perturbations may be caused by the change in microbiota, versus whether the immune perturbation is instead the cause of the identified microbial changes—is often unknown. However the microbiome field is still relatively young compared to the field of immunology, for example, so many more exciting interactions and insights, as well as mechanistic follow-ups that can add indications of directionality, can be expected from this field in coming years.

Meta-analysis can help to identify and strengthen the links between the microbiome and immunomodulation. For example, in a recent meta-analysis spanning 28 studies and 10 diseases, it was discovered that half of genera were actually associated with more than one disease (Duvallet et al., 2017). The authors extended this finding to state that many associations found in case–control studies may therefore be part of a nonspecific, shared response to health versus disease, thus warranting a need for further meta-analyses in order to identify truly disease-specific microbiota. Massive sequencing projects such as the Human Microbiome Project (Turnbaugh et al., 2007) and the Integrative Human Microbiome Project (Proctor et al., 2019), which are both available at **https://hmpdacc.org/**, are great sources of data for both healthy and disease-associated microbiome studies (**Table 1.1**). The American Gut project (McDonald et al., 2018) (**http://humanfoodproject.com/americangut/**) is another planned large microbiome study that is in the data collection phase. Some other published microbiome sequencing data not contained within these databases can be found in the SRA (Leinonen et al., 2011).

Single-Cell Sequencing

The advent of techniques for cell compartmentalization and transcript barcoding has allowed for high-throughput, whole-transcriptome or -epigenome profiling at the resolution of hundreds to hundreds of thousands of individual cells per sample. Many of the data types described above gain significantly enhanced capabilities when analyzed at single-cell resolution because it replaces the need for groups to be separated prior to characterization with the ability to group cells artificially, in many different ways, in post-capture analysis. Such resolution allows heterogeneity within individual cell types to be analyzed as well. Many different forms of single-cell analysis have been developed with varied applications, including single-cell RNA-seq (scRNA-seq) for transcriptional profiling, single-cell ChIP-seq (Grosselin et al., 2019) and single-cell ATAC-seq (J. D. Buenrostro et al., 2018) for epigenome characterization, single-cell DNA-

seq for both microbiome characterization and mutation heterogeneity analysis of cancers (Blainey & Quake, 2014), and CRISP-seq for combining scRNA-seq transcriptional profiling with high-throughput CRISPR screens targeting multiple genetic loci (Jaitin et al., 2016). Each was recently reviewed (Chappell et al., 2018).

scRNA-seq is the most advanced, standardized, and widespread single-cell sequencing technique and is therefore the most amenable currently to systems immunology approaches. Most types of analysis available for bulk RNA-seq are available now for scRNA-seq, including limited methods for batch correction, and many additional analyses are being developed. The typical pipeline of clustering similar cells using machine learning algorithms, such as those within the Seurat package (Macosko et al., 2015; Stuart et al., 2019), followed by performance of differential expression across clusters has been used to identify novel immune cell types in multiple settings (Papalexi & Satija, 2018). This approach is useful for characterizing all populations of cells within blood, bone marrow, tumors, or other tissues, as well as how those populations might change either over time or in a vaccination or activation setting (Tirosh et al., 2016). Pseudotime analysis is another immensely powerful tool that can be performed with single-cell transcriptional profiling datasets. Algorithms for pseudotime analysis include Monocle (Trapnell et al., 2014) and slingshot (Street et al., 2018), among many others (Saelens et al., 2019). With pseudotime analysis, cells can be ordered according to a developmental or differentiation timeline. Afterward, important genes and pathways can be identified, and their expression changes over time after activation or during development, as well as bifurcation points between heterogeneous trajectories, can be analyzed (Drissen et al., 2016; Psaila & Mead, 2019; Schlitzer et al., 2015).

Single-cell sequencing currently comes at the cost of capturing fewer genes or regions per cell, relative to the number of genes or regions captured per whole sample in bulk RNA-, ChIP-, or ATAC-seq of thousands of cells. Thus, direct differential expression comparisons can often be made stronger with the use of bulk sequencing. However, groups of cells from single-

cell sequencing experiments can be combined into pseudobulk profiles to mitigate this drawback. Moreover, the ability to de novo cluster cells that have similar profiles and then make comparisons directly across these groups has immense power for characterizing novel patterns of heterogeneity and cannot be understated. Still, it is worth noting that this resolution does come with a few costs. Computational power and technical expertise requirements for analysis of single-cell sequencing data are quite high. Additionally, certain necessary features for cell type determination that are regularly directly targeted in flow cytometry and cytometry by time-of-flight (CyTOF), as described in the next section, are only infrequently captured in common scRNA-seq approaches. Some others are simply indistinguishable from similar features. For example, CD4, which is necessary for T cell CD4 versus CD8 lineage determination, is often not captured from every CD4+ T cell in scRNA-seq datasets. Moreover, determination of CD45 (*PTPRK* gene) isoform usage in typical scRNA-seq data, which is useful for sub-phenotyping naive versus memory T cells, currently requires a specialized scRNA-seq analysis pipeline (Ntranos et al., 2019).

Simultaneous cellular indexing of transcriptomes and epitopes (CITE-seq) is another major approach that can be used to address the missing marker gene and isoform caveats described above. The technique utilizes user-picked, DNA-barcoded antibodies to add quantification of per-cell protein or isoform expression levels alongside scRNA-seq approaches. CITE-seq's use within large meta-analysis data integration pipelines faces additional barriers, some of which will surely lessen in the coming years. Currently, CITE-seq is in the early stages of commercialization and standardization of the sample handling pipeline and best practices for post-sequencing analysis of CITE-seq data have not yet been fully established. But some technical caveats will surely remain, including the additional planning, reagents, sequencing, and computational costs on the user end, plus the fact that CITE-seq antibodies may vary between studies, adding additional technical variation that meta-analysis approaches will need to take into account.

Not much meta-analysis has been conducted with single-cell data so far of which we are aware, likely due to a lack of standards, methods, and availability of data. Single-cell specific methods for batch effect reduction and dataset integration, such as Harmony (Korsunsky et al., 2019), LIGER (Welch et al., 2019), or the unnamed method internal to Seurat version 3 (Stuart et al., 2019), have been developed and appear to fair relatively well, compared to other available methods, at reducing technical signals that arise from differences in study structure and in sequencing technologies (Tran et al., 2020). However, such methods often specifically target only the dimensionality reductions of single-cell data but not the gene expression counts data (Harmony), or target expression matrices as well but such outputs are explicitly not recommended for use in differential expression due to unpredictable creation of noise between groups (Seurat). Thus, rather than a standard meta-analysis approach, the currently recommended path for combined utilization of multiple single-cell datasets is to utilize batch-correction to perform clustering and identification of cell types in a combined way, followed by separate analyses within individual datasets using the pre-batch-correction data.

Researchers are starting to aggregate single-cell data into new repositories such as the Jingle Bells repository (Ner-Gaon et al., 2017) (**http://jinglebells.bgu.ac.il**), the PanglaoDB database (Franzén et al., 2019) (**https://panglaodb.se/**), and the Single Cell Expression Atlas (Cook et al., 2019) (**https://www.ebi.ac.uk/gxa/sc/home**) (**Table 1.1**). Many other scRNA-seq studies have deposited their data within GEO (Barrett et al., 2013).

Protein Expression at the Single-Cell Level

Few techniques exist with wider usage and applicability in the field of immunology than flow cytometry and fluorescence-activated cell sorting (FACS). These techniques involve incubating cells with fluorophore-labeled antibodies that are specific to the proteins of interest, followed by quantification of fluorescence levels as a proxy for actual protein expression levels. FACS involves the additional step of selecting cells with particular combinations of marker

expression for diversion into collection tubes. The extraction of highly purified cell populations for various experimental and characterization techniques is often carried out with FACS. Flow cytometry has been continuously used for characterizing and identifying heterogeneous expression levels of proteins on the surface of immune cells for more than four decades (McKinnon, 2018).

One of the limitations of flow cytometry is the use of fluorescent markers for quantifying antibody abundance. Overlap of fluorescence spectra has impeded the use of more than 10–15 markers in standard flow cytometry experiments. Marking of individual cell lineages can often take up three or more markers per lineage, leaving even fewer spectra for quantifying proteins of interest. Thus, flow cytometry, while powerful for targeted experiments, has been limited in its ability to characterize multiple cell types.

In CyTOF, a more recently developed technique which utilizes a modified method of mass spectrometry, heavy metals with distinct atomic weights are used instead of fluorescence spectra. Mass cytometry is limited only by the availability of heavy metals of distinct masses, leading to potential utilization of 30–45 markers per panel. CyTOF's expanded feature space can be used to characterize the states of many more distinct cell types within a tissue sample, all in one experiment, similar to single-cell sequencing. The breadth of markers that can be assessed for every cell in CyTOF can be too much for manual marker-by-marker analysis methods that are typically used for flow cytometry analysis. Thus, dimensionality-reduction techniques are often used instead, followed by manual investigation (Kimball et al., 2018). This pipeline is quite similar to scRNA-seq approaches. CyTOF data can be analyzed by other methods that are similar to scRNA-seq as well, such as pseudotime analysis for tracking expression profiles alongside differentiation and activation timelines, and clustering for identifying novel cell subtypes (Setty et al., 2016).

Indeed, scRNA-seq and CyTOF overlap immensely in potential broad expression profiling applications, with users typically deciding between the two based on the limitations

inherent to each technology. While CyTOF is limited to measuring expression of less than 50 user-identified proteins and to proteins for which antibodies exist, single-cell sequencing captures hundreds to thousands of genes per cell. As such, single-cell sequencing approaches have the benefit of not requiring marker selection beforehand and therefore are less user biased, but analysis of the higher-parameter scRNA-seq data requires more computational power and expertise. Conversely, while scRNA-seq currently suffers from loss of information about certain important cell lineage markers, CyTOF can capture this information readily if antibodies for those markers exist.

In the last few years, CyTOF has been used to assess cellular expression and heterogeneity in quite varied applications. For example, CyTOF was used to characterize, throughout pregnancy, the diverse peripheral blood immune cell types, which facilitated the creation of a chronology of pregnancy-associated immune changes and revealed a novel modulation of interleukin-2 signaling in T cells during pregnancy (Aghaeepour et al., 2017). Another application helped to characterize the vast heterogeneity of natural killer (NK) cell phenotypic populations, which were estimated to be 6,000–30,000 per individual studied. Through such applications it was discovered that NK inhibitory receptor expression is largely determined by genetics, whereas NK activation receptor expression is heavily environmentally influenced (Horowitz et al., 2013).

Flow cytometry's relative ease of use due to decades of commercialization has made it a common characterization both for ex vivo samples and for tracking outcomes of in vitro experiments. However, flow cytometry and CyTOF data are rarely submitted to data repositories upon publication. Thus, meta-analyses of these data types are currently hampered by data availability. In recent years, many groups have attempted to fix this (**Table 1.1**). Spidlen et al. (Spidlen et al., 2012) created the first public repository of flow cytometry data in 2012 (**https://flowrepository.org/**), which currently holds 37 annotated flow cytometry datasets. Hu et al. (Hu et al., 2018) recently developed MetaCyto, a tool for automated meta-analysis of both

mass and flow cytometry data. The Cytobank website (https://www.cytobank.org/data-sets.html) contains links to numerous publications that have used CyTOF, and ImmPort (Bhattacharya et al., 2018) currently contains data from 140 flow cytometry and 27 CyTOF experiments.

Part 3 – Fetal Immune Cells & The Layered Transition Hypothesis

Cells of the developing fetal immune system are functionally and transcriptionally unique compared with phenotypically similar cells in the adult. In previous work, it has been shown that midgestational fetal naïve αβ CD4 T cells differ from phenotypically-similar adult naïve T cells in that they are predisposed to promote tolerance and preferentially differentiate into $FoxP3^+$ regulatory T (T_{reg}) cells upon activation (Bronevetsky et al., 2016; Mold et al., 2008, 2010; Ng et al., 2019). More recently, transcriptional and epigenetic programs, that are shared between fetal naive T cells and committed adult Treg cells yet are inactive in adult naive T cells, and that likely contribute to this predisposition, have been described (Ng et al., 2019). Separately, fetal classical ($CD14^+CD16^-$) monocytes have been reported to phosphorylate canonical and noncanonical signal transducers and activators of transcription (STATs) and also to show impaired upregulation of antigen presentation capacity, compared to adult classical monocytes, in responses to cytokine stimulation (Krow-Lucal et al., 2014). Importantly, it has also been demonstrated that the unique nature of fetal naïve T cells is programmed at the level of their hematopoietic progenitors: when transplanted into humanized SCID-hu Thy/Liv mice, human fetal – but not adult – $CD34^+$ HSPCs give rise to naïve CD4 T cells with a transcriptional profile and a predisposition toward T_{reg} differentiation that is similar to that found in primary fetal naïve CD4 T cells (Mold et al., 2010). Thus, T cells and monocytes generated in the second trimester may represent effectors of a fetal immune lineage that is distinct from an adult immune lineage generated later in development. However, the processes by which fetal cells transition to more mature developmental states remain unknown.

The biological processes by which cells that are developmentally restricted to the fetal period are replaced by their counterparts in post-natal life have been the subject of much speculation. Previous studies of B and T lymphocyte lineages in mice and birds, which capitalized on the then emerging technology of flow cytometry for single-cell resolution analysis, suggested the programmed, staged emergence of temporally-restricted immune cell populations throughout gestation and beyond (Havran & Allison, 1988; Hayakawa et al., 1985; Jotereau & Le Douarin, 1982; Kantor et al., 1992; Lalor et al., 1989; Le Douarin & Jotereau, 1975; Montecino-Rodriguez et al., 2006). Three decades ago, several seminal publications proposed a model in which development of the immune system is temporally "layered", positing that fetal and adult lymphoid cells arise from distinct, developmentally restricted lineages of fetal and adult hematopoietic stem cells (HSCs) (Herzenberg & Herzenberg, 1989; Ikuta et al., 1990). While additional evidence for layered immune system development has been generated more recently in mice (Montecino-Rodriguez et al., 2016, 2018; Ramond et al., 2014), and many immune cells generated in the human fetus are distinct from those arising from adult HSCs, including $CD5^+$ B-1 B cells, fetal erythrocytes, microglia, and certain $\gamma\delta$ T cells (Ginhoux et al., 2013; Hadland & Yoshimoto, 2018; Montecino-Rodriguez & Dorshkind, 2012; Stamatoyannopoulos, 2005; Tieppo et al., 2020), it remains unconfirmed whether a similar, layered, developmental transition occurs in humans around the time of birth.

Part 4 – Aims of this Study

Human newborns have impaired responses to vaccines (Siegrist, 2001) and are more susceptible to serious infections, even by microbes that are generally non-pathogenic in older children and adults (Blanchard et al., 2015; Phares et al., 2008). The persistence of fetal tolerogenic immune responses at birth would predictably contribute to these immunological vulnerabilities by blunting anti-vaccine and anti-microbial immune responses. As the newborn period represents an intermediate timepoint between fetal and adult immunity, understanding

the timing and mechanism by which the fetal-to-adult transition occurs could accordingly have significant implications for newborn health. The layered immune system hypothesis (Herzenberg & Herzenberg, 1989) posits that HSPCs giving rise to a tolerogenic fetal immune system are superseded by distinct HSPCs that instead give rise to a protective adult HSPC-derived immune system, and that for a time these immune systems are layered upon each other, exerting opposing influences. Alternatively, the fetal-to-adult transition might instead occur through relatively uniform progressive maturational changes across immune cell populations, leading to a continuous spectrum of intermediate phenotypes (**Fig. 1.5A**). Depending on which of these two models is operative, naïve T cells at the time of full-term birth would be predicted to either consist of a mixture of fetal- and adult-like T cells or a relatively homogenous population of cells with an intermediate newborn phenotype. The dissertation work presented herein was designed to distinguish between these alternatives.

In recent years, rapid advancements in transcriptional profiling have made it feasible to sequence the transcriptomes of tens of thousands of individual cells in a single reaction while concurrent advances in computational and machine learning approaches have facilitated the interpretation of these complex molecular datasets. Here, initial conventional transcriptional profiling approaches (mRNA microarray and quantitative reverse transcription polymerase chain reaction; qRT-PCR) were utilized to identify a transcriptional signature of fetal and adult immune cells, and to demonstrate that immune cells of newborn humans show intermediate and variable expression of this maturation signature. To better understand the maturational state and composition of T cells at the single-cell level during the newborn period, we then leveraged single-cell RNA sequencing technology, a machine-learning approach within our bioinformatics pipeline, and a newly-developed visualization tool (dittoSeq) built for side-by-side analysis of bulk and single-cell RNA sequencing data to directly compare the transcriptomes of fetal, full-term UCB, and adult naïve T cells and $CD34^+$ HSPCs (**Fig. 1.5B**). By doing so, we were able to ask whether fetal-associated features persist in UCB immune cells at birth and, if so, whether T

cells in UCB are composed either of layered waves of cells with either fetal or adult transcriptional programs or of a single population with an intermediate transcriptional program.

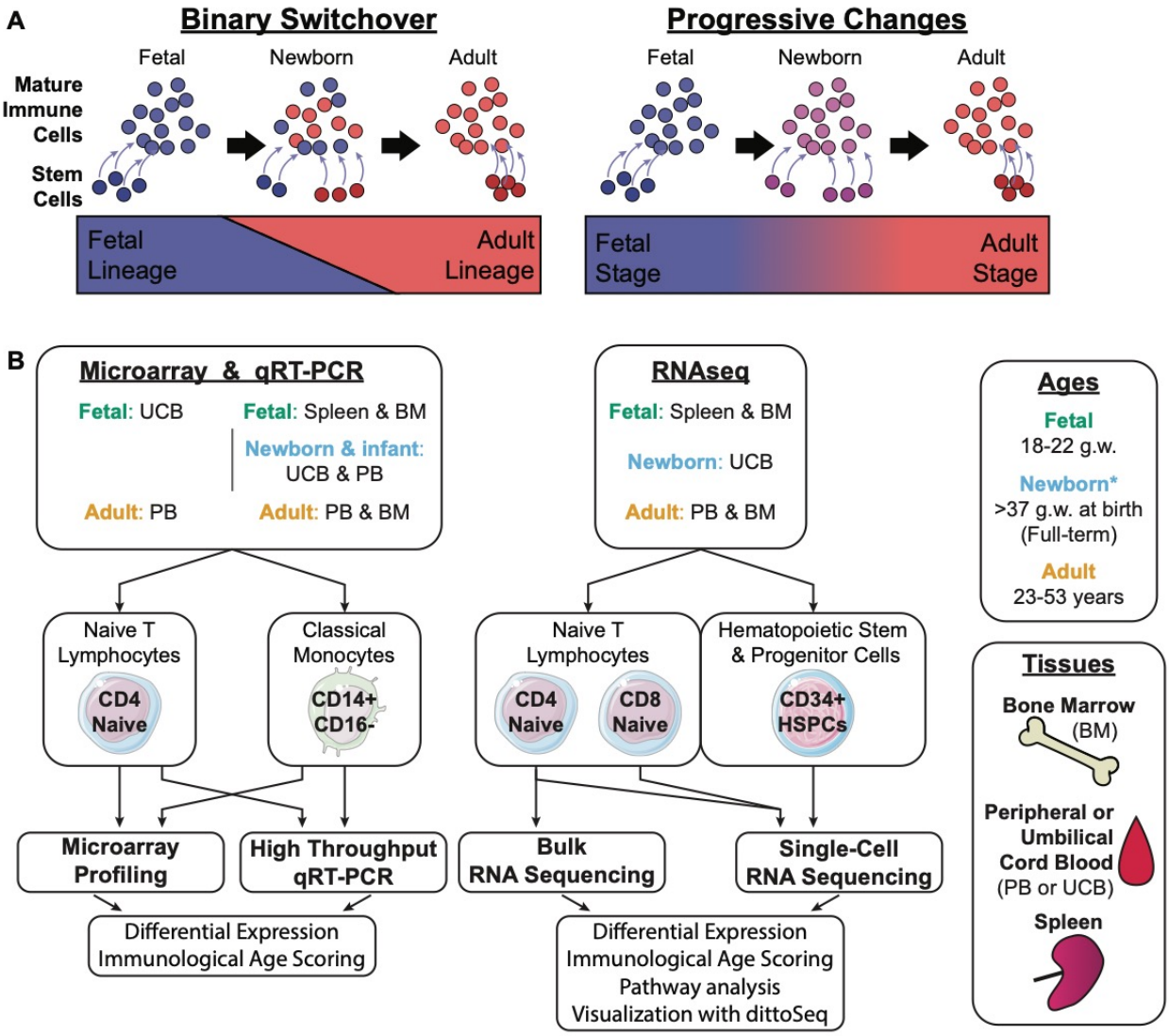


Figure 1.4: Theoretical and experimental overview

A. Theoretical models for how the immune cell fetal-to-adult transition may occur.
Left, the Binary Switchover model in which an initial wave of hematopoiesis emanating from fetal HSPCs gives rise to a fetal layer of differentiated immune cells. A subsequent wave arising from distinct adult HSPCs then results in the adult layer of differentiated immune progeny cells. During the transition period, the co-existence of fetal and adult layers would have intermediate phenotype at the population level, but not at the single-cell level.
Right, the Progressive Change model in which an original population of fetal-phenotype HSPCs undergoes multiple gradual, relatively homogenous, changes that ultimately result in adoption of an adult HSPC phenotype. Throughout this transition, HSPCs that are intermediate between fetal and adult give rise to progeny cells that also have intermediate phenotypic characteristics. In the newborn period, the immune system would have an intermediate phenotype at the population and single-cell level.
B. Experimental overview showing the tissues collected and cells sorted to generate the various datasets of the current study, and methods used to compare fetal, newborn, and adult cells within these datasets. g.w. = gestational weeks. * = all newborn samples were collected at the time of birth from deliveries in this time period except for samples of the URECA cohort which spanned deliveries from 34-42 g.w.

Chapter 2 – Single-cell mapping of progressive fetal-to-adult transition in human naïve T cells

Material for this chapter was modified from a manuscript currently in-press:

Daniel G. Bunis, Yelena Bronevetsky, Elisabeth Krow-Lucal, Nirav R. Bhakta, Charles C. Kim, Srilaxmi Nerella, Norman Jones, Ventura F. Mendoza, Yvonne J. Bryson, James E. Gern, Rachel L. Rutishauser, Chun Jimmie Ye, Marina Sirota, Joseph M. McCune, Trevor D. Burt. "Single-cell mapping of progressive fetal-to-adult transition in human naive T cells," in-press with Cell Reports, 2020.

The transition between fetal and adult transcriptional programs in human T cells and monocytes is incomplete at full-term gestation

We first sought to investigate the extent to which the immune system at full-term gestation has transitioned from a fetal to an adult program. Most transcriptional profiling studies of fetal immunity have been carried out utilizing cells associated with lymphoid organs (e.g., thymus, spleen, gut, mesenteric lymph node, or bone marrow) due to the technical difficulty of sampling fetal peripheral blood (Bronevetsky et al., 2016; Cupedo et al., 2005; Halkias et al., 2019; Krow-Lucal et al., 2014; N. Li et al., 2019; Mold et al., 2008, 2010; Ng et al., 2019). Here, a fetal vs. adult expression signature was first derived using global gene expression analysis in phenotypically-similar populations of sort-purified fetal and adult peripheral blood classical monocytes (HLA-DR$^+$CD14$^+$CD16$^-$) and naïve, non-T$_{reg}$, CD4 T cells (CD3$^+$CD4$^+$CD25$^-$ CD45RA$^+$CD27$^+$). Adult and fetal peripheral blood cells were obtained from adult donor apheresis units and from fetal UCB, respectively (sorting strategy shown in **Fig. 2.1**). To define common hematopoietic developmental programs conserved between myeloid and lymphoid lineages, we identified 159 genes that were significantly differentially expressed in the same direction (false discovery rate (FDR) < 0.05; and expression fold change > 1.5) between fetal and adult cells in both monocytes and naïve T cells. These signature genes were then used to build a qRT-PCR-based developmental stage scoring system for use in full-term UCB and other peripheral blood samples. After validation and optimization testing of qRT-PCR primers, a set of 33 genes shared between monocytes and T cells met criteria to represent a fetal vs. adult transcriptional signature (**Fig. 2.2A**). Principal component analysis (PCA) was used to weigh the expression of signature genes by the degree to which they contribute to distinguishing fetal vs. adult identity for T cells and for monocytes separately (**Fig. 2.3A,B**), and we confirmed that fetal and adult T cell and monocyte populations from prior microarray analyses (Krow-Lucal et al., 2014; Mold et al., 2010) scored appropriately as fetal or adult (**Fig. 2.3C,D**). We further demonstrated that naïve CD4 and

classical monocyte populations sorted from newly-processed fetal tissue (spleen or bone marrow) and adult peripheral blood, subjected to qRT-PCR and analyzed for developmental stage score, also scored appropriately as highly fetal or adult (**Fig. 2.2B**).

To assess the degree of immune maturation (i.e., the degree of transition from fetal to adult gene expression patterns) at full-term gestation, we extended our developmental stage score analysis to naïve CD4 T cells and classical monocytes that were sort-purified from the UCB of a cohort of healthy, vaginally delivered, full-term newborns (>37 weeks gestation age at birth, n=29). Although we observed a wide range of inter-individual variability among samples, this variability across UCB CD4 T cell and monocyte populations was less than the variation between fetal versus adult samples (**Fig. 2.2C**). Notably, some samples had monocytes and T cells with more adult-like gene expression scores while others scored more similarly to fetal samples. We also found that, within a given individual, there was a significant correlation between T cell and monocyte signature scores, suggesting that the fetal-to-adult transition of these hematopoietic lineages may be under control of a shared regulatory mechanism (**Fig. 2.2D**).

We then applied the developmental stage scoring system to analyze sort purified CD4 T cells and classical monocytes from a much larger independent birth cohort of infants born between 34 wks. gestational age (GA; i.e., late pre-term) and 42 wks. GA (i.e., post-term) with a reduced set of immunological markers (see Materials and Methods for details). A wide range of variability at any given gestational age was observed as well as a significant positive correlation between developmental stage score and gestational age at birth for T cells, but not monocytes (**Fig. 2.2E**). The absence of correlation between developmental stage score and gestational age in monocytes may be the result of the relatively short life span of monocytes resulting in increased heterogeneity in samples' developmental stages at any a given time point compared with T cells.

Given that CD4 T cells displayed graded maturation even in a focused window of gestational age ranges at birth, we proceeded to investigate the pace of fetal-to-adult transition specifically in CD4 T cells after birth in a separate cohort of HIV-exposed, but uninfected and

otherwise healthy neonates. Developmental stage scoring analysis performed on naïve CD4 T cells collected longitudinally for the first several months of life revealed a steady transition in developmental stage scores, beginning with a largely fetal signature and transitioning to a more adult-like signature within 1-2 months (**Fig. 2.2F**).

Individual UCB T cells have intermediate transcriptional profiles and are unimodally distributed between fetal and adult cell profiles

While bulk sample analysis revealed that UCB cell transcriptional profiles are generally intermediate between those of fetal and adult cells, such an approach does not distinguish between the two proposed models of fetal-to-adult transition (**Fig. 1.1A**). To determine if UCB cells are either composed of a mixture of fetal- and adult-like cells (i.e., as predicted by the layering hypothesis) or a relatively homogenous population with intermediate gene expression (i.e., consistent with a model of progressive change), we turned to single-cell RNA sequencing (scRNA-seq). Fetal splenic, full-term UCB, and adult peripheral blood naïve, non-T_{reg}, CD4 T cells ($CD4^+CD45RA^+CD27^+CD25^-$) and naïve, non-stem cell memory, CD8 T cells ($CD8^+CD45RA^+CD27^+CCR7^+CD95^-$) were purified by FACS (n=5 per age group; **Fig. 2.4**). Cells from all samples yielding more than 1000 cells were pooled between three lanes for library generation in such a way that CD4 vs. CD8 T cell identity would be known after sample identification and doublet removal with Demuxlet (**Fig. 2.5**). Altogether, we obtained 40,152 T cell transcriptomes representing a median of 1,464 CD4 or CD8 cells per sample (range 777 – 1,829).

Dimensionality reduction analysis and clustering were performed using Seurat (Butler et al., 2018) to query general transcriptional relationships. In uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), cells from distinct ages grouped in distinct regions of the UMAP plot, with fetal cells at the positive end, UCB cells in the middle, and adult cells at the negative end of UMAP_1 (**Fig. 2.6A**). Using Seurat's Louvain clustering algorithm to quantify the

separation within dimensionality reduction space, clustering at low resolution grouped cells of the same age with 92.5% accuracy overall (**Fig. 2.6B**). Notably, and in contradistinction to the layering hypothesis, rather than consisting of a mixture of fetal-clustering and adult-clustering cells, CD4 and CD8 naïve T cells from UCB samples instead clustered mostly (91.9% mean per sample, 77.3% minimum) as a singular population of cells that aggregated between the adult and fetal clusters (**Fig. 2.6C,D**).

To more directly characterize the fetal-to-adult progression of individual cells, a single-cell developmental stage score was derived. Here, we leveraged a robust machine learning approach which could be focused on commonly captured genes, that are part of the fetal-to-adult transition, and that would therefore be less influenced by noise within sparsely captured single-cell transcriptome data. In so doing, we applied the relatively assumption-agnostic random forest machine learning regression method to an initial training set of 10% of the fetal and adult naïve T cells, and an initial set of marker genes that were differentially expressed (log fold change ≥ 0.585, FDR < 5%) between these cells (**Fig. 2.6E**). To narrow down the number of features entered into our final model and thereby reduce the chances of the model becoming overfit to the training set, the iterative random forest model building and feature elimination algorithm from the feseR package (Perez-Riverol et al., 2017) was utilized. The final set of marker genes spanned diverse functionalities including transcriptional and post-translational regulation, signal transduction, and cytoskeletal regulation. Finally, expression of these 17 definitive markers within the training set cells was used to train a random forest regression model. Confirming the model's ability to properly score non-transitioned (i.e., fetal) and fully transitioned (i.e., adult) cells, our single-cell developmental stage score model accurately scored the set of fetal and adult cells which were left out of the model training process (**Fig. 2.6F,G**; AUROC = 0.9996). Importantly, although phenotypic differences, such as those resulting from prior homeostatic proliferation, exist between naïve cells within an individual fetal or adult sample (Mold et al., 2019; van den Broek et al., 2018),

our system consistently scores most naïve T cells of the same fetal or adult sample similarly, indicating that the system is not sensitive to such features.

We next evaluated the scores given to cells from UCB samples, noting four key features. First, both CD4 and CD8 UCB naïve T cells generally received intermediate developmental stage scores (**Fig. 2.6F,G**): whereas 84% of fetal cells were scored 0.1 or lower and 89% of adult cells scored 0.9 or greater, 88% of newborn UCB cells scored between 0.1 and 0.9 (mean = 0.43, standard deviation = 0.24). Secondly, the scores of cells within each individual UCB sample were unimodally distributed, consistent with progressive change, rather than with a binary switch, which would have shown a bimodal distribution (**Fig. 2.6F,G**). Thirdly, variation between samples appears as a shift in the mode of a sample's scores rather than via differences in abundance of very low scoring (i.e., fetal-like) or very high scoring (i.e., adult-like) cells – indicating that variation in progression manifests as relatively homogeneous changes rather than as differences in preponderance of cells with relatively distinct phenotypes. Lastly, the expression pattern among UCB cells for six of the marker genes was more closely aligned with either the fetal cell pattern (*RPS24*, *SOX4*, *RGS1*) or the adult cell pattern (*UBB*, *ACTB*, *HSP90AA1*), suggesting that individual genes that make up the signature undergo transitions in expression with different timing (**Fig. 2.7**). These findings demonstrate that the transition from fetal-to-adult layers during human T cell development occurs as a process of gradual, progressive change and not by layering of a fetal lineage alongside a distinct adult lineage.

Subsets of genes undergo fetal-to-adult transition with different timing

Given the evidence that the fetal-to-adult transition occurs via progressive change between fetal and adult transcriptional programs, we hypothesized that the UCB naïve CD4 T cell transcriptome would share some properties with fetal naïve CD4 T cells and others with adult naïve CD4 T cells. We also hypothesized that some fetal and adult genes may be intermediately

expressed by UCB cells while other genes might be unique to UCB. To address these hypotheses, we utilized bulk RNA sequencing to obtain greater depth and to thus achieve more complete transcriptional profiling of fetal, UCB, and adult naïve CD4 T cells. Both the bulk and single-cell datasets showed similar intermediate placement of UCB cells in dimensionality reduction analysis. In PCA of the bulk RNA-seq dataset, UCB samples clustered at an intermediate point between fetal and adult samples in PC1 (**Fig. 2.8A**), and CD4 and CD8 T cells from UCB similarly clustered intermediately in UMAP_1 between fetal and adult cells in UMAP analysis of the scRNA-seq data (**Fig. 2.6A**). Further, when genes differentially expressed between ages in both these datasets were compared using equivalent differential expression criteria, all but seven genes identified in the single-cell dataset were also identified in the bulk dataset (**Fig. 2.9**). Finally, to address the potential confounder of tissue origin in the RNA-seq datasets (i.e., fetal spleen vs. blood for newborn and adult samples), we utilized the differentially expressed genes from our prior microarray analysis (**Fig. 2.2**) to compare expression between circulating and tissue-derived fetal T cells, and found that genes differentially expressed between circulating fetal versus adult cells could discriminate expression patterns in fetal spleen- versus newborn UCB- and versus adult PB-derived T cells analyzed by bulk RNA-seq (**Fig. 2.10**).

Upon differential expression (log2 fold change ≥ 1.5, FDR ≤ 0.05) analysis of the bulk RNA-seq dataset, 2,201 unique genes were identified as significantly differentially expressed between all pairwise age comparisons: 1,840 genes between fetal and adult naïve T cells; 713 genes between fetal and UCB; and 855 genes between UCB and adult. Out of these pairwise comparisons, 443 genes were upregulated in both fetal and UCB samples but not adult samples, while 154 genes were enriched in both UCB and adult samples but not fetal samples (**Fig. 2.11**). K-means clustering of all differentially expressed genes (**Fig. 2.8B**) revealed a spectrum of gene expression clusters that were distinctively fetal or adult up-regulated with varying degrees of expression in UCB cells. One cluster of fetal up-regulated genes remained fully upregulated in UCB cells (cluster 3) while, by contrast, another cluster of fetal up-regulated genes were partially

down-regulated (cluster 1) and a third cluster was further up-regulated (cluster 2) in UCB T cells. In the other direction, cluster 5 represented adult-upregulated genes that were already partially expressed in UCB T cells, while cluster 4 represented a subset of adult-upregulated genes that showed relatively low expression in UCB T cells. These results strongly suggest that clusters of co-regulated groups of genes, or gene expression "modules," undergo transition from fetal-like expression patterns to adult-like expression patterns with different timing (**Fig. 2.8C**).

Several gene pathways are uniquely enriched within the UCB T cell transcriptome

Gene pathway analysis is frequently used to interpret gene expression data in such a way as to better understand the functional programs that are active in a given cell or population of cells. Comparing gene pathways that are activated in fetal, UCB, and adult samples, particularly in comparison to one another, may provide a deeper understanding of the cellular functional adaptations at each developmental stage. To identify specific fetal- and adult-associated genetic programs expressed in UCB cells, we initially applied unbiased pathway enrichment analysis to fetal-associated cluster genes (clusters 1, 2, and 3) and adult-associated cluster genes (clusters 4 and 5) highlighted in **Figures 2.8B** and **C**. Immune activation pathways, both general (e.g. antigen processing and presentation, Th1, Th2, and Th17 differentiation, intestinal immune network for IgA production, and phagosome pathways) and pathologic (graft-versus-host disease, inflammatory bowel disease, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes mellitus, asthma, allograft rejection, and autoimmune thyroid disease-associated pathways), were found to be enriched in the adult clusters (**Fig. 2.8C**), consistent with previous findings that fetal CD4 T cells have decreased enrichment for genes that determine other effector differentiation fates due to their bias towards T_{reg} differentiation (Mold et al., 2008, 2010; Ng et al., 2019). Also consistent with previous reports that fetal naïve T cells are more highly proliferative compared to

adult naïve T cells (Halkias et al., 2019; Michaëlsson et al., 2006), the cell cycle gene pathway was enriched within fetal-associated gene clusters (**Fig. 2.8C**).

To understand at higher granularity how different pathways are enriched within the distinct stages of the fetal-to-adult transition, unbiased pathway enrichment analysis was next applied to each of the separate pairwise comparison upregulated gene sets (**Fig. 2.12**). Suggestive that some tolerogenic fetal programs might persist at birth, some immune activation pathways were enriched within the adult relative to UCB as well as fetal samples (e.g., asthma, allograft rejection, autoimmune thyroid disease, and intestinal immune network for IgA production). Gene pathways upregulated only in adult samples, but not in UCB or fetal samples, likely are associated with normal developmental changes from birth to adulthood and/or represent adaptations that occur specifically in response to environmental exposures in the post-natal period. IgA production and intestinal immunity can be associated both with normal colonization by microbiota and with immune responses to mucosal infections. Thus, enrichment of the intestinal immune network for IgA production in adult relative to both fetal and UCB samples is consistent with widespread mucosal responses to microbial exposure after birth (Dominguez-Bello et al., 2019; Zhuang et al., 2019).

Pathway analysis also revealed differences in several signaling pathways. In particular, multiple signaling pathways (e.g., Notch, Oxytocin, phospholipase D, and relaxin signaling pathways) were exclusively enriched within UCB samples relative to adult samples while others (e.g., Rap1, Ras, and Wnt signaling pathways) were jointly enriched within both fetal and UCB samples relative to adult samples (**Fig. 2.12B**). Yet other signaling pathways (e.g., AGE-RAGE, MAPK, p53, and TNF signaling pathways) were found to be enriched within fetal samples compared to both UCB and adult samples. Interestingly, the estrogen signaling pathway stood out among these pathway enrichment comparison analyses in that it was enriched in both the “Fetal up vs UCB” and “UCB up vs fetal” gene lists, indicating that fetal and UCB naïve T cells differentially express distinct sets of genes within this single pathway. We examined all genes

contributing to the estrogen signaling pathway enrichments, and found that UCB samples showed higher expression of *BCL2*, *ESR1*, *GABBR1*, and *PLCB1* while the fetal samples showed higher expression of *CREB5*, *FKBP4*, *HBEGF*, *HSPA1A*, *HSPA1B*, *HSPA2*, *HSPA6*, *HSP90AA1*, *RARA*, and *TGFA*. We conclude that, while multiple fetal gene expression pathways persist within UCB at the time of birth, others do not. We also identify multiple signaling pathways that seem to arise exclusively in UCB.

UCB naïve T cells retain a partial expression of a fetal-associated T_{reg} signature

The relative enrichment of several pathways related to inflammatory immune responses in adult samples compared to both fetal and UCB samples likely reflects the longstanding observation that prenatal immunity is largely dominated by tolerogenic and/or less-inflammatory responses (Burt, 2013; Cupedo et al., 2005; Darrasse-Jèze et al., 2005; Krow-Lucal et al., 2014; Michaëlsson et al., 2006; Mold et al., 2008, 2010). Fetal naïve CD4 T cells have been shown to have elevated expression of *IKZF2* (Helios) and other T_{reg}-associated genes, and it is thought that the elevated expression of such genes may contribute to their T_{reg}-differentiation predisposition (Ng et al., 2019). Thus, we next determined whether elevated expression of a published T_{reg}-associated transcriptome (Ng et al., 2019) might be retained in UCB cells. Many T_{reg} up- and down-regulated signature genes were found to remain up- and down-regulated in UCB relative to adult naïve T cells (**Fig. 2.13**). Overall, UCB naïve T cells expressed T_{reg}-upregulated genes more highly, including *IKZF2* and *FOXP3*, and expressed T_{reg}-downregulated genes less highly than adult naïve T cells. Of note, UCB naïve T cells were intermediate in their expression of the overall T_{reg}-associated gene signature between fetal and adult samples, suggesting that they have shut down parts of the program predisposing them toward T_{reg} differentiation.

The transcriptomes of CD34$^+$ HSPCs in UCB are intermediate between fetal and adult HSPCs

It has previously been demonstrated that human fetal CD34$^+$ HSPCs, after transplantation into humanized mice, give rise to CD4 T cells that are transcriptionally similar to primary fetal T cells and that are also predisposed toward T_{reg} differentiation (Mold et al., 2010). It has also been reported that both human embryonic stem cells and fetal CD34$^+$ liver HSPCs, but not CD34+ HSPCs from peripheral blood of adult donors, differentiate *in vitro* into monocytic cells with functional and gene expression profiles that are characteristic of anti-inflammatory M2-type macrophages (Klimchenko et al., 2011). Given the finding that individual UCB naïve T cells express an intermediate transcriptional phenotype, we sought to understand whether CD34$^+$ HSPCs in UCB might also have a transcriptional profile that is intermediate between fetal and adult HSPCs. The answer to this question may also have significant clinical implications as UCB is commonly used as a source of hematopoietic progenitors for therapeutic transplantation in the settings of both primary immunodeficiency and cancer. Thus, given that the marker CD34 is present on multiple early stages in the developmental ontogeny of human HSPCs, and that the entire pool of CD34$^+$ cells is regularly used in clinical transplantation protocols, we included all CD34$^+$ cells in our sequencing approach.

CD34$^+$ HSPCs were sort-purified from fetal bone marrow (BM), full-term UCB, and adult BM (**Fig. 1.1B** & **Fig. 2.4B**). Sorted cells from 7 samples (3 fetal BM, 2 full-term UCB, and 2 adult BM) were pooled for sequencing, then Demuxlet was used to annotate the sample identity of each cell and to identify and remove doublets. 5,183 single-cell transcriptomes, representing 747 median (range 602-898) cells per sample, were obtained. We then employed a combination of bioinformatic techniques to identify cells in similar differentiation states (**Fig. 2.14A,B**; see Materials and Methods for details). This approach led to the identification of 1,408 cells putatively in the earliest stages of hematopoiesis (i.e., hematopoietic stem cells and multipotent progenitor

cells, HSC/MPPs) and 2,583 progenitors differentiating at the ends of three distinct branches of hematopoiesis: 591 putative megakaryocyte-erythroid progenitors (MEPs), 789 granulocyte-monocyte progenitors (GMPs), and 1,203 common lymphoid progenitors (CLPs). Supportive of these cell type assignments, expression patterns of genes canonically associated with these cell types match our annotations (**Fig. 2.15**). Notably, we found that cell type annotation frequencies vary between ages and tissues (**Fig. 2.14C,D**; Chi-squared p-value < 2.2e-16 for all pairwise age comparisons). In particular, only 19 out of 1,203 cells annotated as CLPs originated from UCB samples and more than twice as many HSC/MPP-annotated cells originated from UCB (n = 2) or adult BM (n = 2) samples as were annotated from fetal BM samples (n = 3). These data provide evidence that the abundance of specific developmental intermediates is different between fetal BM, UCB, and adult BM samples, and specifically that CLPs may be of particularly low abundance in UCB.

To compare the transcriptomes of cells across ages, separate developmental stage scoring models were built for each HSPC subtype with sufficient numbers of UCB cells (i.e., HSC/MPP, MEP, GMP). Accuracy of the model for each cell type was confirmed to be high for non-transitioned (i.e., fetal) and fully transitioned (i.e., adult) cells that were not used in training the models (above 0.99 for all cell types; HSC: 0.9901, MEP: 0.9691, GMP: 0.9986) before developmental stage scores of cells were ultimately compared (**Fig. 2.14E-G**). Median scores for all cell types with substantial numbers of cells recovered in UCB were found to be intermediate between those of fetal and adult. The intermediate scoring of progenitor cells indicates that HSPCs in full-term UCB are at an intermediate state along the fetal-to-adult transition, and are thus transcriptionally unique from fetal and adult HSPCs.

Figures

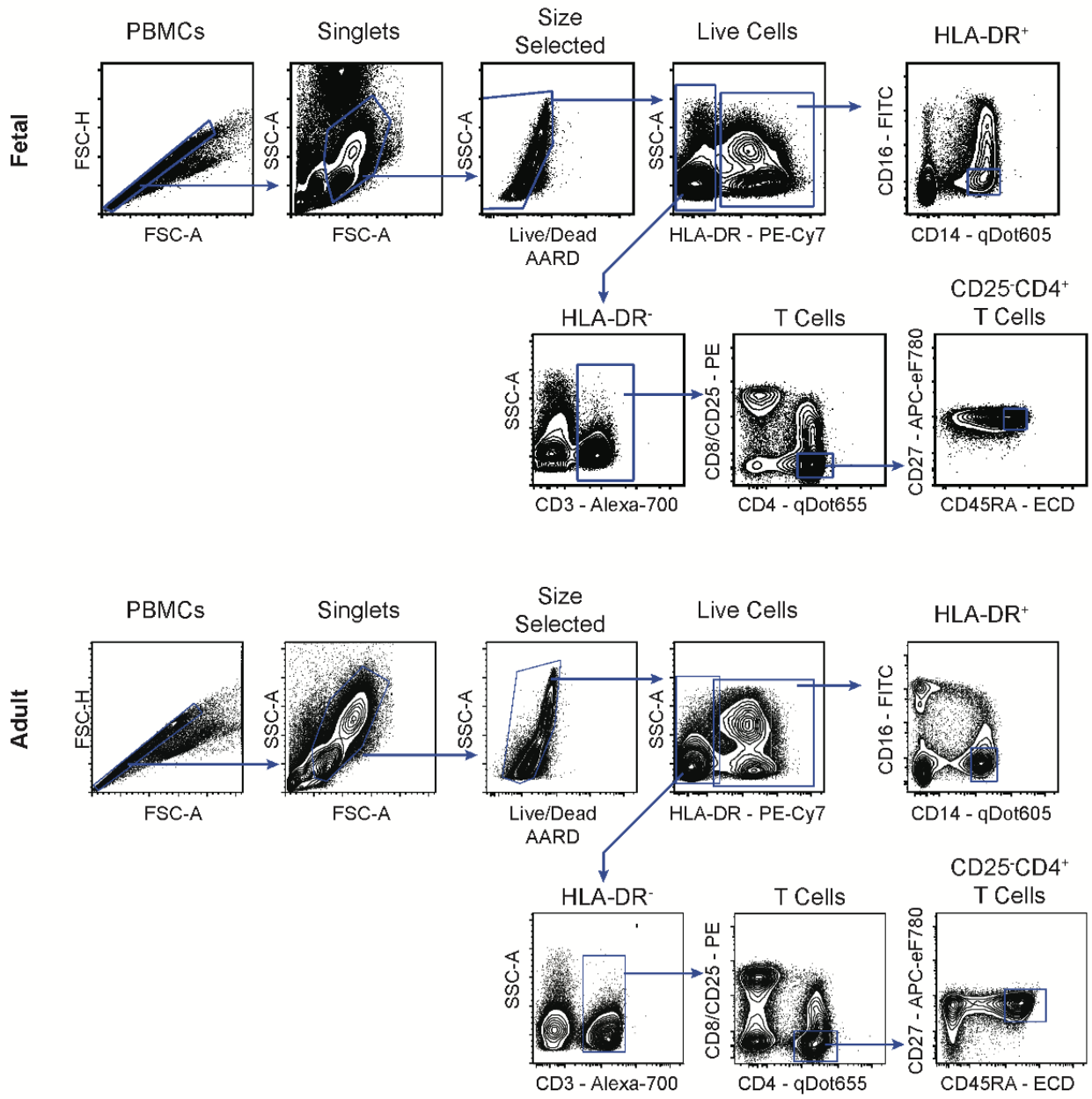


Figure 2.1: Microarray and qRT-PCR samples sorting strategy
Naïve CD4 T cells and classical monocytes were stained for FACS and sorted using the gating strategy shown. Plots shown reflect single representative samples for fetal and adult peripheral blood samples. Polygons represent approximate gates used, and arrows show the gating hierarchy.

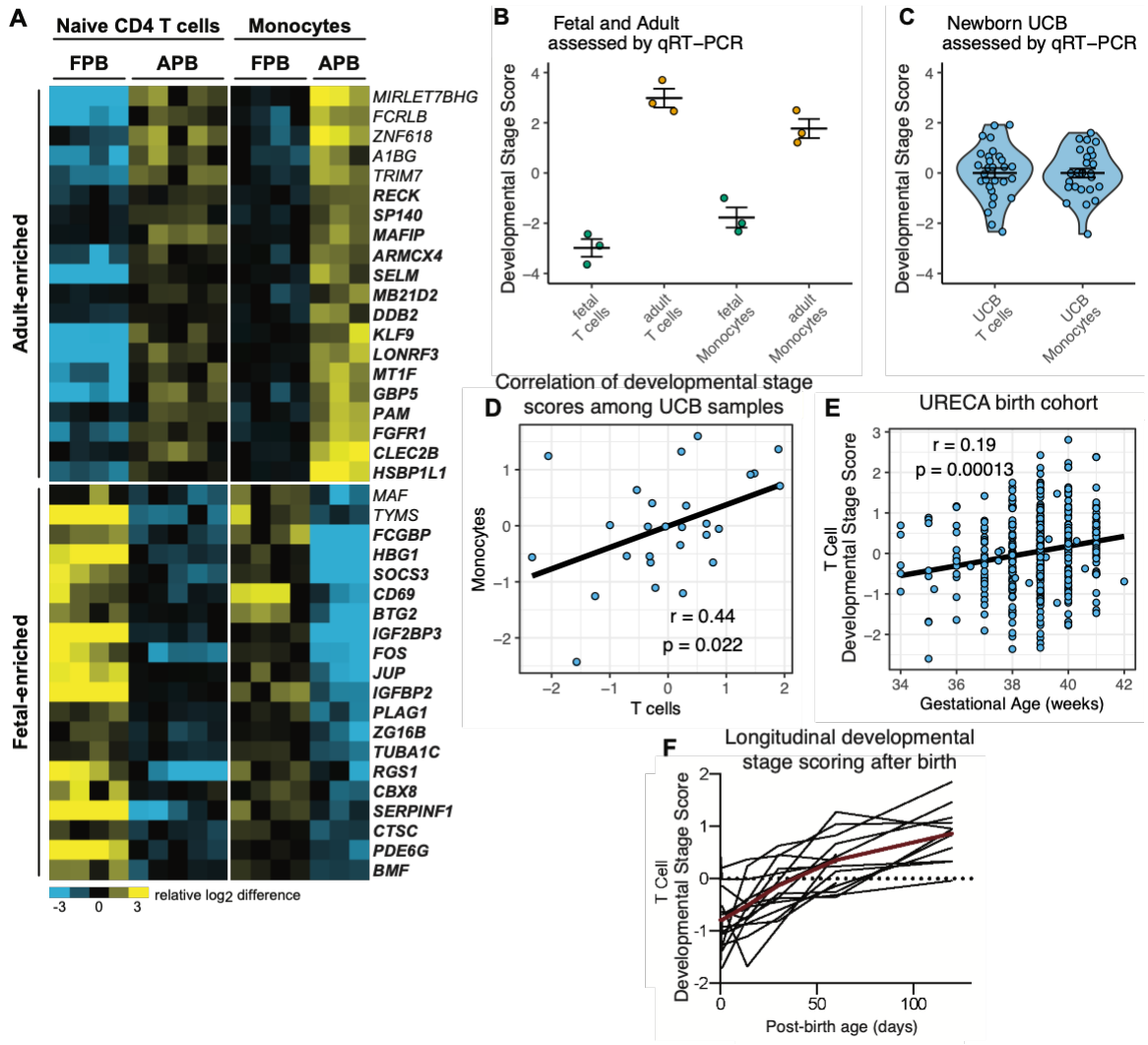


Figure 2.2: Population-level developmental stage scoring places UCB and infant immune cells intermediate between fetal and adult

Fetal and adult peripheral blood (PB) naïve CD4 T cell and classical monocytes were transcriptionally profiled by microarray. Genes differentially expressed across ages within both cell types were used to build a developmental stage score model which was then used to score fetal-to-adult transition progress of new samples.

A. Heatmap showing relative expression of top differentially expressed genes within individual samples. Bolded genes = used for developmental stage scoring.

B&C. Developmental stage scores of naïve CD4 T cells and classical monocytes from fetal and adult samples (B, n=3 each) or from UCB samples (C, T cells n=29. monocytes n=27). Bars = mean and standard error of the mean (SEM)

D. Scatterplot showing correlation between developmental stage scores of UCB naïve CD4 T cells and monocytes from the same samples (n=27). Pearson correlation coefficient (r), associated p-value and least squares regression line are shown.

E. Scatterplot showing correlation between developmental stage score and gestational age from a larger birth cohort for T cells (n = 405). Pearson correlation coefficient (r), associated p-value and least squares regression line are shown.

F. Developmental stage scoring of naïve CD4 T cell samples tracked longitudinally. Black lines = individuals (n=15). Heavier red line = mean value at a given time point.

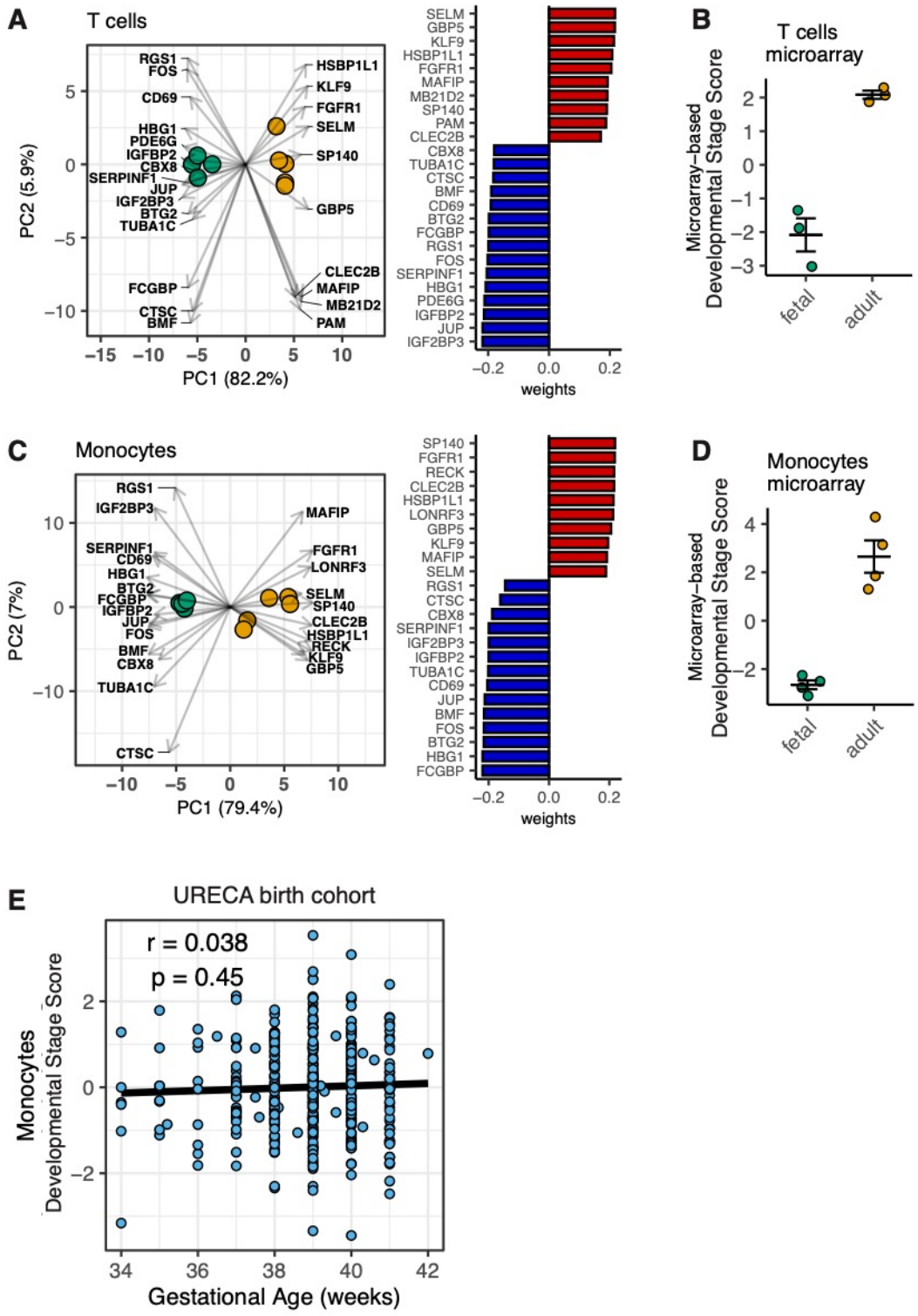


Figure 2.3: Population-level developmental stage score generation and initial validation
33 genes differentially expressed in peripheral naïve CD4 T cells and classical monocytes between fetal versus adult samples were identified and validated for qRT-PCR. Expression of these marker genes in new samples were then assessed by qRT-PCR. After filtering of probes and genes for a given experiment (Materials and Methods), principal components analysis

(PCA) of processed microarray data for these genes was run and PC1 loadings were used to weight expression of each gene in calculating developmental stage scores of samples.
A,C. Signature gene-focused PCA, left, and PC1 loadings for each gene, right, for T cells (A) and monocytes (C).
B,D. Developmental stage scores generated for fetal and adult T cells (B) and monocytes (D) from prior microarray analyses (Krow-Lucal et al., 2014; Mold et al., 2010).
E. Scatterplot showing correlation between developmental stage score and gestational age from UCB of a larger birth cohort for monocytes (n = 389). Line represents least squares linear regression. Pearson correlation coefficient (r) is shown along with the associated p-value.

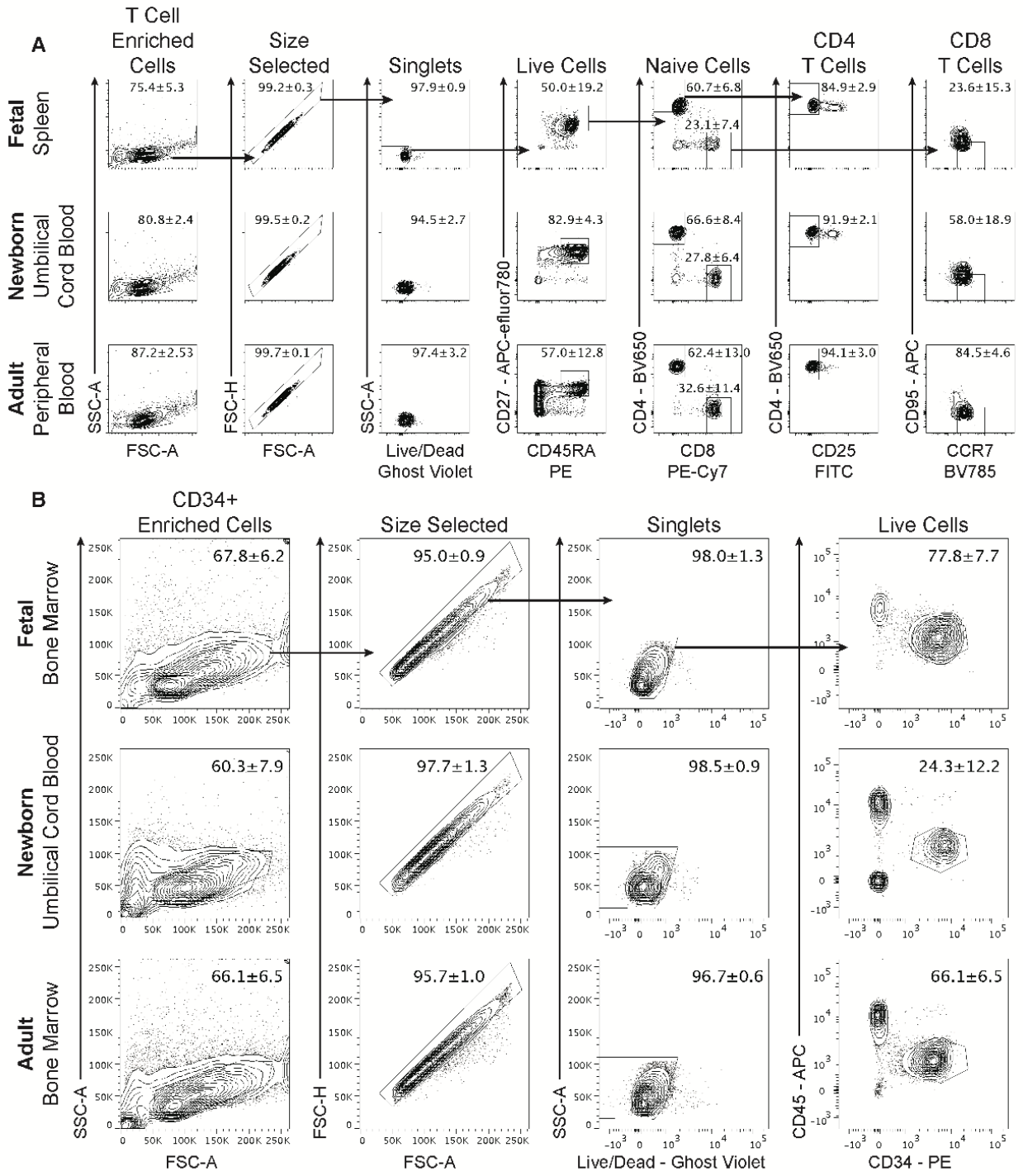


Figure 2.4: Sort strategy for single-cell and bulk RNA-seq
A. Fetal splenic, full-term UCB, and adult peripheral blood CD4 and CD8 T cells were sorted for single-cell and (CD4 only) bulk RNA-sequencing using the strategy shown, after magnetic bead-based T cell pre-enrichment. Plots shown reflect single representative samples for each age. Polygons represent approximate gates used, and arrows in the top example show the gating hierarchy. Numbers next to gates or in the top right of each plot represent mean and standard deviation across all samples of the given age (n = 5 for each age).

B. Fetal bone marrow, full-term UCB, and adult bone marrow CD34+ cells were sorted for single-cell RNA-sequencing using the strategy shown, after magnetic bead-based pre-enrichment. Plots shown reflect single representative samples for each age. Polygons represent approximate gates used, and arrows in the top example show the gating hierarchy. Numbers next to gates or in the top right of each plot represent mean and standard deviation across all samples of the given age (n = 3 for fetal, n = 2 for UCB and adult).

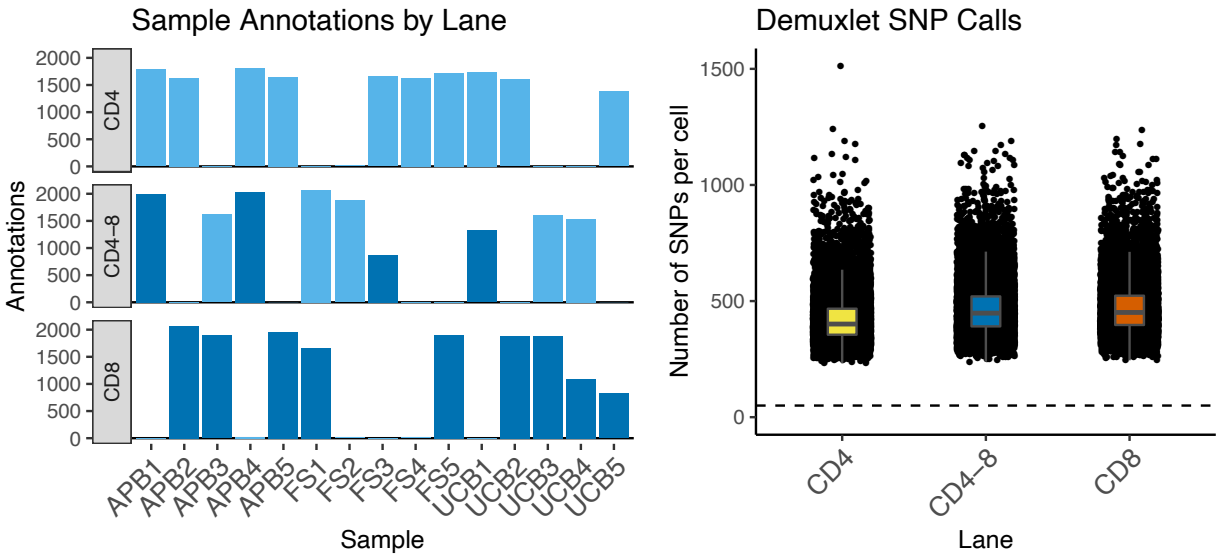


Figure 2.5: Demuxlet accurately identifies single-cell RNA-seq T cells' original samples
Naïve CD4 and CD8 T cells from 15 samples were combined across 3 lanes for droplet-based single-cell RNA-seq library generation. CD4 and CD8 T cells for each sample were run in separate lanes. After sequencing, Demuxlet (Kang et al., 2018) was used to identify original sample identity of sequenced cells based on SNPs captured within the mRNA transcripts.
A. Bar plots showing the number of annotations, per 10X lane, for each sample. Light blue = CD4. Dark blue = CD8.
B. Box plots representing the number of SNPs identified by Demuxlet for each cell. Dots represent individual cells. Dotted line, at 50, represents a baseline number of SNPs for high confidence sample annotation (Kang et al., 2018).

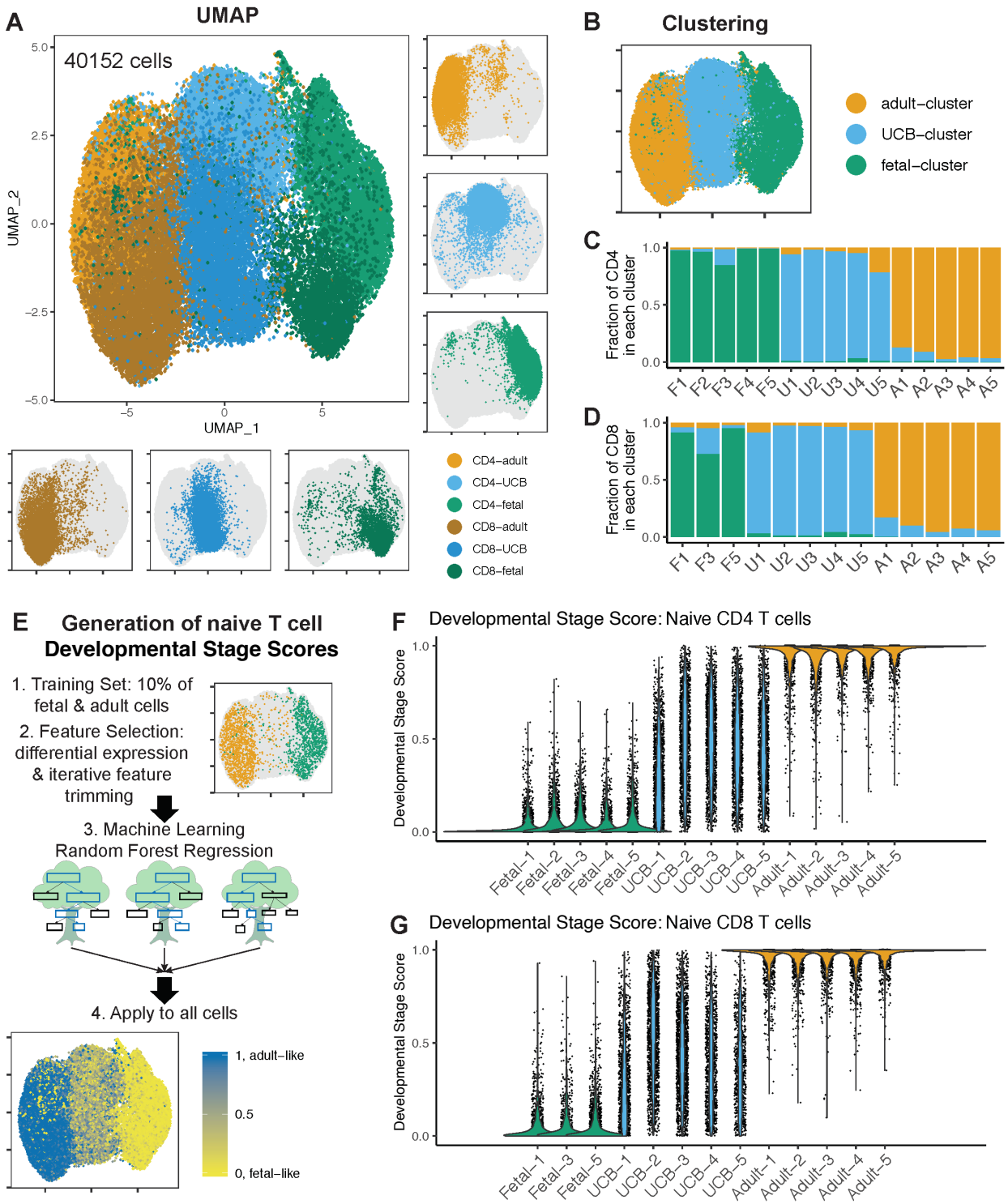


Figure 2.6: Single-cell-level developmental stage scoring places individual UCB naïve T cells intermediate between fetal and adult

Fetal splenic, full-term UCB, and adult peripheral blood naïve CD4 (n = 15) and CD8 (n = 13) T cell samples were profiled by single-cell RNA-sequencing.

A. UMAP plot showing distribution of cells of each age and CD4 versus CD8 lineages. Separate plots along the bottom and right show cells of individual identities.

B. UMAP plot showing Louvain clustering of single-cell profiles.

C,D. Quantification of CD4 (C) and CD8 (D) T cell clustering where each column represents a different sample. F=Fetal, U=UCB, A=Adult
E. Overview of single-cell developmental stage score model generation.
F,G. Developmental stage scoring of individual CD4 (F) and CD8 (G) naïve T cells for each sample. Points = individual cells.

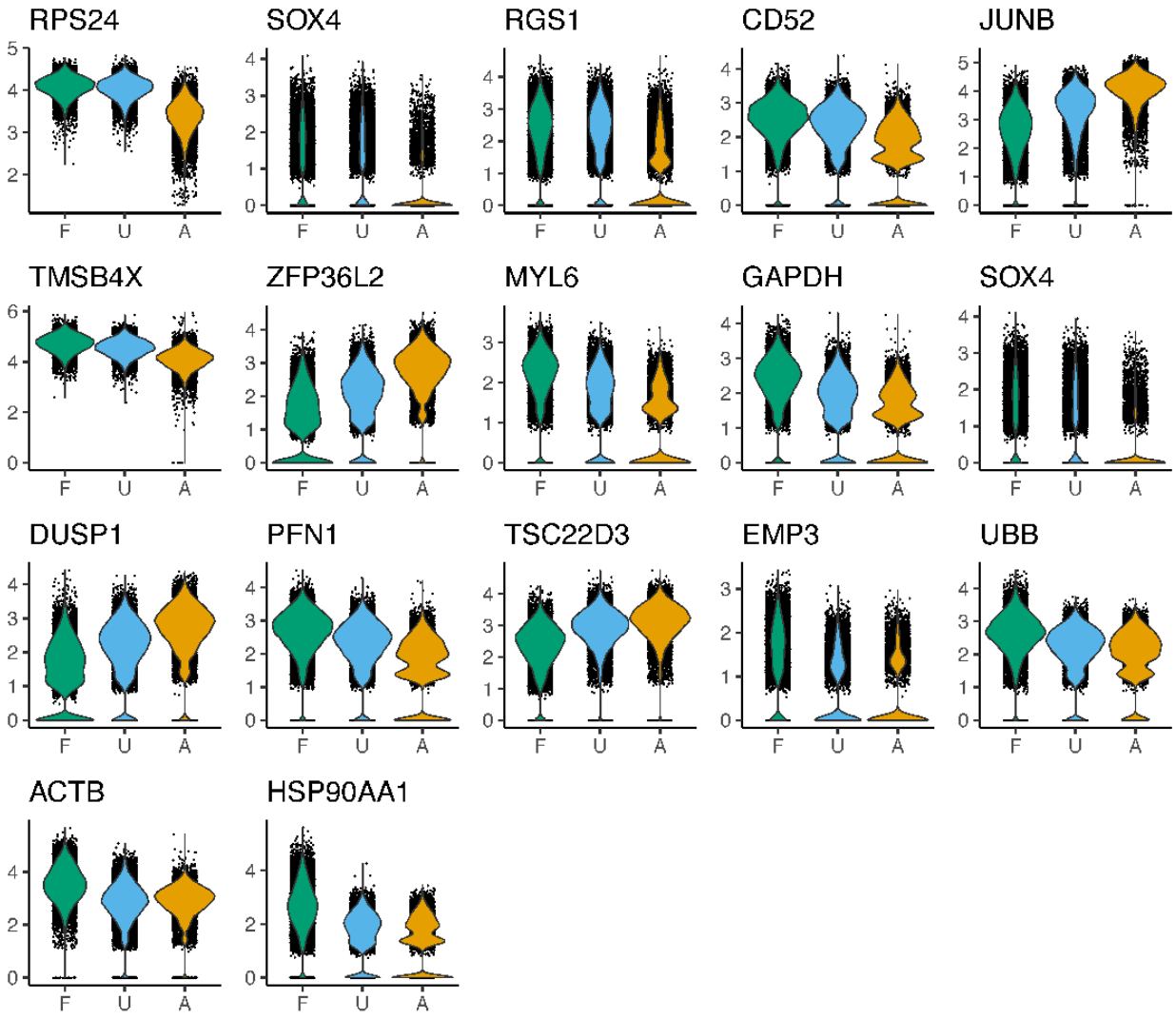


Figure 2.7: Single-cell RNA-seq expression profile of genes used in the single-cell developmental stage score

Violin plots representing log normalized expression, in fetal, UCB, and adult naïve T cells, of each gene used in generating single-cell developmental stage scores. Dots represent expression levels within individual cells. F=Fetal, U=UCB, A=Adult.

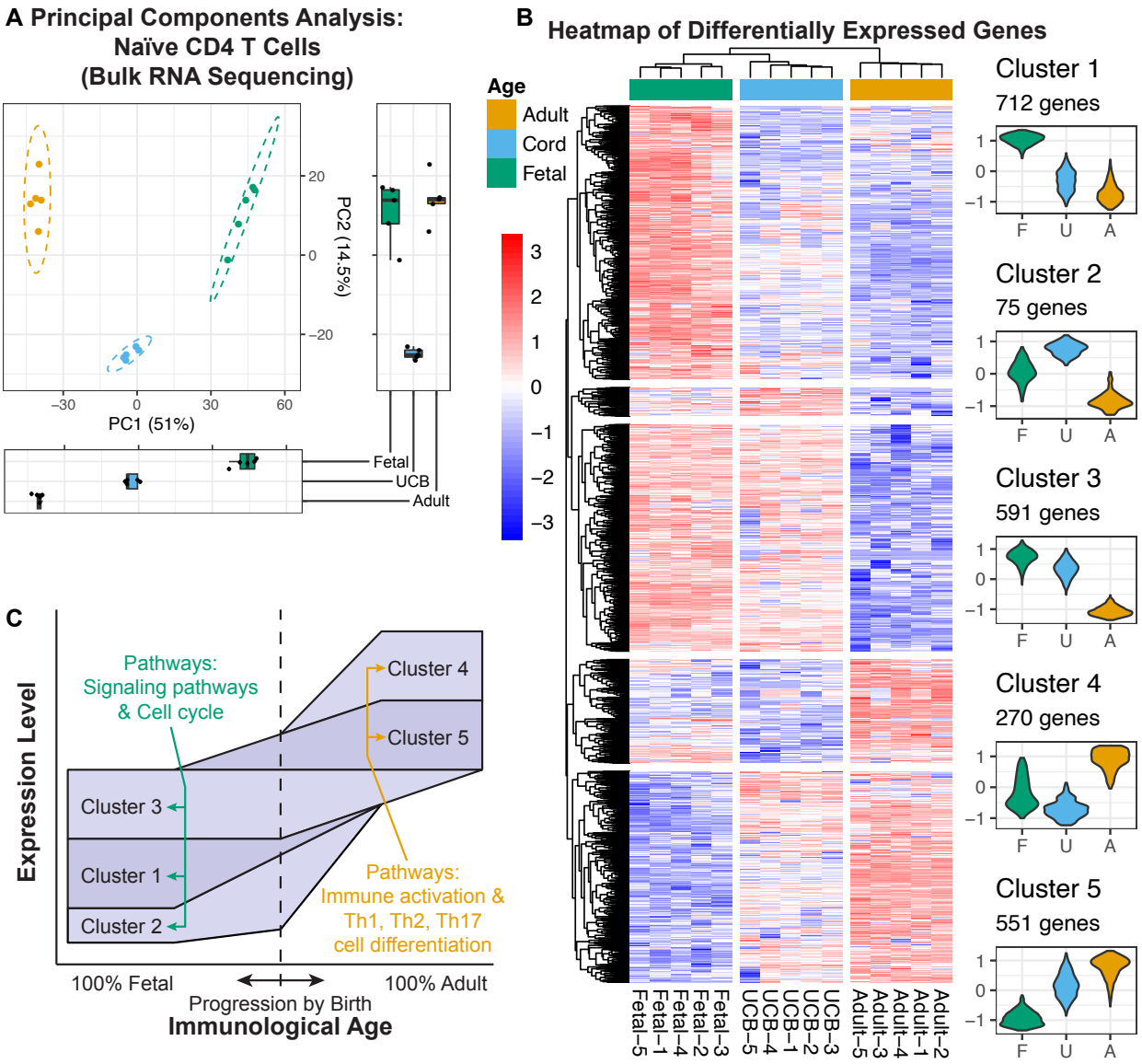


Figure 2.8: Discrete subsets of genes undergo fetal-to-adult transition with varied timing in naïve CD4 T cells

Fetal splenic, newborn UCB, and adult peripheral blood naïve CD4 T cells (n = 5, each) were profiled by bulk RNA-sequencing of 50 thousand cells per sample.

A. Principal components analysis (PCA) plot, with projections into PC1 and 2

B. Left, heatmap showing expression levels of all genes differentially expressed (FDR < 0.05, log2 fold change ≥ 1.5) between ages. Colors = relative, z-score, log-normalized expression across each gene. Clusters obtained from k means clustering with k=5. Right, Log-normalized expression values for each gene summarized, separately for each cluster, by mean z-score across each age. F=Fetal, U=UCB, A=Adult

C. Conceptual overview showing how the expression levels of genes in clusters, from the heatmap in B, fluctuate over the course of the fetal-to-adult transition with distinct timing. Heights of bars represent relative expression level at a given time. Pathways reflect a summary of the pathways enriched within fetal-associated cluster (3 and 4) and adult-associated clusters (1 and 2) genes.

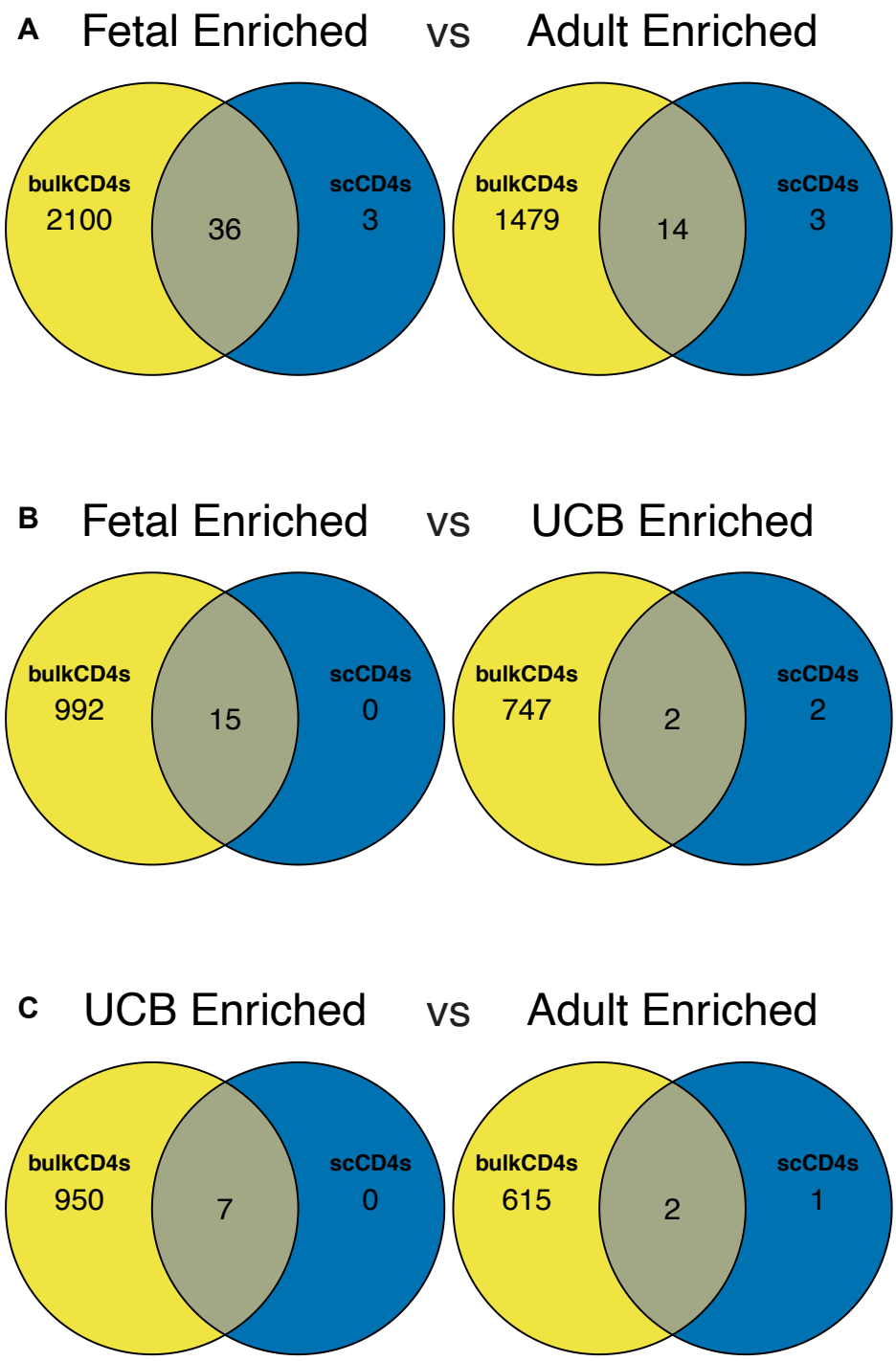


Figure 2.9: Comparison of genes differentially expressed between ages in naïve CD4 T cell single-cell versus bulk RNA-seq datasets
Fetal, UCB, and adult naïve CD4 T cells (n = 21,639 total) samples (n = 5 per age) were compared by both single-cell and bulk RNA-sequencing. For both datasets, differential expression analysis was performed between all pairwise age comparisons with cutoffs FDR < 0.05 and absolute fold change ≥ 1.8. Shown are Venn diagrams representing composition of the resulting gene lists for the fetal versus adult (A), fetal versus UCB (B), and UCB versus adult (C) comparisons.

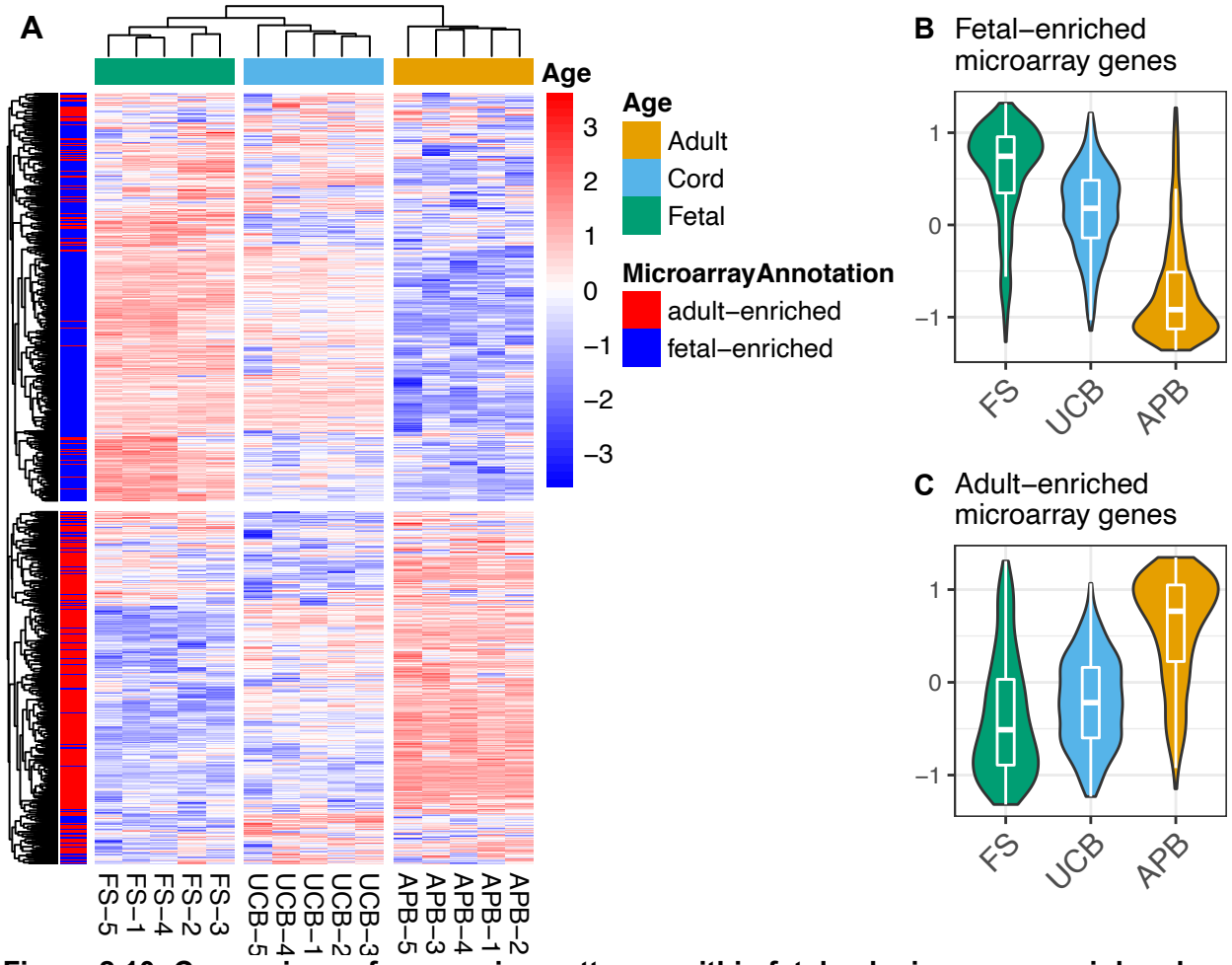


Figure 2.10: Comparison of expression patterns, within fetal splenic versus peripheral blood naïve CD4 T cell bulk RNA-seq, of genes identified in peripheral blood versus adult peripheral blood microarray analysis

Fetal and adult peripheral blood naïve CD4 T cell transcriptomes were initially compared by microarray analysis. Later, fetal splenic (FS), full-term umbilical cord blood (UCB), and adult peripheral blood (APB) naïve CD4 T cell transcriptomes were compared by bulk RNA-seq.
A. heatmap showing expression, within mixed tissue and blood RNA-seq samples, of genes annotated as differentially expressed between fetal and adult peripheral T cells in microarray analysis. Colors represent relative, z-score, log-normalized expression in RNA-seq. Clusters obtained from k means clustering with k=2. Gene annotations reflect gene annotations from peripheral blood microarray analysis: red = adult-enriched, blue = fetal-enriched.
B,C. Log-normalized RNA-seq expression values summarized separately for genes annotated as fetal-enriched (B) and adult-enriched (C) in microarray using the mean, across each age, z-scores for each gene.

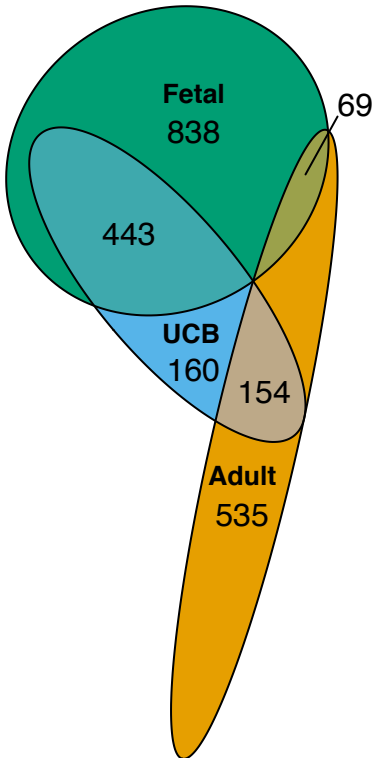
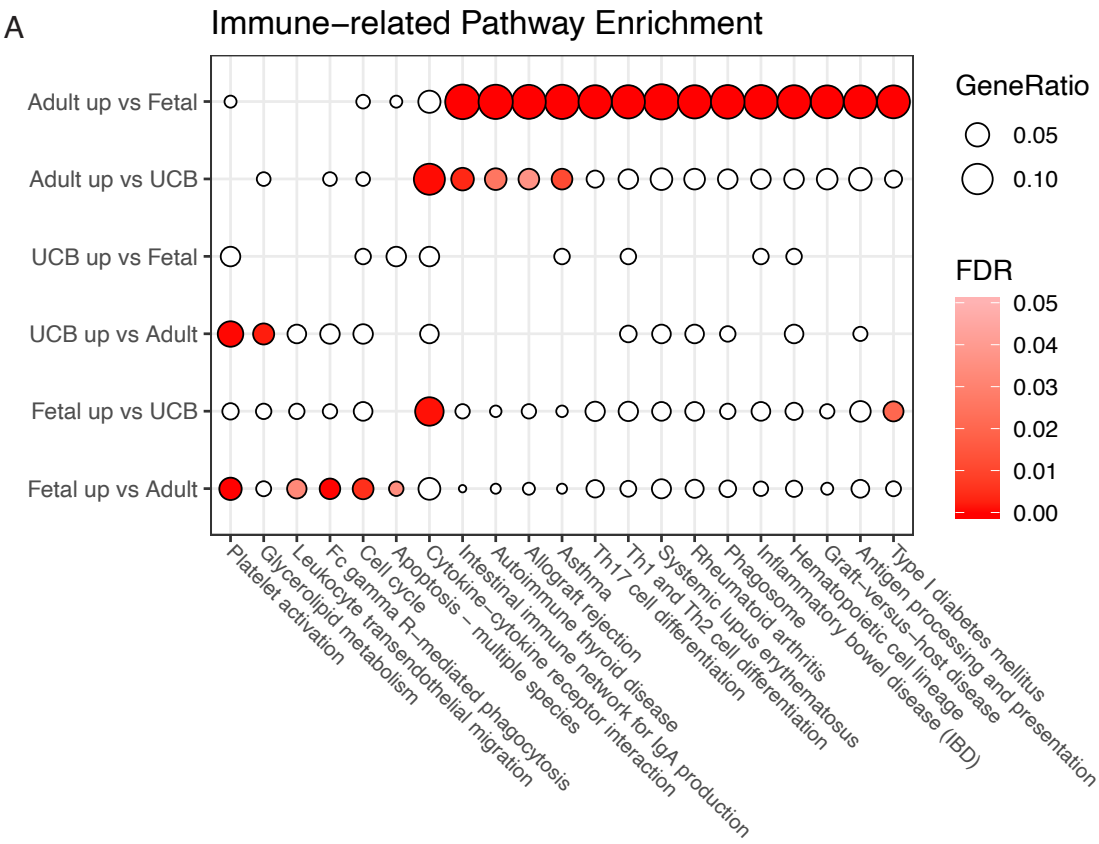


Figure 2.11: Compositional comparison of bulk RNA-seq upregulated gene lists
Fetal splenic, full-term UCB, and adult peripheral blood naïve CD4 T cell transcriptomes were compared by bulk RNA-seq. Pairwise differential expression analysis (FDR < 0.05, log2 fold change ≥ 1.5) was performed across all ages for all non-ribosomal and non-mitochondrial genes. Shown is a Venn diagram comparing the composition of total upregulated gene lists (the union of genes upregulated in each age versus either of the other two ages).


A
Immune−related Pathway Enrichment
Adult up vs Fetal
Adult up vs UCB
UCB up vs Fetal
UCB up vs Adult
Fetal up vs UCB
Fetal up vs Adult
Platelet activation
Glycerolipid metabolism
Leukocyte transendothelial migration
Fc gamma R−mediated phagocytosis
Cell cycle
Apoptosis − multiple species
Cytokine−cytokine receptor interaction
Intestinal immune network for IgA production
Autoimmune thyroid disease
Allograft rejection
Asthma
Th17 cell differentiation
Th1 and Th2 cell differentiation
Systemic lupus erythematosus
Rheumatoid arthritis
Phagosome
Inflammatory bowel disease (IBD)
Hematopoietic cell lineage
Graft−versus−host disease
Antigen processing and presentation
Type I diabetes mellitus
GeneRatio
0.05
0.10
FDR
0.05
0.04
0.03
0.02
0.01
0.00

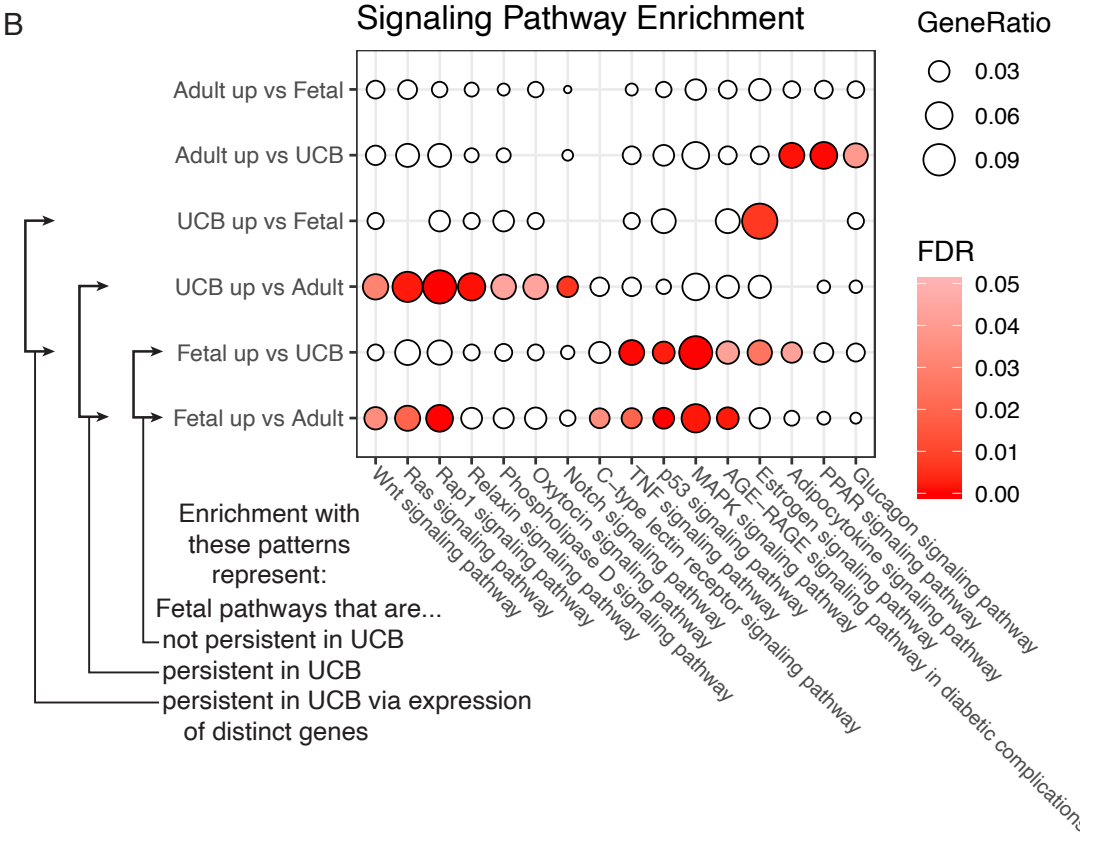

B
Signaling Pathway Enrichment
Adult up vs Fetal
Adult up vs UCB
UCB up vs Fetal
UCB up vs Adult
Fetal up vs UCB
Fetal up vs Adult
Wnt signaling pathway
Ras signaling pathway
Rap1 signaling pathway
Relaxin signaling pathway
Phospholipase D signaling pathway
Oxytocin signaling pathway
Notch signaling pathway
C−type lectin receptor signaling pathway
TNF signaling pathway
p53 signaling pathway
MAPK signaling pathway
AGE−RAGE signaling pathway in diabetic complications
Estrogen signaling pathway
Adipocytokine signaling pathway
PPAR signaling pathway
Glucagon signaling pathway
GeneRatio
0.03
0.06
0.09
FDR
0.05
0.04
0.03
0.02
0.01
0.00
Enrichment with these patterns represent:
Fetal pathways that are...
not persistent in UCB
persistent in UCB
persistent in UCB via expression of distinct genes

Figure 2.12: Distinct immune-related and signaling pathways are enriched within fetal, UCB, and adult naïve CD4 T cells
KEGG Pathway enrichment assessed within naïve CD4 T cell bulk RNA-sequencing differential expression gene sets. Enriched pathways are represented by FDR-corrected p-values (color) and the percentage of genes of a given set that are part of the given pathway (GeneRatio, size). Selected non-infection immune-related (A) and signaling (B) pathways are shown.

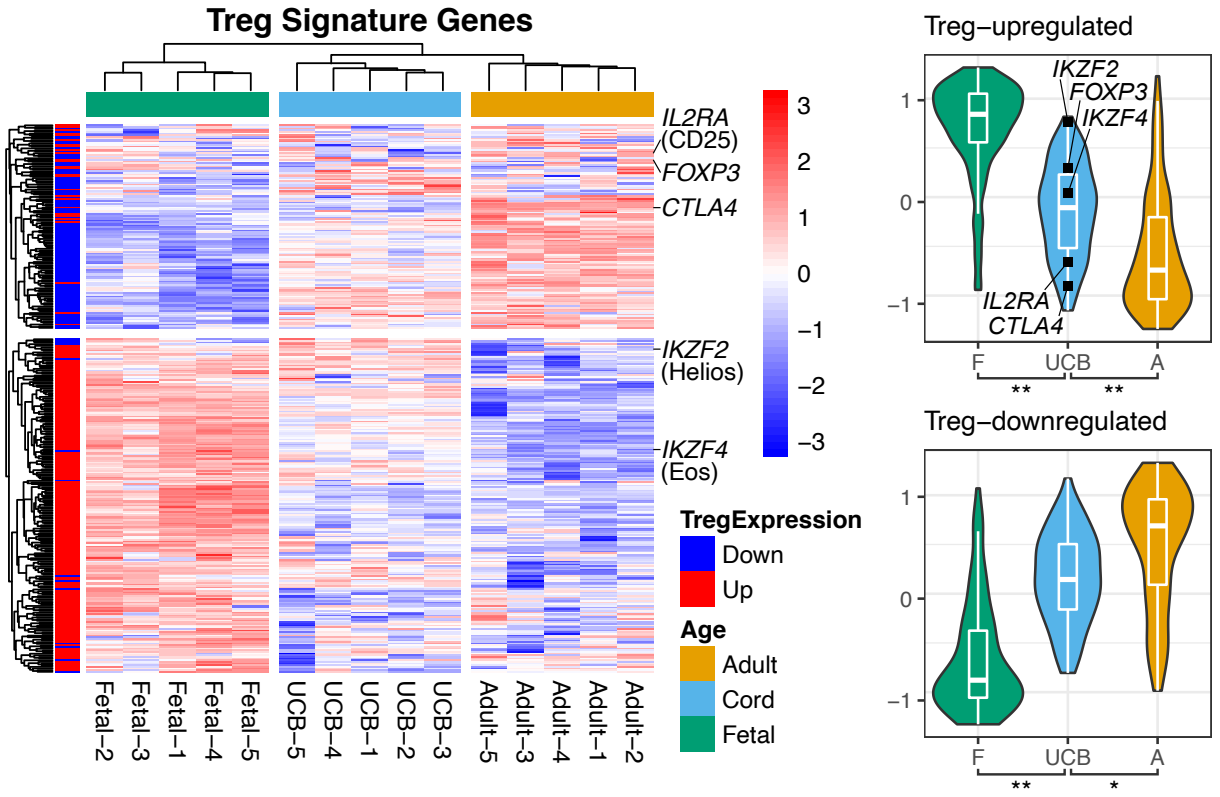


Figure 2.13: T_{reg} signature gene expression is partially maintained in UCB naïve CD4 T cells

A. heatmap showing expression levels of T_{reg} cell signature genes within the fetal, UCB, and adult naïve CD4 T cells characterized by bulk RNA-seq. Gene signature derived from comparison of genes differentially expressed between adult naïve T cells versus fetal and adult T_{regs} by Ng et. al. 2019. Location of selected T_{reg}-associated genes within the heatmap are highlighted. Colors represent relative, z-score, log-normalized expression. Clusters obtained from k means clustering with k=2. Gene annotations: blue=down-regulated, red=up-regulated in T_{reg} signature.

B,C. Log-normalized expression values summarized separately for T_{reg}-upregulated (B) and T_{reg}-downregulated (C) genes by mean z-score expression, across each age, for each gene. Expression level of particular T_{reg}-associated genes are highlighted within UCB samples. F= Fetal, U=UCB, A=Adult. P-values from Mann-Whitney U test, *$p < 1x10^{-7}$, **$p < 1x10^{-15}$.

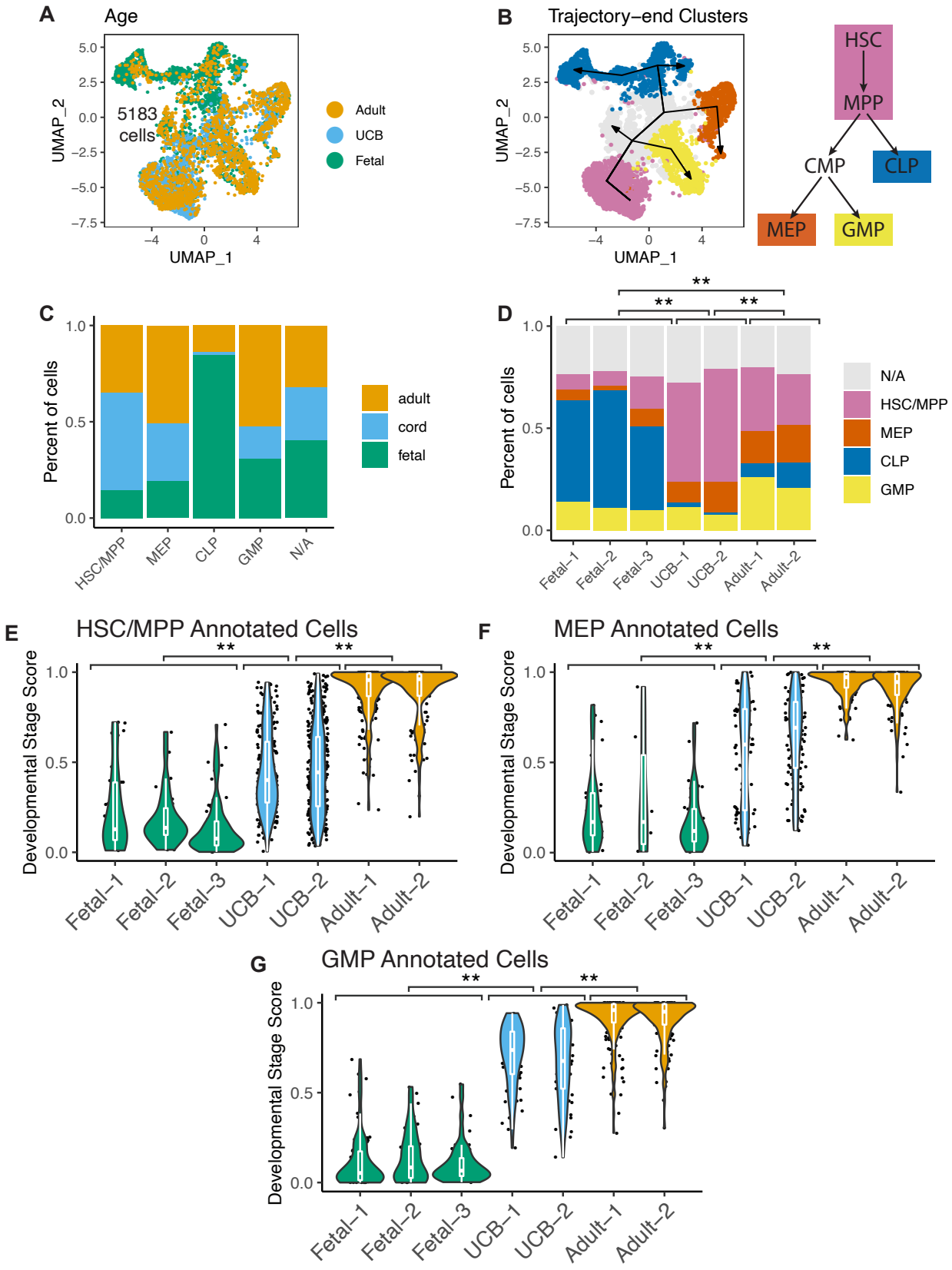

A
Age
5183 cells
UMAP_1
UMAP_2
Adult
UCB
Fetal
B
Trajectory–end Clusters
HSC
MPP
CMP
CLP
MEP
GMP
C
Percent of cells
adult
cord
fetal
HSC/MPP
MEP
CLP
GMP
N/A
D
Percent of cells
N/A
HSC/MPP
MEP
CLP
GMP
Fetal–1
Fetal–2
Fetal–3
UCB–1
UCB–2
Adult–1
Adult–2
E
HSC/MPP Annotated Cells
Developmental Stage Score
F
MEP Annotated Cells
G
GMP Annotated Cells

Figure 2.14: Single-cell developmental stage scoring places UCB HSPCs intermediate between fetal and adult

Fetal bone marrow (BM) (n=3), full-term UCB (n=2), and adult BM (n=2) $CD34^+$ HSPCs were profiled by single-cell RNA-sequencing. Fetal-to-adult transition phenotype of cells were scored, separately for each annotated cell type, via developmental stage.

A. UMAP plot showing cells of each age after batch correction

B. Left, UMAP plot showing final cell type annotations with inferred cluster-based differentiation trajectories overlaid on top. Right, tree diagram showing accepted hematopoietic differentiation trajectories leading to all cell types annotated in this dataset. HSC = hematopoietic stem cell, MPP = multipotent progenitor, CMP = common myeloid progenitor, CLP = common lymphoid progenitor, MEP = megakaryocyte-erythroid progenitor, GMP = granulocyte-monocyte progenitor.

C. Percentages of age identities of cells annotated as each cell type. N/A refers to cells not annotated in trajectory-end cell type clusters.

D. Percentages of cell type annotations per sample. P-values from chi-squared test, $**p < 10^{-15}$.

E,F,G. Developmental stage scoring of individual HSC/MPP (G) MEP (H) and GMP (I) cells for each sample. Points = individual cells. Fetal and adult cells used to build developmental stage score models are not included. P-values from Mann-Whitney U test, $**p < 10^{-15}$.

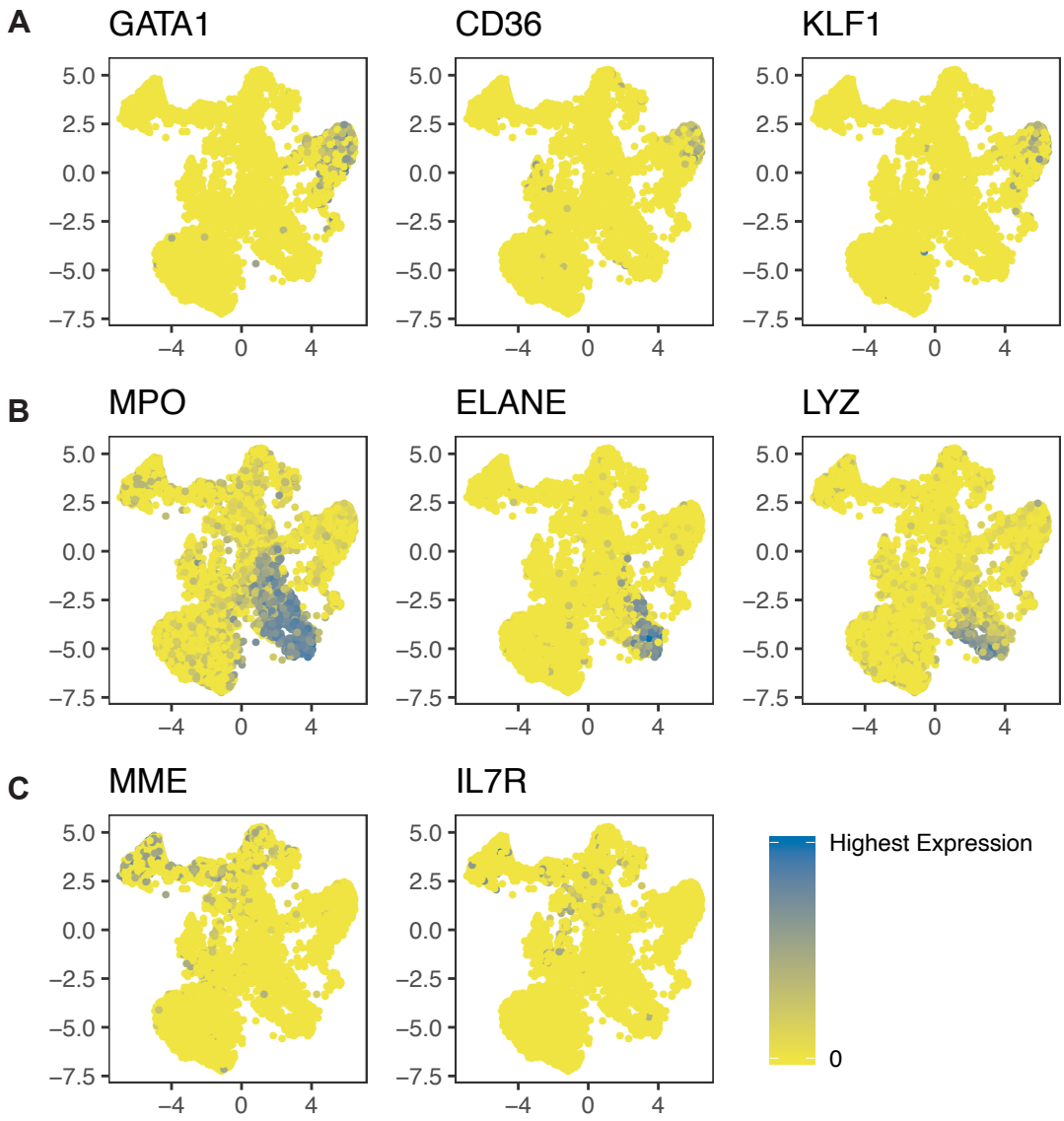


Figure 2.15: Differentiated HSPC cell type annotations express canonical genes
Individual HSPC transcriptomes characterized by single-cell RNA-seq were annotated as stem or progenitor cell types. Shown here is log normalized expression of canonical marker genes for differentiated progenitor cell annotations.
A: Megakaryocyte and erythrocyte progenitor (MEP) genes
B: Granulocyte-monocyte progenitor (GMP) genes
C: Common lymphocyte progenitor (CLP) genes

Chapter 3 – dittoSeq: Universal User-Friendly Single-Cell and Bulk RNA Sequencing Visualization Toolkit

Material for this chapter was modified from a manuscript currently under peer-review:

Daniel G. Bunis, Jared Andrews, Gabriela K. Fragiadakis, Trevor D. Burt, and Marina Sirota. "dittoSeq: Universal User-Friendly Single-Cell and Bulk RNA Sequencing Visualization Toolkit," manuscript under review, 2020.

Motivation

The tools available for analysis of sequencing based transcriptomic data are quite diverse. Unfortunately, so are the data structures for holding such data, yet conversion functions can be unreliable: they often do not convert all fields, and they can cause unintended alterations. Thus, there is need for analysis and visualization tools that are capable of working with all possible data structures. In the case of single-cell next generation sequencing (NGS) analysis—a rapidly growing field with many publicly available datasets and more being generated daily—the demand for universal tools is especially high considering the popularity of the Seurat (Butler et al., 2018; Stuart et al., 2019) analysis package, as well as the plethora of analysis packages in Bioconductor (Amezquita et al., 2020) that generally utilize the SingleCellExperiment (SCE) structure. In order to take advantage of all potential single-cell analysis datasets and all analysis tools that are available, therefore, users often must work with data in both the Seurat and SCE formats. Such work becomes significantly easier with visualization tools that are amenable to either structure. Here we present dittoSeq, a visualization tool that works with multiple data structures and thereby 1) minimizes the need for conversion between structures within complex analysis pipelines, 2) allows comparison across diverse analysis methods directly, and 3) allows standard, out-of-the-box, visualization of pre-processed data, no matter its original structure. Additionally, from the standpoint of novice users, 4) such a structure-agnostic visualization tool reduces the activation energy required for learning new analysis methods by ensuring that, regardless of the new package's required data structure, users would only need to learn the new analysis functions; the visualization methods would carry over. To our knowledge, dittoSeq is the first and only robust visualization tool which natively accepts both SCE and Seurat objects. Additionally, dittoSeq enables side-by-side visualization of single-cell and bulk RNAseq data, is color blind-friendly by default, and is powerfully flexible while remaining accessible and intuitive.

Software Description

dittoSeq is an R package available through Bioconductor via an open source MIT license. Full vignettes are available through Bioconductor, bioconductor.org/packages/dittoSeq/, and on GitHub, github.com/dtm2451/dittoSeq/.

Universal to the most common single-cell and bulk RNAseq data structures in R

dittoSeq was built with enabling side-by-side analysis of single-cell and/or bulk RNAseq data in mind. Thus its visualizations rely on a set of helper functions (gene, isGene, getGenes, meta, metaLevels, isMeta, getMetas, and getReductions) that can retrieve expression, metadata, and dimensionality reduction data directly from both Seurat and SCE objects. dittoSeq allows import of SummarizedExperiment (SE; the Bioconductor storage structure for bulk NGS data) and DGEList data through conversion of such objects into an SCE via an importDittoBulk function. The most common tools used for differential gene expression of bulk RNAseq data are edgeR (Robinson et al., 2010), which uses the DGEList structure, and DESeq2 (Love et al., 2014), which uses a structure that extends the SE structure. Thus, by accepting Seurat and SCE structures natively, and by providing conversion tools for SE and DGEList data, dittoSeq becomes universally applicable to all of the most common RNAseq data structures in R.

Diverse visualizations that are powerfully customizable

Visualizations supported in dittoSeq include dimensionality reduction plots (dittoDimPlot & dittoDimHex), scatterplots (dittoScatterPlot & dittoScatterHex), heatmaps (dittoHeatmap), and percent composition or expression across groups (dittoBarPlot & dittoPlot). All functions work for both bulk and single-cell data and allow customizations via simple, discrete inputs that are robustly-documented for ease of use. Examples of such customizations include: subsetting to certain cells or samples, changing sizes of data points and other representations, title

adjustments, automatic generation of annotations for heatmaps, overlay of trajectory analysis or density gradients onto dimensionality reduction plots, and making plots interactive via ggplotly (Sievert, 2020) conversion. Additionally, because most dittoSeq plot outputs are ggplot objects, extra layers and adjustments can be added manually via standard ggplot code. To make such manual alterations even easier, while also powering the publication-ready nature of dittoSeq plots, all functions also allow output of their underlying data via a 'data.out' input.

Color blindness-friendly by default

dittoSeq utilizes a modified version of the 8-color Okabe-Ito color panel –which is distinguishable by individuals with the most common forms of color blindness (Wong, 2011). By extending this panel to 40 colors, with lighter and darker repeats, we ensure that dittoSeq's default color set is equally accessible to most users, yet also amenable to the many color requirement of complex scRNAseq data. Additional tweaks and options add to the color blindness compatibility of the package as well: default legend adjustments (enlarged keys); optional use of shapes, letter-overlay, and/or faceting, in addition to coloring, when possible; and optional labelling or circling of groups in dimensionality reduction plots.

Example: Visualizing Expression of the Human Pancreas on the Single Cell Level

Figure 3.1 provides an example of how dittoSeq visualizations might be used to explore the cell type specific expression profiles of a human pancreas scRNAseq dataset (Baron et al., 2016) with visualizations that include: a UMAP plot with cell types labeled (**Fig. 3.1a**), a bar graph displaying cell type frequencies within each sample (**Fig. 3.1b**), a violin plot showing expression of a gene of interest across cell types (**Fig. 3.1c**), and a heatmap with metadata annotations (**Fig. 3.1d**). dittoSeq figures like these, each obtained via a single line of code, allow viewing of expression data in multiple ways to power both initial, iterative, data interrogation as well as the production of precisely-tuned, deliberately-labeled, publication-quality figures.

Figure

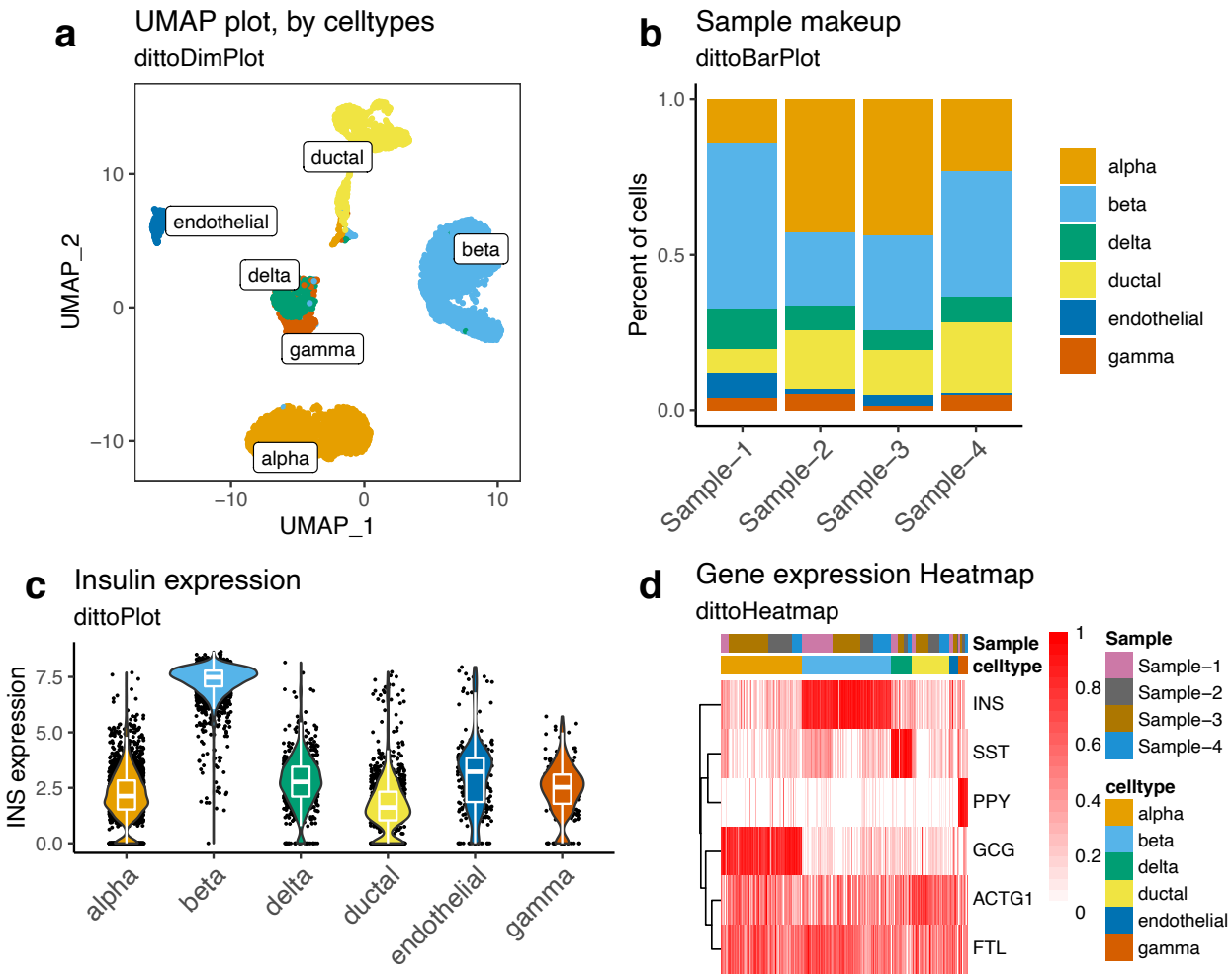


Figure 3.1: dittoSeq offers a plethora of highly-customizable visualization options. Data for these figures comes from Baron M et al. (2016), subset to only some of the most common cell types for simplicity, then processed with a standard Seurat workflow. Plots were made with (a) dittoDimPlot, (b) dittoBarPlot, (c) dittoPlot, and (d) dittoHeatmap.

Chapter 4 – Discussion

Material for this chapter was modified from a manuscript currently in-press:

Daniel G. Bunis, Yelena Bronevetsky, Elisabeth Krow-Lucal, Nirav R. Bhakta, Charles C. Kim, Srilaxmi Nerella, Norman Jones, Ventura F. Mendoza, Yvonne J. Bryson, James E. Gern, Rachel L. Rutishauser, Chun Jimmie Ye, Marina Sirota, Joseph M. McCune, Trevor D. Burt. "Single-cell mapping of progressive fetal-to-adult transition in human naive T cells," in-press with Cell Reports, 2020.

It is clear that several uniquely fetal immune populations can be generated only by progenitors that are present during fetal development, and the interpretation of earlier experiments in mice, birds, and humans favored the possibility that ontogeny of the hematopoietic system is regulated by distinct layered waves of unrelated HSCs and their progeny (Hadland & Yoshimoto, 2018; Herzenberg & Herzenberg, 1989; Ikuta et al., 1990). Here, we have studied the gene expression programs of monocytes, T cells, and HSPCs at varying time points during human ontogeny to characterize the progress of fetal-to-adult immune transition at the time of birth. While the present study is not the first to employ a system which scores an immunological age of human samples – Alpert et al. recently developed an IMM-AGE score based on immune cell frequencies in adults (ages 20-96) – here we uniquely developed cell type-specific models of samples' age- or developmental- associated differences (termed "developmental stage score") for directed application to sorted cell populations. We applied our system to fetal (18-23 gest. wks.), full-term newborn, infant, and adult (ages 27-53) T cells, monocytes, and hematopoietic progenitors. Through single-cell developmental stage score characterization of naïve T cells and HSPCs, we find evidence that human immune ontogeny instead follows a progressive, and not a layered, pattern of transition, during late fetal development (**Fig 4.1**).

A bulk developmental stage scoring model initially revealed that there is a high degree of inter-individual variability at birth, and that variability persists even while developmental stage scores become more adult-like during the first several months of life. While these results do not distinguish between putative models of fetal-to-adult transition, the observation of inter-individual variability in developmental stage score does suggest that some newborns are born with a more fetal-like immune system whereas others are born with a more fully transitioned, adult-like immune system and that these differences continue after birth. These findings are consistent with clinical observations that several years are required for the acquisition of robust, mature immune responses (Dowling & Levy, 2014) and suggest that certain characteristics of the fetal immune system may persist for a considerable period of time after birth. It is well established that some

infants produce less effective responses to vaccination or are more susceptible to infection than others, even in the absence of an identifiable primary immunodeficiency (Borghesi et al., 2017; Newport et al., 2004; Siegrist, 2001). Our findings may suggest that these infants are born with a more fetal-like immune system. Assessing this possibility would require employing developmental stage scores as a tool to stratify infants in studies of immune responses, including vaccine responses, in early life. Inter-individual variability in the transition from fetal-to-adult immunity both before and after birth is likely to be influenced by diverse factors such as “clock” gene expression in HSPCs, heritable genetic traits, microbial colonization of the infant mucosa and skin, and/or nutritional state (Madan et al., 2012; Newport et al., 2004; Palmer, 2011). Of note, the UCB samples studied here were obtained from clinical cohorts in which exhaustive demographic data collection and laboratory screening were not always available. Therefore, the observed inter-individual variability in gene expression may also be influenced by uncontrolled factors, including environmental exposures, maternal medication use, and undiagnosed infection.

To address the question of whether the developmental scoring of individual samples reflected the averaged signal of two distinct cell populations (i.e., fetal vs. adult) in bulk cell lysates or the actual per-cell state of a unimodally distributed population (**Fig. 1.5**), we turned to developmental stage scoring with single-cell resolution. Using this approach, we demonstrate that individual UCB T cells received unimodally-distributed intermediate developmental stage scores and were not composed of a mixture of cells with fetal or adult transcriptional identities. It is important to note that, while most developmental stage score markers denoting extremes of fetal or adult identity are intermediate in expression at the single cell level, some are skewed toward fetal (e.g., RPS24) or adult (e.g., HSP90AA1) expression patterns (**Fig. 2.7**). As these findings occur with the same pattern in the majority of cells analyzed, such a phenotype cannot be explained by a mixture of fetal-layer and adult-layer cells, and instead indicate a transition that includes progression of distinct genetic programs, each with different timing. Further, variation between samples was noted as a shift in the modes of UCB samples’ per-cell scores, rather than

via differences in the abundance of cells with particular ranges of scores—such as very low scoring (i.e., fetal-like) or very high scoring (i.e., adult-like) cells—as we would have otherwise expected if the fetal-to-adult transition occurred via layering of distinct cell populations. Though these data represent a “snapshot” in time along the course of development, they suggest that UCB cells are captured at a single point along a relatively uniform and progressive fetal-to-adult transition. Future analyses including additional intermediate gestational ages could potentially clarify whether changes truly occur progressively between the timepoints we have demonstrated here or, conversely, via multiple step-wise transitions between dominant fetal and newborn, then newborn and adult cell populations.

The findings presented here demonstrate that UCB naïve T cells represent a transitional population with transcriptional features that are both distinct from and intermediate between fetal and adult naïve T cells. Thus, the unique nature of immune responses in neonates and young children, including blunted responses to infection and vaccination, may be related to the transcriptional and functional state of these, and possibly other, immune cells. Further mechanistic studies to understand neonatal transcriptional programs will not only provide insight into early-life immune responses but may ultimately also suggest druggable targets to enhance immune maturation and shift neonatal T cells toward adult-like protective immune responses. Here, unbiased pathway analysis demonstrated that UCB T cells are enriched for both Wnt and Notch signaling pathways compared to adult cells. It should be noted that the proportion of maternally-derived cells in UCB has been shown to be fewer than 0.4% (Mold et al., 2008; Opstelten et al., 2019), thus it is unlikely that maternal cells contributed significantly to these analyses. Notch signaling plays a pivotal regulatory role in multiple aspects of T cell development and lineage differentiation (Vijayaraghavan & Osborne, 2018). The Wnt signaling pathway has been shown to promote differentiation of T_{fh} cells, which are crucial for effective vaccine responses (Loosdregt & Coffer, 2018), and pharmacological inhibition of the Wnt/β-catenin signaling pathway modulates fate decisions between self-renewal and differentiation of long-lived

central memory and stem cell-like memory CD4 T cells (Mavigner et al., 2019). The oxytocin signaling pathway was also upregulated in UCB T cells and, while oxytocin is involved the physiology of parturition, it is expressed by murine fetal and newborn thymic epithelial cells and may play a role in thymocyte development and central tolerance (Geenen et al., 2000).

In several trials, UCB transplantation has been shown to result in overall lower rates of graft-versus-host disease (GVHD) compared to adult BM transplantation (Ballen et al., 2013; Merindol et al., 2011). Direct evidence of a salutary influence of T_{regs} in HSPC transplantation is clearly demonstrated in recent clinical trials where co-transplantation of UCB-derived T_{regs} resulted in a marked reduction in GVHD (Blazar et al., 2018; Elias & Rudensky, 2019; D. H. McKenna et al., 2017). In light of findings presented here that UCB T cells retain a partial T_{reg}-associated transcriptome (**Fig. 2.13**), that variable bulk transcriptional signatures were detected in two distinct hematopoietic lineages (lymphoid and myeloid; **Fig 2.2**), and that fetal HSPCs give rise to T cells with transcriptional and functional characteristics of fetal T cells (Mold et al Science, 2010), we analyzed $CD34^+$ HSPC cell populations with cell type-specific single cell transcriptional developmental stage score analysis to determine whether progenitor cell populations also undergo gradual progressive transition from fetal to adult identity. We found that the developmental age signatures of HSC-MPPs, MEPs, and GMPs from UCB were distributed between fetal and adult cells of the same lineage, suggesting that the intermediate nature of T cell progeny may be reflective of (and potentially programmed by) the intermediate development stage of the HSPCs from which they arise. Along these lines, a recent study comparing naive T cells recovered two months after T-replete umbilical cord blood transplant (UCB-T) versus T-replete bone marrow transplant (BM-T), with naïve T cells from control fetal spleen, newborn UCB, and post-natal (7-40 years old) peripheral blood donors, found that T cells arising from UCB-T were more similar to those of control UCB while those arising from BM-T were more similar to those of post-natal peripheral blood (Hiwarkar et al., 2017). It is unknown whether such results were reflective of new T cells arising from transplanted hematopoietic progenitors versus

expansion of remnant donor T cells. In either case, the retention of newborn-like and adult-like phenotypes in a transplant experiment, in which contributions from environment, diet, tissue, and hormonal milieu are significantly reduced, is consistent with the conclusion that our results characterize cell-intrinsic, developmental stage-associated, phenotypes.

In single-cell analysis of HSPCs, the frequency of HSPC developmental lineage intermediates was variable between fetal bone marrow, newborn UCB, and adult bone marrow samples. In particular, very few CLP cells were annotated within UCB samples. The relative dearth of intermediate progenitors that we observe in UCB compared to adult BM may provide an explanation for clinical observations of slower engraftment of T cells after UCB transplantation compared to adult BM transplants (Komanduri et al., 2007; Servais et al., 2017). Our data suggest that, compared to adult BM, UCB has a smaller population of CLP precursors that can migrate directly to the thymus, In that case, generating a robust population of nascent donor-derived T cells from UCB transplants would be delayed by the need for HSC/MPP cells to transit through the CLP stage prior to initiating thymopoiesis.

Although our data are consistent with a model of gradual, progressive change from fetal to adult immunity at the level of individual cells around the time of birth, multiple limitations associated with human immunological research must be acknowledged. The expected variability inherent in human studies, as revealed in our survey of a healthy full-term birth cohort, raises the concern that a limited number of UCB transcriptomes (n=5 for T cells, n=2 for HSPCs) may not be reflective of the greater human population. Though this concern is considerably mitigated by the consistent intermediacy between fetal and adult samples of UCB samples that were subjected to single-cell developmental stage scoring, it nonetheless underscores the need to extend these findings to a larger and more diverse population of human subjects in the future. It must be noted that tissue-, environment-, diet-, and hormonal milieu- associated differences unavoidably exist between our samples and that such differences would be aggregated, inextricably, alongside age-based differences in our transcriptional data. Due to practical limitations, we primarily used T cells

isolated from spleen for fetal timepoint, from umbilical cord blood for newborn timepoint, and from peripheral adult peripheral blood for adult timepoint. Yet, we were able to compare gene expression in these tissues using a signature derived from microarray profiling of fetal peripheral (umbilical cord) blood to adult peripheral blood, and confirmed that expression patterns remained consistent for these genes in RNAseq comparison of fetal spleen versus adult PB, and that newborn UCB sample T cells exhibit an intermediate expression profile for these genes as well (**Fig. 2.14**). Whereas UCB has often been used for assessment of newborn immune cell phenotypes, it has been shown recently that immune system phenotypes can change drastically between those measured in UCB versus one week later in peripheral blood (Olin et al., 2018). Though the overall increase in mean bulk developmental stage scoring of at-birth UCB versus two-week-old peripheral blood naïve T cells was about the same as the difference between two- versus four-week-old peripheral blood (**Fig. 2.2F**), it will be important to follow up in newborn peripheral blood samples to determine whether Notch signaling, Wnt signaling, and other pathways identified as expressed in UCB remain expressed in the days and weeks after birth, and whether expression of these pathways remains enriched in newborn naïve T cells from babies delivered by Caesarean section in the absence of labor.

In conclusion, we provide evidence at the level of single cells that, in humans, T cell populations transition along a continuum that is characterized by progressive downregulation of fetal genes and upregulation of adult genes during the course of ontogeny from midgestational fetal development to birth and into adulthood (**Fig 4.1**). We find that the transcriptomes of newborn T cells and hematopoietic progenitors possess features that are unique from, and intermediate between, those of their fetal and adult counterparts, and we identify particular pathways enriched in UCB T cells that warrant further study. While our findings do not rule out the concept of “layered” hematopoiesis as invoked in other species (Havran & Allison, 1988; Hayakawa et al., 1985; Herzenberg & Herzenberg, 1989; Ikuta et al., 1990; Jotereau & Le Douarin, 1982; Kantor et al., 1992; Lalor et al., 1989; Le Douarin & Jotereau, 1975; Montecino-Rodriguez et al., 2006, 2016,

2018; Ramond et al., 2014), and at earlier timepoints in humans (Ginhoux et al., 2013; Hadland & Yoshimoto, 2018; Montecino-Rodriguez & Dorshkind, 2012; Stamatoyannopoulos, 2005; Tieppo et al., 2020), they demonstrate that layering of distinct fetal-like and adult-like cell populations is not present at the time of birth in human αß T cells.

Future Directions

One method for further study of how the unique functional properties of UCB HSPCs may affect the outcomes of hematopoietic transplants, would be to assess differences in donor-derived immune cell progeny after transplantation in humanized mice. Single-cell RNA sequencing of bone marrow and secondary lymphoid tissues from humanized mice, after transplantation of CD34+ HSPCs from fetal, newborn, or adult sources, could be utilized to study human immune cell progeny of diverse lineages in a single experiment. Developmental Stage Score analysis of resulting cellular profiles could be used to reveal if transcriptional states of distinct lineages of immune cell progeny reflect the transcriptional states of the original donor material. Additionally, individual immune cells could be extracted from such mice for functional analyses such as the assessment of T_{reg}-differentiation potential of T lymphocytes or antigen presentation machinery upregulation of classical monocytes after *in vitro* stimulation.

Further study of the unique properties of immune cells of all types in the newborn and young infant will clarify whether or not a layered fetal-to-adult transition underlies the ontogeny of other immune cells. Of particular interest will be studies to determine whether the interindividual variation in the progress of fetal-to-adult transition by the time of birth, noted here in both naïve T cells and monocytes (**Fig 2.2**), has a significant effect on the nature of immune responses to vaccines and infections in early life. One idea would be to collect blood and antibody titers, both at birth as well as after vaccination, and assess whether more adult-like baseline Developmental Stage Scoring correlates with higher antibody titers after vaccination. Another important question is whether agonists or antagonists of Notch or Wnt signaling pathways, pathways which are highlighted here as enriched within UCB samples' compared to adult samples' transcriptomes and which are known to play roles in T cell differentiation towards specific fates, might have potential utility as adjuvants for vaccination. Initial steps toward such a goal would be to explore *in vitro* whether such molecules might promote differentiation of

newborn naïve CD4 T cells towards pro-inflammatory fates, and if so, to follow-up *in vivo* in newborn or weeks-old mice.

Such studies may lead to more focused stratification of children according to their individual risk for infection and for inadequate vaccine responses. They may lead to the design of precision therapeutic approaches to boost newborn immunity and to improved public health outcomes in this vulnerable population. They may also highlight methods by which we might promote tolerance in adult immune systems, which could influence future treatments for autoimmunity or other sources of unwanted inflammation, such as organ transplantation.

Figure

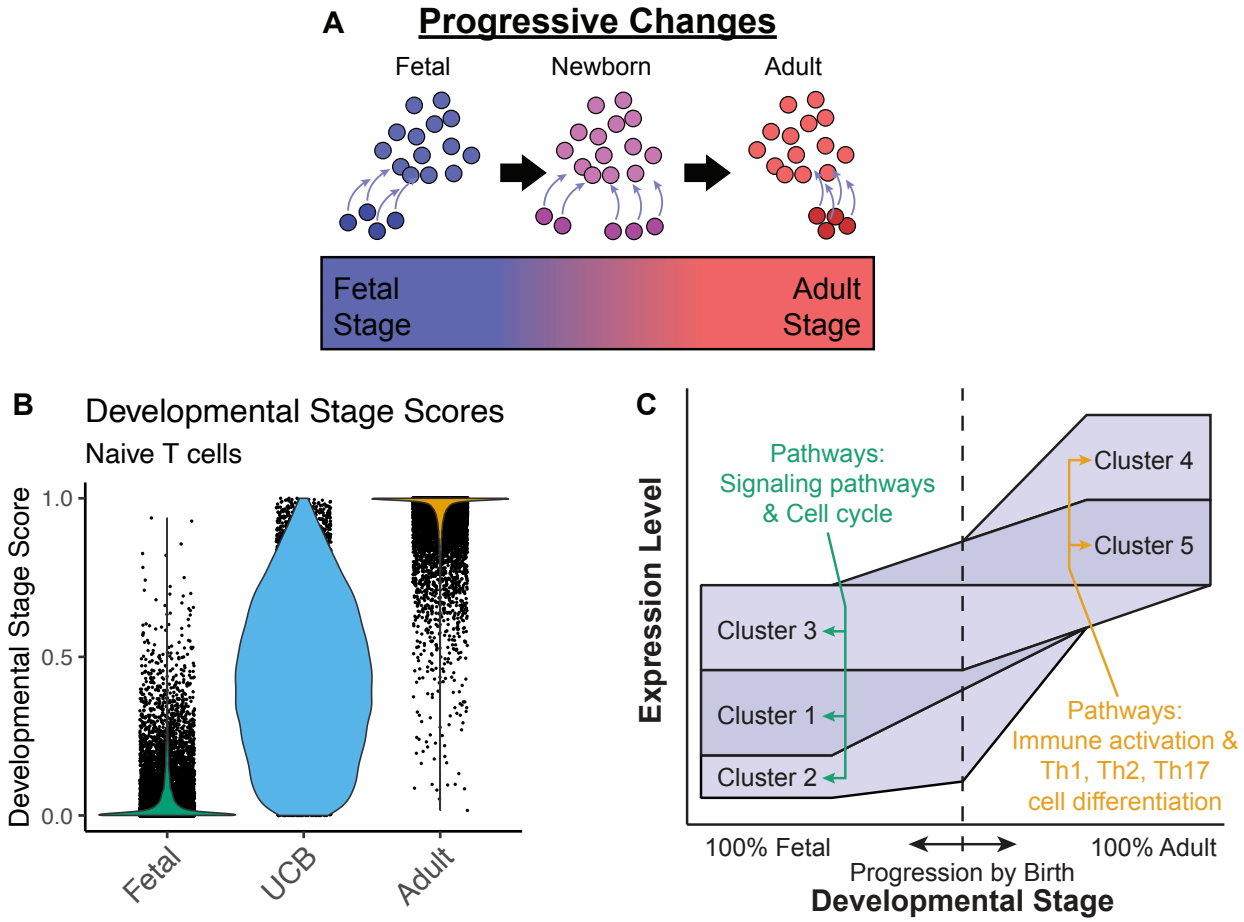


Figure 4.1: Overall Conclusions
We find that the fetal-to-adult transition is best modeled by progressive changes (A) largely based on facts that Developmental Stage Scores of individual UCB cells were generally intermediate between those of their fetal and adult counterparts (B), and that individual groups of genes appear to transition from fetal to adult expression patterns with distinct timing (C).

Materials and Methods

Material for this chapter was modified from a manuscript currently in-press:

Daniel G. Bunis, Yelena Bronevetsky, Elisabeth Krow-Lucal, Nirav R. Bhakta, Charles C. Kim, Srilaxmi Nerella, Norman Jones, Ventura F. Mendoza, Yvonne J. Bryson, James E. Gern, Rachel L. Rutishauser, Chun Jimmie Ye, Marina Sirota, Joseph M. McCune, Trevor D. Burt. "Single-cell mapping of progressive fetal-to-adult transition in human naive T cells," in-press with Cell Reports, 2020.

Tissue collection

Fetal bone marrow (BM), spleen, and umbilical cord blood (UCB; 18-23 weeks GA) were obtained from San Francisco General Hospital (San Francisco, California, USA) after elective termination of pregnancy and written informed consent was obtained with IRB-approval from and under the guidelines of the UCSF Human Research Protection Program. Samples were excluded in the cases of (a) known maternal infection, (b) known intrauterine fetal demise, and/or (c) known or suspected chromosomal abnormality. Samples were fully de-identified, had no associated Personal Heath Information (PHI), and researchers had no access to PHI. Fetal UCB was collected from the umbilical cord after sterile preparation with antiseptic swabs as described previously (Zota et al., 2018). Samples were transported in media on ice, and were processed within 1-2 hours of collection. Due to the inherent technical challenges of obtaining fetal peripheral blood, we limited our analyses to four fetal UCB samples. These were used for microarray analysis. Full-term UCB was obtained in a de-identified manner under the auspices of CHR approved protocols from healthy, full-term infants from San Francisco General Hospital (San Francisco, California, USA) with a gestational age of 37 weeks or greater, and from a subset of children from the Boston metropolitan area who participated in the URECA (Urban Environment and Childhood Asthma) study (Gern et al., 2009). Matched adult peripheral blood and BM samples were obtained from healthy volunteer donors through AllCells (Alameda, CA). UCB and longitudinal post-natal samples were obtained from a cohort of 15 infants that were followed for up to the first 6 months of life by venipuncture under protocols approved by the UCLA CHR. 11 of these infants were HIV-exposed, but not infected (mother HIV+). All infants were healthy at the time of collection and none had congenital CMV infection.

Cell isolation

BM cells were isolated by dispersion of BM in R10 (RPMI-1640 supplemented with 10% FBS with 100U/ml penicillin, 100U/ml streptomycin, 2mM L-Glutamine). Splenic tissue was

digested with collagenase IV and DNAse I in R10 to yield a single cell suspension. Mononuclear cells were isolated from all samples (fetal splenocytes, fetal UCB, newborn UCB and peripheral blood, adult peripheral blood, and fetal and adult BM cells) by density centrifugation of a Ficoll-paque gradient (or Percoll gradient for three fetal splenic T cell samples, which has been previously established to result in identical T cell phenotypes compared with Ficoll (Ng et al., 2019)). All samples were viably cryopreserved prior to use.

Fluorescence activated cell sorting (FACS) for microarray and Fluidigm qRT-PCR

Mononuclear cell preparations were incubated in FACS staining buffer (PBS with 2% FBS and 2 mM EDTA) with fluorochrome-conjugated anti-human surface monoclonal antibodies (mAbs). Antibodies used included: CD3 Alexa-700 (SP34-2, BD Biosciences), CD14 Qdot605 (Tuk4, Invitrogen), CD16 FITC (3G8, BD Pharmingen), HLA-DR PE-Cy7 (G46-6, BD Biosciences), CD4 Qdot655 (S3.5, Life Technologies), CD45RA ECD (3P, Beckman Coulter), CD27 APC-eFluor780 (O323, eBiosciences), CD25 PE(PC61, BD Biosciences), and CD8 PE(SK1, BD Biosciences). All cells were stained with a live/dead marker (Amine-Aqua/AmCyan; Invitrogen) to exclude dead cells from the analysis and sorting. Cells were then sorted by FACS (FACS Aria, BD Biosciences) into PBS. These cells were then re-sorted to a purity of greater than 99% directly into RNAqueous Micro lysis buffer (Ambion – Life Technologies). For Fluidigm qRT-PCR, cells were sorted into Cells Direct 2x Reaction Mix (Life Technologies).

RNA preparation for microarray analysis

RNA was isolated from FACS-sorted samples using the RNAqueous-Micro kit (Life Technologies) and subjected to two rounds of linear amplification using the Aminoallyl MessageAmp II aRNA Amplification kit (Invitrogen). Cy3-coupled aRNA was fragmented and hybridized overnight to a SurePrint G3 Human GE v2 8x60K microarray, which was washed and scanned per the manufacturer's instructions (Agilent Technologies).

Statistical analysis of microarrays

Raw intensities were extracted using Feature Extraction software (Agilent) and log2 transformed. Probes with at least 70% missing data were filtered, then any other missing data was imputed. Data were then quantile normalized using the normalizeBetweenArrays function of the Limma package (Ritchie et al., 2015), followed by filtering of probes without expression above a 2^7 background for every replicate in at least one sample group, followed by median scaling per gene. Differentially expressed genes were identified using Limma (Tusher et al., 2001) in R, and data visualized as heatmaps using custom Perl scripts. After unbiased clustering of fetal and adult gene expression, we identified two adult monocyte microarray samples, APB1 and APB5, that clustered with fetal samples. The corresponding T cells for these samples, on the other hand, clustered appropriately with the other adult samples. Based on pathway analysis of genes driving this aberrant cluster, we hypothesized that these two adult samples had a viral infection at the time of blood draw. We thus removed those two adult samples when performing differential expression analysis for the monocytes. Further, the fetal sample FPB5 was an outlier in clustering and we removed it from further analysis. Thus, for T cell adult vs. fetal microarray analysis, we compared 5 adult and 4 fetal samples, while 3 adult and 4 fetal samples were used in monocytes. All 5 adult and 4 fetal samples were used in subsequent PCA analyses to calculate relative weights of signature genes for developmental stage scoring.

Gene signature derivation for bulk developmental stage score

To derive a broad fetal vs. adult transcriptional signature, we used microarray gene expression data to identify genes that were significantly differentially expressed (FDR <0.05) between fetal and adult samples, and greater than 1.5 fold differentially expressed in the same direction in both the monocytes and the naïve T cells. After removing uninformative genes that include “XLOC”, “LOC1”, “ENST”, “ORF”, “A_19”, “A_24”, or “A_33” we identified 169 genes that

met these criteria. After optimizing PCR primers, we identified a final list of 33 genes for further analysis.

Fluidigm qRT-PCR of signature genes

Classical monocytes and naïve T cells were sorted directly into 96-well plates containing Cells Direct 2x Reaction Mix, with 200 cells per well in replicates of 6. Lysed cells were subjected to reverse transcription and gene-specific pre-amplification using the SuperScript® III CellsDirect™ cDNA Synthesis Kit (Invitrogen). Primers were purchased from IDT and the specific targeting amplification (STA) mix was prepared according to the Fluidigm protocol (Gene Expression Using SsoFast EvaGreen SuperMix with Low ROX on the BioMark™ or BioMark HD System). Unincorporated primers were digested with Exonuclease I (New Englad Biolabs). Amplified cDNA was diluted and combined with 2X SsoFast EvaGreen Supermix with Low ROX (Bio-rad) and 20X DNA Binding Dye Sample Loading Reagent (Fluidigm). Primers were combined with 2X Assay Loading Reagent (Fluidigm) and 1X DNA Suspension Buffer (Teknova). Sample and primer mixes were loaded onto 96.96 Dynamic Array IFC (Fluidigm) and qPCR was performed using the BioMark System (Fluidigm).

Statistical analysis of Fluidigm qPCR for comparison of fetal, newborn, and adult samples.

BioMark qPCR data were analyzed using R software and filtered for quality: Cts greater than 27 (or with amplification quality < 0.5 as determined by Fluidigm Real Time PCR Analysis software) were marked as failed, followed by removal of replicates with greater than 80% of primer assays failed, and of remaining Ct values that were outliers from a normal distribution (two-tailed $p<0.02$). Subjects with fewer than three successful replicate wells were then removed. Next, primers were removed which displayed: greater than or equal to 80% of samples failed, followed by removal of primers with an average Ct of greater than 25.

Normalization between chips was performed by subtracting the mean Ct value for a gene within a chip from all values for that gene on that chip. Principal component analysis was performed on the processed microarray data to generate PC1 loadings for signature genes. Log2 expression values were averaged over replicates; values were then z-score standardized across samples within a gene and multiplied by the PC1 loading values before summation to generate the signature score.

Statistical analysis of Fluidigm qPCR for large, URECA, birth cohort

BioMark qPCR data were analyzed using R software via RStudio (R Core Team, 2019; RStudio Team, 2016) to mitigate technical variation and filter for quality through removal of reactions with amplification quality < 0.5, followed by inter-chip normalization, normalization across samples by endogenous control genes selected to be stable by geNorm (GAPDH, RPL35A, RPL11 for monocytes; GAPDH, RPL35A, B2M for T cells) (Vandesompele et al., 2002), and removal of reactions with un-normalized Cts greater than 27. We then required that there be at least three successful replicate wells for each gene associated with every sample. Genes were removed from the signature in order to maximize the number of samples retained. Principal component analysis was performed on the processed microarray data to generate PC1 loadings for these signature genes. Log2 expression values were averaged over replicates; values were then z-score standardized across samples within a gene and multiplied by the PC1 loading values before summation to generate the signature score.

Cell enrichment and FACS for RNA sequencing.

Mononuclear cell preparations were pre-enriched by immunomagnetic selection with either the EasySep Human CD34 Positive Selection Kit (STEMCELL Technologies) or the EasySep Human T Cell Isolation Kit (STEMCELL Technologies) following manufacturer's protocols. Pre-enriched cells were then incubated in FACS staining buffer (PBS with 2% FBS and 2 mM EDTA) with fluorochrome-conjugated anti-human monoclonal antibodies. Pre-

enriched T cells were stained with CD4-BV650 (SK3, BD Biosciences), CCR7-BV785 (G043H7, Biolegend), CD25-FITC (2A3, BD Biosciences), CD45RA-PE (HI100, BD Biosciences), CD8-PECy7 (RPA-T8, BD Biosciences), CD95-APC (DX2, Invitrogen), and CD27-APC-eFluor780 (O323, Invitrogen). Pre-enriched CD34+ hematopoietic stem and progenitor cells were stained with CD34-PE (581, Biolegend) and CD45-APC (8I30, Tonbo Biosciences). All cells were also stained with a live/dead marker (Ghost Dye Violet 510, Tonbo Biosciences) to exclude dead cells from analysis and sorting. Stained, pre-enriched cells were then sorted by FACS (FACS Aria Fusion, BD Biosciences) into R10 and kept on ice until further use. Sort purity was assessed by running a small fraction of the sorted cells from each sample through the FACS Aria Fusion a second time, after all samples had been sorted. Two cord samples yielded fewer than 1000 naïve CD8 T cells after sorting, and these cells were left out from further processing.

Preparation for single cell RNA-seq library generation

After sort purification, 25 thousand naïve CD4 T cells, naïve CD8 T cells, and CD34+ hematopoietic stem and progenitor cells (or as many as were sorted), from up to 10 distinct samples, were pooled at equal cell concentrations, encapsulated into droplets, and converted into single-cell transcriptome libraries with 10X Chromium 3' v2 chemistry (10X Genomics) as described previously (Kang et al., 2018; Zheng et al., 2017). For bulk RNA-seq library generation, after sort purification, RNA from 50 thousand fetal, cord, and adult naïve CD4 T cells was extracted and purified using RNAqueous Total RNA Isolation columns (Invitrogen AM1912), with addition of QIAshredder homogenization columns (Qiagen 79654) at the cell lysate step. mRNA was enriched with poly-dT beads (NEB E7490) reverse transcribed, ligated to adapters, dual barcoded and amplified using a NEBNext Ultra II Directional RNA Library Prep Kit (NEB E7760) according to the manufacturer's protocol. Completed bulk and single-cell RNA sequencing libraries were assessed for proper sizing via Agilent Bioanalyzer and concentration via qRT-PCR using a NEBNext Library Quantification Kit (NEB E7630). Bulk RNA sequencing

libraries were pooled at equal concentration and sequenced on an Illumina HiSeq via paired end 2x100bp sequencing. Single-cell RNA sequencing libraries were pooled at equal concentration and sequenced on an Illumina Novaseq via paired end 2x150bp sequencing.

Raw sequencing data pre-processing

Sequencing reads from bulk RNA sequencing were assessed for quality using FastQC (Andrews, 2018). Low quality reads and ends of reads were trimmed using Trim Galore! (Krueger, 2019). Reads were aligned to the human genome (hg38) using STAR (Dobin et al., 2013). Aligned reads were then used for two different purposes. First, reads for each sample overlapping with exons were quantified, using featureCounts (Liao et al., 2014) for the purpose of differential gene expression and pathways analysis. Second, genotype information was extracted using the broad Genome Analysis Toolkit (A. McKenna et al., 2010) for use in Demuxlet deconvolution of single cells' sample identities in single cell RNA sequencing datasets. Sequencing reads from single-cell RNA sequencing were quality controlled, matched by cell barcode, aligned to the genome, and overlapping gene features with unique UMIs quantified using cellranger (10X Genomics). Genetic polymorphisms captured within the single cell RNA sequencing reads were then extracted and compared to genotyping information, previously extracted from bulk RNA sequencing reads, in order to match cells to their original samples using Demuxlet (Kang et al., 2018).

Dimensionality reduction analysis and clustering of RNAseq data

Further processing and analysis of the data was carried out in R using RStudio (R Core Team, 2019; RStudio Team, 2016). For bulk RNA sequencing, regularized log (rlog) normalization of gene counts for all genes with more than 10 reads total was performed using DESeq2 (Love et al., 2014). Rlog values for the 2500 genes, captured in at least 4 of 5 samples of each age, with highest coefficient of variation (standard deviation / mean) were used for principal components analysis (PCA), which was calculated with prcomp, a function built into the

base R package (R Core Team, 2019). T cells and hematopoietic stem and progenitor cell single-cell sequencing datasets were analyzed separately. Cells with fewer than 1500 unique molecular identifiers (UMI) captured, fewer than 750 genes captured, or greater than 5 (T cells) or 7 (HSPCs) percent of reads coming from mitochondrial genes were filtered out. Cells determined to be doublets by Demuxlet, as well as any cell assigned to a sample that was not included in the 10X Chromium lane that the cell came from, were also filtered out. After quality control, a standard workflow for dimensionality reduction and clustering analysis in Seurat (Butler et al., 2018) was used for these datasets: Gene counts per cell were log normalized. Top variable genes were selected using the FindVariableGenes function with default settings. For the hematopoietic stem and progenitor cell dataset, cells for the distinct ages were additionally aligned with each other at this step, using the IntegrateData function, in order to enable cell type identification later through joint clustering and trajectory inference analysis (Stuart et al., 2019). Expression (or IntegrateData-adjusted expression) of the identified variable genes was then scaled using the ScaleData function. CellCycle, percentage of mitochondrial reads, and number of UMI were regressed out with the ScaleData function for non-Integrated data. The first 50 principal components (PCs) were calculated based on this scaled expression matrix. Standard deviation of the resulting PCs, and empirical testing by the jack straw method were used to determine how many PCs to carry forward for further dimensionality reduction and clustering. Further dimensionality reduction was carried out with uniform manifold approximation and projection (UMAP) by the umap-learn algorithm (McInnes et al., 2018), as well as clustering by the default Seurat Louvain algorithm with resolution parameter empirically chosen and set to 0.1 for T cells and 1.0 for HSPCs.

Differential expression and pathway analysis of RNAseq data

For the naïve CD4 T cell bulk RNA sequencing dataset, DESeq2 was used for calculating genes differentially expressed between ages (FDR < 0.05 and log2 fold change ≥

1.5, unless otherwise stated in the text). For the naïve CD4 and CD8 T cell single-cell dataset and for our hematopoietic stem and progenitor cell dataset, differentially expressed genes were determined using the Seurat adaptation of MAST (Finak et al., 2015) via the FindMarkers function with cutoffs—FDR < 0.05 and relative log fold change ≥ 0.585 fold—applied externally to the function. Venn diagram comparisons of gene sets were generated with the eulerr package (Larsson, 2019). For unbiased pathway analysis, enrichment tests were run, on given gene sets, for all KEGG pathways based on a hypergeometric distribution, using clusterProfiler (G. Yu et al., 2012) with a false discovery rate cutoff of less than 0.05. In order to facilitate visualization, interpretation, and discussion of data, we focused on signaling pathways by subsetting the full list of enriched pathways to those that included “signal” in their name, or to non-infection immune-related pathways by manual curation based on domain knowledge (**Fig. 2.12**).

Cell type annotation in the hematopoietic stem and progenitor cell single-cell dataset

To facilitate cell type identification, expression differences likely related to sample age or tissue were first mitigated using the IntegrateData batch correction tool. This was followed by combination of an reference-based cell type assignment, via the Bioconductor version of SingleR (Aran et al., 2019), and a trajectory inference, via slingshot (Street et al., 2018), to identify cell types at the start or end of the differentiation tree. SingleR, along with its included Blueprint-ENCODE (Aran et al., 2019; ENCODE Project Consortium, 2012; Martens & Stunnenberg, 2013) reference dataset, was used to initially score our cells, and identified which cluster contained the highest numbers of hematopoietic stem cells (HSCs). This cluster was then used as the starting point (the start.clus parameter) for slingshot trajectory analysis (Street et al., 2018) based on the principal components of the dataset. Clusters before and after the first and last splits in the trajectory map were then merged and annotated based on most frequent

SingleR annotations within the combined clusters. At this point, two separate endpoint clusters branching from the same upstream cluster had been annotated as common lymphoid progenitors (CLP) so these endpoint clusters and their common upstream cluster were merged and all annotated as CLPs.

Developmental stage score generation with machine learning through random forest regression

Random forest regression was chosen as the algorithm for generating a developmental stage score of each single-cell transcriptome due to the relative simplicity and assumption-free nature of this algorithm. First, a randomly selected training set of fetal and adult cells for each target cell type was chosen. For naïve T cells, the training set consisted of a random 10% of fetal and adult naïve CD4 and CD8 T cells. For individual hematopoietic progenitor cell type annotations, the training set consisted of ~30% of fetal and adult cells of those populations, but the exact percentage of fetal and adult cells included in the training set was adjusted in order to account for the fact that there were unequal numbers of fetal and adult cells for each annotation. Heavily imbalanced training sets can reduce the performance of machine learning algorithms, but applying a 50% correction based on such imbalances is recommended (Chicco, 2017). To a apply such a correction, we used the following equations to establish adjusted training cell percentages:

$$Fetal\ training\ percentage = 30\% * \left(0.5 + \frac{\mathrm{a}}{\mathrm{f}+\mathrm{a}}\right)$$

$$Adult\ training\ percentage = 30\% * \left(0.5 + \frac{\mathrm{f}}{\mathrm{f}+\mathrm{a}}\right)$$

in which f equals the number of fetal cells and a equals the number of adult cells with each cell type annotation. After training set selection, differentially express genes, calculated within the training sets, were used as potential features. These potential features were narrowed down to less than 20 markers using the filter.corr, and rfeRF functions of the feseR package (Perez-

Riverol et al., 2017) in order to reduce the potential for model over-fitting when too many features are used for training. With this package, after an initial correlation filter of 0.3, iteration of random forest model generation and evaluation, followed by removal of least important feature, was carried out until a maximum area under the receiver operator curve (AUC) was reached. If the number of features leading to maximal AUC was greater than 20, iteration was continued until the final set of variables for which each of three measures of evaluation (accuracy, specificity, and sensitivity) remained at or above 0.99. Next, a final random forest regression model was generated with the caret package using expression of these final markers within training set cells and the ranger method (Kuhn, 2008; Wright & Ziegler, 2017). Accuracy of the models were confirmed based on AUC within the fetal and adult cells that were left out of the training set before being applied to all cells of the target populations. Resulting developmental stage scores were then stored as a metadata within the Seurat objects for further analysis and visualization.

Visualization of RNA sequencing data

The R package “dittoSeq: User Friendly Single-Cell and Bulk RNA Sequencing Visualization” was created to complement the analysis of the sequencing data in this paper by powering and simplifying the visualization of both bulk and single-cell RNAseq data types. Visualization of the dimensionality reduction analyses, cell clustering, developmental stage scores, gene expression levels, differentiation trajectories, and cell type annotations for RNA-seq datasets were all performed with dittoSeq. The package has been accepted into Bioconductor and is available through Bioconductor at https://bioconductor.org/packages/dittoSeq/ and on GitHub at https://github.com/dtm2451/dittoSeq.The package works directly with single-cell RNA-seq data stored as either Seurat or SingleCellExperiment data structures, and can import bulk RNA-seq data that have been normalized by DESeq2 (Love et al., 2014), or edgeR (Robinson et al.,

2010). It contains functions for overlaying gene expression and cellular meta data on top of gene-by-gene or dimensionality reduction scatterplots; for plotting gene expression or other numerical metadata (for example developmental stage scores), grouped by samples or grouped by any other discrete per-cell metric, in violin and/or box or ridge plot format; for plotting discrete per-cell data, such as single-cell clusters, grouped by samples or grouped by any other discrete per-cell metric, as percent composition within the groupings; for generation of gene expression heatmaps (after log normalization of bulk RNAseq counts data via the DESeq2 normTransform function with default settings); and additionally for import of Demuxlet sample calls, and generation of Demuxlet-associated quality assessment metrics and plots.

Quantification and Statistical Analysis

Statistical analyses of differential gene expression were performed with packages appropriate to each of our particular types of transcriptional data (Limma for microarray, Seurat and MAST for scRNA-seq, and DESeq2 for bulk RNA-seq), as described within associated Method Details sections. Details of all other statistical tests are provided within associated figure legends. All statistical tests were computed in R.

Data and Code Availability

Microarray and RNA-seq counts data have been deposited in GEO, accession TBD. To enable exploration of these datasets by other researchers, raw counts and fully processed objects for the bulk and single-cell RNA-seq data are available on figshare (https://figshare.com/projects/Single-cell_mapping_of_progressive_fetal-to-adult_transition_in_human_hematopoiesis/76143), and all code necessary for recreating the reported analyses and figures within R is available on GitHub (https://github.com/dtm2451/ProgressiveHematopoiesis). The dittoSeq visualization software created to aid visualization of these, and other, RNA-seq data is available through Bioconductor

at https://bioconductor.org/packages/dittoSeq/ and on GitHub at https://github.com/dtm2451/dittoSeq.

REFERENCES

Aghaeepour, N., Ganio, E. A., Mcilwain, D., Tsai, A. S., Tingle, M., Gassen, S. V., Gaudilliere, D. K., Baca, Q., McNeil, L., Okada, R., Ghaemi, M. S., Furman, D., Wong, R. J., Winn, V. D., Druzin, M. L., El-Sayed, Y. Y., Quaintance, C., Gibbs, R., Darmstadt, G. L., … Gaudilliere, B. (2017). An immune clock of human pregnancy. *Science Immunology*, *2*(15), eaan2946. https://doi.org/10.1126/sciimmunol.aan2946

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x

Andrews, S. (2018). *FastQC: a quality control tool for high throughput sequence data.* (0.11.8) [Computer software]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aran, D., Hu, Z., & Butte, A. J. (2017). xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, *18*(1), 220. https://doi.org/10.1186/s13059-017-1349-1

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163. https://doi.org/10.1038/s41590-018-0276-y

Ballen, K. K., Gluckman, E., & Broxmeyer, H. E. (2013). Umbilical cord blood transplantation: The first 25 years and beyond. *Blood*, *122*(4), 491–498. https://doi.org/10.1182/blood-2013-02-453175

Baranzini, S. E., Khankhanian, P., Patsopoulos, N. A., Li, M., Stankovich, J., Cotsapas, C., Søndergaard, H. B., Ban, M., Barizzone, N., Bergamaschi, L., Booth, D., Buck, D., Cavalla, P., Celius, E. G., Comabella, M., Comi, G., Compston, A., Cournu-Rebeix, I., D'alfonso, S., … Oksenberg, J. R. (2013). Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls. *The American Journal of Human Genetics*, *92*(6), 854–865. https://doi.org/10.1016/j.ajhg.2013.04.019

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4), 346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bashford-Rogers, R. J. M., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A., & Kellam, P. (2013). Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Research*, *23*(11), 1874–1884. https://doi.org/10.1101/gr.154815.113

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K. A., Shankar, R. D., Shen-Orr, S. S., Thomson, E., Wiser, J., & Butte, A. J. (2018). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, *5*, 180015. https://doi.org/10.1038/sdata.2018.15

Blainey, P. C., & Quake, S. R. (2014). Dissecting genomic diversity, one cell at a time. *Nature Methods*, *11*(1), 19–21.

Blanchard, A. C., Quach, C., & Autmizguine, J. (2015). Staphylococcal infections in infants: Updates and current challenges. *Clinics in Perinatology*, *42*(1), 119–132, ix. https://doi.org/10.1016/j.clp.2014.10.013

Blazar, B. R., MacDonald, K. P. A., & Hill, G. R. (2018). Immune regulatory cell infusion for graft-versus-host disease prevention and therapy. *Blood*, *131*(24), 2651–2660. https://doi.org/10.1182/blood-2017-11-785865

Borghesi, A., Stronati, M., & Fellay, J. (2017). Neonatal Group B Streptococcal Disease in Otherwise Healthy Infants: Failure of Specific Neonatal Immune Responses. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.00215

Bradley, P., & Thomas, P. G. (2019). Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annual Review of Immunology*, *37*, 547–570. https://doi.org/10.1146/annurev-immunol-042718-041757

Bronevetsky, Y., Burt, T. D., & McCune, J. M. (2016). Lin28b Regulates Fetal Regulatory T Cell Differentiation through Modulation of TGF-β Signaling. *The Journal of Immunology*, *197*(11), 4344–4350. https://doi.org/10.4049/jimmunol.1601070

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., & Greenleaf, W. J. (2018). Integrated Single-Cell Analysis Maps the

Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*. https://doi.org/10.1016/j.cell.2018.03.074

Buenrostro, J., Wu, B., Chang, H., & Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *109*, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burt, T. D. (2013). Fetal Regulatory T Cells and Peripheral Immune Tolerance In Utero: Implications for Development and Disease. *American Journal of Reproductive Immunology*, *69*(4), 346–358. https://doi.org/10.1111/aji.12083

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., & Pritchard, J. K. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*. https://doi.org/10.1038/s41588-019-0505-9

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Chappell, L., Russell, A. J. C., & Voet, T. (2018). Single-Cell (Multi)omics Technologies. *Annual Review of Genomics and Human Genetics*, *19*(1), 15–41. https://doi.org/10.1146/annurev-genom-091416-035324

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*(1), 35. https://doi.org/10.1186/s13040-017-0155-3

Chiu, L., Bazin, T., Truchetet, M.-E., Schaeverbeke, T., Delhaes, L., & Pradeu, T. (2017). Protective Microbiota: From Localized to Long-Reaching Co-Immunity. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.01678

Cook, C. E., Lopez, R., Stroe, O., Cochrane, G., Brooksbank, C., Birney, E., & Apweiler, R. (2019). The European Bioinformatics Institute in 2018: Tools, infrastructure and training. *Nucleic Acids Research*, *47*(D1), D15–D22. https://doi.org/10.1093/nar/gky1124

Cupedo, T., Nagasawa, M., Weijer, K., Blom, B., & Spits, H. (2005). Development and activation of regulatory T cells in the human fetus. *European Journal of Immunology*, *35*(2), 383–390. https://doi.org/10.1002/eji.200425763

Darrasse-Jèze, G., Marodon, G., Salomon, B. L., Catala, M., & Klatzmann, D. (2005). Ontogeny of CD4+CD25+ regulatory/suppressor T cells in human fetuses. *Blood*, *105*(12), 4715–4721. https://doi.org/10.1182/blood-2004-10-4051

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

Davis, M. M., & Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature*, *334*(6181), 395–402. https://doi.org/10.1038/334395a0

Davis, Mark M., Tato, C. M., & Furman, D. (2017). Systems immunology: Just getting started. *Nature Immunology*, *18*(7), 725–732. https://doi.org/10.1038/ni.3768

Ding, Y.-H., Qian, L.-Y., Pang, J., Lin, J.-Y., Xu, Q., Wang, L.-H., Huang, D.-S., & Zou, H. (2017). The regulation of immune cells by Lactobacilli: A potential therapeutic target for anti-atherosclerosis therapy. *Oncotarget*, *8*(35), 59915–59928. https://doi.org/10.18632/oncotarget.18346

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., & Blaser, M. J. (2019). Role of the microbiome in human development. *Gut*, *68*(6), 1108–1114. https://doi.org/10.1136/gutjnl-2018-317503

Dowling, D. J., & Levy, O. (2014). Ontogeny of early life immunity. *Trends in Immunology*, *35*(7), 299–310. https://doi.org/10.1016/j.it.2014.04.007

Drissen, R., Buza-Vidas, N., Woll, P., Thongjuea, S., Gambardella, A., Giustacchini, A., Mancini, E., Zriwil, A., Lutteropp, M., Grover, A., Mead, A., Sitnicka, E., Jacobsen, S. E. W., & Nerlov, C. (2016). Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nature Immunology*, *17*(6), 666–676. https://doi.org/10.1038/ni.3412

Dumas, A., Corral, D., Colom, A., Levillain, F., Peixoto, A., Hudrisier, D., Poquet, Y., & Neyrolles, O. (2018). The Host Microbiota Contributes to Early Protection Against Lung Colonization by Mycobacterium tuberculosis. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02656

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-01973-8

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Elhanati, Y., Sethna, Z., Callan, C. G., Mora, T., & Walczak, A. M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, *284*(1), 167–179. https://doi.org/10.1111/imr.12665

Elias, S., & Rudensky, A. Y. (2019). Therapeutic use of regulatory T cells for graft-versus-host disease. *British Journal of Haematology*, *187*(1), 25–38. https://doi.org/10.1111/bjh.16157

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Fernandez, D. M., Rahman, A. H., Fernandez, N. F., Chudnovskiy, A., Amir, E. D., Amadori, L., Khan, N. S., Wong, C. K., Shamailova, R., Hill, C. A., Wang, Z., Remark, R., Li, J. R., Pina, C., Faries, C., Awad, A. J., Moss, N., Bjorkegren, J. L. M., Kim-Schulze, S., … Giannarelli, C. (2019). Single-cell immune landscape of human atherosclerotic plaques. *Nature Medicine*, *25*(10), 1576–1588. https://doi.org/10.1038/s41591-019-0590-4

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., & Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*. https://doi.org/10.1186/s13059-015-0844-5

Franzén, O., Gan, L.-M., & Björkegren, J. L. M. (2019). PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, *2019*. https://doi.org/10.1093/database/baz046

Geenen, V., Martens, H., Brilot, F., Renard, C., Franchimont, D., & Kecha, O. (2000). Thymic neuroendocrine self-antigens. Role in T-cell development and central T-cell self-tolerance. *Annals of the New York Academy of Sciences*, *917*, 710–723. https://doi.org/10.1111/j.1749-6632.2000.tb05435.x

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., & Quake, S. R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, *32*(2), 158–168. https://doi.org/10.1038/nbt.2782

Gern, J. E., Visness, C. M., Gergen, P. J., Wood, R. A., Bloomberg, G. R., O'Connor, G. T., Kattan, M., Sampson, H. A., Witter, F. R., Sandel, M. T., Shreffler, W. G., Wright, R. J., Arbes, S. J., & Busse, W. W. (2009). The Urban Environment and Childhood Asthma (URECA) birth cohort study: Design, methods, and study population. *BMC Pulmonary Medicine*, *9*, 17. https://doi.org/10.1186/1471-2466-9-17

Ginhoux, F., Lim, S., Hoeffel, G., Low, D., & Huber, T. (2013). Origin and differentiation of microglia. *Frontiers in Cellular Neuroscience*, *7*, 45. https://doi.org/10.3389/fncel.2013.00045

Gowthaman, R., & Pierce, B. G. (2019). TCR3d: The T cell receptor structural repertoire database. *Bioinformatics*, *35*(24), 5323–5325. https://doi.org/10.1093/bioinformatics/btz517

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., Pousse, Y., Reichen, M., Woolfe, A., Brenan, C.,

Griffiths, A. D., Vallot, C., & Gérard, A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature Genetics*, *51*(6), 1060–1066. https://doi.org/10.1038/s41588-019-0424-9

Hadland, B., & Yoshimoto, M. (2018). Many layers of embryonic hematopoiesis: New insights into B-cell ontogeny and the origin of hematopoietic stem cells. *Experimental Hematology*, *60*, 1–9. https://doi.org/10.1016/j.exphem.2017.12.008

Halkias, J., Rackaityte, E., Hillman, S. L., Aran, D., Mendoza, V. F., Marshall, L. R., MacKenzie, T. C., & Burt, T. D. (2019). CD161 contributes to prenatal immune suppression of IFN-γ–producing $PLZF^{+}$ T cells. *The Journal of Clinical Investigation*, *129*(9), 3562–3577. https://doi.org/10.1172/JCI125957

Havran, W. L., & Allison, J. P. (1988). Developmentally ordered appearance of thymocytes expressing different T-cell antigen receptors. *Nature*, *335*(6189), 443–445. https://doi.org/10.1038/335443a0

Hayakawa, K., Hardy, R. R., Herzenberg, L. A., & Herzenberg, L. A. (1985). Progenitors for Ly-1 B cells are distinct from progenitors for other B cells. *The Journal of Experimental Medicine*, *161*(6), 1554–1568. https://doi.org/10.1084/jem.161.6.1554

Heather, J. M., Ismail, M., Oakes, T., & Chain, B. (2018). High-throughput sequencing of the T-cell receptor repertoire: Pitfalls and opportunities. *Briefings in Bioinformatics*, *19*(4), 554–565. https://doi.org/10.1093/bib/bbw138

Heng, T. S. P., Painter, M. W., & Immunological Genome Project Consortium. (2008). The Immunological Genome Project: Networks of gene expression in immune cells. *Nature Immunology*, *9*(10), 1091–1094. https://doi.org/10.1038/ni1008-1091

Herzenberg, L. A., & Herzenberg, L. A. (1989). Toward a layered immune system. *Cell*, *59*(6), 953–954. https://doi.org/10.1016/0092-8674(89)90748-4

Hiwarkar, P., Hubank, M., Qasim, W., Chiesa, R., Gilmour, K. C., Saudemont, A., Amrolia, P. J., & Veys, P. (2017). Cord blood transplantation recapitulates fetal ontogeny with a distinct molecular signature that supports CD4+ T-cell reconstitution. *Blood Advances*, *1*(24), 2206–2216. https://doi.org/10.1182/bloodadvances.2017010827

Horowitz, A., Strauss-Albee, D. M., Leipold, M., Kubo, J., Nemat-Gorgani, N., Dogan, O. C., Dekker, C. L., Mackey, S., Maecker, H., Swan, G. E., Davis, M. M., Norman, P. J., Guethlein, L. A., Desai, M., Parham, P., & Blish, C. A. (2013). Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. *Science Translational Medicine*, *5*(208), 208ra145. https://doi.org/10.1126/scitranslmed.3006702

Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J., Forbes, S., Gilbert, J. G. R., Halls, K., Harrow, J. L., Hart, E., Howe, K., Jackson, D. K., Palmer, S., Roberts, A. N., Sims, S., Stewart, C. A., Traherne, J. A., Trevanion, S., … Beck, S. (2008). Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. *Immunogenetics*, *60*(1), 1–18. https://doi.org/10.1007/s00251-007-0262-2

Hu, Z., Jujjavarapu, C., Hughey, J. J., Andorf, S., Lee, H.-C., Gherardini, P. F., Spitzer, M. H., Thomas, C. G., Campbell, J., Dunn, P., Wiser, J., Kidd, B. A., Dudley, J. T., Nolan, G. P., Bhattacharya, S., & Butte, A. J. (2018). MetaCyto: A Tool for Automated Meta-analysis of Mass and Flow Cytometry Data. *Cell Reports*, *24*(5), 1377–1388. https://doi.org/10.1016/j.celrep.2018.07.003

Iglesia, M. D., Parker, J. S., Hoadley, K. A., Serody, J. S., Perou, C. M., & Vincent, B. G. (2016). Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. *JNCI Journal of the National Cancer Institute*, *108*(11). https://doi.org/10.1093/jnci/djw144

Ikuta, K., Kina, T., MacNeil, I., Uchida, N., Peault, B., Chien, Y., & Weissman, I. L. (1990). A developmental switch in thymic lymphocyte maturation potential occurs at the level of hematopoietic stem cells. *Cell*, *62*(5), 863–874. https://doi.org/10.1016/0092-8674(90)90262-D

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A., & Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, *167*(7), 1883-1896.e15. https://doi.org/10.1016/j.cell.2016.11.039

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, *316*(5830), 1497–1502. https://doi.org/10.1126/science.1141319

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jotereau, F. V., & Le Douarin, N. M. (1982). Demonstration of a cyclic renewal of the lymphocyte precursor cells in the quail thymus during embryonic and perinatal life. *Journal of Immunology (Baltimore, Md.: 1950)*, *129*(5), 1869–1877.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L. A., & Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing

using natural genetic variation. *Nature Biotechnology*, *36*(1), 89–94. https://doi.org/10.1038/nbt.4042

Kantor, A. B., Stall, A. M., Adams, S., Herzenberg, L. A., & Herzenberg, L. A. (1992). Differential development of progenitor activity for three B-cell lineages. *Proceedings of the National Academy of Sciences*, *89*(8), 3320–3324. https://doi.org/10.1073/pnas.89.8.3320

Kennedy, A. E., Ozbek, U., & Dorak, M. T. (2017). What has GWAS done for HLA and disease associations? *International Journal of Immunogenetics*, *44*(5), 195–211. https://doi.org/10.1111/iji.12332

Kim, I. S., Gao, Y., Welte, T., Wang, H., Liu, J., Janghorban, M., Sheng, K., Niu, Y., Goldstein, A., Zhao, N., Bado, I., Lo, H.-C., Toneff, M. J., Nguyen, T., Bu, W., Jiang, W., Arnold, J., Gu, F., He, J., … Zhang, X. H.-F. (2019). Immuno-subtyping of breast cancer reveals distinct myeloid cell profiles and immunotherapy resistance mechanisms. *Nature Cell Biology*, *21*(9), 1113–1126. https://doi.org/10.1038/s41556-019-0373-7

Kimball, A. K., Oko, L. M., Bullock, B. L., Nemenoff, R. A., Dyk, L. F. van, & Clambey, E. T. (2018). A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data. *The Journal of Immunology*, *200*(1), 3–22. https://doi.org/10.4049/jimmunol.1701494

Kim-Hellmuth, S., Bechheim, M., Pütz, B., Mohammadi, P., Nédélec, Y., Giangreco, N., Becker, J., Kaiser, V., Fricker, N., Beier, E., Boor, P., Castel, S. E., Nöthen, M. M., Barreiro, L. B., Pickrell, J. K., Müller-Myhsok, B., Lappalainen, T., Schumacher, J., & Hornung, V. (2017). Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature Communications*, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-00366-1

Klimchenko, O., Di Stefano, A., Geoerger, B., Hamidi, S., Opolon, P., Robert, T., Routhier, M., El-Benna, J., Delezoide, A.-L., Boukour, S., Lescure, B., Solary, E., Vainchenker, W., & Norol, F. (2011). Monocytic cells derived from human embryonic stem cells and fetal liver share common differentiation pathways and homeostatic functions. *Blood*, *117*(11), 3065–3075. https://doi.org/10.1182/blood-2010-07-295246

Komanduri, K. V., St. John, L. S., de Lima, M., McMannis, J., Rosinski, S., McNiece, I., Bryan, S. G., Kaur, I., Martin, S., Wieder, E. D., Worth, L., Cooper, L. J. N., Petropoulos, D., Molldrem, J. J., Champlin, R. E., & Shpall, E. J. (2007). Delayed immune reconstitution after cord blood transplantation is characterized by impaired thymopoiesis and late memory T-cell skewing. *Blood*, *110*(13), 4543–4551. https://doi.org/10.1182/blood-2007-05-092130

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Krow-Lucal, E. R., Kim, C. C., Burt, T. D., & McCune, J. M. (2014). Distinct functional programming of human fetal and adult monocytes. *Blood*, *123*(12), 1897–1904. https://doi.org/10.1182/blood-2013-11-536094

Krueger, F. (2019). *Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* (0.6.2) [Computer software]. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lalor, P. A., Stall, A. M., Adams, S., & Herzenberg, L. A. (1989). Permanent alteration of the murine Ly-1 B repertoire due to selective depletion of Ly-1 B cells in neonatal animals. *European Journal of Immunology*, *19*(3), 501–506. https://doi.org/10.1002/eji.1830190314

Larsson, J. (2019). *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses.* (R package version 5.1.0) [Computer software]. https://cran.r-project.org/package=eulerr

Le Douarin, N. M., & Jotereau, F. V. (1975). Tracing of cells of the avian thymus through embryonic life in interspecific chimeras. *The Journal of Experimental Medicine*, *142*(1), 17–40. https://doi.org/10.1084/jem.142.1.17

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, *3*(9), e161. https://doi.org/10.1371/journal.pgen.0030161

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

Li, N., Unen, V. van, Abdelaal, T., Guo, N., Kasatskaya, S. A., Ladell, K., McLaren, J. E., Egorov, E. S., Izraelson, M., Lopes, S. M. C. de S., Höllt, T., Britanova, O. V., Eggermont, J., Miranda, N. F. C. C. de, Chudakov, D. M., Price, D. A., Lelieveldt, B. P. F., & Koning, F. (2019). Memory CD4 + T cells are generated in the human fetal intestine. *Nature Immunology*, 1. https://doi.org/10.1038/s41590-018-0294-9

Li, Y., & Tollefsbol, T. O. (2011). DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in Molecular Biology (Clifton, N.J.)*, *791*, 11–21. https://doi.org/10.1007/978-1-61779-316-5_2

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liston, A., & Goris, A. (2018). The origins of diversity in human immunity. *Nature Immunology*, *19*(3), 209–210. https://doi.org/10.1038/s41590-018-0047-9

Loosdregt, J. van, & Coffer, P. J. (2018). The Role of WNT Signaling in Mature T Cells: T Cell Factor Is Coming Home. *The Journal of Immunology*, *201*(8), 2193–2200. https://doi.org/10.4049/jimmunol.1800633

López-Santibáñez-Jácome, L., Avendaño-Vázquez, S. E., & Flores-Jasso, C. F. (2019). The Pipeline Repertoire for Ig-Seq Analysis. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.00899

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Madan, J. C., Farzan, S. F., Hibberd, P. L., & Karagas, M. R. (2012). Normal neonatal microbiome variation in relation to environmental factors, infection and allergy. *Current Opinion in Pediatrics*, *24*(6), 753–759. https://doi.org/10.1097/MOP.0b013e32835a1ac8

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., … Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, *39*(10), 1181–1186. https://doi.org/10.1038/ng1007-1181

Martens, J. H. A., & Stunnenberg, H. G. (2013). BLUEPRINT: Mapping human blood cell epigenomes. *Haematologica*, *98*(10), 1487–1489. https://doi.org/10.3324/haematol.2013.094243

Martin, J.-E., Assassi, S., Diaz-Gallo, L.-M., Broen, J. C., Simeon, C. P., Castellvi, I., Vicente-Rabaneda, E., Fonollosa, V., Ortego-Centeno, N., González-Gay, M. A., Espinosa, G., Carreira, P., Camps, M., Sabio, J. M., D'alfonso, S., Vonk, M. C., Voskuyl, A. E., Schuerwegh, A. J., Kreuter, A., … Martin, J. (2013). A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. *Human Molecular Genetics*, *22*(19), 4021–4029. https://doi.org/10.1093/hmg/ddt248

Mavigner, M., Zanoni, M., Tharp, G. K., Habib, J., Mattingly, C. R., Lichterfeld, M., Nega, M. T., Vanderford, T. H., Bosinger, S. E., & Chahroudi, A. (2019). Pharmacological modulation of the Wnt/β-catenin pathway inhibits proliferation and promotes differentiation of long-lived memory CD4+ T-cells in ART-suppressed SIV-infected macaques. *Journal of Virology*. https://doi.org/10.1128/JVI.01094-19

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A. K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., … Knight, R. (2018). American Gut: An Open Platform for Citizen

Science Microbiome Research. *MSystems*, *3*(3). https://doi.org/10.1128/mSystems.00031-18

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018, September 2). *UMAP: Uniform Manifold Approximation and Projection*. Journal of Open Source Software. https://doi.org/10.21105/joss.00861

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKenna, D. H., Sumstad, D., Kadidlo, D., Batdorf, B., Lord, C. J., Merkel, S. C., Koellner, C. M., Curtsinger, J. M., June, C. H., Riley, J. L., Levine, B. L., Miller, J. S., Brunstein, C. G., Wagner, J. E., Blazar, B. R., & Hippen, K. L. (2017). Optimization of cGMP Purification and Expansion of Umbilical Cord Blood-Derived T-Regulatory Cells in Support of First-in-Human Clinical Trials. *Cytotherapy*, *19*(2), 250–262. https://doi.org/10.1016/j.jcyt.2016.10.011

McKinnon, K. M. (2018). Flow Cytometry: An Overview. *Current Protocols in Immunology*, *120*(1), 5.1.1-5.1.11. https://doi.org/10.1002/cpim.40

Merindol, N., Charrier, E., Duval, M., & Soudeyns, H. (2011). Complementary and contrasting roles of NK cells and T cells in pediatric umbilical cord blood transplantation. *Journal of Leukocyte Biology*, *90*(1), 49–60. https://doi.org/10.1189/jlb.0111007

Michaëlsson, J., Mold, J. E., McCune, J. M., & Nixon, D. F. (2006). Regulation of T Cell Responses in the Developing Human Fetus. *The Journal of Immunology*, *176*(10), 5741–5748. https://doi.org/10.4049/jimmunol.176.10.5741

Mingueneau, M., Kreslavsky, T., Gray, D., Heng, T., Cruse, R., Ericson, J., Bendall, S., Spitzer, M. H., Nolan, G. P., Kobayashi, K., Boehmer, H. von, Mathis, D., Benoist, C., Consortium, the I. G., Best, A. J., Knell, J., Goldrath, A., Jojic, V., Koller, D., … Turley, S. (2013). The transcriptional landscape of αβ T cell differentiation. *Nature Immunology*, *14*(6), 619. https://doi.org/10.1038/ni.2590

Mold, J. E., Michaëlsson, J., Burt, T. D., Muench, M. O., Beckerman, K. P., Busch, M. P., Lee, T.-H., Nixon, D. F., & McCune, J. M. (2008). Maternal Alloantigens Promote the Development of Tolerogenic Fetal Regulatory T Cells in Utero. *Science*, *322*(5907), 1562–1565. https://doi.org/10.1126/science.1164511

Mold, J. E., Réu, P., Olin, A., Bernard, S., Michaëlsson, J., Rane, S., Yates, A., Khosravi, A., Salehpour, M., Possnert, G., Brodin, P., & Frisén, J. (2019). Cell generation dynamics underlying naive T-cell homeostasis in adult humans. *PLOS Biology*, *17*(10), e3000383. https://doi.org/10.1371/journal.pbio.3000383

Mold, J. E., Venkatasubrahmanyam, S., Burt, T. D., Michaëlsson, J., Rivera, J. M., Galkina, S. A., Weinberg, K., Stoddart, C. A., & McCune, J. M. (2010). Fetal and Adult Hematopoietic Stem Cells Give Rise to Distinct T Cell Lineages in Humans. *Science*, *330*(6011), 1695–1699. https://doi.org/10.1126/science.1196509

Montecino-Rodriguez, E., Casero, D., Fice, M., Le, J., & Dorshkind, K. (2018). Differential Expression of PU.1 and Key T Lineage Transcription Factors Distinguishes Fetal and Adult T Cell Development. *The Journal of Immunology*, *200*(6), 2046–2056. https://doi.org/10.4049/jimmunol.1701336

Montecino-Rodriguez, E., & Dorshkind, K. (2012). B-1 B cell development in the fetus and adult. *Immunity*, *36*(1), 13–21. https://doi.org/10.1016/j.immuni.2011.11.017

Montecino-Rodriguez, E., Fice, M., Casero, D., Berent-Maoz, B., Barber, C. L., & Dorshkind, K. (2016). Distinct Genetic Networks Orchestrate the Emergence of Specific Waves of Fetal and Adult B-1 and B-2 Development. *Immunity*, *45*(3), 527–539. https://doi.org/10.1016/j.immuni.2016.07.012

Montecino-Rodriguez, E., Leathers, H., & Dorshkind, K. (2006). Identification of a B-1 B cell–specified progenitor. *Nature Immunology*, *7*(3), 293–301. https://doi.org/10.1038/ni1301

Moskowitz, D. M., Zhang, D. W., Hu, B., Saux, S. L., Yanes, R. E., Ye, Z., Buenrostro, J. D., Weyand, C. M., Greenleaf, W. J., & Goronzy, J. J. (2017). Epigenomics of human CD8 T cell differentiation and aging. *Science Immunology*, *2*(8), eaag0192. https://doi.org/10.1126/sciimmunol.aag0192

Ner-Gaon, H., Melchior, A., Golan, N., Ben-Haim, Y., & Shay, T. (2017). JingleBells: A Repository of Immune-Related Single-Cell RNA–Sequencing Datasets. *The Journal of Immunology*, *198*(9), 3375–3379. https://doi.org/10.4049/jimmunol.1700272

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Newport, M. J., Goetghebuer, T., Weiss, H. A., Whittle, H., Siegrist, C.-A., Marchant, A., & MRC Gambia Twin Study Group. (2004). Genetic regulation of immune responses to vaccines in early life. *Genes and Immunity*, *5*(2), 122–129. https://doi.org/10.1038/sj.gene.6364051

Ng, M. S. F., Roth, T. L., Mendoza, V. F., Marson, A., & Burt, T. D. (2019). Helios enhances the preferential differentiation of human fetal CD4+ naïve T cells into regulatory T cells. *Science Immunology*, *4*(41). https://doi.org/10.1126/sciimmunol.aav5947

Ntranos, V., Yi, L., Melsted, P., & Pachter, L. (2019). A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nature Methods*, *16*(2), 163–166. https://doi.org/10.1038/s41592-018-0303-9

Olin, A., Henckel, E., Chen, Y., Lakshmikanth, T., Pou, C., Mikes, J., Gustafsson, A., Bernhardsson, A. K., Zhang, C., Bohlin, K., & Brodin, P. (2018). Stereotypic Immune System Development in Newborn Children. *Cell*, *174*(5), 1277-1292.e14. https://doi.org/10.1016/j.cell.2018.06.045

Opstelten, R., Slot, M. C., Lardy, N. M., Lankester, A. C., Mulder, A., Claas, F. H. J., van Rood, J. J., & Amsen, D. (2019). Determining the extent of maternal-foetal chimerism in cord blood. *Scientific Reports*, *9*(1), 5247. https://doi.org/10.1038/s41598-019-41733-w

Palanichamy, A., Apeltsin, L., Kuo, T. C., Sirota, M., Wang, S., Pitts, S. J., Sundar, P. D., Telman, D., Zhao, L. Z., Derstine, M., Abounasr, A., Hauser, S. L., & von Büdingen, H.-C. (2014). Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Science Translational Medicine*, *6*(248), 248ra106. https://doi.org/10.1126/scitranslmed.3008930

Palmer, A. C. (2011). Nutritionally Mediated Programming of the Developing Immune System. *Advances in Nutrition*, *2*(5), 377–395. https://doi.org/10.3945/an.111.000570

Papalexi, E., & Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, *18*(1), 35–45. https://doi.org/10.1038/nri.2017.76

Patin, E., Hasan, M., Bergstedt, J., Rouilly, V., Libri, V., Urrutia, A., Alanio, C., Scepanovic, P., Hammer, C., Jönsson, F., Beitz, B., Quach, H., Lim, Y. W., Hunkapiller, J., Zepeda, M., Green, C., Piasecka, B., Leloup, C., Rogge, L., … Albert, M. L. (2018). Natural variation

in the parameters of innate immune cells is preferentially driven by genetic factors. *Nature Immunology*, *19*(3), 302–314. https://doi.org/10.1038/s41590-018-0049-7

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M.-P., & Audain, E. (2017). Accurate and fast feature selection workflow for high-dimensional omics data. *PLOS ONE*, *12*(12), e0189875. https://doi.org/10.1371/journal.pone.0189875

Phares, C. R., Lynfield, R., Farley, M. M., Mohle-Boetani, J., Harrison, L. H., Petit, S., Craig, A. S., Schaffner, W., Zansky, S. M., Gershman, K., Stefonek, K. R., Albanese, B. A., Zell, E. R., Schuchat, A., Schrag, S. J., & Active Bacterial Core surveillance/Emerging Infections Program Network. (2008). Epidemiology of invasive group B streptococcal disease in the United States, 1999-2005. *JAMA*, *299*(17), 2056–2065. https://doi.org/10.1001/jama.299.17.2056

Pineda, S., Sigdel, T. K., Liberto, J. M., Vincenti, F., Sirota, M., & Sarwal, M. M. (2019). Characterizing pre-transplant and post-transplant kidney rejection risk by B cell immune repertoire sequencing. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-09930-3

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C., & The Integrative HMP (iHMP) Research Network Consortium. (2019). The Integrative Human Microbiome Project. *Nature*, *569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

Psaila, B., & Mead, A. J. (2019). Single-cell approaches reveal novel cellular pathways for megakaryocyte and erythroid differentiation. *Blood*, *133*(13), 1427–1435. https://doi.org/10.1182/blood-2018-11-835371

R Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Ramond, C., Berthault, C., Burlen-Defranoux, O., de Sousa, A. P., Guy-Grand, D., Vieira, P., Pereira, P., & Cumano, A. (2014). Two waves of distinct hematopoietic progenitor cells colonize the fetal thymus. *Nature Immunology*, *15*(1), 27–35. https://doi.org/10.1038/ni.2782

Relman, D. A., & Lipsitch, M. (2018). Microbiome as a tool and a target in the effort to address antimicrobial resistance. *Proceedings of the National Academy of Sciences*, *115*(51), 12902–12910. https://doi.org/10.1073/pnas.1717163115

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Roskin, K. M., Simchoni, N., Liu, Y., Lee, J.-Y., Seo, K., Hoh, R. A., Pham, T., Park, J. H., Furman, D., Dekker, C. L., Davis, M. M., James, J. A., Nadeau, K. C., Cunningham-Rundles, C., & Boyd, S. D. (2015). IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Science Translational Medicine*, *7*(302), 302ra135. https://doi.org/10.1126/scitranslmed.aab1216

Rotival, M. (2019). Characterising the genetic basis of immune response variation to identify causal mechanisms underlying disease susceptibility. *HLA*, *94*(3), 275–284. https://doi.org/10.1111/tan.13598

RStudio Team. (2016). *RStudio: Integrated Development Environment for R* (1.1.456) [Computer software]. RStudio, Inc. http://www.rstudio.com/

Rubelt, F., Busse, C. E., Bukhari, S. A. C., Bürckert, J.-P., Mariotti-Ferrandiz, E., Cowell, L. G., Watson, C. T., Marthandan, N., Faison, W. J., Hershberg, U., Laserson, U., Corrie, B. D., Davis, M. M., Peters, B., Lefranc, M.-P., Scott, J. K., Breden, F., Prak, E. T. L., & Kleinstein, S. H. (2017). Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature Immunology*, *18*(12), 1274–1278. https://doi.org/10.1038/ni.3873

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Schlitzer, A., Sivakamasundari, V., Chen, J., Sumatoh, H. R. B., Schreuder, J., Lum, J., Malleret, B., Zhang, S., Larbi, A., Zolezzi, F., Renia, L., Poidinger, M., Naik, S., Newell, E. W., Robson, P., & Ginhoux, F. (2015). Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. *Nature Immunology*, *16*(7), 718–728. https://doi.org/10.1038/ni.3200

Schroeder, H. W. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental and Comparative Immunology*, *30*(1–2), 119–135. https://doi.org/10.1016/j.dci.2005.06.006

Servais, S., Hannon, M., Peffault de Latour, R., Socie, G., & Beguin, Y. (2017). Reconstitution of adaptive immunity after umbilical cord blood transplantation: Impact on infectious complications. *Stem Cell Investigation*, *4*(6), 40–40. https://doi.org/10.21037/sci.2017.05.03

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., & Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, *34*(6), 637–645. https://doi.org/10.1038/nbt.3569

Shugay, M., Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., Putintseva, E. V., Pogorelyy, M. V., Nazarov, V. I., Zvyagin, I. V., Kirgizova, V. I., Kirgizov, K. I., Skorobogatova, E. V., & Chudakov, D. M. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology*, *11*(11), e1004503. https://doi.org/10.1371/journal.pcbi.1004503

Siegrist, C.-A. (2001). Neonatal and early life vaccinology. *Vaccine*, *19*(25), 3331–3346. https://doi.org/10.1016/S0264-410X(01)00028-7

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Simeonov, D. R., Gowen, B. G., Boontanrart, M., Roth, T. L., Gagnon, J. D., Mumbach, M. R., Satpathy, A. T., Lee, Y., Bray, N. L., Chan, A. Y., Lituiev, D. S., Nguyen, M. L., Gate, R. E., Subramaniam, M., Li, Z., Woo, J. M., Mitros, T., Ray, G. J., Curie, G. L., … Marson, A. (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*, *549*(7670), 111–115. https://doi.org/10.1038/nature23875

Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H., & Butte, A. J. (2009). Autoimmune Disease Classification by Inverse Association with SNP Alleles. *PLOS Genetics*, *5*(12), e1000792. https://doi.org/10.1371/journal.pgen.1000792

Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N., & Brinkman, R. R. (2012). FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A*, *81A*(9), 727–731. https://doi.org/10.1002/cyto.a.22106

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A. S., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., Amos, C. I., Ardlie, K. G., Barton, A., Bowes, J., Brouwer, E., Burtt, N. P., … Plenge, R. M. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, *42*(6), 508–514. https://doi.org/10.1038/ng.582

Stamatoyannopoulos, G. (2005). Control of globin gene expression during development and erythroid differentiation. *Experimental Hematology*, *33*(3), 259–271. https://doi.org/10.1016/j.exphem.2004.11.007

Stewart, P. A., Welsh, E. A., Slebos, R. J. C., Fang, B., Izumi, V., Chambers, M., Zhang, G., Cen, L., Pettersson, F., Zhang, Y., Chen, Z., Cheng, C.-H., Thapa, R., Thompson, Z., Fellows, K. M., Francis, J. M., Saller, J. J., Mesa, T., Zhang, C., … Haura, E. B. (2019). Proteogenomic landscape of squamous cell lung cancer. *Nature Communications*, *10*. https://doi.org/10.1038/s41467-019-11452-x

Strauli, N. B., & Hernandez, R. D. (2016). Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Medicine*, *8*(1), 60. https://doi.org/10.1186/s13073-016-0314-z

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., & Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*. https://doi.org/10.1186/s12864-018-4772-0

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Theriot, C. M., & Young, V. B. (2015). Interactions Between the Gastrointestinal Microbiome and Clostridium difficile. *Annual Review of Microbiology*, *69*(1), 445–461. https://doi.org/10.1146/annurev-micro-091014-104115

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., … Shmulevich, I. (2018). The Immune Landscape of Cancer. *Immunity*, *48*(4), 812-830.e14. https://doi.org/10.1016/j.immuni.2018.03.023

Tieppo, P., Papadopoulou, M., Gatti, D., McGovern, N., Chan, J. K. Y., Gosselin, F., Goetgeluk, G., Weening, K., Ma, L., Dauby, N., Cogan, A., Donner, C., Ginhoux, F., Vandekerckhove, B., & Vermijlen, D. (2020). The human fetal thymus generates invariant effector γδ T cellsHuman fetal invariant effector γδ thymocytes. *The Journal of Experimental Medicine*, *217*(3). https://doi.org/10.1084/jem.20190580

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (New York, N.Y.)*, *352*(6282), 189–196. https://doi.org/10.1126/science.aad0501

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 12. https://doi.org/10.1186/s13059-019-1850-9

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, *14*(1), 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5116–5121. https://doi.org/10.1073/pnas.091062498

Unanue, E. R., Turk, V., & Neefjes, J. (2016). Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annual Review of Immunology*, *34*(1), 265–297. https://doi.org/10.1146/annurev-immunol-041015-055420

van den Broek, T., Borghans, J. A. M., & van Wijk, F. (2018). The full spectrum of human naive T cells. *Nature Reviews Immunology*, *18*(6), 363–373. https://doi.org/10.1038/s41577-018-0001-y

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., & Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, *3*(7), RESEARCH0034. https://doi.org/10.1186/gb-2002-3-7-research0034

Vijayaraghavan, J., & Osborne, B. A. (2018). Notch and T Cell Function – A Complex Tale. In T. Borggrefe & B. D. Giaimo (Eds.), *Molecular Mechanisms of Notch Signaling* (Vol. 1066, pp. 339–354). Springer International Publishing. https://doi.org/10.1007/978-3-319-89512-3_17

von Büdingen, H.-C., Kuo, T. C., Sirota, M., van Belle, C. J., Apeltsin, L., Glanville, J., Cree, B. A., Gourraud, P.-A., Schwartzburg, A., Huerta, G., Telman, D., Sundar, P. D., Casey, T., Cox, D. R., & Hauser, S. L. (2012). B cell exchange across the blood-brain barrier in multiple sclerosis. *The Journal of Clinical Investigation*, *122*(12), 4533–4543. https://doi.org/10.1172/JCI63842

Vora, B., Wang, A., Kosti, I., Huang, H., Paranjpe, I., Woodruff, T. J., MacKenzie, T., & Sirota, M. (2018). Meta-Analysis of Maternal and Fetal Transcriptomic Data Elucidates the Role of Adaptive and Innate Immunity in Preterm Birth. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.00993

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, *177*(7), 1873-1887.e17. https://doi.org/10.1016/j.cell.2019.05.006

Wong, B. (2011). Points of view: Color blindness. *Nature Methods*, *8*(6), 441–441. https://doi.org/10.1038/nmeth.1618

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yaari, G., & Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, *7*, 121. https://doi.org/10.1186/s13073-015-0243-2

Yu, B., Zhang, K., Milner, J. J., Toma, C., Chen, R., Scott-Browne, J. P., Pereira, R. M., Crotty, S., Chang, J. T., Pipkin, M. E., Wang, W., & Goldrath, A. W. (2017). Epigenetic landscapes reveal transcription factors that regulate $CD8^+$ T cell differentiation. *Nature Immunology*, *18*(5), 573–582. https://doi.org/10.1038/ni.3706

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., Wu, J., Lin, L., Yang, F., Gao, F., Wang, X., Li, T., Zhang, R., Saksena, N. K., Yang, H., … Liu, X. (n.d.). PIRD: Pan Immune Repertoire Database. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz614

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049. https://doi.org/10.1038/ncomms14049

Zhuang, L., Chen, H., Zhang, S., Zhuang, J., Li, Q., & Feng, Z. (2019). Intestinal Microbiota in Early Life and Its Implications on Childhood Health. *Genomics, Proteomics & Bioinformatics*, *17*(1), 13–25. https://doi.org/10.1016/j.gpb.2018.10.002

Zota, A. R., Mitro, S. D., Robinson, J. F., Hamilton, E. G., Park, J.-S., Parry, E., Zoeller, R. T., & Woodruff, T. J. (2018). Polybrominated diphenyl ethers (PBDEs) and hydroxylated PBDE metabolites (OH-PBDEs) in maternal and fetal tissues, and associations with fetal cytochrome P450 gene expression. *Environment International*, *112*, 269–278. https://doi.org/10.1016/j.envint.2017.12.030

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Daniel Burris
1A3CAEB053D74F9...
Author Signature

8/28/2020
Date