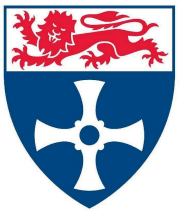
From origin to lifespan:

Unravelling early human developmental haematopoiesis and macrophage heterogeneity

Issac Emmanuel Goh Kai'en

Newcastle University



Thesis submitted to the Faculty Medical Sciences, Newcastle University,

for the degree of Doctor of Philosophy

June 2024

Abstract

This thesis explores early human developmental haematopoiesis, emphasising the integral roles of the extraembryonic yolk sac (YS) in generating the first blood and immune cells, including macrophages, and providing nutritional support to the embryo. Much of our current understanding of early haematopoiesis derives from pivotal studies in model systems. However, the application of single-cell genomics technologies have enabled this study to deliver a detailed multiomic characterisation of the multifunctional role the YS plays in early human development with temporal resolution.

As part of the Human Cell Atlas, I co-led a collaborative, multi-site effort to construct a time-resolved multiomics atlas of the human YS from the 3rd to the 8th post-conception weeks. This atlas integrates protein and gene expression data and is augmented with existing scRNA-seq data from eleven other prenatal tissues, including crucial haematopoietic tissues like the liver, bone marrow, and Aorta-Gonad-Mesonephros (AGM). This integrative approach facilitated the mapping of human early immune and blood development. Our findings elucidate the unique spatiotemporal characteristics of YS haematopoiesis fundamental to the emergence of the first blood and immune cells.

Our analyses reveal YS evolutionarily conserved roles in metabolism, coagulation, vascular development, and early haematopoietic regulation, with these functions gradually ceded as development progresses. We detailed the dynamic emergence and decline of YS Hematopoietic Stem and Progenitor Cells (HSPCs) which originate from YS hemogenic endothelium (HE), illustrating the earliest wave of haematopoiesis. We also identified a YS-specific macrophage production pathway, and a pre-specified, TREM2+ microglia-like macrophage subset across prenatal tissues.

This work illuminates a previously obscure phase of human development, establishing the YS as a vital multifunctional organ. Building this map of early haematopoiesis offers valuable insights into cellular differentiation pathways specific to early life, paving the way for novel tissue engineering strategies and cellular therapeutic avenues.

Acknowledgements

I am profoundly grateful to the tissue donors and their families, whose generosity has made this research possible. I have been extraordinarily fortunate in the environment, people, and resources that have been so generously provided throughout this critical stage of my training.

During my doctoral studies, I have had the privilege of meeting and working with remarkable individuals. I am especially thankful to Professor Muzlifah Haniffah, who, with generous funding from Wellcome, welcomed me as a staff PhD student and served as my supervisor. This role marked a pivotal moment in my development as a scientist, offering me an indispensable opportunity to embark on my academic journey. I am also deeply indebted to Dr. Botting, Dr. Jardine, Dr Popescu, and Dr Green, key figures in refining my scientific skills, and to Dr. Teichmann and team, who generously shared their expertise in computational methods. I extend my gratitude to Professor Bertie Göttgens and Professor Roberts for their insightful consultations and discussions on the origins of haematopoiesis.

I would like to acknowledge my progression panel: Professor Rajan and Dr. Stewart, for their unwavering support and encouragement during challenging times.

Additionally, I am thankful for all the wonderful colleagues at Haniffalab across both the Newcastle and Sanger sites, whose warmth and collaboration have enriched my experience.

My heartfelt thanks go to my family and friends, who have been my steadfast support network. Lastly, a special thank you to Olga Gumenyuk and her family, who have been my anchor, keeping me grounded while I navigated the challenges of a PhD during the COVID-19 pandemic.

Candidate declaration

This doctoral dissertation is the product of my own efforts, and due credit has been given to the works of others. None of the contents have been previously presented for any degree or qualification at this or any other educational institution.

This dissertation originates from a research project that involved collaborative scientific endeavours across various institutions. All analyses in this dissertation were performed by me and includes the compilation of peer-reviewed research and review papers in a way that accentuates my intellectual contributions. The ensuing section provides a catalogue of the articles that form the foundation of this dissertation, along with a description of my individual contributions. Significant inputs from others that are included in this dissertation are acknowledged below:

Droplet-based fetal YS scRNA-seq and CITE-seq datasets were collaboratively generated by Dr Rachel Botting, Dr Emily Stephenson, Justin Engelbert, and myself. Plate-based scRNA-seq data generation, including FACS isolation, were performed by Dr Emily Stephenson, Dr Rachel Botting and Dr Laura Jardine. Light-sheet microscopy was performed by Dr Yorick Gitton, and Megumi Inoue. Immunohistochemistry assays were performed by Dr Meghan Acres and Rowan Coulthard. RNA-sope experiments were conducted by Nana-Jane Chipampe and Kwasi Kwaka. I assembled and analysed the reference scRNA-seq datasets on: human YS , AGM, skin, brain , gonads, thymus, gut, kidneys, liver, spleen, bone marrow, and MLN, and mouse YS, and liver. Exploration and annotation of all scRNA-seq datasets were performed jointly by myself, Rachel Botting, and Dr Laura Jardine. Probabilistic latent-representation projection methods, CITE-seq preprocessing pipelines including denoising with GMM-based methods and clustered gene-set annotation analyses were written and performed by myself and tested by Antony Rose.

List of Publications

* - authors contributed equally

1. **Issac Goh***, Rachel A. Botting*, Antony Rose, Simone Webb, Justin Engelbert, Yorick Gitton, Emily Stephenson, Mariana Quiroga Londoño, Michael Mather, Nicole Mende, Ivan Imaz-Rosshandler, Lu Yang, Dave Horsfall, Daniela Basurto-Lozada, Nana-Jane Chipampe, Victoria Rook, Jimmy Tsz Hang Lee, Mai-Linh Ton, Daniel Keitley, Pavel Mazin, M.S. Vijayabaskar, Rebecca Hannah, Laure Gambardella, Kile Green, Stephane Ballereau1, Megumi Inoue, Elizabeth Tuck, Valentina Lorenzi, Kwasi Kwakwa, Clara Alsinet, Bayanne Olabi, Mohi Miah, Chloe Admane, Dorin-Mirel Popescu, Meghan Acres, David Dixon, Thomas Ness, Rowen Coulthard, Steven Lisgo, Deborah J Henderson, Emma Dann, Chenqu Suo, Sarah J. Kinston, Jong-eun Park, Krzysztof Polanski, John Marioni, Stijn van Dongen, Kerstin B. Meyer, Marella de Bruijn, James Palis, Sam Behjati, Elisa Laurenti, Nicola K. Wilson, Roser Vento-Tormo, Alain Chédotal, Omer Bayraktar, Irene Roberts, Laura Jardine*, Berthold Göttgens*, Sarah A. Teichmann*, Muzlifah Haniffa*. **2023.** Yolk sac cell atlas reveals multiorgan functions during human early development. ***Science***, https://doi.org/10.1101/2022.08.03.502475

   I led the computational analysis, in-silico experimental design and development of novel analytical architectures for this collaborative and interdisciplinary project. I leveraged single cell multiomic data from over 3.6million single cells across multiple datasets and research institutions. I am co-first author for this study which forms the basis of my PhD project and is published in *Science*. The results chapters of my thesis are largely based on the structure and results from this article. (*1*)

2. Chenqu Suo*, Emma Dann*, **Issac Goh**, Laura Jardine, Vitalii Kleshchevnikov, Jong-Eun Park, Rachel A. Botting, Emily Stephenson, Justin Engelbert, Zewen Kelvin Tuong, Krzysztof Polanski, Nadav Yayon, Chuan Xu, Ondrej Suchanek, Rasa Elmentaite, Cecilia Domínguez Conde, Peng He, Sophie Pritchard, Mohi Miah, Corina Moldovan, Alexander S. Steemers, Martin Prete, John C. Marioni, Menna R. Clatworthy, Muzlifah Haniffa, Sarah A. Teichmann. 2022**.** Mapping the developing human immune system across organs. ***Science***, https://doi.org/10.1101/2022.01.17.476665

   I provided analytical architecture design inputs and led the analysis and annotation of all myeloid and progenitor subsets in this data. I led the generation and analysis coordination for the spatial visium data analysed in this comprehensive atlasing project. I am the second author for this study which contributes significantly to the multi-organ integration and atlas used in this project. Results from the myeloid analysis portion of this article are integrated into portions of this thesis. (*2*)

3. Laura Jardine*, Simone Webb***, Issac Goh**, Mariana Quiroga Londoño, Gary Reynolds, Michael Mather, Bayanne Olabi, Emily Stephenson, Rachel A. Botting, Dave Horsfall, Justin Engelbert, Daniel Maunder, Nicole Mende, Caitlin Murnane, Emma Dann, Jim McGrath, Hamish King, Iwo Kucinski, Rachel Queen, Christopher D Carey, Caroline Shrubsole, Elizabeth Poyner, Meghan Acres, Claire Jones, Thomas Ness, Rowen Coulthard, Natalina Elliott, Sorcha O'Byrne, Myriam L. R. Haltalli, John E Lawrence, Steven Lisgo, Petra Balogh, Kerstin B Meyer, Elena Prigmore, Kirsty Ambridge, Mika Sarkin Jain, Mirjana Efremova, Keir Pickard, Thomas Creasey, Jaume Bacardit, Deborah Henderson, Jonathan Coxhead, Andrew Filby, Rafiqul Hussain, David Dixon, David McDonald, Dorin Mirel-Popescu, Monika S. Kowalczyk, Bo Li, Orr Ashenberg, Marcin Tabaka, Danielle Dionne, Timothy L. Tickle, Michal Slyper, Orit Rozenblatt-Rose, Aviv Regev, Sam Behjati, Elisa Laurenti, Nicola K. Wilson*, Anindita Roy*, Berthold Göttgens*, Irene Roberts*, Sarah A. Teichmann*, Muzlifah Haniffa*. **2021.** Blood and immune development in human fetal bone marrow and Down syndrome. ***Nature***, https://10.1038/s41586-021-03929-x

   I supported the computational analysis of this project, and led the development of novel probabilistic projection methodologies and NLP summarisation methods for annotations of regulatory and gene set enrichment methods. The data output from this project is integrated into the multi-organ landscape used in this thesis. (*3*)

4. Mohi Miah***, Issac Goh*,** and Muzlifah Haniffa. **2021.** Prenatal Development and Function of Human Mononuclear Phagocytes. ***frontiers in cell and developmental biology*****,** https://doi.org/10.3389/fcell.2021.649937

   As co-first author of this review article, I conceived, researched and wrote on the development and functional diversification of the mononuclear phagocyte system in humans. I included how recent developments in single-cell technologies and in-siloco modelling are revolutionising our understanding of the prenatal MPS system and highlighted future roles for collaborative studies into such big data. I also designed a web portal instance for quick reference of all markers provided in the review. This review article is integrated into the introduction of chapter 3 in this thesis. (*4*)

5. Gary Reynolds*, Peter Vegh*, James Fletcher*, Elizabeth F. M. Poyner*, Emily Stephenson, **Issac Goh**, Rachel A. Botting, Ni Huang, Bayanne Olabi, Anna Dubois, David Dixon, Kile Green, Daniel Maunder, Justin Engelbert, Mirjana Efremova, Krzysztof Polanski, Laura Jardine, Claire Jones, Thomas Ness, Dave Horsfall, Jim McGrath, Christopher Carey, Dorin-Mirel Popescu, Simone Webb, Xiao-nong Wang, Ben Sayer, Jong-Eun Park, Victor A. Negri, Daria Belokhvostova, Magnus D. Lynch, David McDonald, Andrew Filby, Tzachi Hagai, Kerstin B. Meyer, Akhtar Husain, Jonathan Coxhead, Roser Vento-Tormo, Sam Behjati, Steven Lisgo, Alexandra-Chloé Villani,

Jaume Bacardit, Philip H. Jones, Edel A. O'Toole, Graham S. Ogg, Neil Rajan, Nick J. Reynolds, Sarah A. Teichmann, Fiona M. Watt, Muzlifah Haniffa**. 2021.** Developmental cell programs are co-opted in inflammatory skin disease. ***Science*****,** 371 (6527).

I designed novel analytical architectures published with this study including a NLP summarisation method for gene ontology graph embedding, and generalised linear models for cell-cell correspondence between disease and healthy data. I was additionally involved in drafting the manuscript. (*5*)

6. Dorin-Mirel Popescu*, Rachel A. Botting*, Emily Stephenson*, Kile Green, Simone Webb, Laura Jardine, Emily F. Calderbank, Krzysztof Polanski, **Issac Goh**, Mirjana Efremova, Meghan Acres, Daniel Maunder, Peter Vegh, Yorick Gitton, Jong-Eun Park, Roser Vento-Tormo, Zhichao Miao, David Dixon, Rachel Rowell, David McDonald, James Fletcher, Elizabeth Poyner, Gary Reynolds, Michael Mather, Corina Moldovan, Lira Mamanova, Frankie Greig, Matthew D. Young, Kerstin B. Meyer, Steven Lisgo, Jaume Bacardit, Andrew Fuller, Ben Millar, Barbara Innes, Susan Lindsay, Michael J. T. Stubbington, Monika S. Kowalczyk, Bo Li, Orr Ashenberg, Marcin Tabaka, Danielle Dionne, Timothy L. Tickle, Michal Slyper, Orit Rozenblatt-Rosen, Andrew Filby, Peter Carey, Alexandra Chloé Villani, Anindita Roy, Aviv Regev, Alain Chédotal, Irene Roberts, Berthold Göttgens, Sam Behjati, Elisa Laurenti, Sarah A. Teichmann & Muzlifah Haniffa. **2019.** Decoding human fetal liver haematopoiesis. ***Nature*****,** 574 (7778), 365-371.

   I co-led the execution of computational analyses via pre-written pipelines during the revision process of this study. I additionally hypothesised and performed analysis to show interactions which influence B cell survival in developing tissues and contributed to writing the manuscript. Data from this study was integrated into the multi-organ atlas mentioned throughout this thesis. (*6*)

7. Jong-Eun Park*, Rachel A. Botting, Cecilia Domínguez Conde, Dorin-Mirel Popescu, Marieke Lavaert, Daniel J. Kunz, **Issac Goh**, Emily Stephenson, Roberta Ragazzini, Elizabeth Tuck, Anna Wilbrey-Clark, Kenny Roberts, Veronika R. Kedlian, John R. Ferdinand, Xiaoling He, Simone Webb, Daniel Maunder, Niels Vandamme, Krishnaa T. Mahbubani, Krzysztof Polanski, Lira Mamanova, Liam Bolt, David Crossland, Fabrizio de Rita, Andrew Fuller, Andrew Filby, Gary Reynolds, David Dixon, Kourosh Saeb-Parsy, Steven Lisgo, Deborah Henderson, Roser Vento-Tormo, Omer A. Bayraktar, Roger A. Barker, Kerstin B. Meyer, Yvan Saeys, Paola Bonfanti, Sam Behjati, Menna R. Clatworthy, Tom Taghon, Muzlifah Haniffa, Sarah A. Teichmann. **2020.** A cell atlas of human thymic development defines T cell repertoire formation. ***Science,*** 367 (6480).

   I provided computational input, processed clinical samples and performed the initial mouse to human comparisons and homology mapping which influenced the final analytical conclusions in this study. I additionally generated and maintained the webportal resource published with this study. (*7*)

8. Rasa Elmentaite, Natsuhiko Kumasaka, Kenny Roberts, Aaron Fleming, Emma Dann, Hamish W. King, Vitalii Kleshchevnikov, Monika Dabrowska, Sophie Pritchard, Liam Bolt, Sara F. Vieira, Lira Mamanova, Ni Huang, Francesca Perrone, **Issac Goh Kai'En**, Steven N. Lisgo, Matilda Katan, Steven Leonard, Thomas R. W. Oliver, C. Elizabeth Hook, Komal Nayak, Lia S. Campos, Cecilia Domínguez Conde, Emily Stephenson, Justin Engelbert, Rachel A. Botting, Krzysztof Polanski, Stijn van Dongen, Minal Patel, Michael D. Morgan, John C. Marioni, Omer Ali Bayraktar, Kerstin B. Meyer, Xiaoling He, Roger A. Barker, Holm H. Uhlig, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Matthias Zilbauer, Menna R. Clatworthy, Muzlifah Haniffa, Kylie R. James & Sarah A. Teichmann. **2021.** Cells of the human intestinal tract mapped across space and time. ***Nature,*** 597, 250–255 (2021).

   I provided computational advice to the lead authors for single cell analysis, shared code, and helped draft manuscript methods. I also processed, snap-froze and provided guidance for tissue digestion and acquisition for this study. (*8*)

9. He P, Lim K, Sun D, Pett JP, Jeng Q, Polanski K, Dong Z, Bolt L, Richardson L, Mamanova L, Dabrowska M, Wilbrey-Clark A, Madissoon E, Tuong ZK, Dann E, Suo C, **Goh I,** Yoshida M, Nikolić MZ, Janes SM, He X, Barker RA, Teichmann SA, Marioni JC, Meyer KB, Rawlins EL. **2022.** A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. **Cell**. 8;185(25):4841-4860.e25.

   I provided logistic regression models of data that I had trained from data integrated and annotated by myself to be used as reference in this study. I further provided computational and analytical input. (*9*)

10. Mariana Quiroga Londoño, Nicole Mende, Emily Stephenson, Deena Iskander, Simone Webb, **Issac Goh**, Vijaya Mahalingam Shanmugiah, Anindita Roy, Irene Roberts, Elisa Laurenti, Muzlifah Haniffa, Nicola K Wilson, Berthold Göttgens. **2022.** A PROTEIN‐TRANSCRIPTOME ATLAS OF HAEMATOPOIESIS ACROSS THE HUMAN LIFESPAN. ***Experimental Hematology***. https://doi.org/10.1016/j.exphem.2022.07.222

    I advised the lead author on computational approaches for single cell multi omic data and provided my own code for the purposes of analysis, preprocessing and testing different analytical approaches. (*10*)

11. Natalia Alkon, Wolfgang M. Bauer, Thomas Krausgruber, **Issac Goh**, Johannes Griss, Vy Nguyen, Baerbel Reininger, Christine Bangert, Clement Staud, Patrick M. Brunner, Christoph Bock, Muzlifah Haniffa, Georg Stingl. **2020.** Single-cell analysis reveals innate

lymphoid cell lineage infidelity in atopic dermatitis. ***Journal of Allergy and Clinical Immunology.*** 10.1016/j.jaci.2021.07.025.

I provided single cell analysis computational and analytical advice to the lead authors on this study, provided analytical guidance, assistance in running my code, and helped draft manuscript methods. (*11*)

12. Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, Masahiro Yoshida, Natsuhiko Kumasaka, Katarzyna Kania, Justin Engelbert, Bayanne Olabi, Jarmila Stremenova Spegarova, Nicola K Wilson, Nicole Mende, Laura Jardine, Louis CS Gardner, **Issac Goh,** Dave Horsfall, Jim McGrath, Simone Webb, Michael W Mather, Rik GH Lindeboom, Emma Dann, Ni Huang, Krzysztof Polanski, Elena Prigmore, Florian Gothe, Jonathan Scott, Rebecca P Payne, Kenneth F Baker, Aidan T Hanrath, Ina CD Schim van der Loeff, Andrew S Barr, Amada Sanchez-Gonzalez, Laura Bergamaschi, Federica Mescia, Josephine L Barnes, Eliz Kilich, Angus de Wilton, Anita Saigal, Aarash Saleh, Sam M Janes, Claire M Smith, Nusayhah Gopee, Caroline Wilson, Paul Coupland, Jonathan M Coxhead, Vladimir Yu Kiselev, Stijn van Dongen, Jaume Bacardit, Hamish W King, Anthony J Rostron, A John Simpson, Sophie Hambleton, Elisa Laurenti, Paul A Lyons, Kerstin B Meyer, Marko Z Nikolić, Christopher JA Duncan, Kenneth GC Smith, Sarah A Teichmann, Menna R Clatworthy, John C Marioni, Berthold Göttgens, Muzlifah Haniffa. **2021**. Single-cell multi-omics analysis of the immune response in COVID-19. ***Nature Medicine*****,** 27 (5) 904-916.

I provided conceptual input for computational approaches and analytical code to the lead authors for use in this study. (*12*)

13. Ni Huang, Bayanne Olabi, Chloe Admane, Rachel A. Botting, April Rose Foster, Fereshteh Torabi, Elena Winheim, Dinithi Sumanaweera, **Issac Goh**, Mohi Miah, Emily Stephenson, Win Min Tun, Pejvak Moghimi, Ben Rumney, Peng He, Sid Lawrence, View ORCID ProfileKenny Roberts, Keval Sidhpura, Justin Englebert, Laura Jardine, Gary Reynolds, Antony Rose, View ORCID ProfileClarisse Ganier, Vicky Rowe, Sophie Pritchard, Ilaria Mulas, James Fletcher, Dorin-Mirel Popescu, Elizabeth Poyner, Anna Dubois, Andrew Filby, Steven Lisgo, Roger A. Barker, Jong-Eun Park, Roser Vento-Tormo, Phuong Ahn Le, Sara Serdy, Jin Kim, CiCi Deakin, Jiyoon Lee, Marina Nikolova, Neil Rajan, Stephane Ballereau, Tong Li, View ORCID ProfileJosh Moore, David Horsfall, Daniela Basurto Lozada, Edel A. O'Toole, Barbara Treutlein, Omer Bayraktar, Maria Kasper, Pavel Mazin, Laure Gambardella, Karl Koehler, Sarah A. Teichmann, Muzlifah Haniffa. **2023**. A human prenatal skin cell atlas reveals immune cell regulation of skin morphogenesis. **biorxiv**, 10.1101/2023.10.12.556307.

I provided computational advice to the lead authors, and wrote both code and methodological sections for suitable probabilistically approaches to perform label

harmonisation, transfer learning, and time-conditioned comparative analyses between developing in-vivo scRNA-seq data of fetal skin origin against in-vitro derived hair-bearing organoid systems. (*13*)

List of Figures

Abbreviations

AD	–	atopic dermatitis
AGM	–	aorta-gonad-mesonephros
AML	–	acute myeloid leukaemia
B-ALL	–	B cell acute lymphoblastic leukaemia
BM	–	bone marrow
CITE-seq	–	cellular indexing of transcriptomes and epitopes by sequencing
CLP	–	common lymphoid progenitor
CMP	–	common myeloid progenitor
CNV	–	copy number variation
CyTOF	–	cytometry by time of flight
DC	–	dendritic cell
DEG	–	differentially expressed genes
DSB	–	denoised and scaled by background
EC	–	endothelial cell
ECM	–	extracellular matrix
EHT	–	endothelial cell to haematopoietic transition
EI	–	erythroid island
EL	–	embryonic liver
ELP	–	early lymphoid progenitor
EN	–	elasticnet regression
FACS	–	fluorescence activated cell sorting
FDG	–	force directed graph
FDR	–	false discovery rate
FFPE	–	Formalin-Fixed Paraffin-Embedded
FISH	–	fluorescence in situ hybridisation
FL	–	fetal liver
GAM	–	general additive model
GLM	–	general linear model
GMP	–	granulocyte monocyte precursor
GMM	–	gaussian mixture model
GO	–	gene ontology
GRN	–	gene regulatory network
GWAS	–	genome wide association studies
HCA	–	human cell atlas
HDBR	–	human developmental biology resource
HDCA	–	human developmental cell atlas
HE	–	hemogenic endothelium
HPC	–	high performance cluster

HSC	–	haematopoietic stem cell
HSPC	–	haematopoietic stem and progenitor cells
HSPC (C.)	–	cycling haematopoietic stem and progenitor cells
HVG	–	highly variable gene
ILC	–	innate lymphoid cell
KNN	–	k-nearest neighbour
LMPP	–	lymphoid-primed multipotent progenitor
LR	–	logistic regression
mAb	–	monoclonal antibody
MEM	–	megakaryocyte-erythroid mast cell
MEMP	–	megakaryocyte-erythroid mast cell progenitor
MEP	–	megakaryocyte-erythroid progenitor
MHC	–	major histocompatibility complex
MK	–	megakaryocyte
MOP	–	monocyte progenitor
MPP	–	multipotent progenitor
MPS	–	mononuclear phagocyte system
MSC	–	mesenchymal stem cell
NGS	–	next generation sequencing
PBMC	–	peripheral blood mononuclear cells
PCA	–	principal component analysis
PCW	–	post-conception weeks
pDC	–	plasmacytoid dendritic cell
PS	–	Psoriasis
QC	–	quality control
scATAC-seq	–	single cell assay for transposable-accessible chromatin with sequencing
scNMT-seq	–	single cell nucleosome, methylation and transcription sequencing
scRNA-seq	–	single cell RNA sequencing
SS2	–	Smart-seq2
SNP	–	single nucleotide polymorphism
TF	–	transcription factor
tSNE	–	t-distributed stochastic neighbour embedding
UMAP	–	uniform manifold approximation and projection
UMI	–	unique molecular identifier
YS	–	yolk sac

Table of Contents

1 Literature overview

In this chapter, I provide an overview of current literature and research on single-cell multiomics, the landscape of human developmental embryology, early human hematopoiesis, and the developmental origins and ontogeny of the human mononuclear phagocyte system.

This chapter includes lightly-edited portions of two manuscripts I co-first authored (*1*, *4*). The adapted work results from collaboration within the Haniffa group, led by Professor Muzlifah Haniffa (Biosciences Institute, Newcastle University, and the Wellcome Sanger Institute).

Myself, Dr. Rachel Botting, Dr. Laura Jardine, Dr. Simone Webb, wrote and proofread the original manuscript sections on the human developing yolk sac and its evolutionarily conserved roles. Section 1.1 of this overview contains adaptations from this manuscript, with additional detail around the developmental origins of the human yolk sac added for this section (*1*). Section 1.2 contains adaptations with additional detail from a literature review co-written by Mohi Miah and myself (*4*).

1.1 Single-Cell Multi-Omics: Applications in Developmental Biology and Beyond

The advent of Single-cell RNA sequencing (scRNA-seq) marked a significant milestone in the early 1990s when Iscove *et al.* managed to amplify mRNA from an individual cell, enabling detailed transcriptomic profiling at the single-cell level (*14*). In 2009, work by Tang *et al.* further refined this concept with next-generation sequencing (NGS) technology spurring the current modern development of methodologies that delineate transcriptional diversity and expression patterns within single cells (*15*). This innovation provided a leap beyond bulk RNA-seq methods, which could only offer averaged data from mixed cell populations (*16*, *17*). Modern droplet-based scRNA-seq allows for high-throughput sequencing of individual cells, revealing the intricate details of cellular heterogeneity and uncovering rare cell types and transient cell states (*2*, *6*, *18*, *19*). This is further pronounced by its use in decoding previously understudied tissues, perturbed conditions, and uncharacterised cell states (*1–3*, *5*). For example, a study by Villani *et al.* in 2017 utilised scRNA-seq to redefine the populations of dendritic cells and monocytes in human blood, showcasing the power of this technology to unveil new cellular landscapes (*19*). Similarly, the Haniffa lab used scRNA-seq to map the immune cell landscape in human healthy, diseased and developing skin, contrasting their findings with data derived from other developing tissues as part of the Human Developmental

Cell Atlas (HDCA) and the broader Human Cell Atlas (HCA). This work identified novel cell states and elucidated their roles in disease and development (*1–3*, *5–7*, *13*).

Compared to traditional bulk sequencing methods, single-cell approaches offer a more granular view of cellular diversity and dynamics. Bulk RNA-seq captures the overall RNA content of a cell population, often obscuring the nuances of individual cell types (*16*, *17*). In contrast, scRNA-seq provides insights into the distinct transcriptomic profiles of single cells, facilitating the identification of unique cell populations and developmental trajectories (*20*, *21*). For example, work by Reynolds *et al.* in 2021 demonstrated the discovery of key transcriptional and inferred regulatory mechanisms underlying pathological deregulations of macrophages and vasculature in Atopic Dermatitis (AD) and Psoriatic (PS) skin using the scRNA-seq modality (*5*). This level of detail is crucial for constructing accurate gene regulatory networks and understanding interactions within cellular microenvironments (*5*). While scRNA-seq excels in revealing novel gene expression programs in both developmental and perturbed systems, it lacks resolution in translational proteomic and regulatory epigenetic modalities (*22*, *23*). Organisms have evolved gene expression mechanisms that utilise both functional and copy redundancy, complicating the discovery of causal mechanisms in complex polygenic or dysregulation-related diseases (*24–26*). Functional redundancy, where different genes or proteins can perform similar functions, and redundancy, involving multiple copies or closely related versions of genes, enhance biological robustness and adaptability (*24*). However, these features pose significant challenges for genomics-based disease research. The compensatory mechanisms obscure the direct links between genetic variations and phenotypic outcomes, making it difficult to identify disease-associated genes (*24–27*). By observing the transcriptomic landscape and mapping cDNA read sequences to an annotated genomic reference, it is possible to identify which genes, and to some extent gene isoforms, are active in specific temporal snapshots of individual cells. This ability to recover gene isoform expression is however caveated by the resolution afforded by the depth of sequencing and corresponding length of read sequences (*21*, *28*). Additionally, by comparing the proportions of spliced and unspliced transcript variants, we gain insights into the burst kinetics of transcription and the current expression programs being activated (*29–31*). Our interpretation in this space relies on the assumption that the transcriptional landscape is part of a series of interactions between different modality layers (transcriptomic, proteomic, and epigenetic), corresponding to cell states and their interactions with local environments (*32*, *33*). These expression programs have evolved to address problems of metabolism, local tissue

architectural modifications, and pathogenic threats (*2*, *3*, *5*, *34*). However, scRNA-seq does not capture regulatory mechanisms of translation, which are crucial since proteins are the primary effectors of cellular function. To gain a comprehensive understanding of cellular functions influenced by translational regulation and post-translational modifications, scRNA-seq must be complemented with proteomic techniques. Surface proteome profiling, through techniques such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), narrow down the search space by providing simultaneous quantification of surface protein markers using oligonucleotide-labelled antibody tags, and RNA transcripts. This allows for simultaneous measurement of transcriptomes and surface proteomes in single cells, revealing discrepancies between RNA and protein levels, further enhancing the resolution of cell state identification (*23*, *35*).

Additionally, methodologies that characterise transposase-accessible regions, like single cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq), fill in the gaps regarding transcriptional regulation and gene program landscapes.These techniques provide insights into the chromatin accessibility and regulatory elements that control gene expression, thereby offering a more complete picture of cellular regulatory mechanisms (*22*, *35*, *36*).

The integration of various modalities, including RNA, chromatin accessibility, methylation, and select protein signatures, has enabled a more nuanced definition of cell types and states. Data increasingly derived from scRNA-seq, CITE-seq, and scATAC-seq, provide a comprehensive view of cellular function and regulation (*23*, *35*, *37–39*). However, current cell-type definitions and lineage assignments in humans often rely on frameworks derived from lineage tracing in model organisms and adult cellular profiles (*40*, *41*). These approaches may not fully capture the complexity of prenatal cell types, transient developmental states, and transitional differentiation states (*35*, *40*, *42*, *43*). Therefore, applying well-nuanced, data-driven, multi-omic approaches that account for known cellular ontogeny and standardised ontology definitions is essential to accurately depict the dynamic landscape of cellular development and function (*1–3*, *5–7*, *13*, *35*, *44–46*). Applications of single-cell multi-omics in developmental biology extend to understanding human developmental processes and their implications for health and disease. Historically, developmental biology research has focused on model organisms due to the practical challenges of studying human development directly (*47*). However, single-cell technologies have revolutionised our ability to study human development at high resolution (*47*). By building comprehensive, tissue-specific atlases of human development through large collaborative efforts and consortia

such as the HDCA, we can now trace the progression of haematopoiesis from early embryos from Carnegie Stage (CS) 4, roughly 3 post-conception weeks (PCW), through various stages of embryonic and fetal development (*1–3*, *5–7*, *13*, *35*, *44–46*). These atlases are crucial for identifying the origins of congenital disorders and childhood cancers such as childhood acute myeloid leukaemia (AML) (*48*, *49*), and childhood B-cell Acute Lymphoblastic Leukaemia (B-ALL), which often arise during critical windows of development (*50*). Understanding the molecular mechanisms of human development has far-reaching implications for regenerative medicine, cancer research, and stem cell therapies. For instance, profiling the independent early and definitive waves of haematopoiesis in the embryonic Yolk Sac (YS) and Aorta-Gonad-Mesonephros (AGM), respectively, during development may allow us to disentangle specific causal mechanisms and distinct regulatory programmes underlying how pleiotropic developmental mechanisms are co-opted to modify tissue architecture in various pathological conditions. This includes the pro-angiogenic role of YS-derived microglia and macrophages in the pathogenesis of glioblastoma (*51–53*), as well as the similarly pro-angiogenic role of F13A1+ macrophages, which acquire fetal macrophage transcriptional profiles in AD and PS skin (*5*). Additionally, developmental trajectories and stem cell dynamics provide essential references for engineering human stem-cell-derived models, organoids, and cellular therapies, enhancing our capacity for translational research and therapeutic interventions (*54*, *55*).

1.1.1 Single-Cell Multi-Omic Technologies: A Conceptual Overview

Genomics focuses on the comprehensive characterisation of whole genomes using various advanced techniques. Historically, Sanger sequencing was used for DNA sequencing, but it was limited by high costs and low throughput (*56*). In the early 2000s, high-density DNA microarrays, and the recognition that linkage disequilibrium (LD) allows a sparse set of "tag" SNPs to capture the bulk of common variation, made large-scale genome-wide association studies (GWAS) feasible at a wide scale for the first time (*57–59*). Next-generation sequencing (NGS), which was subsequently made commercially available around 2005, has thereafter revolutionised genomics by enabling massively parallel sequencing, significantly reducing costs and increasing throughput (*60*). NGS enhanced GWAS by enabling sequence-based discovery of novel variants and fine-mapping of association signals, and, when applied to exomes, pinpointed rare, protein-coding mutations to uncover gene associations with various disorders (*61*). For example, the 1000 Genomes Project utilised NGS to provide comprehensive insights into human genetic variation (*62*). Fluorescence in

situ hybridisation (FISH) continues to provide insights into genetic architecture and chromosomal abnormalities by localising specific DNA sequences on chromosomes (*63*). Haplotyping, which determines the specific combination of alleles at multiple loci on a single chromosome inherited together from one parent, is crucial for understanding genetic linkage and trait inheritance (*64*). It plays a significant role in GWAS by identifying haplotypes associated with diseases, thereby pinpointing genomic regions contributing to pathogenic traits (*65*). This approach enhances GWAS resolution and allows for the identification of causal variants that may be missed when examining individual Single Nucleotide Polymorphisms (SNPs) (*66*). Haplotyping also improves the accuracy of imputation, where missing genotype data is inferred based on known haplotypes, further increasing GWAS power (*67*). For example, in studying complex diseases like type 2 diabetes, haplotyping can reveal combinations of genetic variants that together increase disease risk (*65*, *68*).

The decreasing cost of NGS has also enabled whole-genome single-cell sequencing, enabling the study of genomic clonotype variation across single cells including the landscape of cancer cell clonotypes (*69*). This provides insights into cellular heterogeneity and the genetic underpinnings of complex traits and diseases.

Proteomics involves the large-scale study of proteins, particularly their structures and functions. Flow cytometry is a critical technology in proteomics, enabling the detection, quantification, and sorting of single cells based on their protein expression profiles. It uses fluorescently labelled antibodies to target specific proteins on the cell surface or within cells. However, fluorescence-based methods suffer from spectral overlap, and autofluorescence, limiting reliable multiplexing to fewer than 15 markers (*70*). The development of mass cytometry (CyTOF) has further advanced single-cell proteomics and quantification by using heavy metal-labelled antibodies to detect multiple proteins simultaneously, allowing simultaneous detection of both intracellular and extracellular proteins at higher plexity (≥31 markers), improving over the spectral-overlap limitations of fluorescence-based detection methods (*70*). However, because CyTOF vaporizes cells during ionization, it cannot be used for live-cell sorting unlike flow cytometry. For example, Palii et al. (2019) used CyTOF to track changes in transcription factor expression during erythropoiesis, revealing cell-fate decisions in individual cells (*71*). However, despite raising the plex ceiling, CyTOF data still imposes a finite marker limit and modest throughput compared with purely label-free approaches, restricting its utility as a single-cell proteomic technique. Recent advances in single-cell mass spectrometry have added a new dimension to our understanding of cellular

protein function and molecular characteristics. For instance, the Liquid chromatography tandem mass spectrometry analysis (LC-MS/MS) family of techniques allow for the label-free, simultaneous measurement of numerous proteins at single-cell resolution, providing insights into protein dynamics and regulatory networks that cannot be captured by transcriptomics alone (*72*). However, simultaneous scRNA-seq on the same cells is still not possible. Though it is limited to surface proteins and is not a label-free technology, CITE-seq uniquely allows for simultaneous capture of both scRNA-seq data and surface proteome data, albeit with critical experimental caveats (*23*, *35*). Specifically, careful optimization of antibody panels is essential to avoid non-specific binding and ensure accurate protein quantification. Optimization typically involves precise antibody titration, ideally validated via sequencing-based approaches or approximated using flow cytometry (*23*, *73*).

Commercially available antibody panels, such as the TotalSeqA Human Universal Cocktail (v1.0) panel from BioLegend, provide pre-optimized antibody solutions tailored for specific tissue types or experimental contexts. For instance, the TotalSeq™-A (V1) panel contains 154 antibody clones targeting major immune cell lineages, including T cells, B cells, monocytes, and NK cell immune populations with optimization performed primarily on human peripheral blood mononuclear cells (PBMC) by BioLegend using sequencing-based titration methods. Specifically, these antibody clones were titrated and validated through direct sequencing of antibody-derived tags (ADTs) to establish optimal concentrations and ensure accurate quantification of protein expression (*23*, *73*). Consequently, the performance of such pre-optimized panels may vary considerably when applied to other cell types or tissues, requiring additional user-driven optimization. Critical experimental considerations, including thorough Fc receptor blocking, rigorous washing protocols, and stringent standards for cell viability (>95%), are also essential to ensure accurate surface protein profiling and to minimize technical artifacts (*23*, *73*). For instance Stephenson *et al.* characterised the coordinated lymphocyte expansion response and increased platelet activation in blood during COVID-19 pathogenesis by integrating scRNA-seq with CITE-seq data (*12*).

Epigenomics studies modifications to DNA and histone proteins that affect gene expression without altering the underlying DNA sequence. Techniques like scATAC-seq profile chromatin accessibility, providing insights into the regulatory elements that control gene expression (*22*, *36*, *74*). This method uses the Tn5 transposase enzyme to randomly cleave and insert sequencing adapters into accessible regions of the genome, which are then amplified and sequenced (*22*, *36*, *74*). Single-cell reduced representation bisulfite sequencing

(scRRBS-seq) detects DNA methylation at single-base resolution, identifying epigenetic modifications that play crucial roles in development and disease. This technique uses bisulfite treatment to convert unmethylated cytosines to uracil, which are then sequenced to determine methylation patterns (*75*).

The combination of these technologies forms the basis of multi-omics, which integrates genomic, transcriptomic, proteomic, and epigenomic data to provide a comprehensive understanding of cellular function and regulation (*1–3*, *5–7*, *13*, *35*, *44–46*). Techniques like CITE-seq combine RNA sequencing with surface protein profiling, whilst methods such as single-cell Nucleosome, Methylation, and Transcription sequencing (scNMT-seq), and scATAC-seq integrate epigenomic and transcriptomic data to understand the interplay between chromatin state and gene expression (*22*, *35*, *36*, *74*, *76*).

In addition to experimental advancements, computational tools for multi-omic data integration have also progressed significantly. Statistical frameworks and analysis pipelines, such as those developed by Argelaguet *et al*. (MOFA+) and Ashuach *et al*. (MultiVI), facilitate the integration of multi-omic datasets, enabling comprehensive analysis of cellular states and transitions (*77*, *78*). Analytical frameworks like *Seurat* v3 and *Scanpy* have made multi-omic analysis more accessible, incorporating features for integrating and visualising single-cell data from multiple omics layers. These advancements underscore the transformative potential of single-cell multiomics in developmental biology and beyond, paving the way for new discoveries and therapeutic strategies.

Spatial sequencing technologies have advanced our understanding of cellular contexts within tissues. Techniques such as the 10X genomics Visium using Formalin Fixed Paraffin Embedded (FFPE) tissues and fresh frozen spatial transcriptomics (also on the Visium platform) enable the capture of spatial gene expression data by using spatially barcoded arrays to map RNA sequences back to their original tissue locations. However, as of Visium v1, gene expression capture areas are limited to 6.5mm x 6.5mm with a resolving power of 4992 spots, each spot being 55µm in diameter and separated by 100µm centre to centre (*79*). This coarse resolution allows for the study of population and tissue-level expression, but resolving finer cell type contributions requires inference from matched scRNA-seq data. For example, the cell2location method developed by Kleshchevnikov *et al.* uses Bayesian non-negative matrix factorisation to infer fine-grained cell type contributions within Visium spots (*80*).

RNA-scope employs in situ hybridisation techniques to detect RNA molecules within intact tissue sections, providing spatial context to matched (but not simultaneous) gene expression data. RNA-scope enables the detection of RNA molecules at intracellular resolutions, with the HiPlex V2 assay acquiring up to 12 targets in FFPE and 48 targets in fresh frozen sections. However, lacking direct cellular barcoding, quantitative single-cell analysis of RNA-scope data requires additional computational segmentation to acquire cellular RNA transcriptional profiles. Combined, Visium and RNA-scope offer excellent utilities to spatially map coarse-grain tissue expression programs and targeted high-resolution gene expression signatures, respectively. For instance, Calvanese *et al.* 2022, and subsequently, Goh *et al*. 2023 used fresh frozen Visium and RNA-scope to spatially map specific endothelial and haematopoietic cell states in the developing human yolk sac, revealing spatially restricted patterns of cellular availability (*1*, *2*, *46*).

The recent advancement in Slide-tags extends this concept, addressing both resolution and cellular quantification concerns by tagging frozen tissue sections with spatial barcode oligonucleotides (<10μm resolution) before extracting cellular nuclei (*81*). This allows for the spatial resolution of both scATAC-seq and Single Nuclei RNA-seq (snRNA-seq) modalities. Work by Russel *et al*. has shown promising results using this technology to simultaneously spatially resolve the epigenetic, transcriptomic, T-Cell Receptor (TCR) repertoire, and inferred Copy Number Variation (CNV) landscape from human metastatic melanoma samples (*81*). This technology represents a convergence of multiple modalities, addressing key spatial context weaknesses of droplet-based single-cell multiomics methodologies.

1.1.2 Advanced Insights from scRNA-seq and Multi-Modal Data Integration

scRNA-seq has unveiled properties of cellular systems that were previously hidden in bulk analyses. One such property is the identification of rare cell types and transient cell states that play critical roles in development and disease but are often masked in bulk RNA-seq data (*2*, *6*, *16–19*), By profiling individual cells, scRNA-seq can detect these rare populations and provide insights into their unique gene expression profiles and functions. For example, Suo et al. (2022) used scRNA-seq to characterise previously unidentified self-renewing, *CD5*+, *CS27*+, *SPN*+, *CCR10*+ putative B1 cells in human fetal life. This characterisation enabled the isolation of these B1 cells from fetal B cell subsets. This allowed for the characterisation

of their antibody secretion profile via ELISpot experiments, which demonstrated the increased IgM secretion capacity of *CCR10*+ B1 cells. Additionally, *CCR10* expression was hypothesised to play a role in tissue localisation to bone marrow (BM), gut epithelium, and skin keratinocytes via interactions with its ligand, *CCL28 (2)*.

Another property revealed by scRNA-seq is the dynamic nature of cellular differentiation and lineage commitment. By tracking the transcriptomic changes in individual cells over time, we can reconstruct developmental trajectories and identify key regulatory programs that govern cell fate decisions (*29*, *82–84*). This approach has been used to map the differentiation pathways of various cell types, including haematopoietic stem progenitors, providing a deeper understanding of how complex tissues and organs develop (*46*, *85*).

The integration of scRNA-seq with other modalities, such as the aforementioned proteomics and epigenomics technologies, further enhances our understanding of cellular function and regulation. Multi-modal data can reveal the discrepancies between RNA and protein levels, shedding light on post-transcriptional and translational regulatory mechanisms in regulating protein expression and function (*86–88*). Additionally, the combination of scRNA-seq with scATAC-seq allows for the identification of regulatory and promoter elements, and potential transcription factor binding sites that control gene expression. This integrated approach can uncover the gene regulatory networks that drive cellular differentiation in development (*9*, *22*, *36*, *89*, *90*). For instance, Wang *et al. 2022* utilised scRNA-seq alongside chromatin accessibility profiling to dissect the cellular heterogeneity and mechanistic responses to therapy within recurrent glioblastoma. Their integrative approach revealed distinct cellular states, identifying tumour-support niches including brain macrophages and angiogenic signalling in the tumour microenvironment (*91*). Using this information, this study identified potential cell-intrinsic and extrinsic therapeutic targets, underscoring the value of multi-modal single-cell analyses in understanding complex diseases .

Spatial sequencing technologies add another layer of complexity by providing spatial context to gene expression data. Mapping gene expression programs onto tissue architectures allows the study of how cells interact with their microenvironment and how spatial organisation influences cellular function (*80*, *81*, *85*). For instance, spatial transcriptomics has been used to study the organisation of the cutaneous squamous cell carcinoma (cSCC) tumour microenvironment, revealing how different cell states interact and contribute to tumour progression (*92*). Ji *et al.* leveraged this approach to discover key integrin signalling and

tumorigenic features of tumour-specific keratinocytes (TSKs) and (PD-L1/L2)-expressing migratory dendritic cells (migDCs), which contribute to cSCC T-cell exhaustion and immune suppression (*92*).

Currently, many multi-organ, spatially resolved atlases of human development and disease rely heavily on manual dissection, and subsequent digestion to acquire good spatially matched single cell or single nuclei data (*8*, *85*). For instance, work across the HDCA such as work by Popescu *et al.* in acquiring data across human development have largely focused on manually dissecting and digesting individual fetal organ tissues (*6*). In another example, Reynolds *et al.* dissected and separately digested the epidermis and dermis of human healthy, AD, and PS skin prior to performing scRNA-seq (*5*). These manual dissection methods aid computational disentanglement of tissue components and are an effective means to preserve coarse-grain architectural information. However, they are likely to suffer from batch-to-batch inconsistencies in dissection. Methods that preserve the organisational information of tissue architecture, whilst simultaneously capturing multi-modal information such as the aforementioned Slide-tags technology, likely represent a paradigm shift in how we study developing tissues going forward.

In conclusion, integrating data from multiple omics layers provides a comprehensive view of cellular function and regulation that is not possible with a single modality, uncovering the complex regulatory networks that drive cellular behaviour and potentially identify new targets for therapeutic interventions. These advancements underscore the transformative potential of single-cell multi-omics in developmental biology and beyond, paving the way for new discoveries and therapeutic strategies.

1.1.3 Developmental origins of human haematopoiesis

The development of the human blood and immune system begins at various anatomical sites during gestation (*1*, *3*, *46*, *85*). Initially, blood and immune cells differentiate from extra-embryonic YS Haematopoietic Stem Progenitor Cells (HSPCs) and AGM-derived haematopoietic stem cells (HSCs) (*46*). To understand this complex process, it is essential to consider the stages of human prenatal development, which are categorised into three main stages: pre-embryonic, embryonic, and fetal development (*93*, *94*).

Human prenatal development is often described using three complementary systems: Carnegie stages, Post Conception Weeks (PCW), and Post Conception Days (PCD) (*95–97*). These systems offer detailed timelines and developmental milestones, facilitating precise communication in embryology and related fields (Fig 1.2) (*95–97*). Carnegie Stages: The Carnegie staging system is a standardised method of categorising the early stages of embryonic development based on morphological characteristics. This system includes 23 stages, covering the first 8 weeks post-conception. Each stage represents specific developmental milestones, such as the formation of the neural tube or the emergence of limb buds (Fig. 1.1 b) (Fig 1.2) (*95–97*). This detailed morphological classification is invaluable for accurately tracking and comparing embryonic development across different studies. PCW is a timeline that measures the age of the embryo or fetus from the time of conception. This system is widely used in research and clinical settings because it provides a clear chronological framework (*95–97*). For example, the transition from the embryonic to the fetal period occurs at around 8 weeks PCW (*94–97*). Using PCW allows researchers to describe developmental milestones in a temporal context, making it easier to correlate these with clinical observations and interventions (*94*, *97*). PCD is a more precise measurement that counts the exact number of days since conception. This level of detail is particularly useful in early embryonic development when significant changes can occur within short time spans (*94–97*). Reference to the embryonic developmental staging table (Fig. 1.2) provides a comprehensive breakdown of the Carnegie stages, PCD, and PCW, allowing for easy reference and comparison between these different systems.

As an overview of human embryological development, human prenatal development begins with fertilisation, typically occurring in the ampulla of the Fallopian tube. The fertilised egg, or conceptus, undergoes several mitotic divisions without an increase in volume, a process known as cleavage. By the third day or Carnegie stages 1-2 (CS), the conceptus, now a 16-cell morula, reaches the uterus (*93*, *94*, *98*, *99*). The totipotent cells in the morula continue to divide while freely floating in the uterus for several more days. During this time, it utilises nutrients stored in the egg cytoplasm and uterine milk secreted by the endometrium. When the cell count reaches around 100, the cells begin to arrange around a fluid-filled cavity, forming a blastocyst around the fifth day (CS3) (Fig. 1.1 A and B) (Fig 1.2)(*93*, *94*, *98*, *99*). The blastocyst, with its inner cell mass (ICM) and outer trophoblast, begins the process of implantation into the uterine wall by days 6-7 post-conception (CS4). By Days 8-9, the blastocyst fully implants into the endometrium (*93*, *94*, *98*, *99*). During this time, the

trophoblast aids in embedding and secreting human chorionic gonadotropin (hCG) to maintain progesterone production, signalling the corpus luteum to continue producing progesterone, the hormone that sustains pregnancy. Progesterone stimulates the growth of nutrient-rich decidual cells that nourish the early embryo. The trophoblast eventually develops into the chorion, the fetal portion of the placenta (Fig. 1.1A) (*93*, *94*, *98*, *99*).

By the end of the first Post Conception Week (PCW) (CS5) of development, the blastocyst forms a bilaminar embryonic disc consisting of the hypoblast and epiblast, which then gives rise to the primary YS and amnion (Fig. 1.1 A and B) (Fig 1.2) (*93*, *94*, *98*, *99*). The YS, initially providing nutrients and facilitating gas exchange, is the first site of haematopoiesis in humans, generating and supporting the initial embryonic primitive waves of blood and immune cells through endothelial-to-haematopoietic transition (EHT) of specialised hemogenic endothelium (HE) and support niches (Fig. 1.1 A, B, and C, and Fig. 1.2) (Fig 1.2).

Gastrulation during 3PCW transforms the bilaminar disc into three germ layers: ectoderm, mesoderm, and endoderm, critical for establishing the foundation for complex structures and organ systems. The formation of the mesoderm during this process is initiated by the ingress of epiblast cells through the primitive streak, forming an adjacent layer to the hypoblast. The gastrulation process also sees the collapse of the primary YS into small vesicular remnants, subsequently reabsorbed by the embryo. The primary YS is superseded by the secondary YS which we discuss in more detail in our overview of the human YS section below (Fig. 1.1 A and B) (Fig 1.2) (*93*, *94*, *98*, *99*).

By 4PCW (CS10-13), neurulation begins with the development of the neural tube from the ectoderm (CS10-12), which will give rise to the central nervous system. Somite formation occurs (CS11-12) , segmenting the mesoderm into somites, precursors to the vertebral column and skeletal muscles. The primitive heart tube forms and begins beating at 4PCW (CS10). Concurrently, the AGM region forms by 4-5PCW (CS14) (Fig. 1.1 A, B, and C) (Fig 1.2) (*93*, *94*, *98*, *99*). The AGM region emerges from the dorsal aorta as the site of definitive human haematopoiesis. Definitive HSCs arise here through an EHT (*46*). These HSCs migrate to the embryonic liver (EL) (5-6PCW) and eventually colonise the bone marrow (11PCW) (Fig. 1.1 C, Fig 1.2) (Fig 1.2) (*3*). The yolk sac typically regresses around 8PCW (CS23), and is no longer visible by the end of the embryonic period (Fig. 1.1 B and C) (Fig 1.2) (*94*). This marks the start of the fetal developmental period where the human liver

becomes the primary site of fetal hematopoiesis from 9PCW until the mid-second trimester when BM is then established as a major site for lifelong haematopoiesis after 20PCW (Fig 1.3) (*6*).

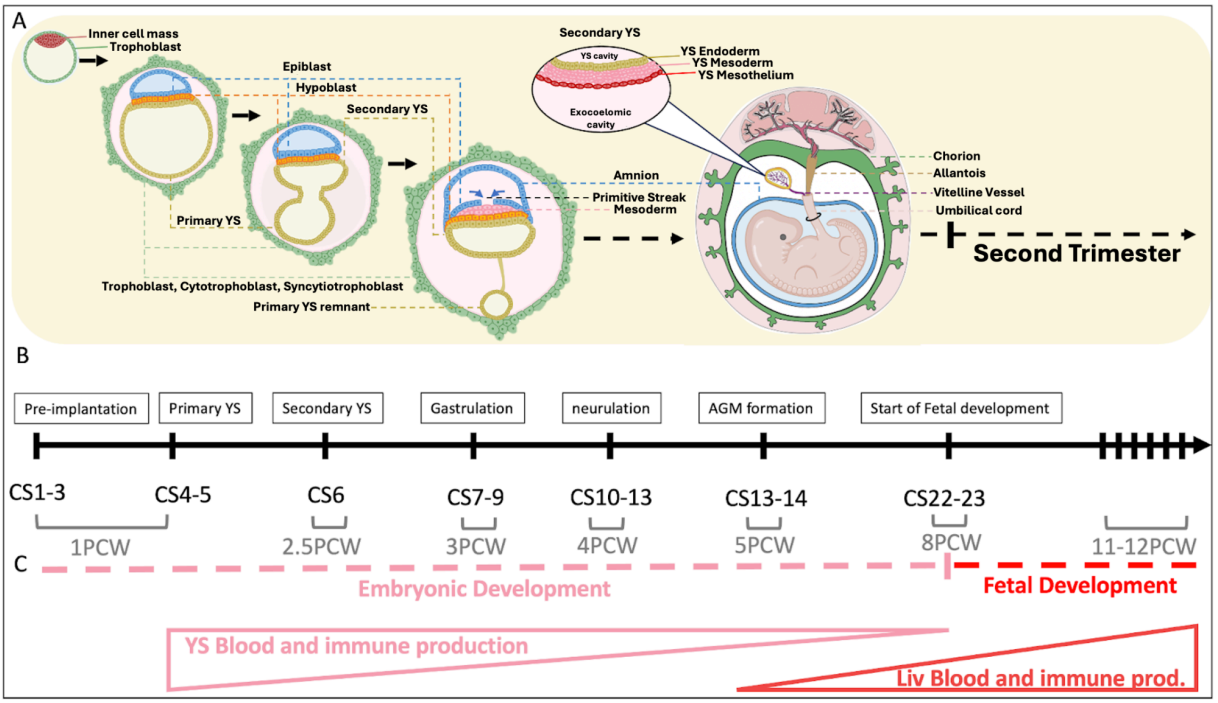


Figure 1.1: Illustration of a timeline highlighting important events of human developmental embryogenesis and the origins of the human yolk sac. *This figure provides a depiction of the human embryonic development timeline and the embryonic origins of the yolk sac, highlighting critical stages and processes. (A) The yolk sac is the first site of primitive hematopoiesis. We illustrate the cascade of embryological processes leading to the formation of the yolk sac, starting from the blastocyst stage. This includes the formation of the bilaminar disk and primary yolk sac, the development of the secondary yolk sac from the hypoblast, the degradation of the primary yolk sac, and the formation of the trilaminar disk initiated by the ingress of epiblast cells through the primitive streak to form the mesoderm during gastrulation. We further depict a zoomed-in view of the endoderm, mesoderm, and mesothelial layers of the secondary yolk sac and its interface to the embryo (~6PCW) via the vitelline duct. (B) We show the developmental timeline in CS and PCW illustrating important developmental events and how they relate to the illustration in A. Importantly, we highlight the formation of the yolk sac by CS4-5, its involution by CS22-23, and the formation of the AGM at 5PCW which signals the start of definitive hematopoiesis in humans. (C) Finally, the timeline categorises embryonic (1 - 8PCW)(pink) and fetal development (8PCW+)(red) and illustrates the relative hematopoietic contribution of the yolk sac (pink) prior to AGM formation, and the embryonic liver (red) post AGM formation. Figure produced using BioRender.com and adapted from Goh and Botting et al, 2023, a research article that I co-first-authored (1, 94).*

Carnegie Stages (CS)	Post-Conception Days	Post-Conception Weeks (PCW)
CS1	1	0
CS2	2-3	0-1
CS3	4-5	0-1
CS4	5-6	0-1
CS5	7-12	1-2
CS6	13-15	2
CS7	16-19	2-3
CS8	19-21	3
CS9	22-23	3
CS10	23-25	3-4
CS11	24-26	4
CS12	26-30	4-5
CS13	28-32	4-5
CS14	32-33	5
CS15	33-36	5-6
CS16	37-40	5-6
CS17	41-43	6
CS18	44-48	6-7
CS19	48-51	6-7
CS20	51-53	7-8
CS21	52-54	7-8
CS22	54-55	8
CS23	56-60	8

Figure 1.2: Embryonic developmental staging table. *This table provides a breakdown of equivalent conversions between CS, PCD, and PCW time points (94–97) .*

1.1.4 The Human Yolk Sac : The site of earliest human hematopoiesis

The YS is an evolutionarily conserved, extraembryonic structure found in fish, birds, and mammals. In phylogenetic terms, the YS is first seen in vertebrates with yolk-rich eggs e.g., birds, reptiles and amphibians, where its role is to extract macronutrients from yolk to sustain the embryo (*100*). The capacity to uptake, transport and metabolise nutrients is retained in both mouse and human YS (*101*). In humans, the primary YS derives from the hypoblast, which produces extraembryonic endoderm that migrates to form Heuser's membrane and the primary YS cavity at around CS4-5 (Fig 1.1 A and B) (Fig 1.2). This structure absorbs nutrients and oxygen through primordial uteroplacental circulation, essential for early embryonic growth. Around 2.5PCW (CS6), a "second wave" of cells migrates from the hypoblast to form the secondary YS, which supersedes the primary YS and persists until 8PCW (CS22-23) (Fig 1.1 A and B, and Fig 1.3) (Fig 1.2) (*93*, *101*). The secondary YS, a more compact structure compared to the primary YS, consists of three layers: an outer mesothelium layer facing the exocoelomic cavity, a middle mesodermal layer, and an inner endodermal epithelium facing the YS cavity (Fig. 1.1 A). The secondary YS surrounds a vitelline fluid-filled cavity and retains the capacity to uptake, transport, and metabolise nutrients.

Haematopoiesis originates in the YS in mammals, birds and some ray-finned fishes (*102*). In the mouse, the YS is pivotal in the first wave of haematopoiesis, producing primitive erythroblasts, megakaryocytes, mast cells, and myeloid cells from embryonic day 7.5 (E7.5) (*102*, *103*). Following the onset of circulation, a second wave of erythro-myeloid and lymphoid-myeloid progenitors arise in the YS and supply the embryo (*104*). In humans, definitive haematopoietic stem cells arise in the AGM region of the dorsal aorta and seed the fetal liver from 5PCW, and later colonise the BM from 11PCW (Fig 1.1 A and B, and Fig 1.3). Recent work by Calvanese *et al. 2022*, Suo *et al. 2022* and Goh and Botting *et al 2023.* have collectively built upon large single-cell multi-omic and spatial atlases, further delineating the role the human YS plays during early blood and immune haematopoiesis (*1*, *46*, *85*). The evidence posited suggests that YS not only provides the first blood cells during development in humans, but is also the site for the earliest emergence of HSPCs. In humans, blood islands, containing primitive erythroblasts that express embryonic globin genes (*HBZ* and *HBE1*), and surrounded by endothelium, termed haemato-endothelium, emerge around

2.5PCW (CS6) (7, 8). The human yolk sac (YS) produces megakaryocytes and mast cells, contributing to the initial differentiation of immune cells, including myeloid cells. Recent work has highlighted its role in seeding developing organs, such as the microglia population in the brain (*6*).

Transplantation studies of human developmental tissues into immunodeficient mice have pinpointed the origin of definitive HSCs within the AGM region of the embryo at CS14 (~5PCW), with equivalent cells subsequently found in the YS at CS16 and the liver from CS17. This sequence was also documented by following the transcriptional signature of definitive HSPCs across organs (*46*). While the process of definitive HSPC emergence from HE has been reconstructed in the human AGM region (*46*, *105*), the emergence of earlier progenitors in the human YS remains under-studied due to the rarity of sample collection and its short developmental window (Fig. 1.1 A and B). Notably, differences exist between erythroid and megakaryocyte (MK) cells derived from YS progenitors and those from later haematopoietic organs (*46*, *105*). The full repertoire of human YS-derived blood and immune cells, their differences from later counterparts, and their contributions to long-lived cells such as tissue-resident macrophages (TRMs) require further investigation.

Despite the importance of the yolk sac (YS) during early human development, research often relies on studying yolk sacs from model organisms like mice, which, although different in some aspects, provide valuable insights. For instance, the timelines of haematopoietic output from various organs differ significantly between mice and humans. In mice, haematopoiesis is dominated by the fetal liver (FL), with the YS primarily producing progenitors and the BM contributing only after birth. In contrast, human fetal bone marrow (FBM) can produce outputs similar to adult BM by early in the second trimester (Fig 1.3)(*3*). Murine models have shown multiple waves of YS-derived erythroblasts. Primitive-wave erythroblasts initially express *Hbb BH1* and *Hba X,* then undergo primitive maturation, switching to *Hbb εy* and *Hba a1/2* haemoglobin expression profiles (*106*, *107*). Immature erythroid progenitors exit the YS, and rapidly mature in other sites as early as E10.5, prior to murine AGM formation, seeding the EL and giving rise to a pro-definitive hematopoietic wave (*Hbb BT1* and *Hbb BS*) mirrored in both liver and YS (*107*). These observations suggest that murine YS-derived haematopoiesis is not spatio-temporally restricted to YS, but occurs in distinct waves, firstly in the YS, and subsequently, in both YS and liver tissues.

In humans, the extent to which YS-derived progenitors contribute to EL hematopoiesis remains unclear. It is also uncertain whether the YS has a more extended phase of contribution in humans compared to mice (*93*). Recent studies suggest more expansive species references, such as rabbits with their greater early gestational similarity to humans, may be more appropriate for understanding the role of the YS in human development (*108*).

Several key questions about human YS haematopoiesis remain unanswered: what is the full repertoire of human YS-derived blood cells, does the YS produce limited progenitors or HSPCs, do YS progenitors/HSPCs contribute to long-lived populations such as TRMs, do YS progenitors/HSPCs arise from HE, and what are the regulatory features of this process.

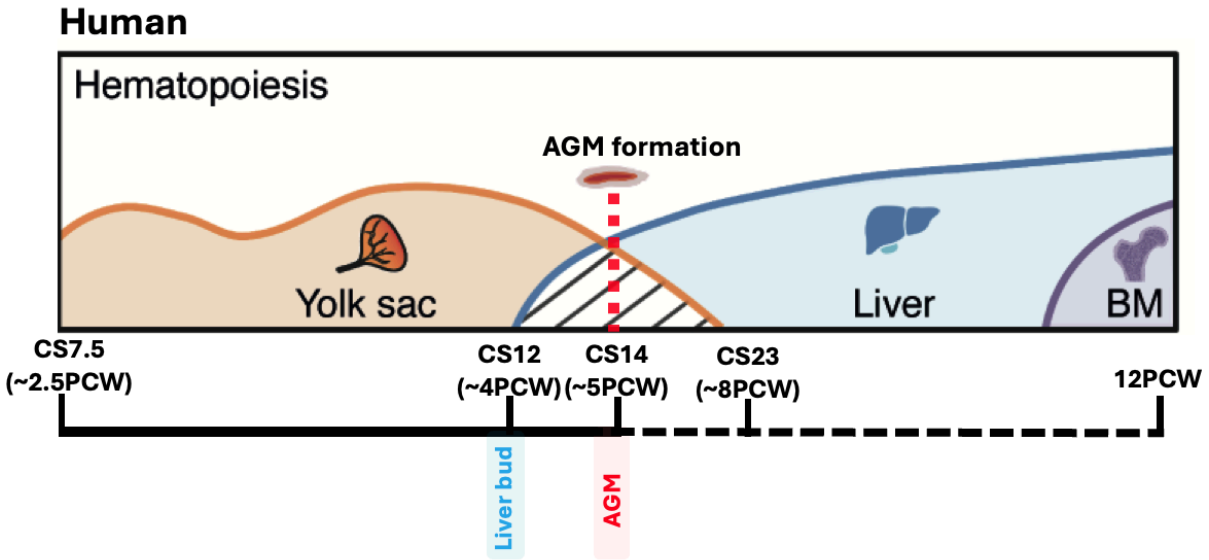


Figure 1.3: Timeline of embryonic and fetal hematopoiesis: *This illustration is a breakdown of human hematopoietic timeline starting in the YS during embryonic development, highlighting the formation of the liver bud from 4PCW, the formation of the AGM from 5PCW, and the gradual cascade of hematopoietic functions between the YS, liver (after 5-8PCW), and BM (>12PCW). The solid line represents the YS-derived primitive hematopoietic timeline prior to AGM formation. The dotted line represents the definitive hematopoietic timeline post AGM formation. Figure adapted from Goh and Botting et al, 2023. (1, 94)*

1.2 Development and functional diversification of the human Mononuclear Phagocyte system and its roles in organising tissue architecture

The human mononuclear phagocyte (MP) system, which includes dendritic cells, monocytes, and macrophages, is a critical regulator of innate and adaptive immune responses. During embryonic development, MPs derive sequentially from YS progenitors, and subsequently from AGM-derived HSCs in fetal liver and bone marrow. As part of the myeloid lineage, MPs maintain tissue homeostasis and confer protective innate immunity in postnatal life. Recent evidence - primarily in animal models - highlight their critical role in coordinating the remodelling, maturation, and repair of target organs during embryonic and fetal development. However, the molecular regulation governing chemotaxis, homeostasis, and functional diversification of resident MP cells in their respective organ systems during development remains elusive.

The below sections summarise current knowledge about the development and functional roles of tissue MPs during human organ development and morphogenesis. We highlight how single-cell multi-omic approaches, bioinformatic advancements, and next-generation ex-vivo culture systems, such as organ-on-chip and organoid models, offer new platforms to study the role of human MPs in development and disease.

1.2.1 Introduction: The Human MP system

The mononuclear phagocyte (MP) system comprises macrophages, monocytes, and their precursors, categorised by their morphology, function, and origin. Originally, macrophages were considered to be differentiated monocytes (*109*). The discovery of dendritic cells (DCs) by Steinman et al. in 1979, characterised by their probing morphology and ability to activate naïve T-cells, expanded the MP system to include these cells (*110*).

In 1882, Elie Metchnikoff proposed that macrophages are crucial for maintaining tissue integrity and homeostasis, requiring them to distinguish between self and non-self, recognise tissue damage, and detect invading pathogens (*111*). Since then, extensive research has refined our understanding of MP functions. While the roles of MPs in post-natal life have been extensively studied, their functional heterogeneity during human fetal and embryonic development remains less understood (*112–114*). Recent studies have identified MPs in the human yolk sac (YS) as early as Carnegie Stage (CS) 7 (3 PCW), within a pathogen-free in

utero environment (*6*, *7*, *47*). Given the early establishment of this MP niche within the YS, well before pathogen exposure or immune challenges, this strongly suggests that beyond classical immune functions, MPs likely perform critical developmental roles, such as coordinating apoptotic cell clearance and extracellular matrix remodeling, to support angiogenesis, organogenesis and tissue morphogenesis, analogous to their known roles in adult wound healing and tissue repair (*112–114*).

1.2.2 Development of the human MP system

Human embryonic hematopoiesis occurs in several transient waves. The generation and differentiation of human primitive HSPCs begin within the YS, leading to the emergence of the first myeloid lineage cells around CS7 (2-3PCW). This includes tissue-resident macrophage populations, such as microglia in the brain, originating from the YS (*112*, *115–117*). In 2022, Dick *et al.* conducted pivotal research on the origins and dynamics of tissue-resident macrophage populations across various murine organs, including the heart, liver, lung, kidney, and brain (104). Using a tamoxifen-inducible Cx3cr1CreER, tdTomato reporter fate mapping mouse model, they identified a conserved population of TLF+ (TIMD4, LYVE1, and/or FOLR2) expressing macrophages in both murine and human single-cell expression profiles. These macrophages, derived from YS progenitors, share a core gene program conserved post-seeding in tissues (*118*). This finding may help trace the developmental origins of macrophages seeded into tissues throughout human fetal life.

Definitive hematopoiesis follows primitive hematopoiesis, beginning with the EHT of HE, giving rise to definitive aorta-gonad-mesonephros (AGM)-derived HSCs from CS14-15 (3-4PCW) (*112*, *114*, *116*). These HSCs enter the circulation and seed the embryonic liver (EL), which produces the first granulocyte-monocyte progenitors (GMPs) and blood monocytes from CS15 (5-6PCW) (*113*, *114*). The fetal liver (FL), which receives seeding from YS MP progenitors and subsequently AGM-derived HSCs, serves as the main hematopoietic organ, generating monocytes and TRMs during embryonic and early fetal development (*6*, *112*, *119*). Additionally, fetal bone marrow (FBM)-derived monocytes contribute to TRM populations from 11-12PCW (*120*).

TRM populations are derived from progenitors that seed tissues during embryonic development, originating from the YS, AGM, liver, or BM. These TRMs also depend on IL-34 signalling for self-renewal. Additionally, TRMs can be derived from circulating

monocytes of liver and BM origin during later stages of development and in inflammation (*121*).

Monocytes represent a highly plastic population that circulates through the blood to surrounding tissues, where they differentiate into macrophages or monocyte-derived dendritic cells (mo-DCs). In adult humans, monocytes mature in the bone marrow (BM) and are released into circulation as CD14++, CD16- classical monocytes. These classical monocytes give rise to CD14+, CD16+ intermediate monocytes and eventually to $CD14^{dim}$, CD16++ non-classical monocytes (*122*).

To delineate the heterogeneity of prenatal monocytes across tissues, Suo *et al. (2022)* integrated single-cell RNA sequencing (scRNA-seq) data from nine developing human organs, including hematopoietic organs (yolk sac, liver, and bone marrow), lymphoid organs (thymus, spleen, and lymph nodes), and non-lymphoid peripheral organs (skin, kidney, and gut) (*85*). They characterised three distinct subsets of monocytes differentially distributed across the BM and other peripheral tissues, based on the expression of *CXCR4*, *CCR2*, or *IL1B*. Their findings indicated that adult classical monocytes most closely resemble the *IL1B*-hi and *CCR2*-hi prenatal monocyte subsets, while no *CXCR4*-hi subsets were observed in non-lymphoid adult tissues. Further, Suo et al. observed that *CXCR4*-hi monocyte subsets were enriched in prenatal BM, whereas *IL1B*-hi monocytes were more prevalent in peripheral organs. *CCR2*-hi subpopulations in the BM expressed proliferation markers, while *CCR2*-hi monocytes in peripheral tissues upregulated *IL1B* and other *TNF-α* signalling genes, indicating a pro-inflammatory profile and potential involvement in peripheral tissue remodelling, angiogenesis, and lymphoid tissue organogenesis (*123–127*). This evidence suggests that the transition of monocytes from the bone marrow to peripheral tissues involves a shift from a *CXCR4*-hi to a *CCR2*-hi state, with *CCR2*-hi monocytes subsequently maturing into *IL1B*-hi monocytes (*85*).

In mouse models, similar processes have been observed. Murine Ly6C- non-classical monocytes differentiate from circulating Ly6C+ monocytes and patrol the vascular system (*112*). In the murine bone marrow, monocyte retention is regulated via *CXCR4* and *CXCL12* interactions between monocytes and stromal cells respectively. Monocytes are retained until CCR2-CCL2 interactions predominate, enabling egress (*128*). This is consistent with the observation of *CXCL12* expression in human prenatal BM fibroblasts and osteoblasts (*2*). In contrast, the developing human liver presents comparatively lower proportions of $CXCR4^{hi}$

monocytes (*2*), suggesting that alternative mechanisms of monocyte retention and release operate in the human developing liver. This is in keeping with reports from murine models (*129*).

Comparison of human fetal and post-natal monocytes has shown that both fetal and adult populations conserve high expression of myeloid and monocyte surface markers CD11b, CD11c, CCR2 and CX3CR1 (*130*).

Upon interferon γ (IFN-γ) stimulation, post-natal monocytes upregulate antigen presentation genes. Conversely, fetal monocytes upregulate genes involved in innate antimicrobial responses (*130*). This differential response highlights the distinct functional roles and developmental adaptations of monocytes at these stages. In fetal monocytes, the upregulation of innate antimicrobial response genes suggests a reliance on the innate immune system, which is crucial given the immature state of the adaptive immune system during developmental periods (*2*, *130*). The acquisition of immune effector roles in fetal life prioritises immediate and broad-spectrum protection to ensure rapid responses to potential pathogens (*130*). While adaptive responses are crucial for clearance of pathogens in the postnatal period, potentially, this shift in fetal monocytes avoids triggering harmful adaptive responses such as anti-self or anti-maternal rejection, which could result in preterm termination (*130*). In contrast, post-natal monocytes' upregulation of antigen presentation genes reflects the maturation of the adaptive immune system, essential for developing specific and long-lasting immune responses (*130*). This shift indicates a transition from innate to adaptive immunity (*2*). Consequently, the functional emphasis of fetal monocytes on innate responses and post-natal monocytes on antigen presentation illustrates the evolving priorities of the immune system from immediate defence mechanisms in utero to a more adaptive response post-birth (*2*, *130*).

Macrophages have been observed both morphologically and transcriptionally from the earliest wave of mouse YS haematopoiesis and as early as CS19 (6PCW) in human YS and decidua (*6*, *131–133*). Hofbauer cells, observed from CS7 (2.5PCW), are used to describe any fetal derived placental macrophage that resides within the placental villous core, amnion, and chorion (*131*, *134*, *135*). Hofbauer cells isolated from the human placenta express CD14, CD163 and secrete anti-inflammatory TGF**β** and IL-10, suggestive of immune-suppressive and pro-vasculo/angiogenic functions (*134*, *136*).

TRMs are adapted to their resident tissues and have been named based on anatomical location, protein, and transcriptional signatures (*137*). The surface markers of macrophages include those shared with monocytes (CD14, CD16, CD68 and CCR5), general macrophage program markers (C1QC, VEGF) and tissue specific markers such as VCAM1 for fetal liver Kupffer cells, and PPARγ for both alveolar and Kupffer macrophages (*114*, *138*, *139*). TRMs can be long-lived and self-renewing following prenatal seeding but can also be replaced by circulating monocytes during inflammation, underscoring their critical role in maintaining tissue-specific immune functions and highlighting the immune system's flexibility and resilience (*140–142*).

Major human dendritic cell (DC) subsets include *AXL*+, *SIGLEC6*+ DCs (ASDC), plasmacytoid DCs (pDC), conventional DC1 (cDC1), DC2 (cDC2), DC3 (cDC3), monocyte-derived DCs (moDC) and Langerhans cells (LCs) (*143*, *144*). Post-natally, DCs function to sense pathogens and activate the adaptive immune system (*145*). DCs can arise from both myeloid and lymphoid progenitors (*146*), with studies showing monocyte participation to the DC pool following inflammation (*147–149*).

Although human DC origin has been attributed to BM-derived HSCs, fetal DCs have been observed as early as 5PCW, suggesting that FL HSCs may also generate DCs (*6*). Compared to adult DCs, fetal DCs possess an immature phenotype but can induce allogeneic T-cell proliferation upon culture (*150*). Their presence during early development is attributed to ensuring tolerogenic responses to self and maternal antigens (*150*).

1.2.3 Functional and organisational diversification of MPs in prenatal life

The distribution of MPs in prenatal organs is carefully coordinated with key timelines of specific organ development, allowing the continuous survival and development during embryonic and fetal life (*130*, *151*). Discoveries including MP origin from progenitors to their post-natal immunological and repair functions have overshadowed the important roles they play during development and in organogenesis (*113*, *114*).

The functional absence of the MP system is embryonically lethal, which illustrates the importance of MPs for survival and development during development, demonstrated by the growth abnormalities, including reduced brain size and enlarged ventricles observed in the

Csf1r double knockout mouse model (*152*, *153*). Recent studies have shown that macrophages are present from as early as CS7 (3PCW) of human life, whilst monocytes and DCs have been observed from CS14 (5PCW) of human life in a relatively pathogen-free in utero environment (*6*, *47*). This raises the hypothesis that MPs may play an important role in tissue modelling and homeostasis in addition to immunity during early prenatal life. During human development, haematopoiesis occurs in both canonical and non-canonical haematopoietic organs. Canonical haematopoietic organs are those traditionally recognised as primary sites of blood cell formation, including YS, liver, BM, and spleen. Non-canonical haematopoietic organs refer to those not traditionally associated with blood cell formation but shown to support haematopoiesis under certain conditions or during specific developmental stages. Examples include thymus, lung, skin, and gut. Here, we synthesise the literature on MP functional diversification in canonical and noncanonical hematopoietic tissues, as well as early fetal tissues that receive hematopoietic seeding and develop TRMs during human development. This section highlights the orchestrated, system-wide nature of blood and immune cell production during human prenatal development.

1.2.4 Yolk Sac : A haematopoietic organ

As previously discussed, the yolk sac (YS) serves as an early site of hematopoiesis. However, the full lineage of YS-derived MPs in humans has not been fully delineated. In mice, hematopoiesis is initiated in the YS around embryonic day 7.5 (E7.5), with detectable erythroid and macrophage progenitors from this early stage (*111*, *112*, *154*). By E8.25, the YS is actively producing multi-lineage erythro-myeloid progenitors (EMPs) in what is known as the second wave of hematopoiesis or the transcend definitive stage (*154*, *155*). These YS macrophages are functionally mature, capable of phagocytosis, and critical for embryonic tissue patterning and apoptosis clearance (*113*). Notably, lineage-tracing studies have shown that these cells are self-renewing and contribute to the adult TRM pools, including microglia in the brain and Kupffer cells in the liver where they are involved in both neural development and maintenance (*155–157*).

In humans, YS macrophages become detectable by CS11 (~4PCW) and YS-derived myeloid progenitors (YSMPs) seed the EL by CS12 (4-5PCW) (*46*). Studies have established that murine YS-macrophages are derived directly from the YS without transitioning through a monocyte intermediate, which distinguishes their lineage from macrophages derived from later hematopoietic sites such as the bone marrow and fetal liver. This direct lineage pathway

has been corroborated by lineage tracing and fate mapping studies, highlighting a unique developmental trajectory for these YS-macrophages compared to bone marrow-derived macrophages (*113*, *151*). However, it remains unclear whether a similar developmental trajectory occurs in humans.

Support for the origin and function of YS-macrophages is bolstered by experimental evidence, including gene expression profiling and in vivo lineage tracing. Notably, we have previously explored the characterisation of a TLF+ (*TIMD4, LYVE1,* and/or *FOLR2*) macrophage population in both murine and human fetal tissues. In the murine model, lineage tracing experiments suggested that these TLF+ macrophages are derived directly from the YS, further substantiating the developmental origins of certain TRM populations (*118*).

Despite this knowledge, the YS origins of different TRM populations, and whether human YS macrophages undergo a monocyte intermediate differentiation route remain poorly understood.

1.2.5 Liver as a haematopoietic organ:

The liver is canonically known as a site of hematopoiesis during fetal life. Kupffer cells, identified by CD163, VCAM1 and CLEC5A in humans (*158*) are the TRMs of the liver, lining the sinusoids. Kupffer cells assist in the proliferation and enucleation of erythroblasts as well as in iron recycling to facilitate erythropoiesis in the FL (*107*, *159*). Definitive erythropoiesis occurs in the human FL and requires HIF2A and EPO expression for progenitor survival. This process is MYB dependent and relies on transcriptional regulators such as SOX6 and BCL11A that down-regulate embryonic (Gower 1 - ζ2ε2, Gower 2 - α2ε2) and fetal globin expression (α2γ2) (*107*, *159–162*). At the stage of active FL haematopoiesis, macrophages can migrate from the sinusoids to the parenchyma to form erythroblastic islands consisting of a central macrophage surrounded by erythroblasts (*163*). The central macrophages express VCAM1, CD163 and EPOR to mediate interactions with early erythroid cells and EPO, stimulating erythroblast enucleation, proliferation, and differentiation (*6*, *163*).

A non-canonical function of human Kupffer cells is to prevent pathogenic accumulation of lipids in the liver. Peroxisome proliferator activated receptor gamma (PPARγ) was identified as an important regulator of macrophage activator programs linked to the fatty acid oxidation function of Kupffer cells (*142*, *164*). PPARγ also regulates pro-proliferative interleukin (IL)-4 driven programs (basophil recruitment) during damage to assist during liver

generation/regeneration (*165*). These studies demonstrate a common non-canonical function between human fetal and post-natal Kupffer cells in coordinating erythropoiesis and restoring damaged tissue during fetal development and post-natal liver regeneration.

Like adult DCs, human FL DCs can migrate to lymph nodes and initiate T-cell proliferation in response to toll-like receptor (TLR) ligation. Additionally, it was observed after TLR ligation, FL cDC2s show markedly reduced TNFα cytokine production when compared to adult DCs to mediate immune suppressive responses during gestation (*165*).

1.2.6 Bone Marrow as a Myelopoietic organ

The BM is the primary hematopoietic organ in adult life. It facilitates the generation of various cell types such as circulating monocytes, granulocytes, lymphocytes, megakaryocytes, osteoclasts, and DCs (*3*). Hematopoiesis begins early in embryonic development, initially occurring in the YS and subsequently transitioning to the FL before establishing in the BM (*2*, *3*). In humans, hematopoietic cells start to populate the BM around 11-12PCW, a critical shift that establishes the BM as the central site for lifelong blood and immune cell production (*3*).

Monocytes originate in the BM and play a crucial role in the innate immune system. These phagocytic immune cells circulate in the blood and migrate into tissues, where they may differentiate into activated monocytes or macrophages (172). The development and mobilisation of monocytes from the BM are tightly regulated processes. In humans, monocytes exit the BM in a CCR2-dependent manner, a chemokine receptor critical for their egress into the bloodstream. Initially, these monocytes are characterised by a high expression of CXCR4, a receptor involved in their retention within the BM. As they mature and prepare for exit, they undergo a transition to a CCR2-high state, facilitating their release into circulation (*2*, *128*).

Human monocytes are broadly classified into three subsets: classical, non-classical, and intermediate, each with distinct functions and roles in immune responses (*122*). Classical monocytes are known to enter tissues and differentiate into activated inflammatory monocytes or macrophages, especially following infection (*121*). These cells play a key role in the body's defence mechanisms against pathogens by engaging in phagocytosis and producing pro-inflammatory cytokines. Non-classical monocytes, also referred to as patrolling monocytes, exhibit a unique behaviour by patrolling the vascular endothelium (*122*, *166*). They are involved in surveying the endothelium for signs of injury or infection and play a role

in the resolution of inflammation and tissue repair (*167*). Intermediate monocytes are less well understood, and their exact function is still a subject of ongoing research. Human intermediate monocytes have been found to roll along the endothelium rather than circulate freely in the blood. Unlike other monocyte subsets, they do not differentiate into tissue-resident macrophages but are thought to have a putative role in surveying for vascular injury and maintaining vascular integrity (*166*).

Osteoclasts are a TRM of the BM, and are identified by protein-based surface markers: CTSK, CALCR, SIGLEC15, ACP5, DCSTAMP, OCSTAMP and TNFRSF11A (*168*). Osteoclasts differentiate from common myeloid progenitors via cytokine-dependent signalling involving M-CSF (macrophage-CSF/CSF-1), and receptor activated NF-κB ligand (RANKL) (*169–171*). They function to assist with the clearance/resorption of bone tissue and are critical in the maintenance, repair, and remodelling of the skeleton (*169*). In studies performed in mice, CSF-1 has been shown to be essential for osteoclast formation, as its absence leads to impaired cell fusion, which is necessary for forming the large, multinucleated cells required for effective bone remodelling (*152*, *172*). In addition to osteoclasts, the BM hosts various resident BM macrophages, which play distinct roles in hematopoiesis and bone homeostasis (*2*, *3*).

Among these macrophage populations, Erythroblastic Island macrophages (EI macs) play a crucial role in hematopoietic regulation. These macrophages determine hematopoietic egress through the phagocytosis of cells not expressing the ‘don’t eat me’ signalling CD47 ligand (*173*, *174*). EI macs are in contact with BM erythroblasts to support erythropoiesis. They contribute to heme synthesis and iron recycling (*175*), erythropoietin (EPO) feedback sensing (*176–178*), and express cytokines including insulin-like growth factor (IGF) and bone morphogenetic protein (*178*, *179*), which promote erythropoiesis. Decreased MPs in the BM affect the HSC niche and induce HSC mobilisation into the blood due to niche collapse, resulting in a decline in erythro-/hematopoiesis due to bone endosteal niche disruptions (*180*).

A study on adult human BM DCs showed reduced activity of canonical DC-functionalities when compared to matched DCs in the peripheral blood (*181*). For example, cDC2s were less able to upregulate T-cell stimulatory molecules such as CD80 upon TLR-triggering when compared to peripheral blood cDC2s (*181*). The BM niche was concluded to be primarily a DC developmental location. Murine BM studies have shown the contribution of DCs in the regulation of haematopoiesis: ablation of murine BM cDCs resulted in HSC mobilisation into

peripheral blood to transiently lodge into other haematopoietic organs such as the spleen. BM cDC ablation in the mice also led to a loss of BM macrophages, increased BM vascular permeability and the expansion of BM endothelial cells, which are required for haematopoietic regulation (*182*). These studies demonstrate the need for MPs in the BM niches during development and how they coincide with post-natal functions to maintain skeletal and BM haematopoietic niches.

1.2.7 Spleen as a haematopoietic organ:

The spleen functions to clear blood-borne pathogens and acts as an early haematopoietic organ during development, bridging erythro-/haematopoiesis between the FL and BM. It is divided into red pulp and white pulp fractions separated by the marginal zone. The macrophages in each zone have specific functions and interactions and originate from the FL and fetal BM (*114*).

Human red pulp macrophages (RPM) form a vast network required for the uptake of senescent red blood cells and iron homeostasis. They express CD163, CD68, and are SPI-C-dependent. RPMs selectively upregulate SPI-C expression, driving HMOX1 expression, which encodes the essential heme recycling enzyme, Heme Oxygenase 1 (HO-1) (*183*, *184*). Splenic monocytes may also express SPI-C when induced by free heme from red-blood cell degradation, generating new RPMs (*183*). Prenatally, RPMs localise in splenic cords and assess the condition of erythrocytes. CD47 expression on erythrocytes inhibits phagocytosis via interaction with the signal regulatory protein α (SIRPα) found on RPMs. Conformational changes to CD47 indicates erythrocyte senescence leading to phagocytosis by fetal RPMs (*185*).

White pulp and germinal centre formation occurs in the presence of CD209 in humans and SIGN-R1+ in murine marginal zone macrophages (MZM) (*186–188*). After birth, MZM and marginal metallophilic macrophage generation are dependent upon the nuclear liver X receptor (LXR) and function to filter the blood as it is released into the marginal zone (*189*). The macrophages act like scavenger cells via scavenger receptors such as MARCO, which recognise non-opsonised molecules and blood-borne antigens. MARCO also directly binds and mediates the phagocytosis of bacteria such as E.coli and S.aureus and works in conjunction with TLRs to mediate pathogen control (*190*).

Splenic pre-follicular DCs secreting CXCL13 and driving B-cell chemotaxis also contribute to white pulp and marginal zone development after birth (*188*). Human fetal spleen cDC1s and cDC2s, observed by 13PCW, have been observed to induce differentiation of T-regulatory (Treg) cells in vitro from adult T-cells. Fetal spleen cDCs also show significantly less pro-inflammatory cytokine production when compared to adult spleen DCs, including increased expression of arginase-2, consistent with the notion of fetal tolerance establishment (*191*).

1.2.8 Thymus MP system

During early human gestation, CD45+ early lymphoid progenitors (ELP) have been reported to colonise the fetal thymus from the FL and BM and give rise to plasmacytoid and conventional DC subsets and T-cells, (*7*, *192*). T-cells undergo positive and negative selection during development. Double positive CD4+/CD8+ cells that do not recognise Major Histocompatibility Complex (MHCs) on thymic epithelial cells (cTECs) and single positive cells that respond to self-antigens are eliminated by apoptosis, a process essential for positive selection (*193*). Negative selection occurs predominantly in the medulla, where medullary TECs (mTECs) express the transcriptional regulator AIRE and present a diverse repertoire of tissue-restricted antigens, directly deleting high-affinity self-reactive thymocytes and promoting the development of $FOXP3^{+}$ regulatory T cells (*7*).

Thymic DCs complement mTEC-mediated tolerance by cross-presenting self-antigens, thereby enhancing the efficiency of clonal deletion (*193*). Thus, although depletion of thymic DCs can increase susceptibility to autoimmunity, the core negative‑selection program driven by mTECs remains intact (*193*). In the fetal thymus, DCs further support tolerance by expressing chemokines such as CCL17, CCL19 and CCL22 to recruit $CD4^{+}$ thymocytes and nascent Tregs into the medulla, while AIRE expression remains confined to mTECs and underpins their dominant role in negative selection (*7*, *192*, *193*).

The early thymic MP system is composed of Mac1+ (CD11b/CD18) thymic macrophages (TMs) observable from 8PCW (*7*), but little is known of their prenatal function. During murine development, TMs phagocytose apoptotic thymocytes, assisting with the clearance of negatively selected cells, carrying out DNA fragmentation via DNase-II-dependent degradation in lysosomes (*194*). DNase-II knockout mice with impaired macrophage function during development displayed reduced brain, kidney and thymic size due to accumulation of

undigested apoptotic cell debris within phagocytes (*194*). A study on E14.5 mice thymi demonstrated CD4+/CD11b+ macrophages exhibiting phagocytosis of apoptotic thymocytes (*195*). RUNX1 knockout mice have impaired macrophage development and display impaired thymic development, including the accumulation of double-negative thymocytes (*196*).

1.2.9 Lung MP system

Macrophages have been detected in the human fetal lung from as early as 5–6PCW, when arterial, capillary, and lymphatic endothelial cells (ECs) are present, along with embryonic erythrocytes and HMOX1+ erythroblasts (*9*, *197*). As development progresses, particularly after 11PCW, the relative numbers of lymphoid and myeloid cells increase, with macrophages, innate lymphoid cells (ILCs), and dendritic, NK, T, and B cells becoming more prominent (*9*, *197*). Despite these observations, the specific interactions and functional specialisations of lung macrophages during fetal life remain a significant area requiring further research.

In the postnatal lung, there are two primary types of TRMs: alveolar macrophages (AMs) and interstitial macrophages (IMs). Human AMs express CD64, CD206 and CD163, FABP4, INHBA, SPP1 and MERTK (*198*, *199*). AMs reside on the luminal surfaces of the alveoli and are in direct contact with commensal bacteria, inhaled particles, and host-epithelial-derived factors such as surfactants. In the post-natal steady state, AMs phagocytose excessive surfactant proteins. Mice and humans lacking AMs due to a dysfunction in GM-CSF signalling, develop pulmonary alveolar proteinosis as a result of defective surfactant clearance, suggesting a vital function for AMs (*200*, *201*). Surfactant production in human fetus' has been observed to start between 22 and 24PCW and needs careful regulation to prevent build up (*202*, *203*). This implies that AMs are likely needed around this time or shortly thereafter to ensure proper surfactant clearance and lung function.

IMs are present between the airways in the lung tissue interstitium. They are involved in tissue remodelling, maintenance, and antigen presentation (*204*). Unlike AMs, IMs isolated from human lungs tend to be smaller and exhibit a greater variability in shape (*205*, *206*). They exhibit lower phagocytic activity but show higher expression levels of surface MHC-II (HLA-DR) (*206*). Research involving lung tissue from patients undergoing surgical resection for lung carcinoma has shown that IMs produce higher levels of matrix metalloproteinases (MMPs) and tissue inhibitors of metalloproteinases (TIMP-1) when stimulated by T cell

membranes, indicating a role in tissue remodelling that AMs do not exhibit under the same conditions (*207*). Furthermore, IMs have been shown to secrete higher levels of cytokines such as IL-6, IL-10, and IL-1 receptor antagonist (IL-1Ra) both at baseline and following stimulation with lipopolysaccharide (LPS), suggesting their significant role in modulating immune responses and maintaining homeostasis in the lung tissue (*206*). IMs have also been shown to interact with DCs to influence airway allergic responses (*208*).

AMs and IMs play distinct but essential roles in the human lung. AMs are crucial for surfactant clearance and pathogen defence, while IMs have been shown to be key players in tissue remodelling and immune regulation (*205*, *206*). The presence of AMs in fetal life coincides with the onset of surfactant production, highlighting their importance in lung development and function (*202*, *203*). The increasing fraction of macrophages and other immune cells after 11PCW underscores their likely significance in lung development (*9*), though much remains to be learned about their specific functions and interactions during these stages.

1.2.10 Gut MP system

The human intestinal tract develops distinct morphological crypt-villus features by 12PCW (*209*). MPs such as CD103+ and CCR7+ DCs and macrophages are observed from as early as 14PCW (*210*). Human intestinal macrophages have been shown to express HLA-DR, CD206 and CD209 (*211*). Intestinal macrophages assist with epithelial homeostasis during development as they do in post-natal life (*114*, *212*). The intestinal epithelium rapidly divides and requires constant and continuous ECM remodelling, which the macrophage provides via Wnt1 signalling and secretion of hepatocyte growth factor (HGF) (*212*, *213*).

Post-natally, intestinal macrophages and DCs can penetrate the epithelium through trans-epithelial dendrites (TEDs). TED formation was found to be dependent upon the expression of CX3CR1 and the membrane ligand fractalkine (CX3CL1). This process allows them to sample and capture luminal bacteria for antigen presentation (*214*, *215*). Intestinal goblet cells assist the transfer of antigens from the intestinal lumen to CD103+ DCs (*216*, *217*). These DCs then migrate from the lamina propria (LP) to the mesenteric lymph nodes (MLN) in a CCR7-dependent manner, or within the Peyer's Patches into T-cell zones (*217*).

The human intestinal cDC populations play a crucial role in regulating gut homeostasis and maintaining tolerance towards gut microbiota. These cDC populations are characterised by their expression of CD103 and Sirpα (*218*). LP DCs are considered tolerogenic and assist with gut homeostasis. CD103+ DCs have been observed to metabolise retinoic acid secreted by the liver to induce homing of protective CCR9+α4+β7+ T and B-cells to the gut (*219*, *220*). Other transcription factors such as transforming growth factor β receptor II (TGFβII) (*221*, *222*) and tumour necrosis factor receptor associated factor 6 (TRAF6) (*223*, *224*) maintain cDC tolerance of gut microbiota. cDC antigen presentation promotes the generation of forkhead box P3+ (FOXP3+) inducible Tregs in the MLN and is key to the development of the symbiotic relationship with microbiota (*225*, *226*). Human fetuses start swallowing amniotic fluid from 8-12PCW with data indicating the possible existence of an in-utero microbiome in the amniotic fluid and fetal gastrointestinal tract. This evidence thus suggests a potential role for cDC education and tolerance in fetal gut development (*227*, *228*). Therefore, further investigation into the mechanisms behind fetal cDC function and tolerance could provide insights into the early development of gut immunity and the establishment of microbiota tolerance.

1.2.11 Skin MP system

Various populations of MPs reside in the skin, including Langerhans cells (LCs) and dermal dendritic cells (DCs). LCs are found in human fetal skin at 4-5 PCW, detected as HLA-DR+ and CD1a+, harbouring a mixed DC/macrophage signature (*229*, *230*). LC precursors were observed to acquire CD1c and langerin expression at 9 PCW and grow in number throughout development (*231*, *232*). The early presence and marker acquisition of these cells suggest they play a crucial role in the initial establishment of the skin's immune environment, preparing the tissue for early immune responses and contributing to tissue homeostasis.

Dermal MPs (macrophages and DCs) are seeded prenatally but are replaced in humans over time by classical monocytes (*233*). Classical human monocytes can also be recruited to replace LCs when they are unable to self-renew (*233*). In humans, LCs express CD1A, CD11B, CD11C, CD207, and MHC class II. Dermal DCs express lower amounts of CD1A, CD1C, CD11B, CD206, CD209, and MHC class II compared to LCs (*234*, *235*). Although CD209, also known as DC-SIGN, is expressed on both dermal DCs and macrophages, recent studies suggest its expression is more pronounced in macrophages, indicating a need for careful marker interpretation.

Dermal macrophages play a crucial role in maintaining skin homeostasis and immune defence. Additionally, they promote the proliferation of fibroblasts in damaged tissue to assist with repair through a TGFα, fibroblast growth factor (FGF), and platelet-derived growth factor-dependent mechanism (*236*, *237*). Although human LCs are prenatally derived and share a similar origin with prenatal TRMs, they have additional 'DC properties' in their ability to migrate to draining lymph nodes and initiate an immune response (*238*). LCs coordinate a state of immune tolerance in the postnatal skin but can instruct the adaptive immune system when skin integrity is compromised (*239*). During the early stages of wound healing, LCs are present as an immune barrier and coordinate with dermal macrophages to promote repair in a fibroblast-dependent manner (*235*, *236*). Little is known about the function of LCs during development, but data suggests they could be involved in ECM remodelling alongside dermal macrophages (*238*). More work is required to fully clarify their role during organogenesis.

Postnatally, dermal DCs function as migratory antigen-presenting cells, crucial for maintaining tolerance to self-antigens and initiating immune responses against pathogens (*240*). The human equivalent of murine dermal cDC1 is the CD141+ DC subset, which co-expresses XCR1, CADM1, CLEC9A, and TLR3. These markers are essential for the DCs migratory and antigen-presenting functions. Migratory and resident CD141+ DCs are found in skin-draining lymph nodes, highlighting their role in connecting peripheral and lymphoid immune responses (*240*).

Developmental macrophage cell programs were recently shown to be co-opted in two common inflammatory skin conditions, psoriasis (PS) and atopic dermatitis (AD) (*241*, *242*). These new observations highlight the importance of developmental pathways in inflammatory disease pathogenesis that could be therapeutically targeted. Understanding the development and diverse functions of skin MPs, including LCs and dermal DCs, is critical for developing therapeutic strategies targeting inflammatory skin conditions and improving tissue damage repair strategies. The dynamic roles these cells play in immune surveillance, tissue repair, and tolerance underscore their significance in both homeostasis and disease. Further research into their developmental pathways and functional mechanisms could reveal novel targets for wound healing, and intervention in inflammatory skin diseases.

1.2.12 The future of studying the MP system and its roles in organising tissue

Studies on human MPs have focused on in vitro culture systems from BM-HSC, peripheral blood monocytes, induced pluripotent stem cells (iPSC) (*243–247*) or ex vivo primary MPs isolated from peripheral tissues (*140*, *248*, *249*). Such culture systems allow the study of MP interactions with specific cells in the tissues through co-culture. Murine iPSC-derived macrophages (iMacs) can differentiate into microglia in co-culture with iPSC-derived neurons. Murine iMacs can be differentiated into functional TRMs of the lung and brain when transplanted in vivo showcasing the remarkable plasticity of MPs (*249*). These studies have provided new insights into the classification and roles of primary MPs but suffer from a failure to allow dissection into how MP ontogeny and functions are shaped by their physiological tissue of residence. Recent developments in human tissue organoid culture systems provide new opportunities to interrogate human MPs. Organoids are 3D culture systems that attempt to model in vivo settings by leveraging the intrinsic ability of cells for self-assembly and organisation. The most common form of organoid culture; spheroids, organise aggregated cells with or without hydrogen scaffold substrates and aim to replicate three key features of a specific tissue: the spatial distribution of cells, the biochemical environment, and its mechanical environment (*250–252*). Organoids can facilitate studies on organogenesis, disease pathophysiology and drug discovery in an ex vivo setting (*253*). A study by Kuo *et al (2019)* used an air-liquid interface to generate patient-derived tumour organoids with preserved immune cell types including CD68+ CD14+ macrophages (*254*). Bourgine *et al. (2018)* used a perfusion bioreactor system to create a BM organoid with a human osteoblastic environment that supports HSC function (*255*). These advancements underscore the potential of organoid systems to revolutionise our understanding of MP ontogeny and function within their native tissue contexts, paving the way for deeper insights into the role MPs play in tissue organisation and homeostasis.

1.2.13 Organ-on-chip systems

Organoids can replicate organ-level function but may lack key chemical, spatial or other tissue physico-biomechanical properties distinct from those in vivo e.g., fluid flow and are labour intensive (*250*, *252*). Microfluidic organ-on-chips (OoCs) enable organoids to be cultured in perfused, multiplexed chips which recapitulate tissue-specific mechanical and biochemical parameters at higher throughput (*250*).

MP migration and role in development and disease have been studied using OoC models (*256*, *257*). Monocytes and macrophages embedded in OoCs respond to hypoxia and injury associated signals, such as MCP-1 and IL-6 by migrating down chemokine gradients (*250*, *258*). These MPs express gene-expression programs similar to those observed to be crucial in development, such as programs involved in angiogenesis (VEGF, COX2, Wnt5a, FGF2), tissue remodelling (MMP9) glucose transport (e.g. solute carrier family 2 member 1) and glycolytic metabolism (enolase 2) (*250*, *259–261*). Other OoC models incorporating MPs have been developed for the spleen (*262*), skin (*261*), bone marrow (*260*), liver (*259*), the feto-maternal interface (*263*) and an inter-connected multi-organ platform (*257*).

Seiber *et al (2018)* developed a BM-on-a-chip model consisting of two media perfused micro-channels filled with BM progenitor, stromal, and endothelial cells. Cells were embedded in a hydroxyapatite scaffold to mimic the 3D physiology of BM (*260*, *264*). The BM OoC successfully demonstrated key expression programs known to be essential for sustaining the BM HSC niche in vivo, with qPCR assays showing upregulation of nestin, osteopontin, VEGF, angiopoietin1 and fibronectin expression (*260*, *264*). CD34+ cells isolated from the OoCs could form Colony Forming Units (CFUs) of erythrocytes, macrophages, granulocyte/macrophage, and granulocyte-erythrocyte-macrophage-megakaryocytes, demonstrating the ability of OoC derived BM HSCs to differentiate into various progenies, and maintain functional HSC niches in vitro for up to 4 weeks (*264*). In a separate study by Chou *et al (2019)*, BM OoCs supported differentiation into myeloid and erythroid-primed lineages, whilst improving maintenance of CD34+ progenitors (*260*). Interestingly, post differentiation, myeloid mobilisation and remodelling were also observed across 4 weeks of OoC culture. Selective drug toxicity and recovery from 5-fluoracil and radiation exposures using BM OoCs demonstrated increased biologically mimetic toxicity responses and recoveries compared to their static gel-spheroid counterparts.

Liver OoCs recapitulate the hepatic lobule by patterning hepatocytes and other associated cells via micropillar arrays, which introduce perfusion to maintain functionality over time (*265*). Liver-on-a-chip platforms have been designed to model and investigate various functions of the liver including metabolism, detoxification, and response to pharmaceutical interventions (*265*, *266*). Groger *et al (2016)* assessed the interaction of circulating monocytes and their ability to trigger tissue repair and the repolarisation of Kupffer cells within a polystyrol scaffold embedded in a liver-on-a-chip (*266*). Inflammation stimulated by TLR1

and 2 agonists and lipopolysaccharide (LPS) in the liver OoC caused the release of pro-inflammatory IL-1β, IL-6 and TNFα within 72 hours and anti-inflammatory cytokine IL-10 post 72 hours. The OoC exhibited similar responses to livers undergoing sepsis, in vivo, which caused hepatocellular dysfunction and cell death. There was also a shift from LPS-induced inflammatory macrophages to regenerative polarisation with the introduction of THP-1 monocytes to the system (*266*). These results demonstrate a crosstalk of the liver microenvironment and immune system with higher-throughput phenotypic readouts when compared to spheroid cultures. (*267*).

A single tissue OoC cannot fully recapitulate physiologically relevant pharmacokinetic properties and toxicity responses across multiple tissues. To achieve crosstalk of organs, long-term multi-OoC cultures have been developed to better mimic these multi-tissue interactions (*268*). A multi-OoC model comprising cardiomyocytes, skeletal muscle and liver was successfully used to study THP-1 macrophage response to drug and inflammatory stimuli (*257*). The multi-OoC model incorporated biological microelectromechanical systems (BioMEMS) to non-invasively measure cardiomyocyte electrical (microelectrode array) and muscle mechanical function (cantilever). Liver function was monitored using biomarker quantification of CYP1A1, 3A4, 2C9, urea and albumin. The system facilitated multi-organ response to the drug amiodarone, and revealed the selective THP-1 monocyte activation and infiltration in the cardiac OoC due to cytokines released by cardiomyocytes. LPS and IFNγ treatment of the chip system elicited a sepsis-like response characterised by TNFα, IL-6 and CCL5 and decreased cardiac, skeletal muscle and liver function.

Taken together, these studies demonstrate the effectiveness of OoC systems to study MPs within tissues, as well as their responses to chemical stimuli (*264–266*, *269*, *270*). However, despite the promising applications of organoids and OoCs, challenges in recapitulating organs to scale with accurate tissue architecture (size, cell number and distribution) remain (*265*, *266*).

1.2.14 Single cell atlases of fetal MP development

To fully establish an atlas of fetal myeloid progenitor (MP) populations and their functional relevance to tissue development, it is essential to transcend traditional murine models and

incorporate advanced biomimetic models such as organoid systems and OoCs that recapitulate human fetal and embryonic development (*2*). Increasingly, international consortia initiatives such as the Human Development Cell Atlas (HDCA) have leveraged high-throughput, unbiased, technologies such as scRNA-seq and spatial techniques to create tissue specific cellular atlases of the developing human (*2*). In a recent HDCA publication, Popescu *et al.* (2019) applied scRNA-seq alongside the spatially resolved Hyperion to define the cellular and spatial composition of the human FL and YS. This study underscored potential signaling pathways through which Kupffer cells may influence B lineage survival in the FL, showcasing the intricate cellular interactions within these early hematopoietic sites, demonstrating the potential for high-throughput omics technologies to inform on the in vivo cellular interactions between MPs and their microenvironment (*6*).

The study by Suo *et al.* provides a comprehensive analysis of macrophage subsets across various prenatal organs and timepoints, integrating scRNA-seq data with spatial transcriptomics using the Visium platform. The organs included in the study were the YS, liver, BM, thymus, spleen, mesenteric lymph node (MLN), skin, kidney, and gut. This integration allowed for the cross-tissue contextualisation of different macrophage subpopulations and their functional roles in tissue-specific microenvironments revealing dynamic changes in macrophage populations over gestation (*2*).

Among the macrophage subsets identified, *LYVE1*^{hi} macrophages were notably enriched in early gestational stages across multiple organs, including the yolk sac, liver, and skin. These cells exhibited high self-renewal potential and expressed proinflammatory markers indicative of angiogenesis and lymphoid tissue morphogenesis. For instance, the proinflammatory phenotype was characterised by the expression of *TNF*, *NF-κB* subunits, and proangiogenic chemokines like *CXCL3*, *CXCL2*, and *CXCL8 (271–274)*. This suggests their involvement in the formation and remodelling of vascular structures during early development. Iron-recycling macrophages were predominantly found in the liver and spleen, with enrichment observed in the later stages of gestation. These cells expressed markers related to iron metabolism, such as *VCAM1* and *HMOX1 (275)*, and were implicated in maintaining iron homeostasis by recycling iron erythroid cells. Their spatial localisation within these organs underscores their role in supporting hematopoiesis and erythrocyte turnover. MHC class II^{hi} macrophages, enriched in the liver and BM, showed an increase in abundance around 12-14PCW. Expression of MHC II molecules on these macrophages is crucial for antigen presentation and the activation of adaptive immune responses (*2*). The spatial transcriptomics data revealed

their colocalization with developing lymphoid structures, suggesting their role in initiating and shaping immune responses.

Kupffer-like macrophages were primarily localised in the liver, consistent with the role of Kupffer cells hepatic phagocytosis. These cells expressed markers such as *APOE*, and *CAV1* indicating their specialisation in maintenance of lipid metabolic byproducts. Microglia-like *TREM2*^{hi} macrophages were only identified in earlier gestation stages prior to 14PCW and were identified in the skin, yolk sac, and liver, with high expression of *TREM2*, and *P2RY12*. However, brain tissue was not included in this study, thus missing a direct comparison to similar stage brain microglia. Osteoclasts, identified by markers such as *ACP5* and *MMP9*, were enriched in BM (*2*).

Proliferating macrophages, which exhibited high levels of proliferation markers like *MKI67* and *TOP2A*, were abundant in the yolk sac and within the *LYVE1*^{hi} subset across various organs. These cells were particularly enriched in early gestational stages, indicating active self-renewal and expansion. Conversely, macrophages enriched in later gestational stages, particularly iron-recycling and MHCII^{hi} macrophages, upregulated genes encoding immune effector functions. This transition coincides with the development of the adaptive immune system and the expansion of lymphoid tissues (*123*).

Overall, the study's findings highlight the diverse roles and spatial dynamics of macrophage subsets in human prenatal development. By integrating single-cell and spatial transcriptomics data, Suo *et al*. provided a comprehensive atlas of macrophage heterogeneity and their functional specialisation across different developing immune tissues and stages of gestation. However, the study did not include fetal brain tissue, which limits direct comparisons between brain microglia and the microglia-like populations identified in other tissues. Addressing this gap in future research could provide an even more complete understanding of macrophage dynamics and development.

Further enriching the field, organoid culture models have also benefited from the recent surge of data and accessibility provided by omics technologies (*276*, *277*). These models leverage data from extensive genomic repositories to enhance their biomimetic capabilities and complexity, effectively mimicking in vivo conditions and developmental trajectories. (*6*, *237*, *238*). For example, Lee *et al. (2020)* developed hair-bearing human skin organoids (*276*). These organoids were characterised by scRNA-seq at 1-week and 1-month timepoints.

Morphological comparisons showed progressive maturation, with the 1-month organoids closely resembling human 18PCW fetal skin (*276*). FGF and a BMP-inhibitor were used to induce differentiation of human pluripotent stem cell (hPSC) spheroids into cranial neural crest cell populations, key in later assembly of the epidermis. Crucially, differential expression analysis of organoid scRNA-seq data revealed key signalling modulators for self-organisation; expression of WNT modulators in the epidermal (*WNT6*, *LEF1*) and dermal (*SFRP2, TCF4, WIF1, APCDD1*) layers highlight critical pathways which may govern the tissue assembly and interactions between the respective layers, mirroring in vivo development. Furthermore, dermal expression of *FGF7* (Keratinocyte Growth Factor), was also identified as a key driver of epidermal stratification in the organoids (*276*). However, the absence of immune cells and MPs in these organoids suggests that they may not capture the full complexity of in vivo tissue dynamics.

Highlighting the critical role of early hematopoiesis, recent advancements in organoid and iPSC technologies provide a pivotal foundation for modelling the earliest stages of human hematopoiesis, particularly from the yolk sac (YS). Recent work by Alsinet *et al.* provides an insight into the process of modelling early YS haematopoiesis through in vitro myelopoiesis from induced pluripotent stem cells (iPSCs), mapping over 470,000 cells via scRNAseq and scATACseq (*89*). Their research delineates the differentiation trajectory of myeloid cells in a manner that reflects early YS hematopoiesis, crucially preceding the emergence of definitive life-long hematopoietic stem cells (HSCs). Employing a refined protocol adapted from van Wilgenburg *et al.* , this approach uses a feeder-free system that shepherds iPSCs through embryoid body formation, myeloid differentiation, and final macrophage production stages (*89*, *278*). This method leverages a cocktail of cytokines, including M-CSF (Macrophage Colony-Stimulating Factor) and IL-3 (Interleukin 3), vital for promoting myeloid lineage commitment and maturation under conditions that mimic the early human YS environment, encouraging the emergence of myeloid-biased progenitors (*89*).

Alsinet *et al.* identified key transcription factors that orchestrate the early stages of myeloid cell fate decisions. These include RUNX1, which is pivotal for the initiation of hematopoiesis (*279*); GATA1, which plays a significant role in erythroid and megakaryocytic lineage differentiation (*280*); and SPI1 (also known as PU.1), which is crucial for the development of various myeloid cells (*281*). These transcription factors were dynamically regulated across the differentiation timeline, highlighting their roles in the stepwise specification and maturation of iPSC-derived myeloid cells. Additionally, RNA velocity and pseudotime analyses were

utilised to reconstruct the developmental pathways from iPSCs towards distinct myeloid lineages. Comparing iPSC-derived myeloid cells to scRNA-seq data of macrophages from the decidual–placental interface, revealed that the myeloid cells derived from iPSCs display transcriptional and epigenetic features characteristic of YS-derived myeloid progenitors. This suggests that the iPSC-derived myeloid cells could serve as a model for the earliest wave of human hematopoiesis observed in the yolk sac (*89*). This aspect of the study is particularly crucial, as it demonstrates the potential of iPSC-based models to recapitulate and study complex biological processes such as the early immune system development, which is otherwise challenging to observe directly in human embryos (*47*).

These insights not only enhance our understanding of human hematopoietic development but also highlight the potential of iPSCs as a powerful tool in disease modelling, drug discovery, and regenerative medicine. Human fetal tissues are rare and unsuitable for perturbation studies, and animal models may not accurately represent human development (*47*, *89*). Organoids such as this offer a unique advantage by allowing gene knockdowns and other manipulations, enabling the demonstration of the crucial roles of specific genes or transcription factors in development. By closely replicating early YS hematopoiesis, this model offers a valuable platform for studying the molecular mechanisms underlying hematopoietic lineage commitment and the functional maturation of myeloid cells during early development.

1.2.15 Future Directions in MP Research

A comprehensive understanding of myeloid progenitor (MP) roles in tissue repair and regeneration holds profound implications for regenerative medicine and therapeutic interventions (*282*). Traditionally, investigations into the key roles of MPs during human organogenesis and organ morphogenesis have depended heavily on animal models, which, while insightful, often fail to capture the complexities of human biological processes (*47*). The study of MPs in human prenatal tissues faces numerous logistical challenges, from ethical considerations to technical limitations in sample collection. However, the advent of organoid and Organ-on-a-Chip (OoC) technologies, empowered by the surge of data from emerging omics technologies and bolstered by international collaborative efforts like those of the Human Development Cell Atlas, has started to forge new experimental and discovery

pathways (*47*). These offer promising platforms that more accurately recapitulate human development and physiological conditions, addressing the longstanding hurdles in the field.

These systems have shown initial success in modelling early yolk sac (YS) hematopoiesis, providing a unique window into the earliest stages of human immune system development. Employing iPSC-derived organoids that closely mimic early YS hematopoiesis allows the delineation of the molecular mechanisms that govern hematopoietic lineage commitment and the functional maturation of myeloid cells during early development (*89*). This not only enhances our understanding of human hematopoietic development but also underscores the potential of iPSCs as powerful tools in disease modelling, drug discovery, and regenerative medicine.

Yet, significant hurdles remain. The successful incorporation of a full complement of immune cells into these models, along with achieving physiologically relevant mechanical forces, vascularisation and perfusion in organoid and OoC systems, are critical milestones yet to be fully realised (*283*). Addressing these challenges will be essential for these innovative models to fully reflect the intricate interplay of cellular processes that occur in vivo and to translate these findings into clinically relevant applications.

1.3 Overview of single cell data acquisition, integration, and analysis

The widespread adoption of droplet-based single-cell RNA sequencing (scRNA-seq) methodologies such as the 10x Genomics scRNA-seq platform has significantly advanced our understanding of cellular heterogeneity across various biological systems. International collaborative efforts, such as the Human Cell Atlas (HCA) and the Human Developmental Cell Atlas (HDCA), have been pivotal in this progress, generating an abundance of publicly available atlas data. These initiatives have established standardised protocols for data processing, transformation, integration, and deposition, facilitating ease of access, peer review, and integration into larger studies.For instance, the HCA consortium mandates that published data be deposited onto large publicly accessible platforms. Raw sequencing data in FASTQ format is typically stored on the EBI-ArrayExpress platform, while processed counts and code repositories are deposited on Zenodo. The EBI BioImage Archive is used for imaging and spatial in-situ imaging data, and GitHub serves as the repository for analysis code. These standardised practices ensure that data is readily available for reuse and further integration, thus promoting transparency and reproducibility in scientific research.The recent surge in the availability of large pan-organ atlas data has made it commonplace to integrate and contextualise new findings within the framework of these publicly accessible datasets. This practice enhances the robustness of scientific outputs by allowing researchers to compare their results against a comprehensive reference. Given the extensive use of publicly available scRNA-seq atlas data, it is crucial to understand the generic tools and methodologies used in tissue preparation for scRNA-seq data generation. This includes recognising the unique features and limitations introduced by various experimental conditions, as well as addressing issues such as batch-to-batch variation. Overcoming these challenges is essential for accurate data interpretation and integration. For the purposes of this thesis, it is important to highlight the experimental and physical limitations of droplet-based scRNA-seq data-capture modalities. These modalities constitute the bulk of the data generated, integrated, and utilised throughout this thesis. Understanding these limitations and sources of experimental variability enables us to incorporate modelling decisions during integrative analyses of these atlases, such as accounting for sources of technical batch effects, these may include experimental batches, sequencing technologies, donor variability, library preparation methods, and sequencing batches, thereby enhancing the accuracy of our interpretations. This brief overview will cover key features of scRNA-seq experimental considerations and will set the stage for the integration and analytical modelling strategies used in our exploration of these

large atlases throughout this thesis. These considerations include: tissue preparation, data generation, technology features, experimental limitations, batch-to-batch variation and interpretation.

1.3.1 Tissue preparation and data generation

scRNA-seq experiments often begin with tissue preparation. Appropriate tissue digestion and preparation protocols are essential to ensure integrity of downstream analyses and interpretation as they may deplete or preclude specific cells. These protocols typically involve various combinations of mechanical dissociation, enzymatic digestion, and enrichment, to obtain cell suspensions suitable for droplet-based scRNA-seq platforms.

In solid tissues, alterations in digestion time, concentrations, and choice of enzymatic agents such as pepsin, collagenase or proteinase K, can sufficiently alter or cleave cell surface antigens. These antigens are often used to isolate and enrich rare or target cell populations (*6*, *13*, *284–287*). Conversely, under-digestion may lead to the inadequate extraction of specific cell populations.

Isolating and studying rare immune cells is often challenging due to their low abundance. Recently, Fluorescence-Activated Cell sorting (FACS) has become a common technique used to isolate and enrich purified fractions of single cells (*288*). In FACS methods, cell suspensions are first tagged with a targeted monoclonal antibody conjugated with a fluorophore. This enables either positive or negative sorting via electrostatic deflection for droplets containing distinct populations. For instance, CD45+ FACSisolation methods enrich for leukocytes which express the CD45 surface marker or live/dead sorting with the DAPI stain. This technique increases the proportion of immune cells in a sample, enabling detailed analysis of rare subsets and improving the detection of immunological events. This approach has been used to study rare immune cell populations in various diseases, including the identification of novel immune cell subsets in cancer and autoimmune diseases (*3*, *5*, *6*, *12*, *288*).

One advantage of using FACS-based enrichment strategies include the ability to design scRNA-seq experiments with built-in proteomic sort metadata which is useful for informing downstream cell type interpretation (*1–3*, *289*). However, potential limitations of this technique include the requirement for pre-established monoclonal antibodies, potential impact

on downstream differential abundance analyses, and the requirement for large starting cell numbers (>10,000). Recently, several statistical and analytical frameworks have sought to account for FACS acquired abundance biases by including detection event counts information in differential abundance analyses (*1–3*, *289*). For instance the MILO package developed by Dann *et al.* performs differential abundance testing using overlapping neighbourhoods on a pre-computed k-nearest neighbour graph (KNN), allowing users to define a FACS correction factor for each sample to be used as a confounding variable (*1*, *2*, *289*).

1.3.2 data generation

scRNA-seq can be performed using a variety of experimental platforms, with plate-based methods such as Smart-seq2 (SS2), and droplet-based methods like 10x Genomics being particularly popular.

The Smart-seq2 (SS2) method is effective at capturing full-length transcripts at single cell resolution. It involves using FACS to sort single cells into 96 or 384 well plates. The cells then undergo reverse transcription, template switching, PCR pre-amplification of cDNA, followed by library preparation and sequencing (*20*). The commercial 10x Genomics protocol, on the other hand, employs a droplet-based microfluidics approach. This method encapsulates individual cells into droplets. Inside these droplets, mRNA from each cell is reverse transcribed on cell-specific DNA barcoded beads. This is followed by library preparation and sequencing (*290*). Droplet-based methods like 10x Genomics offer higher throughput compared to plate-based methods, enabling the analysis of thousands of cells in a single experiment. However, due to the size of these large libraries, the 10x Genomics platform leverages short-read sequencing to improve the cost-effectiveness of sequencing, capturing only short fragments of transcripts on either the 3’ or 5’ end depending on the tagging chemistry used (*290*).

Microfluidic technologies have revolutionised single-cell isolation by enabling the precise handling and processing of individual cells . Devices use microfluidic channels to capture single cells in droplets or wells, allowing for highly efficient, high-throughput processing and minimising sample loss. The 10X Genomics platform, for example, uses a microfluidic device to generate microdroplets allowing the capture of single cells in monodisperse nanolitre aqueous droplets within a continuous oil phase. These droplets contain a predictable poisson distribution of encapsulated single cells, ensuring efficient and scalable manipulation of

thousands of cells. This platform has been instrumental in numerous single-cell studies, including the creation of comprehensive cellular atlases of various tissues (*2*, *3*, *6*, *12*, *13*, *19*, *291*). One significant advantage of microfluidics-based droplet-based scRNA-seq methods, such as the 10X Genomics platform, is the experimentally conditioned, unbiased nature of cell selection within the technical constraints of the technology. However, a notable technical limitation in microfluidic cell isolation solutions is accommodating differences in cell size. The 10X genomics platform is demonstrated to handle a range of cell sizes (*292–294*), however, recent computational fluid dynamic simulations have shown that larger cells (~30μm) such as hepatocytes, keratinocytes, or adipocytes, may delay droplet breakup at the microfluidic junction. This can lead to the formation of satellite droplets, whilst smaller cells are more evenly distributed across droplets (*292–294*).

Cell sizes and clumping may influence the biases caused by droplet generation dynamics, affecting capture and encapsulation efficiencies. Therefore, proper tissue digestion and cell preparation are essential to ensure cells are appropriately suspended. However, alterations in the choice of tissue preparation strategies change the microfluidic inputs and thus potentially skew downstream cellular landscapes and differential abundance analyses. International consortia efforts, such as the HCA, have encouraged the standardisation and reporting of tissue preparation methods (*47*, *295*). leading to more standardised digestion and extraction protocols. Protocols are often deposited on platforms such as protocols.io (*296–298*), allowing free public access to optimised protocols for preparing specific tissues such as the fetal gut and skin (*6*, *13*, *284–286*).

1.3.3 scRNA-seq technology features

Despite their high throughput, microfluidic droplet-based methods have limitations due to their reliance on short-read sequencing. These limitations include incomplete transcript coverage, which can hinder the identification of alternative splicing events and the accurate determination of gene isoforms (*28*, *290*, *299*). Additionally, short-read sequencing struggles with complex genomic regions, such as those with high GC content or repetitive sequences, potentially leading to biases in data interpretation (*300*).

scRNA-seq offers a more granular view of cellular gene expression compared to bulk RNA sequencing (RNA-seq). However, scRNA-seq data are inherently noisier and more variable (*28*, *299*). Technical noise and biological variation, such as stochastic transcription, present

substantial challenges for computational analysis (*35*, *300*). While numerous tools exist for bulk RNA-seq data analysis, they are often not directly applicable to scRNA-seq data. Methods such as differential expression analysis, cell clustering, and gene regulatory network inference must be adapted to handle the unique characteristics of scRNA-seq data (*35*, *87*, *301*, *302*). Due to the high level of technical noise, quality control (QC) is crucial in identifying and removing low-quality scRNA-seq data to achieve robust and reproducible results (*87*). QC measures include avoiding multi-cells or dead cells during the cell capture step and discarding samples with insufficient sequencing depth or low mapping ratios after sequencing (*87*).

To illustrate, droplet-based scRNA-seq methods, such as the 10x Genomics platform, typically generate 20,000 to 50,000 reads per cell depending on the specific chemistry and sequencing depth used. For instance, the 3' V1 and 5' V1 chemistries are recommended to be sequenced to a minimum of 20,000 reads per cell, while the 3' V2 chemistry are recommended to yield a minimum of 50,000 reads per cell (10x Genomics, 2021). These read depths, while high, are often insufficient for capturing lowly expressed transcripts or full-length isoforms, leading to incomplete gene expression profiles (*87*, *300*). Moreover, the coverage bias can significantly affect the identification of alternative splicing events, gene isoform determination, and other complex genomic features (*300*).

Sequencing TCR and B-cell receptor (BCR) repertoires using short-read technologies can be challenging. Accurate determination of TCR and BCR clonotypes requires capturing the full-length variable (V), diversity (D), and joining (J) regions. Short-read sequencing often provides insufficient coverage for these regions, resulting in incomplete or ambiguous clonotype assignments (*300*). The commercial 10x Genomics single cell immune profiling platform addresses this challenge by enriching sequences using the 5’ chemistry. These platforms enable the generation of paired transcriptomics and adaptive immune receptor repertoire (AIRR) data. For example, the 10x Genomics platform facilitates the detailed analysis of immune repertoires through the use of specific chemistries that improve the capture of full-length receptor sequences, thus enhancing the accuracy of clonotype assignments. Analytical frameworks such as scirpy (*303*), Dandelion (*304*), or scRepertoire (*305*) can then be employed to further dissect the immune receptor data, providing insights into the diversity and specificity of the immune response.

In conclusion, while the 10x Genomics platform and similar droplet-based methods have significantly advanced scRNA-seq by providing high-throughput capabilities, they have interpretational limitations due to their reliance on short-read sequencing. These limitations include biases in cell capture based on size, incomplete transcript coverage, and challenges in accurately sequencing TCR and BCR repertoires (*87*, *300*). Proper experimental design and the use of complementary methods, such as integrating with different CITE-seq or scATAC-seq methods which simultaneously profile the surface proteome, and transposase accessible regions, respectively, can help mitigate these limitations, enhancing the reliability and interpretability of the data (*35*, *87*).

1.3.4 Experimental Limitations and analytical considerations

Droplet-based scRNA-seq methods, such as those provided by 10x Genomics, have several experimental limitations. One significant issue is the occurrence of doublets, where two cells are captured in the same droplet, leading to mixed transcriptomes. This phenomenon can confound downstream analyses such as cell type identification and differential expression analysis (*306*). To address this, doublet detection algorithms such as DoubletFinder (*307*) and Scrublet (*308*) can be employed to identify and remove doublets from the data. DoubletFinder and Scrublet both assume that doublets arise randomly, in approximate proportion to cell loading density, and follow characteristic transcriptomic profiles in high-dimensional space. If cell sizes or capture efficiencies differ greatly, or if loading concentrations deviate from the model, doublet calls by these methods can be skewed. Consequently, some true cells may be incorrectly flagged, while genuine doublets might remain undetected. Therefore, these methods should be carefully tuned and interpreted with downstream validation, especially in data which contains heterogeneous cell populations.

Another issue is the presence of empty droplets, which are droplets that do not contain any cells but may still have ambient RNA. These empty droplets can introduce background noise into the dataset. Tools like EmptyDrops (*306*) are designed to detect and exclude these empty droplets from the analysis. Tools such as EmptyDrops assume that empty droplets contain minimal true intracellular RNA and model this background as a distinct distribution. If significant lysis has occurred or if a subset of droplets contains partial cell debris, the boundary between genuine low-RNA cells and empty droplets can become blurred, potentially removing real but low-content cells.

Additionally, biases in transcript coverage, capture efficiency, and sequencing depth are prominent in scRNA-seq experiments. These biases can result in dropout events, where certain transcripts are not detected, and uneven transcript representation across different cells (*300*). Specific QC measures are thus essential to mitigate these biases by identifying and excluding low-quality cells, such as those exhibiting low mapping ratios indicative of RNA degradation or poor capture efficiency (*87*, *299*). Common QC practices involve excluding cells with fewer than 2000 reads and fewer than 200 detected genes, as these metrics reflect insufficient sequencing depth and transcriptional diversity (*21*) Genes expressed in fewer than three cells are typically also removed from data to avoid noise from lowly expressed genes (86). Additionally, cells with high mitochondrial transcript content, typically 10-15% mitochondrial reads, are often excluded in human scRNA-seq analyses, as elevated mitochondrial reads are commonly thought to indicate cell stress, ongoing lysis, or apoptosis. However, this threshold is often debated as mitochondrial content can vary substantially depending on tissue type, cellular bioenergetic profiles, pathological conditions, and developmental stages, which have inherently elevated mitochondrial activity in specific populations (*5*, *309*, *310*). Furthermore, Recent oncological scRNA-seq studies further challenge rigid mitochondrial read thresholds, highlighting that stringent filtering at 10% for mitochondrial reads may inadvertently discard metabolically altered cells, such as those observed in the malignant epithelial compartment of breast and colorectal carcinomas which exhibit intrinsically higher mitochondrial activity crucial for cancer progression (*309*, *311*). Consequently, more flexible or data-driven methods, such as the ddqc framework (*309*), are increasingly recommended. The ddqc framework employs dynamic thresholds accounting for both celltype and tissue stratifications, determined by the median absolute deviation (MAD) across multiple QC metrics including total transcript counts, transcript diversity, mitochondrial reads, and ribosomal gene content, to robustly account for variation across different cell types and tissues, providing enhanced robustness against outliers and preventing inappropriate exclusion of biologically relevant populations (*309–311*).

Handling the high variability and noise in scRNA-seq data requires careful analytical approaches. Techniques such as normalisation, imputation, and dimensionality reduction are crucial for reducing technical noise and highlighting biological signals. Methods like scran (*312*) for normalisation, MAGIC (*313*) for imputation, and Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbour Embedding (t-SNE) for dimensionality reduction are commonly used in scRNA-seq analysis.

Feature selection also plays a crucial role in addressing high dimensionality and noise in scRNA-seq data. The genes chosen for downstream analysis significantly influence clustering accuracy, cell-type identification, and dimensionality reduction performance (*87*). While simple approaches often select highly variable genes (HVGs) based on thresholds for mean expression levels (min_mean and max_mean), this strategy can inadvertently exclude biologically meaningful genes or include genes dominated by technical noise or housekeeping functions. For example, stringent upper bounds might exclude highly expressed housekeeping genes that exhibit variability but reflect constitutive cellular processes rather than relevant biological differences. Conversely, lower thresholds could mistakenly remove genes crucial for identifying rare cell states due to their modest mean expression. An alternative and more principled analytical approach involves explicitly modeling the relationship between gene variance and mean expression across the dataset (*312*). Methods such as scran's *modelGeneVar()* address this by fitting a variance-mean trend to distinguish genuine biological variability from technical noise (*312*). This modeling explicitly estimates a baseline variability, attributable to technical factors or transcriptional bursting, and ranks genes by their deviation from this baseline, thereby identifying those genes whose variation exceeds random or technical expectations. When available, external spike-in controls further refine this estimation by providing a more accurate baseline of purely technical variation (*314*, *315*). Even in their absence, statistical frameworks based on Poisson-distributed noise assumptions (especially pertinent to UMI-based datasets) can be employed to robustly estimate and control for technical noise. By systematically accounting for technical noise, expression level biases, such analytical frameworks improve the reliability of feature selection, thus enhancing the accuracy and interpretability of subsequent biological analyses.

Additionally, after feature selection, integrating multiple datasets from different batches or conditions necessitates batch correction methods to avoid batch effects confounding biological interpretations. We discuss tools to perform batch correction and integration in greater detail in the below section titled "batch-to-batch variation and integration".

Maternal contamination can be a significant issue in scRNA-seq studies, particularly in prenatal samples. To address this, tools such as Souporcell can be used to detect and correct for potential maternal contamination. This tool pools data by donor and uses genotype clustering to determine the likelihood of contamination, cells with high alternate genotype likelihood are identified as potential contaminants and excluded from downstream analysis (*2*, *316*). Here, Souporcell's genotype clustering framework assumes that each cell's genotype

corresponds to one of a small number of discrete genetic backgrounds in the sample, for example, maternal versus fetal. Consequently, if mosaicism, mixed maternal lineages, or incomplete genotyping are present, Souporcell may misassign certain cells or fail to detect partial contamination. Thus, it remains critical to verify putative contaminants through known markers, orthogonal data, or through well-designed experimental means. For instance, by including known populations of genotype or species contaminants and assessing the accuracy and sensitivity of genotype assignment (*308*).

Ambient RNA, which is RNA present in the solution but not inside cells, can contaminate the scRNA-seq data. Sources of ambient RNA can include partial cell lysis during digestion, sorting, or microfluidic processing steps, dying cells releasing free-floating transcripts, or high-expressing cell populations shedding excess RNA into the microfluidic environment (*317*). Cellbender (v0.2.0) is a tool that models and subtracts ambient RNA using a Bayesian generative model, explicitly distinguishing ambient RNA molecules from genuine intracellular transcripts. It removes ambient RNA based on a user-specified expected false discovery rate (FDR) and a fixed number of training epochs. Crucially, parameters such as expected-cells (the anticipated number of cells captured) and total-droplets-included (total droplets analyzed) must be carefully tuned to avoid overcorrection (removing genuine intracellular transcripts) or undercorrection (retaining ambient RNA contaminants) (*317*).

Other tools, such as SoupX which estimates ambient RNA contamination by modelling the ambient RNA profile from empty droplets under the assumption that this ambient signature is uniform across cells (*318*), and DecontX, which employs a hierarchical Bayesian mixture model assuming contamination follows a multinomial distribution across cells (*319*). These ambient-RNA correction methods assume relatively homogeneous contamination profiles across droplets, as well as accurate detection of empty or low-RNA droplets for defining a background. Consequently, substantial deviations, such as partial lysis of distinct subpopulations, can violate these assumptions, risking either underestimation or overestimation of ambient RNA contamination. As such, manual inspection of marker genes in known populations post-correction remains crucial.

The computational methods discussed effectively address several shortcomings of droplet-based scRNA-seq technology, such as empty droplets, doublets, and dropout events, many of which are introduced by physical properties of the microfluidic capture process. Although these methods significantly enhance data quality and reliability, should their

underlying assumptions be violated the performance of these algorithms can degrade. This underscores the importance of cross-validation with known markers, iterative parameter tuning, and replicate experiments to confirm data robustness. Ultimately, while computational strategies greatly mitigate technical challenges, continued advancements in capture technology remain essential to fully overcome these limitations and achieve more accurate and comprehensive single-cell profiles.

1.3.5 Normalisation and transformation

Normalisation is a crucial step in scRNA-seq data analysis to ensure that observed differences in gene expression are due to biological variation rather than technical artefacts (*87*). In scRNA-seq, normalisation adjusts for differences in sequencing depth and technical variability across cells, which is essential for accurate comparisons of gene expression levels (*288*, *290*). This process reduces technical variability by scaling the data to account for differences in sequencing depth and efficiency, ensuring that the gene expression levels are comparable across different cells (*288*, *290*).

Standard procedures for normalisation in scRNA-seq include Counts Per Million (CPM) normalisation and log transformation (*290*, *320*). CPM normalisation scales the raw counts by the total number of reads per cell and then multiplies by one million, providing an easy comparison of gene expression levels across different cells. In scRNA-seq, normalisation strategies like CPM are favoured due to their simplicity and effectiveness in comparing gene expression across individual cells (*320*). This approach is straightforward and provides an easy comparison of gene expression levels across different cells. In contrast, TPM (Transcripts Per Million) normalises gene counts by the length of each gene before scaling them to one million (*321*). This method is designed to account for gene length differences, which is more crucial in bulk RNA-seq where gene length can significantly influence read counts. FPKM (Fragments Per Kilobase of transcript per Million mapped reads) normalises for both gene length and sequencing depth, accounting for the number of fragments mapped to each gene (*321*). In scRNA-seq, the primary interest is often in comparing gene expression levels between individual cells, rather than accounting for gene length differences. Since each cell's transcriptome is unique and the sequencing depth can vary significantly, CPM's simplicity in normalising by cell-specific total counts is advantageous. Gene length

normalisation is less critical in scRNA-seq due to the short read lengths and high dropout rates, where many genes may not be detected in some cells (*300*). In the context of scRNA-seq, this makes relative expression levels of genes within a cell or across cells more relevant rather than absolute quantification. CPM facilitates this by providing a direct measure of gene expression relative to the total transcriptional activity of each cell.

Transformations such as the log transformation, often log2 with an added pseudo-count (e.g., +1), help stabilise variance and improve the normality of the data by compressing extreme values, (e.g down-weighting highly expressed genes whilst proportionally up-weighting lowly expressed genes, thereby linearizing the expression range) improving the ability to detect biologically relevant signals (*87*, *322*). These transformations are essential for downstream analytical and statistical methods such as PCA to derive linearly decomposed representations of the data (*87*), the Wilcoxon rank-sum test and differential expression analysis using tools like DESeq2 and MAST (Model-based Analysis of Single-cell Transcriptomics) (*87*, *323*, *324*). Normalisation and transformation are essential for differential expression testing because they reduce the impact of technical noise and ensure that the statistical methods applied can correctly identify true biological differences (*87*). Without proper normalisation, differences in sequencing depth and efficiency could lead to false positives or mask true biological signals. Log transformation, in particular, reduces the skewness of the data, making it more suitable for statistical tests that assume normally distributed data (*87*, *322*).

The Wilcoxon rank-sum test is a non-parametric test commonly used in scRNA-seq for comparing gene expression between two groups (*87*). It ranks all data points and compares the sum of the ranks between groups. This test does not assume a specific data distribution, making it robust in many scenarios. However, it is often outperformed by more sophisticated methods like MAST (*324*, *325*). MAST is specifically designed for the unique challenges of scRNA-seq data, such as sparsity and heterogeneity. MAST employs a hurdle model that combines logistic regression to model the binary outcome of whether a gene is expressed, and linear regression to model the expression levels among the expressed genes. Historically, this approach has been justified by assumptions of high zero inflation in scRNA-seq data. However, there is now growing consensus in the literature that droplet-based scRNA-seq data are not inherently zero-inflated, and observed zeros primarily reflect genuine biological variability or differences in gene expression abundance across cells, rather than technical artefacts (*326*). Nevertheless, the MAST model remains effective because it explicitly addresses the sparsity and heterogeneity characteristic of scRNA-seq data (*324*). DESeq2,

originally developed for bulk RNA-seq, adapts well to scRNA-seq by using a negative binomial model to account for the discrete nature of count data and inherent variance, with shrinkage estimators improving stability and interpretability (*323*).

Recent benchmarks indicate that methods like MAST perform well on pseudocells of scRNA-seq data at moderate sequencing depth (*325*). MAST, which incorporates zero-inflation models, and edgeR combined with observation weights for zero-inflation (ZINB-WaVE), show strong performance (*324*, *325*, *327*). Parametric differential expression methods such as MAST, ZINB-WaVE, DESeq2, and limmatrend have been shown to surpass the Wilcoxon test, especially when batch covariates are modelled to handle substantial batch effects (*302*, *325*, *328*).

Batch effects can introduce systematic differences that obscure true biological signals and complicate normalisation processes (*35*, *88*). These effects can result from variations in sample preparation, sequencing runs, and other technical factors (*35*, *88*). Although scaling methods such as log-normalization and Z-scoring help in stabilizing variance and standardizing expression levels, they do not correct for the systematic, sample-specific shifts introduced by variations in experimental conditions. Log-normalisation stabilises variance, by compressing high expression values and reducing heteroscedasticity (unequal variance across different levels of expression), thereby making the data distribution more symmetrical and suitable for methods reliant on assumptions of normality and equal variance, such as PCA and clustering. Z-scoring standardises data by centering and scaling, which ensures each gene has a mean of zero and a standard deviation of one, aiding in the comparison across cells (*87*, *322*).

In summary, effective normalisation and transformation strategies are essential in scRNA-seq to correct for technical variability, stabilise variance, and prepare data for accurate biological interpretation and comparison. Methods such as CPM and log-transformation specifically address variability driven primarily by differences in library size or sequencing depth. However, these approaches inherently assume uniformity in gene-specific capture efficiency and transcript complexity across cell types, assumptions which may not hold true in highly heterogeneous datasets. Furthermore, differential expression testing methods such as DESeq2 or MAST rely on statistical assumptions including negative binomial distributions that can be sensitive to deviations caused by unexpected noise patterns, potentially skewing effect sizes and statistical significance. Thus, while normalisation and transformation methods

substantially enhance downstream analyses, careful validation using spike-in controls or replicate samples remains crucial. Proper validation ensures that technical artifacts are effectively minimised without inadvertently obscuring genuine biological differences, ultimately enabling robust and reliable biological inference.

1.3.6 Dimensionality reduction techniques

Dimensionality reduction is a crucial step in the analysis of scRNA-seq data due to the high dimensionality and sparsity inherent in such datasets (*87*, *88*, *329–332*). scRNA-seq experiments typically measure the expression levels of thousands of genes across thousands to millions of individual cells, resulting in datasets with millions of data points. This high-dimensional data poses significant challenges for downstream analysis, including clustering, visualisation, and identifying meaningful biological patterns. Dimensionality reduction techniques simplify these data by transforming them into a lower-dimensional space, which makes subsequent analyses more computationally feasible and interpretable (*87*, *329–332*).

Reducing the dimensionality of scRNA-seq data addresses these challenges in multiple ways. First, it enhances interpretability by projecting high-dimensional gene expression data onto a lower-dimensional space. Underlying biological patterns thus become clearer and easier to visualise and interpret. Linear methods such as PCA achieve this by identifying directions (principal components) along which the variance in the data is maximised. PCA transforms the data into a set of orthogonal principal components, each capturing decreasing proportions of the overall variance. This significantly reduces complexity while highlighting the most salient biological signals, such as clusters of cells with similar gene expression profiles or continuous trajectories representing developmental gradients. In very high-dimensional spaces, such relationships among cells and genes are obscured by redundant or noisy features, PCA simplifies these relationships, enabling intuitive visualization and facilitating meaningful biological interpretations (*87*, *332*). Second, dimensionality reduction methods reduce noise by filtering out less informative or stochastic signals, ensuring analyses focus on robust and consistent patterns of gene expression. Third, reducing dimensionality enhances computational efficiency by lowering the number of features, thus enabling faster and more scalable analyses, particularly important for large single-cell datasets. Finally, dimensionality reduction can assist with feature extraction by highlighting the genes or components that

capture most of the variability in the data, guiding targeted downstream investigations (*87*, *329*).

Several techniques are commonly used for dimensionality reduction in scRNA-seq data, including both linear and non-linear methods. PCA projects high-dimensional scRNA-seq data onto a lower-dimensional space by maximising the variance captured in the first few principal components. This method is widely used due to its simplicity and effectiveness in reducing dimensionality while preserving the global structure of the data (*87*, *329*, *332*). Non-negative Matrix Factorization (NMF) is another linear technique that factors the data matrix into two lower-dimensional matrices with non-negative entries (*333*). It is particularly useful for identifying parts-based representations in the data, making it suitable for uncovering underlying biological processes (*334–336*). On the non-linear side, t-SNE emphasises local structure and preserves the pairwise distances between points. It is widely used for visualisation purposes as it can reveal complex structures in high-dimensional data. However, t-SNE assumes the data lies on a smooth, continuous manifold by converting high-dimensional distances into probabilities that emphasize local pairwise similarities. Because its cost function focuses on preserving these local neighborhoods, t-SNE may sacrifice the global arrangement of clusters, causing distances between well-separated groups to be unreliable. The algorithm is also highly sensitive to hyperparameters, particularly perplexity (which sets the effective neighborhood size), and to the choice of random initialization, so different settings or random seeds can yield markedly different embeddings from the same data. Finally, the axes in a t-SNE plot have no intrinsic scale or orientation, which complicates biological interpretation of inter-cluster distances or comparisons across runs (*337*). Uniform Manifold Approximation and Projection (UMAP) is another non-linear technique that preserves both local and global data structure. It employs the laplacian eigenmaps technique for informative initialisation of the embedding, and provides a balance between maintaining the high-dimensional structure and creating a meaningful low-dimensional representation. Nevertheless, UMAP assumes data lie uniformly on a locally connected Riemannian manifold, where the rule for measuring distances (encoded by the metric tensor) is assumed to remain roughly the same from one small neighborhood to the next. If some cell populations are much more densely sampled than others or if there is substantial noise, UMAP's graph-based embedding can overemphasize these density differences and produce layouts that distort true relationships (*338*). Furthermore, it has been shown that, when randomly initialised, UMAP does not preserve global structure more

effectively than tSNE with random initialisation. This highlights the importance of informative initialisation for the UMAP algorithm (*338*, *339*). As neither t-SNE nor UMAP provide a direct measure of interpretable distance in their 2D embeddings, cluster proximities must be interpreted cautiously when making biological conclusions. Thus, verifying dimensionality reduction results through marker-gene analysis or alternative embedding strategies (e.g., PCA or diffusion maps) is strongly recommended.

Neighbourhood abstraction involves identifying and analysing local structures within the dimensionally reduced embeddings. One common method for neighbourhood abstraction is k-nearest neighbours (kNN). The kNN algorithm identifies the 'k' closest data points (neighbours) to a given data point in a multi-dimensional space (*340*). This method is widely used in various scRNA-seq analyses due to its simplicity and effectiveness in capturing local structures within the data (*341*). This process is critical in scRNA-seq analysis for several reasons. Firstly, by grouping cells into expression-homogeneous neighbourhoods, kNN helps in clustering cells with similar gene expression profiles into biologically relevant clusters (*341*). This helps reduce intra-cluster variability, thereby amplifying inter-cluster expression contrasts, and thus simplifies cell-type annotation both by streamlining differential marker-gene discovery as described section 1.3.5 (*87*, *342*, *343*) and by allowing unlabelled cells to be assigned labels based on proximity to annotated neighbours (21). Secondly, it aids in trajectory inference, which involves mapping cell states in a low-dimensional space where neighbourhood relationships are preserved, helping to understand developmental trajectories and lineage relationships among cells, as seen in methods like Monocle (*344*, *345*).

1.3.6 batch-to-batch variation and integration

Batch effects, which are technical variations introduced during different experimental runs, can significantly distort the results of dimensionality reduction (*35*, *87*, *88*). These effects can cause cells from different batches to cluster separately, obscuring true biological differences. Methods like PCA can be particularly susceptible to batch effects as they may capture the variance introduced by batch differences rather than the biological variability of interest (*346*).Therefore, it is crucial to employ appropriate integration and batch correction techniques before or during dimensionality reduction to ensure that the resulting low-dimensional embeddings accurately reflect the underlying biology.

In scRNA-seq analysis, addressing batch effects and integrating data from multiple experiments are crucial for obtaining accurate biological insights. Various methods have been developed to tackle these issues, each with unique strengths and underlying assumptions. Batch correction methods remove confounding variation in high-dimensional space (*88*) and include methods such as COMBAT which uses a hierarchical empirical Bayes framework to adjust for batch effects, making it suitable for integrating datasets with significant batch variability. ComBat assumes a log-normal distribution of gene expression and models batch effects as linear, additive variations. While effective for certain bulk applications, these assumptions contrast significantly with scRNA-seq data characteristics, which include discrete, highly sparse, and overdispersed count distributions (*347*). Consequently, ComBat struggles to accurately model the complex, nonlinear batch effects common in single-cell experiments. Benchmarking studies on scRNA-seq data highlight these limitations, showing ComBat performing inconsistently and ranking lower than methods specifically developed for single-cell integration (e.g., scVI, Harmony, MNN, Scanorama) (*88*). Additionally, ComBat's use of continuous log-transformed data discards the integer nature of scRNA-seq counts, removing the discrete statistical properties (such as overdispersion) leveraged by specialized single-cell differential expression methods (e.g., DESeq2,, MAST). In contrast, alternative approaches explicitly designed for single-cell data have emerged. The Mutual Nearest Neighbours (MNN) algorithm directly corrects batch effects by pairing cells across batches that are mutual nearest neighbours, using these pairs to compute gene-wise correction vectors, and thereby effectively adjusting for batch-specific technical variation while preserving biological structure (*348*).

Integration methods seek to learn lower-dimensional representations of high-dimensional gene expression vectors that minimise confounding variation (*88*, *349*). Some widely used integration methods include: Harmony (*350*), BBKNN (*351*), Scanorama (*349*), scVI (*330*), and Reciprocal PCA (RPCA) from Seurat v3 (*35*). However, the Harmony, BBKNN, and RPCA methods tend to favour the removal of batch effects over the conservation of biological variation, which can be particularly important in complex scenarios such as integrating pan-organ atlases where donors of individual tissues are not shared across organs (*88*). Additionally, MNN, as well as methods like Harmony and BBKNN, assume that overlapping cell populations across batches can be aligned by matching nearest neighbours or centroid positions in latent space. However, if there is minimal true biological overlap between

batches, such as when integrating distinct developmental time points, over-correction can occur, inadvertently removing biologically meaningful variation.

Harmony corrects batch effects by aligning cell embeddings in a shared space through iterative nearest neighbour graph correction and returns a corrected cell embedding (*350*). This method begins with precomputed PCA embeddings. It uses a soft k-means clustering approach to learn cluster-specific linear correction factors and iteratively adjust the positions of cells to ensure that each cluster has a balanced representation from each batch. In each iteration, Harmony attempts to minimise batch effects by assigning each cell a cluster weighted average correction term according to the soft k-means clustering assignment. This step ensures that cells from different batches are mixed within clusters, which helps mitigate batch effects (*350*). For instance, Korsunsky *et al. (2019)* demonstrated Harmony's effectiveness in large datasets, showing its scalability and efficiency in aligning cell populations across different experimental batches (*88*, *350*). Nonetheless, Harmony primarily targets confounding variation removal and does not explicitly model the underlying biological variability. Additionally, its effectiveness heavily relies on hyperparameter tuning, particularly the "theta" parameter controlling batch mixing, where suboptimal values may distort genuine biological variation or inadequately correct batch effects (*88*). BBKNN (Batch Balanced K-Nearest Neighbors) balances batch contributions by modifying the kNN graph to ensure that each cell's neighbourhood includes cells from all batches and returns a corrected kNN graph object. This approach directly operates in the latent space generated by other dimensionality reduction techniques, such as PCA. Polański *et al. (2020)* highlighted BBKNN's straightforward implementation and computational efficiency in balancing batch effects (*351*). However, BBKNN performance also depends heavily on parameter selection, notably the neighbourhood size (k), making it vulnerable to biological distortion if improper values are chosen. Consequently, robust parameter tuning is critical to the performance of both BBKNN, and Harmony, as is appropriate experimental design, such as including matched donor batches across experimental sites or sequencing technologies, to facilitate accurate biological integration and minimize potential confounding effects. Nevertheless, like Harmony, BBKNN does not explicitly model the underlying biological structure of the data (*88*).

RPCA, introduced in Seurat V3, is designed for batch correction by first employing PCA on individual datasets to reduce dimensionality. Following this, the identification of mutual nearest neighbours (MNNs) across datasets aligns similar cells from different batches. This

alignment effectively integrates data and preserves biological variability. However, like other integration approaches, RPCA can misalign dissimilar populations if certain cell types are absent in some batches or if multiple biological states are confounded with experimental batches. Such scenarios can lead to spurious mixing of distinct cell states or incomplete correction due to the inability to identify suitable matching neighbours. The integration process in Seurat's RPCA adjusts PCA embeddings based on identified MNNs, minimising batch effects while maintaining true biological signals (*35*). This ensures that biological variability is carefully preserved while correcting for batch effects in scRNA-seq data.

Scanorama, demonstrated effectively by Hie *et al. (2019)*, also addresses the challenge of integrating diverse datasets using mutual nearest neighbours to align datasets. Scanorama outputs corrected count matrices (batch correction) and embeddings (integration), making it suitable for various downstream analyses (*349*). It is designed to be scalable and flexible, capable of handling large datasets efficiently, thereby providing a nuanced solution for complex integration tasks while preserving biological signals.

Other methods, such as COMBAT and scVI, also output corrected counts matrices (*330*, *347*). scVI performs both batch correction and integration tasks, integrating batch correction directly into the Variational Autoencoder (VAE) framework whilst learning the underlying biological structure. Unlike methods that perform batch correction and dimensionality reduction as separate steps, scVI models batch-specific distributions during the encoding process, incorporating batch annotations to adjust the latent space. This hierarchical Bayesian approach allows scVI to capture both biological and batch-specific variations, resulting in robust and biologically meaningful latent representations (*330*).

Compared to other methods, scVI has been shown to outperform in both batch correction and preserving biological signals (*87*, *88*, *330*). For example, Korsunsky et al. (2019) and Polański *et al*. (2020) highlighted the efficiency and computational simplicity of Harmony and BBKNN, respectively. However, benchmarking studies have shown that scVI excels in maintaining cell type separation while integrating batches (*87*, *88*, *330*). Even so, scVI's performance relies on sufficient cell numbers and representative sampling of each batch. If a batch lacks certain cell populations or ages, the model could over-correct genuine differences, particularly if scVI attributes them to batch rather than biology. As with other tools, tuning of parameters such as the number of latent dimensions, dropout rates, thorough marker-based validation, and well-designed experiments (e.g., including matched donors across tissues,

sequencing batches, and sequencing chemistries) are necessary to distinguish true biological signals from confounded technical artifacts.

1.3.7 VAE based integration

Variational Autoencoders (VAEs) in scVI model gene expression data as a Z-inflated negative binomial distribution. The encoder, a neural network with layers including fully connected (dense) layers, batch normalisation layers, and activation functions like ReLU (Rectified Linear Unit), transforms the input data to produce the mean (μ) and variance (σ^2) of the latent variable distribution. These latent variables z are sampled from a Gaussian distribution $N(\mu,\sigma^2)$ (*352*). The decoder, also a neural network, transforms z back to the original data space using a softmax transformation to model gene expression probability. The VAE's objective function includes a reconstruction loss, measuring the difference between input and reconstructed data using a negative log-likelihood function adapted to the Z-inflated negative binomial distribution, and a KL divergence term, regularising the divergence between the learned latent distribution and a prior distribution. This combined loss is optimised during training (*330*). However, like most VAE-based frameworks (*330*, *353–355*), scVI's generative assumptions may not perfectly capture all outliers or overdispersed genes, and model performance often depends on carefully tuned hyperparameters (e.g., number of latent dimensions, dropout_rate, learning rate, number of epochs). Systematic optimisation of these hyperparameters is strongly recommended, ideally, this might be conducted by evaluating how robustly the model recovers established biological signals, such as ground-truth cell-type labels or known expression patterns, across various parameter settings. For example, sensitivity to parameters such as dropout_rate (e.g., variations of ±0.1 or ±0.2) or the number of neural network layers could substantially impact latent-space representations and downstream interpretations. Moreover, VAEs typically require large cell numbers to robustly learn complex distributions, and overcorrection or merging of rare subpopulations can occur if batch effects are over-adjusted. While scVI adeptly models gene expression, these non-linear transformations produce latent components that are not directly interpretable at the gene level, unlike PCA or other linear factor methods where loadings can be readily inspected. Consequently, changes in the latent space cannot be trivially attributed to specific genes due to the chain of non-linear encodings and decodings. Thus, thorough validation, by examining reconstruction errors, verifying known marker genes, and exploring multiple hyperparameter settings remains crucial to ensure meaningful biological signals are preserved.

Linearly Decoded Variational Autoencoders (LDVAEs) are a variant where the decoder is constrained to be linear. The encoding process in LDVAEs is similar, but the decoder performs a linear transformation Wz+b, where W is a weight matrix and b is a bias vector, instead of using non-linear transformations. This linear decoder retains interpretability while simplifying the model (*331*). In LDVAEs, gene factor embeddings store the linear coefficients associated with each gene, aiding in interpreting how each gene contributes to the latent variables. Svensson *et al. (2022)* demonstrated that these embeddings could identify key genes in cell differentiation processes, highlighting the interpretability of the linear decoder. Cell factor embeddings contain the latent embeddings of the cells, representing the low-dimensional latent variables learned by the model.

LDVAEs bridge the gap between the interpretability of linear factor models and the efficiency of VAEs. By using a linear reconstruction function, LDVAEs maintain the interpretability of factor models while benefiting from the scalability and efficiency of VAEs. This approach allows for efficient low-dimensional representations of scRNA-seq data, identifying the relationship between cell representation coordinates and gene weights via a factor model (*331*).

In scRNA-seq analysis, using LDVAE for cell type classification has several advantages, including interpretability of gene contributions, computational efficiency, and scalability. Valentine Svensson *et al. (2018)* highlighted that simpler models like LDVAE could efficiently handle large-scale single-cell datasets, providing interpretable results. However, LDVAEs have limitations, such as limited expressiveness and potentially lower reconstruction quality compared to non-linear decoders, by enforcing a strictly linear decoding (Wz+b), LDVAEs may not fully capture complex, non-linear gene interactions, potentially overlooking subtle variations in the data, thus limiting their “expressiveness” (*331*). Additionally, LDVAEs still require careful hyperparameter tuning similar to scVI, and if not validated, they risk omitting important gene-gene relationships that non-linear models could detect. For classification tasks, latent variables from LDVAE can be directly used in logistic regression due to their interpretability and lower-dimensional nature, providing computational efficiency and insights into gene contributions (*1*, *2*, *87*). While standard VAE latent variables can also be used for classification, their non-linear nature complicates interpretation, though they may provide more accurate representations if the data structure is complex (*87*, *330*, *331*).

A hybrid combinatorial approach to cell state classification leverages the strengths of both LDVAE and VAE (*1*). This method uses low-dimensional representations from LDVAE as input into elastic net logistic regression classifiers and aggregates predictions across clusters defined by VAE-based approaches using majority voting. This hybrid method combines the interpretability and efficiency of LDVAE with the expressive power of VAE-based clustering, allowing clusters defined by VAE to provide nuanced biological insights while the linear factors from LDVAE facilitate interpretable classification (*1*). This has a distinct advantage over training logistic regression directly on gene expression information, as it leverages the low-dimensional, interpretable latent space of LDVAE for classification, enhancing both computational efficiency and biological interpretability. Compared to methods that directly apply logistic regression to high-dimensional gene expression data, this hybrid approach mitigates the curse of dimensionality, improves classification performance, and provides biologically meaningful insights by combining the non-linear clustering power of VAE with the interpretability and computational advantages of LDVAE latent representations (*1*, *343*, *356*).

scANVI (Single-cell ANnotation using Variational Inference) is a method designed for label prediction through transfer learning, employing a neural network classifier to leverage annotated datasets for the prediction of cell types in new, unlabeled datasets. This approach integrates scVI's generative model with a classification network, effectively combining unsupervised and supervised learning to enhance label transfer accuracy (*343*). While scANVI can produce highly accurate label predictions, the interpretability of the model is more challenging compared to LDVAE due to its complex, non-linear nature (*87*, *330*, *331*).

In contrast, the hybrid method, which leverages low-dimensional representations from LDVAE as input into elastic net logistic regression classifiers and aggregates predictions across clusters defined by VAE-based approaches, offers enhanced interpretability. This approach combines the strengths of both models: the interpretability and efficiency of LDVAE with the expressive power of VAE-based clustering. By defining clusters using VAE and applying majority voting for predictions, the hybrid method facilitates nuanced biological insights while maintaining the linear factors from LDVAE for more straightforward interpretation of gene contributions (*1*). Nevertheless, the hybrid approach carries notable assumptions and limitations. It presumes that LDVAE’s linear embeddings preserve all biologically relevant variation required for classification and that VAE-derived clusters accurately reflect underlying cell states. Conflicts may arise if subtle, non-linear patterns

captured by VAE are lost within LDVAE's linear factor space, potentially masking rare or transitional subpopulations during majority voting. Additionally, implementing both generative models along with a classifier introduces increased computational costs, and optimally balancing hyperparameters across both frameworks can be challenging.

Using LDVAE with a linear decoder offers a balance between interpretability and computational efficiency, making it a suitable choice for tasks like cell type classification in scRNA-seq data (*331*). While standard VAEs provide more expressive power, the simplicity and interpretability of LDVAEs can be advantageous, particularly when combined with logistic regression for classification tasks. The hybrid combinatorial method combines the strengths of both approaches, leveraging the expressiveness of VAEs and subsequent derived clusters, and the interpretability of LDVAEs to enhance both accuracy and biological insight in scRNA-seq analysis. Moreover, scVI's sophisticated modelling of batch-specific distributions corrects for batch effects, providing robust and biologically meaningful latent representations. Compared to methods like Harmony, BBKNN, and Scanorama, scVI demonstrates superior performance in both batch correction and preserving biological signals, as evidenced by benchmarking studies (*88*, *330*, *331*, *347*, *349–351*).

1.3.8 Clustering and annotation

Clustering is a crucial step in scRNA-seq analysis, aimed at grouping cells with similar gene expression profiles into distinct clusters. These clusters often correspond to different cell types, states, or subpopulations, enabling the exploration of cellular heterogeneity within a tissue or organism. Effective clustering relies on the choice of dimensionality reduction, distance metrics, and clustering algorithms (*87*, *290*).

Dimensionality reduction techniques such as Principal Component Analysis (PCA), UMAP , t-SNE, and single-cell Variational Inference (scVI) are commonly used to reduce the high-dimensional gene expression data into a more manageable and interpretable form (*87*, *330*). PCA is often used as an initial step to retain most of the variance in fewer dimensions, while UMAP and t-SNE provide non-linear transformations that can capture complex, non-linear relationships in the data (*87*). scVI leverages a Bayesian framework within a

variational autoencoder to model both biological and technical variations, providing robust and biologically meaningful low-dimensional representations (*330*).

Once the data has undergone pre-processing and dimensional reduction, various clustering methods are applied to group cells with similar transcriptional profiles. Among the commonly used methods are partition-based, optimisation-based, and density-based algorithms (*87*, *290*, *357*).

K-means clustering is a widely used partition-based method. It involves dividing the data into a predefined number of clusters (k). Initially, k centroids are randomly placed in the data space. Each data point is then assigned to the nearest centroid, and the centroids are recalculated as the mean position of all points in the cluster. This process iterates until the centroids stabilise, minimising within-cluster variation (*357*, *358*). Despite its simplicity and efficiency, k-means clustering requires the number of clusters to be known beforehand, which can be a limitation when the actual number of cell types is unknown (*357*).

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular density-based clustering method. It defines clusters as regions in the data space with a high density of points. DBSCAN discards points in low-density regions as noise and identifies clusters based on the local density of data points. Unlike k-means, DBSCAN can detect clusters of arbitrary shapes, making it suitable for capturing complex structures in scRNA-seq data (*359*).

Graph-based clustering methods, such as the Louvain (*360*) and Leiden (*361*) algorithms, have become prominent in scRNA-seq analysis due to their scalability and ability to handle large datasets. Louvain clustering assigns data points to communities based on local modularity, a measure of the density of links within a community compared to links between communities (*362*). The algorithm iterates by treating communities as nodes and reassigning them to maximise modularity until no further improvement is possible. Louvain clustering does not require a predetermined number of clusters, but relies on a termination condition to conclude the clustering process (*360*). Additionally, the Louvain method also has several limitations. Firstly, Louvain's local, greedy optimization of modularity at each step does not guarantee a globally optimal partition, and subtle community structures can be lost when clusters are repeatedly merged yielding arbitrarily badly connected communities (*361*). Second, imbalanced cell numbers or cluster-dependent connectivity can lead to suboptimal

partitioning in single-cell datasets, where rare or transitional subtypes may be absorbed into larger, more dominant clusters. Third, the algorithm depends heavily on a user-defined resolution parameter, and if set too high, biologically distinct populations may be over-split into fragmented clusters, if set too low, functionally distinct subpopulations may be merged. Finally, the iterative, multi-level Louvain process can produce slightly non-deterministic outcomes across runs, reflecting its reliance on initialization and update order. These factors collectively pose challenges for scRNA-seq clustering, where accurate identification of cell subtypes depends on fine-grained resolution and biologically appropriate partitions.

By contrast, the Leiden algorithm was introduced to address the major limitations of the Louvain method, particularly its propensity for suboptimal local merges and its sensitivity to initialization. Leiden ensures that all subsets of communities are optimally assigned locally, resulting in more coherent clusters and faster convergence (*361*). Nonetheless, some constraints remain shared by both algorithms, such as the continued reliance on resolution parameters that can still over-segment or merge subpopulations if misconfigured. Although Leiden typically mitigates these issues more effectively than Louvain, careful parameter selection and biological validation remain paramount to achieving biologically meaningful clusters in single-cell datasets. This method has been integrated into popular scRNA-seq analysis tools such as Scanpy and Seurat, underscoring its effectiveness and widespread adoption (*35*, *363*). Unlike Louvain, Leiden can partition the network more effectively, ensuring that the identified clusters are well-connected and biologically meaningful (*361*).

Following the identification of clusters, the next step is to annotate them, assigning biological semantics such as cell types or states. This process often involves identifying differentially expressed genes (DEGs) that are significantly upregulated or downregulated in specific clusters (*87*). The scVI framework can be used to derive DEGs through its Bayesian modelling approach, which accounts for both biological and technical noise, providing robust estimates of gene expression differences (*330*).

Annotating clusters with biological meaning can be challenging, especially for bioinformaticians without extensive domain specific biological expertise. To address this, several automated annotation tools have been developed, such as scCATCH, Garnett, SingleR, and CellTypist (*356*, *364–366*). These tools use reference datasets and machine learning algorithms to assign cell type identities based on gene expression profiles. For example, SingleR compares cluster expression profiles to reference datasets of sorted cell populations,

while Garnett uses a hierarchical classification framework, and scCATCH employs a consensus approach based on known marker genes. CellTypist utilises a library of pre-trained logistic regression models and can use majority voting of pre-computed clusters for accurate and comprehensive annotations (*356*, *364–366*).

Annotation is a critical step in scRNA-seq analysis, as it provides insights into the cellular composition of the dataset and the success of the experiment. High-quality annotations reveal the presence of diverse and well-characterised cell types, which are essential for downstream analysis and biological interpretation .

1.3.9 Trajectory inference

Trajectory inference in scRNA-seq offers profound insights into cellular differentiation and development by mapping potential differentiation trajectories among cell types and cell states. These computational methods encompass a range of techniques each predicated on specific assumptions about cellular transitions.

Methods such as Monocle3, force-directed graphs (FDG) using ForceAtlas2, diffusion pseudotime, and partition-based graph abstraction (PAGA) are founded on the premise of gradual and continuous changes in gene expression (*82*, *84*, *344*, *367*, *368*). These changes guide the organisation of cells within and between temporal snapshots based on similarities in gene expression, thus forming the basis for constructing pseudotime trajectories. However, these approaches assume that transitions between cell states are smooth and that intermediate states are present in the data. If differentiation is highly asynchronous or if certain key transitional states are undersampled, pseudotime approximations may be skewed or incomplete

Monocle 3 significantly automates and refines the process of trajectory inference. It utilises community detection methods, primarily the Louvain algorithm, to group cells from trajectories of interest. Principal graphs are then constructed for these clusters within reduced-dimension spaces like UMAP, to define the developmental trajectories that cells follow in a pseudotemporal order. This framework facilitates dynamic differential gene expression analysis across pseudotime trajectories. Differential gene expression analysis can then be performed using generalised linear models that accommodate the high frequency of zero counts in scRNA-seq data, or graph-autocorrelation analysis for genes that significantly

vary across a pseudotime axis, providing insights into gene expression variations over pseudotime, rather than across static clusters (*367*).

ForceAtlas2, employed in FDG, uses solutions derived from physics simulations where nodes (cells) repel each other like charged particles, and edges (expression similarities) attract, this then converges into a balanced state reflecting the inherent data structure in an intuitive visual manner (*84*). Diffusion pseudotime treats data as if undergoing a diffusion process, constructing pseudotime trajectories by simulating a random walk through data points, ideal for capturing continuous developmental stages (*368*). PAGA, on the other hand, abstracts connectivity between cell clusters rather than individuals, offering a clear view of large-scale structures and potential bifurcations, beneficial for complex topologies (*82*). However, each of these graph-based or physics-based algorithms presumes relatively dense sampling and stable correlation structures among transitioning cells. Large gaps or abrupt lineage branches may be masked if the model imposes a single, continuous manifold. Parameter choices , for example, knn size in PAGA, or minimum edge weight, can drastically alter the inferred topology.

Contrasting these methods, RNA velocity and CellRank are methods that incorporate assumptions about the burst kinetics of RNA transcription, focusing on the proportions of spliced and unspliced RNA to provide a refined temporal framework for organising cells (*29*, *83*). RNA velocity specifically attempts to predict relationships between cells and likely future cell-states by analysing the dynamic distribution between unspliced and spliced mRNAs captured. This adds a directional component to trajectory inference, known as a directional transition matrix which highlights localised probabilities of potential transition between cells, thereby adding a temporal dimension to static single-cell data (*29*). RNA velocity assumes that splicing kinetics and transcript degradation rates remain sufficiently consistent to be estimated globally. If certain genes exhibit highly atypical splicing patterns or deviate significantly from the standard kinetic model, velocity vectors can become imprecise and lead to misleading transition inferences. Additionally, datasets biased toward fully differentiated states may undersample crucial intermediate populations, limiting the reliability of pseudotime ordering.

CellRank builds on this concept by integrating Markov chain models to probabilistically estimate the likelihood of cells transitioning to specific future states, starting with a coarse-grained inference of cell fate probabilities informed by RNA velocity or other lineage indicators (*83*). This initial analysis sets the stage for a more refined estimation process that

considers the entire transcriptional complexity of the data. Using the generalised Perron-Frobenius operator, CellRank then quantitatively estimates the probabilities of each cell transitioning into terminal states, effectively mapping out the landscape of cell fate decisions across a continuum of pseudotime. The refined understanding of cell dynamics is further enhanced by CellRank's use of real Schur decomposition in its eigendecomposition approach to the transition matrix. This method, suitable for non-symmetric matrices typical in outputs from methods such as RNA velocity, facilitates the stable calculation of eigenvalues and eigenvectors (*369*, *370*). By analysing these, CellRank distinguishes between transient and terminal states. Transient states, identified through eigenvectors associated with eigenvalues less than one, represent intermediate phases that cells transiently occupy before reaching their terminal states. This insight is critical for understanding the progression of cells through various stages of differentiation. In addition, CellRank leverages changes in gene expression correlated with transition probabilities to identify key genes that influence cell fate decisions, providing valuable insights into the regulatory mechanisms driving cellular development (*83*).

1.3.10 Regulatory inference

Understanding the regulatory mechanisms and cell-cell communication in single-cell RNA sequencing (scRNA-seq) datasets is critical for inferring key regulator drivers of cellular functions and interactions within complex tissues. Regulatory inference aims to predict the underlying gene regulatory networks (GRNs) that control gene expression within cells using scRNA-seq data.

One notable tool in this domain is pySCENIC (*371*, *372*), a Python implementation of the Single-Cell Regulatory Network Inference and Clustering (SCENIC) workflow. The pipeline infers GRNs from scRNA-seq data in three main steps. First, pySCENIC constructs clusters of co-expressed gene modules across cells, typically using correlation-based approaches like the GRNBoost2 algorithm, a tree-based method that infers gene co-expression relationships (*371*). Second, it identifies transcription factors (TFs) that regulate these modules by scanning for enriched TF binding motifs in the upstream regulatory regions of the genes within each module, typically 500-1000 base pairs upstream of the gene. This motif enrichment analysis predicts which TFs are likely to regulate the genes in each module. Finally, pySCENIC scores the activity of these TFs (regulons) in individual cells using AUCell (Area Under the Curve

for Cells). This step measures the activity of the regulons across all cells, enabling the identification of unique regulatory landscapes amongst populations of cells (*371*, *372*).

However, like most correlation-based approaches, pySCENIC presumes that co-expression of a TF and its target genes indicates a direct regulatory relationship, which can overlook indirect or non-causal interactions. Furthermore, it does not explicitly incorporate post-transcriptional regulation, chromatin constraints, or combinatorial control by multiple TFs. As a result, pySCENIC can produce false-positive associations if genes correlate for reasons unrelated to TF binding. Motif enrichment also relies on the assumption that annotated cis-regulatory elements are actively engaged in the cell type of interest, which may not hold if epigenetic states vary or if critical enhancers lie far from the gene body. One way to address these limitations is to integrate or cross-validate pySCENIC's predictions with epigenomic data. For example, as previously discussed in Section 1.1, scATAC-seq can directly profile open TN5 accessible chromatin regions, where TF binding is more likely to occur. Confirming that putative binding motifs inferred by pySCENIC align with accessible regions identified via scATAC-seq, can reduce false positives and better pinpoint functionally relevant TF target interactions. Moreover, combining these two omics layers helps disambiguate whether certain genes are merely co-expressed or if they have a genuine regulatory relationship reinforced by accessible binding sites in promoter or enhancer regions, thereby refining regulatory inference hypotheses by revealing physical accessibilities of proposed regulatory elements.

1.3.11 Cell-Cell communication inference

Cell-cell communication inference tools analyse scRNA-seq data to predict potential protein-protein interactions between different cell types, focusing on ligand-receptor pairs and downstream signalling pathways. Two widely used tools in this field are CellPhoneDB and NicheNet. CellPhoneDB is a curated database and analytical framework that predicts cell-cell interactions based on the expression of ligand-receptor pairs (*373*, *374*). It contains a comprehensive database of experimentally validated, and predicted ligand-receptor pairs, curated from literature and various databases. Using permutation testing, CellPhoneDB assesses the significance of interactions between cell types, highlighting pairs that are significantly expressed in those putatively communicating (*373*).

NicheNet extends the functionality of tools like CellPhoneDB by incorporating downstream signalling and regulatory network information. It starts with identifying potential ligand-receptor pairs, similar to CellPhoneDB. NicheNet then incorporates signalling pathways and transcriptional regulation, predicting the effect of ligand-receptor interactions on target gene expression. By integrating GRNs, NicheNet can predict the impact of cell-cell communication on recipient cells, offering a more comprehensive view of how intercellular signalling influences the transcriptional states of recipient cells (*374*).

However, while effective at inferring axes of cell-cell communication and thus generating hypotheses, several caveats must be acknowledged when using tools such as CellPhoneDB and NicheNet. First, these methods rely on known or predicted ligand-receptor pairs and assume that mRNA abundance accurately reflects surface protein expression and functional signalling. In reality, post-translational modifications, subcellular localization, and spatiotemporal dynamics often diverge significantly from transcriptomic readouts, and interactions requiring multi-protein complexes, proteolytic cleavage, or specific receptor subunits may be overlooked.

Second, the simultaneous detection of a ligand in one population and its cognate receptor in another does not by itself demonstrate that the recipient cells mount a response. Current methods offer only partial solutions to this problem. Methods such as NicheNet attempt to infer downstream activation by scoring the enrichment of ligand-linked target-gene programmes, but these predictions depend on curated regulatory networks and assume that transcriptional changes are a direct consequence of the proposed ligand. Rapid post-translational events, transient phosphorylation cascades, or non-transcriptional responses are therefore missed. Multimodal approaches yield intermediate evidence that, while still correlative, provide greater biological resolution. For instance, CITE-seq provides a more direct readout through the detection of receptor proteins on the cell surface, strengthening the inference beyond mRNA co-expression alone. Observing a correlation between high ligand expression in sender cells and elevated levels of both the receptor protein and downstream pathway gene/protein signatures in receiver cells can build a stronger, albeit still circumstantial, case for a functional interaction within the sampled tissue. Nevertheless, such evidence is correlational, dependent on the quality of antibody panels for CITE-seq, and cannot definitively establish causality or rule out confounding factors such as non-functional co-expression or convergent signalling pathways (*374*).

To establish functional consequences of ligand receptor interactions, inferred hypotheses can be tested with orthogonal experimental strategies. Methodologies applicable directly to ex vivo tissues are often more accessible than in vivo models and allow for validation in a cellular context that more closely approximates the native state than simplified in vitro cultures. For example, mass cytometry–based Phospho-CyTOF enables the quantification of phosphorylation states of key signalling proteins in cells from freshly isolated tissue, providing a snapshot of signal transduction in situ. Its utility, however, is limited by the availability of validated antibodies and potential artefacts from epitope masking during fixation (*70*, *375*). In contrast, investigating causal links between specific signalling components and their downstream effects often involves perturbation-based approaches, which typically necessitate a shift to in vitro models. Technologies like Perturb-seq, which couple CRISPR-based gene knockouts with scRNA-seq, can directly link genetic modifications to their transcriptional consequences (*376*). For instance, demonstrating that the deletion of a receptor substantially attenuates a downstream gene programme provides evidence for its functional role in that signalling cascade. By incorporating a time-course element into the experimental design, such perturbation studies can be further extended to dissect the temporal dynamics of the signalling response. However, these experiments are typically performed within the simplified context of a cell culture or cell line system that may not fully recapitulate the native tissue microenvironment (*376–378*). In summary, integrating large-scale inference tools such as CellPhoneDB and NicheNet with targeted, hypothesis-driven functional assays yields a robust, mechanism-driven strategy for validating predicted cell-cell communication networks. Nevertheless, this approach is constrained by the accuracy of computational predictions, the specificity of assay reagents, and the reductionist nature of in vitro systems. Therefore, iterative testing in more physiologically relevant contexts, such as organoids or in vivo models, is critical to fully corroborate and refine inferred signalling interactions.

1.4 Thesis structure

This thesis is thematically organised around the involvement of the human embryonic yolk sac (YS) within various broad developmental roles, contextualised against the backdrop of comparative and functional organ systems across developmental time and species. The following section is a Methods chapter where I elaborate on the different methodological frameworks I have applied, including a brief section on optimisation strategies for different

bioinformatic tools used in transfer learning and integration. This chapter details the methodological frameworks for various bioinformatic tools employed in this thesis, including techniques for scRNA-seq, CITE-seq, and multi-omic data integration. The chapter also covers the application of transfer learning to enhance the accuracy of cell type annotation and the integration of diverse datasets to create comprehensive cellular atlases.

In Results Chapter 1, I describe the unbiased and exploratory approach applied using different single-cell, spatial, and multi-omic technologies to decode early human YS haematopoiesis. Results Chapter 2 focuses on the vital, evolutionarily conserved functional roles the human YS plays during development. This chapter highlights how these roles support embryo survival, including nutrient transfer, early haematopoiesis, and the establishment of the initial immune system. In Results Chapter 3, I explore the YS's roles in the developmental diversification, functions, and origins of macrophages across human life. This chapter expands on the life-long impact of YS-derived macrophages on tissue-resident populations in different tissues. The integration of single-cell and spatial data has revealed distinct developmental trajectories and functional specialisations of macrophages originating from the YS. The thesis concludes with a final discussion chapter that synthesises the findings from the results chapters. This discussion integrates the insights gained from the study of the human embryonic YS and its roles in haematopoiesis, immune development, and tissue homeostasis. The chapter also considers the broader implications of these findings for understanding human development and disease.

2 Materials and methods

In this chapter I describe the methods towards generating, analysing, and interpreting single-cell, spatial, and multi-omic atlas of the developing human YS.

The work outlined in this chapter is the result of a multinational, multisite collaboration between the Haniffa group led by Professor Muzlifah Haniffa (Biosciences Institute, Newcastle University and the Wellcome Sanger Institute), members of the Wellcome sanger institute, and members of Sorbonne Université.

This method chapter is a lightly-edited version of the manuscript which I co-first authored (*1*), including additional details and optimisation parameters for specific computational methods.

Below I delineate and highlight respective work which I either assisted in, or did not personally conduct.

Sample acquisition methods, wet-lab protocols, including antibody panels and FACS were adapted from work by Dr Rachel Botting, Dr Emily Stephenson, and Dr Laura Jardine. RNA-scope methods and protocols were adapted from work by Nana-Jane Chipampe and Kwasi Kwaka (Wellcome Sanger Institute). Light-sheet microscopy methods were adapted from work by Dr Yorick Gitton, and Megumi Inoue, led by Professor Alain Chédotal (Sorbonne Université, INSERM, CNRS, Institut de la Vision). Antony Rose assisted in the integration and annotation of CITE-seq data using *TotalVI*. I performed all other analyses independently and wrote all computational sections in this method chapter.

2.1 Data generation

2.1.1 Ethics and sample acquisition

Tissues were obtained from the MRC–Wellcome Trust-funded Human Developmental Biology Resource (HDBR; http://www.hdbr.org) with appropriate written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (18/NE/0290). HDBR is regulated by the UK Human Tissue Authority (HTA; www.hta.gov.uk) and operates in accordance with the relevant HTA Codes of Practice. Tissues used for light-sheet fluorescence microscopy were obtained through INSERM's HuDeCA Biobank and made available in accordance with the French bylaw (Good practice concerning the conservation, transformation and transportation of human tissue to be used therapeutically, published on December 29, 1998). Permission to use human tissues was obtained from the French agency for biomedical research (Agence de la Biomédecine, Saint-Denis La Plaine, France).

2.1.2 Fetal developmental stage assignment

Embryos were staged using the Carnegie staging method (*379*). A piece of skin or chorionic villi tissue was collected from each sample to perform quantitative PCR karyotyping of sex chromosomes and autosomal chromosomes 13, 15, 16, 18, 21, and 22 for the most commonly seen chromosomal abnormalities. No abnormalities were detected.

2.1.3 Processing samples for imaging and single-cell sequencing

Tissues were transported in phosphate-buffered saline (PBS) on ice, were dissected within 24 hours, and were processed immediately (<1 hour after dissection). For formalin-fixation and paraffin-embedding, samples were immediately placed in 10% (w/v) formalin. Processing and embedding were performed by NovoPath, Newcastle upon Tyne NHS Trust. For RNAscope, samples were snap-frozen in an isopentane bath in liquid nitrogen prior to embedding in optimal cutting temperature (OCT) compound. Single-cell suspensions were generated by dicing tissue into segments <1 mm^3, followed by enzymatic digestion for 30 min at 37°C with intermittent shaking. Digestion media was 1.6 mg/ml collagenase type IV (Worthington) in RPMI (Sigma-Aldrich) supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS; Gibco), 100 U/ml of penicillin (Sigma-Aldrich), 0.1 mg/ml of streptomycin

(Sigma-Aldrich), and 2 mM L-glutamine (Sigma-Aldrich). Digested tissue was passed through a 100-μm filter, and cells were collected by centrifugation (500*g* for 5 min at 4°C). Cells were treated with 1X RBC lysis buffer (eBioscience) for 5 min at room temperature and washed once with Flow Buffer (PBS containing 5% (v/v) FBS and 2 mM EDTA) before counting. Processing for scRNA-seq was continued promptly on fresh cells, for other uses (including CITE-seq) cells were collected by centrifugation (500*g* for 5 min at 4°C) and resuspended in 10% (v/v) DMSO in FBS for freezing. For light-sheet fluorescence microscopy (LSFM), tissues were fixed in 4% PFA and dissected. Gestational age was then estimated as previously described (*380*).

2.1.4 Processing of single-cell suspensions for scRNA-seq

Immediately following isolation and counting, cells were collected by centrifugation (500g for 5 min at 4°C) and resuspended in a residual buffer. Three microliters of CD45 BUV395 (clone: HI30, BD Biosciences) was added to the resuspended cells and incubated on ice in the dark for 30 min, washed with Flow Buffer and resuspended at ~1×10^7 cells/ml. Immediately prior to sorting, cells were passed through a 35-μm filter (Falcon) and DAPI (Sigma-Aldrich) was added at a final concentration of 3 μM. Flow sorting was performed on a BD FACSAria Fusion instrument using DIVA v.8, and data were analyzed using FlowJo (v.10.4.1, BD Biosciences). Cells were gated to exclude dead cells and doublets, and then isolated for scRNA-seq analysis (droplet-based 10x Genomics, or plate-based Smart-seq2) using a 100-μm nozzle. For droplet-based scRNA-seq, $CD45^+$ and $CD45^-$ cells were sorted into separate chilled fluorescence-activated cell sorting (FACS) tubes coated with FBS and prefilled with 500 μl of sterile PBS. For plate-based scRNA-seq, $CD45^-AF^+SSC^{++}$ single

cells were index-sorted into 96-well LoBind plates (Eppendorf) containing 10 µl of lysis buffer (TCL (Qiagen) + 1% (v/v) β-mercaptoethanol) per well.

2.1.5 Library preparation and sequencing of scRNA-seq and CITE-seq samples

For the droplet-based scRNA-seq experiments, cell suspensions isolated by FACS were counted and loaded onto the 10X Genomics Chromium Controller to achieve a maximum yield of 10,000 cells per reaction. 5′ V1 kits were used and sequencing libraries were generated according to the manufacturer's protocols. Libraries were sequenced using either an Illumina HiSeq 4000 or NovaSeq 6000 to generate at least 50,000 raw reads per cell.

For the plate-based scRNA-seq experiments, the frozen cell lysates were thawed on ice for 1 min. Purified cDNA was generated and amplified using a modified Smart-seq 2 protocol described in Villani et al (*19*). Sequencing libraries were then generated using Illumina Nextera XT kits with v2 index sets A, B, C and D. 384 cells were pooled and were sequenced using a HiSeq 4000 to generate at least 1×10^6 raw reads per cell.

For the CITE-seq experiments, frozen cells were thawed, counted, and pooled. Fc blocking reagent (Biolegend) was added to the cell pools and left to incubate at room temperature for 10 min. 500nL of CD34 APC/Cy-7 (clone: 581, Biolegend) was then added to the Fc-blocked cells and left to incubate in the dark and on ice for 10 min. Concurrently, the pre-titrated, pre-optimized TotalSeqA Cocktail (Human Universal Cocktail v1.0; BioLegend) (Fig. A10) containing 154 antibodies targeting immune cell surface markers (*73*) was centrifuged at 14,000*g* for 1 min (*381*). Flow buffer was then added to reconstitute before incubating for 5 min at room temperature. The resuspended antibody cocktail was then centrifuged at 14,000*g* for a further 10 min before adding to the cells. The cells and CITE-seq antibody cocktail were then left to incubate for 30 min in the dark and on ice. After this time, the cells were washed

twice with Flow buffer and resuspended in a final concentration of 50 μg/ml of 7 AAD (Thermo Fisher Scientific) in Flow buffer.

Live, single cells or live, single $CD34^-$ cells and live, single $CD34^+$ cells (for the CITE-seq experiments) were then isolated by FACS into 500 μl of PBS in FACS tubes coated with FBS. Cells were then counted and submitted to the CRUK CI Genomics Core Facility for subsequent processing using 10x Genomics protocols and sequencing. Single-cell gene expression and cell surface protein libraries were generated using Single cell 3′ v3 kits according to the manufacturer's protocol. Libraries were sequenced using a NovaSeq 6000 to achieve a minimum of 20,000 reads per cell for gene expression and 5000 reads per cell for cell-surface protein.

2.2 Data pre-processing

2.2.1 Alignment, quality control, filtering, and preprocessing of scRNA-seq and CITE-seq data

scRNA-seq expression data (including droplet-based and plate-based) were mapped with CellRanger (version 3.0.2) to a human reference genome and low-quality cells expressing <2000 reads, <200 genes, and >20% mitochondrial reads were filtered out of the data based on heuristic assessment of QC scatter plots via *sc.pp.calculate_qc_metrics* and *sc.pl.scatter* in Scanpy (*363*) (v1.9.0). Genes expressed in fewer than three cells were also removed.

2.2.2 Doublet detection and background noise reduction

For droplet-based scRNA-seq data, the following additional QC steps were performed. Scrublet (*308*) v0.2.3 was applied to each sequencing lane for doublet detection, and clusters with >(Median+(1.48*MAD)) (MAD: Median absolute deviation) of the median cluster doublet detection score were removed. Ambient RNA was removed with Cellbender (v0.2.0) with fdr=0.01 and epochs=150 (*382*). To determine likelihood of maternal contamination, data were pooled by donor and submitted to Souporcell (v2.4.0) at genotype clusters c=1 and c=2 models to represent likelihood of no maternal contamination and possible maternal contamination, respectively. The optimal model was identified via BIC (Bayesian Information Criterion), where we observed a smaller BIC index at c=2 in one donor (F37, Female, 5 post conception weeks (PCW)). Cells from the F37 alternate genotype were identified as potential maternal contaminants, composed mainly of monocytes (n=149), and monocyte–macrophage intermediates (mono–mac int.) (n=25), and excluded from downstream analysis.

For CITE-seq data, FASTQ alignment was performed for multiplexed RNA lanes with CellRanger (v4.0.0) and GRCh38-2020-A reference genome, and for multiplexed protein lanes with CITE-seq-Count (v1.4.3). Lanes with cells pooled from multiple donors were deconvoluted using Souporcell singularity image at https://github.com/wheaton5/souporcell. Low quality cells expressing <200 genes and >20% mitochondrial reads were removed and doublets were removed by applying Scrublet v0.2.3 to each sequencing lane and then removing clusters with >(Median+(1.48*MAD)) of the median cluster doublet detection score. CITE-seq protein data underwent QC and preprocessing as previously described (*3*), (i.e., cells were first filtered to intersect barcodes with counterpart CITE-seq RNA data, then unmapped antibodies were filtered out and then protein cells were filtered for low quality by cells with <30 proteins and expressing >5000 reads).

scRNA-seq count matrix transformation, normalization, and preprocessing were performed using Scanpy (*363*) in python (v3.8.6). We normalized raw gene counts using the *sc.pp.normalize_total* function (*target_sum = 10e4*) from and performed *ln(x+1)* transformation. Reported expression values were normalized, log-transformed, and mean-centred and variance scaled using the *sc.pp.scale* function independently for each analysis.

For CITE-seq data count matrix transformation, we first performed denoised and scaled by background (DSB)-normalization originally developed by Mulè et al (2022) (*383*) and applied a Gaussian Mixture Model (GMM) for background non-specific binding signal regression per sample (*383*). For the first step, a modified DSB-normalization approach used in our previous study (*3*) was constructed. For each CITE-seq lane, low quality/empty droplets were identified as droplets under the largest UMI peak which had a value <1.96*standard deviations (std) of the mean UMI counts value (mu_UMI) per sample. In this approach, we

define the largest UMI peak as the most frequent low-UMI count peak, rather than the highest UMI count value under the assumption that empty droplets contain lower UMI counts. Peak detection was conducted using the *scipy.signal.find_peaks* function. The number of peak detection bins were dynamically estimated as (3.322*log(X)), where X was the total number of droplets. The factor 3.322 is derived as (1/log10(2)). This makes (3.322 * log10(X)) mathematically equivalent to Sturges' rule (k = 1 + log2(X)), which is a widely used heuristic for dynamically determining the number of histogram bins based on sample size (*384*). The model iterated through a series of 20 prominence intervals (0-20) with widths (0-10) where peaks detected < (mu_UMI-(1.96*std)) were retained as empty droplet peaks. In cases where no empty droplet peaks were detected, the empty droplet threshold was taken to be <(mu_UMI-(1.96*std)). The estimated empty droplets matrix was then taken into downstream DSB normalization (*383*) in the same way as our previous study (*3*). While this approach successfully captures ambient antibody noise (*383*), it assumes that most droplets are empty and that the ambient signal follows a distinct low-UMI distribution. Deviations from these assumptions, such as unusual loading concentrations, excessive ambient antibody carryover, or atypical droplet encapsulation efficiencies, may shift the distribution and limit the accuracy of empty-droplet identification. For the second step of CITE-seq matrix transformation, we trained a GMM to model the variance of protein expression levels in each cell. We used the sckitlearn (v.1.1.3) *sklearn.mixture.GaussianMixture* module to fit 20 models with an increasing number of cell clusters *k* (between *c*=2 and *c*=21) to represent expression patterns of each protein by cell. The optimal model was identified using BIC ($BIC_i = 2L_i + k_i \log n$) and AIC (Akaike information criterion) ($AIC_i = 2L_i + 2k_i$) metrics where *k* is the number of GMM cell protein expression clusters, n is the number of cells in the sample and L is the model log likelihood. Lower values of AIC and BIC indicate better model performance, balancing model fit and complexity. In cases where two or more models had equivalent likelihood values, the model with fewer parameters (smaller k) was selected, prioritizing parsimony. The mean

expression values of GMM clusters with lowest expression values from each GMM model were interpreted as mean background expression per protein. Background-signal regression was performed separately for each protein using a Generalized Linear Model (GLM), implemented with the *statsmodels (v0.13.5)* library. Specifically, the GLM modeled observed protein expression levels per cell (dependent variable) as a function of standardized per-cell background scores (independent variable), explicitly estimating and removing the ambient antibody-derived technical component as: Protein Expression (per cell) ~ Background Score (per cell). Per-cell background scores were defined by taking the exponent of each protein background mean ($m\mu\_bg$), dividing it by the exponent of the protein expression in each cell (x), and then scaling the resulting ratio to a 0–1 distribution by subtracting the minimum score and dividing by the maximum score, where score is the ratio $\exp(m\mu\_bg) / \exp(x)$. Background scores inversely correlate with the magnitude of background expression per cell. (Background Score = (score - $\min(e^{mu\_bg}/e^{x})$)/$\max(e^{mu\_bg}/e^{x})$). The per-cell background signal regressed counts were used for subsequent analyses, interpretation and visualization (Fig. 2.1). Cells comprising the empty droplet matrix were removed and were not considered for downstream analyses. Post-DSB-GMM, assessment was performed by qualitatively inspecting the expression of known marker proteins in each cell population before and after correction. For each population, the proportion of cells expressing proteins not expected to mark that population was quantified, and the resulting decrease in this false-positive fraction was used to indicate background removal, while verifying that canonical marker protein signals remained stable. However, this qualitative, marker-based evaluation is not sufficiently robust as it lacks quantitative metrics for reproducibility, sensitivity to subtle expression changes, and an objective scale for comparing signal retention versus background removal. In future work, denoising efficacy will be quantified for each marker–population pair by calculating pre- and post-normalization mean expression levels, computing their ratio, and measuring the change in the fraction of cells expressing marker proteins. This objective

assessment will simultaneously capture signal preservation of known populations, and background removal (Fig. 2.1).

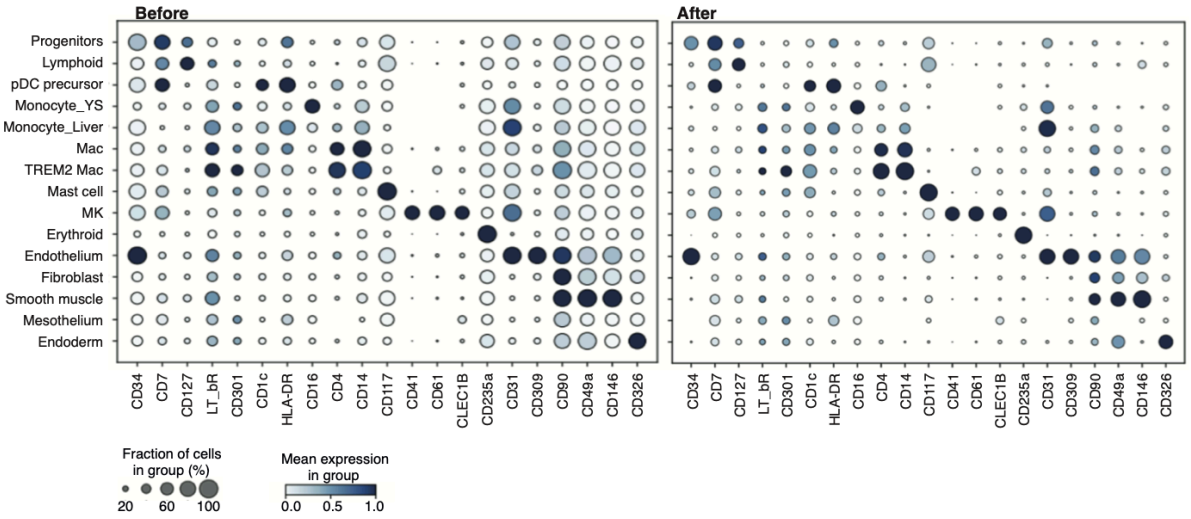


Figure 2.1: DSB normalisation with GMM test: *Dot plots illustrating the mean expression (color scale) and the fraction of cells expressing each protein (dot size) of curated marker proteins for populations using raw YS CITE-seq data (left) and after DSB-GMM normalisation (383) removal (right). Adapted from Goh and Botting et al, 2023 (1).*

2.3 Data analysis and robustness comparisons

2.3.1 Integration and batch correction of scRNA-seq and CITE-seq datasets

For integration of newly generated yolk sac (YS) scRNA-seq data with external datasets, CellRanger count was first reapplied for the alignment of CS10/CS11 and CS14 embryonic YS scRNA-seq data previously acquired (*46*, *385*). The following steps were then followed for the total integrated YS droplet-based scRNA-seq dataset. Highly variable gene (HVG) selection was performed using the *sc.pp.highly_variable_genes* function (min_mean=0.001, max_mean=10) based on normalized dispersion and used for downstream embedding. Dimensionality reduction and batch correction was carried out using the scVI module within *scvi-tools* (v0.19.0) (*330*, *386*) (HVG = 7500, dropout_rate=0.2, n_layer=2) with biological replicate taken as the technical covariate. To ensure model performance was optimal for each independent analysis, scVI was benchmarked against the python implementation of Harmony (*350*) (*Harmonypy* v0.0.5) (Fig. 2.2) at various theta values between 1 and 20. kBET (*387*) and Silhouette scores (*sklearn.metric.sil_score*) were computed for each iteration between donor covariates and compared to the scVI integration. For integration of adult scRNA-seq data, publicly available single-cell and single-nuclei RNA-seq data of 20 healthy adult tissues were integrated using scVI (HVG=1500, layers=1). Batch correction was conducted on donors, single cell or single nuclei, data source, number of genes, total counts, percentage of mitochondrial genes and ribosomal genes. Please see Fig. A9 for information regarding external single-cell RNA sequencing (scRNA-seq) datasets that have been incorporated and integrated in this study.

For multimodal integration of CITE-seq datasets, we integrated both RNA and protein modalities using the *totalVI* module in *scvi-tools* (v0.19.0). Cells lacking either RNA or protein measurements were excluded prior to totalVI integration. Although *totalVI* can accommodate unimodal query datasets by imputing them in the joint latent space (*330*, *386*,

388), only intersecting cells were retained here to ensure that each cell in the CITE-seq analysis had directly measured surface-protein and scRNA-seq profiles for robust validation of population-specific protein markers. We performed *sc.pp.highly_variable_genes* function on RNA modality (HVG=4000) accounting for FACS sampling and donor technical covariates. We then generated a multi-modal totalVI VAE latent representation following totalVI pipeline (*330*). To compare bioconservation between Totalvi and scVI, the data was first annotated using an elastic-net classifier trained on a joint scVI embedding of the yolk sac scRNA-seq atlas and CITE-seq scRNA counts (described section 2.3.2). For benchmarking, a separate scVI model was independently trained on CITE-seq scRNA alone to derive an RNA-only latent embedding. Leiden clusters derived from both embeddings (TotalVI and scVI) were annotated by majority vote of those per-cell labels, with manual refinement via marker differential expression (described section 2.3.2), and cluster separation was quantified by the global silhouette index. In this comparison, totalVI achieved a marginally higher global silhouette index (0.347 vs. 0.344 for scVI), however, the 0.003 difference falls within the calculated standard errors (±0.013 for totalVI; ±0.016 for scVI). Consequently, this result does not constitute significant evidence that totalVI produces better-defined cluster boundaries. TotalVI was chosen because its multimodal embedding integrates both transcriptomic and proteomic information, enabling direct validation of surface-protein markers against scRNA-seq–defined populations. Future analyses will report bootstrap-derived 95 percent confidence intervals and perform paired tests on per-cell silhouette scores to rigorously assess any performance differences (Fig. 2.3).

All scVI VAE and ldVAE models trained on the YS and integrated atlases are available on our data portal (see data availability) and will facilitate transfer learning for future reference mapping of scRNA-seq data with single cell architectural surgery (scArches) (*18*)

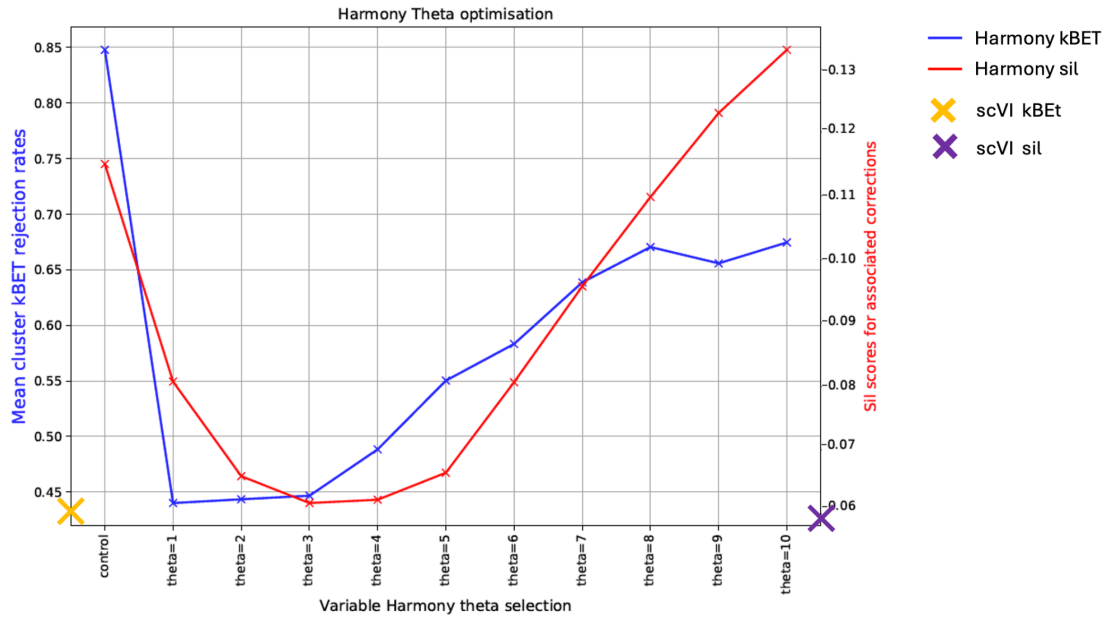


Figure 2.2: Effect of differential Harmony theta values on measured kBET and Sil scores. *We observe that harmony theta = 3 displays the optimum kBET vs Sil score in our data. kBET and Sil scores were computed from 50 PCs for each cell state and the mean value per theta variable taken. Sil scores used euclidean distances between batch covariates as the metric for measurement. Yellow X mark indicates the mean cluster kBet score from scVI embeddings, and purple X mark indicates the mean cluster Sil score for the same.*

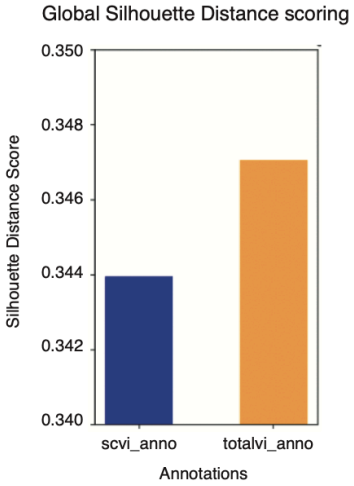


Figure 2.3: TotalVI compared against scVI in CITE-seq data: *Bar chart displaying global silhouette distance scores on cluster-derived annotations between SCVI and totalVI annotations. This figure is adapted from Goh and Botting et al, 2023 (1).*

2.3.2 Clustering and annotation of scRNA-seq and CITE-seq data

Clustering of scRNA-seq and CITE-seq datasets was performed using the Leiden algorithm (*361*) (*sc.tl.leiden*) with a resolution parameter of *res*=1.5 (CITE-seq res=3) on a k-nearest neighborhood graph (k=30 for scRNA-seq and k=15 for CITE-seq) unless specified otherwise. To assess the degree of homogeneity of different populations and the effect of graph complexity on population annotations, for instance iPSC-derived macrophages and endoderm fractions, we computed specificity metrics (Adjusted Rand index (ADJ Rand) and Adjusted mutual information (MI) scores) (Fig. 2.4, A and B (Left)) and population homogeneity metrics (Within-Cluster-Sum of Squared Errors (WSS) and Silhouette index (SI)) (Fig. 2.4, A and B (Right)) across decreasing KNN graph complexity (K number of nearest neighbours = 5-50). We assessed the degree of population homogeneity by performing the following analyses: We first confirmed that our clustering approach was robust and the Macrophage annotation in iPSC data and endoderm in YS data was stable across different levels of K-nearest neighbour (KNN) graph complexity. We iterated K (number of nearest neighbours) in KNN across intervals of decreasing complexity between 5-50 nearest neighbours (based on 30 latent SCVI VAE components). We then performed community detection using leiden clustering (res = 5) and UMAP on each K interval model. Clusters were labelled using majority voting. Adjusted Rand index (ADJ Rand) and Adjusted mutual information (MI) scores were computed between new clusters and original annotations. Cluster MI and Rand scores indicate how the complexity of KNN models affect mixing (specificity) of annotations. ADJ Rand and MI scores were generated globally and for newly derived Macrophage/Endoderm clusters respectively. (Fig. 2.4 A and B (Left)). To model the degree of homogeneity across the iPSC macrophages and YS endoderm, we computed Within-Cluster-Sum of Squared Errors (WSS) and Silhouette index (SI) on UMAP embeddings across K intervals within all clusters (global) and Macrophage/Endoderm

populations. (Fig. 2.4, A and B)(Right). WSS indicates the variation within each cluster (homogeneity) whilst SI measures how similar a cluster is to itself compared to other clusters (cohesion). In these analyses, we assume that the graph-based Leiden clustering accurately captures local transcriptional similarities. Varying k between 5 to 50 tests how robustly cells cluster at different neighborhood scales. The majority voting label assignment assumes each cluster is relatively homogeneous in identity. If multiple rare cell types are pooled in one cluster, the majority label would overshadow them and thus result in lower MI and ADJ Rand scores when compared to the original labels.

We also assume that loss of distinct cell-type clusters would negatively impact silhouette scores (driving them down) and WSS scores (driving them up) calculated on the original labels in UMAP embeddings, indicating an increasing overlap between distinct communities in the resultant embedding space. However, as UMAP is a non-linear dimensionality reduction technique, it may distort certain nearest-neighbour relationships. WSS was employed on the UMAP coordinates primarily as a comparative, heuristic measure of cluster tightness in the 2D embedding. By consistently applying the same embedding approach and parameter settings across all K intervals, we assume that changes in WSS reflect relative shifts in cluster compactness. Additionally, direct measurements in embedding space could be used to confirm these results, mitigating the risk of misinterpretation due to UMAP's non-linear transformations.

In cases where datasets are compared probabilistically, or where new classifications have been made, an implementation of low-dimensional ElasticNet regression (EN) (described in the "Cell state predictions using probabilistic low-dimensional ElasticNet regression" section of the manuscript methods) was used to first classify individual cells where a model-specific decision threshold of 0.9 was used for classification tasks. Cells with a predicted probability above 0.9 were assigned the cell‑state labels output by the EN classifier trained on the YS

scRNA-seq atlas. Clusters were then assigned a class if the count of the most frequent projected label exceeded (mean + (1 × std)) of the predicted label counts in that cluster (where "mean" and "std" are calculated across counts of all labels within the cluster). Otherwise, if no representative label could be defined by this majority voting rule, the cluster was flagged for manual review. Resultant cell state classifications were further manually checked using differentially expressed genes using the *sc.tl.rank_genes_groups* function in Scanpy which performed a two-sided Wilcoxon rank-sum test for genes expressed in >25% of cells, with a log-transformed fold change cut-off of 0.25. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Annotation of YS and liver CITE-seq data was performed by training an EN model using YS and Embryonic Liver (EL) scRNA-seq datasets as references respectively. These labels were then distributed by majority voting onto Leiden clusters derived from CITE-seq data. The resultant cluster annotations were validated using the same markers identified in matched RNA data and underwent additional manual annotation where required. For differential expression testing of surface proteins in multi-modal CITE-seq data, we conducted a one_vs_all DE test using the *vae.differential_expression module* within totalVI (Bayes factor>3, Median LFC>0.25). Marker proteins and corresponding populations were subsequently grouped by hierarchical clustering using complete‑linkage agglomerative clustering on distances defined by Pearson correlation of their median expression profiles across clusters, as implemented in the *sc.tl.dendrogram* functionality within Scanpy. Differential expression testing between cell states in the integrated 12 fetal organ atlas was carried out on using a *scVI* integrated latent VAE representation with a one_vs_all DE test using the *vae.differential_expression module* within scVI (v0.19.0) (Bayes factor>3, Median LFC >4). Bayesian differential expression testing between cell-states in the integrated 12 fetal organ atlas was carried out using the *vae.differential_expression* module within the *scVI-tools* package. This method leverages repeated sampling from the posterior variational distribution to estimate variability and effect

sizes robustly, providing Bayesian inference of state-specific gene expression differences. Significant differential expression was defined using a Bayes Factor >3 and a median log-fold change (LFC) >4.

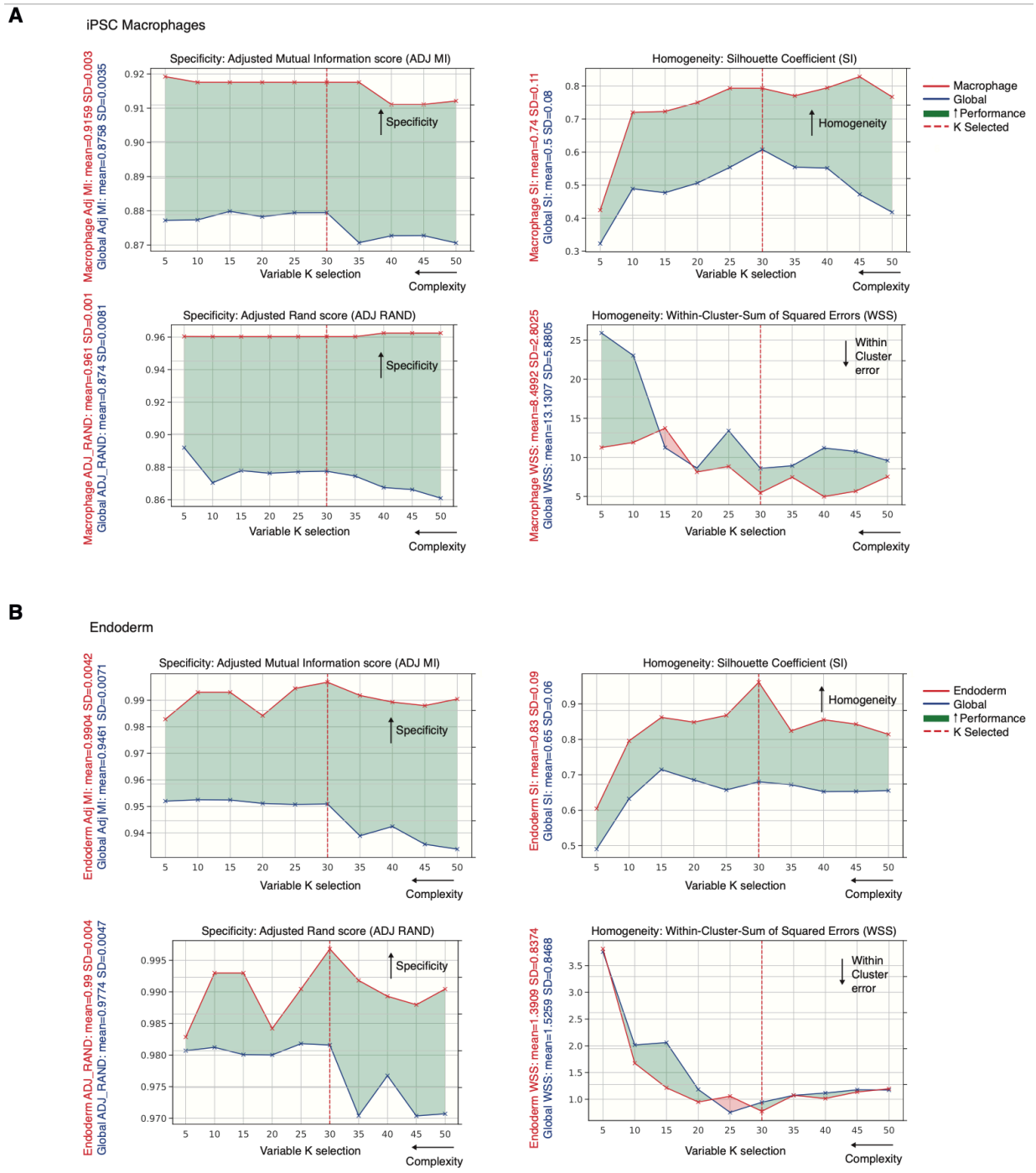


Figure 2.4: kNN graph complexity effect on population heterogeneity: *(A) Left: Metrics of iPSC-derived macrophage population specificity, adjusted mutual information (MI) score (top) and adjusted rand (ADJ RAND) (bottom) of iPSC-derived macrophage population across decreasing graph complexity compared against all other populations. Right: Metrics of population homogeneity, silhouette coefficient index (SI) (top) and within squared sum error (WSS) (bottom) of iPSC-derived macrophage population across decreasing graph complexity . (B) Plots as described in A for the endoderm population compared against all other populations in the YS scRNA-seq data. This figure is adapted from Goh and Botting et al, 2023 (1).*

2.3.3 Dimensionality reduction and marker expression visualisation

For visualisation, the uniform manifold approximation (UMAP) algorithm was run using the *sc.tl.umap* function in Scanpy. Dot-plots and violin plots were produced in Scanpy and all gene expression values displayed were normalised, log-transformed, and scaled as described in the preprocessing section unless otherwise stated. Dot plots that display data from multiple datasets employ independent log-normalisation, variance scaling, and min–max standardisation to a distribution of 0-1 per dataset unless otherwise stated. Force-directed graphs (FDGs) computed with the *sc.tl.draw_graphs* function in Scanpy using the Force Atlas2 parameter were used to infer trajectories. Partition-based graph abstraction (PAGA) were computed on the k-nearest neighbour graphs and overlaid onto FDGs where nodes represented the centroid of each cell state cluster and the thickness of edges represented the similarity between cell states. FDG embeddings assume a continuous manifold and can obscure global relationships if local neighborhoods dominate. Likewise, PAGA presupposes that k-nearest neighbors accurately capture transitions between related cell types, and if certain rare states are under-sampled, the resulting edges can be incomplete or misleading.

Consequently, we interpret all visualizations alongside marker-based cell identities to avoid misclassifying spatial proximity as true biological similarity.

Proportion line graphs for specific populations (e.g., erythroid cells) enriched in specific genes (e.g., *HBZ*) using the *sc.tl.score_genes* function in Scanpy were produced using Matplotlib (v3.6.2). To track temporal changes in the fraction of cells expressing each gene, we grouped cells from each population into discrete age bins. Within each age bin, we calculated the proportion of cells expressing detectable levels (>0 log-normalized counts) of the target gene. Enriched cells were defined as cells with a positive enrichment score using the *sc.tl.score_genes* function, whereby, for each cell, it subtracts the mean log-normalized expression of 200 reference genes (randomly sampled from 50 expression-level bins) from the average log-normalized expression of the target gene set. Proportions of enriched cells with enrichment score >0 in each cell type compartment were then plotted as a discrete time-series across gestational age to visualize differential enrichment of cells expressing the genes of interest. Data point sizes represented enriched cell counts. To aid interpretation, an ordinal scale of representative cell counts was included as a legend in the plots.

Proportions of specific populations (e.g., macrophages) enriched in specific gene modules (e.g., Pre-AGM module) were visualized using violin graphs produced using Matplotlib and *Seaborn* (v0.12.1) python libraries. To ensure background expression profiles were accounted for, we segregated our population of interest and computed changes in relative population proportion enriched in each gene module. Significant module enrichment was defined as described above. Enriched cells from each cell type compartment were then graphed across organs. Enrichment scores were standardized to the median by subtraction of the median and subsequent division by MAD.

2.3.4 Differential abundance testing and FACS correction

We tested for differential cell-state abundance across gestation using the *Milo* framework (*289*), correcting for CD45 positive and negative FACS isolation strategies using a previously published technique (*85*). Where FACS correction was applied, we calculated a FACS isolation correction factor for each sample *s* sorted with gate *i* as ($fs = \log(p_i S/S_i)$) where p_i is the true proportion of cells from gate *i* and S represents the total number of cells from both gates. The *milopy.core.make_nhoods* function (prop = 0.05) was applied to randomly sample and compute KNNs for 5% of cells in the data. Each sampled cell was then refined by selecting the cell closest to the neighborhood median position, defining neighborhoods as partially overlapping sets comprising each index cell and its *k* nearest neighbors. This approach yielded a compact yet comprehensive coverage of the manifold. Neighborhood labels were determined by majority voting of cell labels by frequency in each neighborhood (>50%). The YS scRNA-seq data was then split into five age bins (3PCW, 4PCW, 5PCW, 7PCW, and 8PCW) and cell counts were modeled as a negative binomial generalized linear model (NB-GLM) with Benjamini–Hochberg weighted correction as previously described (*85*). Significantly differentially abundant neighborhoods were detected by SpatialFDR– (<0.1, logFC <0) for early enriched neighborhoods and (<0.1, logFC >0) for late neighborhoods.

Beeswarm plots were generated using the *ggplot2* library (v3.4.2). Each node represents an independent neighborhood of cells derived from the KNN graph. The *x*-axis position of each node represents the fold-change (positive/negative) associated with the distribution of age groups present in each neighborhood where larger proportions of older groups in a given neighborhood encourages a positive fold change and vice versa. Colored nodes represent neighborhoods with significant enrichment (P<0.05 spatial FDR) and the intensity represents the degree of significance.

2.3.5 Clustered gene-set enrichment analysis

We used the *FindConservedMarkers* function in Seurat (v3.1) with Bonferroni-corrected FDR-adjusted *P*-values to identify markers consistently differentially expressed ($P < 0.05$) between paired cell-state comparisons, specifically between YS endoderm and EL hepatocyte populations, and between YS endoderm and mouse gastrulation-derived endoderm populations in the scRNA-seq data. Markers were submitted for gene set enrichment ranking and analysis using the Enrichr tool as implemented in the *GSEApy* (v1.0) package to query the Gene Ontology (GO) Biological Process database (GO_BP_2022). Using the enrichrR package (v3.0) in R, enrichment was first computed by Fisher exact test for randomly sampled genes to derive a mean rank and standard deviation to estimate background for each ontological term accessed. A *z*-score for deviation of each term to its background rank was then used to rank output genesets. We derived statistical significance (Fisher exact test <0.05, ranked by *z*-score) for each gene set enrichment and performed Markov clustering (MCL) using the MCL (v1.0) package in R to derive network neighborhoods based on geneset intersect. Gene set clusters were annotated using the AutoAnnotate function in the RCy3 (v2.16) package and clusters were ranked by the mean *z*-score of all gene sets within each cluster and manually curated based on biological significance. We used the Cytoscape software (v3.9.1) to visualize clusters.

As with most enrichment analyses, our approach relies on assumptions about curated databases, specifically that GO_BP_2022 accurately reflects true functional categories. Another critical assumption is that observed overlaps between differentially expressed genes (DEGs) and gene sets genuinely reflect biological relationships, rather than correlated gene expression unrelated to the functional pathways of interest. Additionally, multiple-testing corrections help mitigate false positives, though they do not fully address interdependencies

among related gene sets, which could inflate significance estimates for large gene clusters. Consequently, manual curation remains a necessary step in our analysis to ensure biological plausibility and relevance of the enriched functions.

2.3.6 Cell state predictions using probabilistic low-dimensional ElasticNet regression

Label transfer class assignments and median probability of class correspondence between gene expression matrices in single cell datasets were carried out using a logistic regression (LR) framework, as previously described (*3*), using a similar workflow to the *CellTypist* tool (*389*).

Raw scRNA-seq datasets being compared were first concatenated, normalized, and log-transformed, as described in preprocessing. HVG selection was performed (min_mean=0.001, max_mean=10) for embeddings by dispersion. HVG expression matrices were used as training inputs for models unless otherwise stated. For models trained in combined low-dimensional representations, linear VAE latent representations were computed using the *LDVAE* module within *scvi-tools* (hidden layers=256, dropout-rate=0.2, reconstruction-loss=negative binomial) with donor, dataset origin, and organ information taken as technical covariates. Where PCs were used as input for training, harmony batch-corrected PCs (c=100pcs) were used, using Harmonypy (v0.0.9) with technical covariates as described above. Harmony runs were iterated through theta=1:20 and resultant embeddings benchmarked using kBET and silhouette scores between technical covariates where a low kBET rejection rate and corresponding high silhouette score denoted the optimal theta parameter.

ElasticNet regression (EN) LR models were built utilising the "sklearn.linear_model.LogisticRegression" module in the sklearn package (v0.22). The

models were trained using either gene expression data or SCVI batch-corrected low-dimensional LDVAE representation of the training data with regularisation parameters (L1-ratio and alpha) tuned using the GridSearchCV function in sklearn (v1.1.3). The test grid was designed with five l1_ratio intervals (0, 0.2, 0.4, 0.6, 0.8, 1), five alpha (inverse of regularisation strength) intervals (0.2, 0.4, 0.6, 0.8, 1) at five train-test splits and three repeats for cross-validation. The unweighted mean over the weighted mean squared errors (MSEs) of each test fold (the cross-validated MSE) was used to determine the optimal model. This logistic regression approach assumes that each cell type can be discriminated by linear combinations of HVGs, scVI, or, LDVAE latent dimensions, and that the training data adequately represent all cell states in the validation set. If a novel or rare cell population exists in the target dataset that is absent in the reference, the classifier may assign them to the closest known state. Moreover, EN's penalty encourages sparsity but can over-penalize highly correlated features, potentially removing relevant genes in large multi-gene modules. Hence, manual verifications of known populations against known markers remains a crucial final check.

The resultant model was used to predict the probability of correspondence between trained labels and precomputed clusters in the target dataset. For dataset comparison tasks where predesignated labels already existed in the target dataset, the median probability of training label assignment per predesignated class was computed and visualised as a heatmap. We show how this method for label transfer compares to PAGA-based label transfer reported in our previous publication (*6*) (Fig. 2.5 A). Genes (or features) predicted to be significantly discriminatory in each LR model were evaluated using an "impact" score. We define "impact" for each gene (or latent dimension, if training is done on low-dimensional embeddings) as e^beta, where beta is the logistic regression coefficient. This exponentiated coefficient is interpreted as the odds ratio for that feature's contribution to classification. Features were

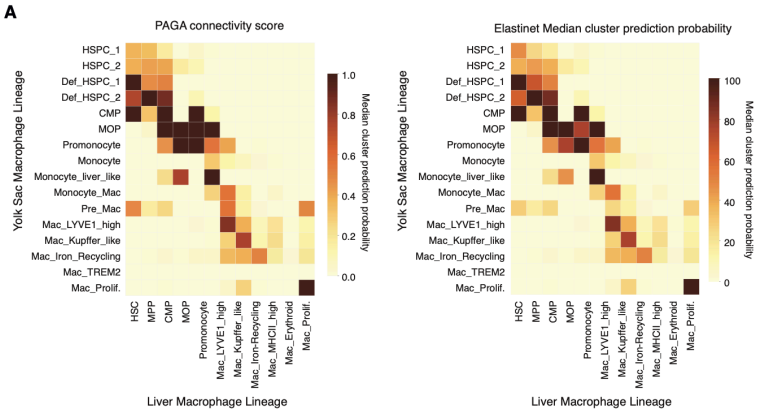
ranked by descending impact scores per gene (I_g). We then employed a rank-based approach to test statistical significance, the survival function (sf) for each gene's impact score was computed over all features, where sf(I_g) is the proportion of features with impact scores greater than or equal to I_g. The p-value for each feature was taken as sf(I_g), and a gene (or latent dimension) was considered significantly impactful at $p < 0.05$ (Fig. 2.5 B).

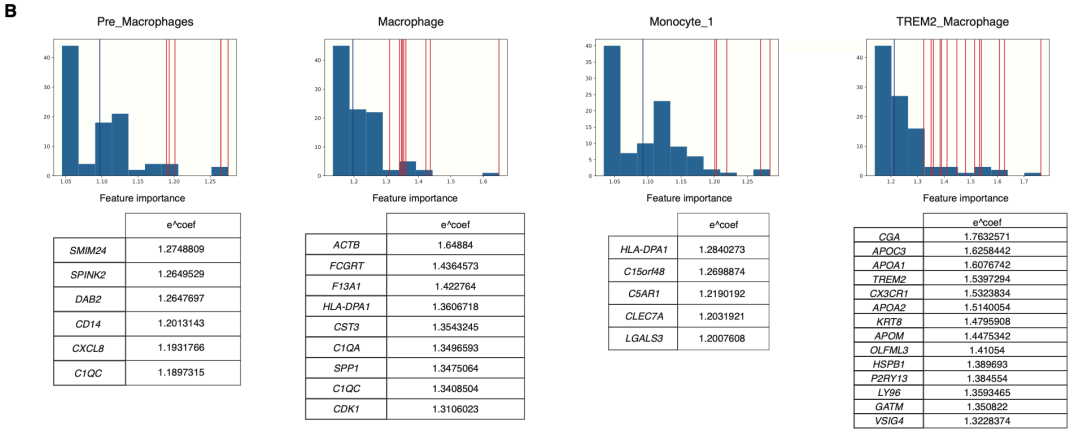
For classification tasks, a model-specific decision threshold of 0.9 was used to determine predicted labels. Clusters derived from the scVI VAE integration (see integration and batch correction section) were then assigned classes if the majority projected label had a label count distribution of >(median + (1.48*MAD)) of label counts per cluster. Resultant cell state classifications were further manually checked using differentially expressed genes. Further assessment of the predicted cluster labels was carried out by computing the adjusted Rand index and mutual information scores from the modules 'sklearn.metrics.adjusted_rand_score' and 'sklearn.metrics.mutual_info_score' between the original cluster labels and predicted cluster labels in each dataset. This methodology was applied to classify and annotate several external datasets including the scRNA-seq human gastrulation data (*390*), the human AGM data (*46*), the human EL data and human fetal skin data (*85*), as well as the human YS and liver CITE-seq data.

An implementation of the EN workflow described above, in conjunction with the SAMap (Self-Assembling Manifold mapping) workflow (v1.0.7) (*391*), was used to classify and probabilistically compare cell states across the human YS scRNA-seq data and the mouse gastrulation YS data. A gene–gene sequence homology graph weighted by human and mouse sequence similarity was first constructed using the SAMAP tool. Reciprocal BLAST mapping using the tblastx tool between the entire mouse and human transcriptomes for significant homology (E-value$<10^{-6}$) was supplied. The resultant SAM object returned c=300

species-stitched PC components for the top 3000 paired genes. These PC components were used to train the cross-species EN model as a classification task described above.

LR models and weights trained on the YS and integrated fetal atlases are available via our interactive web portal in ".sav" format (see data availability) and will facilitate future use of our YS atlas for label transfer and and to rapidly annotate scRNA-seq datasets using the Python package CellTypist (v.0.1.9) (*389*).





	e^coef
SMIM24	1.2748809
SPINK2	1.2649529
DAB2	1.2647697
CD14	1.2013143
CXCL8	1.1931766
C1QC	1.1897315

	e^coef
ACTB	1.64884
FCGRT	1.4364573
F13A1	1.422764
HLA-DPA1	1.3606718
CST3	1.3543245
C1QA	1.3496593
SPP1	1.3475064
C1QC	1.3408504
CDK1	1.3106023

	e^coef
HLA-DPA1	1.2840273
C15orf48	1.2698874
C5AR1	1.2190192
CLEC7A	1.2031921
LGALS3	1.2007608

	e^coef
CGA	1.7632571
APOC3	1.6258442
APOA1	1.6076742
TREM2	1.5397294
CX3CR1	1.5323834
APOA2	1.5140054
KRT8	1.4795908
APOM	1.4475342
OLFML3	1.41054
HSPB1	1.389693
P2RY13	1.384554
LY96	1.3593465
GATM	1.350822
VSIG4	1.3228374

Figure 2.5: Probabilistic EN label transfer *(A) Left: Heat map illustrating median PAGA connectivity scores based on abstracted neighborhood distances (AGA) between YS and corresponding liver macrophage cell-states. Right: ldVAE ElasticNet LR median cluster projection probabilities between YS labels and corresponding clusters in the fetal liver. The model is trained on the YS subset of a jointly integrated 12-organ ldVAE latent representation (C=0.2, L1_ratio=0.05, R2=0.86, RMSE=0.13). (B) Histograms illustrating key discriminatory features predicted by logistic regression (top) between YS macrophage cell states weighted by impact on model prediction (e^beta). Red lines indicate positions of significant features (P<0.05 of 1-cumulative distribution function) in the histogram. This figure is adapted from Goh and Botting et al, 2023 (1).*

2.3.7 Differential lineage priming and progenitor cell fate predictions

The CellRank package (v1.5.1) was used to define and rank fate probabilities of terminal state transitions across annotated hematopoietic lineages in the YS and iPSC scRNA-seq datasets. In the YS data, cell clusters broadly annotated to be in the myeloid lineage were first subsetted from the YS data. After refinement, DCs were excluded from this subset. We did not identify any DCs in the <CS14 (pre-agm) myeloid lineage. Macrophage trajectory inference was then constructed across the myeloid subset (fig. A8E). Cells were divided by donors aged <CS14 and >CS14 (post-agm) and trajectory inference recomputed on new embeddings. First-order-kinetics matrices were imputed for each dataset using the *pp.moments* function (n_pcs=20, n_neighbours=30) in the scVelo package (v0.2.4). A Cytotrace pseudotime for state transitions across each dataset was then computed to direct graph-edges towards estimated neighborhood regions of increasing differentiation using the Cytotrace kernel provided within the CellRank package. The resultant KNN and Cytotrace pseudotime were used to compute a probability transition matrix with the *compute*

_transition_matrix command in Cytotrace. Neighborhoods of cells representing terminal states of differentiation were identified using true Schur matrix eigen decomposition of the transition matrix *compute_schur* (n_components=20, method=brandts), followed by the *compute_macrostates* (n_states=10) command in Cytotrace. The resultant terminally differentiated cell states were then manually selected if multiple terminal states were identified per lineage. Fate absorption probabilities were then computed across all cells terminating at each prespecified terminal cell state neighborhood using the *compute_absoprtion_probabilities* command in CellRank. To visualize each cell's fate probability, we first computed a circular embedding using the *pl.circular_projection* function from CellRank. In this layout, uncommitted cells lie near the center of the circle, and, depending on the model's confidence, a cell is placed closer to the corresponding terminal node as its probability of differentiating toward that state increases. Next, to highlight HSPC distributions along each predicted lineage trajectory, we overlaid a kernel density estimate (KDE) on the same circular coordinates (KDE calculated using the *tl.embedding.density* function in Scanpy), which highlights hotspots where HSPCs accumulate near particular differentiation endpoints.

For HSPC lineage priming analyses which included the respective embryonic erythroid and erythroid terminal states, embryonic erythroid states were defined as any erythroid cell with a *HBZ* expression *z*-score > 0, and erythroid as any erythroid cell with individual expression module *z*-score of *HBA1*, *HBA2, HBG1, HBG2, HBD* > 0 (expression module scores were computed using the *sc.tl.score_genes* function in Scanpy as described in methods section 2.3.3).

2.3.8 pySCENIC for regulon analysis

The pySCENIC package (v0.9.19) was used to identify transcription factors (TFs) and their target genes in the YS and iPSC scRNA-seq datasets. The ranking database (hg38 refseq-r80 500bp_up_and_100bp_down_tss.mc9nr.feather), motif annotation database (motifs-v9-nr.hgnc-m0.001-o0.0.tbl) and list of TFs (lambert2018.txt) were used. An adjacency matrix of TFs and their targets was generated. TF activity from the AUcell output was modeled along diffusion pseudotime rankings of each trajectory and used to train a nonlinear Generalized Additive Model (nlGAM) using the pyGAM.LinearGAM model to identify TF modules which significantly changed across each lineage pseudotime. A gridsearch of between 50 and 200 splines were calculated. Significantly changing TF regulons across pseudotime were classified with a $P<0.05$ and reported in **Fig. 5, D and H)**. Regulon matrix heatmaps were plotted using the *Seaborn* (v0.12.1) package in Python. Regulon scores were variance-scaled and min–max-standardized with a distribution of 0-1.

2.3.9 Cell-cell interaction predictions using CellPhoneDB

To assign putative cell–cell interactions within the YS scRNA-seq dataset, we used CellPhoneDB (v2.1.2). Log-transformed, normalized, and scaled gene expression values for all cell states were exported. CellPhoneDB was run using the statistical method using the receptor-ligand database (v2.0.0) with a significance *P*-cut-off of 0.05. Outputs were ranked by log-mean expression for interactions between cell types of interest in each analyses and plotted as a *z*-scored heatmap to show standard deviations from mean for each receptor–ligand pair.

2.4 Imaging and spatially resolved sequencing

2.4.1 Hiplex RNAscope

Human YS tissue (8PCW) was frozen in OCT compound (Tissue-Tek). 12-plex smFISH was performed using the RNAscope HiPlex v2 assay (ACD, Bio-Techne) on three cryosections (10 μm) per manufacturer's instructions, using the standard pretreatment for freshly frozen samples and permeabilized with Protease III, for 15 min at room temperature. The imaging cycles, primary probes and label fluorophores were: *Cycle1_KLRB1_AlexaFluor488*, *Cycle1_CD1C_Dylight550*, *Cycle1_IL7R_Dylight650*, *Cycle1_SPINK2_AlexaFluor750*, *Cycle2_P2RY12_AlexaFluor488*, *Cycle2_TNFA_Dylight550*, *Cycle2_LGALS3_Dylight650*, *Cycle2_IL33_AlexaFluor750*, *Cycle3_PLVAP_AlexaFluor488*, *Cycle3_SPINK1_Dylight550*, *Cycle3_C1QA_Dylight650*, *Cycle3_ACTA2_AlexaFluor750*, *Cycle4_P2RY12_Opal570 and Cycle4_IBA1_Cy5*. Slides were counterstained with DAPI and coverslipped for imaging.

For protein validation, slides were fixed with 4% (w/v) paraformaldehyde (PFA) for 60 min at room temperature and then washed and dehydrated in an ethanol gradient (50 to 100%) for 5 min each. Sections were treated with Protease III (ACD, Bio-Techne) for 15 min at room temperature, then washed with PBS prior to blocking in 10% (v/v) normal donkey serum containing 1% (w/v) Triton X-100 and 0.2% (w/v) gelatin for 60 min at room temperature. Primary antibodies were incubated at 4°C overnight, then washed three times for 20 min each with a wash buffer (0.1% (w/v) Triton X-100 in PBS). Slides were blocked with HRP Block (ACD, Bio-Techne) for 60 min at room temperature, and washed with ACD Wash Buffer (ACD, Bio-Techne) prior to addition of secondary antibody and incubation for 60 min at room temperature. Slides were washed three times for 20 min each (0.1% (w/v) Triton X-100 in PBS). TSA-Opal570 was added for 10 min at room temperature, then washed three times with ACD Wash Buffer. Slides were counterstained with DAPI and coverslipped for imaging.

Imaging was performed on a custom two-camera spinning disk confocal microscope built around a Crest Optics X-light v3 module by Cairn Research, a scientific equipment manufacturer. The instrument was controlled using the Micro-Manager software (*392*). All imaging was performed in spinning disk confocal mode with a 40X water immersion objective (NA 1.15, 180nm/pixel) and 1.5-µm *z*-step using Prime BSI Express (Teledyne Photometrics) camera.

2.4.2 RNAscope image analysis

Before each imaging experiment, a slide covered in a sparse layer of 0.5-µm Tetraspeck beads was imaged in all channels. The bead images in all channels were then registered against the beads in the DAPI channel and their respective affine transforms were saved.

After imaging, each individual tile was *z*-projected with a maximum intensity projection, then the channels were transformed using the saved affine transforms. The projected, transformed tiles were saved back to a temporary directory along with a bigstitcher-compatible XML file. The BigStitcher software (*393*) was then used to stitch the transformed tiles together and the final stitched image exported for further analysis.

All imaging cycles for a given tissue section were registered in two steps. First, we used feature registration algorithm implemented in Python via OpenCV-contrib library (version 4.3.0) (*394*) to compute an affine transformation of DAPI channel from cycle r>1 (moving image) with respect to DAPI channel from the first cycle r=1 (reference image). Key points were detected using the FAST feature detector, whose surrounding areas were described using the DAISY feature descriptor, while the FLANN-based matcher was used to find correspondences between pairs of key points from reference and moving images and filter out unreliable points. The remaining key points were processed using the RANSAC-based

algorithm that aligns them and estimates affine transformation parameters with four degrees of freedom.

For the second registration step, a nonlinear registration algorithm based on Farneback optical-flow available in Python via OpenCV library was used to achieve more accurate registration by warping images locally. Specifically, local warping was computed using the DAPI channel, from cycle r>1 with respect to the corresponding channel of the first round. The computational pipeline implementing these registration steps was optimized so that it could be performed efficiently on large images. The corresponding code for feature registration is available at github.com/BayraktarLab/feature_reg, while the code for optical-flow registration at github.com/BayraktarLab/opt_flow_reg.

2.4.3 Immunohistochemistry

Formalin-fixed, paraffin-embedded blocks of YS 4-8PCW, EL 7-8PCW, and healthy adult liver were sectioned at 4-μm thickness onto slides coated with 3-aminopropyltriethoxysilane (APES).

For hematoxylin and eosin staining, slides were dewaxed in xylene and rehydrated through graded ethanol, as previously described (*6*). Rehydrated slides were incubated for 5 min in Mayer's hematoxylin (Dako, Agilent), rinsed in tap water and then differentiated for 2 s in acid alcohol before washing in tap water followed by Scott's tap water substitute (Leica Biosystems). Sections were counterstained in triple eosin (Dako, Agilent) for 5 min before being rinsed in tap water, dehydrated through graded ethanol (70% to 99%), and then placed in xylene before mounting with DPX (Dako, Agilent).

For immunohistochemistry (IHC), dewaxing, rehydration, and staining was performed using the Discovery Ultra auto Stainer and kits (Ventana, Roche) following the manufacturer's protocols. Slides were counterstained with one drop of hematoxylin II (Ventana, Roche) for 8 min, rinsed with Reaction Buffer and one drop of Bluing reagent (Dako, Agilent) added for 4 min. The slide was then rinsed with a Reaction buffer, before being dehydrated by hand through graded ethanol (70% to 99%), placed in xylene and mounted with DPX (Dako, Agilent).

Rabbit polyclonal anti-human alpha-1-fetoprotein (AFP; Agilent) staining was performed by NovoPath, Newcastle upon Tyne NHS Trust, using a proprietary method.

For the Martius Scarlet eBlue (MSB) stain, slides were dewaxed in xylene and rehydrated through graded ethanol as previously published (*6*). Rehydrated slides were placed in Bouins' fixative (Atom Scientific) for 1 hour at 60°C, washed in running water, incubated in Weigert's solution (Atom Scientific) for 10 min and washed in water. Slides were differentiated in 0.9% ethanol for 1-2 s before rinsing in tap water followed by Scott's tap water substitute (Leica Biosystems), distilled water and finally 95% ethanol. Slides were then incubated stepwise in Martius yellow (3 min) (Atom Scientific), Brilliant crystal scarlet (6 min) (Atom Scientific), and 50% (v/v) Methyl blue (2 min) (Atom Scientific), washing with distilled water between each stain. Slides were washed in tap water, rapidly dehydrated (2-3 min) through graded ethanol (70 to 99%), then placed in xylene before mounting with DPX mountant (Dako, Agilent).

All slides were imaged at 20X magnification on a NanoZoomer S360 (Hamamatsu) digital slide scanner. MSB stained images were deconvolved into respective Martius yellow, crystal scarlet and methyl blue channels using the Colour Deconvolution plugin (v1.8) (Masson

Trichrome) in FIJI with thresholds set using the Otsu method. Pseudocolors for each deconvolved channel were then assigned as in **Fig. 2C.**

2.4.4 ASGR1 and CD34 immunofluorescence microscopy

YS sections were baked onto slides for 2 hours at 60°C before being dewaxed in xylene and rehydrated through graded ethanol as previously described (*6*). Slides were washed with distilled water then placed in a pressure cooker with boiling citrate buffer pH 6 (10 mM citric acid (Sigma), 0.05% v/v Tween 20 (Sigma) in DI water) for 2 min for antigen retrieval. Slides were then washed for 3 min with distilled water followed by 3 min in PBS (Sigma). Sections were blocked with 20% (v/v) goat serum (R&D Systems) for 45 min at room temperature. Primary antibodies were diluted in blocking solution, added to the sections and incubated for 1 hour at room temperature. Slides were washed twice for 3 min each in a wash buffer (0.1% (w/v) Triton X (Sigma) in PBS), then twice for 3 min each in PBS. Secondary antibodies were diluted in blocking solution, added to section and incubated for 2 hours at room temperature. The wash step was repeated and then 300 nM DAPI (Sigma) was added. Slides were incubated for 5 min before washing with PBS. Slides were then mounted with ProLong™ Diamond Antifade (Thermofisher) and imaged on a Zeiss Axioimager with Zeiss ZEN pro software.

2.4.5 SMA and LYVE1/CD34 immunofluorescence microscopy

PFA-fixed YS was cryoprotected with sucrose 10%, embedded in gelatin-sucrose solution (7.5% x/v gelatin (VWR 24350.262), 10% w/v sucrose (VWR27478.296), in 0.12M PBS), frozen at -50℃, then sectioned at 14µm. Slides were stored at -80℃ until use, dried for 30 min, then blocked with PBS Gelatin Triton (0.2% w/v gelatin, 0.25% Triton X-100 (Sigma-Aldrich) in PBS) for 1 hour. Primary antibodies were diluted in blocking solution ,

added to the sections, and incubated overnight. Slides washed with PBS three times at 10-min intervals. Secondary antibodies were diluted in blocking solution and added to sections to incubate for 2 hours. Hoechst 33258 (Sigma-Aldrich) was added to the secondary antibody solution. Sections were washed with PBS three times at 10-min intervals, and coverslips were mounted with Mowiol (Calbiochem). Sections were imaged at 20X magnification on Leica DM6000 widefield microscope with MetaMorph software. Brightness and contrast were adjusted and a scale bar was added with FIJI (*395*).

2.4.6 Light-sheet fluorescence microscopy

Candidate antibodies were screened by immunofluorescence on cryosections obtained from OCT-embedded specimens as previously described (*6*, *380*). Routine light-sheet immunofluorescence microscopy (LSFM) was then performed on floating whole-mount yolk sacs as previously described, with primary antibody incubation reduced to 10 days and secondary reduced to 2 days, both at 37°C to preserve tissue integrity. Yolk sacs were embedded in 1.5% agarose blocks prior to solvent-based clearing as previously described (*380*). YS retained its spherical shape throughout the procedure. Imaging was performed as previously described in dibenzyl ether with a Miltenyi Biotec Ultramicroscope Blaze (sCMOS camera 5.5MP controlled by Inspector Pro 7.3.2 acquisition software), which generates light sheets at excitation wavelengths of 488, 561, 640, and 785 nm. Objective lenses of 4X magnification (MI Plan 4X NA0.35) and 12X magnification (MI Plan NA 0.53) were used. Imaris (v9.8, BitPlane) was used for image conversion, processing, and video production. Blender 3.0 was used to edit videos and add text. All raw image data are available on request (A.C. and M.H.).

3 Results chapter 1: The human YS; a multi-functional site of early haematopoiesis

The work outlined in this chapter is the result of a multinational, multisite collaboration between the Haniffa group led by Professor Muzlifah Haniffa (Biosciences Institute, Newcastle University and the Wellcome Sanger Institute), members of the Wellcome sanger institute, and members of Sorbonne Université.

Samples were collected by Dr Rachel Botting, Dr Emily Stephenson, and Dr Dorin Popescu. Droplet-based fetal YS scRNA-seq and CITE-seq datasets were collaboratively generated by Dr Rachel Botting, Dr Emily Stephenson, Justin Engelbert, and myself. Plate-based scRNA-seq data generation, including FACS isolation, were performed by Dr Emily Stephenson, Dr Rachel Botting and Dr Laura Jardine. RNA-sope experiments were conducted by Nana-Jane Chipampe and Kwasi Kwaka (Wellcome Sanger Institute). Light-sheet microscopy was performed by Dr Yorick Gitton, and Megumi Inoue, led by Professor Alain Chédotal (Sorbonne Université, INSERM, CNRS, Institut de la Vision).

This chapter is a lightly-edited version of the manuscript which I co-first authored which was recently published in the journal *Science (1)*. I assisted in planning and organising the tissue preparation for scRNA-seq and CITE-seq data generation for some EL and YS, analysed, integrated, and annotated SS2, scRNA-seq and CITE-seq data. TotalVI integration of CITE-seq data was jointly performed by Antony Rose and myself. I organised and oversaw this process including downstream clustering, and initial manual annotations of the CITE-serq data. Additionally, I performed logistic regression-based probabilistic celltype correspondence analyses between YS scRNA-seq and CITE-seq data, and between YS and EL scRNA-seq data.

3.1 Introduction

In this chapter I describe the collaborative work towards generating and annotating a single-cell, spatial, and multi-omic atlas of the developing human YS.

3.1.1 Human embryonic YS haematopoiesis

The yolk sac (YS) is an evolutionarily conserved extraembryonic structure important for nutritional support as well as the initial site of haematopoiesis in humans. In many mammals, the YS fuses with the chorion, forming an early placenta that facilitates nutrient transfer during organogenesis. In humans there is no yolk and no fusion with the chorion: the YS floats within the extraembryonic cavity tethered only to the embryo via the vitelline circulation. These differences have led to debate about whether the human YS is vestigial or whether it serves a specific function during embryonic development.

The primary human YS is derived from the hypoblast at around the time of embryo implantation (Carnegie stage 4, CS4; ~1PCW) (*93*, *101*). A secondary YS beneath the embryonic disc supersedes the primary structure at around CS6 (~2.5PCW) and persists until 8PCW (*93*, *101*). The secondary YS surrounds a vitelline fluid-filled cavity with three tissue compartments: mesothelium (which is a mesoderm-derived epithelial layer), mesoderm (which contains an array of cell types, including endothelial cells, blood cells and smooth muscle), and endoderm. A comparative study in mouse, chick and later human YS, has inferred from whole tissue RNA sequencing and mass spectrometry of coelomic fluid that the YS may support haematopoiesis and nutrient transfer in all three species (*101*).

The human YS is the dominant site of embryonic human developmental haematopoiesis with blood islands containing primitive erythroblasts detectable from 2.5PCW (CS6), providing the first blood and immune cells during development (*396*, *397*). While model organism data (mouse and chicken) demonstrate that haemato-endothelium arises from mesoderm(*398–400*), some evidence suggests it arises from extraembryonic endoderm in humans (*401*). YS also produces MK, mast cells and myeloid cells (*6*).

Definitive HSCs arise from the AGM region of the dorsal aorta at CS14 (~5PCW), subsequently entering YS, EL, and later, FBM (Fig 1.3)(*105*, *402*, *403*). Though the human developmental YS plays a vital role in the developmental relay of haematopoiesis, studies into this transient state of haematopoiesis have mainly been conducted in model organisms

(*398–400*). The longevity of YS-derived HSPCs, the full repertoire of human YS-derived HSPCs, blood and immune cells, whether they differ from later cell counterparts, and how they contribute to long-lived cells requires further clarification.

3.2 Methods

To investigate the functional role, hematopoietic repertoire and supportive features for embryonic survival of the human embryonic YS, single cell suspensions of YS membrane, contents, and vitelline duct dissections were produced via enzymatic digestion via collagenase following mechanical dissociation (Fig. 3.1). Cell suspensions were sorted by FACS for live/dead (DAPI), CD45+ and CD45- fractions. Live isolates were counted and loaded onto the 10X Genomics Chromium Controller to achieve a maximum yield of 10,000 cells as per 10X genomics 5' V1 reaction kits recommendations (Fig. 3.1).

For plate-based SS2 experiments, FACS isolation for live/dead (DAPI), CD45+ and CD45- fractions were subjected to a modified Smart-seq2 protocol as previously described by Villani *et al. (19)* (Fig. 3.1).

For the CITE-seq experiments frozen cell suspensions were thawed, counted and pooled. Cells were labelled using the commercially 10X genomics ADT cocktail. Cell suspensions were sorted by FACS for live/dead (DAPI), CD34+ and CD34- fractions. Live isolates were counted and loaded onto the 10X Genomics Chromium Controller according to the 10X genomics 3' V3 kit specifications (Fig. 3.1).

For methods on the Hiplex RNA-scope and Light-sheet fluorescence microscopy which were performed by collaborators as named in the introduction to this chapter, and for further details on tissue processing methodology, including specific reagent concentrations, blocking strategies, and imaging platforms, please refer to the materials and methods section of this thesis.

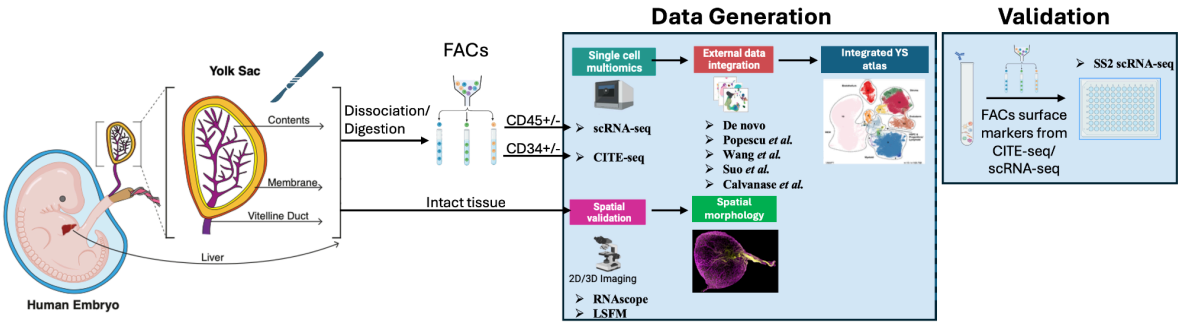


Figure 3.1: Schematic overview of YS and matched embryonic liver tissue acquisition and single cell experimental outline. *This cartoon depicts the generation of scRNA-seq and CITE-seq data from dissections of the membrane, vitelline duct, and contents of the human secondary YS, the validation of broad cell types and scRNA-seq and CITE-seq derived key marker genes using plate-based SS2 scRNA-seq of select cell types, and spatial data generation via LSFM (shown) and RNA-scope technologies. Spatial data illustrated was performed by collaborative efforts as named in the introduction. Figure produced using BioRender.com and adapted from Goh et al, 2023 , a publication I co-first authored and the primary work for which this thesis is based on.*

Publicly available scRNA-seq datasets were integrated with the de novo produced YS and EL scRNA-seq and CITE-seq data, and used for the transfer learning and contextualisation of the YS data within an extended pan-organ developmental atlas. scRNA-seq datasets generated from previous studies from our research group and wider collaborations were integrated with publicly available data culminating in 12 distinct fetal and embryonic tissues including: YS, AGM, EL/FL, BM, brain, skin, kidney, gonads, gut, Mesenteric Lymph Nodes (MLN), spleen, and thymus. External datasets were re-mapped with CellRanger to a reference genome and computationally processed in line with the YS and EL scRNA-seq generated for this work (Fig. 3.1 - 3.2).

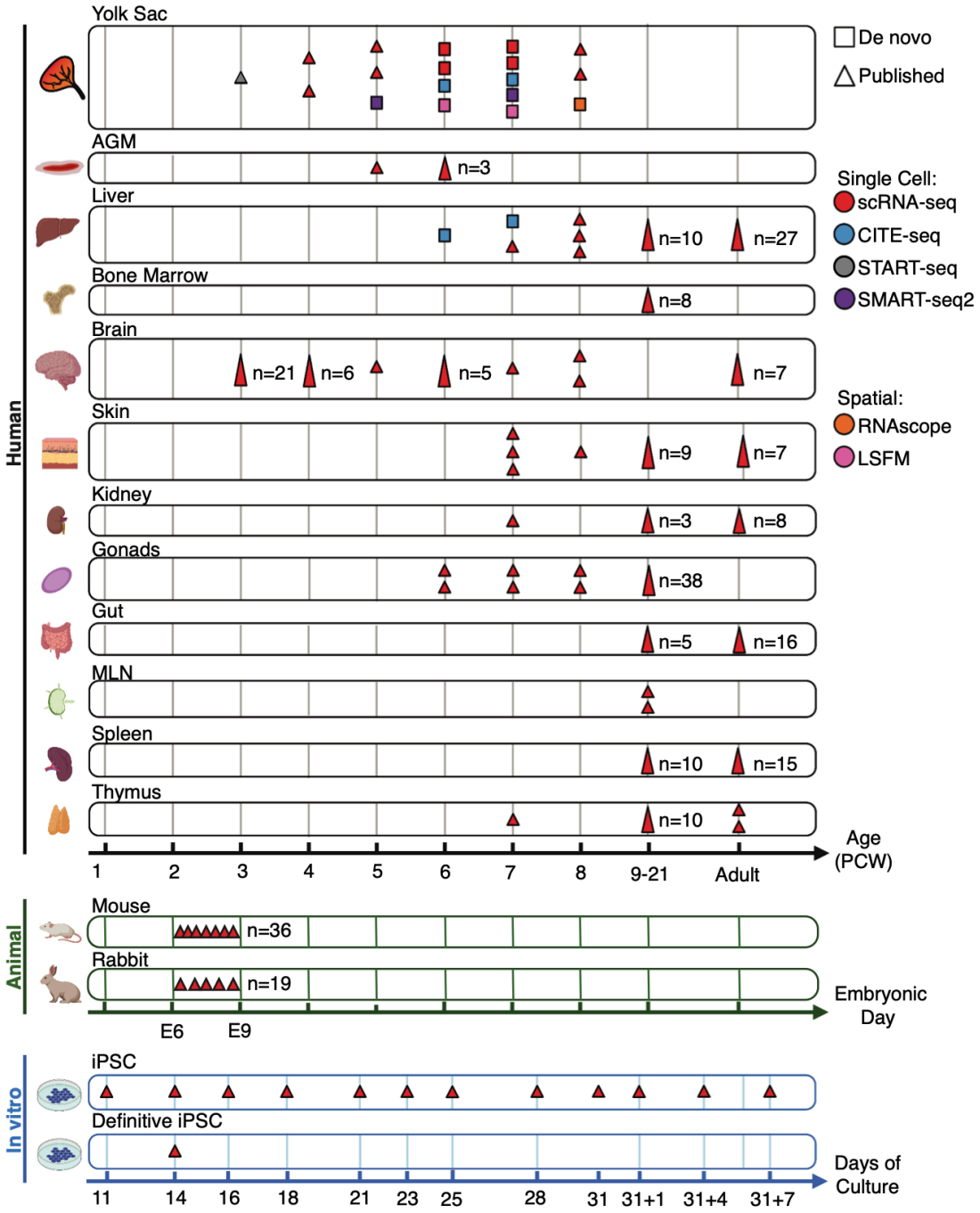


Figure 3.2: Illustration summarising the developmental timeline of different tissues and data capture modalities used in this study. *These include the De novo YS and EL scRNA-seq and CITE-seq data generated for the recently published work I co-first-authored (squares). Triangles represent published data: YS (6, 46, 85, 385), AGM (46), FL (6), BM (3), brain (404), skin (85), kidney (405), gonads (44), gut, MLN, spleen, and thymus scRNA-seq data, published available murine and rabbit data (406), and two published iPSC culture systems representing in vitro-derived early and definitive hematopoiesis (46, 89). Colour indicates assay used. Figure adapted from Goh et al, 2023 .*

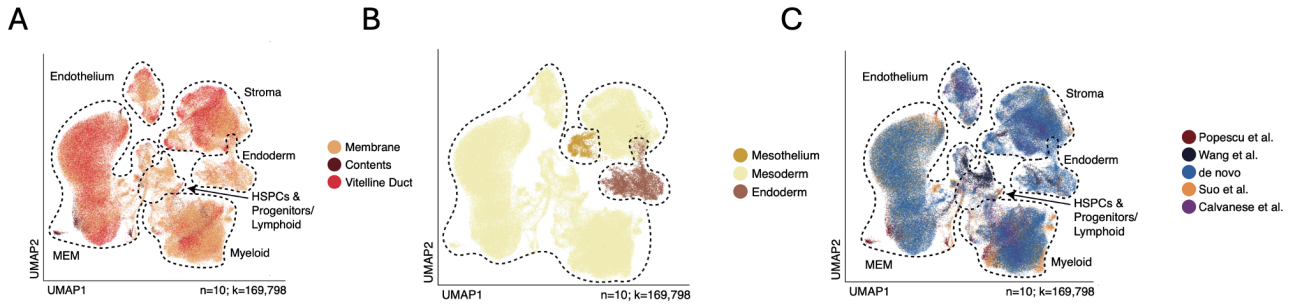


Figure 3.3: UMAP of YS scRNA-seq data integrated across newly generated and published data: *A) UMAP coloured by dissected region (membrane, contents, and vitelline duct). B) UMAP coloured by annotated YS layers (Mesothelium, Mesoderm, and Endoderm) C) UMAP coloured by data source integrated (6, 46, 85, 385). Figure adapted from Goh and Botting et al, 2023 (1).*

To integrate YS scRNA-seq datasets from public sources including from CS10/CS11 and CS14 timepoints (*6*, *46*, *85*, *385*), data were remapped with a common reference genome (GRCh38-2020-A). Low quality cells were removed from concatenated data (<2000 reads, <200 genes, and >20% mitochondrial reads). Doublet detection was performed with dynamic thresholds and likelihood of maternal contamination was estimated, these cells were not excluded from analyses till after annotation in order to ascertain possible specific cellular contributions to doublets or maternal contamination (see methods). Data was then normalised, and log-transformed (see methods). Highly variable gene (HVG) selection was performed and dimensionality reduction and integration was carried out using the scVI module within scvi-tools (v0.19.0) (*330*) with biological replicate taken as the technical covariate. To ensure model performance was optimal for each independent analysis, scVI was benchmarked against the python implementation of Harmony (*350*) (*Harmonypy* v0.0.5) at various theta values between 1 and 20. kBET (*387*) and Silhouette scores (*sklearn.metric.sil_score*) were computed for each iteration between donor covariates and compared to the scVI integration (see methods section for integration and benchmarking). Output latent representation was then clustered on scRNA-seq and CITE-seq datasets was performed using the Leiden algorithm (*361*) (*sc.tl.leiden*) with a resolution parameter of *res*=1.5 (CITE-seq res=3) on a k-nearest neighborhood graph (n=30 for scRNA-seq and n=15 for CITE-seq). For visualisation, the uniform manifold approximation (UMAP) algorithm was utilised to create 2 dimensional visual representations of each displayed data subset.

To align YS scRNA-seq data cell state annotations with the multi-modal CITE-seq data with surface proteomic profiling, we employed a ldVAE-based ElasticNet LR framework for transfer learning and probabilistic label transfer. scRNA-seq modalities from both data were first concatenated, and linear VAE latent representations were computed using the LDVAE module within *scvi-tools* (hidden layers=256, dropout-rate=0.2, reconstruction-loss=negative binomial) with donor, chemistry, and dataset of origin taken as technical covariates. We then applied the ElasticNet LR framework with the ldVAE embeddings of the scRNA-seq subset serving as reference as described in the methods section. This outputs per-cell probabilistic cell-state assignments within the scRNA-seq modality of the CITE-seq data. Majority voting for these probabilistic labels was then carried out upon independently computed clusters from *TotalVI* derived embeddings of CITE-seq data using both protein and RNA modalities. Resultant cell state classifications in CITE-seq data are shown (Fig. 3.4, A and B).

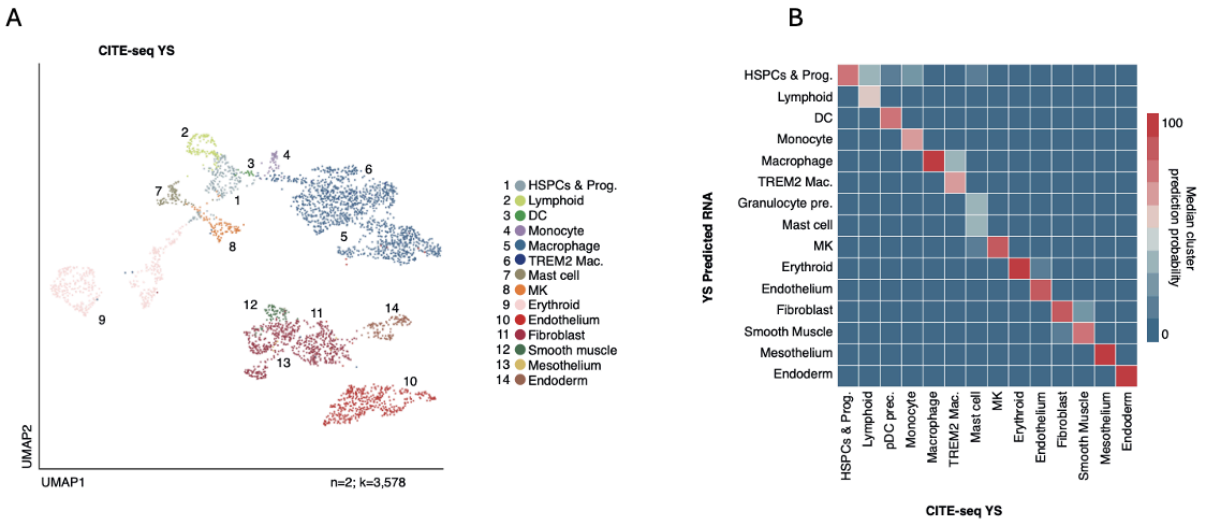


Figure 3.4: UMAP and label transfer of integrated CITE-seq data: *A) UMAP of YS cells from CITE-seq data, intersected between RNA and protein modalities batch corrected using TotalVI from n=2 biologically independent samples (n=2, c=3,578). Colours represent broad cell states. (B) Heatmap of class prediction probabilities output from an elasticnet logistic regression model trained on YS scRNA-seq cell states (y-axis) and projected onto corresponding cluster-derived cell states in YS CITE-seq data (x-axis). Colour scale indicates median probabilities (see methods).*

Further details on integration methods, annotation strategy, tissue processing methodology, including specific reagent concentrations, blocking strategies, and imaging platforms can be found in the materials and methods section of this thesis.

3.3 Results

3.3.1 A single-cell atlas of early haematopoiesis

We performed droplet-based single cell RNA-sequencing (scRNA-seq) to profile human YS, and integrated with four external datasets to yield 169,798 high-quality cells from 10 samples spanning 4-8PCW (CS10-CS23) (Fig. 3.1 - 3.3 A to C; Fig. 3.5 A and B; Fig. A1, A to C). We illustrate this process in the schematic (Fig 3.1).

Integrating the YS data across public data from four sources (Fig. 3.3), we then applied graph-based leiden clustering which yielded 39 cell types grouped into 15 broad categories including hematopoietic cells, endoderm, mesoderm, and mesothelium (Fig. 3.5B, and Fig. A1 D). Clustering and annotation recovered cell states corresponding to germ layers of origin (mesothelium, mesoderm, and endoderm), with each layer delineating transcriptionally distinct populations with mesothelium and endoderm comprising of largely transcriptionally homogenous populations, and mesoderm containing populations of transcriptionally heterogenous cell states including the vasculature, and hematopoietic lineages (Fig. 3.5B, and Fig. A1 A to D).

Three-dimensional visualisation of the YS by light sheet microscopy marked the $CD34^{hi}LYVE1^{lo}$ vitelline artery and $CD34^{lo}LYVE1^{hi}$ vitelline vein contiguous with a branching network of $CD34^{lo}LYVE1^{hi}$ vessels (Fig. 3.5C ; fig. A3F). The $CD34^{lo}LYVE1^{hi}$*IL33*$^{+}$ vessels were situated within the mesoderm, a distinct layer beneath the ASGR1^{+}*SPINK1*$^{+}$ endoderm (Fig. 3.5, C and D; and Fig. A3, F to I; and A4B). RNAscope performed on YS showed *ACTA2*$^{+}$ smooth muscle cells formed a sublayer between mesoderm and endoderm (Fig. 3.5D; and fig. A4B). Macrophages (*C1QA*$^{+}$*CD1C*$^{+/-}$) and a small number of dendritic cells (DCs) (*C1QA*$^{-}$*CD1C*$^{+}$) were identified within the mesoderm (Fig. 3.5D; and fig. A4, A and B).

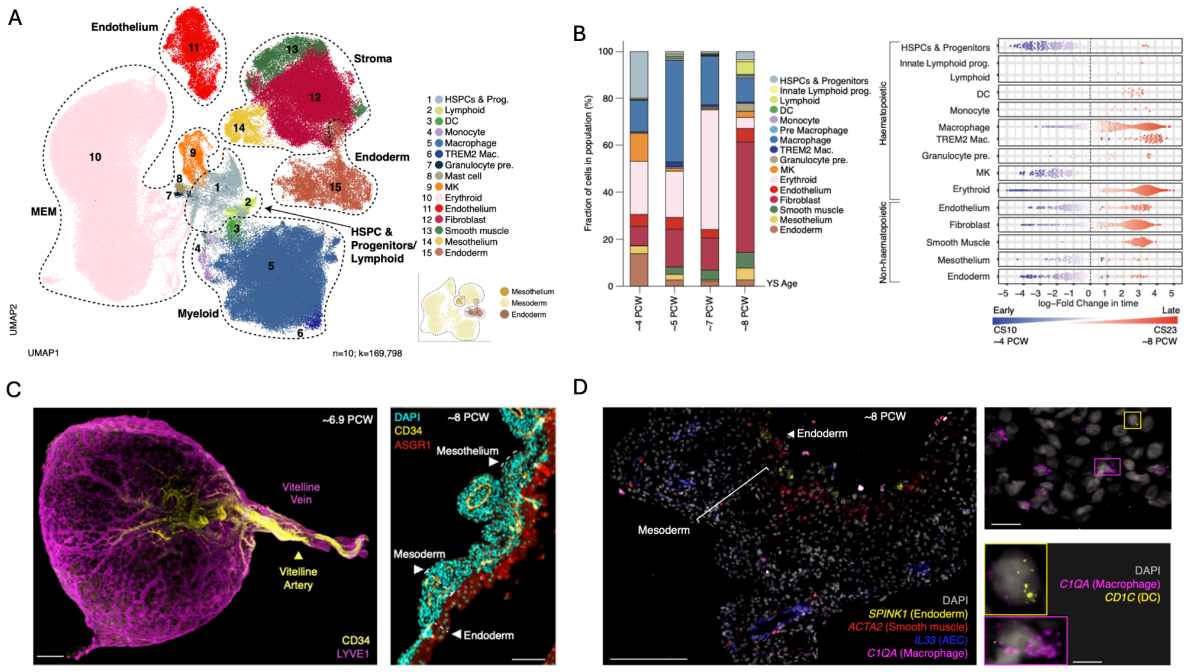


Figure 3.5: A single cell atlas of embryonic YS: *(**A**) UMAP visualisation of YS scRNA-seq data (n=10; c=169,798), colours represent broad cell states: DC: dendritic cell, Mac: macrophage, MEM: megakaryocyte–erythroid–mast cell lineage, MK: megakaryocyte, pre.: precursor.* ***(B)*** *Left**:** Bar graph showing the proportion representation of cell states in YS scRNA-seq data by gestational age. Right: Milo beeswarm plot of YS scRNA-seq neighborhood differential abundance across time. Blue/red neighborhoods are significantly enriched earlier/later in gestation respectively. Color intensity denotes degree of significance.* ***(C)*** *Left: light-sheet fluorescence microscopy of $CD34^+$ and $LYVE1^+$ vascular structures in YS (representative ~6.9PCW sample; scale bar: 500 µm; movie S1). Right: Immunofluorescence of an ~8PCW YS highlighting endoderm with anti-ASGR1 (red) and endothelium with anti-CD34 (yellow), costained with DAPI (blue). Scale bar: 100 µm.* ***(D)*** *RNAscope of YS (representative 8PCW sample). Left: endoderm (SPINK1; yellow), smooth muscle (ACTA2; red), AEC (IL33; blue), and macrophages (C1QA; magenta) (scale bar: 200 µm). Right: DCs (CD1C; yellow box) and macrophages (C1QA; magenta box) (scale bar: 50 µm). Figure adapted from Goh and Botting et al, 2023 (1).*

Comparing cell state proportions across time in the YS revealed that the most prevalent hematopoietic cell types in early YS (CS10; ~4PCW) were HSPCs, erythroid cells, macrophages, and megakaryocytes. Both HSPCs and MKs proportionately diminished

thereafter, whereas erythroid cells and macrophages were sustained. DCs and *TREM2*$^{+}$ macrophages did not emerge until >6PCW (Fig. 3.5B). The ratio of hematopoietic to nonhematopoietic cells was around 3:1 in early YS (CS10; ~4PCW), with endoderm relatively abundant (Fig. 3.5B). The ratio approached 1:3 in late YS (CS22-23; ~8PCW) due to expansion of fibroblasts (Fig. 3.5B). The transcriptional profile of MKs was consistent across gestation, but both erythroid cells and macrophages had early and late gestation-specific molecular states, suggesting dual waves of production (Fig. 3.5B; fig. A4C).

Key differentially expressed marker genes for each population were validated by plate-based scRNA-seq (SS2), and by CITE-seq profiling (Fig. 3.6 A and B; Fig. 3.7 A and B; Fig. 3.4A and B; Fig. A1, B to F) (see methods). We used the term "HSPC" for cells collectively based on their expression of a core HSPC signature (e.g., *CD34*, *SPINK2*, and *HLF*) without implying long-term repopulating capacity or multilineage potential. With comparison datasets, unless otherwise specified, we adopted published annotations (*2*). Surface protein expression from CITE-seq of n=2 YS cell suspensions (fig. A1, G and H) identified combinatorial antigens for cell purification and functional characterization (Fig. A2A).

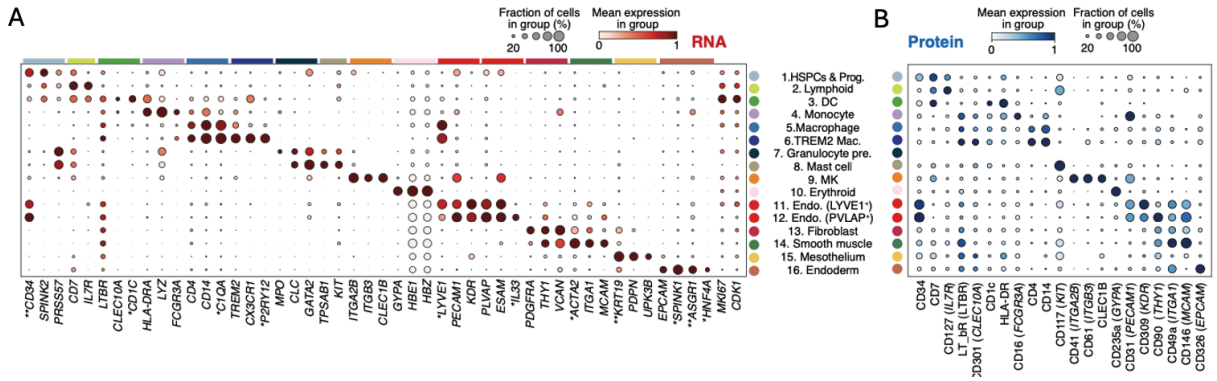


Figure 3.6: Validated cell type marker genes: *(A): Dot plot showing the mean expression (color) and proportion of cells expressing genes (dot size) of broad cell states in YS scRNA-seq data. (B): Equivalent protein expression (color) and proportion of cells expressing proteins (dot size) from YS CITE-seq data (n=2; c=3,578). Equivalent gene names are in parentheses. * indicates genes validated by RNAscope and ** indicates proteins validated by IHC/IF. Data are variance-scaled and min-max-standardised. Figure adapted from Goh and Botting et al, 2023 (1).*

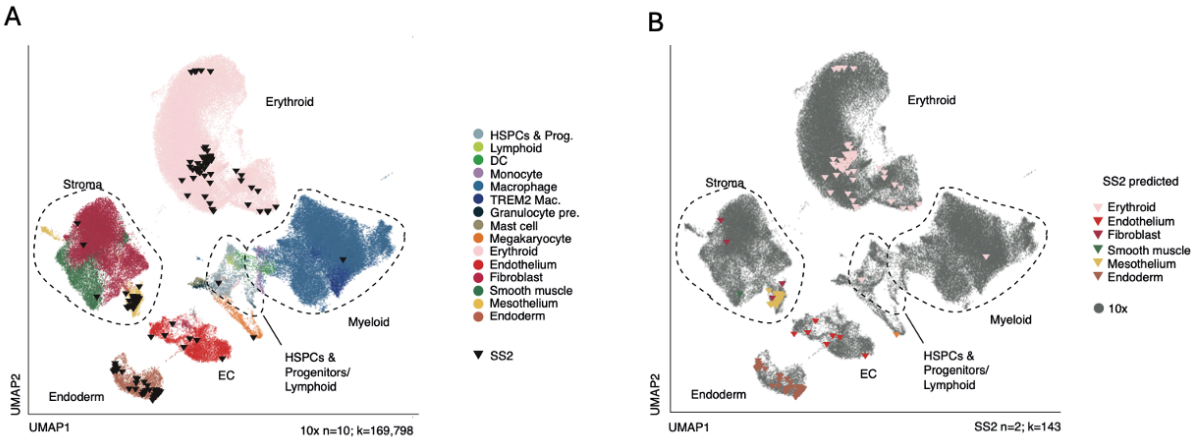


Figure 3.7: SS2 cell state validation: *(A): UMAP visualisation of the YS cells (n=10, c=169,798) integrated with c=143 cells from plate-based sequencing (SS2) cells (triangles) FACS isolated from n=2 individual donors (5-7PCW). Colours indicate cell states in droplet based scRNA-seq (10x). (B): UMAP visualisation of the YS cells shown in (A), with colours highlighting cell states in plate-based scRNA-seq (SS2). Figure adapted from Goh and Botting et al, 2023 (1).*

To contextualise the early hematopoietic landscape, we generated matched EL scRNA-seq and CITE-seq data (fig. 3.8, A to E and A3, A to C), which confirmed the presence of discrete B cell progenitor stages only in the liver (fig. 3.8D). Around half of YS lymphoid cells were innate lymphoid progenitors, which showcased inferred transcriptional proximity to natural

killer (NK) and innate lymphoid cell (ILC) precursor states on force-directed graph (FDG) visualisation (fig. A3D). A small population of cells were termed “Lymphoid B lineage” due to their expression of *CD19*, *CD79B*, and *IGLL1*. These cells did not express the typical B1 markers *CD5*, *CD27*, or *CCR10* however. Given the absence of distinct B cell progenitor stages and their later emergence (>5PCW), these may constitute migratory B cells of fetal liver origin (fig. A3, D and E).

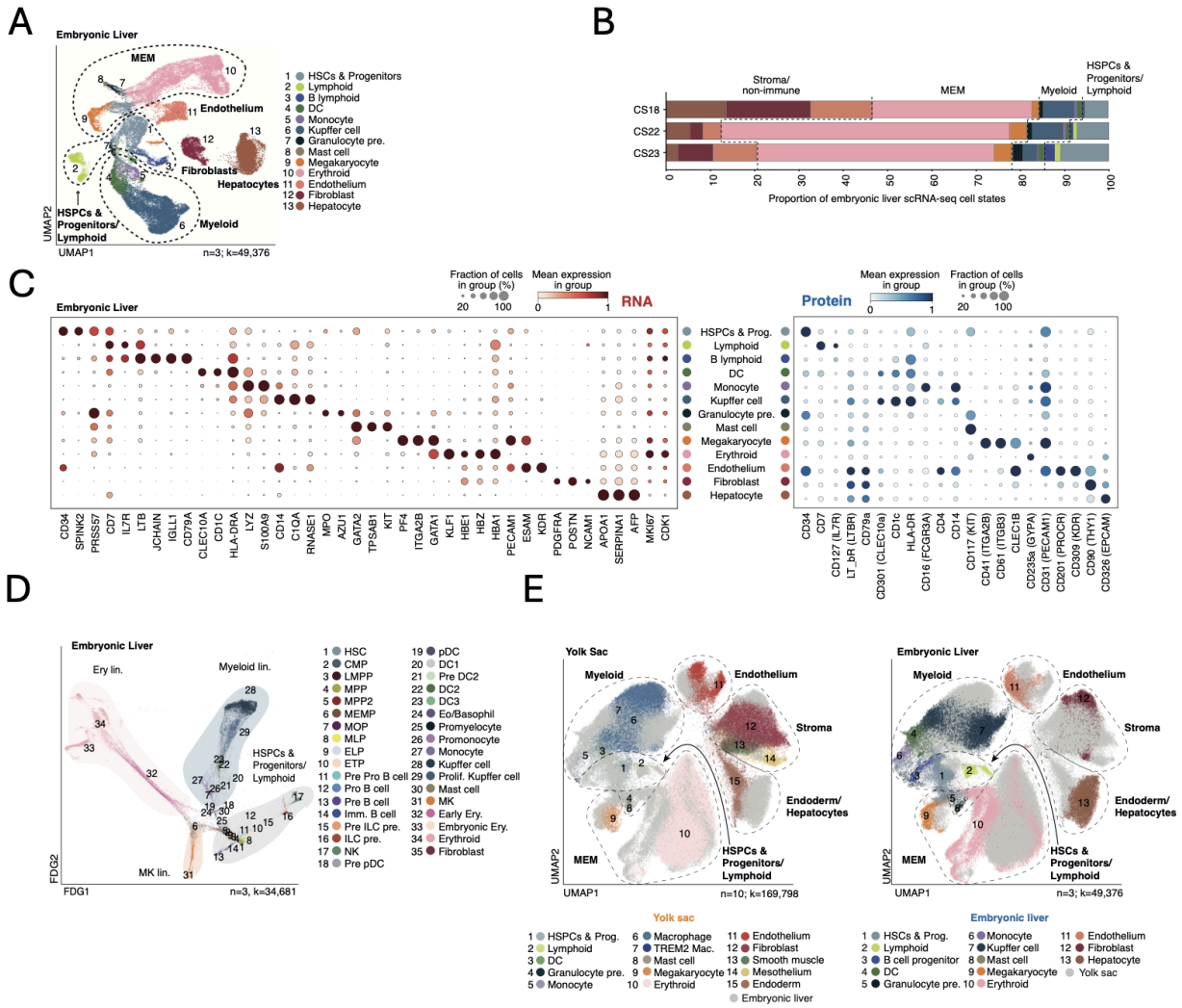

A
Embryonic Liver
MEM
Endothelium
Fibroblasts
Hepatocytes
HSPCs & Progenitors/ Lymphoid
Myeloid
UMAP1
UMAP2
n=3; k=49,376
1 HSCs & Progenitors
2 Lymphoid
3 B lymphoid
4 DC
5 Monocyte
6 Kupffer cell
7 Granulocyte pre.
8 Mast cell
9 Megakaryocyte
10 Erythroid
11 Endothelium
12 Fibroblast
13 Hepatocyte
B
Stroma/ non-immune
MEM
Myeloid
HSPCs & Progenitors/ Lymphoid
CS18
CS22
CS23
Proportion of embryonic liver scRNA-seq cell states
C
Embryonic Liver
Fraction of cells in group (%)
Mean expression in group
RNA
Protein
HSPCs & Prog.
Lymphoid
B lymphoid
DC
Monocyte
Kupffer cell
Granulocyte pre.
Mast cell
Megakaryocyte
Erythroid
Endothelium
Fibroblast
Hepatocyte
CD34
SPINK2
PRSS57
CD7
IL7R
LTB
JCHAIN
IGLL1
CD79A
CLEC10A
CD1C
HLA-DRA
LYZ
S100A9
CD14
C1QA
RNASE1
MPO
AZU1
GATA2
TPSAB1
KIT
PF4
ITGA2B
GATA1
KLF1
HBE1
HBZ
HBA1
PECAM1
ESAM
KDR
PDGFRA
POSTN
NCAM1
APOA1
SERPINA1
AFP
MKI67
CDK1
CD127 (IL7R)
LT_bR (LTBR)
CD79a
CD301 (CLEC10a)
CD1c
HLA-DR
CD16 (FCGR3A)
CD4
CD14
CD117 (KIT)
CD41 (ITGA2B)
CD61 (ITGB3)
CLEC1B
CD235a (GYPA)
CD31 (PECAM1)
CD201 (PROCR)
CD309 (KDR)
CD90 (THY1)
CD326 (EPCAM)
D
Embryonic Liver
Ery lin.
Myeloid lin.
HSPCs & Progenitors/ Lymphoid
MK lin.
FDG1
FDG2
n=3, k=34,681
1 HSC
2 CMP
3 LMPP
4 MPP
5 MPP2
6 MEMP
7 MOP
8 MLP
9 ELP
10 ETP
11 Pre Pro B cell
12 Pro B cell
13 Pre B cell
14 Imm. B cell
15 Pre ILC pre.
16 ILC pre.
17 NK
18 Pre pDC
19 pDC
20 DC1
21 Pre DC2
22 DC2
23 DC3
24 Eo/Basophil
25 Promyelocyte
26 Promonocyte
27 Monocyte
28 Kupffer cell
29 Prolif. Kupffer cell
30 Mast cell
31 MK
32 Early Ery.
33 Embryonic Ery.
34 Erythroid
35 Fibroblast
E
Yolk Sac
Myeloid
Endothelium
Stroma
Endoderm/ Hepatocytes
MEM
HSPCs & Progenitors/ Lymphoid
UMAP1
UMAP2
n=10; k=169,798
Embryonic Liver
HSCs & Progenitors/ Lymphoid
n=3; k=49,376
Yolk sac
1 HSPCs & Prog.
2 Lymphoid
3 DC
4 Granulocyte pre.
5 Monocyte
6 Macrophage
7 TREM2 Mac.
8 Mast cell
9 Megakaryocyte
10 Erythroid
11 Endothelium
12 Fibroblast
13 Smooth muscle
14 Mesothelium
15 Endoderm
Embryonic liver
Embryonic liver
1 HSCs & Prog.
2 Lymphoid
3 B cell progenitor
4 DC
5 Granulocyte pre.
6 Monocyte
7 Kupffer cell
8 Mast cell
9 Megakaryocyte
10 Erythroid
11 Endothelium
12 Fibroblast
13 Hepatocyte
Yolk sac

Figure 3.8: Analysis of human yolk sac and embryonic liver: *(A) UMAP visualisation of the cell states identified in the embryonic liver (EL) scRNA-seq dataset from n=3 independent biological repeats (c=49,376, CS18-23). Colors represent cell states. DC: dendritic cell; MEM: megakaryocyte-erythroid-mast cell lineage; pre.: precursor. (B) Stacked bar chart displaying the proportional representation of broad cell states in EL scRNA-seq data by sample. Colours match cell states shown in (A). (C) Dot plot showing the mean expression (by colour) and the fraction of cells expressing each gene or protein (by dot size) of broad cell state-defining genes in EL scRNA-seq data (n=3, c=49,376) (left), and their protein counterparts in EL CITE-seq dataset (n=9 biologically independent samples, c=57,310, CS16-17PCW) (right). Data are min-max-standardised with a distribution of 0-1. (D) Force-directed graph (FDG) visualisation of hematopoietic cell states identified in EL scRNA-seq from n=3 biologically independent donors (c=34,681). Colors represent cell states and clouds represent lineages. CMP: common myeloid progenitor; DC: dendritic cell; ELP: early lymphoid progenitor; Eo./Baso.: eosinophil/basophil; Ery.: erythroid; ETP: early thymic progenitor; HE: hemogenic endothelium; HSC: hematopoietic stem cell; HSPC: hematopoietic stem and progenitor cell; ILC: innate lymphoid cell; LMPP: lymphoid-primed multipotent progenitor; Mac: macrophage; MEM: megakaryocyte-erythroid-mast cell lineage; MEMP: megakaryocyte-erythroid-mast cell progenitor; MK: megakaryocyte; MLP: multi-lymphoid progenitor; Mono: monocyte; MOP: monocyte progenitor; MPP: multipotent progenitor; Neut: neutrophil; NK: natural killer cell; pDC: plasmacytoid DC; pre.: precursor; prog.: progenitor; prolif.: proliferating. (F) UMAP visualisation of the merged YS and EL scRNA-seq data coloured by cell state and tissue (Left, YS, n=10, c=169,798; Right, EL, n=3, c=49,376). Figure adapted from Goh and Botting et al, 2023 (1).*

3.3.4 Early versus definitive hematopoiesis in YS and liver

Human YS hematopoietic progenitors spanned two groups: HSPCs characterised by *SPINK2*, *CYTL1*, and *HOXB9* expression and cycling HSPCs characterised by cell cycle-associated genes such as *MKI67* and *TOP2A* (fig. A5A). Using markers recently associated with early (*DDIT4*, *SLC2A3*, *RGS16*, and *LIN28A*) and definitive (*KIT*, *ITGA4*, *CD74*, and *PROCR*) HSPCs (*46*), we identified early and definitive fractions within both HSPCs and cycling HSPCs (Fig. 3.9, A to C). Early and definitive HSPCs expressed canonical HSPC genes such as *SPINK2*, *HOPX*, and *HLF* (Fig. 3.9A), but diverged in expression of genes involved in multiple processes such as enzymes (*GAD1*), growth factors (*FGF23*), adhesion molecules (*SELL*), and patterning genes (*HOXA7*) (fig. A5C). Using an LR model trained on early liver progenitor states (methods 2.3.6), YS HSPCs had a high median probability of matching liver HSCs, but this probability was higher for definitive than for early HSPCs (fig. A5B). Similarly using this LR model, early cycling YS HSPCs distributed their classifier‐predicted probabilities between liver MPP and CMP, whereas definitive cycling HSPCs showed a higher logistic regression–derived probability of classifying as MPP (Fig. A5B). Differential protein expression in YS CITE-seq data indicated that CD122, CD194 (CCR4), and CD357 mark early HSPCs whereas CD44, CD48, CD93, and CD197 (CCR7) mark definitive HSPCs (fig. A5D), in keeping with the reported use of CD34 and CD44 to segregate early and definitive-type HSPCs by FACS (*407*). We confirmed that an iPSC-derived culture system reported to generate definitive HSPCs did express RNA markers characteristic of definitive HSPCs (*46*) (Fig. 3.9A), but HSPCs derived from an iPSC-derived culture system optimised for macrophage production (*89*) did not express these definitive markers but expressed HSPC canonical markers (Fig. 3.9A, 5.11B).

To assess cross tissue HSPC heterogeneity, we integrated HSPCs across hematopoietic organs (Fig. 3.9C and fig. A5E). By kernel density estimation score (KDE) on integrated UMAP embeddings, YS definitive HSPCs qualitatively overlapped with definitive HSPCs from age-matched liver (Fig. 3.9C and fig. A5E). However, because UMAP can introduce two-dimensional projection artifacts and is highly sensitive to parameters (e.g., n_neighbors) that balance local and global structure, these visual overlaps may not accurately reflect true biological relationships, as such, further evaluations will be required and may include quantitative measures such as silhouette scores, measured directly on scVI embeddings to more reliably assess cross-tissue cluster separation between HSPCs. From exclusively early HSPCs at ~3PCW, we observed rapid accumulation of definitive HSPCs after AGM development CS14 (~5PCW), likely accounting for the increase in the YS HSPC/progenitor fraction at 8PCW (Figs. 3.5B and 3.9B and fig. A5F).

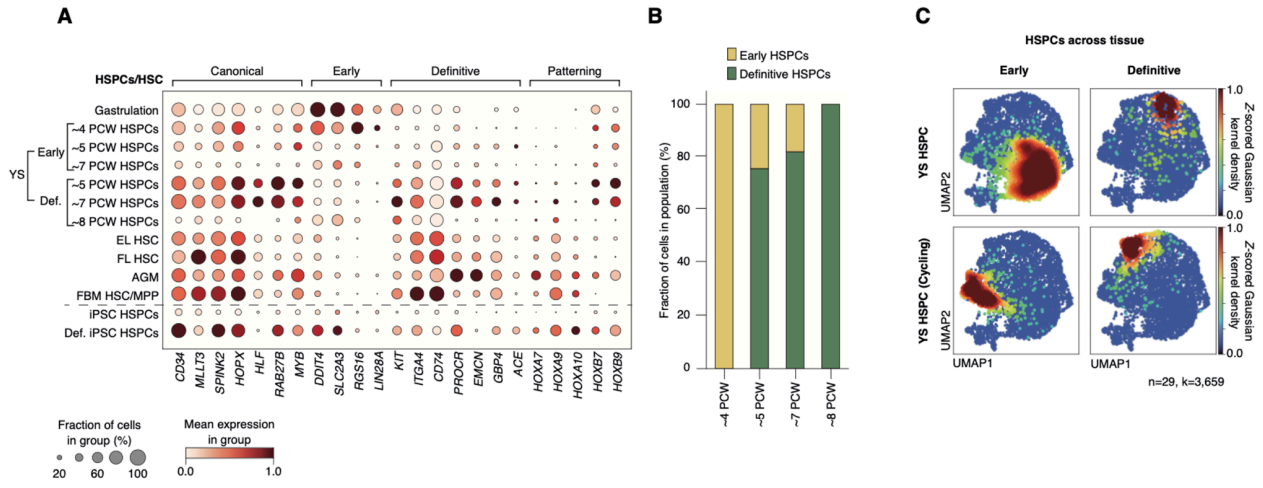


Figure 3.9: Early versus definitive hematopoiesis in YS and liver: *(A) Dot plot showing mean expression (color) and proportion of cells expressing selected HSPC genes (dot size) in HSPCs from YS (main and gastrulation (390)), liver (EL and fetal/FL (6)), AGM ((408), BM (3)) and iPSC cultures (iPSC (89) and definitive iPSC (46)). (B) Bar chart showing proportion of early (yellow) to definitive HSPCs (green) in the YS scRNA-seq data grouped by gestational age. (C) Density plots showing YS HSPC (top) and cycling HSPC (bottom) with early (left) and definitive signatures (right) in an integrated landscape as per A. Color: population z-scored KDE. Tissue contributions are shown in Fig. A5E. Figure adapted from Goh and Botting et al, 2023 (1).*

Next, we examined the transition from YS to liver hematopoiesis. Prior to AGM, the human EL is macroscopically pale, suggesting that erythropoiesis predominantly occurs in the YS (Fig. 3.10A). We tracked the proportional representation of hemoglobin (Hb) subtypes over time as a proxy for YS versus EL contributions. *HBZ* and *HBE1* (genes for Hb Gower 1) were restricted to YS erythroblasts, whereas *HBG1* (which forms fetal Hb/HbF in combination with an alpha chain) was expressed in fetal liver erythroblasts (*107*, *409–412*)(Fig. 3.10B and fig. A5H). The sustained *HBZ* production in YS for several days prior to liver bud formation (4PCW) was consistent with a scenario where YS supports initial erythropoiesis. At 7PCW, the EL contained both *HBZ* and *HBG1* (Fig. 3.10B), in keeping with previous studies of Hb switching (*397*). By 8PCW, EL erythroblasts expressed *HBZ*-repressors and were *HBG1*-dominant, as we have previously shown (*6*). By contrast, the mouse liver was macroscopically red prior to AGM maturation (Fig. 3.10A). Tracking Hb subtype usage in the mouse, we noted two waves of pre-AGM erythropoiesis: an initial wave with *Hbb-y* and *Hba*-x - *Hba-a1/2*), and a second wave mirrored in both YS and torso/liver (*Hbb-bt1* and *Hbb-bs*) (Fig. A5, G to I). Thus, there is a species-specific difference in YS erythropoiesis, and a rapid shift in Hb usage following AGM development in humans.

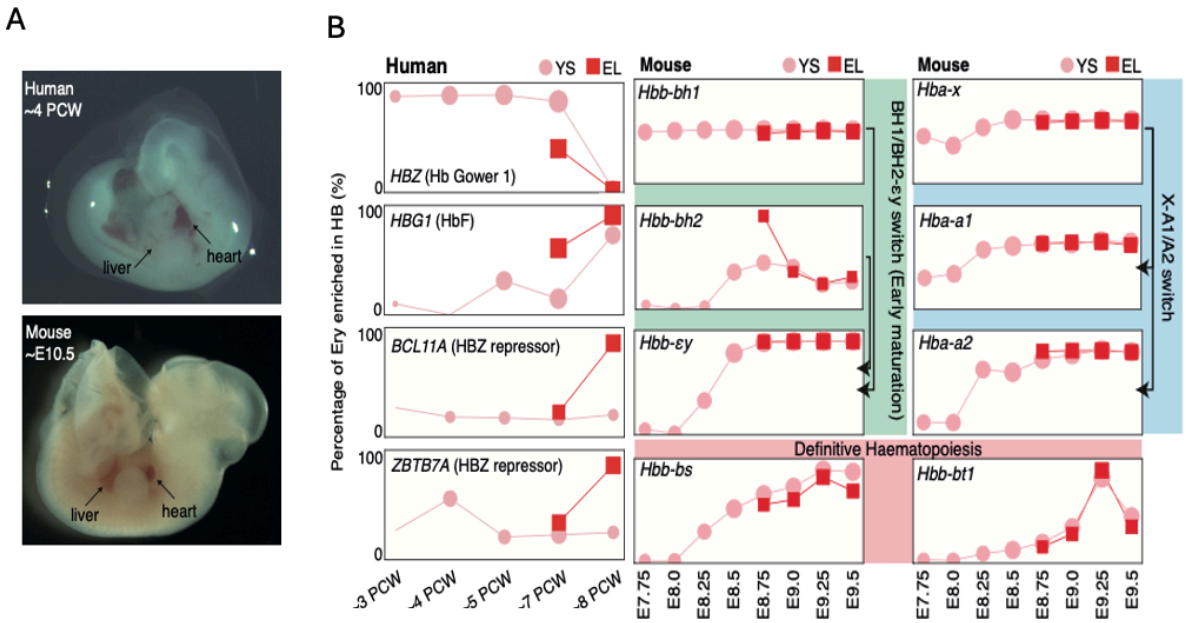


Figure 3.10: Mouse vs human erythropoietic waves: *(A) Representative image of whole ~4PCW/CS12 human (top; n=4) and ~E10.5/CS12 mouse embryo (bottom). Scale bars: 1 mm. (B) Left column: Line graphs showing the relative change in expression of Gower 1/2 globin HBE1, Gower 2 globins HBA1/2 and definitive globin HBG2 in human erythroid cells from YS (pink) and matched embryonic liver (EL) (red) over gestational age. Central and right columns: Globin expression in mouse erythroid cells (406) including HBB BH1, εγ, X, A2, BT1 and BS. Pink lines: mouse YS scRNA-seq data. Red lines: aged-matched mouse torso scRNA-seq data. Mouse hemoglobins implicated in primitive maturation, definitive hematopoiesis, and a switch between the two are grouped. The y-axis represents the proportion of erythroid lineage cells. Figure adapted from Goh and Botting et al, 2023 (1).*

We examined data from human gastrulation (~2-3PCW) and CS10-11 (~4PCW)—timepoints prior to AGM-HSPC formation—to explore the differentiation potential of early HSPCs. At gastrulation, the YS hematopoietic landscape had a tripartite differentiation structure, with erythroid, MK, and myeloid differentiation (Fig. 3.11). This structure was also observed in mouse YS (fig. A6, A and B).

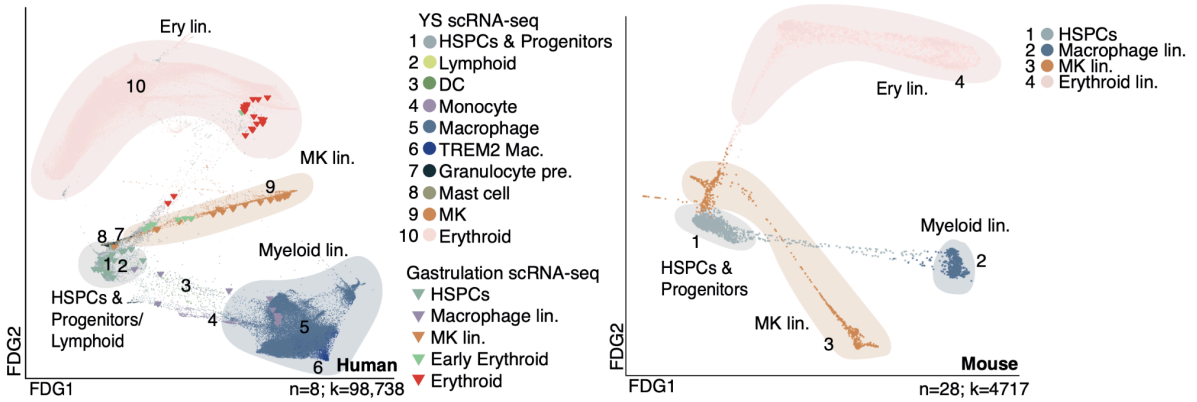


Figure 3.11: Human vs mouse YS haematopoietic products: *FDG of hematopoietic cell states in the YS scRNA-seq data (n=8, c=98,738; dots) integrated with human gastrulation (390) scRNA-seq data (n=1, c=91; triangles) (left), and equivalent cell states in the mouse gastrulation scRNA-seq dataset (406) (n=28, c=4,717; dots) (right). Colours represent cell states and clouds mark lineages. Figure adapted from Goh and Botting et al, 2023 (1).*

3.3.5 Hematopoietic waves in human yolk sac

Differential-fate-prediction analysis using the CellRank package (v1.5.1; described in methods section 2.3.7), predicted that early HSPCs pre-AGM at CS10-11 (~4PCW) were myeloid-biased, consistent with previous observations (*117*). However, the abundance of differentiating erythroid and MK cells at CS10-11 suggested that an earlier wave of erythroid/MK production had occurred (Fig. 3.12A and Fig. A6C). Post-AGM, the model predicted that remaining early HSPCs were erythroid and MK-biased, whereas definitive HSPCs were lymphoid- and MK-biased (Fig. 3.12A). This was in keeping with the first appearance of YS lymphoid cells (ILC progenitors, NK cells, and B lineage cells) post CS14 (Fig. 3.5B). Differential-fate-prediction analyses using CellRank, suggested that iPSC-derived HSPCs were embryonic erythroid-, myeloid-, and MK-biased, whereas definitive iPSC-derived HSPCs were lymphoid-, MK-, erythroid-, and myeloid-primed, consistent with

the predicted lineage potential of their in vivo early and definitive counterparts (Fig. 3.12B and fig. A6D).

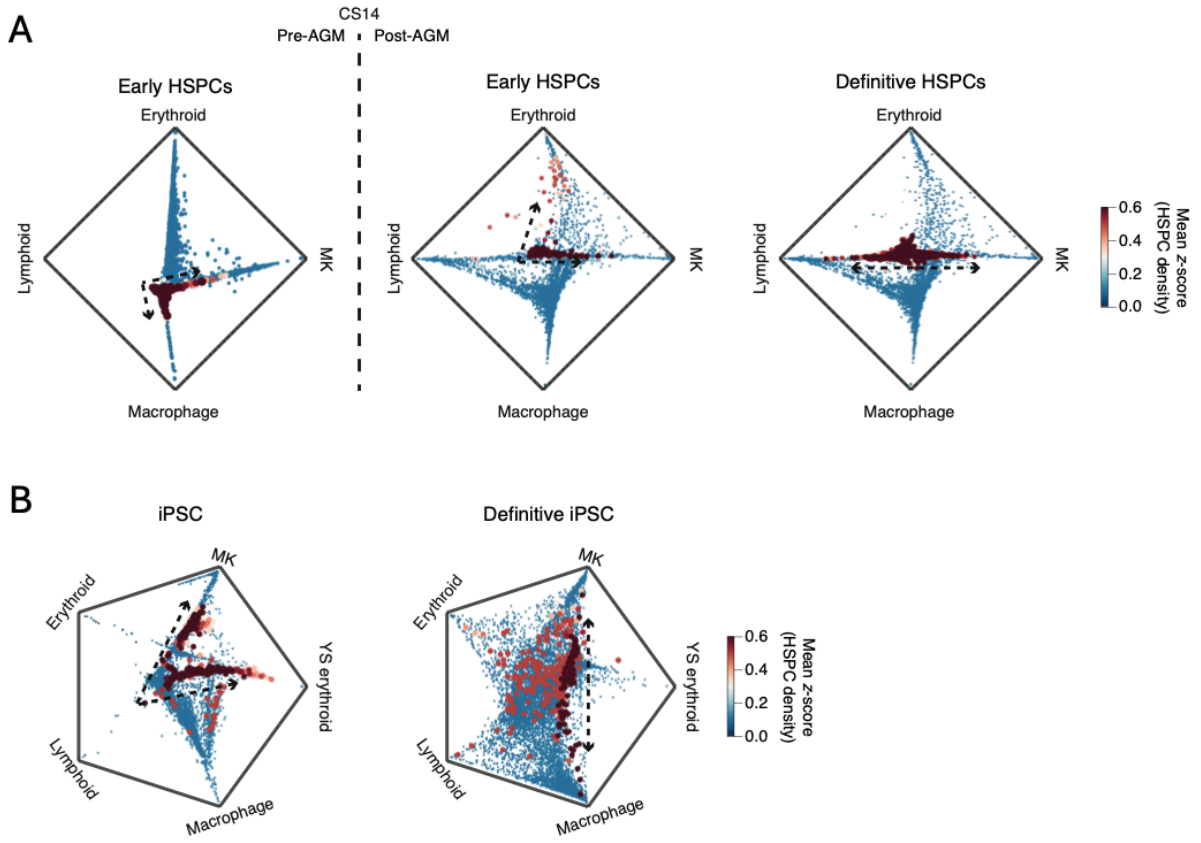


Figure 3.12: Human YS HSPC lineage transition probabilities: *(A) Radial plots showing lineage transition probabilities between YS early HSPCs, pre-AGM (CS10-11) (left) and YS early and definitive HSPCs, post-AGM (>CS14) (right). Colour indicates HSPC density as z-scored KDE, while position of HSPC population densities indicate respective lineage priming probability between macrophage, lymphoid (NK and B lineage), erythroid, and MK terminal states. Arrows indicate proposed lineage priming of HSPCs based on KDE. (B) Radial plots showing relative lineage transition probabilities between iPSC-derived HSPCs (left) and definitive iPSC-derived HSPCs (right). Interpretation as in A, with addition of embryonic erythroid terminal state. Figure adapted from Goh and Botting et al, 2023 (1).*

3.3.6 The lifespan of yolk sac HSPCs

HSPCs arise from HE in the aorta, YS, BM, placenta, and embryonic head in mice (*413–417*). In human AGM, definitive HSPCs emerge from *IL33*$^{+}$*ALDH1A1*$^{+}$ arterial endothelial cells (AECs) via *KCNK17*$^{+}$*ALDH1A1*$^{+}$ HE (*105*). Dissecting YS endothelial cell (EC) states in greater detail, the broad category of PLVAP^{+} ECs included AECs and HE, whereas LYVE1^{+} ECs encompassed sinusoidal, immature, and VWF-expressing ECs (Fig. 3.13, A and B; fig. A7, A and B). HE was a transient feature of early YS (Fig. 4A).

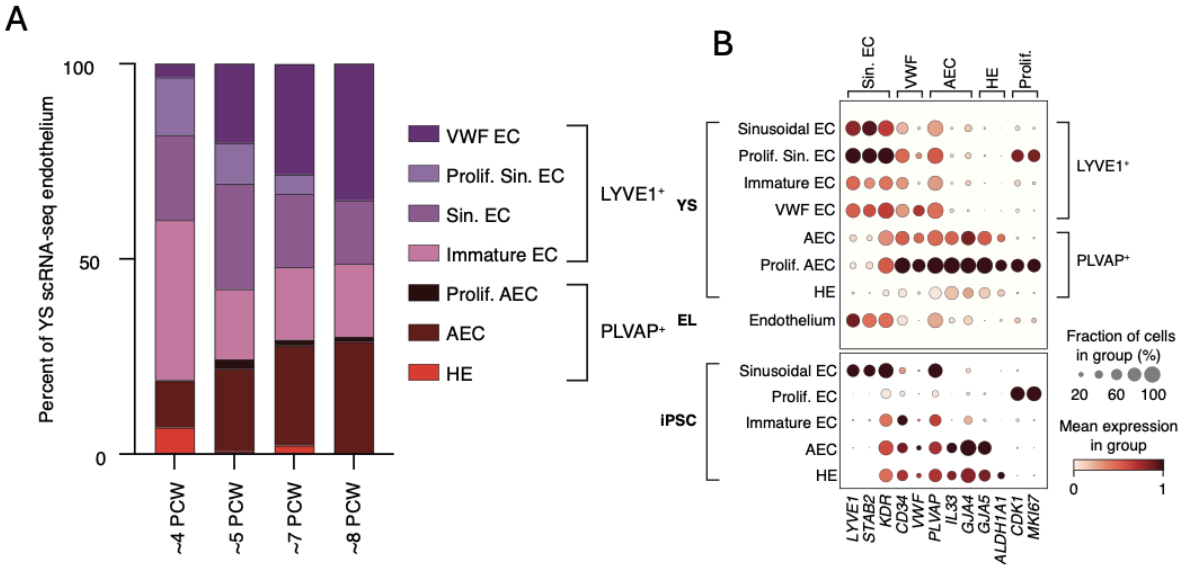


Figure 3.13: YS EC populations: *(A) Bar chart showing the relative proportions of YS endothelial cell (EC) subsets by age (PCW). EC: endothelial cells, AEC: arterial endothelial cells, Sin. EC: sinusoidal endothelial cells, and HE: hemogenic endothelium. (B) Dot plot showing the mean expression (colour scale) and the fraction of cells expressing each gene (dot size) of genes distinguishing endothelial cell subsets in YS (main), matched EL scRNA-seq and iPSC (20) scRNA-seq datasets. Data are min-max-standardised with a distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).*

Along inferred trajectories, YS HSPCs appeared to arise from AECs via HE as in AGM (*46*), sequentially upregulating expected genes such as *ALDH1A1* (*418*) (Fig. 3.14, A to C). The same EC intermediate states and transition points were identified in both iPSC culture systems (Fig. 3.14, A to C). In keeping with their more recent endothelial origin, we found that YS HSPCs and AGM HSPCs, but not EL or fetal BM HSPCs retained an EC gene signature characterised by the expression of *KDR*, *CDH5*, *ESAM*, and *PLVAP* (Fig. 3.14D).

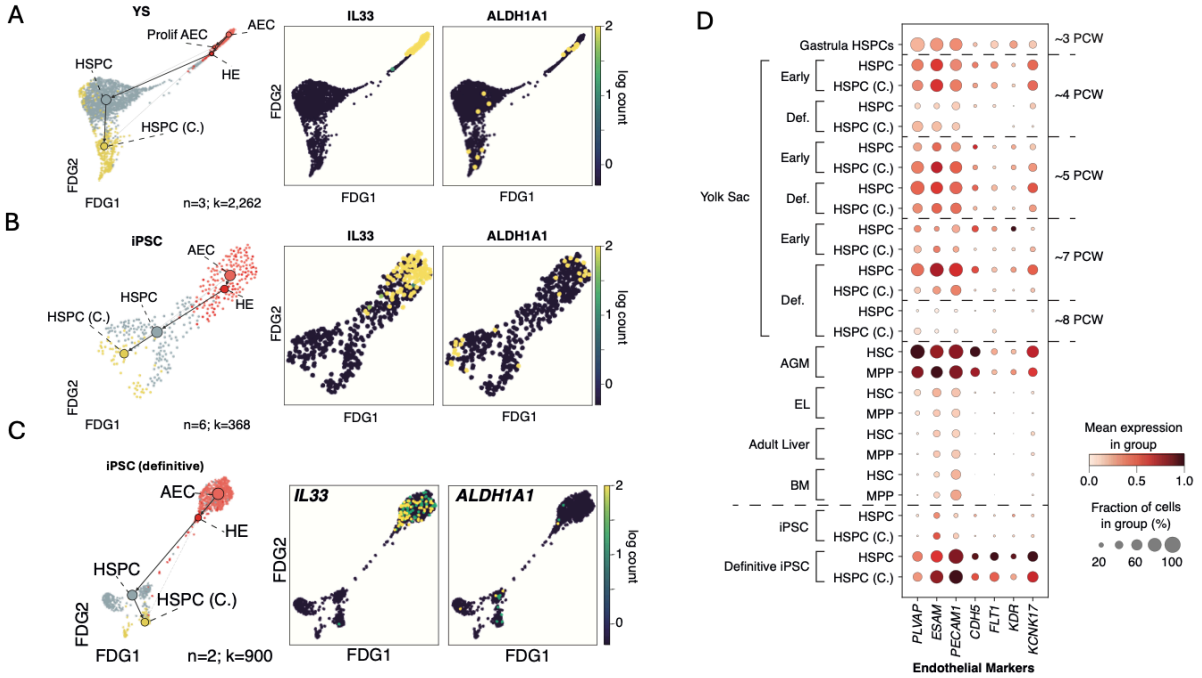


Figure 3.14: YS HSPC differentiation from HE: *(A) FDG overlaid with PAGA showing trajectory of HE transition to HSPC and HSPC (C.) (cycling HSPC) in YS scRNA-seq data (n=3; CS10, 11 and 14; c=2,262), (B) iPSC-derived HSPC scRNA-seq data (n=7, c=437)(89), and (C) Force directed graph overlaid with partition-based approximate graph abstraction (PAGA) map showing the trajectory of HE transition to HSPC in definitive iPSC scRNA-seq data (46) (n=3, c=2262), with feature plots of key genes (IL33, ALDH1A1) involved in endothelial to hemogenic transition. Dot plot showing the mean expression (color scale) and proportion of cells expressing EC-associated genes (dot size) in HSPCs across gestational age (PCW). HSPCs are derived from YS (including gastrulation), AGM (46), matched EL (embryonic liver), FL (fetal liver) (6), fetal BM (3), iPSC-derived HSPC (89) and definitive iPSC-derived HSPC (46) scRNA-seq datasets. Figure adapted from Goh and Botting et al, 2023 (1).*

Receptor–ligand interactions capable of supporting HSPC expansion and maintenance in YS were predicted using CellPhoneDB (*373*) and compared to independent predictions in fetal BM (adapted from Jardine et al. 2021 (3); Fig. A7I) (*3*). We identified YS ECs, fibroblasts, smooth muscle cells, and endoderm as likely interacting partners (Fig. 3.15, A to C). YS ECs, were predicted to maintain and support the HSPC pool (*419*) through the production of stem cell factor (*KITLG*) and NOTCH1/2 (*NOTCH1/2)*, although transcripts expressed diverged between tissues (*DLL1* and *JAG1* in YS and *DLK1*, *JAG1/2*, *NOV*, and *DLL4* in BM (Fig. 3.15A, and A7I). YS endoderm was predicted to support HSC pool expansion (*420*) through WNT5A (*WNT5A*) signaling to FZD3 (*FZD3*). *WNT5A* was also expressed by a wide range of BM stromal cell types, but BM HSPCs were predicted to respond via FZD6 (*FZD6*) rather than FZD3 (Fig. 3.15A, and A7I). All YS stromal fractions contributed to extracellular matrix, which provides a substrate for adhesion but also modifies HSPC function, with FN1 (*FN1*) from all fractions, potentially expanding the HSPC pool and Vitronectin (*VTN*) from endoderm, contributing to long-term HSC-like quiescence (Fig. 3.15, A and C) (*421*, *422*). Although BM HSPCs were also predicted to adhere to extracellular matrix proteins, the

integrins and matrix constituents differed. YS endoderm was predicted to form unique interactions with HSPCs via EPO (*EPO)*, which may influence the fate of differentiating progenitors (*423*), and THPO (*THPO)*, which supports HSC quiescence and adhesion in BM (*424*). No BM stromal source of *EPO* or *THPO* transcripts were detectable in our data however (Fig. A4H) (*3*, *6*). Thus, these anatomically different hematopoietic tissues use similar pathways to support HSPCs, albeit with tissue-specific components.

Predicted YS HSPC receptor to stromal ligand interactions diminished between CS17-CS23 (4-8PCW), including loss of cytokine and growth factor support and loss of TFGB1 (*TFGB1)*, WNT, and NOTCH2 signals (Fig. 3.15B; fig. A7E). In many interactions, there was reduction in HSPC receptor expression as well as stromal ligand expression (Fig. 3.15B; and fig. A7E), yet ligands were still expressed in age-matched liver and AGM stromal cells (fig. A7F). Adhesive interactions in YS were also predicted to be significantly modulated (fig. A7, F and G). Although aged-matched liver provided opportunities for adhesion with stromal cells, the AGM did not (fig. A7F). YS interactions gained between CS17 and CS23 included endoderm-derived *IL13* signalling to the *TMEM219*-encoded receptor implicated in the induction of apoptosis (Fig. 3.15A). Although limited conclusions can be made from studying cells that passed quality control for cell viability, we did observe upregulation in pro-apoptotic gene scores in late-stage YS HSPCs, both early and definitive (fig. A7H).

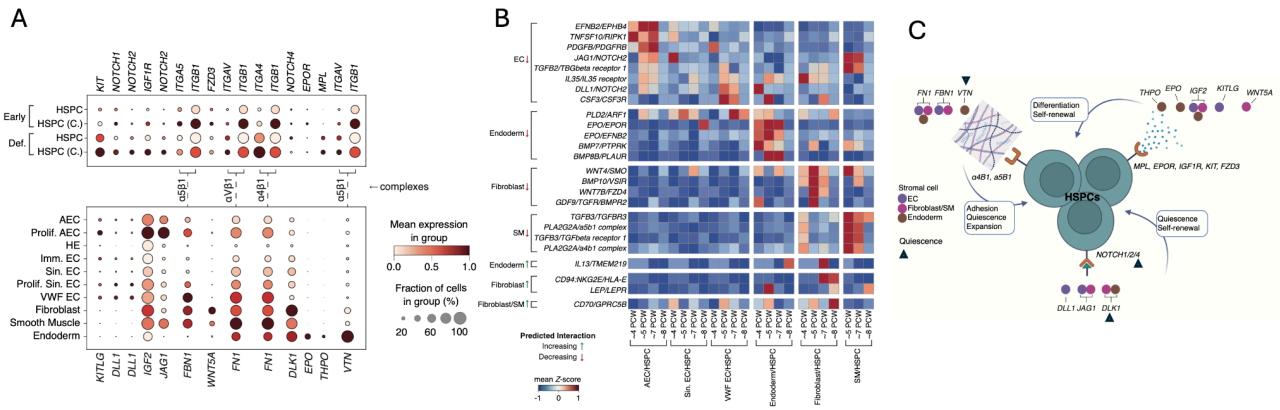


Figure 3.15: YS HSPC extrinsic regulators: *(A) Dot plot of the mean expression (colour scale) and the fraction of cells expressing each gene (dot size) of curated genes predicted by CellphoneDB to form statistically significant (P<0.05) protein–protein interactions between HSPCs (top plot) and stromal cells (bottom plot) across all time gestational points. Brackets indicate which protein counterparts form complexes. Data are log-normalised, variance-scaled, and min–max-standardised with a distribution of 0-1. (B) Heatmap showing curated and statistically significant (P<0.05) CellphoneDB-predicted interactions between YS HSPCs and stromal cells that change across gestation. The colour scale indicates relative CellphoneDB-predicted interaction strength as Z-scores, derived from averaged ligand and receptor gene expression across interacting cell populations. (C) Schematic of selected and statistically significant (P<0.05) CellphoneDB-predicted interactions between YS HSPCs and endoderm, fibroblasts (Fib), smooth muscle cells (SMC), or EC derived from scRNA-seq data. Interactions are grouped by predicted receptor to ECM interactions, ligand–receptor interactions, and surface-bound ligand–receptor interactions. Functional annotations delineate known biological roles of receptor–ligand interactions, ligands/receptors which may induce quiescence are indicated with a blue triangle. Receptors and ligands in italics significantly decrease at CS17-23 (6-8PCW). Figure adapted from Goh and Botting et al, 2023 (1).*

Despite marked change in the stromal environment of later stage YS, the proportion of HSPC to cycling HSPC remained stable (fig. A5F). Differential lineage priming analysis revealed that very few HSPCs remained in CS22-23 (8PCW) YS and most cells were terminally differentiated (fig. A6C). Thus, it is likely that an early burst of early HSPC production arises from transient YS HE, a later influx of definitive HSPCs derives from AGM, and a loss of

stromal support between 6-8PCW, results in apoptosis and depletion of remaining HSPCs by terminal differentiation.

3.4 Chapter 1 discussion

This study presents the first comprehensive single-cell multiomic atlas of human embryonic YS hematopoiesis, employing simultaneous transcriptomic and proteomic profiling techniques. It provides insights into the emergence and differentiation of the first YS HSPCs from specialised HE. By integrating our temporal snapshots of YS-derived cell states with publicly available scRNA-seq references, a major strength of our approach lies in enabling direct comparisons across organs and developmental windows. The results delineate the rapid establishment of an early blood and immune cell repertoire, including erythroid, megakaryocyte, and myeloid cells, generated within the YS during the initial weeks of development.

One significant discovery from our study is the spatio-temporal switch in globin gene expression between the YS and liver in human embryos. Our data highlight the transition from embryonic (*HBZ, HBE1*) to fetal and adult globin gene expression (*HBA1/2, HBG*1/2) in humans (Fig. A5H), which contrasts with the distributed erythropoietic waves observed in murine models. This divergence underscores species-specific differences in early erythropoiesis, an important consideration since much of our understanding of primitive hematopoiesis stems from model organisms such as mice. The regulatory factors *ZBTB7A* and *BCL11A*, inferred to control embryonic *HBZ* gene expression and initiate the switch to *HBA* globin production, are of particular interest. Given the evolutionary paralogous relationship between *HBZ* and *HBA*, these factors could serve as potential therapeutic targets for alpha thalassemia. Reactivating *HBZ* expression in patients with defective *HBA* genes could potentially compensate for the loss, alleviating the severity of the disease. This strategy presents a novel therapeutic avenue for managing severe alpha thalassemia. We discuss potential strategies in more detail in the overall discussion chapter. However, while our integrated dataset broadly spans important developmental periods, the specific window capturing the human embryonic-to-fetal globin switch remains narrowly represented, with relatively few developmental stages, tissues, and donor samples included (Fig. 3.2). This limited representation may not adequately reflect rare transitional states or inter-individual variation occurring precisely at the time of the globin switch. Moreover, although simultaneous scRNA-seq and CITE-seq multiomic profiling robustly captures cell identities and states, it does not directly quantify subcellular protein-level expression (e.g., HBZ, HBG1) or validate mechanisms of transcription factor regulation (e.g., transcriptional repression by BCL11A and ZBTB7A). Comprehensive validation would require additional

assays targeting intracellular proteins and chromatin states, such as intracellular CITE-seq and ATAC-seq, and immunofluorescence imaging. Thus, while the globin switch observations are compelling, additional experiments are needed to validate whether these transitions can be therapeutically manipulated in pathology (e.g., alpha thalassemia).

This study also delineates the stromal support required for YS HSPC maintenance, showing the crucial interactions between endothelial cells, fibroblasts, smooth muscle cells, and endoderm in supporting the HSPC pool. These inferred interactions include key receptor-ligand pairs such as *KITLG* and *NOTCH1/2,* which are vital for HSPC expansion and maintenance, and require further proteomic profiling for full validation (e.g, CITE-seq and immunofluorescence quantification for ligand and receptor protein presence). Understanding these interactions at a detailed level provides insight into how the YS creates a supportive niche for HSPC development and may help us design more effective biomimetic culture systems to more faithfully recapitulate YS haematopoiesis *in vitro*.

Furthermore, though incremental, this research highlights for the first time the evolutionary conservation of YS differentiation restriction in both human YS and mouse gastrulation, which results in the production of myeloid, erythroid, and megakaryocyte populations, but notably, no lymphoid populations prior to AGM establishment. This indicates a conserved mechanism for YS generation of specific lineages necessary for early development, while the lymphoid lineage production is deferred to later stages.

We also characterise the pre-AGM temporal switches between bipartite differentiation profiles of erythropoiesis and myelopoiesis, myelopoiesis and megakaryopoiesis in the YS, with lymphoid lineage differentiation capacity only observed after AGM formation. This delineates the YS role in producing vital blood and immune cells with temporal resolution. By profiling these transitions, we provide a detailed map of how hematopoiesis transitions from the YS as the primary hematopoietic organ to the liver. In particular, we propose that further multi-omic comparisons with ex vivo and iPSC-derived models may uncover precisely how the YS environment orchestrates these early waves. However, as with the globin switch, our dataset provides relatively few snapshots specifically around these pre-AGM transitions. Capturing additional developmental timepoints, coupled with assays targeting the epigenetic mechanisms (e.g, ATAC-seq) orchestrating distinct YS hematopoietic waves and transition to liver hematopoiesis, will be essential to fully understand how the embryo governs early lineage specification and spatiotemporal hematopoietic transitions (e.g., from the YS to liver).

Finally, for the first time, we demonstrate that the YS pre-specifies a microglia-like *TREM2*+ macrophage population. This finding will be further explored in Results Chapter 3.

4 Results chapter 2: Multiorgan functions of human YS

The work outlined in this chapter is the result of collaboration as described in chapter 1 introduction.

IHC imaging experiments were organised by Dr Rachel Botting with assistance from myself. The experiments and imaging were carried out by Rowen Coulthard (Translational and Clinical Research Institute, Newcastle University), and Dr Meghan Acres (Previously a member of Haniffa lab Biosciences Institute, Newcastle University). Mouse and Rabbit data were mapped to a reference genome, and provided by the Göttgens lab (Department of Haematology, Wellcome-MRC Cambridge Stem Cell Institute).

This chapter is a lightly-edited version of a manuscript which I co-first authored (*1*). I assisted in planning and organising the generation of the IHC validation data after identifying key markers in scRNA-seq and CITE-seq data corresponding to hypothesised functions of the YS. Additionally, I performed all analyses for this section including identifying, annotating and visualising conserved and differential gene programs expressed between stromal states of the YS and EL. I also performed the cross-species sequence homology-based integration, transfer learning, and label transfer between human and mouse YS and gastrulation scRNA-seq data.

4.1 Introduction

In this chapter I describe the non-hematopoietic functions of the human embryonic YS, vital for providing nutritional and metabolic support for the early embryo.

4.1.1 YS stromal cell-states and their functions

The YS has the capacity to uptake, transport, and metabolise nutrients in both mouse and human YS, highlighting its fundamental role in vertebrate embryonic development .

The human yolk sac (YS) is a multifunctional organ crucial for early embryonic development, serving as the initial site for haematopoiesis by producing the first blood and immune cells, including erythroid, megakaryocytic, and myeloid lineages. Beyond its pivotal role in blood cell formation, the YS is integral to several biosynthetic, metabolic, and nutritional functions. It facilitates the transfer of nutrients to the developing embryo, producing cholesterol and macromolecule transport proteins, which are essential for nutrient delivery within the extraembryonic space (*101*). These transport mechanisms ensure that vital nutrients are readily available to support the rapid growth and development of the embryo underscoring the multifaceted role the YS plays in supporting embryonic development through both nutritional and hematopoietic mechanisms.

Despite its importance, many vital functions of the YS remain unexplored in humans, warranting further research into its complex roles during early development.

4.2 Methods

4.2.1 Cross species probabilistic projection and label transfer

In this chapter we devise an implementation of the EN workflow described in the methods section for probabilistic label transfer and transfer learning between human YS scRNA-seq data and mouse gastrulation data. This was used to create cross-species label assignments used for the cross-species comparisons seen in this chapter including the evolutionary conserved gene program analysis.

An implementation of the SAMap (Self-Assembling Manifold mapping) workflow (v1.0.7) (*391*) was used to construct a gene–gene sequence homology graph weighted by human and mouse sequence similarity. Reciprocal BLAST mapping was then performed between the entire mouse and human transcriptomes for the detection of significant homology between sequences of each species. The resultant SAM object returned three hundred species-stitched PC components for the top 3000 paired genes. The human subset of this cross-species PC embedding was then used as reference for input into the EN logistic regression workflow we describe in methods. Cluster labels in mouse data were assigned by majority voting of predicted human YS labels with manual checking of key markers (see methods).

4.2.2 Cross species clustered gene-set enrichment analysis

To identify gene programs that were conserved between human and mouse YS, we implemented a modified version of the clustered gene-set enrichment analysis method described in methods. To adequately align and interpret enriched genesets from this analysis between species, only strict 1:1 homologs between human and mouse were considered. This decision was based on the rationale that 1:1 homologs are more likely to retain conserved biological functions across species, thereby allowing more accurate inference of shared gene programs. Expanding to include all orthologs (including 1:many or many:many relationships) would increase gene coverage but at the cost of functional ambiguity, as genes with shared ancestry may have undergone evolutionary divergence, leading to distinct roles in each species. Thus, this approach prioritised functional specificity over gene-space coverage, with the aim of reducing false-positive attributions of conserved gene programs due to non-equivalent ortholog mapping. While this conservative strategy may exclude certain relevant genes that have no clear 1:1 homolog, potentially omitting some species-specific biology, this trade-off was considered acceptable to ensure robustness in interpreting cross-species conservation of gene sets. We ranked conserved markers between the endoderm cell state in YS scRNA-seq data against endoderm in the mouse gastrulation scRNA-seq data using the FindConservedMarkers function in Seurat (v3.1) with Bonferroni corrected FDR adjusted P-values. Markers were submitted for gene set enrichment ranking and analysis using the Enrichr tool as implemented in the GSEApy (v1.0) package to query the Gene Ontology (GO) Biological Process database (GO_BP_2022). Please see the methods section for the downstream processes including the graph embedding and markov clustering of the resultant GSEA outputs which are identical from this point to the run in human data.

4.3 Results

4.3.1 YS stromal cell-states and their functions

The YS stroma consists of five main cell type populations, the Endothelial Cells (ECs), mesothelium, endoderm, fibroblasts, and smooth muscle (as previously described in results chapter 1). During embryonic development, the YS stromal landscape undergoes significant changes. The endoderm population shrinks significantly by 8PCW, while the fibroblast population increases in proportion at the same time (Fig. 3.5B). This period coincides with the involution of the human YS structure, which is particularly interesting as we have previously identified extrinsic modulators of hematopoietic stem and progenitor cell (HSPC) maintenance produced by the endoderm.

In our data, YS endoderm coexpressing *APOA1/2*, *APOC3*, and *TTR*, was present from gastrulation at ~2-3PCW (*390*) (Fig. 4.1 ; fig. A4D). Strikingly, this transcriptomic profile of YS endoderm closely resembled that of fetal and embryonic liver hepatocytes, which were previously characterised by Popescu *et al.* in 2019 (*6*). YS endoderm expressed higher levels of serine protease 3 (*PRSS3*), glutathione S-transferase alpha 2 (*GSTA2*), and multi-functional protein galectin 3 (*LGALS3*), compared to EL hepatocytes, whereas hepatocytes expressed a more extensive repertoire of detoxification enzymes, including alcohol and aldehyde dehydrogenases and cytochrome P450 enzymes (fig. A4D).

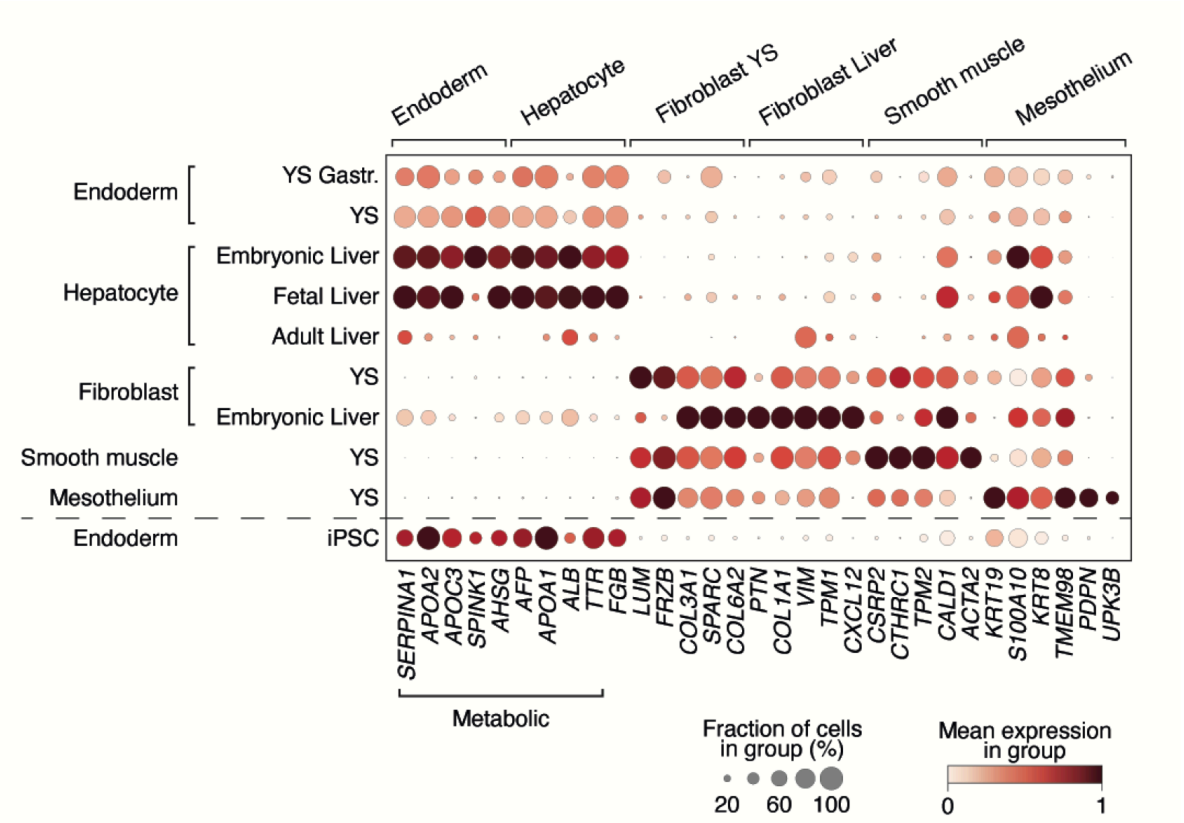


Figure 4.1: YS stromal cell-states and their functions: *Dot plot showing the expression level (colour scale) and percent expression (dot size) of DEGs in YS (main and gastrulation (gastr.) data) stromal cell states, matched embryonic, fetal and adult liver stromal cell states and iPSC stromal cell states (89) (Table S3, 20, 7). All datasets independently scaled to max value=10 and then combined, except YS and matched EL scRNA-seq data which were scaled together. Genes involved in endoderm metabolic function are grouped. Figure adapted from Goh and Botting et al, 2023 (1).*

4.3.2 YS endoderm functional programs

Given the striking similarity in gene signatures between YS endoderm and EL hepatocytes, we sought to determine whether these cell states also shared functional gene modules which would indicate conserved functional roles across these tissue compartments. Our analyses identified both conserved and differential gene modules enriched in both endoderm and hepatocytes including gene modules representing metabolic, nutrient transport, and coagulation functions (Fig. 4.2).

Both Endoderm and liver hepatocyte cell states shared gene modules implicated in coagulation and lipid metabolism (Fig. 4.2), which were also conserved in mouse and rabbit extraembryonic endoderm (fig. 4.5A to C; A4F and G). Crucially, our analysis also identified differential gene modules for metabolic preferences, with glycolytic processes being predominant in the YS endoderm, and acetyl-CoA metabolism, indicative of oxidative phosphorylation preference, in hepatocytes. This could underscore specialised bioenergetic and metabolic adaptations of each cell type to their respective developmental tissue contexts.

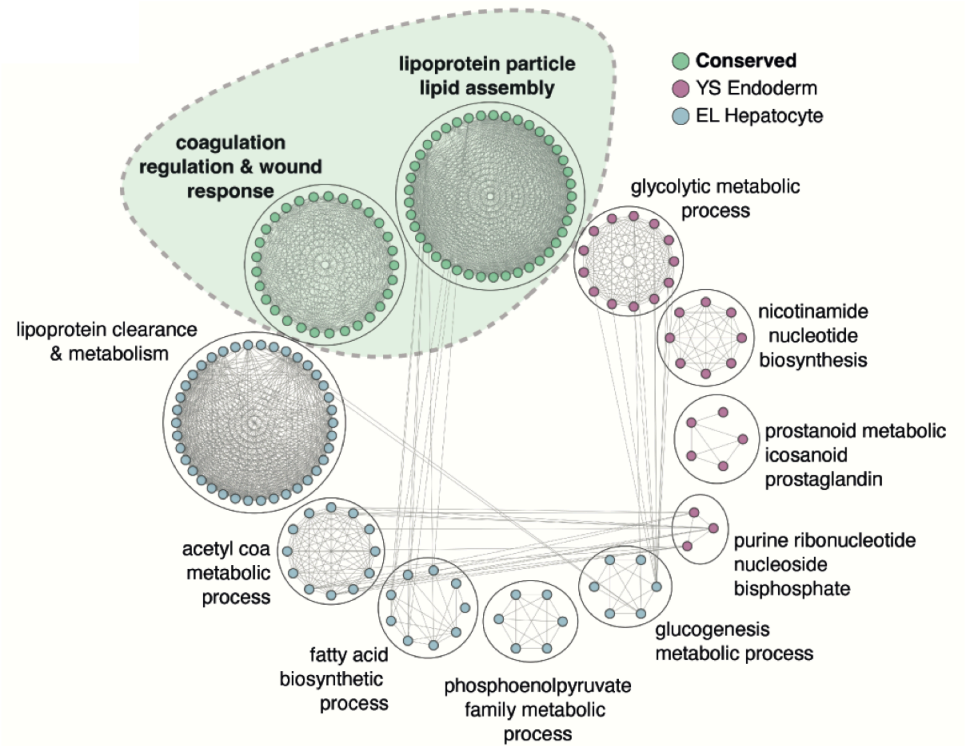


Figure 4.2: Conserved and gene expression programs between YS and liver: *Clustered gene set analysis plot of the significant pathways upregulated in YS endoderm (pink), embryonic liver (EL) hepatocytes (blue) and conserved across both tissues (green). Lines indicate connected nodes of expression (Table S21). Figure adapted from Goh and Botting et al, 2023 (1).*

The expression of transport proteins (alpha-fetoprotein and albumin), a protease inhibitor (alpha-1-antitrypsin), erythropoietin (EPO), and coagulation proteins (thrombin, prothrombin, and fibrin) were validated in human YS endoderm and EL hepatocytes (Fig. 4.3 and fig. A4G).

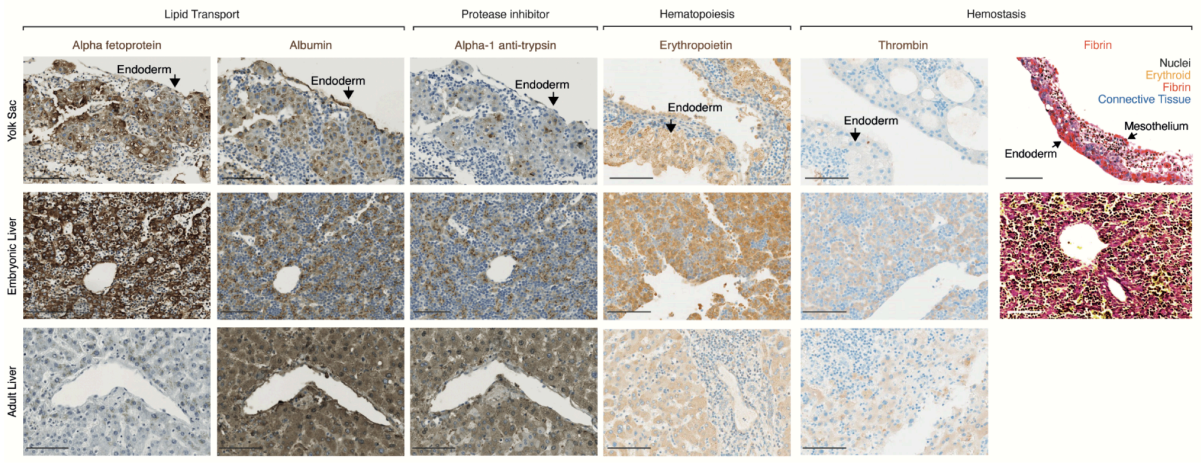


Figure 4.3: IHC images of YS endoderm functions: *Left and middle: IHC staining of alpha fetoprotein (AFP), albumin (ALB) and alpha-1 antitrypsin (SERPINA1) in 8PCW YS, 8PCW EL and healthy adult liver. Representative images from 1 of n=5 biological independent YS (4-8PCW), 1 of n=3 biologically independent ELs (7-8PCW) and 1 of n=3 biologically independent adult livers. Scale bar=100µm. Right: MSB-stained 8PCW EL (representative of n=3 biologically independent samples) and 4PCW YS (representative of n=3 biologically independent samples). Nuclei (grey), erythroid (yellow), fibrin (red), and connective tissue (blue). See 'Immunohistochemistry' section in Methods for details regarding pseudo-colouring shown. Scale bars=100um. Figure adapted from Goh and Botting et al, 2023 (1).*

From the earliest timepoints, YS endoderm expressed genes for anticoagulant proteins antithrombin III (*SERPINC1*) and protein S (*PROS1*) and components of the tissue factor-activated extrinsic coagulation pathway—thrombin (*F2*), factor VII (*F7*), and factor X *(F10)* (Fig. 4.4), confirmed at the protein level for thrombin (Fig. 4.3). Intrinsic pathway

factors VIII, IX, XI, and XII (*F8*, *F9*, *F11*, and *F12*) were minimally expressed in YS, but were expressed by EL hepatocytes (Fig. 4.4). Tissue factor, antithrombin III and fibrinogen subunits were also expressed in mouse extraembryonic endoderm and rabbit YS endoderm (fig. A4E). Embryonic lethality of homozygous-null mice lacking prothrombin, thrombin, and coagulation factor V prior to liver synthetic function (i.e., at E9-12) implies functional relevance of YS expression (Fig. 4.4) (*425*, *426*), whereby coagulant and anticoagulant pathways develop in parallel to balance hemostasis.

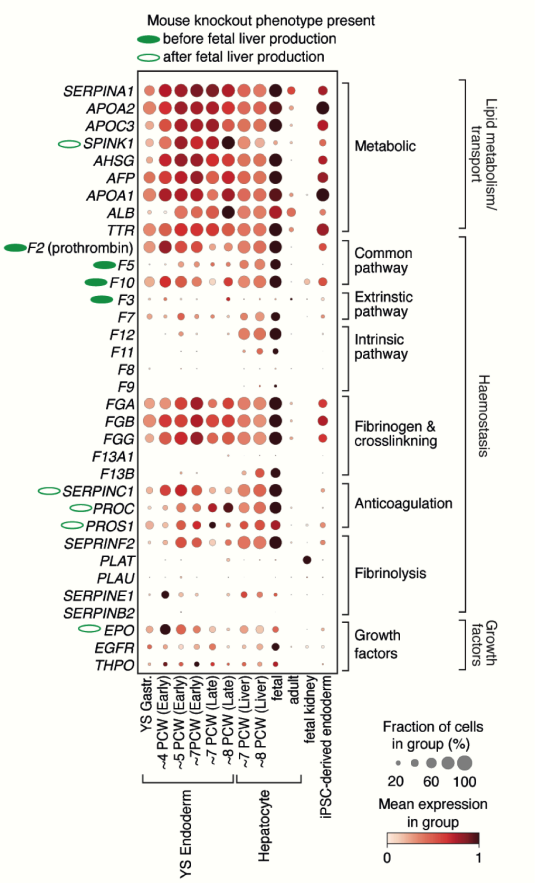


Figure 4.4: YS endoderm haemostasis functions: *Dot plot showing the expression level (colour scale) and percent expression (dot size) of haemostasis factors expressed by YS endoderm (main and gastrula data), embryonic, fetal and adult liver hepatocytes, and endoderm from fetal kidney (55). Grouped by pathway and role. All datasets independently scaled to max value=10 and then combined, except YS and matched EL scRNA-seq data which were scaled together. Figure adapted from Goh and Botting et al, 2023 (1).*

4.3.3 Evolutionarily conserved functions of the YS endoderm

YS endoderm cells expressed EPO and THPO that are critical for erythropoiesis and megakaryopoiesis (Fig. 4.3 and 4.4; fig. A4H). In mouse development, EPO is produced by fetal liver and is only essential for definitive and the later stages of primitive erythropoiesis, with *Epo*/*Epor*-knockout mice dying at around E13 (*427*). An EPO source prior to liver development is therefore likely not needed in mice. Accordingly, EPO has not been found in mouse YS (*428*) (Fig. 4.5A). In parallel to human YS, rabbit YS endoderm also produced EPO at gestational stages preceding liver development (Fig. 4.5A). Cross-species conserved gene set analysis reveals gene modules implicated in coagulation and lipid metabolism are evolutionarily conserved between human and mouse YS endoderm (Fig. 4.5 B).

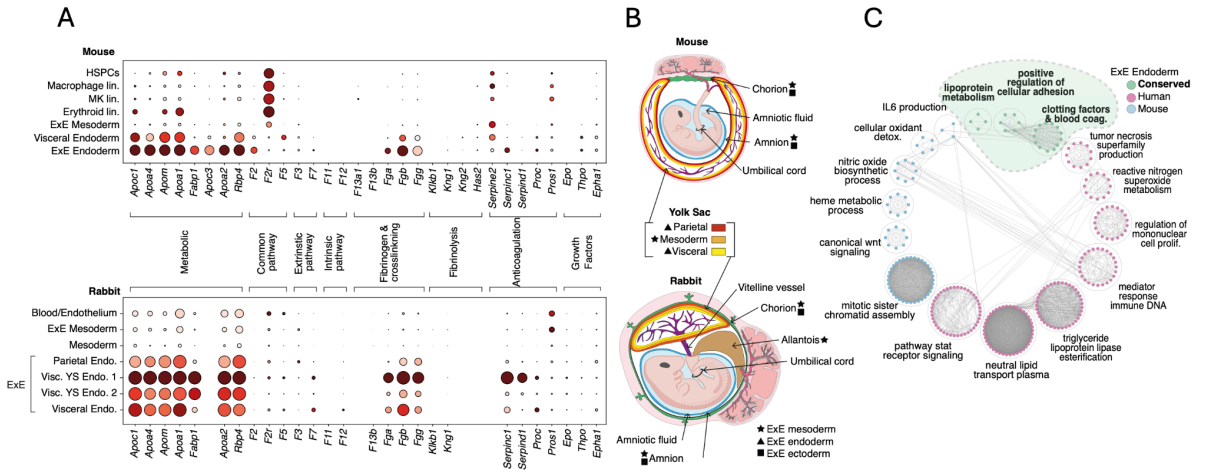


Figure 4.5: Evolutionary conserved functions of YS endoderm: *(A) Dot plots showing the mean expression (color scale) and the fraction of cells expressing each gene (by dot size) of clotting and soluble factors in relevant cell states from mouse gastrulation scRNA-seq data (406) (top) and from rabbit scRNA-seq data (108) (bottom). Data are min-max-standardised with a distribution of 0-1. (B)illustrations of developing mouse (~E9.5) and rabbit (~GD9) embryos. Text legends indicate corresponding extraembryonic anatomical regions between species, whereas shapes indicate germ layer origin of the anatomical regions. Star: mesoderm; triangle: endoderm; and square: ectoderm. Layers of the YS are delineated by color. Red: parietal; orange: mesoderm; and yellow: visceral. (C) Flower plot of the significant gene sets enriched in YS endoderm (pink), mouse extraembryonic (ExE) endoderm (blue), and conserved between species (green). Nodes indicate significantly enriched gene sets (Q-value <0.05), whereas edges between nodes represent gene overlap between gene sets. Annotated grouping circles indicate Markov cluster neighbourhoods of gene expression modules which share high gene-set similarities. Figure adapted from Goh and Botting et al, 2023 (1).*

4.3.4 YS loss of endodermal metabolic function over time

To better understand and contextualise both the unique and conserved functional roles of the human embryonic YS, we compiled a 12-organ integrated human fetal atlas spanning 3-19PCW (c=3.12 million, n=150; Fig. A8, G and H). We observed that EPO and THPO production were restricted to YS and liver (fig. A4H), specifically to YS endoderm and liver hepatocytes (fig. A4I). Differentially expressed genes between early and late YS endoderm

revealed active retinoic acid and lipid metabolic processes until 7PCW, after which genes associated with cell stress and death were expressed (Fig. 4.6).

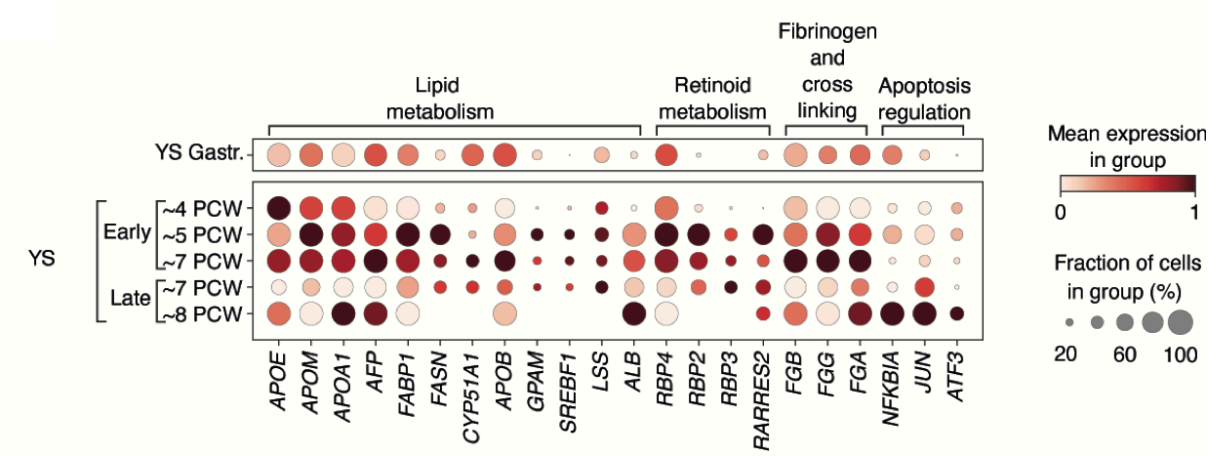


Figure 4.6: YS loss of endodermal function over time: *Dot plot showing the mean expression (color) and proportion of cells expressing Milo-derived DEGs across gestation (dot size) in YS endoderm. Genes are grouped by function. Figure adapted from Goh and Botting et al, 2023 (1).*

A decline in the proportion of YS endoderm cells producing EPO and the general decline of YS endoderm proportions (Fig 3.4B) was compensated by onset of EPO production by hepatocytes at 7PCW (fig. A4J). Thus, the human YS plays a critical role supporting hematopoiesis, metabolism, coagulation, and erythroid cell mass regulation before these functions are taken over by the embryonic/fetal liver, and then ultimately, by the adult liver (metabolism and coagulation), bone marrow (BM) (hematopoiesis), and kidney (erythroid cell mass regulation) (Fig. 4.7).

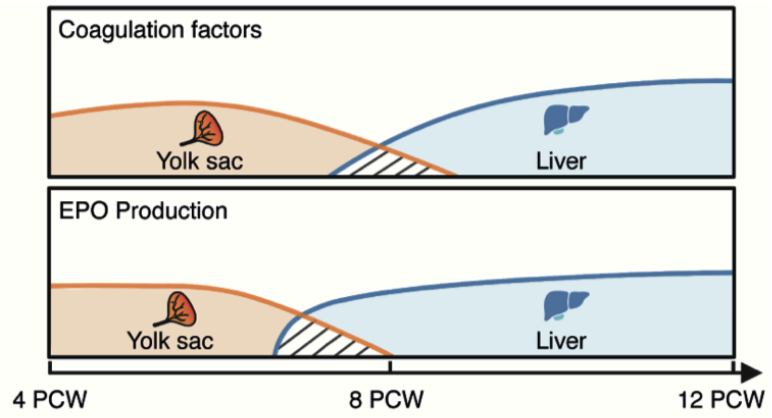


Figure 4.7: Timeline of YS contributions to coagulation and EPO: *Schematic of the relative contributions of YS (orange), liver (blue), to coagulation factor, and EPO production in the first trimester of human development. Figure adapted from Goh and Botting et al, 2023 (1).*

4.4 Chapter 2 discussion

This chapter offers a pioneering single-cell transcriptomic evaluation of the human embryonic yolk sac (YS) stroma, elucidating its multifunctional roles beyond haematopoiesis, specifically in nutritional, erythropoietic, and metabolic support during early development. By leveraging techniques such as cross-species probabilistic projection, label transfer, and clustered gene-set enrichment analyses, we identified conserved and differential gene programs between human and mouse YS, highlighting the evolutionary significance of the embryonic YS. Our comprehensive single-cell RNA sequencing (scRNA-seq) analysis revealed five main stromal cell populations in the human YS: endothelial cells, mesothelium, endoderm, fibroblasts, and smooth muscle cells. Notably, dynamic changes in these populations, particularly the reduction of endoderm and the increase in fibroblasts around 8 post-conception weeks (PCW), coincide with structural changes in the YS as it begins to involute. This observation underscores the critical role of the endoderm in producing modulators of HSPC maintenance.

Our data demonstrate that YS endoderm expresses genes crucial for metabolic processes, including lipid and glucose metabolism, akin to hepatic functions. The presence of transport proteins like alpha-fetoprotein (*AFP*) and albumin (*ALB*), and coagulation factors such as thrombin and fibrin, highlights the YS role in supporting hemostasis and nutrient transport. These findings are validated by immunohistochemistry (IHC) and align with known functions of EL hepatocytes. The cross-species analysis reveals that gene modules implicated in coagulation and lipid metabolism are conserved between human and mouse YS endoderm. This evolutionary conservation underscores the fundamental roles of the YS in early development across vertebrates. Interestingly, erythropoietin (EPO) production is identified in human and rabbit YS but not in mouse, indicating species-specific adaptations in YS functions. This suggests that the rabbit, with its greater gestational similarity to humans, may be a more appropriate model for understanding the role of the YS in human development (*108*). However, the utility of rabbit models is partially constrained by the incomplete annotation of the rabbit genome, limiting detailed genetic and molecular comparisons. Despite these insights, our study is limited by the dynamic nature of YS involution and challenges associated with sample availability at precise gestational intervals (Fig 3.2). As a result, transient cell states may remain undersampled. Additionally, while scRNA-seq strongly suggest diverse endodermal functions (e.g., EPO production, lipid transport), definitive

validation, such as functional knockdown of putative regulatory genes or in vivo assays, is necessary to confirm that these transcripts directly translate into embryonic nutrient supply and erythrocyte generation. Future experiments employing spatial transcriptomics or targeted proteomics could further clarify the compartmentalization and functional roles of these stromal subsets within the YS environment.

Our integrative analysis across multiple developmental stages shows a temporal decline in YS endoderm functions, particularly in EPO production, which shifts to liver hepatocytes by 7PCW. This functional transition from YS to liver, and eventually to adult organs like bone marrow and kidneys, highlights the critical but transient role the YS plays in early development. The timeline illustrates the progressive handover of haematopoietic, metabolic, and coagulation functions from the YS to other organs. Insights gained from this study lay a foundation for understanding the multifunctional roles of the human YS and its evolutionary conservation. The identification of conserved gene programs offers potential targets for further research into developmental biology and evolutionary studies. Future studies employing spatial transcriptomics and advanced imaging techniques could provide deeper insights into the architectural and functional dynamics of the YS and its interactions with other developing organs.

Our findings emphasise the YS significance beyond haematopoiesis, highlighting its critical roles in erythropoietic and metabolic support, and nutrient transport. This comprehensive analysis sets the stage for future investigations into the complex interplay of cellular functions during early human development.

5 Results chapter 3: Macrophage ontogeny and functions across human life

The work outlined in this chapter is the result of collaboration as described in chapter 1.

This chapter is a lightly-edited version of a manuscript which I co-first authored (*1*). I performed all the analyses for this section, including assembling the 12-organ fetal developmental atlas and training the EN classifiers for harmonisation and probabilistic comparisons of YS cell states within the context of the 12-tissue reference developmental atlas.

5.1 Introduction

In this chapter, I describe the myelopoietic functions of the human embryonic yolk sac (YS) and the potential contributions of YS-derived myeloid progenitors to TRMs during development.

5.2 Methods

5.2.3 Assembling 12 tissue pan organ developmental atlas

To appropriately contextualise cell states identified in our YS scRNA-seq data, we aimed to compile a comprehensive atlas of human developmental data, including key sites involved in the cascade of haematopoietic functions, such as the AGM, liver, and bone marrow. Additionally, we sought to integrate data from tissues known to receive contributions from YS myeloid progenitors, such as specific macrophage populations in the skin and brain.

In total, we integrated 3.12 million cells from 150 donors. The organs included in this study are as follows: YS with 10 donors and 169,494 cells; AGM with 4 donors and 12,248 cells; skin with 13 donors and 178,563 cells; brain with 72 donors and 2.16 million cells; gonads with 44 donors and 14,244 cells; thymus with 11 donors and 104,251 cells; gut with 5 donors and 79,435 cells; kidneys with 4 donors and 26,372 cells; liver with 14 donors and 210,549 cells; spleen with 10 donors and 127,186 cells; bone marrow with 8 donors and 93,677 cells; and MLN with 2 donors and 6,039 cells.

For the process of batch correction and integration, we followed the framework outlined in the methods section. Where it was possible to acquire raw sequencing files, data from each source was re-mapped with a common reference genome (GRCh38-2020-A) in line with the YS scRNA-seq data. Low-quality cells were removed from concatenated data if they had fewer than 2,000 reads, fewer than 200 genes, or more than 20% mitochondrial reads. Doublet detection was performed with dynamic thresholds, and the likelihood of maternal contamination was estimated as described in the methods section. The data was then normalised and log-transformed. Highly variable gene (HVG) selection was performed, and dimensionality reduction and integration were carried out using the scVI module within scvi-tools (v0.19.0), treating data source and chemistry as technical covariates (HVG = 10,000, dropout_rate = 0.2, n_layer = 2). Leiden clustering was performed on the output scVI embedding as described in the methods section.

To harmonise annotations across the atlas, a second ldVAE model with the same parameters was trained. The EN LR framework described in the methods section was then trained on the ldVAE latent representation of a labelled subset of nine developmental organs from work by Suo et al. This model was used to probabilistically assign labels across all tissues, and these labels were then subjected to majority voting by the earlier derived scVI clusters. The resultant harmonised labels were manually checked against author pre-assigned labels to ensure reasonable alignment. ldVAE modelling also allowed for interpretability and manual screening of specific gene features used to classify each label.

Finally, to assign the likelihood of correspondence between YS cell states and cells from other tissues, the EN LR framework was trained on the labelled YS subset of the ldVAE latent representation and projected back onto the 12-organ atlas (Fig 5.1).

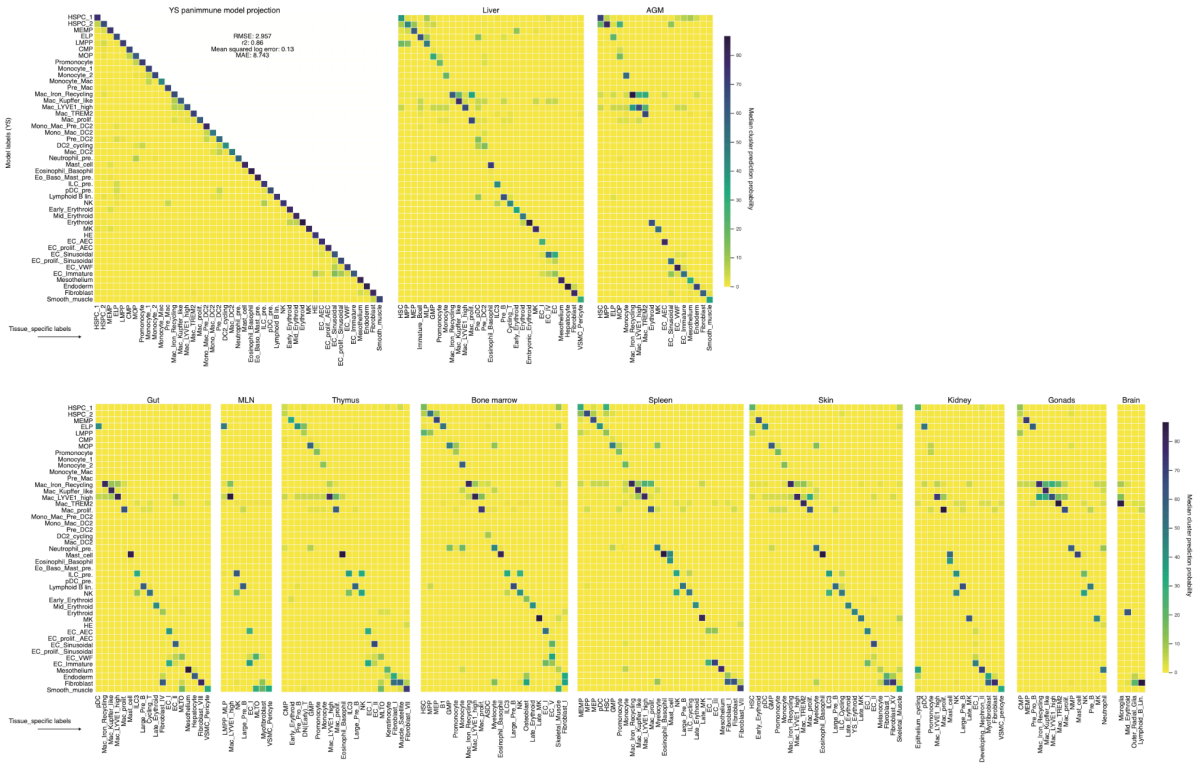


Figure 5.1: Logistic regression model of the fetal pan-organ variational landscape: *Heat maps illustrating ldVAE ElasticNet LR median cluster projection probabilities between YS labels (y-axis) and corresponding clusters (x-axis) in eight of the 12-organ fetal atlas (YS, liver, AGM, gut, MLN, thymus, FBM, spleen, skin, kidney, gonads, and brain). The model is trained on the YS subset of a jointly integrated 12-organ ldVAE latent representation (C=0.2, L1_ratio=0.05, R2=0.86, RMSLE=0.13). Only cell types which had a max probability >0.05 for any YS label are shown in the heatmap. Figure adapted from Goh and Botting et al, 2023 (1).*

5.3 Results

5.3.1 Macrophage subsets in human yolk sac and prenatal organs

Although YS hematopoietic progenitors are restricted to a short time window in early gestation, mouse models suggest that they contribute to long-lived macrophage populations in some tissues (*115*). By scRNA-seq, transcriptionally similar macrophage populations can be identified in YS and fetal brain prior to the emergence of definitive HSPCs (*117*). In our previous work, c=6682 YS macrophages resolved into two subgroups (*6*). By contrast, our integrated dataset of c=45,118 YS macrophages in the current study revealed a greater heterogeneity including pre-macrophages, *C1QA/B/C* and *MRC1*-expressing macrophages, and a rare *TREM2*$^{+}$ macrophage subset (**fig. 5.2**). Promonocytes expressing *HMGB2*, *LYZ*, and *LSP1* and monocytes expressing *S100A8*, *S100A9* and *MNDA* were also detected (**fig. 5.2**).

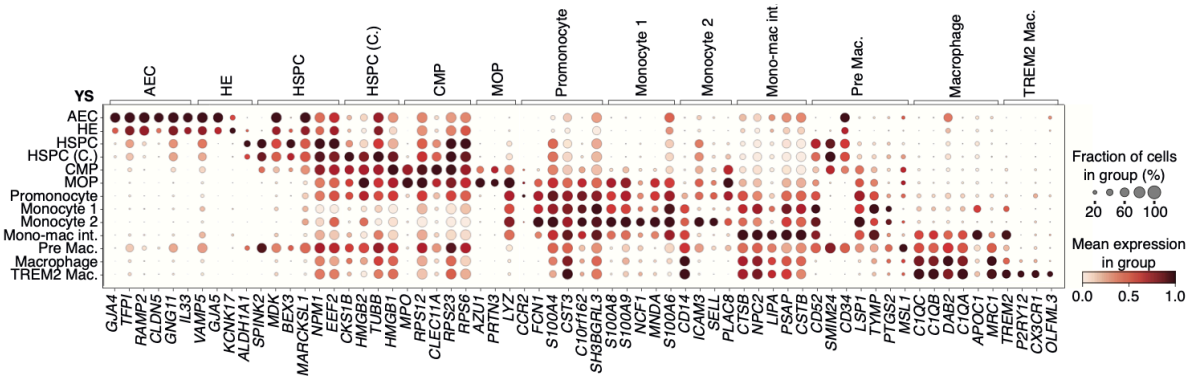


Figure 5.2: YS macrophage subsets: *Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of significant differentially expressed myeloid lineage gene markers (P<0.05) in the YS scRNA-seq. Differential expression was derived via a two-sided Wilcoxon rank-sum test (thresholded at expression in >25% of class, LFC >0.25 and Benjamini–Hochberg corrected P<0.05). Data are min-max-standardized with a distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).*

Monocytes were observed only after liver development and AGM-derived hematopoiesis at CS14 (~5PCW), but pre-macrophages and macrophages formed as early as CS10 (~4PCW)

(**Fig. 5.3 and fig. A8B).** Although the potential of early YS HSPCs to differentiate into monocytes has been demonstrated in vitro (*117*), there were too few promonocytes and monocyte progenitors in our data prior to CS14 to reliably confirm this potential.

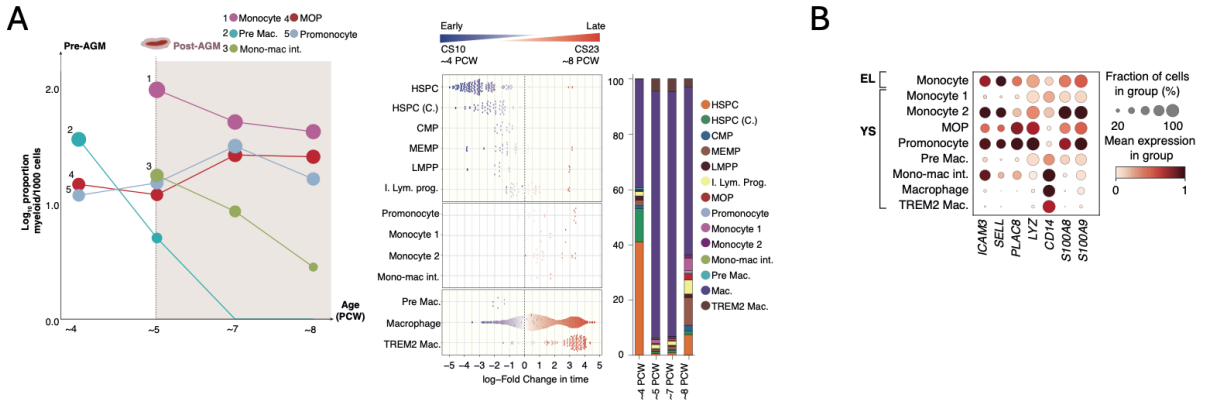


Figure 5.3: YS macrophage population distribution across time: *(A) Left: Line graph of monocyte and macrophage proportions in YS scRNA-seq across time. Dashed line indicates pre- and post-AGM stages. Middle: Milo beeswarm plot showing differential abundance of YS scRNA-seq myeloid neighbourhoods across time. Color shows degree of enrichment (blue: early, red: later). Right: Bar chart of YS scRNA-seq myeloid cell state proportions across time. Mono–mac int. monocyte macrophage intermediate. (B) Dot plot showing the mean expression (color) and proportion of cells expressing monocyte marker genes (dot size) in EL monocytes and YS myeloid cell states. Genes include YS vs EL monocyte DEGs and established monocyte markers. Figure adapted from Goh and Botting et al, 2023 (1).*

We identified two populations of YS monocytes, which diverged in expression of adhesion molecules. YS Monocyte 2 expressed adhesion molecules *ICAM3*, *SELL*, and *PLAC8* (Fig. 5.3), which were also expressed on fetal liver but not YS HSPCs (fig. A5C). YS Monocyte 2 had a high probability of class prediction against FL monocytes (fig. A8C). Thus, Monocyte 2 is likely a recirculating FL monocyte, although sequential waves of monocytopoiesis

occurring within the YS cannot be excluded. YS CITE-seq data was used to identify discriminatory markers (CD15 and CD43 for Monocyte 1; CD9 and CD35 for Monocyte 2) and provide protein-level validation for differential expression of *SELL* (CD62L) and CD14 (fig. A8D).

The YS pre-macrophage uniquely expressed high levels of *PTGS2*, *MSL1*, and *SPIA1*, as well as expressing progenitor genes (*SPINK2*, *CD34*, and *SMIM24*), macrophage genes (*C1QA* and *MRC1*), and *CD52*, which is typically associated with monocytes (fig. A8A).

5.3.2 A monocyte independent accelerated route to macrophage differentiation

This YS pre-macrophage rapidly declined by 5PCW (Fig. 5.3) and had no equivalent in EL (fig. A8C), KNN graph-based FDG and partition-based graph abstraction (PAGA) suggested a direct monocyte-independent trajectory to YS macrophages prior to CS14 (Fig. 5.4). In this pre-AGM trajectory, a transition from HSPC to pre-macrophages, then macrophages (nodes 1, 5, and 6 in Fig. 5.4 upper panel) fit with our predictions that pre-AGM HSPCs exhibit myeloid bias (Fig. 3.11). After CS14, there was a clear differentiation trajectory from cycling HSPC (x (C.)) to monocytes and monocyte–macrophages (nodes 1-7 in Fig. 5.4 lower panel). After CS14, 15.33% of this macrophage pool was proliferating and CellRank RNA state transition analysis was in keeping with active self-renewal (fig. A8E).

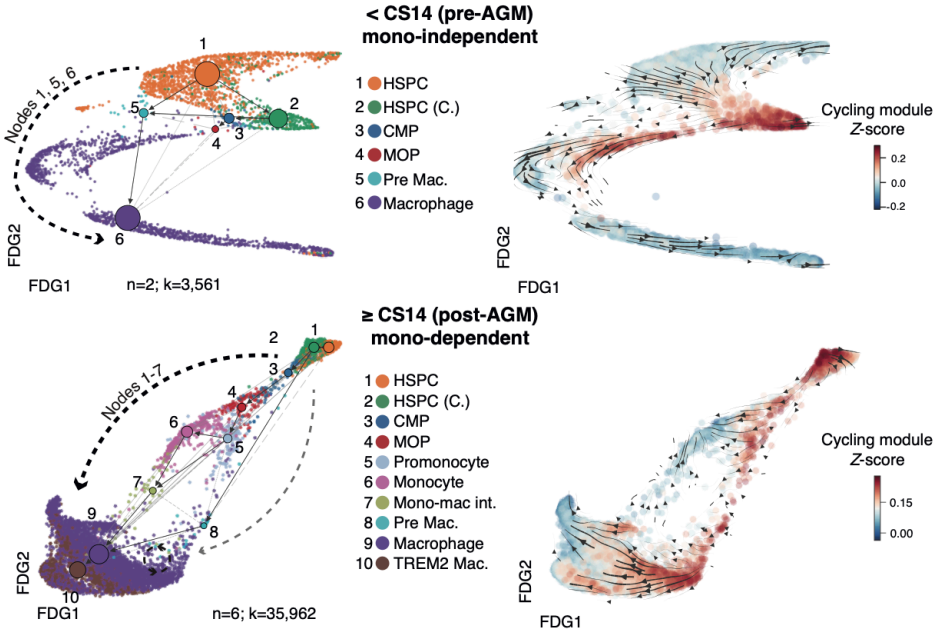


Figure 5.4: Monocyte-independent accelerated macrophage production in YS: *Left: FDG of macrophage trajectory in YS scRNA-seq, colored by cell state, overlaid with PAGA showing monocyte-independent <CS14 (pre-AGM; n=2; c=3,561; top) and monocyte-dependent trajectories >CS14 (post-AGM; n=6; c=35,962; bottom). Right: FDG overlaid with scVelocity directionality, colored by cell cycle gene enrichment (GO:000704 module). Figure adapted from Goh and Botting et al, 2023 (1).*

Using PySCENIC, YS pre-macrophages were predicted to employ a group of transcription factors (TFs), including FLI1 and MEF2C, that have been reported in the differentiation of multiple lineages (*429*, *430*). By contrast, the monocyte-dependent route (CMPs, monocyte progenitor (MOP), promonocytes and monocytes) relied on recognized myeloid transcription factors such as SPI1, CEBPA, and IRF8 (Fig. 5D).

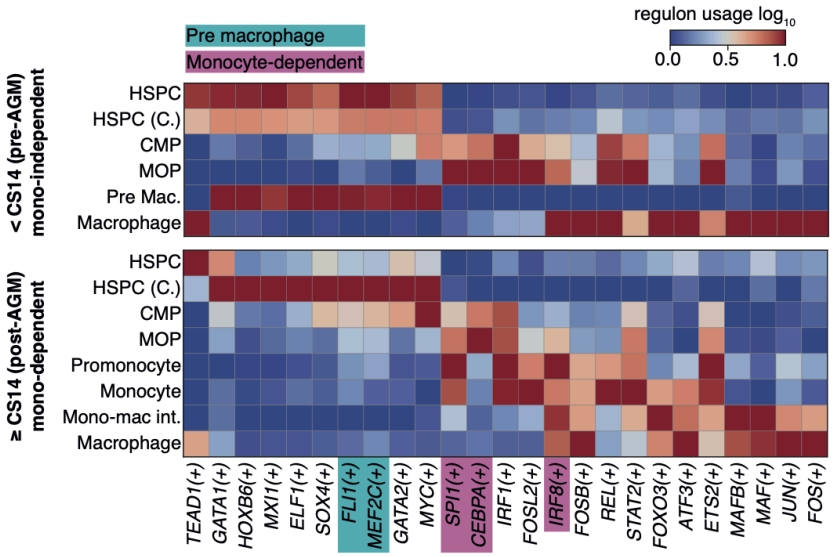


Figure 5.5: Regulons of monocyte independent and dependent macrophage production: *Heatmap of regulons associated with trajectories in Fig. 5.4. TFs discussed in text highlighted (turquoise: pre-macrophage; purple: monocyte-dependent). Figure adapted from Goh and Botting et al, 2023 (1).*

5.3.3 YS prespecification of a microglia-like TREM2+ Macrophage population

TREM2$^+$ macrophages expressed microglia-associated transcripts *CX3CR1*, *OLFML3*, and TREM2 and were observed in YS only after CS14 (Fig. 5.2 to 5.4, and 5.6A). By PAGA and CellRank state transition analysis, TREM2$^+$ macrophages were closely aligned to the self-renewing macrophage population (Fig. 5.4 and fig. A8E). YS TREM2$^+$ macrophages were located adjacent to the mesothelium, in a region enriched by EC (fig. A8F). CellPhoneDB predicted interactions between TREM2$^+$ macrophages and VWF$^+$ EC, via IL-8 (*CXCL8*) and NRP1 (*NRP1*), both of which are involved in angiogenic pathways (*431*, *432*) (Fig. 5.6B). TREM2$^+$ macrophages also expressed the purinergic receptor P2Y purinoceptor 12 (*P2RY12)*, which supports trafficking towards ATP/ADP-expressing ECs, as reported in the mouse CNS (*433*), (*155*) (Fig. 5.6A).

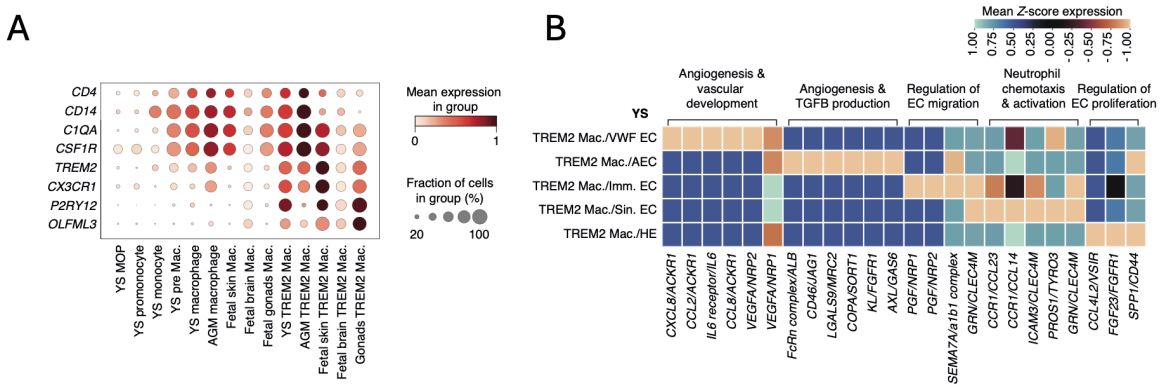


Figure 5.6: TREM2 macrophage functions: *(A) Dot plot showing the mean expression (color) and proportion of cells expressing macrophage and microglia marker genes (dot size) in myeloid cell states in YS, AGM (46), skin (85), gonad (44), and brain (404) fetal scRNA-seq datasets. (B) Heatmap of significant (P<0.05) CellphoneDB-predicted interactions between YS scRNA-seq TREM2^{+} macrophages and ECs. Color represents z-scored expression of gene pairs, brackets indicate top curated interactions for cell-state pairs. Figure adapted from Goh and Botting et al, 2023 (1).*

To establish whether *TREM2*$^{+}$ macrophages are present in other fetal tissues, we assembled an integrated 12-organ developmental atlas (fig. 5.7A). We resolved six macrophage fractions based on harmonized cross-tissue definitions from our recent prenatal immune analysis (by label transfer): pre-macrophages and TREM2^{+} macrophages (as in our cluster-driven annotations), as well as LYVE1^{hi}, Kupffer-like, iron-recycling, and proliferating macrophages (*85*) (Fig. 5.7 A to C; and Fig. A8, C and J). TREM2 is implicated in lipid sensing by anti-inflammatory tissue macrophages in the adult human and mouse (*409–411*), but we observed the highest expression of *TREM2* in macrophages bearing a "microglia-like" signature in developing tissues including YS, skin (as previously reported (*85*)), gonads (as previously reported (*44*)), brain, and AGM, but not BM, liver, kidney, thymus, mesenteric lymph nodes (MLNs), or gut (Fig 5.7C; and Fig. A8J).

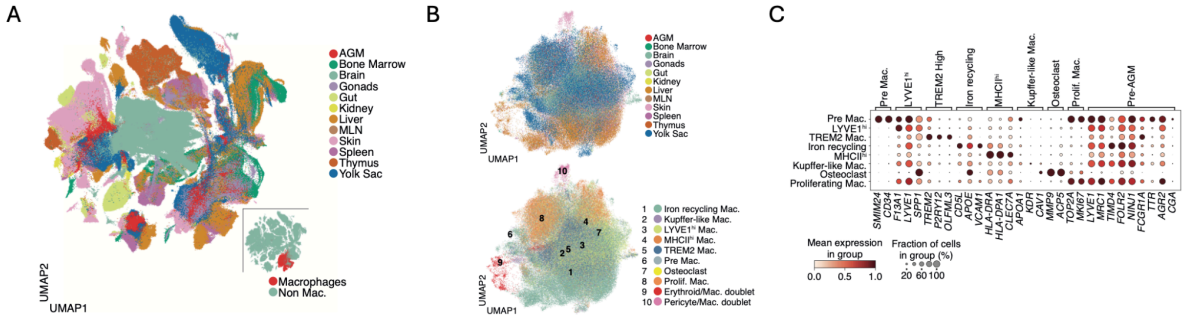


Figure 5.7: Fetal 12 organ macrophage atlas: *(A) UMAP of the integrated 12-organ fetal atlas (c=3.12×10^6, n=150), coloured by organ. Inset indicates the position of macrophages (teal) and non-macrophages (red). Organs include: YS (n=10, c=169,494), AGM (n=4, c=12,248), skin (n=13, c=178,563), brain (n=72, c=2.16×10^6), gonads (n=44, c=14,244), thymus (n=11, c=104,251), gut (n=5, c=79,435), kidneys (n=4, c=26,372), liver (n=14, c=210,549), spleen (n=10, c=127,186), bone marrow (n=8, c=93,677), and MLN (n=2, c=6039). (B) Feature plot showing VAE latent-space derived UMAP representation of macrophages across the integrated 12 organ fetal atlas coloured by organ (top) and by annotated heterogenous macrophage substates (bottom). (C) Dot plot showing the mean expression (colour scale) and the fraction of cells expressing each gene (dot size) of marker genes by macrophage subsets across the 12-organ fetal atlas. Data are min-max-standardized with a distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).*

5.3.4 YS macrophage signature predicts TRM contributions in fetal and adult life

Next, we asked whether transcriptional features of pre-AGM macrophages could be used to evaluate YS macrophage contribution to developing tissues. In our 12-organ macrophage dataset, pre-AGM macrophages were compared against post-AGM macrophages in an integrated variational-autoencoder (VAE) latent space using a Bayesian differential expression approach. The most predictive pre-AGM macrophage features comprised nine genes, including five genes in common with a "TLF$^+$ signature" identified from cross-tissue analysis of mouse macrophages (*LYVE1*, *TIMD4*, *FOLR2*, *MRC1*, and *NINJ1*) (*118*) (fig. 5.8A). By KDE, macrophages significantly enriched in our pre-AGM module colocalized with LYVE1^{hi}

macrophages from gonads, liver, skin, and AGM, and with all macrophage fractions from the YS (Fig 5.8B and C; fig. 5.9A; and fig. 5.7B).

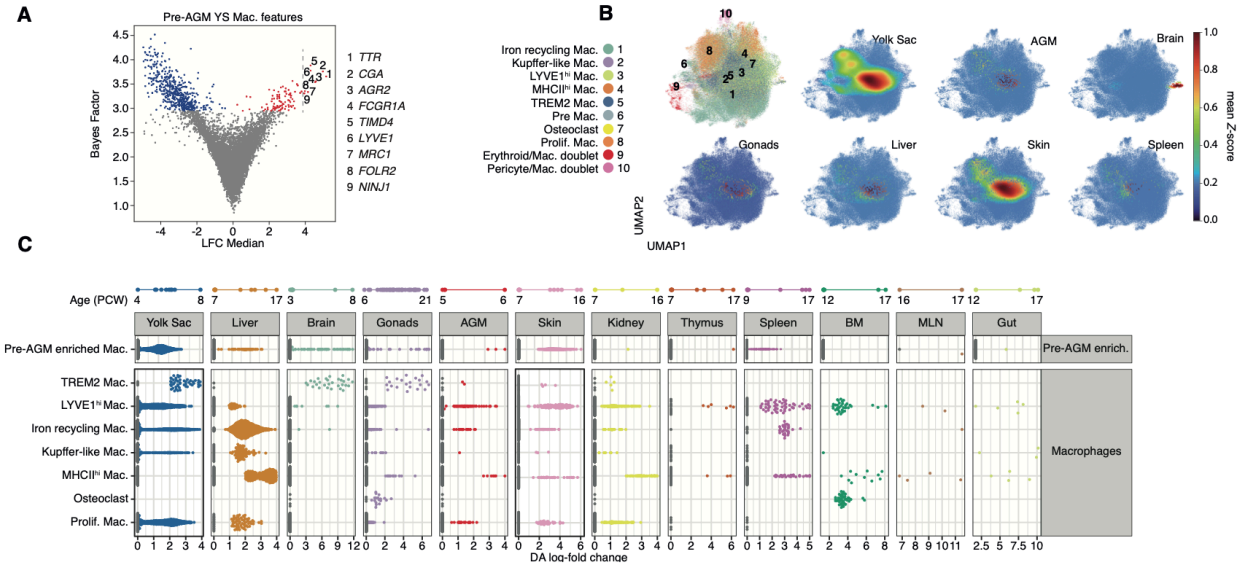


Figure 5.8: Pre-AGM macrophage markers and differential abundance: *(A) Volcano plot displaying the top nine significant differentially expressed genes (Bayes factor >3, Median LFC >4) of YS Macrophages arising pre-AGM (<CS14) against macrophages from all timepoints in the 12 organ-atlas, determined by scVI variational autoencoder differential feature selection. (B) UMAP (top left) and density plots showing macrophages from the 12-organ atlas colored by cell type (UMAP) or z-scored kernel density estimation (KDE) score of pre-AGM macrophage gene module enrichment (see methods), in AGM, brain, gonads, liver, skin, spleen and YS (Density plots). (C) Milo beeswarm plot showing neighborhood differential abundance of macrophages enriched in pre-AGM macrophage module, and other macrophage subtype modules across organs and gestational time. Colored neighborhoods are significantly enriched with positive fold changes (SpatialFDR<0.1, logFC>0) denoting significant presence at given time points and colors denote organs of origin. Per-organ sampled age ranges (PCW) are displayed at the top of the plot. Figure adapted from Goh and Botting et al, 2023 (1).*

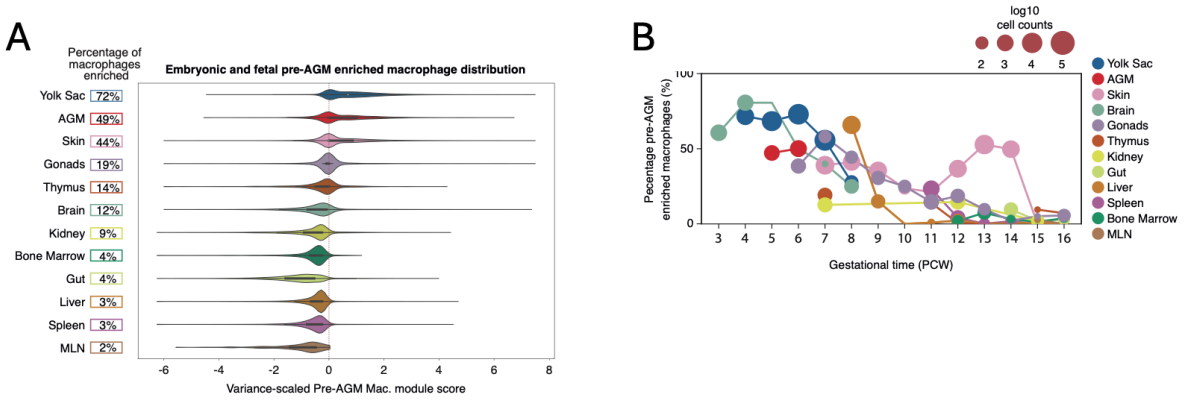


Figure 5.9: Cross tissue abundance of pre-AGM enriched macs: *(A) (left) Percentages of all macrophages positively enriched in Pre-AGM gene-module per organ in the fetal 12-organ atlas. (Middle) Violin plot showing distribution of the macrophages enriched in pre-AGM macrophage gene module. Violins show the median scaled module scores ((module_score-median)/MAD(module_score)). Each Violin indicates the distribution of enrichment scores across macrophages. The dotted red line indicates the threshold for positive enrichment (module_score>0). (B) Line graph showing the relative change in proportion of macrophages enriched in the pre-AGM macrophage module across gestational age. Dot size represents log10 cell counts and color represents organ. Figure adapted from Goh and Botting et al, 2023 (1).*

The proportion of pre-AGM module-enriched macrophages trended downwards over time, even in the brain (Fig. 5.9B).

By transcriptome alone, it was not possible to separate dilution by influx of non-YS macrophages from transcriptional adaptation to the tissue environment. With this caveat, we assembled a 20-organ, cross-tissue integrated landscape of adult tissue macrophages using publicly available single-cell data from the Human Cell Atlas and Tabula sapiens (Fig. 5.10, A to C). Fat, vasculature, muscle, brain, and bladder had the highest proportion of macrophages enriched in the pre-AGM signature (Fig. 5.10C).

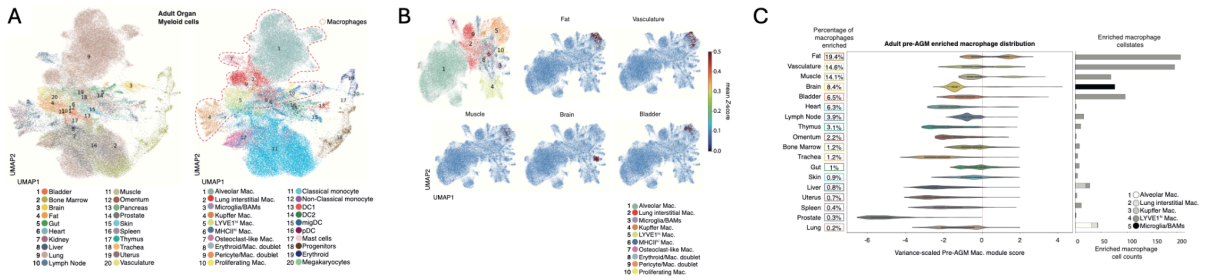


Figure 5.10: Adult 20 organ macrophage atlas:*(A) Left: Feature plot UMAP of the integrated 20-organ adult macrophage lineage atlas (n=65, c=94,286), coloured by organ. Right: UMAP colored by annotated cell-states, red line indicates macrophage clusters. Organs include: bladder (n=3, c=2831), fat (n=2, c=1667), heart (n=7, c=486), pancreas (n=2, c=1030), omentum (n=4, c=103), prostate (n=2, c=356), thymus (n=3, c=508), vasculature (n=2, c=1602) (291), bone marrow (n=10, c=9408), gut (n=13, c=1678), (356), lung (n=19, c=46805), lymph node (n=,15 c=2507), muscle (n=13, c=2553), spleen (n=15, c=13643), skin (n=7, c=1088), trachea (n=6, c=1376), uterus (n=6, c=632), (291, 389), brain (n=7, c=832)(434), kidney (n=9, c=358) (405), liver (n=12, c=4823) (291, 389, 435). (B) UMAP (top left) and density plots showing macrophages from the adult atlas coloured by cell type. UMAPs of the top five organs with the largest proportion of pre-AGM enriched macrophages coloured by z-scored kernel density estimation (KDE) score of pre-AGM gene module enrichment (see methods). (C) (left) Percentages of all macrophages positively enriched in Pre-AGM gene-module per organ in the adult 20-organ atlas. (Middle) Violin plot showing distribution of the macrophages enriched in pre-AGM macrophage gene module. Violins show the median scaled module scores ((module_score-median)/MAD(module_score)). Each violin indicates the distribution of enrichment scores across macrophages. The dotted red line indicates the threshold for positive enrichment (module_score>0). (Right) Bar plot illustrating the cell-state distribution of positively enriched macrophage subsets per organ by percentage. Figure adapted from Goh and Botting et al, 2023 (1).*

5.3.5 iPSC culture system recapitulate YS accelerated macrophage differentiation

To assess the fidelity of *in vitro* iPSC models in recapitulating the *in vivo* YS macrophage differentiation processes, we integrated our YS gene expression data with scRNA-seq data

from iPSC-derived macrophage differentiation (n=19; c=50,512) (*89*), following the refinement of iPSC-derived cell-state annotations (Fig. 5.11, A and B; and Fig 5.12).

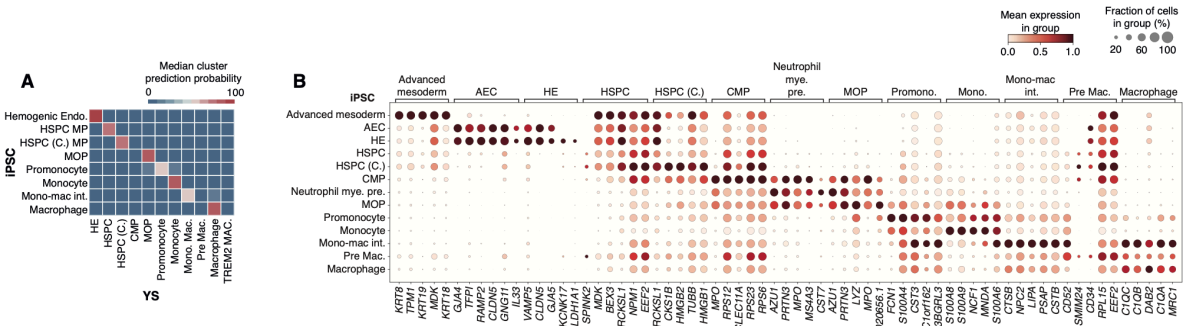


Figure 5.11: iPSC macrophage differentiation:*(A) Heatmap of class prediction probabilities for a logistic regression model (Elasticnet) trained on YS scRNA-seq myeloid cell states (x-axis) projected onto equivalent cell states in iPSC scRNA-seq data (y-axis), where the iPSC system was optimized for macrophage production (89). "Monocyte1" and "Monocyte2" are grouped into the "Monocyte" category. Colour scale indicates median probabilities. (B) Dot plot showing the mean expression (colour scale) and the fraction of cells expressing each gene (dot size) of myeloid lineage genes in myeloid-lineage cell states from the iPSC scRNA-seq (89) dataset. Data are min-max-standardised with a distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).*

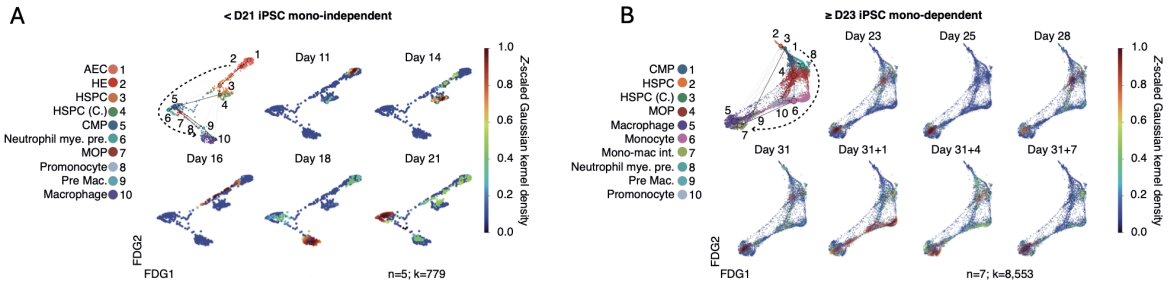


Figure 5.12: iPSC monocyte independent and dependent macrophage differentiation: *Density plots showing the distribution of transitioning iPSC-derived scRNA-seq (89) HSPCs and macrophage lineage cells from <D21 (n=5; c=779) (left) and >D23 (n=7; c=8553) (right) in the integrated FDG embedding. Color of cells represents the z-scored kernel density estimation (KDE) score for each timepoint. Figure adapted from Goh and Botting et al, 2023 (1).*

Non-adherent, *CD14*-expressing cells appearing after week 2 of differentiation expressed *C1QA*, *C1QB*, and *APOC1* in keeping with a macrophage identity, while C*D14*, *CD52*, *FCN1*, and *S100A8/9*-expressing monocytes only emerged after week 3 (Fig. 5.13, and Fig 5.11A).

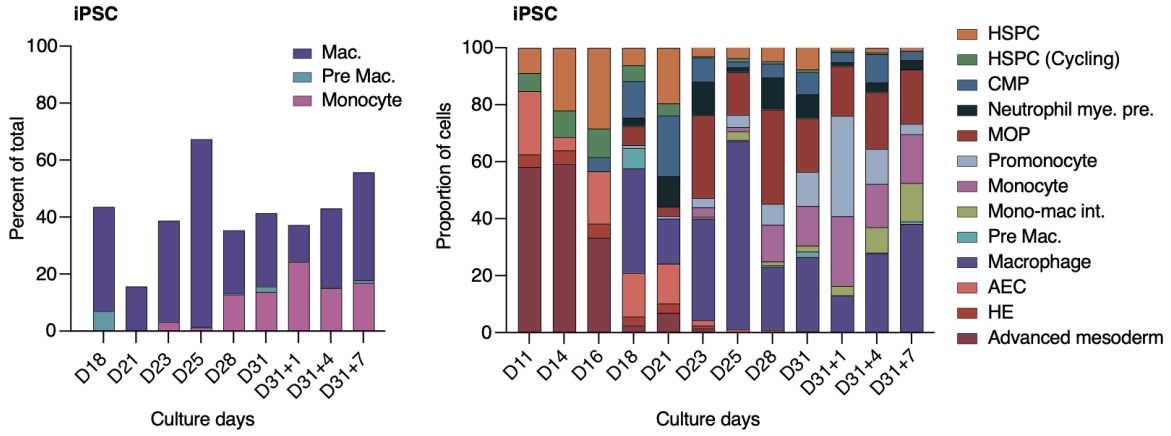


Figure 5.13: iPSC culture proportions by day: *Left: Stacked bar plot displaying the percent of monocytes, pre-macrophages, and macrophages found in iPSC cultures (89) by day. Right: Stacked bar plot displaying the proportion of cell states found in iPSC cultures by day. Figure adapted from Goh and Botting et al, 2023 (1).*

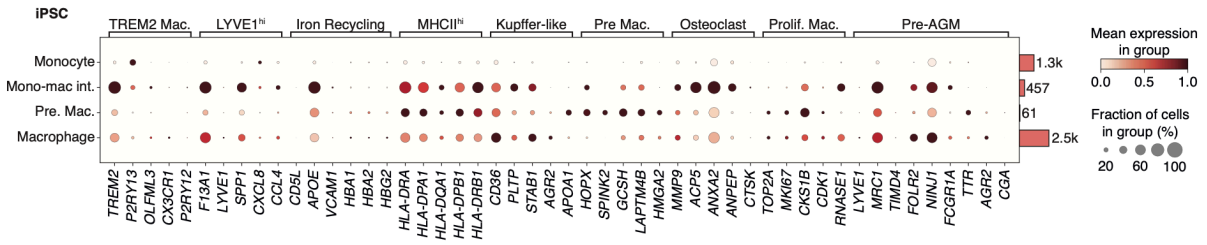


Figure 5.14: iPSC macrophages do not recapitulate YS macrophage heterogeneity: *Dot plot showing the mean expression (colour scale) and the fraction of cells expressing each gene (dot size) of macrophage subtype-defining gene sets in iPSC macrophage cell states. Data are min-max-standardised with a distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).*

Prior to monocyte emergence, a monocyte-independent macrophage differentiation trajectory was observed, consistent with previous observations (*89*) (Fig. 5.15A; and Fig. 5.12, A and B). TF regulatory profiles of iPSC-derived macrophage differentiation were consistent with the both pre-macrophage and monocyte-dependent TF profiles inferred from our YS data, including usage of *MEF2C*, *SPI1*, *CEBPA*, and *IRF8* in iPSC-derived pre-macrophages (Fig. 5.15B). However, neither iPSC culture system could recapitulate the heterogeneity of macrophages seen in native tissues (fig. *5.14*), suggesting that interactions with stromal cells, such as ECs, may be required to acquire specific molecular profiles.

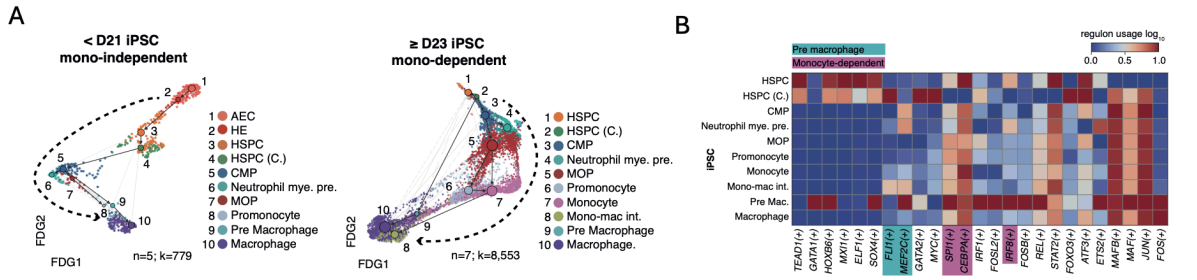


Figure 5.15: iPSC monocyte independent and dependent macrophage differentiation *(A) FDG of macrophage trajectory in iPSC scRNA-seq (89), colored by cell state, overlaid with PAGA showing monocyte-independent <D21 (n=5; c=779; left) and monocyte-dependent >D21 (n=7; c=8,553; right) transitions. (B) Heatmap of regulons associated with iPSC macrophage trajectories shown in (A). TFs discussed in text are highlighted as in (Fig. 5.5). Figure adapted from Goh and Botting et al, 2023 (1).*

5.4 Chapter 3 discussion

While substantial progress has been made in understanding macrophage ontogeny through bone marrow and animal models, this study presents the first observation of an accelerated macrophage differentiation route in humans, offering a novel insight into macrophage development and ontogeny. This pathway, discovered within the YS, bypasses the monocyte stage, transitioning directly from HSPCs to pre-macrophages and subsequently to mature macrophages. This route is significant as it suggests a mechanism of early macrophage specification crucial for the development of YS-derived TRMs during early gestation.

The temporal dynamics of this process are noteworthy. Pre-macrophages appear as early as Carnegie Stage (CS) 10 (~4PCW) and decline rapidly by CS14 (~5PCW). This timing indicates a tightly regulated window for accelerated macrophage differentiation. Gene expression profiling of these pre-macrophages reveals a unique set of genes, including PTGS2, MSL1, and SPIA1, which are associated with both progenitor and macrophage-specific functions. Additionally, these cells express CD52, typically associated with monocytes, indicating a transitional state.

Trajectory analysis via CellRank supports a monocyte-independent differentiation trajectory, suggesting that these pre-macrophages progress directly to mature macrophages without passing through a conventional monocyte stage. This finding is further corroborated by the absence of equivalent populations in the EL prior to CS14 and successful *in vitro* recapitulation of this accelerated differentiation route. iPSC-derived macrophage differentiation systems demonstrate that non-adherent, *CD14*, *C1QA/B/C* expressing cells appear before monocytes, expressing macrophage markers and mimicking the *in vivo* pre-macrophage to macrophage accelerated transition. However, these systems fail to capture the full heterogeneity of YS-derived macrophages including TREM2+ macrophages, suggesting that additional stromal interactions may be necessary for complete differentiation. The transcriptional regulation of these pre-macrophages includes transcription factors such as *FLI1* and *MEF2C*, distinct from the *SPI1*, *CEBPA*, and *IRF8* profiles observed in monocyte-dependent macrophage differentiation. It is also important to acknowledge that trajectory inference methods, including CellRank, inherently rely on assumptions such as consistent splicing kinetics and comprehensive sampling of transitional states. These conditions are particularly challenging to satisfy in rare tissues like the YS, where gaps or undersampling in datasets can introduce biases in trajectory and regulatory inference. Thus,

there remains a critical need for robust, orthogonal model systems, like refined iPSC-based platforms, that can more rigorously replicate the complexities and nuances of YS myelopoiesis. Future work could integrate epigenetic profiling (e.g., ATAC-seq) with trajectory methods to pinpoint exactly how these regulatory factors drive accelerated macrophage specification. Additionally, refining velocity-model assumptions (e.g., gene-specific splicing rate parameters) or sampling more YS donors at 3–5 PCW could help resolve whether these accelerated trajectories are indeed ubiquitous across individuals or partly shaped by limited sample availability.

The pre-specification of microglia-like macrophages in the YS has significant evolutionary implications. We observe these cells in various tissues such as the gonads, skin, YS, and brain, and exhibit pro-angiogenic properties, potentially vital for early vascular development and tissue patterning. One hypothesis is that the YS orchestrates a cascade of signalling events to guide angiogenesis. One potential signalling pathway involves our observation of the endoderm expressing high levels of fibrinogen and F2. These interact with protease-activated receptors (PAR1-4) on VWF^{+} ECs and AECs (*425*, *436*). This interaction produces ATP and ADP, released via $PANX1^{+}$ pannexin channels, recruiting microglia-like macrophages via purinergic P2RY12 receptors responding to these nucleotides (*433*, *437*, *438*). The pro-angiogenic functions of these microglia-like macrophages are underscored by their presence near vascular structures (Fig. A8F), suggesting a role in angiogenesis. These cells express TREM2 and ITGAM (the latter additionally responding to fibrinogen) (*439*) and migrate towards purigenic gradients via purinergic receptors (*433*, *437*, *438*). This signalling axis may facilitate vascular growth and patterning by utilising TREM2 microglia-like macrophages to guide angiogenesis. The interaction between microglia-like macrophages and ECs, potentially involving the exchange of VEGFB and CXCL8, supports the hypothesis that these macrophages contribute to the establishment and maintenance of a functional vascular network (Fig. 5.16) (*433*, *437*, *438*). This function is critical for embryonic survival, as evidenced by the murine lethality observed in PAR1-3/thrombin deficiency models post-AGM (E10.5). Additionally, PAR1-3/thrombin deficiency murine models also produce altered YS vasculature, further supporting the hypothesis (*425*).

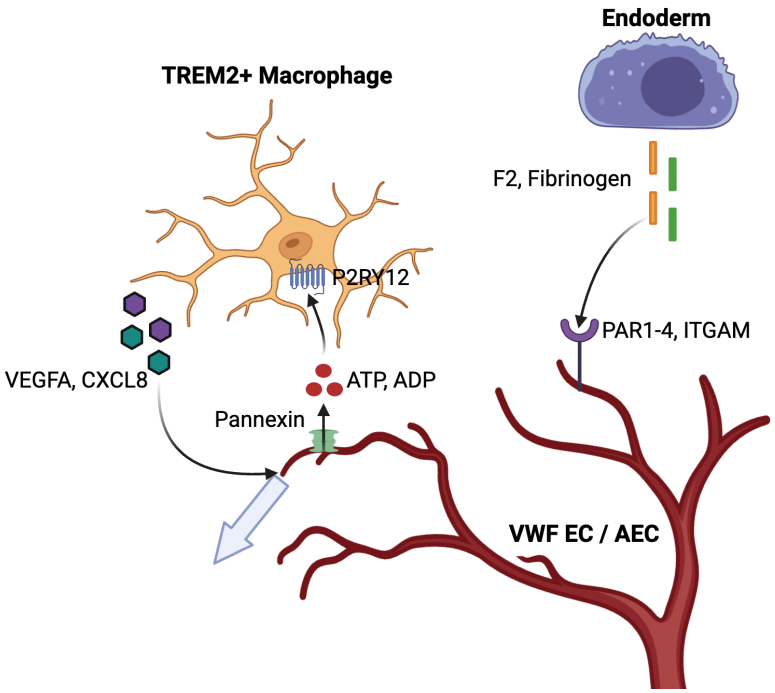


Figure 5.16: Hypothesis for pre-specification of TREM2+ microglia-like cells in YS: *This cartoon illustrates the potential signalling cascade between endoderm, TREM2+ macrophages, and ECs which may functionally allow the YS to orchestrate angiogenesis. PANX1 is a channel protein of the pannexin family. Figure produced using BioRender.com*

As discussed in the introduction chapter, seminal work by Dick et al. explored the orchestrational dynamics and origins of tissue-resident macrophage populations across various murine organs, including the heart, liver, lung, kidney, and brain. Using a fate mapping mouse model, they identified a highly conserved population of TLF+ (TIMD4, LYVE1, and/or FOLR2) expressing macrophages in both murine and human single-cell expression profiles. These macrophages corresponded to YS-derived myeloid progeny from the murine fate mapping study. This research suggests that macrophages originating from the YS share a conserved core gene program that remains intact post-seeding in tissues.

In our work, comparing pre-AGM YS macrophages to post-AGM macrophages from various tissues allowed us to recover an extended pre-AGM YS macrophage expression profile. This included *TIMD4*, *LYVE1*, and *FOLR2*, as well as additional markers such as *TTR, CGA, AGR2,* and *FCGR1A*.

Though we couldn't discount the caveat of non-YS macrophages transcriptionally adapting to the tissue environments, this approach enabled the identification of macrophages across various fetal tissues that retain the YS-derived transcriptional signature. Employing this

pre-AGM module enabled the tracing of the presence and contribution of YS-derived macrophages in developing fetal tissues. We observed enrichment for these modules in liver, brain, gonads, skin, and spleen. However, the global proportion of these macrophages trended down over time, potentially indicating replacement by later, AGM-derived macrophage subsets. Additionally, with these caveats, we employed the same method to delineate YS-derived macrophages in adult tissues and observed an enrichment in fat, vasculature, muscle, and brain tissue compartments. The use of this gene module further supports the hypothesis that YS-derived macrophages follow a conserved differentiation program, establishing themselves as TRMs capable of self-renewal and long-term maintenance in tissues.

In summary, the identification and characterisation of an accelerated macrophage differentiation route and the pre-specification of microglia-like macrophages underscore the complexity and adaptability of the myeloid lineage during human development. Further research into the molecular cues and stromal interactions governing these processes will 1. enhance our understanding of TRM origins and biology, and 2. improve vascularisation efforts in bioengineering and regenerative medicine.

7 Overall Discussion

In this chapter I discuss and conclude the research and overall results reported throughout this thesis.

The work outlined in this chapter is the result of collaboration as described in chapter 1 introduction.

Section 7.1 in this chapter contains lightly-edited portions of the discussion chapter of the manuscript which I co-first authored (*1*).

7.1 Discussion

Using single cell multiomic and imaging technologies we delineate the dynamic composition and functions of human YS from the 3rd post conception week (PCW), when the three embryonic germ layers form, to the 8th PCW when the majority of organ structures are already established (*412*). We detail how the YS endoderm shares metabolic and biosynthetic functions with liver and erythropoiesis-stimulating functions with liver and kidney. In part, this shared functionality may relate to a common role in creating a niche for haematopoiesis (*440*). Unlike in mice, where primitive erythroid progenitors mature in the YS (prior to circulation being established) but erythromyeloid progenitors can exit the YS and mature in the fetal liver, we show that active differentiation of erythroid and macrophage cells occurs for several weeks in human YS, prior to liver handover. The multi-organ functions, including extended haematopoiesis, of human YS may be an evolutionary adaptation to the longer gestation in humans. Previous human studies, primarily based on colony assays, have suggested a transition from YS to liver haematopoiesis at 5PCW, but did not analyse YS samples beyond this point, leaving a gap in our understanding of later stages of YS function after this point (*441*).

The developmental window investigated here encompasses haematopoiesis from HSPCs arising both within the YS and within the embryo proper. We reconstruct YS HSPCs

emergence from a temporally-restricted HE, featuring similar transition states and molecular regulation to AGM HSPCs. By gastrulation (CS7; ~3PCW), YS HSPCs already differentiate into primitive erythroid, MK and myeloid lineages. Building on a recent compilation of gene scorecards that characterise primitive and definitive HSPCs (*46*), we were able to parse the two fractions and document transition to definitive HSPC-dominance after CS14 (~5PCW). This separation also allowed us to identify a primitive HSPC bias towards myeloid, erythroid and megakaryocyte lineages and a definitive HSPCs bias towards megakaryocyte and lymphoid lineages. Both primitive and definitive HSPCs in the YS became more quiescent and upregulated apoptosis-related genes between CS17 and CS23 (~6-8PCW). Stromal cell ligands predicted to support HSPCs were markedly disrupted during this time, suggesting that the barriers to the survival of YS HSPCS may be extrinsic.

Primitive HSPCs uniquely employ an accelerated route to macrophage production independent of monocytes. The monocyte-dependent route may provide more tunable macrophage production via circulating innate immune cells to facilitate macrophage regeneration in response to tissue damage, inflammation or infection. While both primitive and definitive HSPCs, 'accelerated' and monocyte-dependent macrophages were recapitulated during *in vitro* differentiation of iPSCs, which recapitulate key stages of differentiation between HSPCs to macrophages, TREM2^{+} macrophages were not. TREM2^{+} macrophages, which are transcriptionally aligned with brain microglia, fetal skin, testes and AGM TREM2^{+} macrophages, were predicted to interact with endothelial cells, potentially supporting angiogenesis as described in mouse brain (*404*). Benchmarking of *in vitro* cultures against *in vivo* cell states and trajectories can facilitate more faithful replication of early blood and immune cells. The accurate representation of rare or subtle populations, when benchmarking *in vitro* cultures, highlights the critical role of analytical methods. As discussed previously (section 1.3.4), robust feature selection methods that systematically model

variance-mean relationships can significantly improve sensitivity to rare cell states, enhancing the biological interpretability of developmental trajectories. Moreover, integrating datasets using methods such as scVI (section 1.3.7) strongly depends on careful hyperparameter optimization. Future studies should incorporate rigorously assessment of feature selection methodologies, and integration method hyperparameter sensitivity against ground-truth biological labels or known marker expression patterns to ensure that rare populations and subtle transcriptional states are reliably preserved.

There is a growing appreciation of the potentially life-long consequences of early developmental processes. Our study illuminates a previously obscure phase of human development, where vital organismal functions are delivered by a transient extraembryonic organ employing non-canonical cellular differentiation paths. It will be fascinating to explore how these processes may impact on tissue homeostasis and disease across the human lifespan.

7.2 Therapeutic applications of single-cell discoveries

While the YS atlas described in this thesis provides a range of biological insights into critical stages of embryonic hematopoiesis and survival, the contextualisation of this data against a larger backdrop of multi-developmental stage, tissue, adult, and animal atlases also provides key therapeutic insights. Two main directions include: the discovery of novel regulatory and gene targets utilising alternate developmental programs to reverse disease pathogenesis, and the improvement of biomimetic tissue engineering protocols.

7.2.1 Guiding novel therapeutic target development

Discoveries in the single-cell landscape have revolutionised our understanding of cellular heterogeneity and gene regulation, enabling the identification of novel therapeutic targets (*54*, *86*, *442*, *443*). Mapping the detailed molecular profiles of individual cells, enables the discovery of specific pathways and regulatory networks that are perturbed in disease progression and treatment resistance.

The recent approval of CRISPR-Cas9 gene editing therapy for β-thalassemia highlights the potential of advanced genetic technologies in treating genetic disorders. This therapy targets the Beta globin gene mutations responsible for the disease, introducing beneficial genetic changes that promote the production of functional haemoglobin. On 16 November 2023, the UK Medicines and Healthcare Products Regulatory Agency (MHRA) became the first to approve Casgevy (exagamglogene autotemcel), a CRISPR-Cas9-based gene editing therapy, for treating patients with transfusion-dependent β-thalassemia and sickle cell disease in patients aged ≥12 years with recurrent vaso-occlusive crises. This therapy works by editing the HSCs of the patient to reactivate fetal haemoglobin (HbF) production. The primary targets are the

BCL11A gene enhancer elements, which are key regulators that suppress HbF production after birth. By disrupting these regulatory elements, the therapy reactivates the Gamma globin genes, leading to increased production of HbF, which can compensate for the defective Beta globin and ameliorate the symptoms of β-thalassemia. However, no analogous strategy currently exists for α-thalassemia.

CRISPR-Cas9 technology, which stands for Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9, allows for precise editing of the genome by creating double-strand breaks at specific locations, which are then repaired by the cell's own machinery. This technology consists of two key components: the Cas9 protein, which acts as molecular scissors, and a guide RNA (gRNA) that directs Cas9 to the exact location in the genome that needs to be edited. The mechanism behind CRISPR-Cas9 involves the design of a gRNA that is complementary to the target DNA sequence. When introduced into the cell, the gRNA binds to the target DNA, and the Cas9 protein induces a double-strand break. This break can be repaired through two main pathways: non-homologous end joining (NHEJ) or homology-directed repair (HDR). NHEJ often results in insertions or deletions (indels) that can disrupt gene function, making it useful for gene knockouts. HDR, on the other hand, uses a donor template to precisely repair the break, allowing for the insertion of new genetic material. This precision makes CRISPR-Cas9 an invaluable tool for correcting genetic mutations, as seen in the treatment for β-thalassemia.

In α-thalassemia, unlike the fetal Gamma globin genes at the Beta globin locus (*HBG1*, *HBG2*), the expression of the embryonic Zeta globin gene (*HBZ*) is almost never re-activated to physiologically relevant levels by cis- or trans-acting mutations, or during stress erythropoiesis. Taken together, this suggests that the embryonic *HBZ* gene is more deeply silenced than the fetal Gamma globin genes. This deeper silencing of *HBZ* during development occurs via both transcriptional and post-transcriptional mechanisms.

Experimentally, it has been demonstrated that *HBZ* transcriptional silencing involves interactions between its promoter and a specific silencer element located in the gene's 3′ flanking region, resulting in substantial transcriptional repression during definitive erythropoiesis (*90*, *444*). However, this transcriptional silencing is incomplete, and residual *HBZ* mRNA undergoes rapid degradation due to instability mediated by its 3′ untranslated region (UTR). Compared to alpha globin mRNA, *HBZ* mRNA has a lower affinity for a sequence-specific messenger ribonucleoprotein (mRNP) stability complex, as quantified by electrophoretic mobility shift assays (EMSA) revealing a six-fold difference in apparent dissociation constants, which contributes significantly to its reduced stability and thus low expression (*444*). Thus, both robust transcriptional repression and inherent post-transcriptional instability contribute to the more profound and sustained silencing of *HBZ* relative to fetal Gamma globin genes. Nevertheless, if the *HBZ* gene could be effectively stabilized or transcriptionally reactivated by circumventing these mechanisms, it could complement and compensate for otherwise detrimental mutations in the adult alpha globin genes, providing a therapeutic avenue for severe transfusion-dependent α-thalassemia.

Our study sheds light on distinct spatio-temporal switches from zeta to alpha globin expression between the YS and the developing liver by 7PCW. We highlight how the *HBZ* gene is normally regulated in primitive erythropoiesis and how its silenced state is established and maintained in definitive AGM-derived erythropoiesis. This transition is inferred to be regulated by key transcription factors BCL11A and ZBTB7A, which attenuate the production of *HBZ* while promoting alpha globin expression during the shift from embryonic YS-derived erythropoiesis to post-AGM definitive erythropoiesis.

These findings are corroborated by King *et al.* (2021) (*90*), who independently identified these transcriptional regulators governing the primitive erythropoietic switch using scATAC-seq. Their research demonstrated that partial reactivation of the embryonic *HBZ* gene

can be achieved through acetylation and inhibition of histone deacetylases. The key targets identified include the same *BCL11A* and *ZBTB7A* transcriptional repressors, which play crucial roles in the regulation of globin gene expression.

Given these insights, it is realistic to consider a therapeutic strategy using CRISPR to target not only *BCL11A*, as in the current Casgevy approach, but also *ZBTB7A*. Silencing both transcriptional repressors could reactivate the *HBZ* gene, potentially rescuing abnormalities arising from mutations in adult alpha globin genes. However, while this initial strategy may effectively increase *HBZ* transcription, challenges related to post-transcriptional regulation and subsequent mRNA instability could still limit meaningful protein production. Additionally, dual silencing of *BCL11A* and *ZBTB7A* in HSPCs, or downstream erythro-myeloid and megakaryocyte-erythroid progenitors, carries uncertain risks, as both factors regulate additional critical processes (e.g., immune function, cell proliferation). ZBTB7A, for instance, also regulates myeloid differentiation and tumor suppressor pathways (*445*), while BCL11A plays established roles in lymphoid lineage development (*446*). The downstream impacts of these edits thus remain unknown without extensive validation using in vivo and ex vivo models. Computational tools like GEARs (*447*), which leverage large gene regulatory networks (GRNs) to predict perturbation effects, could help anticipate off-target changes but are limited by incomplete knowledge of lineage-specific regulation and may not fully capture emergent cellular states. Therefore, future studies employing multi-omic assays and functional transplantation models will be essential to confirm safety and efficacy in reactivating embryonic globins without compromising off-target hematopoietic lineages.

7.2.2 Hypothesis free mechanistic target identification

As publicly accessible single-cell multi-omic atlases rapidly expand, it is becoming increasingly important to refine methods for hypothesis-free discovery of perturbed mechanisms in disease. These data-driven approaches can uncover novel therapeutic targets and pathways that might otherwise elude conventional hypothesis-driven research, thereby providing a more comprehensive understanding of complex biological contexts in disease, and facilitating the identification and development of targeted intervention strategies.

The developmental data gathered in this study, including rare human fetal tissues from the YS and a multi-organ atlas, represent unique snapshots distinct from typical healthy or diseased states. These developmental snapshots offer valuable insights into novel co-opted disease mechanisms (*5*), filling critical knowledge gaps and providing mechanistic clues crucial for targeting perturbed conditions in disease states, and insights into disease pathology. For example, in the results, we discuss the developmental origins and characterization of human YS EHT, a process absent in adult life and unique to early embryonic development, which underlies the earliest wave of haematopoiesis. Comparing the distinct erythroid states that emerge from this early embryonic wave, to later emerging alpha globin-producing, fetal erythroid cells, revealed the aforementioned transcriptional regulators BCL11A and ZBTB7A as potential therapeutic targets in α-thalassemia, an insight only possible through this early embryonic context.

Early progress toward integrating single-cell data in drug repurposing has emerged in methods like Drug2Cell (*55*), which links drug-target databases such as ChEMBL (*448*) to scRNA-seq expression data. By analyzing gene set enrichment profiles for each drug's predicted targets, Drug2Cell identifies compounds whose pharmacological actions may preferentially impact specific cell populations at single-cell resolution, potentially revealing unexpected cellular targets or off-target interactions that are not apparent in bulk analyses. However, while promising, this approach heavily relies on established drug-target associations, thereby missing more complex pharmacological interactions such as pro-drug activation, polypharmacology, and indirect or secondary cellular responses. Additionally, methods primarily based on known targets often overproduce candidate drugs and yield false positives, complicating interpretability and downstream validation. To better model a wider spectrum of complex drug-cell interactions, high-throughput complementary screening assays in cell lines and ex-vivo culture systems (such as organ-on-chip or organoid models) are

necessary. Such methods could include CRISPR-based perturbation screens (e.g., Perturb-seq (*376*)), drug-induced transcriptome profiling (e.g., Drug-seq (*449*)), and drug-target affinity validation assays (e.g., DARTS) (*449*)). Given these interpretive challenges, a systematic framework is ultimately required to further enhance annotation and clustering of Drug2Cell-derived predictions, for example, one strategy might be to organize candidate drugs according to shared mechanisms of action (MOA), structural similarities, target gene programs, or known side-effect profiles, thereby enabling more robust prioritization and validation.

In parallel, foundation models for single-cell data are emerging as potentially powerful tools to refine these predictions in a more data-driven manner. Models such as scGPT (*450*) and GeneFormer (*451*) are large-scale transformer-based architectures pre-trained on extensive corpora of single-cell transcriptomic profiles (30, and 33 million cells respectively), enabling them to learn nuanced patterns of gene co-expression and potential regulatory interactions from diverse cell states. These models typically accept scRNA-seq data as input, and output low-dimensional embeddings capturing relationships among cells and genes. These embeddings can then be leveraged to interpret gene-expression shifts given specific biological inputs or perturbations, thereby enabling the prediction of key gene-gene relationships underpinning specific contexts, such as shifts between healthy and disease cell states. These transformer-based methods employ attention mechanisms, akin to those used in natural language processing, to identify which genes are most relevant for predicting the expression state of other genes. Meanwhile, scVI (*330*), based on the VAE architecture, represents another foundational approach for generative modeling of single-cell data that can capture complex transcriptional variability across biological conditions. Importantly, these foundation models can be accessed or explored through evolving resources such as the HuggingFace (*452*), and CZI Virtual Cell (*453*) platforms, which aim to make model usage more available across communities.

Despite their promise, it is important to recognize that these foundation models are trained on heterogeneous datasets and may not automatically capture every biological state of interest, particularly those that are rare or underrepresented in public atlases. Additionally, because each transformer-based foundation model only attends to a fixed number of genes per cell (e.g., context lengths of 512 genes for scGPT and 2,048 genes for GeneFormer), any regulatory interactions involving genes outside that window are effectively invisible to the model, limiting its ability to learn long-range or genome-wide relationships. This constraint

makes strategic inclusion of diverse cellular states, and careful verification of gene subsets selected in per cell context windows of known cell-type populations (for example, by inspecting gene-level embeddings to identify genes most frequently attended to and cross-referencing these against validated gene regulatory networks), critically important to verify that the model captures a wide spectrum of relevant regulatory programs during training and inference. As such, fine-tuning becomes a crucial step, allowing users to adapt pre-trained models to specific tasks and data. In fine-tuning, additional classification or regression heads can be attached to the model, while select layers (or the entire model) are unfrozen, enabling their weights to be updated according to new inputs. Specifically, classification tasks can target discrete states, for example, distinguishing pathological, from healthy cell populations in scRNA-seq data. Regression tasks, on the other hand, use continuous labels. For example, during fine-tuning, a regression head is trained to minimize a continuous loss (e.g., MSE, Kullback–Leibler divergence) between predicted and observed protein levels, using embeddings inferred from scRNA-seq inputs together with paired protein measurements from a diverse CITE-seq atlas. At inference, this head maps embeddings of new scRNA-seq profiles to protein abundance predictions. However, his approach depends heavily on the representativeness and diversity of the reference atlas and does not explicitly account for any additional cell or context specific post-transcriptional regulation that may affect protein expression in unseen samples. These approaches also extend to perturbational settings such as the fine-tuning on Perturb-seq (*376*) data. In this scenario, fine-tuning may employ a masked language modeling (MLM) objective in which a subset of genes, including those targeted by knockouts or chemical treatments, are masked, and the model is trained to reconstruct their expression shifts from the remaining gene context and any perturbation metadata. This strategy was demonstrated in the GeneFormer perturbation fine-tuning strategy, where the authors employed a masked language modeling (MLM) objective, that was used to learn how transcriptional profiles change following specific perturbations. In principle, this may enable generalization to broader disease-relevant contexts, assuming these gene-gene relationships remain consistent across diverse conditions. However, experimental validation remains essential to confirm true causal regulatory interactions. Ultimately, the success of fine-tuning or training these foundation models to predict expression shifts in unique disease-relevant contexts depends on the breadth and diversity of the scRNA-seq reference atlases available. Because many disease-relevant mechanisms involve rare cell states that are underrepresented or missing in standard public datasets, integrating additional, diverse single-cell resources is essential to capture the full spectrum of biological variation.

The growing size of publicly accessible single-cell multi-omic atlases, including the developing human datasets described here, offers potential for extending these foundation models into underexplored areas such as fetal and embryonic contexts. Incorporating fetal-specific scRNA-seq data through continuous learning or fine-tuning strategies, may accelerate the pinpointing of cell states and gene programs consistently implicated in both developmental and disease-relevant transitions, offering insights into mechanistic programming of specific disease contexts. For instance, integrating, and fine-tuning on scRNA-seq profiles from embryonic yolk sac, fetal skin, and psoriatic skin could accelerate uncovering how fetal angiogenic and inflammatory programs are co-opted in macrophages across both developmental and disease contexts, a phenomenon first highlighted by Reynolds et al. (2021) (*5*).

Ultimately, the use of fine-tuned single-cell foundation models could expedite the discovery of novel therapeutic avenues, as these architectures may be leveraged to predict transcriptional responses to genetic or chemical perturbations in a wide degree of contexts. However, as the associations predicted by these foundation models do not necessarily imply causal regulation, predictions must still be rigorously validated through experimental and functional assays such as the aforementioned, CRISPR perturbation screens, drug-response validations in organoid models, and targeted mechanistic studies.

7.2.2 Guiding the improvement of biomimetic tissue engineering efforts

Human induced pluripotent stem cell (hiPSC)-derived in vitro culture systems that replicate definitive hematopoiesis, producing both definitive-like hematopoietic stem and progenitor cells (HSPCs) and definitive hematopoietic products, have been largely successful (*46*, *454*, *455*). However, it remains unclear whether these populations represent 'true' long-term repopulating HSCs derived from hiPSCs (*454*, *456*, *457*). Additionally, the *in vitro* recapitulation of the unique hematopoietic profiles of primitive and early yolk sac (YS) hematopoiesis remains underexplored.

Several attempts have been made to replicate YS hematopoiesis and myelopoietic processes in vitro using hiPSCs (*89*, *458*, *459*). The ultimate aim of these efforts is the in vitro generation and selective production of either primitive or definitive types of hematopoiesis. The ability to selectively produce progenitors skewed towards either type of hematopoiesis presents an opportunity to generate progenitors with specific differentiation biases.

While recent efforts have claimed to recapitulate both primitive and definitive hematopoiesis, generating culture conditions that reliably replicate only primitive hematopoietic waves remains challenging. Hislop *et al.* developed a human embryoid model named heX-Embryoid from iPSCs that exhibits self-organising cellular programs reminiscent of embryogenesis. This includes the formation of amniotic cavity and bilaminar disc morphologies, as well as the generation of an anterior hypoblast pole and posterior domain. Notably, these embryoids show distinct waves of hematopoiesis, featuring erythroid-, megakaryocyte-, myeloid-, and lymphoid-like cells. The erythroid cells in this culture system initially express globin genes associated with primitive haematopoiesis (*HBE1*), but shift towards a more definitive globin expression profile by day 21 of culture, expressing a higher proportion of gamma globin genes (*459*). Similarly, Tamaoki *et al.* explored the hematopoietic potential of HSPCs generated from hiPSCs. Their study demonstrated that GFP+CD34+CD43+ cells sorted on day 13 represent primitive type HSPCs, which can be induced into erythroid and myeloid lineage cells. Erythroid cells from this time point express primitive globin genes (*HBZ* and *HBE1*), indicating primitive haematopoiesis. Furthermore, CD5+CD7+ T cell progenitors were detected by day 17, and CD4+CD8+ double-positive T cells were detected by day 27 of T cell differentiation, indicating that their culture system can produce definitive HSPCs from day 17 (*458*).

The yolk sac atlas we have generated contextualises YS HSPCs prior to AGM formation, against AGM-derived HSCs from various haematopoietic tissues, including the liver and bone

marrow. Work by Alsinet *et al.* in modelling early YS haematopoiesis through in vitro myelopoiesis from hiPSCs under conditions that mimic the early human YS environment showed that myeloid cells derived from this iPSC culture system displayed transcriptional and epigenetic features characteristic of YS-derived myeloid progenitors (*89*). By contextualising this data against the YS atlas and wider AGM-derived HSC reference, we show that prior to day 21 of culture, this system faithfully recapitulates the YS accelerated route to macrophage differentiation. However, this monocyte-independent pathway is lost after day 31 of culture, in preference of monocyte-derived macrophage differentiation routes.

Our work highlights potential sources of extrinsic HSPC maintenance by endothelial and endodermal populations in the YS, which are gradually ceded. As it remains to be observed if all YS stromal populations are faithfully reproduced in current YS mimetic culture systems, the introduction of predicted HSPC maintenance factors, including *KITLG*, *NOTCH1/2*, *DLL1*, *JAG1,* and *WNT5A* may be crucial for maintaining YS HSPCs and primitive haematopoiesis in these in vitro systems. Our findings underscore the resource value of the YS atlas and suggest that integrating these extrinsic maintenance factors may enhance the maintenance of primitive hematopoietic waves, thereby advancing the field of hematopoietic tissue engineering.

7.3 Future study design considerations

7.3.1 Cell Type Representation

One of the paramount challenges in single-cell sequencing is ensuring comprehensive recovery of all cell types within a tissue, especially when studying changes of specific populations across tissues and development. As discussed in the introduction chapter, this issue is exacerbated by the relationship between cell size and capture efficiency, leading to limitations inherent to droplet-based platforms like 10x Genomics, which may lead to the underrepresentation or omission of rare but functionally significant cell populations. To address this, advancements in enrichment and isolation technologies such as FACS to enhance the isolation and enrichment of specific cells or nuclei populations from large amounts of banked and preserved tissue, such as FFPE samples or frozen tissue sections is needed.

Moreover, recent advancements in spatial technologies, such as Slide-tags, present promising solutions to the challenge of cell type representation. Slide-tags tag frozen tissue sections with spatial barcode oligonucleotides at resolutions of less than 10 micrometres before extracting cellular nuclei. This method circumvents the issues related to cell size by focusing on nuclei, allowing for high-resolution simultaneous spatial profiling of both scATAC-seq and snRNA-seq modalities. This technology integrates multiple modalities, and uses nuclei as input, addressing key capture-efficiency related weaknesses inherent in microfluidic droplet-based single-cell multiomics methodologies (*81*).

7.3.2 Relative Abundance Bias

Single-cell and spatial RNA sequencing experiments inherently impose constraints on the number of cells that can be profiled, leading to relative abundance bias. This bias arises because an increase in one population may cause others to appear to decrease proportionally, regardless of their actual abundance, complicating the interpretation of changes in cell populations. The challenge is further amplified when studying developing organs, which change rapidly in size and cellular composition. Fetal tissues, in particular, are difficult to acquire and rare, resulting in variations in timepoints and temporal snapshots. This scarcity

can lead to incomplete temporal data, which exacerbates relative abundance bias as different developmental stages might be unequally represented.

Spatially resolved techniques such as Slide-tags offer promising solutions by enabling quantification of cell numbers, tissue size, and the spatial context, providing potentially more accurate estimates of cell abundance relative to tissue area. However, even though Slide-tags typically recovers a comparable number of nuclei per experiment as conventional snRNA-seq (e.g., 10X Genomics), it remains susceptible to similar biases, particularly regarding relative abundance, when considering highly imbalanced frequency of cell populations and transient states, particularly in developmental data (*81*). In Slide-tags, this limitation is compounded by the in situ spatial barcoding strategy, whereby only those nuclei and transcripts in close proximity to the barcoded oligonucleotides on the slide are captured, which can introduce further biases if dissociation is incomplete or transcripts are located distally within large or complex tissue morphologies, thus highlighting the importance of robust dissociation optimisation when using Slide-tags to quantify cell abundances.

Moreover, Slide-tags, like other single-nucleus RNA sequencing (snRNA-seq) approaches, exclusively captures nuclear transcripts and therefore omits cytoplasmic mRNAs. This omission can be significant as Bakken et al. 2018 (*460*) demonstrated in matched cortical murine datasets, scRNA-seq typically yields on the order of 11,000 detected genes per cell, whereas snRNA-seq captures closer to 7,000. One key factor is the heavy reliance on intronic reads in snRNA-seq (>50% of reads), compared to scRNA-seq, where <30% of reads align to intronic regions. Consequently, fully spliced transcripts exported to the cytoplasm are systematically underrepresented. However, nuclei are more resistant to mechanical stress than whole cells, facilitating the analysis of large or fragile cell populations and allowing the use of archived frozen tissues that cannot easily yield viable whole-cell suspensions. Excluding highly labile, immediate-early genes in the cytoplasm, can also avoid artifacts introduced by

cell-dissociation stress, particularly for immediate-early genes that may spike artificially under enzymatic or mechanical disruption. However, short-lived cytoplasmic transcripts, or those crucial for translation-level regulation, may be missed entirely. Lastly, these constraints have direct implications for computational methods like RNA velocity (*29*). RNA velocity relies on the ratio of unspliced (nascent) to spliced (mature) transcripts to infer transcriptional dynamics or "directionality" within single cells. In nuclei-based data, the absence of exported mature mRNA distorts this ratio, complicating efforts to faithfully reconstruct temporal or developmental trajectories. For Slide-tags users, this means that while spatial resolution and compatibility with fragile or archived tissues are major strengths, the nuclear-only scope of data collection can skew transcript abundance estimates, potentially biasing any downstream analysis reliant on cytoplasmic mRNA signals. Therefore, these considerations should be taken into account when designing and interpreting Slide-tags experiments especially when interpreting gene regulation and cell-state dynamics.

7.3.3 Inferring causality in gene regulatory networks

Understanding causality within gene regulatory networks in scRNA-seq data is inherently challenging due to the complexity of biological systems and the observational nature of the data, necessitating experimental validation to infer causative relationships. The rarity of human developmental tissues further complicates this, making in vitro models of developing human organs invaluable. These models provide a controlled environment to study the dynamics of gene regulatory programs and their impact on cellular phenotypes, enabling the capture of consistent temporal snapshots. Often derived from hiPSCs, these models can recapitulate key aspects of human organogenesis, allowing for the manipulation and observation of specific transcription factors (TFs) and signalling pathways.

One powerful approach to inferring causality in these systems is Perturb-seq, a technique that combines CRISPR-based gene perturbations with single-cell RNA sequencing. Perturb-seq enables the systematic disruption of individual or multiple genes and the subsequent analysis

of resulting transcriptional changes at single-cell resolution (*376*). This approach allows for the identification of key TFs and other regulatory elements driving gene expression changes and cell fate decisions.

This methodology, if applied to hiPSC models of haematopoiesis, has important implications for studies across the distributed network of developmental haematopoiesis. Capturing consistent temporal snapshots enables the identification of intermediate states of regulation, revealing the dynamic nature of gene regulation and cellular differentiation. This provides insights into the timing and sequence of key regulatory events that drive development and potentially, disease progression.

7.4 Conclusion

The findings presented in this thesis significantly advance our understanding of the multifaceted role the human YS plays in early hematopoiesis, revealing intricate details about its metabolic, biosynthetic, and erythropoiesis-stimulating functions. By elucidating the dynamic transitions and regulatory mechanisms governing YS-derived HSPC maintenance, as well as unique macrophage differentiation routes, we have filled a critical gap left by previous studies that did not extend beyond the early stages of YS function. This work not only enhances our comprehension of early human development but also provides a valuable resource for future research aimed at decoding complex developmental processes. The insights gained here have profound implications for the development of novel therapeutic strategies, including the potential use of CRISPR-Cas9 technology to target specific regulatory elements, elucidated in this thesis, to treat alpha thalassemia genetic disorders and improve biomimetic tissue engineering protocols. Ultimately, this study underscores the importance of early developmental stages in shaping lifelong health, offering new avenues for disease prevention and treatment. As we continue to explore the YS contributions to human development, we open the door to innovative approaches in regenerative medicine and a deeper understanding of the cellular origins of health and disease.

8 References

1. I. Goh, R. A. Botting, A. Rose, S. Webb, J. Engelbert, Y. Gitton, E. Stephenson, M. Quiroga Londoño, M. Mather, N. Mende, I. Imaz-Rosshandler, L. Yang, D. Horsfall, D. Basurto-Lozada, N.-J. Chipampe, V. Rook, J. T. H. Lee, M.-L. Ton, D. Keitley, P. Mazin, M. S. Vijayabaskar, R. Hannah, L. Gambardella, K. Green, S. Ballereau, M. Inoue, E. Tuck, V. Lorenzi, K. Kwakwa, C. Alsinet, B. Olabi, M. Miah, C. Admane, D.-M. Popescu, M. Acres, D. Dixon, T. Ness, R. Coulthard, S. Lisgo, D. J. Henderson, E. Dann, C. Suo, S. J. Kinston, J.-E. Park, K. Polanski, J. Marioni, S. van Dongen, K. B. Meyer, M. de Bruijn, J. Palis, S. Behjati, E. Laurenti, N. K. Wilson, R. Vento-Tormo, A. Chédotal, O. Bayraktar, I. Roberts, L. Jardine, B. Göttgens, S. A. Teichmann, M. Haniffa, Yolk sac cell atlas reveals multiorgan functions during human early development. *Science* **381**, eadd7564 (2023).

2. C. Suo, E. Dann, I. Goh, L. Jardine, V. Kleshchevnikov, J.-E. Park, R. A. Botting, E. Stephenson, J. Engelbert, Z. K. Tuong, K. Polanski, N. Yayon, C. Xu, O. Suchanek, R. Elmentaite, C. D. Conde, P. He, S. Pritchard, M. Miah, C. Moldovan, A. S. Steemers, M. Prete, J. C. Marioni, M. R. Clatworthy, M. Haniffa, S. A. Teichmann, Mapping the developing human immune system across organs, *bioRxiv* (2022)p. 2022.01.17.476665.

3. L. Jardine, S. Webb, I. Goh, M. Quiroga Londoño, G. Reynolds, M. Mather, B. Olabi, E. Stephenson, R. A. Botting, D. Horsfall, J. Engelbert, D. Maunder, N. Mende, C. Murnane, E. Dann, J. McGrath, H. King, I. Kucinski, R. Queen, C. D. Carey, C. Shrubsole, E. Poyner, M. Acres, C. Jones, T. Ness, R. Coulthard, N. Elliott, S. O'Byrne, M. L. R. Haltalli, J. E. Lawrence, S. Lisgo, P. Balogh, K. B. Meyer, E. Prigmore, K. Ambridge, M. S. Jain, M. Efremova, K. Pickard, T. Creasey, J. Bacardit, D. Henderson, J. Coxhead, A. Filby, R. Hussain, D. Dixon, D. McDonald, D.-M. Popescu, M. S. Kowalczyk, B. Li, O. Ashenberg, M. Tabaka, D. Dionne, T. L. Tickle, M. Slyper, O. Rozenblatt-Rosen, A. Regev, S. Behjati, E. Laurenti, N. K. Wilson, A. Roy, B. Göttgens, I. Roberts, S. A. Teichmann, M. Haniffa, Blood and immune development in human fetal bone marrow and Down syndrome. *Nature* **598**, 327–331 (2021).

4. M. Miah, I. Goh, M. Haniffa, Prenatal development and function of human mononuclear phagocytes. *Front. Cell Dev. Biol.* **9**, 649937 (2021).

5. G. Reynolds, P. Vegh, J. Fletcher, E. F. M. Poyner, E. Stephenson, I. Goh, R. A. Botting, N. Huang, B. Olabi, A. Dubois, D. Dixon, K. Green, D. Maunder, J. Engelbert, M. Efremova, K. Polański, L. Jardine, C. Jones, T. Ness, D. Horsfall, J. McGrath, C. Carey, D.-M. Popescu, S. Webb, X.-N. Wang, B. Sayer, J.-E. Park, V. A. Negri, D. Belokhvostova, M. D. Lynch, D. McDonald, A. Filby, T. Hagai, K. B. Meyer, A. Husain, J. Coxhead, R. Vento-Tormo, S. Behjati, S. Lisgo, A.-C. Villani, J. Bacardit, P. H. Jones, E. A. O'Toole, G. S. Ogg, N. Rajan, N. J. Reynolds, S. A. Teichmann, F. M. Watt, M. Haniffa, Developmental cell programs are co-opted in inflammatory skin disease. *Science* **371** (2021).

6. D.-M. Popescu, R. A. Botting, E. Stephenson, K. Green, S. Webb, L. Jardine, E. F. Calderbank, K. Polanski, I. Goh, M. Efremova, M. Acres, D. Maunder, P. Vegh, Y. Gitton, J.-E. Park, R. Vento-Tormo, Z. Miao, D. Dixon, R. Rowell, D. McDonald, J. Fletcher, E. Poyner, G. Reynolds, M. Mather, C. Moldovan, L. Mamanova, F. Greig, M. D. Young, K. B. Meyer, S. Lisgo, J. Bacardit, A. Fuller, B. Millar, B. Innes, S. Lindsay, M. J. T. Stubbington, M. S. Kowalczyk, B. Li, O. Ashenberg, M. Tabaka, D. Dionne, T. L. Tickle, M. Slyper, O. Rozenblatt-Rosen, A. Filby, P. Carey, A.-C. Villani, A. Roy, A. Regev, A. Chédotal, I. Roberts, B. Göttgens, S. Behjati, E. Laurenti, S. A. Teichmann, M. Haniffa, Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019).

7. J.-E. Park, R. A. Botting, C. Domínguez Conde, D.-M. Popescu, M. Lavaert, D. J. Kunz, I. Goh, E. Stephenson, R. Ragazzini, E. Tuck, A. Wilbrey-Clark, K. Roberts, V. R. Kedlian, J. R. Ferdinand, X. He, S. Webb, D. Maunder, N. Vandamme, K. T. Mahbubani, K. Polanski, L. Mamanova, L. Bolt, D. Crossland, F. de Rita, A. Fuller, A. Filby, G. Reynolds, D. Dixon, K. Saeb-Parsy, S. Lisgo, D. Henderson, R. Vento-Tormo, O. A. Bayraktar, R. A. Barker, K. B. Meyer, Y. Saeys, P. Bonfanti, S. Behjati, M. R. Clatworthy, T. Taghon, M. Haniffa, S. A. Teichmann, A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367** (2020).

8. R. Elmentaite, N. Kumasaka, K. Roberts, A. Fleming, E. Dann, H. W. King, V. Kleshchevnikov, M. Dabrowska, S. Pritchard, L. Bolt, S. F. Vieira, L. Mamanova, N. Huang, F. Perrone, I. Goh Kai'En, S. N. Lisgo, M. Katan, S. Leonard, T. R. W. Oliver, C. E. Hook, K. Nayak, L. S. Campos, C. Domínguez Conde, E. Stephenson, J. Engelbert, R. A. Botting, K. Polanski, S. van Dongen, M. Patel, M. D. Morgan, J. C. Marioni, O. A. Bayraktar, K. B. Meyer, X. He, R. A. Barker, H. H. Uhlig, K. T. Mahbubani, K. Saeb-Parsy, M. Zilbauer, M. R. Clatworthy, M. Haniffa, K. R. James, S. A. Teichmann, Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).

9. P. He, K. Lim, D. Sun, J. P. Pett, Q. Jeng, K. Polanski, Z. Dong, L. Bolt, L. Richardson, L. Mamanova, M. Dabrowska, A. Wilbrey-Clark, E. Madissoon, Z. K. Tuong, E. Dann, C. Suo, I. Goh, M. Yoshida, M. Z. Nikolić, S. M. Janes, X. He, R. A. Barker, S. A. Teichmann, J. C. Marioni, K. B. Meyer, E. L. Rawlins, A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell* **185**, 4841–4860.e25 (2022).

10. M. Q. Londoño, N. Mende, E. Stephenson, D. Iskander, S. Webb, I. Goh, V. M. Shanmugiah, A. Roy, I. Roberts, E. Laurenti, M. Haniffa, N. K. Wilson, B. Göttgens, 3166 – a protein‐transcriptome atlas of haematopoiesis across the human lifespan. *Exp. Hematol.* **111**, S127–S128 (2022).

11. N. Alkon, W. M. Bauer, T. Krausgruber, I. Goh, J. Griss, V. Nguyen, B. Reininger, C. Bangert, C. Staud, P. M. Brunner, C. Bock, M. Haniffa, G. Stingl, Single-cell analysis reveals innate lymphoid cell lineage infidelity in atopic dermatitis. *J. Allergy Clin. Immunol.* **149**, 624–639 (2022).

12. E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, K. Bach, W. Sungnak, K. B. Worlock, M. Yoshida, N. Kumasaka, K. Kania, J. Engelbert, B. Olabi, J. S. Spegarova, N. K. Wilson, N. Mende, L. Jardine, L. C. S. Gardner, I. Goh, D. Horsfall, J. McGrath, S. Webb, M. W. Mather, R. G. H. Lindeboom, E. Dann, N. Huang, K. Polanski, E. Prigmore, F. Gothe, J. Scott, R. P. Payne, K. F. Baker, A. T. Hanrath, I. C. D. Schim van der Loeff, A. S. Barr, A. Sanchez-Gonzalez, L. Bergamaschi, F. Mescia, J. L. Barnes, E. Kilich, A. de Wilton, A. Saigal, A. Saleh, S. M. Janes, C. M. Smith, N. Gopee, C. Wilson, P. Coupland, J. M. Coxhead, V. Y. Kiselev, S. van Dongen, J. Bacardit, H. W. King, Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration, A. J. Rostron, A. J. Simpson, S. Hambleton, E. Laurenti, P. A. Lyons, K. B. Meyer, M. Z. Nikolić, C. J. A. Duncan, K. G. C. Smith, S. A. Teichmann, M. R. Clatworthy, J. C. Marioni, B. Göttgens, M. Haniffa, Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).

13. N. H. Gopee, N. Huang, B. Olabi, C. Admane, R. A. Botting, A. R. Foster, F. Torabi, E. Winheim, D. Sumanaweera, I. Goh, M. Miah, E. Stephenson, W. M. Tun, P. Moghimi, B. Rumney, P. He, S. Lawrence, K. Roberts, K. Sidhpura, J. Englebert, L. Jardine, G. Reynolds, A. Rose, C. Ganier, V. Rowe, S. Pritchard, I. Mulas, J. Fletcher, D.-M. Popescu, E. Poyner, A. Dubois, A. Filby, S. Lisgo, R. A. Barker, J.-E. Park, R. Vento-Tormo, P. A. Le, S. Serdy, J. Kim, C. Deakin, J. Lee, M. Nikolova, N. Rajan, S. Ballereau, T. Li, J. Moore, D. Horsfall, D. B. Lozada, E. A. O'Toole, B. Treutlein, O. Bayraktar, M. Kasper, P. Mazin, L. Gambardella, K. Koehler, S. A. Teichmann, M.

Haniffa, A human prenatal skin cell atlas reveals immune cell regulation of skin morphogenesis, *bioRxiv* (2023)p. 2023.10.12.556307.

14. G. Brady, M. Barbara, N. Iscove, Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. (1990).

15. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, M. A. Surani, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

16. J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, J. M. Trent, Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).

17. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

18. J. Cao, D. R. O'Day, H. A. Pliner, P. D. Kingsley, M. Deng, R. M. Daza, M. A. Zager, K. A. Aldinger, R. Blecher-Gonen, F. Zhang, M. Spielmann, J. Palis, D. Doherty, F. J. Steemers, I. A. Glass, C. Trapnell, J. Shendure, A human cell atlas of fetal gene expression. *Science* **370** (2020).

19. A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, S. Lazo, L. Jardine, D. Dixon, E. Stephenson, E. Nilsson, I. Grundberg, D. McDonald, A. Filby, W. Li, P. L. De Jager, O. Rozenblatt-Rosen, A. A. Lane, M. Haniffa, A. Regev, N. Hacohen, Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356** (2017).

20. S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, R. Sandberg, Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

21. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

22. J. D. Buenrostro, B. Wu, H. Y. Chang, W. J. Greenleaf, ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).

23. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

24. G. M. Edelman, J. A. Gally, Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13763–13768 (2001).

25. Y.-A. Kim, S. Wuchty, T. M. Przytycka, Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput. Biol.* **7**, e1001095 (2011).

26. J. R. Shaffer, E. Feingold, M. L. Marazita, Genome-wide association studies: prospects and challenges for oral health. *J. Dent. Res.* **91**, 637–641 (2012).

27. E. Cano-Gamez, G. Trynka, From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).

28. A. Haque, J. Engel, S. A. Teichmann, T. Lönnberg, A practical guide to single-cell

RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

29. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, P. V. Kharchenko, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

30. S. Luo, Z. Zhang, Z. Wang, X. Yang, X. Chen, T. Zhou, J. Zhang, Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model. *R Soc Open Sci* **10**, 221057 (2023).

31. W. Tang, A. C. S. Jørgensen, S. Marguerat, P. Thomas, V. Shahrezaei, Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. *Bioinformatics* **39** (2023).

32. A. Mukherjee, S. Abraham, A. Singh, S. Balaji, K. S. Mukunthan, From Data to Cure: A Comprehensive Exploration of Multi-omics Data Analysis for Targeted Therapies. *Mol. Biotechnol.*, doi: 10.1007/s12033-024-01133-6 (2024).

33. S. Rajderkar, I. Barozzi, Y. Zhu, R. Hu, Y. Zhang, B. Li, A. Alcaina Caro, Y. Fukuda-Yuzawa, G. Kelman, A. Akeza, M. J. Blow, Q. Pham, A. N. Harrington, J. Godoy, E. M. Meky, K. von Maydell, R. D. Hunter, J. A. Akiyama, C. S. Novak, I. Plajzer-Frick, V. Afzal, S. Tran, J. Lopez-Rios, M. E. Talkowski, K. C. K. Lloyd, B. Ren, D. E. Dickel, A. Visel, L. A. Pennacchio, Topologically associating domain boundaries are required for normal genome function. *Commun Biol* **6**, 435 (2023).

34. S. J. Waddell, S. J. Popper, K. H. Rubins, M. J. Griffiths, P. O. Brown, M. Levin, D. A. Relman, Dissecting interferon-induced transcriptional programs in human peripheral blood cells. *PLoS One* **5**, e9753 (2010).

35. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

36. T. Zhu, K. Liao, R. Zhou, C. Xia, W. Xie, ATAC-seq with unique molecular identifiers improves quantification and footprinting. *Commun Biol* **3**, 675 (2020).

37. V. Svensson, R. Vento-Tormo, S. A. Teichmann, Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

38. V. M. Peterson, K. X. Zhang, N. Kumar, J. Wong, L. Li, D. C. Wilson, R. Moore, T. K. McClanahan, S. Sadekova, J. A. Klappenbach, Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).

39. "Scalable Approaches for Inferring Chromatin States and Lineages of Human Cells," (2020).

40. L. Pan, P. Parini, R. Tremmel, J. Loscalzo, V. M. Lauschke, B. A. Maron, P. Paci, I. Ernberg, N. S. Tan, Z. Liao, W. Yin, S. Rengarajan, X. Li, SCA Consortium, Single Cell Atlas: a single-cell multi-omics human cell encyclopedia. *Genome Biol.* **25**, 104 (2024).

41. Y. Kim, I. Kim, K. Shin, A new era of stem cell and developmental biology: from blastoids to synthetic embryos and beyond. *Exp. Mol. Med.* **55**, 2127–2137 (2023).

42. B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, A. F. Schier, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

43. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).

44. R. Vento-Tormo, L. Garcia-Alonso, V. Lorenzi, C. Mazzeo, C. Sancho-Serra, K. Roberts, J. Engelbert, J. Alves-Lopes, M. Marečková, R. Botting, T. Li, B. Crespo, S. van Dongen, V. Kiselev, E. Prigmore, A. Moffett, M. Herbert, O. A. Bayraktar, A. Surani, M. Haniffa, Single-cell roadmap of human gonadal development. doi: 10.21203/rs.3.rs-496470/v1 (2021).

45. E. Braun, M. Danan-Gotthold, L. E. Borm, K. W. Lee, E. Vinsland, P. Lönnerberg, L. Hu, X. Li, X. He, Ž. Andrusivová, J. Lundeberg, R. A. Barker, E. Arenas, E. Sundström, S. Linnarsson, Comprehensive cell atlas of the first-trimester developing human brain. *Science* **382**, eadf1226 (2023).

46. V. Calvanese, S. Capellera-Garcia, F. Ma, I. Fares, S. Liebscher, E. S. Ng, S. Ekstrand, J. Aguadé-Gorgorió, A. Vavilina, D. Lefaudeux, B. Nadel, J. Y. Li, Y. Wang, L. K. Lee, R. Ardehali, M. L. Iruela-Arispe, M. Pellegrini, E. G. Stanley, A. G. Elefanty, K. Schenke-Layland, H. K. A. Mikkola, Mapping human haematopoietic stem cells from haemogenic endothelium to birth. *Nature* **604**, 534–540 (2022).

47. M. Haniffa, D. Taylor, S. Linnarsson, B. J. Aronow, G. D. Bader, R. A. Barker, P. G. Camara, J. G. Camp, A. Chédotal, A. Copp, H. C. Etchevers, P. Giacobini, B. Göttgens, G. Guo, A. Hupalowska, K. R. James, E. Kirby, A. Kriegstein, J. Lundeberg, J. C. Marioni, K. B. Meyer, K. K. Niakan, M. Nilsson, B. Olabi, D. Pe'er, A. Regev, J. Rood, O. Rozenblatt-Rosen, R. Satija, S. A. Teichmann, B. Treutlein, R. Vento-Tormo, S. Webb, Human Cell Atlas Developmental Biological Network, A roadmap for the Human Developmental Cell Atlas. *Nature* **597**, 196–205 (2021).

48. A. Cazzola, G. Cazzaniga, A. Biondi, R. Meneveri, S. Brunelli, E. Azzoni, Prenatal Origin of Pediatric Leukemia: Lessons From Hematopoietic Development. *Front Cell Dev Biol* **8**, 618164 (2020).

49. M. A. Karalexi, N. Dessypris, X. Ma, L. G. Spector, E. Marcotte, J. Clavel, M. S. Pombo-de-Oliveira, J. E. Heck, E. Roman, B. A. Mueller, J. Hansen, A. Auvinen, P.-C. Lee, J. Schüz, C. Magnani, A. M. Mora, J. D. Dockerty, M. E. Scheurer, R. Wang, A. Bonaventure, E. Kane, D. R. Doody, NARECHEM-ST Group, FRECCLE Group, F. Erdmann, A. Y. Kang, C. Metayer, E. Milne, E. T. Petridou, Age-, sex- and disease subtype-related foetal growth differentials in childhood acute myeloid leukaemia risk: A Childhood Leukemia International Consortium analysis. *Eur. J. Cancer* **130**, 1–11 (2020).

50. T. R. Jackson, R. E. Ling, A. Roy, The Origin of B-cells: Human Fetal B Cell Development and Implications for the Pathogenesis of Childhood Acute Lymphoblastic Leukemia. *Front. Immunol.* **12**, 637975 (2021).

51. L. M. Joseph, R. G. Toedebusch, E. Debebe, A. H. Bastian, C. A. Lucchesi, S. Syed-Quadri, L. A. Wittenburg, X. Chen, F. J. Meyers, C. M. Toedebusch, Microglia-Derived Olfactomedin-like 3 Is a Potent Angiogenic Factor in Primary Mouse Brain Endothelial Cells: A Novel Target for Glioblastoma. *Int. J. Mol. Sci.* **23** (2022).

52. G. Wang, K. Zhong, Z. Wang, Z. Zhang, X. Tang, A. Tong, L. Zhou, Tumor-associated microglia and macrophages in glioblastoma: From basic insights to therapeutic opportunities. *Front. Immunol.* **13**, 964898 (2022).

53. S. Brandenburg, A. Müller, K. Turkowski, Y. T. Radev, S. Rot, C. Schmidt, A. D. Bungert, G. Acker, A. Schorr, A. Hippe, K. Miller, F. L. Heppner, B. Homey, P. Vajkoczy, Resident microglia rather than peripheral macrophages promote vascularization in brain tumors and are source of alternative pro-angiogenic factors. *Acta Neuropathol.* **131**, 365–378 (2016).

54. B. Van de Sande, J. S. Lee, E. Mutasa-Gottgens, B. Naughton, W. Bacon, J. Manning, Y. Wang, J. Pollard, M. Mendez, J. Hill, N. Kumar, X. Cao, X. Chen, M. Khaladkar, J. Wen, A. Leach, E. Ferran, Applications of single-cell RNA sequencing in drug discovery and development. *Nat. Rev. Drug Discov.* **22**, 496–520 (2023).

55. K. Kanemaru, J. Cranley, D. Muraro, A. M. A. Miranda, S. Y. Ho, A. Wilbrey-Clark, J. Patrick Pett, K. Polanski, L. Richardson, M. Litvinukova, N. Kumasaka, Y. Qin, Z. Jablonska, C. I. Semprich, L. Mach, M. Dabrowska, N. Richoz, L. Bolt, L. Mamanova, R. Kapuge, S. N. Barnett, S. Perera, C. Talavera-López, I. Mulas, K. T. Mahbubani, L. Tuck, L. Wang, M. M. Huang, M. Prete, S. Pritchard, J. Dark, K. Saeb-Parsy, M. Patel, M. R. Clatworthy, N. Hübner, R. A. Chowdhury, M. Noseda, S. A. Teichmann, Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–810 (2023).

56. F. Sanger, S. Nicklen, A. R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

57. E. M. Brown, B. J. Barratt, "The HapMap– A haplotype map of the human genome" in *Bioinformatics for Geneticists* (John Wiley & Sons, Ltd, Chichester, UK, 2007), pp. 33–58.

58. J. N. Hirschhorn, M. J. Daly, Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).

59. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).

60. E. R. Mardis, Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).

61. S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, M. J. Bamshad, Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).

62. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

63. D. Pinkel, T. Straume, J. W. Gray, Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 2934–2938 (1986).

64. S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, D. Altshuler, The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

65. E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. B. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A. Hitman, A. D. Morris, A. S. F. Doney, Wellcome Trust Case Control Consortium (WTCCC), M. I. McCarthy, A. T. Hattersley, Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).

66. K. L. McNally, K. L. Childs, R. Bohnert, R. M. Davidson, K. Zhao, V. J. Ulat, G. Zeller, R. M. Clark, D. R. Hoen, T. E. Bureau, R. Stokowski, D. G. Ballinger, K. A. Frazer, D. R. Cox, B. Padhukasahasram, C. D. Bustamante, D. Weigel, D. J. Mackill, R. M. Bruskiewich, G. Rätsch, C. R. Buell, H. Leung, J. E. Leach, Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences* **106**,

12273–12278 (2009).

67. J. Marchini, B. Howie, Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

68. B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. A. McCarroll, C. Langenberg, O. M. Hofmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E. Boerwinkle, L. L. Bonnycastle, K. Bengtsson Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. F. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U. Jackson, P. R. V. Johnson, T. Jørgensen, W. H. L. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. B. Perry, A.-K. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. W. Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. M. van Dam, T. W. van Haeften, T. van Herpt, J. V. van Vliet-Ostaptchouk, G. B. Walters, M. N. Weedon, C. Wijmenga, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. J. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. A. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, B. O. Boehm, H. Campbell, M. J. Daly, A. T. Hattersley, F. B. Hu, J. B. Meigs, J. S. Pankow, O. Pedersen, H.-E. Wichmann, I. Barroso, J. C. Florez, T. M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M. I. McCarthy, MAGIC investigators, GIANT Consortium, Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).

69. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

70. S. C. Bendall, E. F. Simonds, P. Qiu, E.-A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er, S. D. Tanner, G. P. Nolan, Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).

71. C. G. Palii, Q. Cheng, M. A. Gillespie, P. Shannon, M. Mazurczyk, G. Napolitani, N. D. Price, J. A. Ranish, E. Morrissey, D. R. Higgs, M. Brand, Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate. *Cell Stem Cell* **24**, 812–820.e5 (2019).

72. S. Lee, H. M. Vu, J.-H. Lee, H. Lim, M.-S. Kim, Advances in Mass Spectrometry-Based Single Cell Analysis. *Biology* **12** (2023).

73. TotalSeq™-A Human Universal Cocktail, V1.0. https://www.biolegend.com/en-ie/products/totalseq-a-human-universal-cocktail-v1-20321.

74. S. Preissl, R. Fang, H. Huang, Y. Zhao, R. Raviram, D. U. Gorkin, Y. Zhang, B. C. Sos, V. Afzal, D. E. Dickel, S. Kuan, A. Visel, L. A. Pennacchio, K. Zhang, B. Ren, Author Correction: Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 1015 (2018).

75. Y. Hu, K. Huang, Q. An, G. Du, G. Hu, J. Xue, X. Zhu, C.-Y. Wang, Z. Xue, G. Fan,

Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 88 (2016).

76. S. J. Clark, R. Argelaguet, C.-A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, W. Reik, scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

77. R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, O. Stegle, MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

78. T. Ashuach, M. I. Gabitto, R. V. Koodli, G.-A. Saldi, M. I. Jordan, N. Yosef, MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).

79. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

80. V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, O. A. Bayraktar, Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

81. A. J. C. Russell, J. A. Weir, N. M. Nadaf, M. Shabet, V. Kumar, S. Kambhampati, R. Raichur, G. J. Marrero, S. Liu, K. S. Balderrama, C. R. Vanderburg, V. Shanmugam, L. Tian, J. B. Iorgulescu, C. H. Yoon, C. J. Wu, E. Z. Macosko, F. Chen, Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* **625**, 101–109 (2024).

82. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

83. M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, D. Pe'er, F. J. Theis, CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).

84. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).

85. C. Suo, E. Dann, I. Goh, L. Jardine, V. Kleshchevnikov, J.-E. Park, R. A. Botting, E. Stephenson, J. Engelbert, Z. K. Tuong, K. Polanski, N. Yayon, C. Xu, O. Suchanek, R. Elmentaite, C. Domínguez Conde, P. He, S. Pritchard, M. Miah, C. Moldovan, A. S. Steemers, P. Mazin, M. Prete, D. Horsfall, J. C. Marioni, M. R. Clatworthy, M. Haniffa, S. A. Teichmann, Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).

86. M. M. Gladka, B. Molenaar, H. de Ruiter, S. van der Elst, H. Tsui, D. Versteeg, G. P. A. Lacraz, M. M. H. Huibers, A. van Oudenaarden, E. van Rooij, Single-Cell Sequencing of the Healthy and Diseased Heart Reveals Cytoskeleton-Associated Protein 4 as a New Modulator of Fibroblasts Activation. *Circulation* **138**, 166–180 (2018).

87. L. Heumos, A. C. Schaar, C. Lance, A. Litinetskaya, F. Drost, L. Zappia, M. D. Lücken, D. C. Strobl, J. Henao, F. Curion, Single-cell Best Practices Consortium, H. B. Schiller, F. J. Theis, Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).

88. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, F. J. Theis, Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

89. C. Alsinet, M. N. Primo, V. Lorenzi, E. Bello, I. Kelava, C. P. Jones, R. Vilarrasa-Blasi, C. Sancho-Serra, A. J. Knights, J.-E. Park, B. S. Wyspianska, G. Trynka, D. F. Tough, A. Bassett, D. J. Gaffney, D. Alvarez-Errico, R. Vento-Tormo, Robust temporal map of human in vitro myelopoiesis using single-cell genomics. *Nat. Commun.* **13**, 2885 (2022).

90. A. J. King, D. Songdej, D. J. Downes, R. A. Beagrie, S. Liu, M. Buckley, P. Hua, M. C. Suciu, A. Marieke Oudelaar, L. L. P. Hanssen, D. Jeziorska, N. Roberts, S. J. Carpenter, H. Francis, J. Telenius, A.-A. Olijnik, J. A. Sharpe, J. Sloane-Stanley, J. Eglinton, M. T. Kassouf, S. H. Orkin, L. A. Pennacchio, J. O. J. Davies, J. R. Hughes, D. R. Higgs, C. Babbs, Reactivation of a developmentally silenced embryonic globin gene. *Nat. Commun.* **12**, 4439 (2021).

91. L. Wang, J. Jung, H. Babikir, K. Shamardani, S. Jain, X. Feng, N. Gupta, S. Rosi, S. Chang, D. Raleigh, D. Solomon, J. J. Phillips, A. A. Diaz, A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets. *Nat Cancer* **3**, 1534–1552 (2022).

92. A. L. Ji, A. J. Rubin, K. Thrane, S. Jiang, D. L. Reynolds, R. M. Meyers, M. G. Guo, B. M. George, A. Mollbrink, J. Bergenstråhle, L. Larsson, Y. Bai, B. Zhu, A. Bhaduri, J. M. Meyers, X. Rovira-Clavé, S. T. Hollmig, S. Z. Aasi, G. P. Nolan, J. Lundeberg, P. A. Khavari, Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 497–514.e22 (2020).

93. C. Ross, T. E. Boroviak, Origin and function of the yolk sac in primate embryogenesis. *Nat. Commun.* **11**, 3760 (2020).

94. G. C. Schoenwolf, S. B. Bleyl, P. R. Brauer, P. H. Francis-West, *Larsen's Human Embryology* (Elsevier Health Sciences, 2014).

95. J. Kerwin, Y. Yang, P. Merchan, S. Sarma, J. Thompson, X. Wang, J. Sandoval, L. Puelles, R. Baldock, S. Lindsay, The HUDSEN Atlas: a three-dimensional (3D) spatial framework for studying gene expression in the developing human brain. *J. Anat.* **217**, 289–299 (2010).

96. Carnegie stage table. https://embryology.med.unsw.edu.au/embryology/index.php/Carnegie_stage_table.

97. S. Flierman, M. Tijsterman, M. Rousian, B. S. de Bakker, Discrepancies in Embryonic Staging: Towards a Gold Standard. *Life* **13** (2023).

98. A. M. Carter, A. C. Enders, Placentation in mammals: Definitive placenta, yolk sac, and paraplacenta. *Theriogenology* **86**, 278–287 (2016).

99. A. Malassiné, J. L. Frendo, D. Evain-Brion, A comparison of placental development and endocrine functions between the human and mouse model. *Hum. Reprod. Update* **9**, 531–539 (2003).

100. H. W. Mossman, *Vertebrate Fetal Membranes: Comparative Ontogeny and Morphology, Evolution, Phylogenetic Significance, Basic Functions, Research Opportunities* (Rutgers University Press, 1987).

101. T. Cindrova-Davies, E. Jauniaux, M. G. Elliot, S. Gong, G. J. Burton, D. S. Charnock-Jones, RNA-seq reveals conservation of function among the yolk sacs of human, mouse, and chicken. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4753–E4761 (2017).

102. T. Yamane, Mouse Yolk Sac Hematopoiesis. *Front. Cell Dev. Biol.* **0** (2018).

103. J. Palis, J. Malik, K. E. McGrath, P. D. Kingsley, Primitive erythropoiesis in the mammalian embryo. *Int. J. Dev. Biol.* **54**, 1011–1018 (2010).

104. G. Canu, C. Ruhrberg, First blood: the endothelial origins of hematopoietic progenitors. *Angiogenesis* **24**, 199–211 (2021).

105. Y. Zeng, J. He, Z. Bai, Z. Li, Y. Gong, C. Liu, Y. Ni, J. Du, C. Ma, L. Bian, Y. Lan, B. Liu, Tracing the first hematopoietic stem cell generation in human embryo by single-cell RNA sequencing. *Cell Res.* **29**, 881–894 (2019).

106. P. D. Kingsley, J. Malik, R. L. Emerson, T. P. Bushnell, K. E. McGrath, L. A. Bloedorn, M. Bulger, J. Palis, "Maturational" globin switching in primary primitive erythroid cells. *Blood* **107**, 1665–1672 (2006).

107. J. Palis, Primitive and definitive erythropoiesis in mammals. [Preprint] (2014). https://doi.org/10.3389/fphys.2014.00003.

108. M.-L. N. Ton, D. Keitley, B. Theeuwes, C. Guibentif, J. Ahnfelt-Rønne, T. K. Andreassen, F. J. Calero-Nieto, I. Imaz-Rosshandler, B. Pijuan-Sala, J. Nichols, È. Benito-Gutiérrez, J. C. Marioni, B. Göttgens, Rabbit Development as a Model for Single Cell Comparative Genomics, *bioRxiv* (2022)p. 2022.10.06.510971.

109. G. Meuret, Origin, ontogeny, and kinetics of mononuclear phagocytes. *Adv. Exp. Med. Biol.* **73 PT-A**, 71–81 (1976).

110. R. M. Steinman, G. Kaplan, M. D. Witmer, Z. A. Cohn, Identification of a novel cell type in peripheral lymphoid organs of mice. V. Purification of spleen dendritic cells, new surface markers, and maintenance in vitro. *J. Exp. Med.* **149**, 1–16 (1979).

111. A. I. Tauber, Metchnikoff and the phagocytosis theory. *Nat. Rev. Mol. Cell Biol.* **4**, 897–901 (2003).

112. F. Ginhoux, S. Jung, Monocytes and macrophages: developmental pathways and tissue homeostasis. *Nat. Rev. Immunol.* **14**, 392–404 (2014).

113. G. Hoeffel, J. Chen, Y. Lavin, D. Low, F. F. Almeida, P. See, A. E. Beaudin, J. Lum, I. Low, E. C. Forsberg, M. Poidinger, F. Zolezzi, A. Larbi, L. G. Ng, J. K. Y. Chan, M. Greter, B. Becher, I. M. Samokhvalov, M. Merad, F. Ginhoux, C-Myb(+) erythro-myeloid progenitor-derived fetal monocytes give rise to adult tissue-resident macrophages. *Immunity* **42**, 665–678 (2015).

114. G. Hoeffel, F. Ginhoux, Ontogeny of Tissue-Resident Macrophages. *Front. Immunol.* **6**, 486 (2015).

115. F. Ginhoux, M. Guilliams, Tissue-Resident Macrophage Ontogeny and Homeostasis. *Immunity* **44**, 439–449 (2016).

116. A. Ivanovs, S. Rybtsov, E. S. Ng, E. G. Stanley, A. G. Elefanty, A. Medvinsky, Human haematopoietic stem cell development: from the embryo to the dish. *Development* **144**, 2323–2337 (2017).

117. Z. Bian, Y. Gong, T. Huang, C. Z. W. Lee, L. Bian, Z. Bai, H. Shi, Y. Zeng, C. Liu, J. He, J. Zhou, X. Li, Z. Li, Y. Ni, C. Ma, L. Cui, R. Zhang, J. K. Y. Chan, L. G. Ng, Y. Lan, F. Ginhoux, B. Liu, Deciphering human macrophage development at single-cell resolution. *Nature* **582**, 571–576 (2020).

118. S. A. Dick, A. Wong, H. Hamidzada, S. Nejat, R. Nechanitzky, S. Vohra, B. Mueller, R. Zaman,

C. Kantores, L. Aronoff, A. Momen, D. Nechanitzky, W. Y. Li, P. Ramachandran, S. Q. Crome, B. Becher, M. I. Cybulsky, F. Billia, S. Keshavjee, S. Mital, C. S. Robbins, T. W. Mak, S. Epelman, Three tissue resident macrophage subsets coexist across organs with conserved origins and life cycles. *Sci Immunol* **7**, eabf7777 (2022).

119. M. Haniffa, V. Bigley, M. Collin, Human mononuclear phagocyte system reunited. *Semin. Cell Dev. Biol.* **41**, 59–69 (2015).

120. D. A. Hume, W. Allan, J. Golder, R. W. Stephens, W. F. Doe, H. S. Warren, Preparation and characterization of human bone marrow-derived macrophages. *J. Leukoc. Biol.* **38**, 541–552 (1985).

121. K. Murphy, Weaver C. Janeway's immunobiology. *New York London: Garland Science* (2016).

122. L. Ziegler-Heitbrock, P. Ancuta, S. Crowe, M. Dalod, V. Grau, D. N. Hart, P. J. M. Leenen, Y.-J. Liu, G. MacPherson, G. J. Randolph, J. Scherberich, J. Schmitz, K. Shortman, S. Sozzani, H. Strobl, M. Zembala, J. M. Austyn, M. B. Lutz, Nomenclature of monocytes and dendritic cells in blood. *Blood* **116**, e74–80 (2010).

123. K. C. M. Jeucken, J. J. Koning, R. E. Mebius, S. W. Tas, The Role of Endothelial Cells and TNF-Receptor Superfamily Members in Lymphoid Organogenesis and Function During Health and Inflammation. *Front. Immunol.* **10**, 2700 (2019).

124. E. Fahey, S. L. Doyle, IL-1 Family Cytokine Regulation of Vascular Permeability and Angiogenesis. *Front. Immunol.* **10**, 1426 (2019).

125. E. Voronov, Y. Carmi, R. N. Apte, The role IL-1 in tumor-mediated angiogenesis. *Front. Physiol.* **5**, 114 (2014).

126. Y. Wang, J. Xu, X. Zhang, C. Wang, Y. Huang, K. Dai, X. Zhang, TNF-α-induced LRG1 promotes angiogenesis and mesenchymal stem cell migration in the subchondral bone during osteoarthritis. *Cell Death Dis.* **8**, e2715 (2017).

127. E. Giraudo, L. Primo, E. Audero, H.-P. Gerber, P. Koolwijk, S. Soker, M. Klagsbrun, N. Ferrara, F. Bussolino, Tumor Necrosis Factor-α Regulates Expression of Vascular Endothelial Growth Factor Receptor-2 and of Its Co-receptor Neuropilin-1 in Human Vascular Endothelial Cells*. *J. Biol. Chem.* **273**, 22128–22135 (1998).

128. S. Z. Chong, M. Evrard, S. Devi, J. Chen, J. Y. Lim, P. See, Y. Zhang, J. M. Adrover, B. Lee, L. Tan, J. L. Y. Li, K. H. Liong, C. Phua, A. Balachander, A. Boey, D. Liebl, S. M. Tan, J. K. Y. Chan, K. Balabanian, J. E. Harris, M. Bianchini, C. Weber, J. Duchene, J. Lum, M. Poidinger, Q. Chen, L. Rénia, C.-I. Wang, A. Larbi, G. J. Randolph, W. Weninger, M. R. Looney, M. F. Krummel, S. K. Biswas, F. Ginhoux, A. Hidalgo, F. Bachelerie, L. G. Ng, CXCR4 identifies transitional bone marrow premonocytes that replenish the mature monocyte pool for peripheral responses. *J. Exp. Med.* **213**, 2293–2314 (2016).

129. P. Rantakari, N. Jäppinen, E. Lokka, E. Mokkala, H. Gerke, E. Peuhu, J. Ivaska, K. Elima, K. Auvinen, M. Salmi, Fetal liver endothelium regulates the seeding of tissue-resident macrophages. *Nature* **538**, 392–396 (2016).

130. E. R. Krow-Lucal, C. C. Kim, T. D. Burt, J. M. McCune, Distinct functional programming of human fetal and adult monocytes. *Blood* **123**, 1897–1904 (2014).

131. R. Vento-Tormo, M. Efremova, R. A. Botting, M. Y. Turco, M. Vento-Tormo, K. B. Meyer, J.-E. Park, E. Stephenson, K. Polański, A. Goncalves, L. Gardner, S. Holmqvist, J. Henriksson, A. Zou, A. M. Sharkey, B. Millar, B. Innes, L. Wood, A. Wilbrey-Clark, R. P. Payne, M. A.

Ivarsson, S. Lisgo, A. Filby, D. H. Rowitch, J. N. Bulmer, G. J. Wright, M. J. T. Stubbington, M. Haniffa, A. Moffett, S. A. Teichmann, Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).

132. L. Banaei-Bouchareb, M. Peuchmaur, P. Czernichow, M. Polak, A transient microenvironment loaded mainly with macrophages in the early developing human pancreas. *J. Endocrinol.* **188**, 467–480 (2006).

133. D. A. Menassa, D. Gomez-Nicola, Microglial Dynamics During Human Brain Development. *Front. Immunol.* **9**, 1014 (2018).

134. L. Reyes, T. G. Golos, Hofbauer Cells: Their Role in Healthy and Complicated Pregnancy. *Front. Immunol.* **9**, 2628 (2018).

135. J. R. Thomas, A. Appios, X. Zhao, R. Dutkiewicz, M. Donde, C. Y. C. Lee, P. Naidu, C. Lee, J. Cerveira, B. Liu, F. Ginhoux, G. Burton, R. S. Hamilton, A. Moffett, A. Sharkey, N. McGovern, Phenotypic and functional characterization of first-trimester human placental macrophages, Hofbauer cells. *J. Exp. Med.* **218** (2021).

136. E. L. Johnson, R. Chakraborty, Placental Hofbauer cells limit HIV-1 replication and potentially offset mother to child transmission (MTCT) by induction of immunoregulatory cytokines. *Retrovirology* **9**, 101 (2012).

137. M. Guilliams, F. Ginhoux, C. Jakubzick, S. H. Naik, N. Onai, B. U. Schraml, E. Segura, R. Tussiwand, S. Yona, Dendritic cells, monocytes and macrophages: a unified nomenclature based on ontogeny. *Nat. Rev. Immunol.* **14**, 571–578 (2014).

138. K. Asada, S. Sasaki, T. Suda, K. Chida, H. Nakamura, Antiinflammatory Roles of Peroxisome Proliferator–activated Receptor γ in Human Alveolar Macrophages. *Am. J. Respir. Crit. Care Med.* **169**, 195–200 (2004).

139. F. Ginhoux, M. Guilliams, S. Naik, *Dendritic Cell and Macrophage Nomenclature and Classification* (Frontiers Media SA, 2016).

140. N. McGovern, A. Schlitzer, M. Gunawan, L. Jardine, A. Shin, E. Poyner, K. Green, R. Dickinson, X.-N. Wang, D. Low, K. Best, S. Covins, P. Milne, S. Pagan, K. Aljefri, M. Windebank, D. Miranda-Saavedra, A. Larbi, P. S. Wasan, K. Duan, M. Poidinger, V. Bigley, F. Ginhoux, M. Collin, M. Haniffa, Human dermal $CD14^+$ cells are a transient population of monocyte-derived macrophages. *Immunity* **41**, 465–477 (2014).

141. S. A. MacParland, J. C. Liu, X.-Z. Ma, B. T. Innes, A. M. Bartczak, B. K. Gage, J. Manuel, N. Khuu, J. Echeverri, I. Linares, R. Gupta, M. L. Cheng, L. Y. Liu, D. Camat, S. W. Chung, R. K. Seliga, Z. Shao, E. Lee, S. Ogawa, M. Ogawa, M. D. Wilson, J. E. Fish, M. Selzner, A. Ghanekar, D. Grant, P. Greig, G. Sapisochin, N. Selzner, N. Winegarden, O. Adeyi, G. Keller, G. D. Bader, I. D. McGilvray, Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).

142. C. L. Scott, F. Zheng, P. De Baetselier, L. Martens, Y. Saeys, S. De Prijck, S. Lippens, C. Abels, S. Schoonooghe, G. Raes, N. Devoogdt, B. N. Lambrecht, A. Beschin, M. Guilliams, Bone marrow-derived monocytes give rise to self-renewing and fully differentiated Kupffer cells. *Nat. Commun.* **7**, 10321 (2016).

143. M. Collin, V. Bigley, Human dendritic cell subsets: an update. *Immunology* **154**, 3–20 (2018).

144. A. Dzionek, A. Fuchs, P. Schmidt, S. Cremer, M. Zysk, S. Miltenyi, D. W. Buck, J. Schmitz, BDCA-2, BDCA-3, and BDCA-4: three markers for distinct subsets of dendritic cells in human peripheral blood. *J. Immunol.* **165**, 6037–6046 (2000).

145. A. I. Tauber, L. Chernyak, *Metchnikoff and the Origins of Immunology: From Metaphor to Theory* (Oxford University Press, 1991).

146. F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, J. D. McPherson, L. D. Stein, Y. Dror, J. E. Dick, Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).

147. A. Coillard, E. Segura, In vivo Differentiation of Human Monocytes. *Front. Immunol.* **10**, 1907 (2019).

148. S. Tamoutounour, M. Guilliams, F. Montanana Sanchis, H. Liu, D. Terhorst, C. Malosse, E. Pollet, L. Ardouin, H. Luche, C. Sanchez, M. Dalod, B. Malissen, S. Henri, Origins and functional specialization of macrophages and of conventional and monocyte-derived dendritic cells in mouse skin. *Immunity* **39**, 925–938 (2013).

149. E. Segura, M. Touzot, A. Bohineust, A. Cappuccio, G. Chiocchia, A. Hosmalin, M. Dalod, V. Soumelis, S. Amigorena, Human inflammatory dendritic cells induce Th17 cell differentiation. *Immunity* **38**, 336–348 (2013).

150. N. McGovern, J. K. Y. Chan, F. Ginhoux, Dendritic cells in humans--from fetus to adult. *Int. Immunol.* **27**, 65–72 (2015).

151. K. E. McGrath, J. M. Frame, K. H. Fegan, J. R. Bowen, S. J. Conway, S. C. Catherman, P. D. Kingsley, A. D. Koniski, J. Palis, Distinct Sources of Hematopoietic Progenitors Emerge before HSCs and Provide Functional Blood Cells in the Mammalian Embryo. *Cell Rep.* **11**, 1892–1904 (2015).

152. R. Rojo, A. Raper, D. D. Ozdemir, L. Lefevre, K. Grabert, E. Wollscheid-Lengeling, B. Bradford, M. Caruso, I. Gazova, A. Sánchez, Z. M. Lisowski, J. Alves, I. Molina-Gonzalez, H. Davtyan, R. J. Lodge, J. D. Glover, R. Wallace, D. A. D. Munro, E. David, I. Amit, V. E. Miron, J. Priller, S. J. Jenkins, G. E. Hardingham, M. Blurton-Jones, N. A. Mabbott, K. M. Summers, P. Hohenstein, D. A. Hume, C. Pridans, Deletion of a Csf1r enhancer selectively impacts CSF1R expression and development of tissue macrophage populations. *Nat. Commun.* **10**, 3215 (2019).

153. M. A. De Groote, L. Johnson, B. Podell, E. Brooks, R. Basaraba, M. Gonzalez-Juarrero, GM-CSF knockout mice for preclinical testing of agents with antimicrobial activity against Mycobacterium abscessus. *J. Antimicrob. Chemother.* **69**, 1057–1064 (2014).

154. J. Palis, S. Robertson, M. Kennedy, C. Wall, G. Keller, Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073–5084 (1999).

155. F. Ginhoux, M. Greter, M. Leboeuf, S. Nandi, P. See, S. Gokhan, M. F. Mehler, S. J. Conway, L. G. Ng, E. R. Stanley, I. M. Samokhvalov, M. Merad, Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330**, 841–845 (2010).

156. C. Schulz, E. Gomez Perdiguero, L. Chorro, H. Szabo-Rogers, N. Cagnard, K. Kierdorf, M. Prinz, B. Wu, S. E. W. Jacobsen, J. W. Pollard, J. Frampton, K. J. Liu, F. Geissmann, A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science* **336**, 86–90 (2012).

157. E. Gomez Perdiguero, K. Klapproth, C. Schulz, K. Busch, E. Azzoni, L. Crozet, H. Garner, C. Trouillet, M. F. de Bruijn, F. Geissmann, H.-R. Rodewald, Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature* **518**, 547–551 (2015).

158. É. González-Domínguez, R. Samaniego, J. L. Flores-Sevilla, S. F. Campos-Campos, G. Gómez-Campos, A. Salas, V. Campos-Peña, Á. L. Corbí, P. Sánchez-Mateos, C. Sánchez-Torres,

CD163L1 and CLEC5A discriminate subsets of human resident and inflammatory macrophages in vivo. *J. Leukoc. Biol.* **98**, 453–466 (2015).

159. J. Palis, Interaction of the Macrophage and Primitive Erythroid Lineages in the Mammalian Embryo. *Front. Immunol.* **7**, 669 (2016).

160. V. G. Sankaran, Targeted therapeutic strategies for fetal hemoglobin induction. *Hematology Am. Soc. Hematol. Educ. Program* **2011**, 459–465 (2011).

161. "Transcription Factors Associated with Gamma-globin Expression in Human Adult Definitive Erythropoiesis Before and After Induction by Hydroxyurea," (2016).

162. C. Cantù, R. Ierardi, I. Alborelli, C. Fugazza, L. Cassinelli, S. Piconese, F. Bosè, S. Ottolenghi, G. Ferrari, A. Ronchi, Sox6 enhances erythroid differentiation in human erythroid progenitors. *Blood* **117**, 3669–3679 (2011).

163. W. Li, Y. Wang, H. Zhao, H. Zhang, Y. Xu, S. Wang, X. Guo, Y. Huang, S. Zhang, Y. Han, X. Wu, C. M. Rice, G. Huang, P. G. Gallagher, A. Mendelson, K. Yazdanbakhsh, J. Liu, L. Chen, X. An, Identification and transcriptome analysis of erythroblastic island macrophages. *Blood* **134**, 480–491 (2019).

164. W. Luo, Q. Xu, Q. Wang, H. Wu, J. Hua, Effect of modulation of PPAR-gamma activity on Kupffer cells M1/M2 polarization in the development of non-alcoholic fatty liver disease. *Sci. Rep.* **7**, 44612 (2017).

165. B. Daniel, G. Nagy, A. Horvath, Z. Czimmerer, I. Cuaranta-Monroy, S. Poliska, Others, The IL-4/STAT6/PPAR? signaling axis is driving the expansion of the RXR heterodimer cistrome, providing complex ligand responsiveness in macrophages. *Nucleic Acids Res.* **46**, 4425–4439 (2018).

166. C. Auffray, D. Fogg, M. Garfa, G. Elain, O. Join-Lambert, S. Kayal, S. Sarnacki, A. Cumano, G. Lauvau, F. Geissmann, Monitoring of blood vessels and tissues by a population of monocytes with patrolling behavior. *Science* **317**, 666–670 (2007).

167. P. B. Narasimhan, P. Marcovecchio, A. A. J. Hamers, C. C. Hedrick, Nonclassical Monocytes in Health and Disease. *Annu. Rev. Immunol.* **37**, 439–456 (2019).

168. M. M. Weivoda, C. K. Chew, D. G. Monroe, J. N. Farr, E. J. Atkinson, J. R. Geske, Others, Identification of osteoclast-osteoblast coupling factors in humans reveals links between bone and energy metabolism. *Nat. Commun.* **11**, 87 (2020).

169. J. H. Park, N. K. Lee, S. Y. Lee, Current Understanding of RANK Signaling in Osteoclast Differentiation and Maturation. *Mol. Cells* **40**, 706–713 (2017).

170. D. Rolph, H. Das, Transcriptional Regulation of Osteoclastogenesis: The Emerging Role of KLF2. *Front. Immunol.* **11** (2020).

171. G. J. Atkins, P. Kostakis, B. Pan, A. Farrugia, S. Gronthos, A. Evdokiou, Others, RANKL expression is related to the differentiation state of human osteoblasts. *J. Bone Miner. Res.* **18**, 1088–1098 (2003).

172. C. E. Jacome-Galarza, G. I. Percin, J. T. Muller, E. Mass, T. Lazarov, J. Eitler, Others, Developmental origin, functional maintenance and genetic rescue of osteoclasts. *Nature* **568**, 541–545 (2019).

173. B. Sacchetti, A. Funari, S. Michienzi, S. Di Cesare, S. Piersanti, I. Saggio, Others, Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic

microenvironment. *Cell* **131**, 324–336 (2007).

174. R. G. Witt, B. Wang, Q. H. Nguyen, C. Eikani, A. N. Mattis, T. C. MacKenzie, Depletion of murine fetal hematopoietic stem cells with c-Kit receptor and CD47 blockade improves neonatal engraftment. *Blood Adv.* **2**, 3602–3607 (2018).

175. M. J. Leimberg, E. Prus, A. M. Konijn, E. Fibach, Macrophages function as a ferritin iron source for cultured human erythroid precursors. *J. Cell. Biochem.* **103**, 1211–1218 (2008).

176. L. Lifshitz, G. Tabak, M. Gassmann, M. Mittelman, D. Neumann, Macrophages as novel target cells for erythropoietin. *Haematologica* **95**, 1823–1831 (2010).

177. A. May, L. M. Forrester, The erythroblastic island niche: modeling in health, stress, and disease. *Exp. Hematol.* **91**, 10–21 (2020).

178. K. Sawada, S. B. Krantz, E. N. Dessypris, S. T. Koury, S. T. Sawyer, Human colony-forming units-erythroid do not require accessory cells, but do require direct interaction with insulin-like growth factor I and/or insulin for erythroid development. *J. Clin. Invest.* **83**, 1701–1709 (1989).

179. M. Liu, X. Jin, X. He, L. Pan, X. Zhang, Y. Zhao, Macrophages support splenic erythropoiesis in 4T1 tumor-bearing mice. *PLoS One* **10**, e0121921 (2015).

180. S. Kaur, L. J. Raggatt, L. Batoon, D. A. Hume, J. P. Levesque, A. R. Pettit, Role of bone marrow macrophages in controlling homeostasis and repair in bone and bone marrow niches. *Semin. Cell Dev. Biol.* **61**, 12–21 (2017).

181. N. van Leeuwen-Kerkhoff, K. Lundberg, T. M. Westers, S. Kordasti, H. J. Bontkes, M. Lindstedt, Others, Human Bone Marrow-Derived Myeloid Dendritic Cells Show an Immature Transcriptional and Functional Profile Compared to Their Peripheral Blood Counterparts and Separate from Slan+ Non-Classical Monocytes. *Front. Immunol.* **9** (2018).

182. J. Zhang, T. Supakorndej, J. R. Krambs, M. Rao, G. Abou-Ezzi, R. Y. Ye, Others, Bone marrow dendritic cells regulate hematopoietic stem/progenitor cell trafficking. *J. Clin. Invest.* **129**, 2920–2931 (2019).

183. M. Haldar, M. Kohyama, A. Y. So, W. Kc, X. Wu, C. G. Briseno, Others, Heme-mediated SPI-C induction promotes monocyte differentiation into iron-recycling macrophages. *Cell* **156**, 1223–1234 (2014).

184. S. Q. Nagelkerke, C. W. Bruggeman, den H. Jmm, M. Epj, T. K. van den Berg, R. van Bruggen, Others, Red pulp macrophages in the human spleen are a distinct cell population with a unique expression of Fc-gamma receptors. *Blood Adv.* **2**, 941–953 (2018).

185. Y. Murata, T. Kotani, H. Ohnishi, T. Matozaki, The CD47-SIRPalpha signalling system: its physiological roles and therapeutic application. *J. Biochem.* **155**, 335–344 (2014).

186. A. Endo, S. Ueno, S. Yamada, C. Uwabe, T. Takakuwa, Morphogenesis of the spleen during the human embryonic period. *Anat. Rec.* **298**, 820–826 (2015).

187. B. Steiniger, N. Ulfig, M. Risse, P. J. Barth, Fetal and early post-natal development of the human spleen: from primordial arterial B cell lobules to a non-segmented organ. *Histochem. Cell Biol.* **128**, 205–215 (2007).

188. G. Pirgova, A. Chauveau, A. J. MacLean, J. G. Cyster, T. I. Arnon, Marginal zone SIGN-R1(+) macrophages are essential for the maturation of germinal center B cells in the spleen. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12295–12305 (2020).

189. N. Ag, J. A. Guillen, G. Gallardo, M. Diaz, J. V. de la Rosa, I. H. Hernandez, Others, The

nuclear receptor LXRalpha controls the functional specialization of splenic macrophages. *Nat. Immunol.* **14**, 831–839 (2013).

190. Z. Kellermayer, V. Fisi, M. Mihalj, G. Berta, J. Kobor, P. Balogh, Marginal Zone Macrophage Receptor MARCO Is Trapped in Conduits Formed by Follicular Dendritic Cells in the Spleen. *J. Histochem. Cytochem.* **62**, 436–449 (2014).

191. N. McGovern, A. Shin, G. Low, D. Low, K. Duan, L. J. Yao, Others, Human fetal dendritic cells promote prenatal T-cell immune suppression through arginase-2. *Nature* **546**, 662–666 (2017).

192. E. Patel, B. Wang, L. Lien, Y. Wang, L. J. Yang, J. S. Moreb, Others, Diverse T-cell differentiation potentials of human fetal thymus, fetal liver, cord blood and adult bone marrow CD34 cells on lentiviral Delta-like-1-modified mouse stromal cells. *Immunology* **128**, 497–505 (2009).

193. C. Audiger, M. J. Rahman, T. J. Yun, K. V. Tarbell, S. Lesage, The Importance of Dendritic Cells in Maintaining Immune Tolerance. *J. Immunol.* **198**, 2223–2231 (2017).

194. K. Kawane, H. Fukuyama, H. Yoshida, H. Nagase, Y. Ohsawa, Y. Uchiyama, Others, Impaired thymic development in mouse embryos deficient in apoptotic DNA degradation. *Nat. Immunol.* **4**, 138–144 (2003).

195. E. Esashi, T. Sekiguchi, H. Ito, S. Koyasu, A. Miyajima, Cutting Edge: A possible role for CD4+ thymic macrophages as professional scavengers of apoptotic thymocytes. *J. Immunol.* **171**, 2773–2777 (2003).

196. G. Putz, A. Rosner, I. Nuesslein, N. Schmitz, F. Buchholz, AML1 deletion in adult mice causes splenomegaly and lymphomas. *Oncogene* **25**, 929–939 (2006).

197. A. Sountoulidis, S. Marco Salas, E. Braun, C. Avenel, J. Bergenstråhle, J. Theelke, M. Vicari, P. Czarnewski, A. Liontos, X. Abalo, Ž. Andrusivová, R. Mirzazadeh, M. Asp, X. Li, L. Hu, S. Sariyar, A. Martinez Casals, B. Ayoglu, A. Firsova, J. Michaëlsson, E. Lundberg, C. Wählby, E. Sundström, S. Linnarsson, J. Lundeberg, M. Nilsson, C. Samakovlis, A topographic atlas defines developmental origins of cell heterogeneity in the human embryonic lung. *Nat. Cell Biol.* **25**, 351–365 (2023).

198. E. Mitsi, R. Kamng'ona, J. Rylance, C. Solorzano, J. Jesus Reine, H. C. Mwandumba, Others, Human alveolar macrophages predominately express combined classical M1 and M2 surface markers in steady state. *Respir. Res.* **19**, 66 (2018).

199. C. Morse, T. Tabib, J. Sembrat, K. L. Buschur, H. T. Bittar, E. Valenzi, Others, Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **54** (2019).

200. A. Ferretti, J. R. Fortwendel, S. A. Gebb, R. A. Barrington, Autoantibody-Mediated Pulmonary Alveolar Proteinosis in Rasgrp1-Deficient Mice. *J. Immunol.* **197**, 470–479 (2016).

201. B. C. Trapnell, J. A. Whitsett, K. Nakata, Pulmonary alveolar proteinosis. *N. Engl. J. Med.* **349**, 2527–2539 (2003).

202. D. Hashimoto, A. Chow, C. Noizat, P. Teo, M. B. Beasley, M. Leboeuf, Others, Tissue-resident macrophages self-maintain locally throughout adult life with minimal contribution from circulating monocytes. *Immunity* **38**, 792–804 (2013).

203. H. M. Izquierdo, P. Brandi, M.-J. G—mez, R. Conde-Garrosa, E. Priego, M. Enamorado, Others, Von Hippel-Lindau Protein Is Required for Optimal Alveolar Macrophage Terminal

Differentiation, Self-Renewal, and Function. *Cell Rep.* **24**, 1738–1746 (2018).

204. R. van Furth, Z. A. Cohn, The origin and kinetics of mononuclear phagocytes. *J. Exp. Med.* **128**, 415–435 (1968).

205. M. Fathi, A. Johansson, M. Lundborg, L. Orre, C. M. Sköld, P. Camner, Functional and morphological differences between human alveolar and interstitial macrophages. *Exp. Mol. Pathol.* **70**, 77–82 (2001).

206. J. Hoppstädter, B. Diesel, R. Zarbock, T. Breinig, D. Monz, M. Koch, A. Meyerhans, L. Gortner, C.-M. Lehr, H. Huwer, A. K. Kiemer, Differential cell reaction upon Toll-like receptor 4 and 9 activation in human alveolar and lung interstitial macrophages. *Respir. Res.* **11**, 124 (2010).

207. S. Ferrari-Lacraz, L. P. Nicod, R. Chicheportiche, H. G. Welgus, J. M. Dayer, Human lung tissue macrophages, but not alveolar macrophages, express matrix metalloproteinases after direct contact with activated T lymphocytes. *Am. J. Respir. Cell Mol. Biol.* **24**, 442–451 (2001).

208. D. Bedoret, H. Wallemacq, T. Marichal, C. Desmet, F. Quesada Calvo, E. Henry, Others, Lung interstitial macrophages alter dendritic cell functions to prevent airway allergy in mice. *J. Clin. Invest.* **119**, 3723–3738 (2009).

209. R. K. Montgomery, A. E. Mulberg, R. J. Grand, Development of the human gastrointestinal tract: twenty years of progress. *Gastroenterology* **116**, 702–731 (1999).

210. S. F. Stras, L. Werner, J. M. Toothaker, O. O. Olaloye, A. L. Oldham, C. C. McCourt, Others, Maturation of the Human Intestinal Immune System Occurs Early in Fetal Development. *Dev. Cell* **51**, 357–373 (2019).

211. A. Bujko, N. Atlasy, L. Ojb, L. Richter, S. Yaqub, R. Horneland, Others, Transcriptional and functional profiling defines human small intestinal macrophage subsets. *J. Exp. Med.* **215**, 441–458 (2018).

212. F. D'Angelo, E. Bernasconi, M. Schafer, M. Moyat, P. Michetti, M. H. Maillard, Others, Macrophages promote epithelial repair through hepatocyte growth factor secretion. *Clin. Exp. Immunol.* **174**, 60–72 (2013).

213. D. Ortiz-Masia, J. Cosin-Roger, S. Calatayud, C. Hernandez, R. Alos, J. Hinojosa, Others, Hypoxic macrophages impair autophagy in epithelial cells through Wnt1: relevance in IBD. *Mucosal Immunol.* **7**, 929–938 (2014).

214. A. Vallon-Eberhard, L. Landsman, N. Yogev, B. Verrier, S. Jung, Transepithelial pathogen uptake into the small intestinal lamina propria. *J. Immunol.* **176**, 2465–2469 (2006).

215. M. Chieppa, M. Rescigno, H. Ayc, R. N. Germain, Dynamic imaging of dendritic cell extension into the small bowel lumen in response to epithelial cell TLR engagement. *J. Exp. Med.* **203**, 2841–2852 (2006).

216. K. A. Knoop, K. G. McDonald, S. McCrate, J. R. McDole, R. D. Newberry, Microbial sensing by goblet cells controls immune surveillance of luminal antigens in the colon. *Mucosal Immunol.* **8**, 198–210 (2015).

217. C. C. Bain, A. Schridde, Origin, Differentiation, and Function of Intestinal Macrophages. *Front. Immunol.* **9** (2018).

218. P. B. Watchmaker, K. Lahl, M. Lee, D. Baumjohann, J. Morton, S. J. Kim, Others, Comparative transcriptional and functional profiling defines conserved programs of intestinal DC differentiation in humans and mice. *Nat. Immunol.* **15**, 98–108 (2014).

219. G. Bakdash, L. T. Vogelpoel, T. M. van Capel, M. L. Kapsenberg, E. C. de Jong, Retinoic acid primes human dendritic cells to induce gut-homing, IL-10-producing regulatory T cells. *Mucosal Immunol.* **8**, 265–278 (2015).

220. M. M. Roe, S. Swain, T. A. Sebrell, M. A. Sewell, M. M. Collins, B. A. Perrino, Others, Differential regulation of CD103 (alphaE integrin) expression in human dendritic cells by retinoic acid and Toll-like receptor ligands. *J. Leukoc. Biol.* **101**, 1169–1180 (2017).

221. R. Ramalingam, C. B. Larmonier, R. D. Thurston, M. T. Midura-Kiela, S. G. Zheng, F. K. Ghishan, Others, Dendritic cell-specific disruption of TGF-beta receptor II leads to altered regulatory T cell phenotype and spontaneous multiorgan autoimmunity. *J. Immunol.* **189**, 3878–3893 (2012).

222. D. BauchŽ, J. C. Marie, Transforming growth factor ?: a master regulator of the gut microbiota and immune cell interactions. *Clin Transl Immunology* **6**, e136 (2017).

223. D. Han, M. C. Walsh, P. J. Cejas, N. N. Dang, Y. F. Kim, J. Kim, Others, Dendritic cell expression of the signaling molecule TRAF6 is critical for gut microbiota-dependent immune tolerance. *Immunity* **38**, 1211–1222 (2013).

224. M. C. Walsh, J. Lee, Y. Choi, Tumor necrosis factor receptor- associated factor 6 (TRAF6) regulation of development, function, and homeostasis of the immune system. *Immunol. Rev.* **266**, 72–92 (2015).

225. D. Esterhazy, J. Loschko, M. London, V. Jove, T. Y. Oliveira, D. Mucida, Classical dendritic cells are required for dietary antigen-mediated induction of peripheral T(reg) cells and tolerance. *Nat. Immunol.* **17**, 545–555 (2016).

226. A. J. Stagg, A. L. Hart, S. C. Knight, M. A. Kamm, The dendritic cell: its role in intestinal inflammation and relationship with gut bacteria. *Gut* **52**, 1522–1529 (2003).

227. M. K. A. 2nd, Romano-Keeler J, Zackular JP, Moore DJ, Brucker RM, Hooper C, et al. Bacterial DNA is present in the fetal intestine and overlaps with that in the placenta in mice. *PLoS One* **13**, e0197439 (2018).

228. R. W. Walker, J. C. Clemente, I. Peter, L. Rjf, The prenatal gut microbiome: are we colonized with bacteria in utero? *Pediatr. Obes.* **12 Suppl 1**, 3–17 (2017).

229. C. A. Foster, K. A. Holbrook, A. G. Farr, Ontogeny of Langerhans cells in human embryonic and fetal skin: expression of HLA-DR and OKT-6 determinants. *J. Invest. Dermatol.* **86**, 240–243 (1986).

230. S. Carpentier, T.-P. Vu Manh, R. Chelbi, S. Henri, B. Malissen, M. Haniffa, Others, Comparative genomics analysis of mononuclear phagocyte subsets confirms homology between lymphoid tissue-resident and dermal XCR1(+) DCs in mouse and human and distinguishes them from Langerhans cells. *J. Immunol. Methods* **432**, 35–49 (2016).

231. C. Schuster, C. Vaculik, C. Fiala, S. Meindl, O. Brandt, M. Imhof, Others, HLA-DR+ leukocytes acquire CD1 antigens in embryonic and fetal human skin and contain functional antigen-presenting cells. *J. Exp. Med.* **206**, 169–181 (2009).

232. C. Schuster, C. Vaculik, M. Prior, C. Fiala, M. Mildner, W. Eppel, Others, Phenotypic characterization of leukocytes in prenatal human dermis. *J. Invest. Dermatol.* **132**, 2581–2592 (2012).

233. B. Malissen, S. Tamoutounour, S. Henri, The origins and functions of dendritic cells and macrophages in the skin. *Nat. Rev. Immunol.* **14**, 417–428 (2014).

234. M. Merad, F. Ginhoux, M. Collin, Origin, homeostasis and function of Langerhans cells and other langerin-expressing dendritic cells. *Nat. Rev. Immunol.* **8**, 935–947 (2008).

235. F. O. Nestle, P. Di Meglio, J. Z. Qin, B. J. Nickoloff, Skin immune sentinels in health and disease. *Nat. Rev. Immunol.* **9**, 679–691 (2009).

236. J. Etich, M. Koch, R. Wagener, F. Zaucke, M. Fabri, B. Brachvogel, Gene Expression Profiling of the Extracellular Matrix Signature in Macrophages of Different Activation Status: Relevance for Skin Wound Healing. *Int. J. Mol. Sci.* **20** (2019).

237. B. A. Shook, R. R. Wasko, G. C. Rivera-Gonzalez, E. Salazar-Gatzimas, F. Lopez-Giraldez, B. C. Dash, Others, Myofibroblast proliferation and heterogeneity are supported by macrophages during skin repair. *Science* **362** (2018).

238. L. Furio, I. Briotet, A. Journeaux, H. Billard, J. Peguet-Navarro, Human langerhans cells are more efficient than CD14(-)CD1c(+) dermal dendritic cells at priming naive CD4(+) T cells. *J. Invest. Dermatol.* **130**, 1345–1354 (2010).

239. J. Seneschal, R. A. Clark, A. Gehad, C. M. Baecher-Allan, T. S. Kupper, Human epidermal Langerhans cells maintain immune homeostasis in skin by activating skin resident regulatory T cells. *Immunity* **36**, 873–884 (2012).

240. M. Haniffa, A. Shin, V. Bigley, N. McGovern, P. Teo, P. See, Others, Human tissues contain CD141hi cross-presenting dendritic cells with functional homology to mouse CD103+ nonlymphoid dendritic cells. *Immunity* **37**, 60–73 (2012).

241. L. Chorro, A. Sarde, M. Li, K. J. Woollard, P. Chambon, B. Malissen, Others, Langerhans cell (LC) proliferation mediates neonatal development, homeostasis, and inflammation-associated expansion of the epidermal LC network. *J. Exp. Med.* **206**, 3089–3100 (2009).

242. G. Reynolds, P. Vegh, J. Fletcher, P. Efm, E. Stephenson, I. Goh, Others, Poised cell circuits in human skin are activated in disease.

243. F. Sallusto, A. Lanzavecchia, Efficient presentation of soluble antigen by cultured human dendritic cells is maintained by granulocyte/macrophage colony-stimulating factor plus interleukin 4 and downregulated by tumor necrosis factor alpha. *J. Exp. Med.* **179**, 1109–1118 (1994).

244. M. Merad, M. G. Manz, H. Karsunky, A. Wagers, W. Peters, I. Charo, Others, Langerhans cells renew in the skin throughout life under steady-state conditions. *Nat. Immunol.* **3**, 1135–1141 (2002).

245. C. Caux, C. Dezutter-Dambuyant, D. Schmitt, J. Banchereau, GM-CSF and TNF-alpha cooperate in the generation of dendritic Langerhans cells. *Nature* **360**, 258–261 (1992).

246. L. F. Poulin, M. Salio, E. Griessinger, F. Anjos-Afonso, L. Craciun, J. L. Chen, Others, Characterization of human DNGR-1+ BDCA3+ leukocytes as putative equivalents of mouse CD8alpha+ dendritic cells. *J. Exp. Med.* **207**, 1261–1271 (2010).

247. S. Balan, C. Arnold-Schrauf, A. Abbas, N. Couespel, J. Savoret, F. Imperatore, Others, Large-Scale Human Dendritic Cell Differentiation Revealing Notch-Dependent Lineage Bifurcation and Heterogeneity. *Cell Rep.* **24**, 1902–1915 (2018).

248. M. Haniffa, M. Collin, F. Ginhoux, Identification of human tissue cross-presenting dendritic cells: A new target for cancer vaccines. *Oncoimmunology* **2**, e23140 (2013).

249. K. Takata, T. Kozaki, L. Czw, M. S. Thion, M. Otsuka, S. Lim, Others,

Induced-Pluripotent-Stem-Cell-Derived Primitive Macrophages Provide a Platform for Modeling Tissue-Resident Macrophage Differentiation and Function. *Immunity* **47**, 183–198 (2017).

250. A. Shanti, J. Teo, C. Stefanini, In Vitro Immune Organs-on-Chip for Drug Development: A Review. *Pharmaceutics* **10** (2018).

251. N. Iakobachvili, P. J. Peters, Humans in a Dish: The Potential of Organoids in Modeling Immunity and Infectious Diseases. *Front. Microbiol.* **8**, 2402 (2017).

252. J. Kim, B. K. Koo, J. A. Knoblich, Human organoids: model systems for human biology and medicine. *Nat. Rev. Mol. Cell Biol.* **21**, 571–584 (2020).

253. E. Fiorini, L. Veghini, V. Corbo, Modeling Cell Communication in Cancer With Organoids: Making the Complex Simple. *Front Cell Dev Biol* **8**, 166 (2020).

254. J. T. Neal, X. Li, J. Zhu, V. Giangarra, C. L. Grzeskowiak, J. Ju, Others, Organoid Modeling of the Tumor Immune Microenvironment. *Cell* **175**, 1972–1988 (2018).

255. P. E. Bourgine, T. Klein, A. M. Paczulla, T. Shimizu, L. Kunz, K. D. Kokkaliaris, Others, In vitro biomimetic engineering of a human hematopoietic niche with functional properties. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5688–E95 (2018).

256. E. Biselli, E. Agliari, A. Barra, F. R. Bertani, A. Gerardino, A. De Ninno, Others, Organs on chip approach: a tool to evaluate cancer -immune cells interactions. *Sci. Rep.* **7**, 12737 (2017).

257. T. Sasserath, J. W. Rumsey, C. W. McAleer, L. R. Bridges, C. J. Long, D. Elbrecht, Others, Differential Monocyte Actuation in a Three-Organ Functional Innate Immune System-on-a-Chip. *Adv. Sci.* **7**, 2000323 (2020).

258. Y. Liu, X. Gao, Y. Miao, Y. Wang, H. Wang, Z. Cheng, Others, NLRP3 regulates macrophage M2 polarization through up-regulation of IL-4 in asthma. *Biochem. J* **475**, 1995–2008 (2018).

259. K. J. Jang, M. A. Otieno, J. Ronxhi, H. K. Lim, L. Ewart, K. R. Kodella, Others, Reproducing human and cross-species drug toxicities using a Liver-Chip. *Sci. Transl. Med.* **11** (2019).

260. D. B. Chou, V. Frismantas, Y. Milton, R. David, P. Pop-Damkov, D. Ferguson, Others, Human bone marrow disorders recapitulated in vitro using organ chip technology.

261. G. Sriram, P. L. Bigliardi, M. Bigliardi-Qi, Full-Thickness Human Skin Equivalent Models of Atopic Dermatitis. *Methods Mol. Biol.* **1879**, 367–383 (2019).

262. L. G. Rigat-Brugarolas, A. Elizalde-Torrent, M. Bernabeu, M. De Niz, L. Martin-Jaular, C. Fernandez-Becerra, Others, A functional microengineered model of the human splenon-on-a-chip. *Lab Chip* **14**, 1715–1724 (2014).

263. L. Richardson, J. Gnecco, T. Ding, K. Osteen, L. M. Rogers, D. M. Aronoff, Others, Fetal Membrane Organ-On-Chip: An Innovative Approach to Study Cellular Interactions. *Reprod. Sci.* **27**, 1562–1569 (2020).

264. S. Sieber, L. Wirth, N. Cavak, M. Koenigsmark, U. Marx, R. Lauster, Others, Bone marrow-on-a-chip: Long-term culture of human haematopoietic stem cells in a three-dimensional microfluidic environment. *J. Tissue Eng. Regen. Med.* **12**, 479–489 (2018).

265. K. Rennert, S. Steinborn, M. Groger, B. Ungerbock, A. M. Jank, J. Ehgartner, Others, A microfluidically perfused three dimensional human liver model. *Biomaterials* **71**, 119–131 (2015).

266. M. Groger, K. Rennert, B. Giszas, E. Weiss, J. Dinger, H. Funke, Others, Monocyte-induced

recovery of inflammation-associated hepatocellular dysfunction in a biochip-based human liver model. *Sci. Rep.* **6**, 21868 (2016).

267. R. S. Hotchkiss, S. Opal, Immunotherapy for sepsis--a new approach against an ancient foe. *N. Engl. J. Med.* **363**, 87–89 (2010).

268. Y. Zhao, R. K. Kankala, S.-B. Wang, A.-Z. Chen, Multi-Organs-on-Chips: Towards Long-Term Biomedical Investigations. *Molecules* **24**, 675 (2019).

269. M. Hulsmans, F. Sam, M. Nahrendorf, Monocyte and macrophage contributions to cardiac remodeling. *J. Mol. Cell. Cardiol.* **93**, 149–155 (2016).

270. S. D. Prabhu, N. G. Frangogiannis, The Biological Basis for Cardiac Repair After Myocardial Infarction: From Inflammation to Fibrosis. *Circ. Res.* **119**, 91–112 (2016).

271. J. Heidemann, H. Ogawa, M. B. Dwinell, P. Rafiee, C. Maaser, H. R. Gockel, M. F. Otterson, D. M. Ota, N. Lugering, W. Domschke, D. G. Binion, Angiogenic effects of interleukin 8 (CXCL8) in human intestinal microvascular endothelial cells are mediated by CXCR2. *J. Biol. Chem.* **278**, 8508–8515 (2003).

272. X. Yang, P. Lu, C. Fujii, Y. Nakamoto, J.-L. Gao, S. Kaneko, P. M. Murphy, N. Mukaida, Essential contribution of a chemokine, CCL3, and its receptor, CCR1, to hepatocellular carcinoma progression. *Int. J. Cancer* **118**, 1869–1876 (2006).

273. F. Hua, Y. Tian, CCL4 promotes the cell proliferation, invasion and migration of endometrial carcinoma by targeting the VEGF-A signal pathway. *Int. J. Clin. Exp. Pathol.* **10**, 11288–11299 (2017).

274. E. C. Keeley, B. Mehrad, R. M. Strieter, CXC chemokines in cancer angiogenesis and metastases. *Adv. Cancer Res.* **106**, 91–111 (2010).

275. N. Sukhbaatar, T. Weichhart, Iron Regulation: Macrophages in Control. *Pharmaceuticals* **11** (2018).

276. J. Lee, C. C. Rabbani, H. Gao, M. R. Steinhart, B. M. Woodruff, Z. E. Pflum, Others, Hair-bearing human skin generated entirely from pluripotent stem cells. *Nature* **582**, 399–404 (2020).

277. S. Kanton, M. J. Boyle, Z. He, M. Santel, A. Weigert, F. Sanchis-Calleja, Others, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).

278. B. van Wilgenburg, C. Browne, J. Vowles, S. A. Cowley, Efficient, long term production of monocyte-derived macrophages from human pluripotent stem cells under partly-defined and fully-defined conditions. *PLoS One* **8**, e71098 (2013).

279. A. D. Friedman, Cell cycle and developmental control of hematopoiesis by Runx1. *J. Cell. Physiol.* **219**, 520–524 (2009).

280. R. Ferreira, K. Ohneda, M. Yamamoto, S. Philipsen, GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.* **25**, 1215–1227 (2005).

281. H. Chen, D. Ray-Gallet, P. Zhang, C. J. Hetherington, D. A. Gonzalez, D. E. Zhang, F. Moreau-Gachelin, D. G. Tenen, PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**, 1549–1560 (1995).

282. T. A. Wynn, K. M. Vannella, Macrophages in Tissue Repair, Regeneration, and Fibrosis. *Immunity* **44**, 450–462 (2016).

283. C. M. Leung, P. de Haan, K. Ronaldson-Bouchard, G.-A. Kim, J. Ko, H. S. Rho, Z. Chen, P. Habibovic, N. L. Jeon, S. Takayama, M. L. Shuler, G. Vunjak-Novakovic, O. Frey, E. Verpoorte, Y.-C. Toh, A guide to the organ-on-a-chip. *Nature Reviews Methods Primers* **2**, 1–29 (2022).

284. B. Burja, D. Paul, A. Tastanova, S. G. Edalat, R. Gerber, M. Houtman, M. Elhai, K. Bürki, R. Staeger, G. Restivo, R. Lang, S. Sodin-Semrl, K. Lakota, M. Tomšič, M. P. Levesque, O. Distler, Ž. Rotar, M. D. Robinson, M. Frank-Bertoncelj, An Optimized Tissue Dissociation Protocol for Single-Cell RNA Sequencing Analysis of Fresh and Cultured Human Skin Biopsies. *Front Cell Dev Biol* **10**, 872688 (2022).

285. M. Zilbauer, K. R. James, M. Kaur, S. Pott, Z. Li, A. Burger, J. R. Thiagarajah, J. Burclaff, F. L. Jahnsen, F. Perrone, A. D. Ross, G. Matteoli, N. Stakenborg, T. Sujino, A. Moor, R. Bartolome-Casado, E. S. Bækkevold, R. Zhou, B. Xie, K. S. Lau, S. Din, S. T. Magness, Q. Yao, S. Beyaz, M. Arends, A. Denadai-Souza, L. A. Coburn, J. T. Gaublomme, R. Baldock, I. Papatheodorou, J. Ordovas-Montanes, G. Boeckxstaens, A. Hupalowska, S. A. Teichmann, A. Regev, R. J. Xavier, A. Simmons, M. P. Snyder, K. T. Wilson, Gut Cell Atlas Consortium, Human Cell Atlas Gut Biological Network Consortium, A Roadmap for the Human Gut Cell Atlas. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 597–614 (2023).

286. V. Garcia-Flores, Y. Xu, E. Pusod, R. Romero, R. Pique-Regi, N. Gomez-Lopez, Preparation of single-cell suspensions from the human placenta. *Nat. Protoc.* **18**, 732–754 (2023).

287. M. Bozzo, S. Candiani, M. Schubert, Whole mount in situ hybridization and immunohistochemistry for studying retinoic acid signaling in developing amphioxus. *Methods Enzymol.* **637**, 419–452 (2020).

288. B. Hwang, J. H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

289. E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan, J. C. Marioni, Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).

290. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, S. A. Teichmann, The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).

291. Tabula Sapiens Consortium*, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, W. Harper, M. Hemenez, R. Ponnusamy, A. Salehi, B. A. Sanagavarapu, E. Spallino, K. A. Aaron, W. Concepcion, J. M. Gardner, B. Kelly, N. Neidlinger, Z. Wang, S. Crasta, S. Kolluru, M. Morri, A. O. Pisco, S. Y. Tan, K. J. Travaglini, C. Xu, M. Alcántara-Hernández, N. Almanzar, J. Antony, B. Beyersdorf, D. Burhan, K. Calcuttawala, M. M. Carter, C. K. F. Chan, C. A. Chang, S. Chang, A. Colville, S. Crasta, R. N. Culver, I. Cvijović, G. D'Amato, C. Ezran, F. X. Galdos, A. Gillich, W. R. Goodyer, Y. Hang, A. Hayashi, S. Houshdaran, X. Huang, J. C. Irwin, S. Jang, J. V. Juanico, A. M. Kershner, S. Kim, B. Kiss, S. Kolluru, W. Kong, M. E. Kumar, A. H. Kuo, R. Leylek, B. Li, G. B. Loeb, W.-J. Lu, S. Mantri, M. Markovic, P. L. McAlpine, A. de Morree, M. Morri, K. Mrouj, S. Mukherjee, T. Muser, P. Neuhöfer, T. D. Nguyen, K. Perez, R. Phansalkar, A. O. Pisco, N. Puluca, Z. Qi, P. Rao, H. Raquer-McKay, N. Schaum, B. Scott, B. Seddighzadeh, J. Segal, S. Sen, S. Sikandar, S. P. Spencer, L. C. Steffes, V. R. Subramaniam, A. Swarup, M. Swift, K. J. Travaglini, W. Van Treuren, E. Trimm, S. Veizades, S. Vijayakumar, K. C. Vo, S. K. Vorperian, W. Wang, H. N. W. Weinstein, J. Winkler, T. T. H. Wu, J. Xie, A. R. Yung, Y. Zhang, A. M. Detweiler, H. Mekonen, N. F. Neff, R. V. Sit, M. Tan, J. Yan, G. R. Bean, V. Charu, E. Forgó, B. A. Martin, M. G. Ozawa, O. Silva, S. Y. Tan, A. Toland, V. N. P. Vemuri, S. Afik, K. Awayan, O. B. Botvinnik, A. Byrne, M. Chen, R. Dehghannasiri, A. M. Detweiler, A. Gayoso, A. A. Granados, Q. Li, G. Mahmoudabadi, A. McGeever, A. de Morree, J. E. Olivieri, M. Park, A. O. Pisco, N. Ravikumar, J. Salzman, G. Stanley, M. Swift, M. Tan, W. Tan, A. J. Tarashansky, R. Vanheusden, S. K. Vorperian, P. Wang, S. Wang, G. Xing, C. Xu, N. Yosef, M. Alcántara-Hernández, J. Antony, C.

K. F. Chan, C. A. Chang, A. Colville, S. Crasta, R. Culver, L. Dethlefsen, C. Ezran, A. Gillich, Y. Hang, P.-Y. Ho, J. C. Irwin, S. Jang, A. M. Kershner, W. Kong, M. E. Kumar, A. H. Kuo, R. Leylek, S. Liu, G. B. Loeb, W.-J. Lu, J. S. Maltzman, R. J. Metzger, A. de Morree, P. Neuhöfer, K. Perez, R. Phansalkar, Z. Qi, P. Rao, H. Raquer-McKay, K. Sasagawa, B. Scott, R. Sinha, H. Song, S. P. Spencer, A. Swarup, M. Swift, K. J. Travaglini, E. Trimm, S. Veizades, S. Vijayakumar, B. Wang, W. Wang, J. Winkler, J. Xie, A. R. Yung, S. E. Artandi, P. A. Beachy, M. F. Clarke, L. C. Giudice, F. W. Huang, K. C. Huang, J. Idoyaga, S. K. Kim, M. Krasnow, C. S. Kuo, P. Nguyen, S. R. Quake, T. A. Rando, K. Red-Horse, J. Reiter, D. A. Relman, J. L. Sonnenburg, B. Wang, A. Wu, S. M. Wu, T. Wyss-Coray, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

292. Y. Pan, W. Cao, Y. Mu, Q. Zhu, Microfluidics Facilitates the Development of Single-Cell RNA Sequencing. *Biosensors* **12** (2022).

293. K. Jannati, M.-H. Rahimian, M. Raisee, A. Jafari, Simulation-based insights into cell encapsulation dynamics in droplet microfluidics. *Phys. Fluids* , doi: 10.1063/5.0203089 (2024).

294. K. Danielski, Guidance on Processing the 10x Genomics Single Cell Gene Expression Assay. *Methods Mol. Biol.* **2584**, 1–28 (2023).

295. A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, Human Cell Atlas Meeting Participants, The Human Cell Atlas. *Elife* **6** (2017).

296. J. Fletcher, R. Botting, E. Stephenson, P. Vegh, M. Haniffa, Human skin single cell dissociation v1. *protocols.io*, doi: 10.17504/protocols.io.ripd4dn (2018).

297. J. Park, Cambridge Fetal Thymus Dissociation v1. *protocols.io*, doi: 10.17504/protocols.io.wedfba6 (2018).

298. E. S. Patel, B. Wang, L. Lien, Y. Wang, L.-J. Yang, J. Moreb, L.-J. Chang, Diverse T‐cell differentiation potentials of human fetal thymus, fetal liver, cord blood and adult bone marrow CD34 cells on lentiviral Delta‐like‐1‐modified mouse stromal cells. *Immunology* **128** (2009).

299. O. Stegle, S. A. Teichmann, J. C. Marioni, Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).

300. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, S. A. Teichmann, The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).

301. T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, S. A. Teichmann, Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).

302. J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. La Manno, M. A. Skinnider, G. Courtine, Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).

303. G. Sturm, T. Szabo, G. Fotakis, M. Haider, D. Rieder, Z. Trajanoski, F. Finotello, Scirpy: a

Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* **36**, 4817–4818 (2020).

304. C. Suo, K. Polanski, E. Dann, R. G. H. Lindeboom, R. Vilarrasa-Blasi, R. Vento-Tormo, M. Haniffa, K. B. Meyer, L. M. Dratva, Z. K. Tuong, M. R. Clatworthy, S. A. Teichmann, Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nat. Biotechnol.* **42**, 40–51 (2024).

305. N. Borcherding, N. L. Bormann, G. Kraus, scRepertoire: An R-based toolkit for single-cell immune receptor analysis. *F1000Res.* **9**, 47 (2020).

306. A. T. L. Lun, S. Riesenfeld, T. Andrews, T. P. Dao, T. Gomes, participants in the 1st Human Cell Atlas Jamboree, J. C. Marioni, EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).

307. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329–337.e4 (2019).

308. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).

309. A. Subramanian, M. Alperovich, Y. Yang, B. Li, Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biol.* **23**, 267 (2022).

310. D. Osorio, J. J. Cai, Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).

311. J. Yates, A. Kraft, V. Boeva, Filtering cells with high mitochondrial content depletes viable metabolically altered malignant cell populations in cancer single-cell studies. *Genome Biol.* **26**, 91 (2025).

312. A. T. L. Lun, K. Bach, J. C. Marioni, Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

313. D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716–729.e27 (2018).

314. L. Lin, M. Song, Y. Jiang, X. Zhao, H. Wang, L. Zhang, Normalizing single-cell RNA sequencing data with internal spike-in-like genes. *NAR Genom. Bioinform.* **2**, lqaa059 (2020).

315. A. T. L. Lun, F. J. Calero-Nieto, L. Haim-Vilmovsky, B. Göttgens, J. C. Marioni, Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* **27**, 1795–1806 (2017).

316. H. Heaton, A. M. Talman, A. Knights, M. Imaz, D. J. Gaffney, R. Durbin, M. Hemberg, M. K. N. Lawniczak, Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).

317. S. J. Fleming, M. D. Chaffin, A. Arduini, A.-D. Akkad, E. Banks, J. C. Marioni, A. A. Philippakis, P. T. Ellinor, M. Babadi, Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).

318. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet-based

single-cell RNA sequencing data. *Gigascience* **9** (2020).

319. S. Yang, S. E. Corbett, Y. Koga, Z. Wang, W. E. Johnson, M. Yajima, J. D. Campbell, Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).

320. M. D. Luecken, F. J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

321. Y. Zhao, M.-C. Li, M. M. Konaté, L. Chen, B. Das, C. Karlovich, P. M. Williams, Y. A. Evrard, J. H. Doroshow, L. M. McShane, TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J. Transl. Med.* **19**, 269 (2021).

322. C. Ahlmann-Eltze, W. Huber, Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).

323. M. Love, S. Anders, W. Huber, Differential analysis of count data – the DESeq2 package. *Genome Biol.* (2013).

324. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

325. H. C. T. Nguyen, B. Baik, S. Yoon, T. Park, D. Nam, Benchmarking integration of single-cell differential expression. *Nat. Commun.* **14**, 1570 (2023).

326. V. Svensson, Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).

327. L. Alessandrì, M. Arigoni, R. Calogero, "Differential Expression Analysis in Single-Cell Transcriptomics" in *Single Cell Methods: Sequencing and Proteomics*, V. Proserpio, Ed. (Springer New York, New York, NY, 2019), pp. 425–432.

328. D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.-P. Vert, Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **10**, 646 (2019).

329. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

330. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

331. V. Svensson, A. Gayoso, N. Yosef, L. Pachter, Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).

332. K. Tsuyuzaki, H. Sato, K. Sato, I. Nikaido, Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol.* **21**, 9 (2020).

333. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

334. S. Sun, Y. Chen, Y. Liu, X. Shang, A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data. *BMC Syst. Biol.* **13**, 28 (2019).

335. J. A. I. Johnson, A. P. Tsang, J. T. Mitchell, D. L. Zhou, J. Bowden, E. Davis-Marcisak, T. Sherman, T. Liefeld, M. Loth, L. A. Goff, J. W. Zimmerman, B. Kinny-Köster, E. M. Jaffee, P.

Tamayo, J. P. Mesirov, M. Reich, E. J. Fertig, G. L. Stein-O'Brien, Inferring cellular and molecular processes in single-cell data with non-negative matrix factorization using Python, R and GenePattern Notebook implementations of CoGAPS. *Nat. Protoc.* **18**, 3690–3731 (2023).

336. C.-Y. Wang, Y.-L. Gao, X.-Z. Kong, J.-X. Liu, C.-H. Zheng, Unsupervised Cluster Analysis and Gene Marker Extraction of scRNA-seq Data Based On Non-Negative Matrix Factorization. *IEEE J Biomed Health Inform* **26**, 458–467 (2022).

337. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** (2008).

338. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv [stat.ML]* (2018). http://arxiv.org/abs/1802.03426.

339. D. Kobak, G. C. Linderman, Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).

340. O. Kramer, "K-Nearest Neighbors" in *Dimensionality Reduction with Unsupervised Nearest Neighbors*, O. Kramer, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp. 13–23.

341. V. Y. Kiselev, T. S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).

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

343. C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, N. Yosef, Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).

344. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

345. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

346. S. C. Hicks, F. W. Townes, M. Teng, R. A. Irizarry, Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).

347. Y. Zhang, G. Parmigiani, W. E. Johnson, ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).

348. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

349. B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

350. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

351. K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, J.-E. Park, BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).

352. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, *arXiv [stat.ML]* (2013). http://arxiv.org/abs/1312.6114v11.

353. D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, T. Nguyen, Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.* **12**, 1029 (2021).

354. Y. Choi, R. Li, G. Quon, siVAE: interpretable deep generative models for single-cell transcriptomes. *Genome Biol.* **24**, 29 (2023).

355. C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, O. Winther, scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).

356. C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, J. L. Jones, S. A. Teichmann, Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).

357. S. Zhang, X. Li, J. Lin, Q. Lin, K.-C. Wong, Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA* **29**, 517–530 (2023).

358. S. Na, L. Xumin, G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm" in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics* (IEEE, 2010), pp. 63–67.

359. A. Peyvandipour, A. Shafi, N. Saberian, S. Draghici, Identification of cell types from single cell data using stable clustering. *Sci. Rep.* **10**, 12349 (2020).

360. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

361. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 1–12 (2019).

362. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 026113 (2004).

363. F. A. Wolf, F. Alexander Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. [Preprint] (2018). https://doi.org/10.1186/s13059-017-1382-0.

364. X. Shao, J. Liao, X. Lu, R. Xue, N. Ai, X. Fan, scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23**, 100882 (2020).

365. D. Aran, A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, A. R. Abate, A. J. Butte, M. Bhattacharya, Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

366. H. A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).

367. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

368. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly

reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

369. B. Reuter, K. Fackeldey, M. Weber, Generalized Markov modeling of nonreversible molecular kinetics. *J. Chem. Phys.* **150**, 174103 (2019).

370. B. Reuter, M. Weber, K. Fackeldey, S. Röblitz, M. E. Garcia, Generalized Markov state modeling method for nonequilibrium biomolecular dynamics: Exemplified on Amyloid β conformational dynamics driven by an oscillating electric field. *J. Chem. Theory Comput.* **14**, 3579–3594 (2018).

371. N. Kumar, B. Mishra, M. Athar, S. Mukhtar, Inference of Gene Regulatory Network from Single-Cell Transcriptomic Data Using pySCENIC. *Methods Mol. Biol.* **2328**, 171–182 (2021).

372. S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, S. Aerts, SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

373. M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. https://doi.org/10.1101/680926.

374. R. Browaeys, W. Saelens, Y. Saeys, NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).

375. J. Glassberg, A. H. Rahman, M. Zafar, C. Cromwell, A. Punzalan, J. J. Badimon, L. Aledort, Application of phospho-CyTOF to characterize immune activation in patients with sickle cell disease in an ex vivo model of thrombosis. *J. Immunol. Methods* **453**, 11–19 (2018).

376. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, A. Regev, Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).

377. J. M. Replogle, T. M. Norman, A. Xu, J. A. Hussmann, J. Chen, J. Z. Cogan, E. J. Meer, J. M. Terry, D. P. Riordan, N. Srinivas, I. T. Fiddes, J. G. Arthur, L. J. Alvarado, K. A. Pfeiffer, T. S. Mikkelsen, J. S. Weissman, B. Adamson, Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).

378. T. M. Norman, M. A. Horlbeck, J. M. Replogle, A. Y. Ge, A. Xu, M. Jost, L. A. Gilbert, J. S. Weissman, Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).

379. T. Strachan, S. Lindsay, D. I. Wilson, *Molecular Genetics of Early Human Development* (Taylor & Francis, 1997).

380. M. Belle, D. Godefroy, G. Couly, S. A. Malone, F. Collier, P. Giacobini, A. Chédotal, Tridimensional Visualization and Analysis of Early Human Development. *Cell* **169**, 161–173.e12 (2017).

381. *TotalSeq™-A Antibodies and Cell Hashing with 10x Single Cell 3' Reagent Kit v3.1 (Dual Index) Protocol*.

382. S. J. Fleming, J. C. Marioni, M. Babadi, CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. https://doi.org/10.1101/791699.

383. M. P. Mulè, A. J. Martins, J. S. Tsang, Normalizing and denoising protein expression data from droplet-based single cell profiling. https://doi.org/10.1101/2020.02.24.963603.

384. H. A. Sturges, The choice of a class interval. *J. Am. Stat. Assoc.* **21**, 65–66 (1926).

385. H. Wang, J. He, C. Xu, X. Chen, H. Yang, S. Shi, C. Liu, Y. Zeng, D. Wu, Z. Bai, M. Wang, Y. Wen, P. Su, M. Xia, B. Huang, C. Ma, L. Bian, Y. Lan, T. Cheng, L. Shi, B. Liu, J. Zhou, Decoding Human Megakaryocyte Development. *Cell Stem Cell* **28**, 535–549.e8 (2021).

386. A. Gayoso, R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi, V. Svensson, E. da Veiga Beltrame, V. Kleshchevnikov, C. Talavera-López, L. Pachter, F. J. Theis, A. Streets, M. I. Jordan, J. Regier, N. Yosef, A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).

387. M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, F. J. Theis, A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).

388. A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, N. Yosef, Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).

389. C. D. Conde, C. Domínguez Conde, C. Xu, L. B. Jarvis, T. Gomes, S. K. Howlett, D. B. Rainbow, O. Suchanek, H. W. King, L. Mamanova, K. Polanski, N. Huang, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, M. R. Clatworthy, K. Saeb-Parsy, J. L. Jones, S. A. Teichmann, Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture in humans. https://doi.org/10.1101/2021.04.28.441762.

390. R. C. V. Tyser, E. Mahammadov, S. Nakanoh, L. Vallier, A. Scialdone, S. Srinivas, Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285–289 (2021).

391. A. J. Tarashansky, J. M. Musser, M. Khariton, P. Li, D. Arendt, S. R. Quake, B. Wang, Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10** (2021).

392. A. Edelstein, N. Amodaj, K. Hoover, R. Vale, N. Stuurman, Computer Control of Microscopes Using μManager. [Preprint] (2010). https://doi.org/10.1002/0471142727.mb1420s92.

393. D. Hörl, F. Rojas Rusak, F. Preusser, P. Tillberg, N. Randel, R. K. Chhetri, A. Cardona, P. J. Keller, H. Harz, H. Leonhardt, M. Treier, S. Preibisch, BigStitcher: reconstructing high-resolution image datasets of cleared and expanded samples. *Nat. Methods* **16**, 870–874 (2019).

394. M. Gataric, J. S. Park, T. Li, V. Vaskivskyi, J. Svedlund, C. Strell, K. Roberts, M. Nilsson, L. R. Yates, O. Bayraktar, M. Gerstung, PoSTcode: Probabilistic image-based spatial transcriptomics decoder, *bioRxiv* (2021)p. 2021.10.12.464086.

395. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

396. M. Tavian, M. F. Hallais, B. Péault, Emergence of intraembryonic hematopoietic precursors in the pre-liver human embryo. *Development* **126**, 793–803 (1999).

397. C. Peschle, A. R. Migliaccio, G. Migliaccio, M. Petrini, M. Calandrini, G. Russo, G. Mastroberardino, M. Presta, A. M. Gianni, P. Comi, Embryonic----Fetal Hb switch in humans:

studies on erythroid bursts generated by embryonic progenitors from yolk sac and liver. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 2416–2420 (1984).

398. S. I. Nishikawa, S. Nishikawa, M. Hirashima, N. Matsuyoshi, H. Kodama, Progressive lineage analysis by cell sorting and culture identifies FLK1 VE-cadherin cells at a diverging point of endothelial and hemopoietic lineages. [Preprint] (1998). https://doi.org/10.1242/dev.125.9.1747.

399. S. J. Kinder, T. E. Tsang, G. A. Quinlan, A. K. Hadjantonakis, A. Nagy, P. P. Tam, The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. [Preprint] (1999). https://doi.org/10.1242/dev.126.21.4691.

400. P. D. F. Murray, *The Development in Vitro of the Blood of the Early Chick Embryo* (1932).

401. J. Palis, M. C. Yoder, Yolk-sac hematopoiesis: the first blood cells of mouse and man. *Exp. Hematol.* **29**, 927–936 (2001).

402. E. Dzierzak, A. Medvinsky, The discovery of a source of adult hematopoietic cells in the embryo. *Development* **135**, 2343–2346 (2008).

403. S. O'Byrne, N. Elliott, S. Rice, G. Buck, N. Fordham, C. Garnett, L. Godfrey, N. T. Crump, G. Wright, S. Inglott, P. Hua, B. Psaila, B. Povinelli, D. J. H. F. Knapp, A. Agraz-Doblas, C. Bueno, I. Varela, P. Bennett, H. Koohy, S. M. Watt, A. Karadimitris, A. J. Mead, P. Ancliff, P. Vyas, P. Menendez, T. A. Milne, I. Roberts, A. Roy, Discovery of a CD10-negative B-progenitor in human fetal life identifies unique ontogeny-related developmental programs. *Blood* **134**, 1059–1071 (2019).

404. U. C. Eze, A. Bhaduri, M. Haeussler, T. J. Nowakowski, A. R. Kriegstein, Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* **24**, 584–594 (2021).

405. B. J. Stewart, J. R. Ferdinand, M. D. Young, T. J. Mitchell, K. W. Loudon, A. M. Riding, N. Richoz, G. L. Frazer, J. U. L. Staniforth, F. A. Vieira Braga, R. A. Botting, D.-M. Popescu, R. Vento-Tormo, E. Stephenson, A. Cagan, S. J. Farndon, K. Polanski, M. Efremova, K. Green, M. Del Castillo Velasco-Herrera, C. Guzzo, G. Collord, L. Mamanova, T. Aho, J. N. Armitage, A. C. P. Riddick, I. Mushtaq, S. Farrell, D. Rampling, J. Nicholson, A. Filby, J. Burge, S. Lisgo, S. Lindsay, M. Bajenoff, A. Y. Warren, G. D. Stewart, N. Sebire, N. Coleman, M. Haniffa, S. A. Teichmann, S. Behjati, M. R. Clatworthy, Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).

406. G. L. Website, Tracking Early Mammalian Organogenesis – Prediction and Validation of Differentiation Trajectories at Whole Organism Scale, *ExtendedMouseAtlas*. https://marionilab.github.io/ExtendedMouseAtlas/.

407. Y. Zhu, T. Wang, J. Gu, K. Huang, T. Zhang, Z. Zhang, H. Liu, J. Tang, Y. Mai, Y. Zhang, Y. Li, Y. Feng, B. Kang, J. Li, Y. Shan, Q. Chen, J. Zhang, B. Long, J. Wang, M. Gao, D. Zhang, M. Zhou, X. Zhong, J. Chen, D. Pei, J. Nie, B. Liu, G. Pan, Characterization and generation of human definitive multipotent hematopoietic stem/progenitor cells. *Cell Discov* **6**, 89 (2020).

408. E. I. Crosse, S. Gordon-Keylock, S. Rybtsov, A. Binagui-Casas, H. Felchle, N. C. Nnadi, K. Kirschner, T. Chandra, S. Tamagno, D. J. Webb, F. Rossi, R. A. Anderson, A. Medvinsky, Multi-layered Spatial Transcriptomics Identify Secretory Factors Promoting Human Hematopoietic Stem Cell Development. *Cell Stem Cell* **27**, 822–839.e8 (2020).

409. W. Liu, O. Taso, R. Wang, S. Bayram, A. C. Graham, P. Garcia-Reitboeck, A. Mallach, W. D. Andrews, T. M. Piers, J. A. Botia, J. M. Pocock, D. M. Cummings, J. Hardy, F. A. Edwards, D. A. Salih, Trem2 promotes anti-inflammatory responses in microglia and is suppressed under

pro-inflammatory conditions. *Hum. Mol. Genet.* **29**, 3224–3248 (2020).

410. D. A. Jaitin, L. Adlung, C. A. Thaiss, A. Weiner, B. Li, H. Descamps, P. Lundgren, C. Bleriot, Z. Liu, A. Deczkowska, H. Keren-Shaul, E. David, N. Zmora, S. M. Eldar, N. Lubezky, O. Shibolet, D. A. Hill, M. A. Lazar, M. Colonna, F. Ginhoux, H. Shapiro, E. Elinav, I. Amit, Lipid-Associated Macrophages Control Metabolic Homeostasis in a Trem2-Dependent Manner. [Preprint] (2019). https://doi.org/10.1016/j.cell.2019.05.054.

411. Y. Wang, M. Cella, K. Mallinson, J. D. Ulrich, K. L. Young, M. L. Robinette, S. Gilfillan, G. M. Krishnan, S. Sudhakar, B. H. Zinselmeyer, D. M. Holtzman, J. R. Cirrito, M. Colonna, TREM2 Lipid Sensing Sustains the Microglial Response in an Alzheimer's Disease Model. [Preprint] (2015). https://doi.org/10.1016/j.cell.2015.01.049.

412. C. Peschle, F. Mavilio, A. Carè, G. Migliaccio, A. R. Migliaccio, G. Salvo, P. Samoggia, S. Petti, R. Guerriero, M. Marinucci, Haemoglobin switching in human embryos: asynchrony of zeta----alpha and epsilon----gamma-globin switches in primitive and definite erythropoietic lineage. *Nature* **313**, 235–238 (1985).

413. T. Jaffredo, R. Gautier, A. Eichmann, F. Dieterlen-Lièvre, Intraaortic hemopoietic cells are derived from endothelial cells during ontogeny. *Development* **125**, 4575–4583 (1998).

414. L. Yvernogeau, R. Gautier, L. Petit, H. Khoury, F. Relaix, V. Ribes, H. Sang, P. Charbord, M. Souyri, C. Robin, T. Jaffredo, In vivo generation of haematopoietic stem/progenitor cells from bone marrow-derived haemogenic endothelium. *Nat. Cell Biol.* **21**, 1334–1345 (2019).

415. Z. Li, Y. Lan, W. He, D. Chen, J. Wang, F. Zhou, Y. Wang, H. Sun, X. Chen, C. Xu, S. Li, Y. Pang, G. Zhang, L. Yang, L. Zhu, M. Fan, A. Shang, Z. Ju, L. Luo, Y. Ding, W. Guo, W. Yuan, X. Yang, B. Liu, Mouse Embryonic Head as a Site for Hematopoietic Stem Cell Development. [Preprint] (2012). https://doi.org/10.1016/j.stem.2012.07.004.

416. J. M. Frame, K. H. Fegan, S. J. Conway, K. E. McGrath, J. Palis, Definitive Hematopoiesis in the Yolk Sac Emerges from Wnt-Responsive Hemogenic Endothelium Independently of Circulation and Arterial Identity. *Stem Cells* **34**, 431–444 (2016).

417. K. E. Rhodes, C. Gekas, Y. Wang, C. T. Lux, C. S. Francis, D. N. Chan, S. Conway, S. H. Orkin, M. C. Yoder, H. K. A. Mikkola, The emergence of hematopoietic stem cells is initiated in the placental vasculature in the absence of circulation. *Cell Stem Cell* **2**, 252–263 (2008).

418. R. Thambyrajah, M. Mazan, R. Patel, V. Moignard, M. Stefanska, E. Marinopoulou, Y. Li, C. Lancrin, T. Clapes, T. Möröy, C. Robin, C. Miller, S. Cowley, B. Göttgens, V. Kouskoff, G. Lacaud, GFI1 proteins orchestrate the emergence of haematopoietic stem cells through recruitment of LSD1. *Nat. Cell Biol.* **18**, 21–32 (2016).

419. J. Fröbel, T. Landspersky, G. Percin, C. Schreck, S. Rahmig, A. Ori, D. Nowak, M. Essers, C. Waskow, R. A. J. Oostendorp, The Hematopoietic Bone Marrow Niche Ecosystem. *Front Cell Dev Biol* **9**, 705410 (2021).

420. B. Murdoch, K. Chadwick, M. Martin, F. Shojaei, K. V. Shah, L. Gallacher, R. T. Moon, M. Bhatia, Wnt-5A augments repopulating capacity and primitive hematopoietic development of human blood stem cells in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3422–3427 (2003).

421. J. Shen, Y. Zhu, S. Zhang, S. Lyu, C. Lyu, Z. Feng, D. L. Hoyle, Z. Z. Wang, T. Cheng, Vitronectin-activated αvβ3 and αvβ5 integrin signalling specifies haematopoietic fate in human pluripotent stem cells. *Cell Prolif.* **54**, e13012 (2021).

422. P. Zhang, C. Zhang, J. Li, J. Han, X. Liu, H. Yang, The physical microenvironment of hematopoietic stem cells and its emerging roles in engineering applications. *Stem Cell Res. Ther.*

10, 327 (2019).

423. A. S. Eisele, J. Cosgrove, A. Magniez, E. Tubeuf, S. Tenreira Bento, C. Conrad, F. Cayrac, T. Tak, A.-M. Lyne, J. Urbanus, L. Perié, Erythropoietin directly remodels the clonal composition of murine hematopoietic multipotent progenitor cells. *Elife* **11** (2022).

424. H. Yoshihara, F. Arai, K. Hosokawa, T. Hagiwara, K. Takubo, Y. Nakamura, Y. Gomei, H. Iwasaki, S. Matsuoka, K. Miyamoto, H. Miyazaki, T. Takahashi, T. Suda, Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche. *Cell Stem Cell* **1**, 685–697 (2007).

425. J. Xue, Q. Wu, L. A. Westfield, E. A. Tuley, D. Lu, Q. Zhang, K. Shim, X. Zheng, J. E. Sadler, Incomplete embryonic lethality and fatal neonatal hemorrhage caused by prothrombin deficiency in mice. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 7603–7607 (1998).

426. W. Ruf, N. Yokota, F. Schaffner, Tissue factor in cancer progression and angiogenesis. [Preprint] (2010). https://doi.org/10.1016/s0049-3848(10)70010-4.

427. H. Wu, X. Liu, R. Jaenisch, H. F. Lodish, Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. *Cell* **83**, 59–67 (1995).

428. I. Hirano, N. Suzuki, The Neural Crest as the First Production Site of the Erythroid Growth Factor Erythropoietin. *Front Cell Dev Biol* **7**, 105 (2019).

429. C. Gekas, K. E. Rhodes, L. M. Gereige, H. Helgadottir, R. Ferrari, S. K. Kurdistani, E. Montecino-Rodriguez, R. Bassel-Duby, E. Olson, A. V. Krivtsov, S. Armstrong, S. H. Orkin, M. Pellegrini, H. K. A. Mikkola, Mef2C is a lineage-restricted target of Scl/Tal1 and regulates megakaryopoiesis and B-cell homeostasis. *Blood* **113**, 3461–3471 (2009).

430. E. Suzuki, S. Williams, S. Sato, G. Gilkeson, D. K. Watson, X. K. Zhang, The transcription factor Fli-1 regulates monocyte, macrophage and dendritic cell development in mice. *Immunology* **139**, 318–327 (2013).

431. M. L. Petreaca, M. Yao, Y. Liu, K. DeFea, M. Martins-Green, Transactivation of Vascular Endothelial Growth Factor Receptor-2 by Interleukin-8 (IL-8/CXCL8) Is Required for IL-8/CXCL8-induced Endothelial Permeability. *Mol. Biol. Cell*, doi: 10.1091/mbc.e07-01-0004 (2007).

432. Z. Lyu, H. Jin, Z. Yan, K. Hu, H. Jiang, H. Peng, H. Zhuo, Effects of NRP1 on angiogenesis and vascular maturity in endothelial cells are dependent on the expression of SEMA4D. *Int. J. Mol. Med.* **46**, 1321–1334 (2020).

433. K. Bisht, K. A. Okojie, K. Sharma, D. H. Lentferink, Y.-Y. Sun, H.-R. Chen, J. O. Uweru, S. Amancherla, Z. Calcuttawala, A. B. Campos-Salazar, B. Corliss, L. Jabbour, J. Benderoth, B. Friestad, W. A. Mills 3rd, B. E. Isakson, M.-È. Tremblay, C.-Y. Kuan, U. B. Eyo, Capillary-associated microglia regulate vascular structure and function through PANX1-P2RY12 coupling in mice. *Nat. Commun.* **12**, 5289 (2021).

434. A. C. Yang, R. T. Vest, F. Kern, D. P. Lee, C. A. Maat, P. M. Losada, M. B. Chen, M. Agam, N. Schaum, N. Khoury, K. Calcuttawala, R. Pálovics, A. Shin, E. Y. Wang, J. Luo, D. Gate, J. A. Siegenthaler, M. Windy McNerney, A. Keller, T. Wyss-Coray, A human brain vascular atlas reveals diverse cell mediators of Alzheimer's disease risk, *bioRxiv* (2021)p. 2021.04.26.441262.

435. A. Brazovskaja, T. Gomes, C. Körner, Z. He, T. Schaffer, J. C. Eckel, R. Hänsel, M. Santel, T. Denecke, M. Dannemann, M. Brosch, J. Hampe, D. Seehofer, G. Damm, J. G. Camp, B. Treutlein, Cell atlas of the regenerating human liver after portal vein embolization, *bioRxiv*

(2021). https://doi.org/10.1101/2021.06.03.444016.

436. N. Tyagi, A. M. Roberts, W. L. Dean, S. C. Tyagi, D. Lominadze, Fibrinogen induces endothelial cell permeability. *Mol. Cell. Biochem.* **307**, 13–22 (2008).

437. S. G. Utz, P. See, W. Mildenberger, M. S. Thion, A. Silvin, M. Lutz, F. Ingelfinger, N. A. Rayan, I. Lelios, A. Buttgereit, K. Asano, S. Prabhakar, S. Garel, B. Becher, F. Ginhoux, M. Greter, Early Fate Defines Microglia and Non-parenchymal Brain Macrophage Development. *Cell* **181**, 557–573.e18 (2020).

438. A. Fantin, J. M. Vieira, G. Gestri, L. Denti, Q. Schwarz, S. Prykhozhij, F. Peri, S. W. Wilson, C. Ruhrberg, Tissue macrophages act as cellular chaperones for vascular anastomosis downstream of VEGF-mediated endothelial tip cell induction. *Blood* **116**, 829–840 (2010).

439. D. Davalos, J. K. Ryu, M. Merlini, K. M. Baeten, N. Le Moan, M. A. Petersen, T. J. Deerinck, D. S. Smirnoff, C. Bedard, H. Hakozaki, S. Gonias Murray, J. B. Ling, H. Lassmann, J. L. Degen, M. H. Ellisman, K. Akassoglou, Fibrinogen-induced perivascular microglial clustering is required for the development of axonal damage in neuroinflammation. *Nat. Commun.* **3**, 1227 (2012).

440. S. Chou, H. F. Lodish, Fetal liver hepatic progenitors are supportive stromal cells for hematopoietic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **107** (2010).

441. G. Migliaccio, A. R. Migliaccio, S. Petti, F. Mavilio, G. Russo, D. Lazzaro, U. Testa, M. Marinucci, C. Peschle, Human embryonic hemopoiesis. Kinetics of progenitors and precursors underlying the yolk sac----liver transition. [Preprint] (1986). https://doi.org/10.1172/jci112572.

442. N. Abdelfattah, P. Kumar, C. Wang, J.-S. Leu, W. F. Flynn, R. Gao, D. S. Baskin, K. Pichumani, O. B. Ijare, S. L. Wood, S. Z. Powell, D. L. Haviland, B. C. Parker Kerrigan, F. F. Lang, S. S. Prabhu, K. M. Huntoon, W. Jiang, B. Y. S. Kim, J. George, K. Yun, Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target. *Nat. Commun.* **13**, 767 (2022).

443. C. A. Lareau, K. R. Parker, A. T. Satpathy, Charting the tumor antigen maps drawn by single-cell genomics. *Cancer Cell* **39**, 1553–1557 (2021).

444. S. A. Liebhaber, J. E. Russell, Expression and developmental control of the human alpha-globin gene cluster. *Ann. N. Y. Acad. Sci.* **850**, 54–63 (1998).

445. L. Hartmann, S. Dutta, S. Opatz, S. Vosberg, K. Reiter, G. Leubolt, K. H. Metzeler, T. Herold, S. A. Bamopoulos, K. Bräundl, E. Zellmeier, B. Ksienzyk, N. P. Konstandin, S. Schneider, K.-P. Hopfner, A. Graf, S. Krebs, H. Blum, J. M. Middeke, F. Stölzel, C. Thiede, S. Wolf, S. K. Bohlander, C. Preiss, L. Chen-Wichmann, C. Wichmann, M. C. Sauerland, T. Büchner, W. E. Berdel, B. J. Wörmann, J. Braess, W. Hiddemann, K. Spiekermann, P. A. Greif, ZBTB7A mutations in acute myeloid leukaemia with t(8;21) translocation. *Nat. Commun.* **7**, 11733 (2016).

446. Y. Yu, J. Wang, W. Khaled, S. Burke, P. Li, X. Chen, W. Yang, N. A. Jenkins, N. G. Copeland, S. Zhang, P. Liu, Bcl11a is essential for lymphoid development and negatively regulates p53. *J. Exp. Med.* **209**, 2467–2483 (2012).

447. Y. Roohani, K. Huang, J. Leskovec, Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935 (2024).

448. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–7 (2012).

449. M. Y. Pai, B. Lomenick, H. Hwang, R. Schiestl, W. McBride, J. A. Loo, J. Huang, Drug

affinity responsive target stability (DARTS) for small-molecule target identification. *Methods Mol. Biol.* **1263**, 287–298 (2015).

450. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, B. Wang, scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).

451. C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, P. T. Ellinor, Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

452. The AI community building the future. https://huggingface.co.

453. AI-Supported Virtual Cells Platform - CZI. https://virtualcellmodels.cziscience.com/.

454. A.-A. Olijnik, A. Rodriguez-Romera, Z. C. Wong, Y. Shen, J. S. Reyat, N. J. Jooss, J. Rayes, B. Psaila, A. O. Khan, Generating human bone marrow organoids for disease modeling and drug discovery. *Nat. Protoc.*, doi: 10.1038/s41596-024-00971-7 (2024).

455. S. Demirci, J. J. Haro-Mora, A. Leonard, C. Drysdale, D. Malide, K. Keyvanfar, K. Essawi, R. Vizcardo, N. Tamaoki, N. P. Restifo, J. F. Tisdale, N. Uchida, Definitive hematopoietic stem/progenitor cells from human embryonic stem cells through serum/feeder-free organoid-induced differentiation. *Stem Cell Res. Ther.* **11**, 493 (2020).

456. S. Demirci, A. Leonard, J. F. Tisdale, Hematopoietic stem cells from pluripotent stem cells: Clinical potential, challenges, and future perspectives. *Stem Cells Transl. Med.* **9**, 1549–1557 (2020).

457. A. Fidanza, L. M. Forrester, Progress in the production of haematopoietic stem and progenitor cells from human pluripotent stem cells. *J Immunol Regen Med* **13**, 100050 (2021).

458. N. Tamaoki, S. Siebert, T. Maeda, N.-H. Ha, M. L. Good, Y. Huang, S. K. Vodnala, J. J. Haro-Mora, N. Uchida, J. F. Tisdale, C. L. Sweeney, U. Choi, J. Brault, S. Koontz, H. L. Malech, Y. Yamazaki, R. Isonaka, D. S. Goldstein, M. Kimura, T. Takebe, J. Zou, D. F. Stroncek, P. G. Robey, M. J. Kruhlak, N. P. Restifo, R. Vizcardo, Self-organized yolk sac-like organoids allow for scalable generation of multipotent hematopoietic progenitor cells from induced pluripotent stem cells. *Cell Rep Methods* **3**, 100460 (2023).

459. J. Hislop, Q. Song, K. Keshavarz F, A. Alavi, R. Schoenberger, R. LeGraw, J. J. Velazquez, T. Mokhtari, M. N. Taheri, M. Rytel, S. M. Chuva de Sousa Lopes, S. Watkins, D. Stolz, S. Kiani, B. Sozen, Z. Bar-Joseph, M. R. Ebrahimkhani, Modelling post-implantation human development to yolk sac blood emergence. *Nature* **626**, 367–376 (2024).

460. T. E. Bakken, R. D. Hodge, J. A. Miller, Z. Yao, T. N. Nguyen, B. Aevermann, E. Barkan, D. Bertagnolli, T. Casper, N. Dee, E. Garren, J. Goldy, L. T. Graybuck, M. Kroll, R. S. Lasken, K. Lathia, S. Parry, C. Rimorin, R. H. Scheuermann, N. J. Schork, S. I. Shehata, M. Tieu, J. W. Phillips, A. Bernard, K. A. Smith, H. Zeng, E. S. Lein, B. Tasic, Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648 (2018).

9 Appendices

This section contains supplementary appendices including figures A1-A9, and captions for figures A1-A9.

Appendix Figure 1

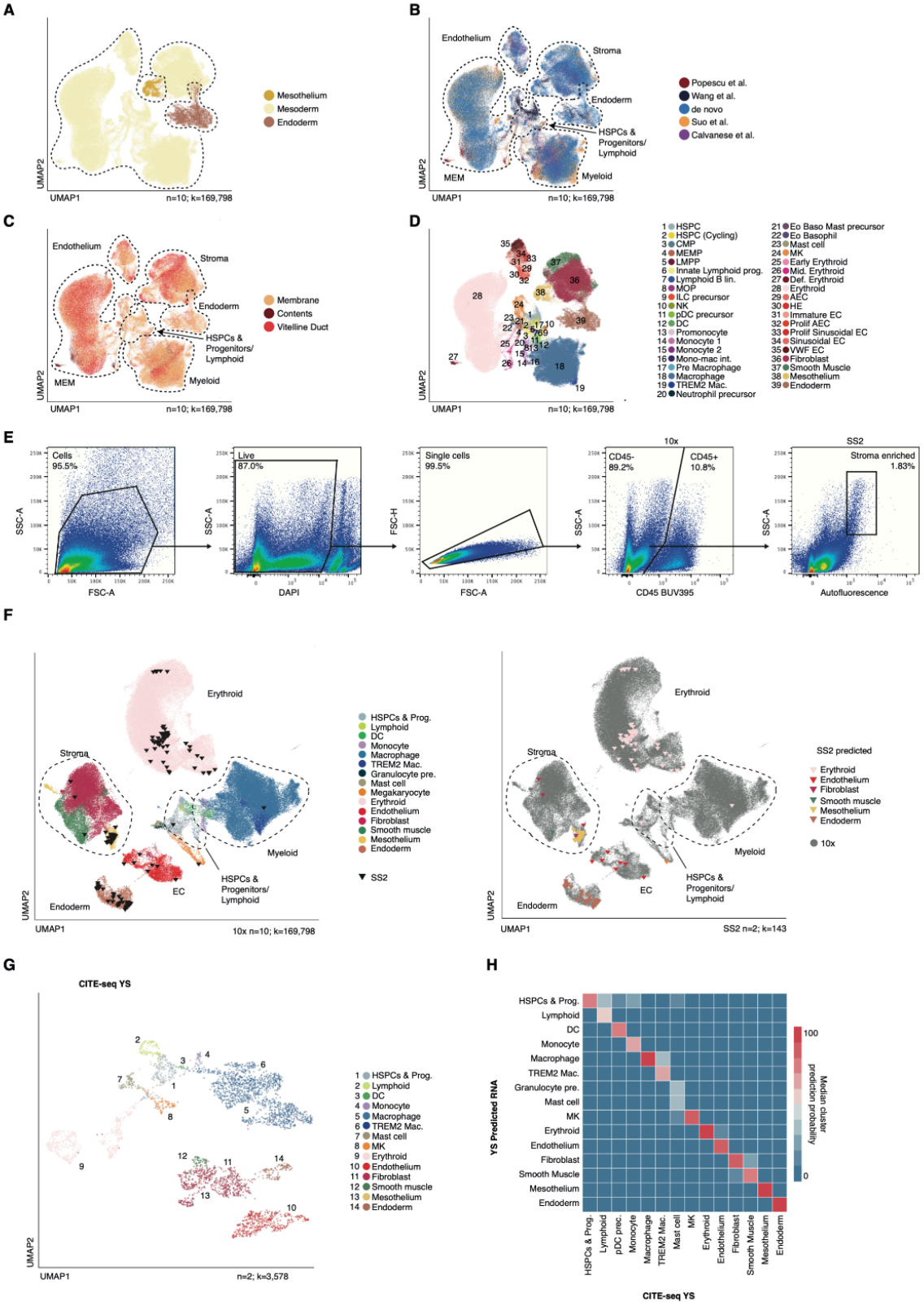


Fig. A1. A single-cell atlas of the human yolk sac. *(A to D) UMAP visualization of scRNAseq data shown in Fig. 3.4 , C colored according to yolk sac (YS) tissue layer (A), data source (B), anatomical structure (C), and refined annotation (D). HSPC:= hematopoietic stem/progenitor cell,; CMP:= common myeloid progenitor,; MEMP:= megakaryocyte–erythroid–mast cell progenitor,; LMPP:= lymphoid–primed multipotent progenitor,; Prog.:= progenitor,; MOP:= monocyte progenitor,; ILC:= innate lymphoid cell,; NK:= natural killer cell,; pDC:= plasmacytoid DC,; pre.:= precursor,; DC:= dendritic cell,; Mac.:= macrophage,; Mono–mac int.: = monocyte macrophage intermediate,; Eo: eosinophil; Baso:= eosinophil basophil,; MK:= megakaryocyte,; AEC:= arteriolar endothelial cell,; HE:= hemogenic endothelium,; EC:= endothelial cell. (E) FACS gating strategy used to sort YS CD45+/–+/– fractions for plate–based scRNA–seq. Sequential gates show selection of CD45––SSChi autofluorescent cells, enriched in non–erythroid stromal populations. Representative gating from n=2 independent samples (5–7PCW). (F) UMAP visualization of the YS cells shown in Fig. 1C (n=10, c=169,798) integrated with c=143 cells from plate–based sequencing (SS2) cells (triangles) FACS–isolated from n=2 individual donors (5–7PCW). Left: colors indicate cell states in droplet based scRNA–seq (10X). Right; colors indicate predicted cell states in plate–based scRNA–seq (SS2). (G) UMAP of YS cells from CITE–seq data, intersected between RNA and protein modalities batch corrected using TotalVI from n=2 biologically independent samples (n=2, c=3,578). Colors represent broad cell states. (H) Heatmap of class prediction probabilities for a logistic regression model (Elasticnet) trained on YS scRNA–seq cell states (y–axis) and projected onto corresponding cluster–derived cell states in YS CITE–seq data (x–axis). Color scale indicates median*

probabilities. Figure adapted from Goh and Botting et al, 2023 (1).

Appendix Figure 2

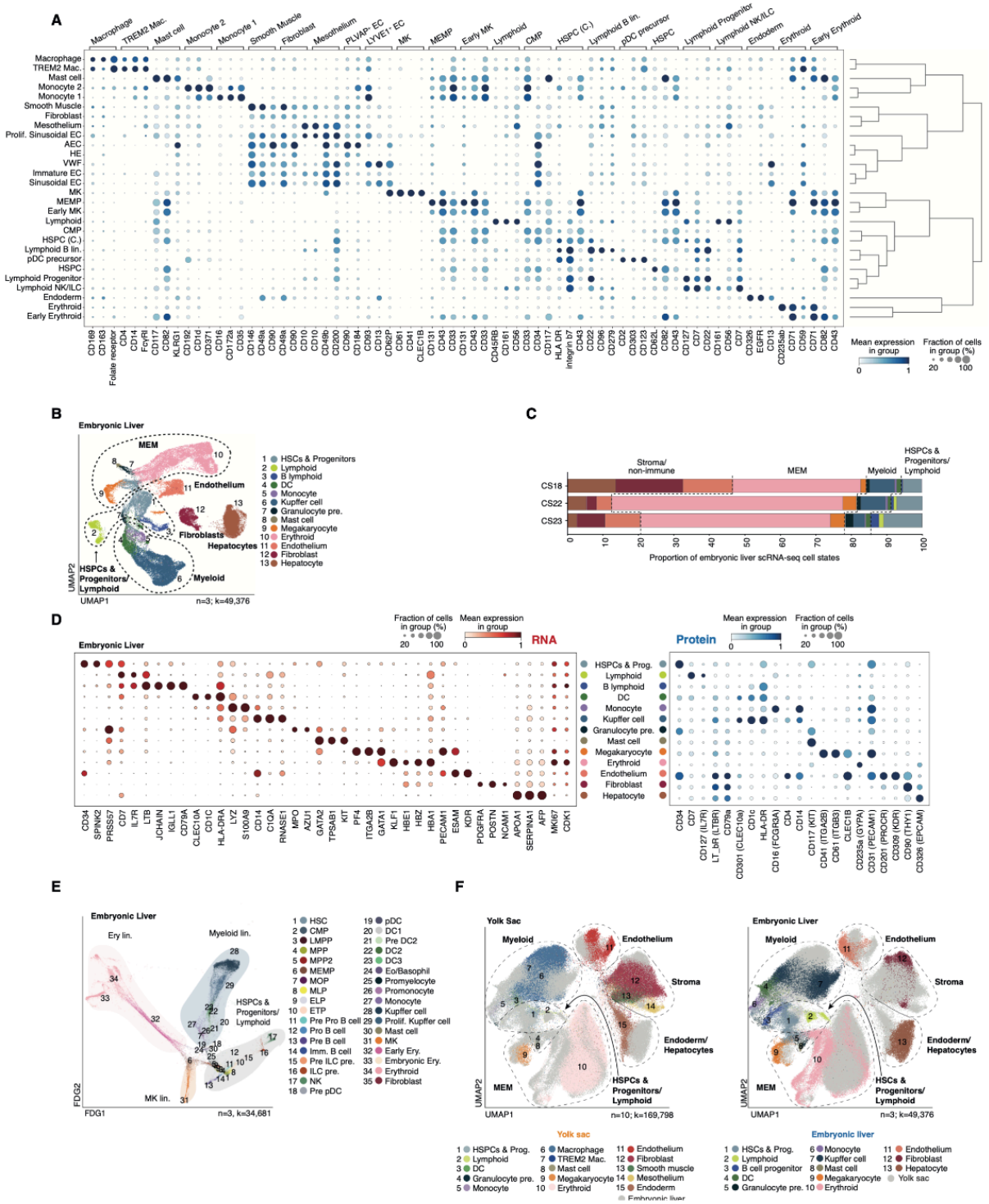


Fig. A2. CITE-seq analysis of human yolk sac and embryonic liver. *(A) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each protein (dot size) of surface proteins in YS CITE-seq data analyzed using totalVI. Cell states were grouped using a hierarchical dendrogram. Marker proteins were derived using a one-vs-all TotalVI differential expression test (P<0.05, Bayes factor >0.95, Median LFC >0.51) (more details on this approach are available in the revised methods section under 'Clustering and annotation of scRNA-seq and CITE-seq data'). Data are min-max-standardized with a distribution of 0-1. . (B) UMAP visualization of the cell states identified in the embryonic liver (EL) scRNA-seq dataset from n=3 independent biological repeats (c=49,376, CS18-23). Colors represent cell states. DC: dendritic cell; MEM: megakaryocyte-erythroid-mast cell lineage; pre.: precursor . (C) Stacked bar chart displaying the proportional representation of broad cell states in EL scRNA-seq data by sample. Colors match cell states shown in (B). (D) Dot plot showing the mean expression (by color) and the fraction of cells expressing each gene or protein (by dot size) of broad cell state-defining genes in EL scRNA-seq data (n=3, c=49,376) (left), and their protein counterparts in EL CITE-seq dataset (n=9 biologically independent samples, c=57,310, CS16-17PCW) (right). Data are min-max-standardized with a distribution of 0-1. (E) Force-directed graph (FDG) visualization of hematopoietic cell states identified in EL scRNA-seq from n=3 biologically independent donors (c=34,681). Colors represent cell states and clouds represent lineages. CMP:common myeloid progenitor; DC: dendritic cell; ELP: early lymphoid progenitor; Eo./Baso:eosinophil/basophil; Ery.: erythroid; ETP: early thymic progenitor; HE: hemogenic endothelium; HSC: hematopoietic stem cell; HSPC: hematopoietic stem progenitor cell; ILC: innate lymphoid cell; LMPP: lymphoid-primed multipotent progenitor; Mac: macrophage; MEM: megakaryocyte-erythroid-mast cell lineage; MEMP: megakaryocyte-erythroid-mast cell progenitor; MK: megakaryocyte; MLP: multi-lymphoid progenitor; Mono: monocyte; MOP: monocyte progenitor; MPP: multipotent progenitor; Neut: neutrophil; NK: natural killer cell; pDC: plasmacytoid DC;pre.: precursor;*

prog.: progenitor; prolif.: proliferating. (F) UMAPvisualization of the merged YS and EL scRNA-seq data shown in Fig. 1B and fig. S3B,respectively, colored by cell state and tissue (Left, YS, n=10, c=169,798; Right, EL, n=3,c=49,376). Figure adapted from Goh and Botting et al, 2023 (1).

Appendix Figure 3

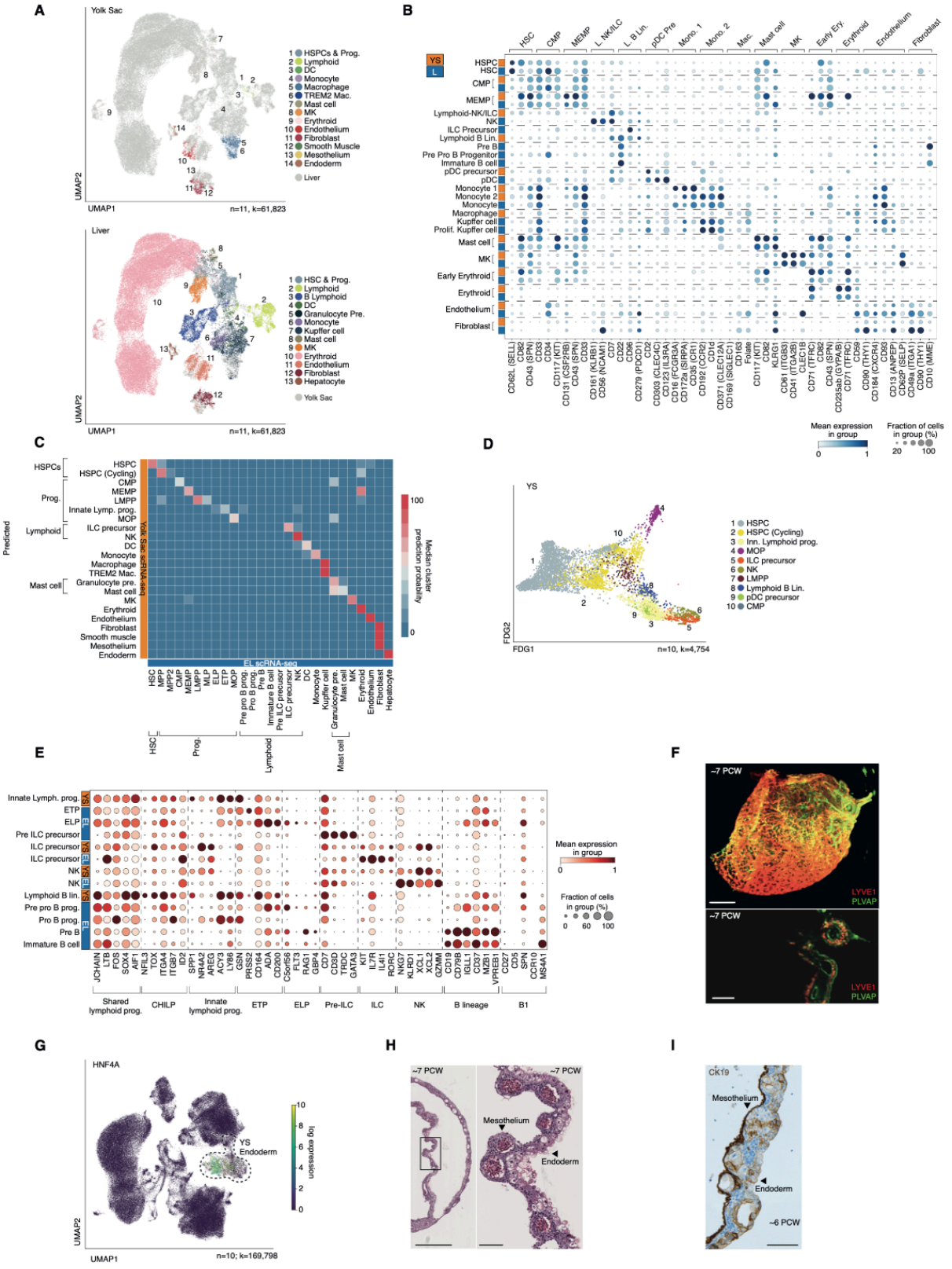


Fig. A3. Comparison between yolk sac and embryonic liver cell states and yolk sac anatomy. *(A) UMAP visualization of combined YS (top) and liver (bottom) CITE-seq datasets (n=11, c=61,823). Colors represent cell states. Colors represent cell states (B) Dot plot*

showing the mean expression (by color) and the fraction of cells expressing each protein (by dot size) of proteins derived from fig. S2A for each matched refined cell state in the YS and liver CITE-seq datasets. The equivalent gene symbol is shown in parentheses when different from the protein. HSC: hematopoietic stem cell; HSPC: hematopoietic stem and progenitor cell; CMP: common myeloid progenitor; MEMP: megakaryocyte–erythroid–mast cell progenitor; pDC:plasmacytoid dendritic cell; MK: megakaryocyte (Data are min-max-standardized with a distribution of 0-1). (C) Heatmap of class prediction probabilities for a logistic regression model (Elasticnet) trained on embryonic liver (EL) scRNA-seq cell states (x-axis) and projected onto corresponding YS scRNA-seq cell states (y-axis). Color scale indicates median probabilities. Brackets indicate broad cell-state groups as shown in (A). (D) Force directed graph (FDG) visualization of lymphoid cell states in the YS scRNA-seq dataset (n=10, c=4754). (E) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (by dot size) of lymphoid marker genes in lymphoid lineage cell states in the YS and EL scRNA-seq datasets. Data are min-max-standardized with a distribution of 0-1. (F) Representative images of 3D (top) and 2D z-stack (bottom) light-sheet fluorescence microscopy images of 7-PCW YS stained with anti-LYVE1 (red) and anti-PLVAP antibodies (green). Scale bars: 700 μm (top) and 100 μm (bottom). (G) Feature plot of HNF4A expression in YS scRNA-seq data (Fig. 1C) log-normalized and scaled max expression value=10 (n=10, c=169,798). (H) Histology of a 7-PCW YS, demonstrated by hematoxylin and eosin staining of a formalin-fixed paraffin-embedded tissue section. A representative image from one of n=4 biologically independent samples (4-8PCW) is shown, with mesothelium and endoderm marked by arrows. Scale bars: 1 mm (left) and 100 μm (right). (I) Immunohistochemistry stain for CK19 (brown) in a 7-PCW YS, representative image from one of n=4 biologically independent samples. Scale bar: 100 μm. Figure adapted from Goh and Botting et al, 2023 (1).

Appendix Figure 4

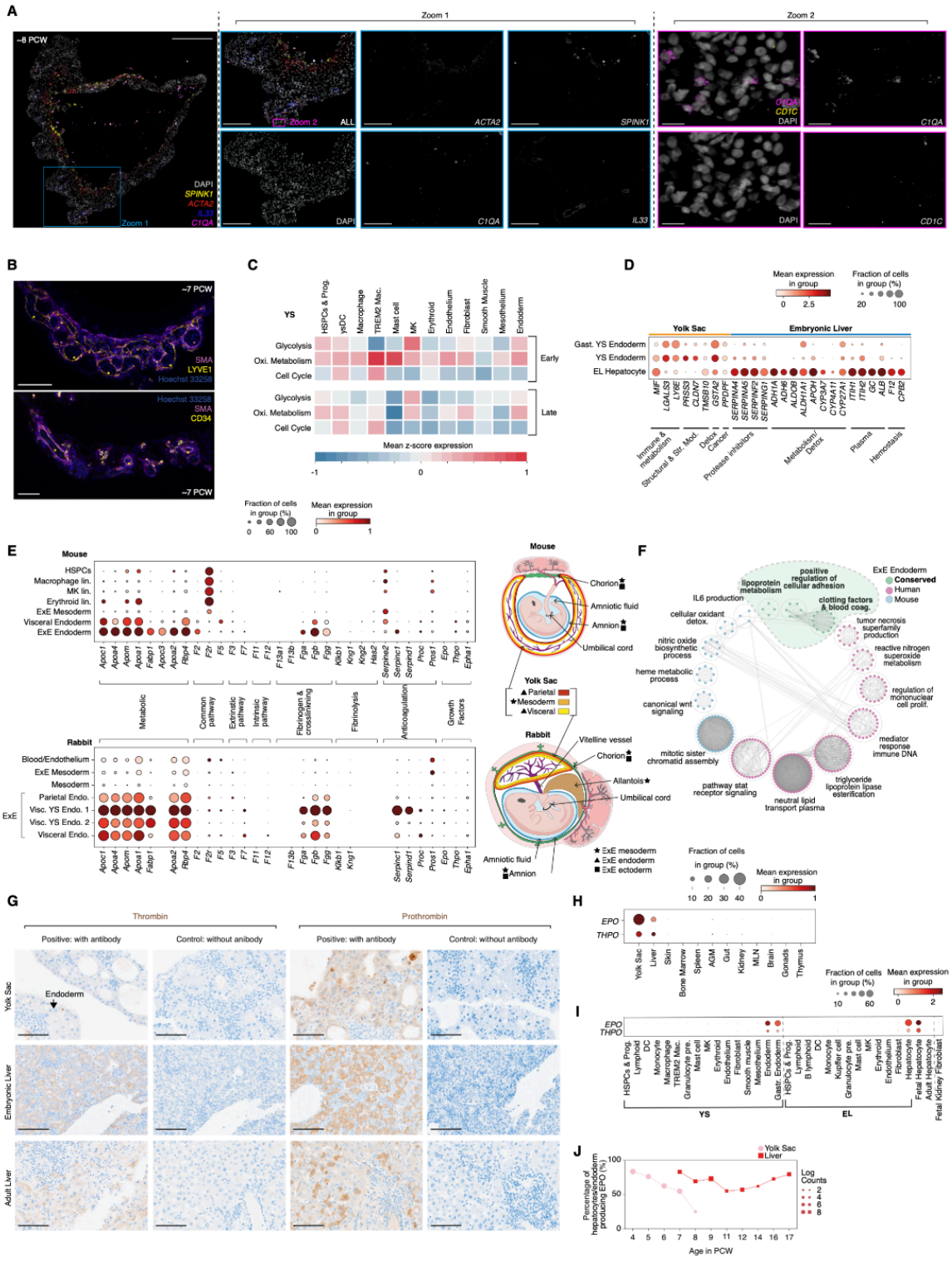


Fig. A4. Multiorgan functions of the yolk sac. *(A) RNAscope imaging of an 8-PCW YS with probes specific for endoderm (SPINK1), smooth muscle (ACTA2), arteriolar endothelial cells (IL33), macrophages (C1QA), and dendritic cells (CD1C; right "Zoom 2" panels only), costained with DAPI. Blue box ("Zoom 1") and magenta box ("Zoom 2") indicate ROIs shown in Fig. 1F. "Zoom 1" and "Zoom 2" panels show greyscale images of the individual channels in Fig. 1F. Left scale bar: 500 μm; middle "Zoom 1" scale bars: 200 μm; right "Zoom 2" scale bars: 50 μm . (B) Lightsheet immunofluorescence imaging of 7-PCW YS stained with anti-SMA (magenta) and anti-LYVE1 (yellow; top) or anti-CD34 (yellow, bottom) antibodies, costained with Hoechst 33268 (blue) and imaged using confocal microscopy. Scale bar: 100 μm . * indicates vessels. (C) Heatmap of z-normalized GO geneset module scores between early and late predicted Milo neighborhoods for cell cycle (GO:0022402), oxidative metabolism (GO:0045333), and glycolysis (GO:0006096), subtracted by the mean expression of 200 randomly sampled genes at 50 bins . (D) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (by dot size) of selected DEGs between YS endoderm (main and gastrula (gast.) data) and embryonic liver (EL) hepatocytes (data scaled max_value=10, gastrulation data scaled independently). Brackets indicate curated enriched GO annotations for each set of genes. Str. Mod.: Structural modification. (E) Left: Dot plots showing the mean expression (color scale) and the fraction of cells expressing each gene (by dot size) of clotting and soluble factors in relevant cell states from mouse gastrulation scRNA-seq data (75) (top) and from rabbit scRNA-seq data (55) (bottom). Data are min-max-standardized with a distribution of 0-1. Right: illustrations of developing mouse (~E9.5) and rabbit (~GD9) embryos. Text legends indicate corresponding extraembryonic anatomical regions between species, whereas shapes indicate germ layer origin of the anatomical regions. Star: mesoderm; triangle: endoderm; and square: ectoderm. Layers of the YS are delineated by color. Red: parietal; orange: mesoderm; and yellow: visceral. (F) Flower plot of the significant gene sets enriched in YS endoderm (pink), mouse*

extraembryonic (ExE) endoderm (blue), and conserved between species (green). Nodes indicate significantly enriched gene sets (Q-value <0.05), whereas edges between nodes represent gene overlap between gene sets. Annotated grouping circles indicate Markov cluster neighborhoods of gene expression modules which share high gene-set similarities. (G) IHC antibody staining of thrombin (F2) (column 1), prothrombin (column 3), and the respective controls without antibody (columns 3 and 4) in 7-PCW YS (top), 7-PCW EL (middle), and healthy adult liver (bottom). Representative images from 1 of n=3, 3, and 3 biologically independent YS (4-7PCW), ELs (7-12PCW) and adult livers, respectively. Protein (brown) and nuclei (blue). Scale bar: 100 μm (H) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of Endoderm-derived soluble factors between YS endoderm (main and gastrula (gast.) data) and embryonic liver (EL) hepatocytes, and non-immune cells from skin, bone marrow, spleen, aorta–gonad–mesonephros (AGM), gut, kidney, mesenteric lymph nodes (MLN), brain, gonads, and thymus. Data are min-max-standardized with a distribution of 0-1. (I) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of soluble factors in YS (main and gastrula), EL, fetal/adult liver, and fetal kidney scRNA-seq select cell states (each dataset scaled max_value=10 independently then combined except YS and EL scRNA-seq). (J) Line graph showing the relative change in proportion of endoderm and hepatocyte cell states in YS and liver, respectively (y-axis), enriched in expression of EPO. Pink: human YS main and gastrulation scRNA-seq data. Red: matched Liver scRNA-seq data. Figure adapted from Goh and Botting et al, 2023 (1).

Appendix Figure 5

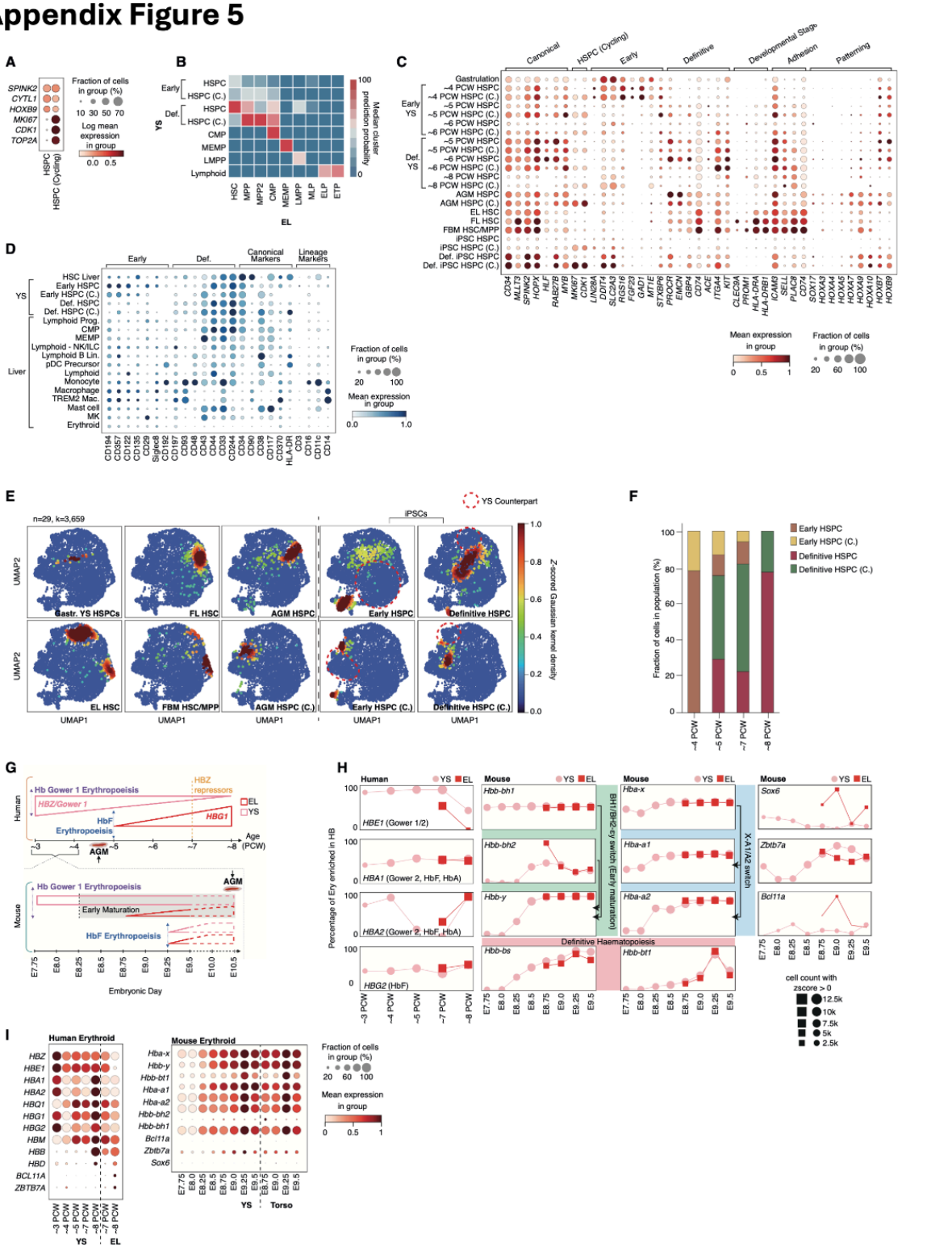


Fig. A5. Early versus definitive hematopoiesis in yolk sac and liver. *(A) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of genes distinguishing YS HSPC from YS HSPC (C.) (cycling HSPC) in the YS (main) scRNA-seq data (data scaled max_value=10). (B) Median ElasticNet logistic regression class prediction probabilities for a model trained on EL progenitor scRNA-seq cell states (x-axis)*

projected onto YS scRNA-seq cell states (y-axis). (C) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of canonical, cycling HSPC-specific, early, definitive, developmental-stage specific, adhesion and patterning HSC markers expressed between YS HSPCs (split by HSPC/ cycling HSPC and early/definitive) across time including gastrulation (67), AGM HSPC (66), matched EL HSC, FL HSC (6), fetal BM HSC/MPP (3), iPSC-derived HSPC (89), and definitive iPSC-derived HPSC (46). Data are min-max-standardized with a distribution of 0-1. (D) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each protein (dot size) of differentially expressed proteins between early and definitive HSPCS, alongside canonical HSC markers and lineage markers, in selected cell states from YS and liver CITE-seq data. Data are min-max-standardized with a distribution of 0-1. (E) Density plots showing the distribution of indicated HSPC populations in the integrated UMAP landscape of HSPC/HSCs from the following scRNAseq datasets: YS (n=10, c=2,597), YS gastrulation (390) (n=1, c=23), AGM (408) (n=3, c=182), matched embryonic liver (EL) (n=3, c=412), fetal liver (FL) (n=14, c=242), fetal bone marrow (3) (FBM) (n=9, c=92), iPSC-derived HSPC (n=12, c=355) (89) and definitive iPSC-derived HSPC (n=2, c=273) (46). The color of HSC/HSPCs represents the Z-scored kernel density estimation (KDE) score for each population. Red dashed lines indicate the embedded positions of corresponding YS HSPC populations as shown in Fig. 3C. (F) Bar chart showing the proportional representation of early YS HSPC and cycling HSPC to definitive YS HSPC and cycling HSPC in the main and gastrulation YS scRNA-seq data (grouped by gestational age in PCW). (G) Schematic diagram showing the relative timescales of early and definitive erythropoiesis in human and mouse, and contributions of AGM, EL, and YS to this process. (H) Left column: Line graphs showing the relative change in expression of Gower 1/2 globin HBE1, Gower 2 globins HBA1/2 and definitive globin HBG2 in human erythroid cells from YS (pink) and matched embryonic liver (EL) (; red) over gestational age. Central and right columns: Globin expression in mouse erythroid cells (406) including HBB BH1, εγ, X, A2, BT1 and BS. Pink lines: mouse YS scRNA-seq data. Red lines: aged-matched mouse torso scRNA-seq data. Mouse hemoglobins implicated in primitive maturation, definitive hematopoiesis, and a switch between the two are grouped. The y-axis represents the proportion of erythroid lineage cells. (I) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of Hb genes in erythroid lineage cells, grouped by gestational age. Left: human YS scRNA-seq (main and gastrulation) and EL data. Right: mouse YS and aged-matched torso scRNA-seq data (406). Data are min-max-standardized with a

distribution of 0-1. Figure adapted from Goh and Botting et al, 2023 (1).

Appendix Figure 6

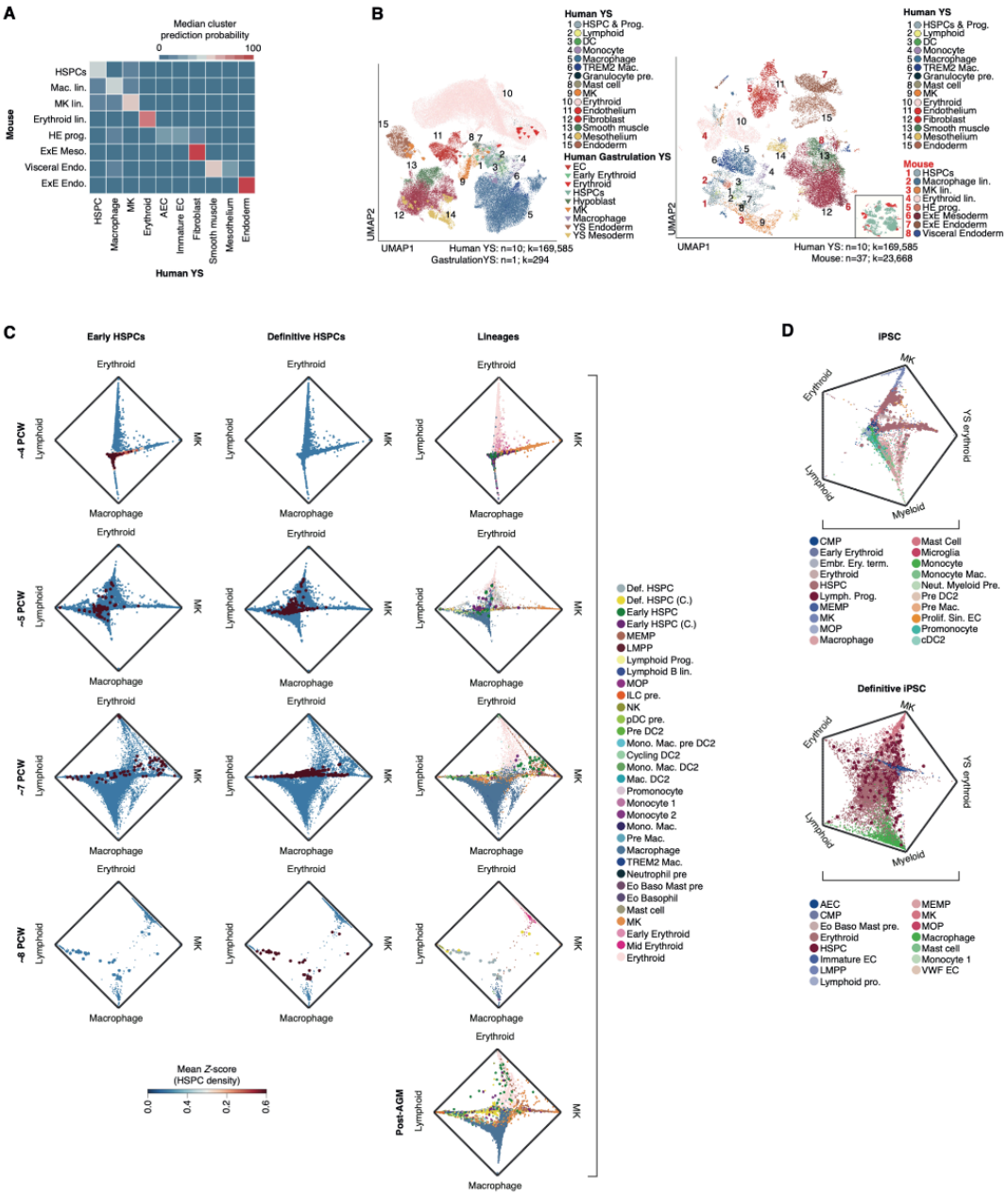


Fig. A6. Hematopoietic waves in human yolk sac. *(A) Heatmap of class prediction probabilities for a LR model (Elasticnet) trained on human YS scRNA-seq cell states (x-axis) projected onto corresponding mouse extraembryonic cell states from mouse gastrulation dataset (75) (y-axis). Prog.: progenitor; AEC: arteriolar endothelial cell; EC: endothelial cell; HSPC: hematopoietic stem and progenitor cell. Color scale indicates median probabilities. (B) Left: UMAP visualization of matched hematopoietic cell states in human YS scRNA-seq (dots) as shown in Fig. 1C (n=10, c=169,798) integrated with human gastrulation (CS7) scRNA-seq data (14) (triangles) (n=1, c=91). Lin.: lineage; pre.: precursor; DC: dendritic cell; MK: megakaryocyte; EC: endothelial cell. Right: UMAP visualization of human YS scRNA-seq (as shown in Fig. 1C) and equivalent mouse gastrulation extraembryonic cell states (75) (n=36, c=139,331). Insert highlighting location of mouse cell states within UMAP. Colors represent cell states. ExE.: extra embryonic; lin.: lineage; HE: hemogenic endothelium. Inset colored by species (mouse: red; human: teal). (C) Radial plots showing relative probabilities of lineage-state transition between HSPCs and lineage-specific cell states starting from early HSPC (left) and definitive HSPC (middle). Right-hand column shows cell state annotations. Plots are segregated by gestational stages between CS10-11, CS14-15, CS17-18 and CS22-23. Color indicates the HSPC population density as a z-scored kernel density estimation (KDE) score and the position of HSPC population densities indicate respective lineage priming probability between macrophage, lymphoid (NK and B lineage), erythroid and MK terminal states. (D) Radial plots showing relative probabilities of lineage-state transition between iPSC-derived HSPCs from the culture protocol optimized for macrophage differentiation (top) and from the definitive iPSC culture protocol (bottom). Color indicates the cell-state annotation and the position of each population indicates respective lineage priming probability between myeloid, lymphoid erythroid, embryonic erythroid and MK terminal states. Figure adapted from Goh and Botting et al, 2023 (1).*

Appendix Figure 7

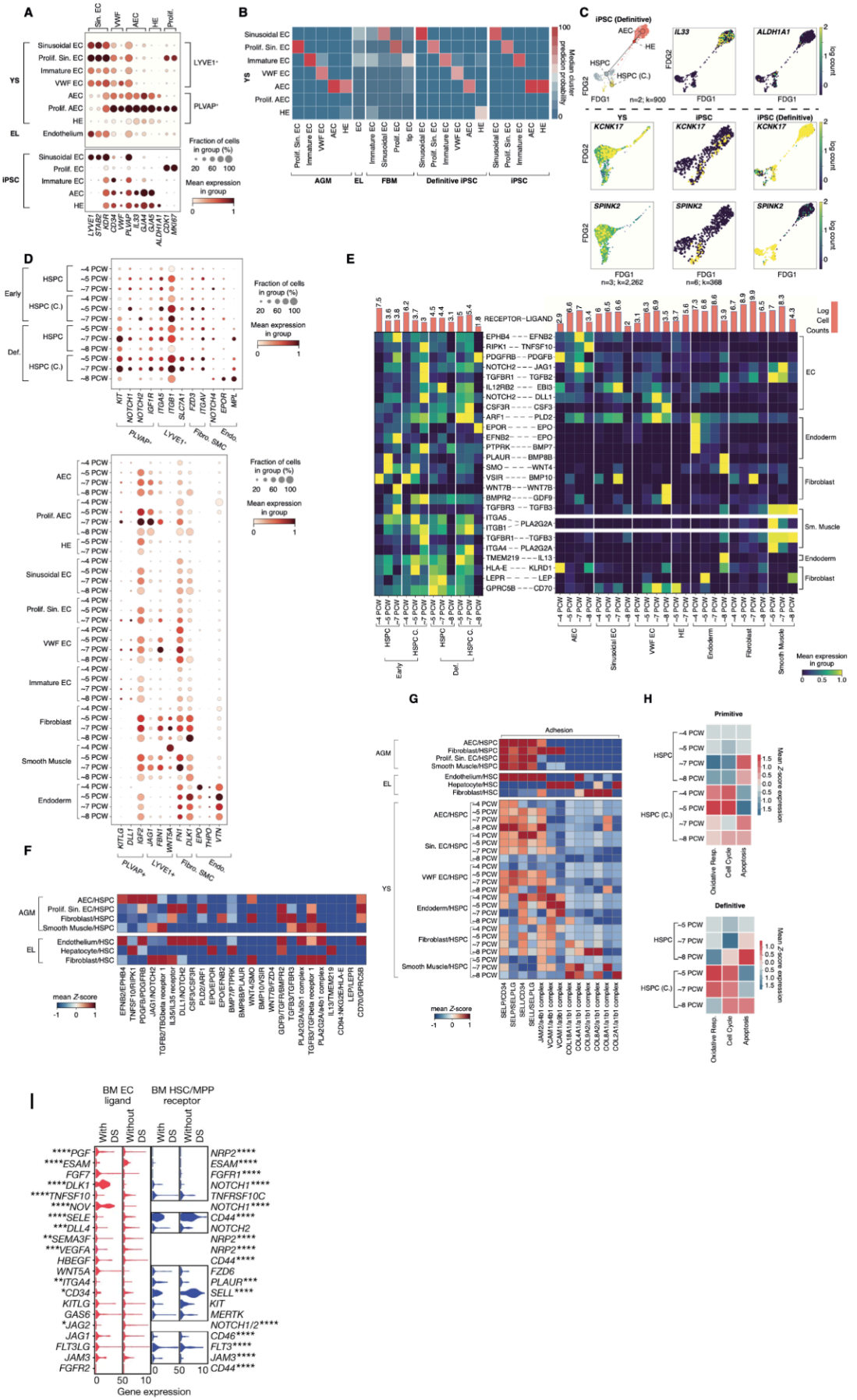


Fig. A7. The lifespan of yolk sac HSPCs. *(A) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of genes distinguishing endothelial cell subsets in YS (main), matched EL scRNA-seq and iPSC (20) scRNA-seq datasets. Data are min-max-standardized with a distribution of 0-1. (B) Heatmap of class prediction probabilities for a logistic regression model (ElasticNet) trained on YS endothelial cell (EC) states (y-axis) projected onto EC states in AGM (32, 76), matched EL, and FBM (35), iPSC (20) and definitive iPSC (12) (x-axis) (vmin=0, vmax=1). Color scale indicates median probabilities. (C) Top: Force directed graph overlaid with partition-based approximate graph abstraction (PAGA) map showing the trajectory of HE transition to HSPC in definitive iPSC scRNA-seq data (12) (n=3, c=2262) with feature plots of key genes (IL33, ALDH1A1) involved in endothelial to hemogenic transition. Bottom: Feature plots of key genes in endothelial to hemogenic transition (SPINK2, KCNK17) in YS, iPSC and definitive iPSCs scRNA-seq trajectories. (D) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of genes predicted by CellphoneDB to form statistically significant (P<0.05) protein-protein interactions between HSPCS (top plot) and stromal cells (bottom plot) arranged by gestational age. Brackets indicate genes which form complexes (data scaled max_value=10). Data are min-max-standardized with a distribution of 0-1. Fibro.: fibroblast, Endo.: endoderm. (E) Heatmaps showing mean (standardized) expression of curated and statistically significant (P<0.05) CellphoneDB putative receptor ligand interactions which change across time. Left: HSPC receptors; right: stromal ligands. Both are grouped by gestational age. Data are min-max-standardized with a distribution of 0-1 (standard-scale='var'). Log normalized cell counts are shown as a barplot above each cell-state column. Sm.: smooth (F) Heatmap showing relative mean expression Z-scores of curated CellphoneDB putative receptor ligand interaction matching S7E between stromal and HSC/HSPC subsets in AGM (top) and liver (bottom). Sin.: sinusoidal (G) Heatmap showing relative mean expression z-scores of curated and statistically significant*

(P<0.05) CellphoneDB putative curated functional adhesion receptor ligand interactions between AGM (top), YS (middle) and EL (bottom) stromal subsets vs HSPC across gestation. (H) Heatmaps showing mean z-scored expression of metabolic (GO-ontology, GO-0045333), cell cycle (GO-ontology, GO:0022402) and apoptosis (GO-ontology GO:0006915) modules for early (left) and definitive (right) YS HSPC and cycling HSPC across gestational age (Module enrichment computed against 200 randomly sampled background genes at 50 bins). Resp.: respiration. (I) Violin plots of gene expression in HSC/MPPs and pooled endothelial cells from FBM scRNA-seq datasets from fetuses with Down syndrome (n = 4; k = 105 HSC/MPPs; k = 111 endothelial cells) and fetuses without Down syndrome (n = 9; k = 92 HSC/MPPs; k = 938 endothelial cells) (3). Genes shown have a significant receptor–ligand interaction in FBM without Down syndrome predicted by CellPhoneDB analysis. Figure A7 A-G were adapted from Goh and Botting et al, 2023 (1), and Figure A7 I, was adapted from Jardine et al, 2021 (3).

Appendix Figure 8

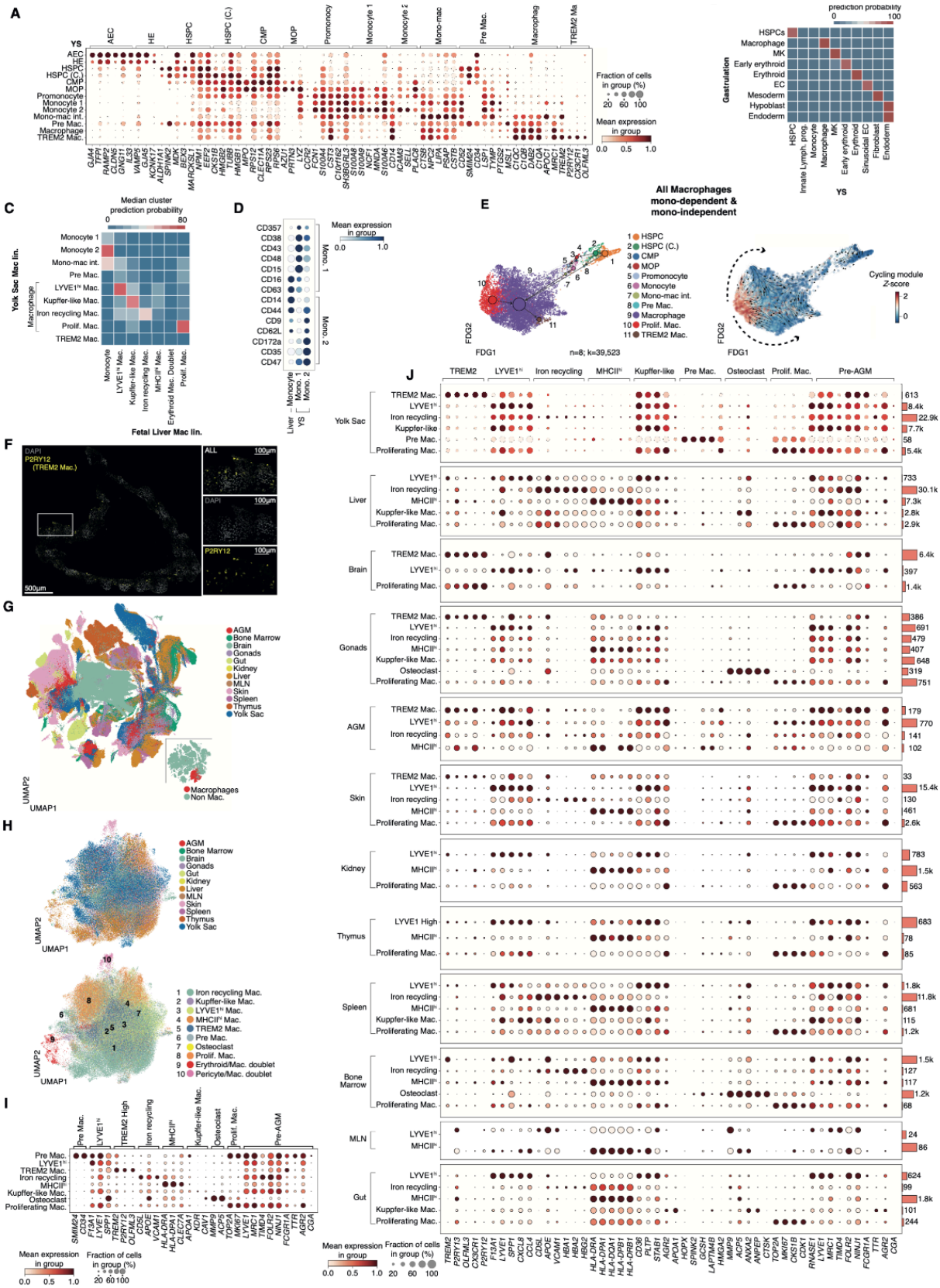


Fig. A8. Macrophage subsets in human yolk sac and prenatal organs. *(A) Dot plot showing*

the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of significant differentially expressed myeloid lineage gene markers (P<0.05) in the YS scRNA-seq. Differential expression was derived via a two-sided Wilcoxon rank-sum test (thresholded at expression in >25% of class, LFC >0.25 and Benjamini–Hochberg corrected P<0.05). Data are min-max-standardized with a distribution of 0-1. (B) Heatmap of class prediction probabilities for a LR model (Elasticnet) trained on YS scRNA-seq cell states (x-axis) and projected onto cell states in human gastrulation scRNA-seq data (390) (y-axis). This LR was performed after reannotating human gastrula data in-house. Color scale indicates median probabilities. (C) Heatmap of class prediction probabilities for a LR model (Elasticnet) trained on expanded YS scRNA-seq Macrophage lineage cell states (y-axis) and projected onto Macrophage lineage cell states in human Fetal liver scRNA-seq data (7-17PCW) (390) (x-axis). Brackets indicate macrophage fractions resolved from the YS macrophages shown in Fig. 5C. Color scale indicates median population probability between 0-80. (D) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each protein (dot size) of proteins in monocytes from EL and matched YS CITEseq datasets. Differentially expressed proteins and proteins matched to RNA markers are shown. (E) Left: Force directed graph (FDG) visualization overlaid with directional partition-based approximate graph abstraction (PAGA) map showing the trajectory of macrophage differentiation is YS scRNAseq data (n=8, c=39,523, CS10-CS23). Right: FDG visualization colored by z-score of enrichment in cycling module (GO:0007049) genes and overlaid with arrows inferred from a CellRank state transition matrix indicating the trend of trajectory. Dashed arrows indicate predicted trajectories of cycling macrophages into macrophage and TREM2 macrophage populations. (Module enrichment computed against 200 randomly sampled background genes at 50 bins). (F) Immunofluorescence images of an 8-PCW YS stained with anti-P2RY12 antibody to demarcate TREM2 macrophages and anti-IBA1 antibody to demarcate macrophages, co-stained with DAPI (n=1). White box indicates ROI

shown. Scale bars: 500 µm and 100 µm (inset). (G) UMAP of the integrated 12-organ fetal atlas (c=3.12×10^6, n=150), colored by organ. Inset indicates the position of macrophages (teal) and non-macrophages (red). Organs include: YS (n=10, c=169,494), AGM (n=4, c=12,248), skin (n=13, c=178,563), brain (n=72, c=2.16×10^6), gonads (n=44, c=14,244), thymus (n=11, c=104,251), gut (n=5, c=79,435), kidneys (n=4, c=26,372), liver (n=14, c=210,549), spleen (n=10, c=127,186), bone marrow (n=8, c=93,677), and MLN (n=2, c=6039). (H) Feature plot showing VAE latent-space derived UMAP representation of macrophages across the integrated 12 organ fetal atlas colored by organ (top) and by annotated heterogenous macrophage substates (bottom). (I) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of marker genes by macrophage subsets across the 12-organ fetal atlas. Data are min-max-standardized with a distribution of 0-1. (J) Dot plot showing the mean expression (color scale) and the fraction of cells expressing each gene (dot size) of marker genes by macrophage subsets across the 12-organ fetal atlas per organ. Data are min-max-standardized with a distribution of 0-1. The cell counts for each macrophage subset are displayed in the bar graphs. Figure adapted from Goh and Botting et al, 2023 (1).

Datasets	Authors	DOI link	Biological replicates and number cells	Data access link	Stage/Age	Sequencing type
Yolk sac (published previously by our group)	Popescu et al, Nature, 2019	https://doi.org/10.1038/s41586-019-1652-y	n=3, c=14,079	https://developmentcellatlas.ncl.ac.uk/datasets/hca_liver/	4-7PCW	10x
Yolk sac external	Calvanese et al Nature, 2022	https://doi.org/10.1038/s41586-022-04571-x	n=2, c=6,256	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162950	4.5PCW	10x
Yolk sac gastrulation	Tyser et al, Nature, 2021	https://doi.org/10.1038/s41586-021-04158-y	n=1, c=1,195	http://www.human-gastrula.net/	CS7	START-seq
AGM external	Calvanese et al Nature, 2022	https://doi.org/10.1038/s41586-022-04571-x	n=4, c=12,248	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162950	4.5 - 6PCW	10x

Fetal bone marrow	Jardine and Webb et al, Nature, 2021	https://doi.org/10.1038/s41586-021-03929-x	n=9, c=103,228	https://fbm.cellatlas.io/	12-17PCW	10x
Fetal kidney	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=7, c=6,847	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	7-12PCW	10x
Fetal brain	Eze et al, Nature Neuroscience, 2021	https://doi.org/10.1038/s41593-020-00794-1	n=10, c=289,000	https://cells-test.gi.ucsc.edu/?ds=early-brain	CS12-22	10x
Fetal skin	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=15, c=185,582	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	7-16PCW	10x
Embryoinc liver	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=3, c=49376	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	CS18 - CS23	10x
Fetal liver (published previously by our group)	Popescu et al, Nature, 2019	https://doi.org/10.1038/s41586-019-1652-y	n=11, c=140,000	https://developmentcellatlas.ncl.ac.uk/datasets/hca_liver/	7-17PCW	10x
Adult liver	Sharma et al, Cell, 2020	https://doi.org/10.1016/j.cell.2020.08.040	n=15, c=31492	https://data.mendeley.com/datasets/6wmzcskt6k/1	48-77 (reduce n=17 to n=15)	10x
Mouse gastrulation	Pijuan-Sala et al, Nature, 2019	https://doi.org/10.1038/s41586-019-0933-9	n=36, c=430,339	https://marionilab.github.io/ExtendedMouseAtlas/#Explore-the-data	E6.5-E9.5	10x
Human iPSC	Alsinet et al, BioRxiv, 2021	https://doi.org/10.1101/2021.11.17.469005	n=19, c=50,512	https://www.hipimmuneatlas.org/	D0-D31+7	10x
Fetal gonads	Garcia-Alonso et al, Nature Portfolio, 2021	https://doi.org/10.21203/rs.3.rs-496470/v1	n=44, c=19,538	In pre-print	6-21PCW	10x
Human iPSC (definitive)	Calvanese et al Nature, 2022	https://doi.org/10.1038/s41586-022-04571-x	n=4, c=12956	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162950	D14	10x
Fetal Thymus	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=11, c=104,251	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	7-17PCW	10x
Fetal Spleen	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=10, c=127,186	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	9-17PCW	10x
Fetal mesenteric lymph node	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=2, c=6,039	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	16-17PCW	10x
Fetal gut	Suo et al, BioRxiv, 2022	https://doi.org/10.1101/2022.01.17.476665	n=5, c=79,435	https://developmentcellatlas.cellgeni.sanger.ac.uk/fetal-immune	12-17PCW	10x
Fetal Brain	http://linnarssonlab.org/	http://linnarssonlab.org/	n=26, c=1,665,937	In pre-print	3-6PCW	10x
Adult liver	Brazovskja et al BioRxiv, 2022	DOI: 10.1101/2021.06.03.444016	n=3, c=924	In pre-print	Adult	10x

Adult bone marrow	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=10, c=9408	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult gut	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=13, c=1401	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult liver	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=7, c=2728	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult lung	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=10, c=25433	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult lymph node	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=12, c=1380	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult muscle	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=5, c=214	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult spleen	Domi’nguez Conde et al Science, 2022	DOI: 10.1126/science.abl51	n=12, c=8060	https://www.tissueimmunecellatlas.org/#datasets	Adult	10x
Adult gut	Elemntaite et al Nature, 2021	DOI: 10.1038/s41586-021-03852-1	n=3, c=380	https://www.ebi.ac.uk/biostudies/arrayexpress	Adult	10x
Adult uterus	Garcia-Alonso et al Nature Genetics, 2021	DOI: 10.1038/s41588-021-00972-2	n=5, c=458	https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10287	Adult	10x
Adult heart	Litvi–ukov‡ et al Nature, 2020	DOI: 10.1038/s41586-020-2797-4	n=6, c=474	https://www.heartcellatlas.org/#DataSources	Adult	10x
Adult lung	Madissoon et al BioRxiv, 2021	DOI: 10.1101/2021.11.26.470108	n=7, c=13200	https://5locationslung.cellgeni.sanger.ac.uk	Adult	10x
Adult trachea	Madissoon et al BioRxiv, 2021	DOI: 10.1101/2021.11.26.470108	n=, c=679	https://5locationslung.cellgeni.sanger.ac.uk	Adult	10x
Adult skin	Reynolds et al Science, 2021	DOI: 10.1126/science.aba6500	n=5, c=262	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8142	Adult	10x
Adult kidney	Stewart et al, Science, 2019	DOI: 10.1126/science.aat503	n=8, c=357	https://www.kidneycellatlas.org	Adult	10x
Adult badder	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=3, c=2831	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult fat	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=1667	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult heart	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=1, c=12	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult liver	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=1171	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult lung	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=3, c=8272	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x

Adult lymph node	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=3, c=1127	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult muscle	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=3, c=822	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult pancreas	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=1030	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult prostate	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=356	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult skin	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=826	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult spleen	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=3, c=5583	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult thymus	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=505	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult trachea	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=697	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult uterus	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=1, c=174	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult vasculature	The Tabula Sapiens Consortium, Science, 2022	DOI: 10.1126/science.abl489	n=2, c=1602	https://tabula-sapiens-portal.ds.czbiohub.org	Adult	10x
Adult brain	Yang et al, BioRxiv, 2021	DOI: 10.1101/2021.04.26.441262	n=7, c=832	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163577	Adult	10x

Fig. A9. External datasets table *A table containing all external datasets collected and used throughout this study.*

CITE-seq antibody details				
ADT_name	**Clone**	**Specificity**	**Barcode**	**DNA Barcode**
ADT_A0145	Ber-ACT8	CD103	145	GACCTCATTGTGAAT
ADT_A0095	RTK4530	IgG2b, k Isotype Ctrl	95	GATTCTTGACGACCT
ADT_A0167	E11	CD35	167	ACTTCCGTCGATCTT
ADT_A0215	11C1	CD268 (BAFF-R, BAFFR)	215	CGAAGTCGATCCGTA
ADT_A0085	BC96	CD25	85	TTTGTCCTGTACGCC
ADT_A0161	ICRF44	CD11b	161	GACAAGTGATCTGCA
ADT_A0147	DREG-56	CD62L	147	GTCCCTGCAACTTGA
ADT_A0007	29E.2A3	CD274 (B7-H1, PD-L1)	7	GTTGTCCGACAATAC

ADT_A0087	UCHL1	CD45RO	87	CTCCGAATCATGTTG
ADT_A0159	L243	HLA-DR	159	AATAGCGAGCAAGTA
ADT_A0149	HP-3G10	CD161	149	GTACGCAGTCCTTCT
ADT_A0080	RPA-T8	CD8a	80	GCTGCGCTTTCCATT
ADT_A0050	HIB19	CD19	50	CTGGGCAATTACTCG
ADT_A0140	G025H7	CD183	140	GCGATGGTAGATTAT
ADT_A0352	AER-37 (CRA-1)	Fc epsilon RI alpha	352	CTCGTTTCCGTATCG
ADT_A0148	G043H7	CD197 (CCR7)	148	AGTTCAGTCAACCGA
ADT_A0083	3G8	CD16	83	AAGTTCACTCTTTGC
ADT_A0375	M1310G05	IgG Fc	375	CTGGAGCGATTAGAA
ADT_A0070	GoH3	CD49f	70	TTCCGAGGATGATCT
ADT_A0423	590H11G1E3	MERTK	423	TCCTGCATGTACCCA
ADT_A0100	2H7	CD20	100	TTCTGGGTCCCTAGA
ADT_A0088	EH12.2H7	CD279 (PD-1)	88	ACAGCGCCGTATTTA
ADT_A0132	AY13	EGFR	132	GCTTAACATTGGCAC
ADT_A0063	HI100	CD45RA	63	TCAATCCTTCCGCTT
ADT_A0236	RTK2071	IgG1, k Isotype Ctrl	236	ATCAGATGCCCTCAT
ADT_A0143	G034E3	CD196	143	GATCCCTTTGTCACT
ADT_A0126	clone 7	CD133	126	TGGTAACGACAGTCC
ADT_A0188	ASL-32	CD66a/c/e	188	GGGACAGTTCGTTTC
ADT_A0367	TS1/8	CD2	367	TACGATTTGTCAGGG
ADT_A0240	RTK4174	IgG2c, k Isotype Ctrl	240	TCCAGGCTAGTCATT
ADT_A0101	9E2	CD335	101	ACAATTTGAACAGCG
ADT_A0205	15-2	CD206 (MMR)	205	TCAGAACGTCTAACT
ADT_A0123	9C4	CD326 (Ep-CAM)	123	TTCCGAGCAAGTATC
ADT_A0206	7-239	CD169	206	TACTCAGCGTGTTTG
ADT_A0353	HIP8	CD41	353	ACGTTGTGGCCTTGT
ADT_A0371	P1E6-C5	CD49b	371	GCTTTCTTCAGTATG
ADT_A0370	201A	CD303	370	GAGATGTCCGAATTT
ADT_A0369	TS2/16	CD29	369	GTATTCCCTCAGTCA
ADT_A0366	12G5	CD184 (CXCR4)	366	TCAGGTCCTTTCAAC
ADT_A0374	MEM-108	CD98	374	GCACCAACAGCCATT
ADT_A0363	G077F6	CD124	363	CCGTCCTGATAGATG
ADT_A0199	7C9	Siglec-8	199	CTTCTCCTCAGCAAT
ADT_A0362	7D4-6	CD309/VEGFR2	362	TTCACGCAGTAAGAT
ADT_A0357	CD43-10G7	CD43	357	GATTAACCAGCTCAT
ADT_A0351	BV10A4H2	CD135 (Flt-3/Flk-2)	351	CAGTAGATGGAGCAT
ADT_A0207	8F9	CD370 (CLEC9A/DNGR1)	207	CTGCATTTCAGTAAG
ADT_A0196	HIR2	CD235ab	196	GCTCCTTTACACGTA
ADT_A0181	Bu32	CD21	181	AACCTAGTAGTTCGG
ADT_A0176	A1	CD39	176	TTACCTGGTATCCGT
ADT_A0163	M80	CD141	163	GGATAACCGCGCTTT

ADT_A0162	10.1	CD64	162	AAGTATGCCCTACGA
ADT_A0359	HB15e	CD83	359	CCACTCATTTCCGGT
ADT_A0372	VI-PL2	CD61	372	AGGTTGGAGTAGACT
ADT_A0358	GHI/61	CD163	358	GCTTCTCCTTCCTTA
ADT_A0360	108-17	CD357 (GITR)	360	ACCTTTCGACACTCG
ADT_A0373	5A6	CD81	373	GTATCCTTCCTTGGC
ADT_A0169	F38-2E2	CD366 (Tim-3)	169	TGTCCTACCCAACTT
ADT_A0168	QA17A04	CD57	168	AACTCCCTATGGAGG
ADT_A0171	C398.4A	CD278 (ICOS)	171	CGCGCACCCATTAAA
ADT_A0156	DX2	CD95	156	CCAGCTCATTAGAGC
ADT_A0146	FN50	CD69	146	GTCTCTTGGCTTAAA
ADT_A0005	2D10	CD80	5	ACGAATCAATCTGTG
ADT_A0055	MI15	CD138	55	ACTCTTTCGTTTACG
ADT_A0027	113-16	CD70	27	CGCGAACATAAGAAG
ADT_A0069	RCR-401	CD201 (EPCR)	69	GTTTCCTTGACCAAG
ADT_A0136	MHM-88	IgM	136	TAGCGAGCCCGTATA
ADT_A0361	p282 (H19)	CD59	361	AATTAGCCGTCGAGA
ADT_A0071	L291H4	CD194	71	AGCTTACCTGCACGA
ADT_A0034	UCHT1	CD3	34	CTCATTGTAACTCCT
ADT_A0046	SK1	CD8	46	GCGCAACTTGATGAT
ADT_A0165	1D11	CD314 (NKG2D)	165	CGTGTTTGTTCCTCA
ADT_A0006	IT2.2	CD86	6	GTCTTTGTCAGTGCA
ADT_A0141	J418F1	CD195	141	CCAAAGTAAGAGCCA
ADT_A0160	L161	CD1c	160	GAGCTACTTCACTCG
ADT_A0386	CD28.2	CD28	386	TGAGAACGACCCTAA
ADT_A0387	1D3	TSLPR (TSLP-R)	387	CAGTCCTCTCTGTCA
ADT_A0023	SKII.4	CD155 (PVR)	23	ATCACATCGTTGCCA
ADT_A0390	A019D5	CD127 (IL-7Ra)	390	GTGTGTTGTCCTATG
ADT_A0066	CD7-6B7	CD7	66	TGGATTCCCGGACTT
ADT_A0061	104D2	CD117	61	AGACTAATAGCTGAC
ADT_A0062	HI10a	CD10	62	CAGCCATTCATTAGG
ADT_A0052	P67.6	CD33	52	TAACTCAGGGCCTAT
ADT_A0029	BJ40	CD48	29	CTACGACGTAGAAGA
ADT_A0024	TX31	CD112 (Nectin-2)	24	AACCTTCCGTCTAAG
ADT_A0073	IM7	CD44	73	TGGCTTCAGGTCCTA
ADT_A0033	HI186	CD52	33	CTTTGTACGAGCAAA
ADT_A0394	CY1G4	CD71	394	CCGTGTTCCTCATTA
ADT_A0393	S-HCL-1	CD22	393	GGGTTGTTGTCTTTG
ADT_A0392	W6D3	CD15	392	TCACCAGTACCTAGT
ADT_A0124	WM59	CD31	124	ACCTTTATGCCACGG
ADT_A0081	M5E2	CD14	81	TCTCAGACCTCCGTA
ADT_A0089	A15153G	TIGIT	89	TTGCTTACCGCCAGA

ADT_A0102	BM16	CD294	102	TGTTTACGAGAGCCC
ADT_A0048	2D1	CD45	48	TCCCTTGCGATTTAC
ADT_A0072	RPA-T4	CD4	72	TGTTCCCGCTCAACT
ADT_A0054	581	CD34	54	GCAGAAATCTCCCTT
ADT_A0189	C1.7	CD244 (2B4)	189	TCGCTTGGATGGTAG
ADT_A0047	5.1H11	CD56	47	TCCTTTCCTGATAGG
ADT_A0144	J252D4	CD185	144	AATTCAACCGTCGCC
ADT_A0385	TS1/18	CD18	385	TATTGGGACACTTCT
ADT_A0214	FIB504	integrin b7	214	TCCTTGGATGTACCG
ADT_A0219	GIR-208	CD119 (IFN-g R a chain)	219	TGTGTATTCCCTTGT
ADT_A0131	16G5	Cadherin 11	131	CGTTGCCATTAACCA
ADT_A0180	ML5	CD24	180	AGATTCCTTCGTGTT
ADT_A0398	9-4D2-1E4	CD115	398	AATCACGGTCCTTGT
ADT_A0135	67A4	CD324 (E-Cadherin)	135	ATCCTTCTCCCTTTC
ADT_A0164	51.1	CD1d	164	TCGAGTCGCTTATCA
ADT_A0242	K036C2	CD192	242	GAGTTCCCTTACCTG
ADT_A0172	9F.8A4	CD275 (B7-H2, B7-RP1, ICOSL)	172	GTTAGTGTTAGCTTG
ADT_A0068	43A3	CD105	68	ATCGTCGAGAGCTAG
ADT_A0400	BV9	CD144	400	TCCACTCATTCTGTA
ADT_A0397	5E8	CD193	397	ACCAATCCTTTCGTC
ADT_A0383	JS11	CD55	383	GCTCATTACCCATTA
ADT_A0187	CB3-1	CD79b	187	ATTCTTCAACCGAAG
ADT_A0166	6/40c	CD66b	166	AGCTGTAAGTTTCGG
ADT_A0064	6H6	CD123	64	CTTCACTCTGTCAGG
ADT_A0053	S-HCL-3	CD11c	53	TACGCCTATAACTTG
ADT_A0218	AK4	CD62P (P-Selectin)	218	CCTTCCGTATCCCTT
ADT_A0216	HIP1	CD42b	216	TCCTAGTACCGAAGT
ADT_A0217	HA58	CD54	217	CTGATAGACTTGAGT
ADT_A0428	9F4	Tim-4	428	CGTCATATAGTATGG
ADT_A0427	94b/FOLR2	Folate Receptor beta	427	TGTGGCTAGTCAGTT
ADT_A0382	MEM-166	CD177	382	AGTATGGAGCCATAT
ADT_A0245	STA	CD106	245	TCACAGTTCCTTGGA
ADT_A0233	MHN3-21	Notch 3	233	CTATTGGACGTATCT
ADT_A0401	H037G3	CD301	401	ACCTAGAAATCAGCA
ADT_A0404	H5C6	CD63	404	GAGATGTCTGCAACT
ADT_A0248	HAE-1f	CD62E	248	CTCCCTGTGGCTTAA
ADT_A0384	IA6-2	IgD	384	CAGTCTCCGTAGAGT
ADT_A0419	3F3	CD72	419	CAGTCGTGGTAGATA
ADT_A0446	VIMD2	CD93	446	GCGCTACTTCCTTGA
ADT_A0409	17G10.2	CD85g (ILT7)	409	TGTCAGTTCCTATGA
ADT_A0408	15-414	CD172a	408	CGTGTTTAACTTGAG
ADT_A0405	HTA125	CD284 (TLR4)	405	GCTTAGCTGTATCCG

ADT_A0406	12C2	CD304	406	GGACTAAGTTTCGTT
ADT_A0407	5-271	CD36	407	TTCTTTGCCTTGCCA
ADT_A0244	CBR-IC2/2	CD102	244	TGACCTTCCTCTCCT
ADT_A0224	IP26	TCR alpha/beta	224	CGTAACGTAGAGCGA
ADT_A0246	TU27	CD122	246	TCATTTCCTCCGATT
ADT_A0237	G0114F7	IgG1, λ Isotype Ctrl	237	GGGAGCGATTCAACT
ADT_A0247	1A1	CD267 (TACI)	247	AGTGATGGAGCGAAC
ADT_A0155	H4A3	CD107a	155	CAGCCCACTGCAATA
ADT_A0134	P1H12	CD146	134	CCTTGGATAACATCA
ADT_A0213	MHN1-519	Notch 1	213	AATCTGTAGTGCGTT
ADT_A0575	TS2/7	CD49a	575	ACTGATGGACTCAGA
ADT_A0579	HI9a	CD9	579	GAGTCACCAATCTGC
ADT_A0581	3C10	TCR V alpha7.2	581	TACGAGCAGTATTCA
ADT_A0583	B3	TCR Vg9	583	AAGTGATGGTATCTG
ADT_A0576	9F10	CD49d	576	CCATTCAACTTCCGG
ADT_A0572	1D9-M12	C5L2	572	ACAATTTGTCTGCGA
ADT_A0574	HI264	CD235a	574	AGAGTATGTATGGGA
ADT_A0569	5D3	CD338 (ABCG2)	569	TAAGACTTGGCCGTC
ADT_A0577	AD2	CD73	577	CAGTTCCTCAGTTCG
ADT_A0578	HM47	CD79a	578	CTTATCACCCGCTTT
ADT_A0598	S16017E	CD110	598	TGTTGTAAGATGCCA
ADT_A0597	9E9A8	CD209 (DC-SIGN)	597	TCACTGGACACTTAA
ADT_A0410	HB-7	CD38	410	CCTATTCCGATTCCG
ADT_A0447	OX-104	CD200	447	CACGTAGACCTTTGC
ADT_A0592	DX27	CD158b (KIR2DL2/L3, NKAT2)	592	GACCCGTAGTTTGAT
ADT_A0805	TX25	CD226 (DNAM-1)	805	AGACCAACTCATTCA
ADT_A0804	K041E5	CD186	804	GACAGTCGATGCAAC
ADT_A0600	UP-R1	CD158f	600	AAAGTGATGCCACTG
ADT_A0599	DX9	CD158e1 (KIR3DL1, NKB1)	599	GGACGCTTTCCTTGA
ADT_A0364	WM15	CD13	364	TTTCAACGCCCTTTC
ADT_A0586	TREM-26	CD354	586	TAGCCGTTTCCTTTG
ADT_A0582	B6	TCR Vd2	582	TCAGTCAGATGGTAT
ADT_A0588	33.1 (Ab33)	CD202b (Tie2/Tek)	588	CGATCCCTTACCTAT
ADT_A0175	NK92.39	CD96	175	TGGCCTATAAATGGT
ADT_A0420	HP-MA4	CD158	420	TATCAACCAACGCTT
ADT_A0801	P30-15	CD337 (NKp30)	801	AAAGTCACTCTGCCG
ADT_A0820	2E1B02	CD130	820	CACGAGAATTTCAGT
ADT_A0822	NY2	CD142	822	CACTGCCGTCGATTA
ADT_A0829	509f6	CD307	829	TCACGCAGTCCTCAA
ADT_A0830	162.1	CD319 (CRACC)	830	AGTATGCCATGTCTT
ADT_A0821	67D2	CD164	821	GAGGCACTTAACATA
ADT_A0814	HD30	CD205 (DEC-205)	814	CTATCGTTTGATGCA

ADT_A0026	CC2C6	CD47	26	GCATTCTGTCACCTA
ADT_A0590	NKTA255	CD305	590	ATTTCCATTCCCTGT
ADT_A0433	8C11	CD325	433	CCTTCCCTTTCCTCT
ADT_A0819	UV4	CD126	819	TGATGGGAGCTTATC
ADT_A0817	W7C5	CD109	817	CACTTAACTCTGGGT
ADT_A0591	15C4	LOX-1	591	ACCCTTTACCGAATA
ADT_A0853	50C1	CD371 (CLEC12A)	853	CATTAGAGTCTGCCA
ADT_A0845	3B2/TA8	CD99	845	ACCCGTCCCTAAGAA
ADT_A0060	5E10	CD90 (Thy1)	60	GCATTGTACGATTCA
ADT_A0844	MEM-55	CD45RB	844	AGATGGGACTCACCA
ADT_A0858	TRA-2-10	CD46	858	ACAGTACGACCTTCT
ADT_A0866	AYP1	CLEC1B	866	TGCCAGTATCACGTA
ADT_A0008	24F.10C12	CD273	8	TCAACGCTTGGCTAG
ADT_A0170	MIH26	BTLA	170	GTTATTGGACTAAGG
ADT_A0138	UCHT2	CD5	138	CATTAACGGGATGCC
ADT_A0154	O323	CD27	154	GCACTCCTGCATGTA
ADT_A0153	SA231A2	KLRG1	153	CTTATTTCCTGCCCT
ADT_A0931	1C1	CD131	931	CTGCATGAGACCAAA
ADT_A0920	ASL-24	CD82	920	TCCCACTTCCGCTTT
ADT_A0934	HSL96	CD179a	934	TAGATGGGATTCCGG
ADT_A0935	LN2	CD74	935	CTGTAGCATTTCCCT
ADT_A0932	31G4D8	Lymphotoxin b Receptor (LT-bR)	932	CCTCTATTCAGAGCA

Fig. A10. CITE-seq antibody details *A table containing all Abs used in the Cite-seq experiments carried out in this study corresponding to the TotalseqA panel (Human Universal Cocktail; V1; Biolegend)*