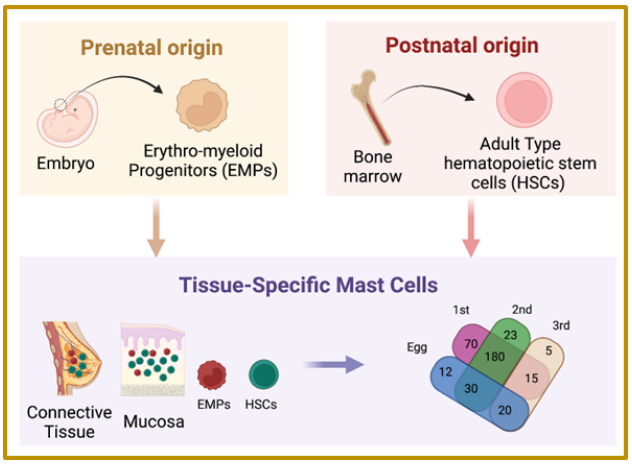
Exploring Myeloid Stem Cell Differentiation Pathways and Gene Expression Across Adult and Human Embryonic Yolk Sac Datasets

Ashley Bielawski[1], Chase Lindeboom[3], Christian Rizza[4], Mariana Sierra[2], and Paula Wu[1]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan
[2]Department of Neurology, University of Michigan
[3]Department of Biological Chemistry, University of Michigan
[4]Department of Pathology, University of Michigan

1 Abstract

Mast cells are myeloid immune cells made by hematopoiesis, the process by which the body generates new blood cells from stem cells, and are found in connective tissue and mucosa, playing a key role in innate immunity St. John et al. (2023). They are known to arise from two distinct, lineage-specific progenitor populations - embryonic and definitive, which may have different transcriptional programs that could determine the cell fate. In this project, we first conducted clustering, differential gene expression and functional enrichment analysis to explore the differentially expressed genes related to transcriptional control. Initial clustering results entailed the heterogeneity in mast cell phenotypes could be attributed to differences in the developmental origins. GSEA results revealed enrichments in STAT5A-related hematopoiesis markers in Cluster 7, transcriptional regulator ATF3 that promotes mast cell maturation in Cluster 9, and developmental transcription factors DIDO1/SUPTH20 in yolk sac clusters. Furthermore, utilizing Monocle 3, we were able to determine mast cell progenitor differentiation trajectories. We found that Monocle 3 was suitable for generating trajectory pathways to show start and end state for hematopoiesis with mast cell and mast cell progenitors, but is limited due to its inability to generate cell fate at a single-cell resolution. Also, looking at expression of DIDO1 in the yolk sac dataset showed no correlation based on pseudotime and appears to be present at the beginning and end stage of development.



2 Introduction

Mast cells are myeloid, tissue-resident immune cells distributed throughout the body and most prominently found in connective tissue or mucosa (Kolkhir et al. (2022)). They play an important role in innate immunity; when encountering foreign antigens, mast cells will release mediators including histamine that invoke an inflammatory response (Gilfillan et al. (2011)). This typically provides the first line of defense against the potential pathogen invader. While mast cells are vital for healthy immunity, they are also thought to play key roles in a variety of diseases and disorders such as mastocytosis (Valent et al. (2017)), mast cell activation syndrome, arthritis (Crisp et al. (1984)), and allergic reactions including anaphylaxis (Galli and Tsai (2012)).

Mast cells, like other myeloid cells, arise through hematopoeisis, which is the generation of white and red blood cells (Jagannathan-Bogdan and Zon (2013)). Hematopoeisis occurs in waves over the course of vertebrate development. During embryogenesis, "primitive" erythroid progenitor populations first emerge from the embryonic yolk sac and give rise to primitive erythrocytes. This is shortly followed by the production of yolk sac-derived erythro-myeloid progenitor cells (EMPs) which historically were thought to primarily give rise to erythrocytes, megakaryocytes, and macrophages (Moore and Metcalf (1970); Frame et al. (2013)). Finally, during definitive hematopoiesis, adult hematopoietic stem cells (HSCs) emerge from the aorta-gonad-mesonephros (Boisset et al. (2010)). These HSCs are self-renewing and capable of differentiating into all blood cellular components, including mast cells (Lee and Hong (2020)). As the embryo develops, the HSCs will eventually move to the bone marrow where they will remain and repopulate blood cell populations throughout the remainder of the vertebrate's life.

In mice, it has been observed that initially during embryogenesis, the population of mast-cells originates from yolk sac erythro-myeloid progenitors (Sonoda et al. (1983), Gentek et al. (2018)). When definitive hematopoiesis begins, the yolk-sac derived mast cells are gradually diluted out by mast cells originating from HSCs (Gentek et al. (2018)). While both EMPs, and HSCs are capable of giving rise to mast cells, they are also both capable of differentiating into other types of cells. EMPs can differentiate into erythroid cells, macrophages, or megakaryocytes (Frame et al. (2013)), while HSCs can become any blood cell through a series of hierarchical differentiations (Chotinantakul and Leeanansaksiri (2012)). Deter-

mination of the ultimate cellular fate of these progenitor cells is thought to be regulated by signaling pathways that adjust the transcriptional profiles of the progenitor cells (Lin et al. (2015)). While both EMPs and HSCs can potentially differentiate into mast cells, it is unknown whether both progenitors rely on the same transcriptional programs to determine an ultimate mast cell fate.

In our study, we took scRNA data from both human embryonic yolk sac data Goh et al. (2023)) and human adult tissue data (Tauber et al. (2023)) and subset the data for hematopoietic stem and progenitor cells, common myeloid progenitors, EO/baso/mast precursors, and mast cells. We then loaded them into Seurat objects, performed quality control, and pre-processed the data including normalization and setting covariates. We then integrated the Seurat objects together and clustered the cells from the integrated Seurat object. Finally, we performed gene-set enrichment analysis to identify putative transcriptional signatures corresponding to embryonic mast cells as well as pseudo time analysis to determine mast cell progenitor differentiation trajectories.

3 Methods

An overview of the workflow is illustrated in Figure 1. Please see our up-to-date codes here: Github link

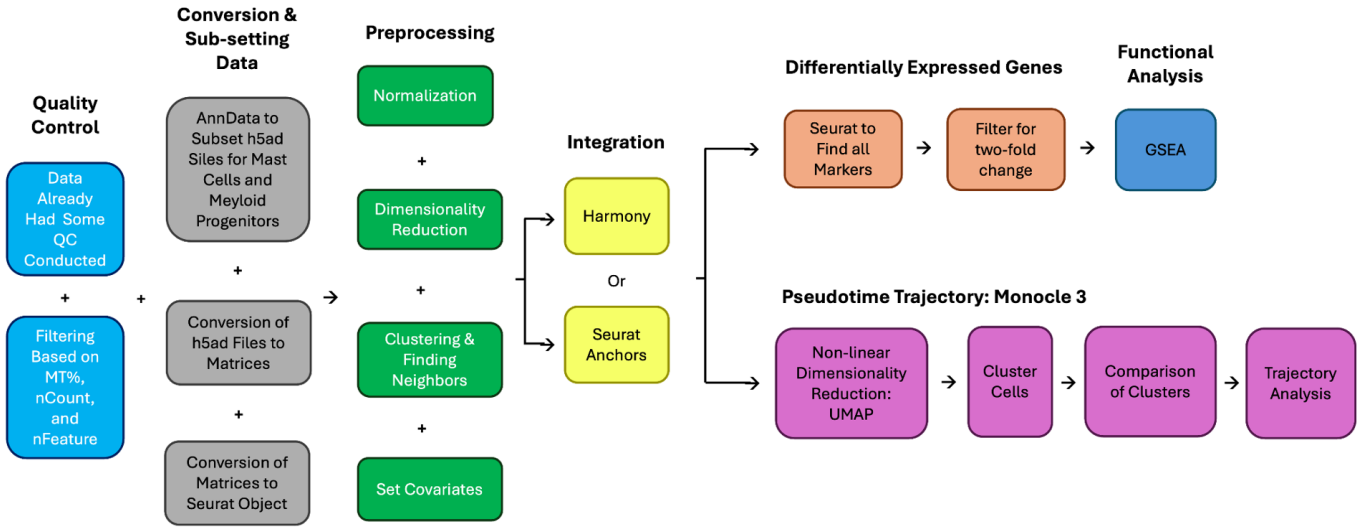


Figure 1: Analysis Pipeline Overview

3.1 Data Acquisition & Preparation

To explore the research questions, we acquired both the yolk sac and the adult data by directly downloading the files from corresponding websites (Tauber et al. (2023); Goh et al. (2023)). Both datasets were stored in the h5ad file format. However, attempting to directly load these files into the R Studio environment not only resulted in session crashes due to

their exceptionally large size, but also created conflicts in environmental variables. Despite several technical difficulties, we managed to first subset the mast cell data from the original datasets and then convert into Seurat objects and bypassed some constraints and obstacles. These steps were done in Python.

To accomplish this, we read the file using the AnnData package and subset the data set using Panda based on predetermined criteria. Once the subsets were created, we saved it as separate h5ad files and used another pipeline to turn them into matrices, features, and barcodes that R can read and process (Sanbomics (2023)). Later in R, we constructed the Seurat objects (Satija et al. (2015)) using the aforementioned matrices and features extracted from the h5ad files.

During our preliminary data exploration and quality control assessments, we noticed that the gene identifiers in the adult dataset remained in their original Ensembl ID ("ENSG") format. For better downstream interpretability and analysis, we proactively converted these identifiers to their corresponding gene names using the AnnotationDbi and org.Hs.eg.db library. During this process, we identified a subset of genes that were not assignable to recognized gene names. We opted to exclude these non-mappable entries. For the duplicate genes, we annotated them by appending "_dup" followed by a sequentially incrementing numeral to distinguish them unambiguously within our dataset. The yolk sac dataset did not undergo this process since the gene names have already been annotated.

3.2 Differential Gene Expression & Functional Annotation

3.2.1 Data Preprocessing & Quality Control

% Mitochondrial DNA content was calculated as a function of total reads for each dataset in Seurat to check for cell viability and technical artifacts using the "PercentageFeatureSet" function. Additionally, sample complexity was determined by taking the log10 ratio of features over total RNA counts. Cells that contained between 500 to 7000, at least 500 total transcripts, and less than 2% mitochondrial content were included in our aggregated dataset for downstream analysis. Samples containing less than 10 cells were excluded from each dataset.

3.2.2 Normalization, Batch correction, and Clustering

Metadata from the adult dataset contained a developmental stage corresponding to age. Similarly, yolk sac data consisted of cells originating from four developmental stages. Both datasets also contained metadata describing anatomical origins of each cell comprising the samples referred to as "component" for the yolk sac and "tissue_in_publication" for adult tissues. Both datasets also were annotated with mitochondrial DNA content and "sex" for each sample. All 4 of these categories: component/tissue, developmental stage, sex, and mitochondrial DNA content were set as covariates during dataset normalization with the SCTransform function. PCA was performed on the normalized counts from SCTransformed data. The PCA reduction was then used in the runUMAP and for clustering via FindNeighbors() and dataset integration performed in harmony.

A Seurat object was then passed to the SCTransform function for normalization with covariates set as developmental stage, sex, tissue, and percent mitochondrial DNA content.

PCA was performed for n = 30 PCs using the RunPCA() function in Seurat. We used the dimensionality reduction calculated by RunPCA() to find clusters via the FindNeighbors() and RunUMAP function. The UMAP reduction was then integrated by using the Harmony Package setting Tau = 300 based on the assumption that cells from the embryonic dataset should cluster together. FindClusters was then performed on the integrated dataset to visualize and compare results of integration.

3.2.3 Differential Gene Expression and Functional Analysis

Identification of differentially expressed genes (DEGs) was done using the Seurat function FindAllMarkers() to find differences between Harmony clusters. We filtered for p-values < 0.05, and further filtered by log fold change of >1.0 or <1.0 for upregulated and downregulated genes, respectively (See Figure S3). We then took the top DEGs (up to 500 for each upregulated and downregulated), and performed functional enrichment analysis using GSEA (Subramanian et al. (2005); Liberzon et al. (2011)) analyzing for both gene ontology and transcription factor targets individually, taking the top 10 of both with an FDR < 0.05.

3.3 Pseudotime Trajectory Analysis

3.3.1 Data Acquisition & Preparation

The same packages and pipeline on Python was used as described above, only four more cell types were added to the data. In order to look at cell fate of mast cells, we accepted hematopoietic stem and progenitor cells (HSPCs), common myeloid progenitors (CMP), EO/baso/mast precursors (EBP), and mast cells.

3.3.2 Data Preprocessing & Quality Control

For pseudotime analysis, since only yolk sac data was used we accepted cells with under 10% mitochondrial content, genes with total transcript counts of at least 800, and feature counts of at least 500 to filter out low-quality reads.

3.3.3 Dimensionality Reduction & Clustering

Before looking for cluster structures we used FindNeighbors and FindClusters functions (resolution = .9) before remapping using Uniform Manifold Approximation and Projection (UMAP) (dims = 1:30, n.neighbors = 50).

3.3.4 Integration

For the dataset taken to Monocle 3 for trajectory analysis we used both Seurat Anchors based on Canonical Correlation Analysis (CCA) and Harmony which batch relies on shared nearest neighbor (SNN) graphs to integrate and correct for batch effect based on fetal ID/Component (based on testing of three different integration metrics, see Figure S9). We used a data set for the yolk sac sample which included hematopoietic stem and progenitor cells (HSPCs), other myeloid progenitors, and mast cells. For the yolk sac dataset run through Seurat we

needed to remove one sample that had under 30 cells and set k.weight to 50 (based on testing of three different k.weights 20,50, and 70, see Figure S6).

3.3.5 Pseudotime Trajectories

Pseudotemporal ordering of hematopoietic stem cells, myeloid progenitors, and mast cells was done with Monocle 3, which employs a technique called reverse graph embedding to infer cell fate. Each cell was assigned a pseudotime value based on its position along the trajectory (we selected the primitive mast cell cluster as the start of the trajectory to guide Monocle 3 through our biological knowledge of hematopoiesis). Our integrated seurat object was first converted to a celldataset object that could be used on Monocle 3. Then, using UMAP reduced dimension we were able to construct a pseudotime trajectory of our cells. Then, we looked at differentially expressed genes that change as a result of pseudotime to see if there were any genes that may be significant for early hematopoiesis.

4 Results

4.1 Summary Statistics of the Data

	Yolk Sac	**Adult**
Age Range	4-8 PCW	33-69 years-old
Number of Samples	10 Embryos	9 human donors; 24 organs
Number of Cells	354 MC, 762 MEMP, 25 CMP,	2238 MC
	3140 HSPC, 56 EBP	

*MC = Mast Cell, MEMP = Megakaryocyte-Erythroid-Mast Cell Progenitor, CMP = Common Myeloid Progenitors

HSPC = Hematopoietic Stem Cell, EBP = EO/Baso/Mast Precursors

Table 1: Summary of Data

Summary statistics for individual data sets is shown in Table 1. Both yolk sac and adult datasets are scRNA-seq collected by 10x Genomics using both 3' and 5' assays. More specifically, the embryonic cells dataset contains human yolk sac data points collected from 10 embryos at 4-8 post conception weeks. In the subset of our interest, it contains 354 mast cells, 762 megakaryocyte-erythroid-mast cell progenitors, 3140 hematopoietic stem cells, 25 common myeloid progenitors, and 56 EO/baso/mast precursors. Similarly, the adult dataset, after subsetting, comprises 2238 mast cells sourced from 24 organs across 9 donors ages 33-69 years-old.

4.2 Data Integration

After loading data into Seurat quality control, subsetting, and preprocessing were performed on each dataset (Figure 1). To better visualize total RNA content and library complexity for each dataset we began by plotting histograms of log10 Genes per UMI calculated by (# of Features / # of UMIs per cell). Doing so revealed that there was a bimodal distribution in the dataset derived from adult tissues (Figure S1). Mitochondrial DNA content was also visualized given a threshold value of $< 2\%$ (Figure S2) Plotting features vs. counts in a scatter plot showed that unprocessed counts from the adult dataset came from two separate sequencing techniques. 10x 3' v3 and 10x 5' v3.

This revealed that the 10x 5' v3 experiment performed in the adult dataset included much longer features which is characteristic of this particular sequencing type. Reads from the embryonic dataset were much closer in depth to the 3' experiment performed (Figure S2a). Based on these characteristics, an upper threshold was set at an RNA count of 7000 maximum to simplify downstream integration of the two datasets. Additionally, mitochondrial DNA was given a threshold value of $< 2\%$ and replotted in a features vs. counts plot after processing (Figure S2b).

4.3 Batch Effect Correction

After dataset integration we plotted UMAP embeddings to assess batch effect correction and better understand manifold projections in low dimensional space (Figure 2a) Subdividing the graph by batch showed some correction to each individual to the experiment (Figure 3). We also used an orthogonal method, RPCA, in Seurat as a point of comparison. In some cases over correction occurred where the smaller embryonic batch of cells was completely mixed with the larger adult dataset (Figure S4). Unfortunately, this correction leads to substantial information loss. Thus, we decided upon using a conservative approach to avoid overcorrection in an attempt to preserve the biological heterogeneity.

Harmony clustering also revealed two distinct clusters enriched for embryonic mast cell populations (Figure 2b). Also notable was the apparent heterogeneity and overlapping nature of each harmony cluster. By projecting tissue of origin onto the UMAP we observed an approximate separation between Mast Cells within the Skin (Figure 4) and other tissues including the bladder, small and large intestine, the lymph node occurring across UMAP2. Clustering revealed 10 distinct clusters for both embryonic and adult populations. Interestingly, Cluster 0 and Cluster 1 were highly enriched for Mast Cells originating from the embryonic dataset showed similar clustering and appeared to be more closely related to Mast cells in the skin. However, there was a minor population that appeared to be more closely related to adult type mast cells located in the intestine and lymph nodes. It is important to note that Mast Cells from the embryonic datasets originate from different developmental time points. It may be that this minor population originates from definitive hematopoiesis. Together these data support the hypothesis that the heterogeneity in Mast Cell phenotype may be partly dependent on developmental origin.

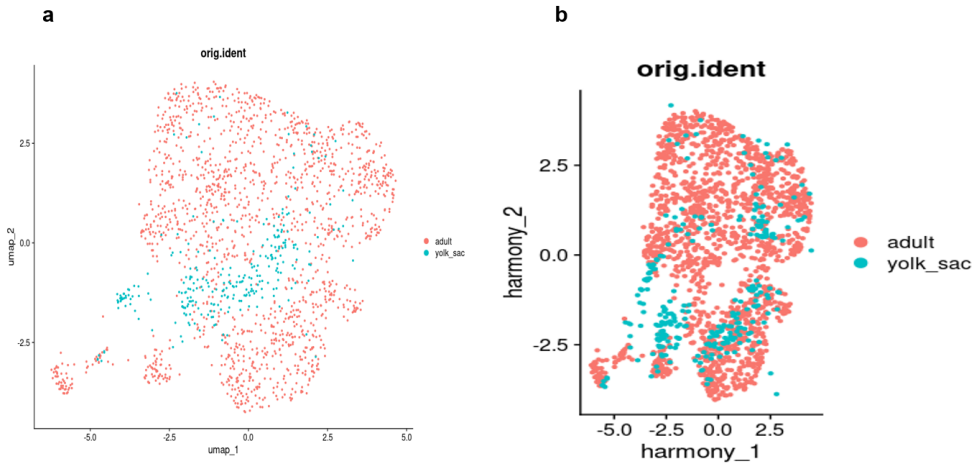


Figure 2: Clustering in Lower Dimension

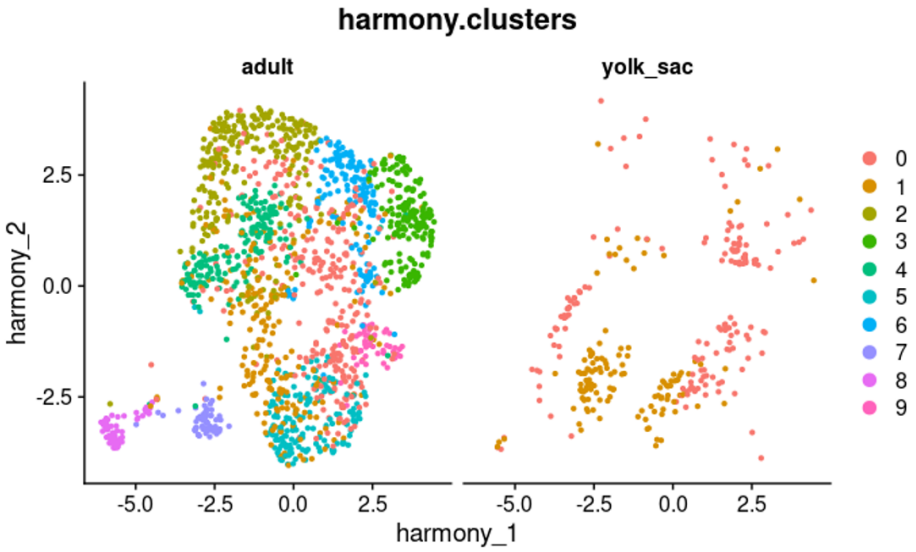


Figure 3: Clustering in different Individual

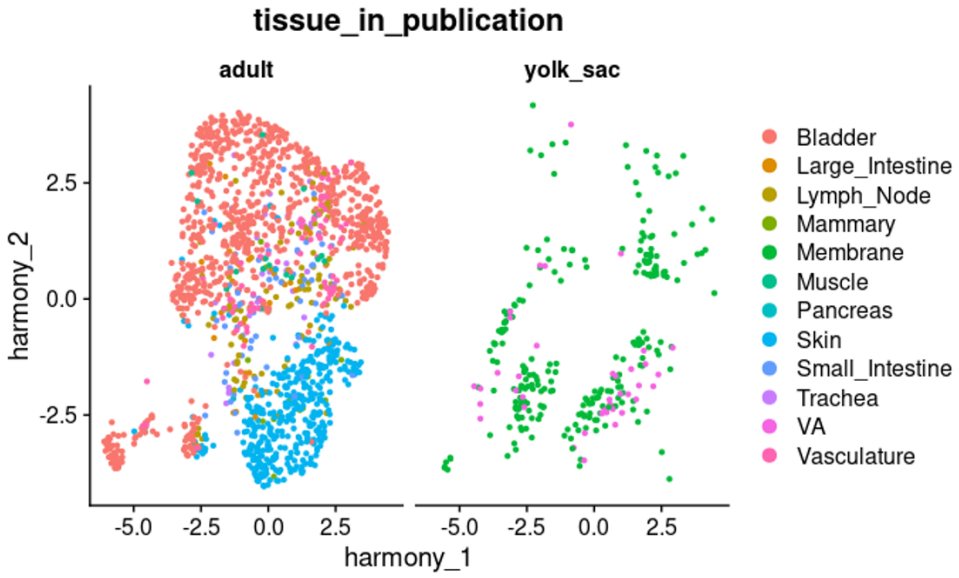


Figure 4: Clustering in different Tissues

4.4 Differentially Expressed Genes & Functional Analysis

Using GSEA and the Molecular Signatures Database (Subramanian et al. (2005), Liberzon et al. (2011)) , we identified the top 10 upregulated GO terms and top 10 upregulated transcription factor (TF) targets for each of the 10 previously identified clusters (results summarized in Table 2, full list of identified GO terms and TF targets can be found in our Github linked in the methods). From our functional analysis, we made a few interesting observations. For example, cluster 5 is enriched for GO terms associated with cell adhesion, so this cluster may be contributing to the development of the vascular system. Cluster 7 is enriched for STAT5A, which is a TF that targets BATF. Interestingly, in mice BATF promotes an inflammatory Mast Cell phenotype (Tomar et al. (2021)). Additionally, cluster 8 and 9 which appear more closely related to cluster 7. For example, TFEB in cluster 8 is a transcription factor associated with lysosomal biogenesis, innate immune activation and Mast Cell degranulation. ATF3 is also involved in Mast Cell survival and negatively regulates gene expression of two critical Mast Cell related cytokines IL6 and IL4 by binding to their gene promoters (Gilchrist et al. (2010)). Together cluster 7, 8, and 9 might represent a more inflammatory/activated Mast Cell phenotype.

In addition to performing functional analysis on the 10 integrated clusters, we separately performed the DEG and functional analysis for the yolk sac cells belonging to cluster 0 and cluster 1 (yolk sac cluster 0 and yolk sac cluster 1 respectively) comparing them individually against all other cells within the integrated dataset. Yolk sac cells in both cluster 0 and 1 were both enriched for the TF targets of DIDO1 which promotes self-renewal in embryonic stem cells via SOX2 and OCT4 (Liu et al. (2014)). They were also both enriched for the TF targets of SUPTH20, which has been associated with Hematopoetic differentiation (Gomes et al. (2002)).

	GO Terms	**Upregulated TFs**
Cluster 0	Apoptosis, Cellular transport, Autophagy,	ELF2, RAG1, ZNFs, E2F2, NAB2, TEAD2
Cluster 1	Mitochondrial terms, Metabolic process	BARX2, **CEBPZ**, **CREB3L4**, CIITA, ZNF, SKIL, SETD1A, SUPT20H, TASOR
Cluster 2	n/a	PSMB5, FOXE1, SETD7, ASH1L, NFE2L1, CDC73, ATXN7L3
Cluster 3	n/a	ETF, PAX7, ETS2, POU2AF1, HDAC4, PAX3, RYBP, **MEF2C**
Cluster 4	n/a	n/a
Cluster 5	Cell adhesion, Kinase activity, Circulatory sys dev.	SKIL, ZNF, **MEF2**
Cluster 6	Mesenchyme migration	HSF, ID1, LMP1, SRF, TEF1, SRF
Cluster 7	Lymphocyte activation, Immune response	**STAT5A**, ZNF, MAML1, ISRE, TGIF
Cluster 8	Cell death, Apoptosis, Leukocyte differentiation	TFEB, NAB2, HES2, BACH2, ZNF, HDGF, KLF7, ARNT
Cluster 9	Cellular stress, Programmed cell death, Apoptosis	ATXN7L, ZNF, GtF2A2, RAG1, PSMB5, NAB2, ATF3, FOXE1
Yolk Sac Cluster 0	Cellular transport: endosome, vesicle membrane	**DIDO1**, ELF2, **SUPT20H**, ZNF711, SKIL, ATF6, CEBPZ, DACH1
Yolk Sac Cluster 1	Mitochondrion, cellular structure, chromatin remodeling	**SUPT20H**, **DIDO1**, SKIL, ZINF711, **ADA2**, BRCA2, SETD1A, ZNF407, SALL4, ZFHX3

Table 2: GSEA Functional Analysis Summary

4.5 Pseudotime Trajectories

In order to map hematopoiesis of mast cells, we included cells that are expected to be involved in this process based on previous knowledge of known mast cell biology (Figure 5a) and also added the addition of MEMPs which are close in relation to mast cells and give rise to platelets and red blood cells, but are not known to directly differentiate to mast cells. Adding the MEMPs allowed our group to see if pseudotime could infer that the MEMPs follow a different differentiation trajectory and allow for enough cell types to be pooled to do integration and clustering, due to limited sample size. When looking specifically at the distribution of cells within each sample we noticed that the two earliest stages of development seemed to have the highest percentage of primitive HSPCs (Figure 5b). This distribution could be a result of early development having more primitive HSPCs or due to batch effects since those two groups were collected together. Early on we noticed that our seurat clusters were separating based on fetal ID as opposed to cell type (Figure S12) so we employed two different integration techniques to correct this issue. We first started with seurat anchors, but noticed that the cells clustered together so much that it was hard to distinguish cell types, potentially due to limited sample size (Figure S5). When taken to Monocle 3 the trajectory plots didn't make biological sense (Figure S7) but the expected expression based on cell type was in line with expected results (Figure S8) so we knew this was an issue with the clustering after integration.

We then tested with our unintegrated datasets to make sure that Monocle 3 could show a good trajectory based on expected biology and with the unintegrated data we were able to see a more clear trajectory plot (Figure S10,Figure S11). Following these results we decided to use another integration method, Harmony, which is known for its ability to limit batch effect and produced new UMAPs (Figure 5c) which separate by cell type better than the ones integrated using Seurat Anchors. We did notice that there were two clusters of cells that were not as well defined (clusters 10 and 11 on Figure S13) and did write in our source code a method to exclude it, but decided to omit this step due to limited sample size. Following integration with Harmony we proceeded with the programing we used on Monocle 3 to generate pseudotime trajectories (Figure 6a,b) and were able to get a trajectory that more or less followed the unintegrated data and had the mast cells and MEMPs with the highest pseudotime (most differentiated) over the EBPs or the HSPCs (Figure 6d). We had run the code through pseudotime without MEMPs, but due to limited samples decided not to continue with the analysis in this way, but in the future would like to pursue the trajectory analysis using only mast cells, mast cell progenitors, and HSPCs with a larger sample size. When looking specifically at the pseudotime based on Carnegie stage of development, we did see a general trend of the later stages of development having a higher pseudotime with the exception of CS17 and CS18, but this could be due to limited sample size and high variability (Figure 6c).

Following the trajectory analysis we decided to look at the expression of a few genes that are related to cell fate in hematopoiesis and one gene that was upregulated in DEG analysis conducted with the yolk sac and adult dataset. During this analysis we noticed that canonical marker for mast cells KIT and MITF, were most highly expressed in the cluster of cells annotated as mast cell, while markers for more pluripotent myeloid cells, CD34 and RUNX1 were more highly expressed in the clusters related to myeloid progenitors and HSPCs

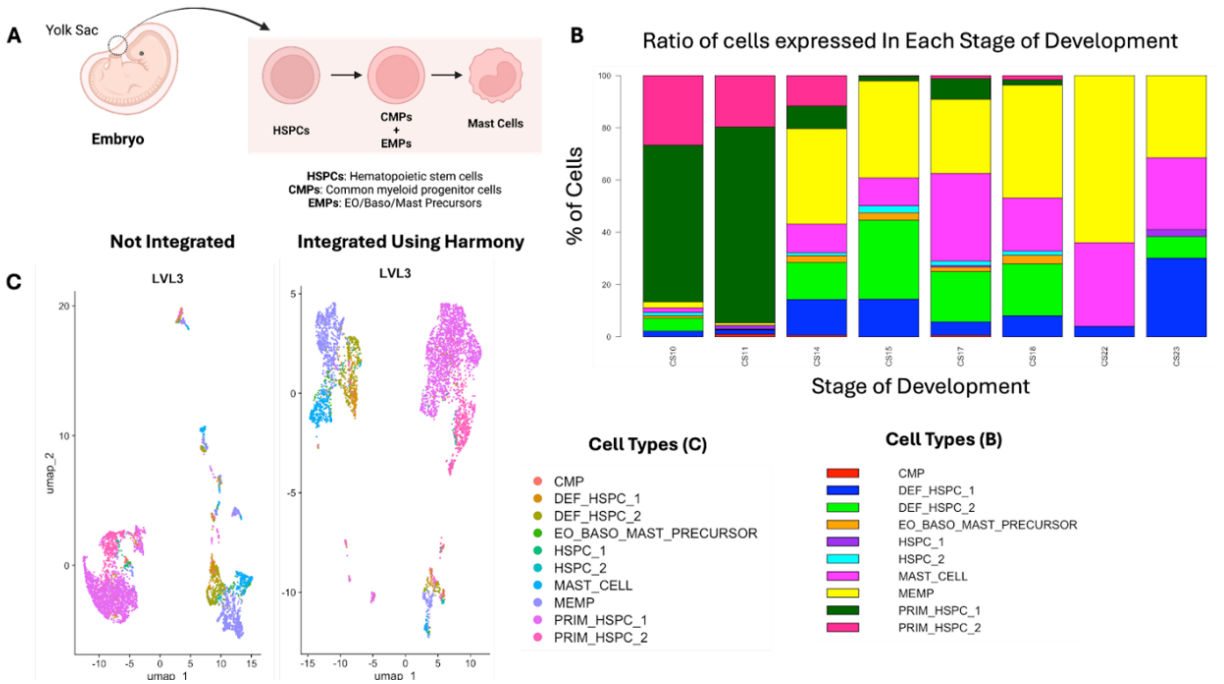


Figure 5: Overview of clustering and cell types in yolk sac data for Pseudotime Trajectory. A) Mast cell differentiation overview. B) Overview of the % of each cell type based on Carnegie stage of development. C) UMAPs of clustering before and after integration.

(Figure 7a,b). When looking at the gene upregulated in the DEGs analysis, we didn't see any major correlation with pseudotime (Figure 7b).

In order to look more closely at the mast cells we selected a subset of cells in the upper right hand corner of our UMAP where the mast cells, MEMPs, EBPs, and some definitive HSPCs were located. Using this subset of cells we remapped for three markers that were associated with early, middle, and late stage differentiation to mast cells (CD34, ITGB37, KIT). With the subset of cells we could see more clearly that KIT was highly expressed in the mast cells, ITGB37 was expressed fairly evenly throughout all of the cell types, and CD34 was expressed in the earlier cell fates (HSPCs, MEMPs, and EBPs) but almost not expressed in the mast cells (Figure 8a). We then also looked for a few of the previous makers we looked at in our heat maps (Figure 7b) as a function of developmental stage within the subset and saw no correlation between Carnegie stage of development and any of the markers chosen (Figure 8b). One of the markers we looked at, AARS1 (alanyl-tRNA synthetase 1: thought to play a crucial role in protein synthesis), was one of the most upregulated genes that changed as a function of pseudotime (Table S1) but seemed to also be independent of stage of development. Part of the reason for this discrepancy could also be related to the limited sample sizes and high variability in gene expression across the cells.

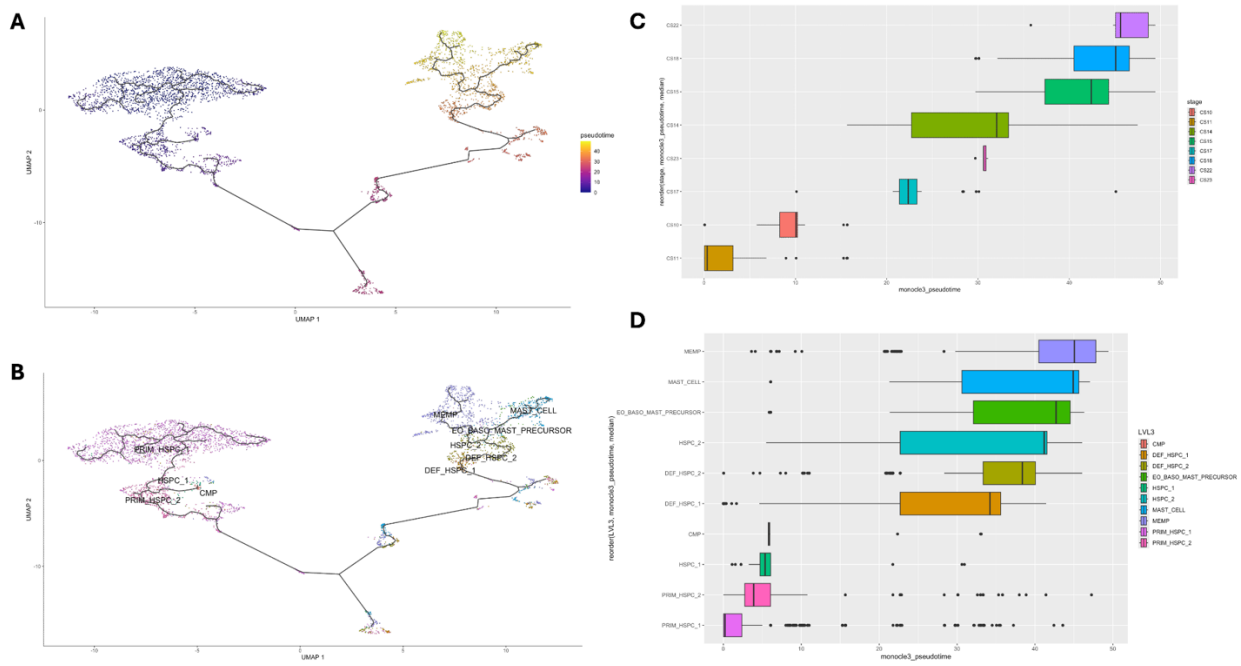


Figure 6: Pseudotime with Harmony clustering and associated barplots to show distribution of cells. A) Pseudotime Trajectory Plot B) Pseudotime Trajectory plot annotated C) Bar plots with pseudotime of carnegie developmental stage D) Bar plots with cells type to show how Monocle 3 annotated their pseudotime (cell fate)

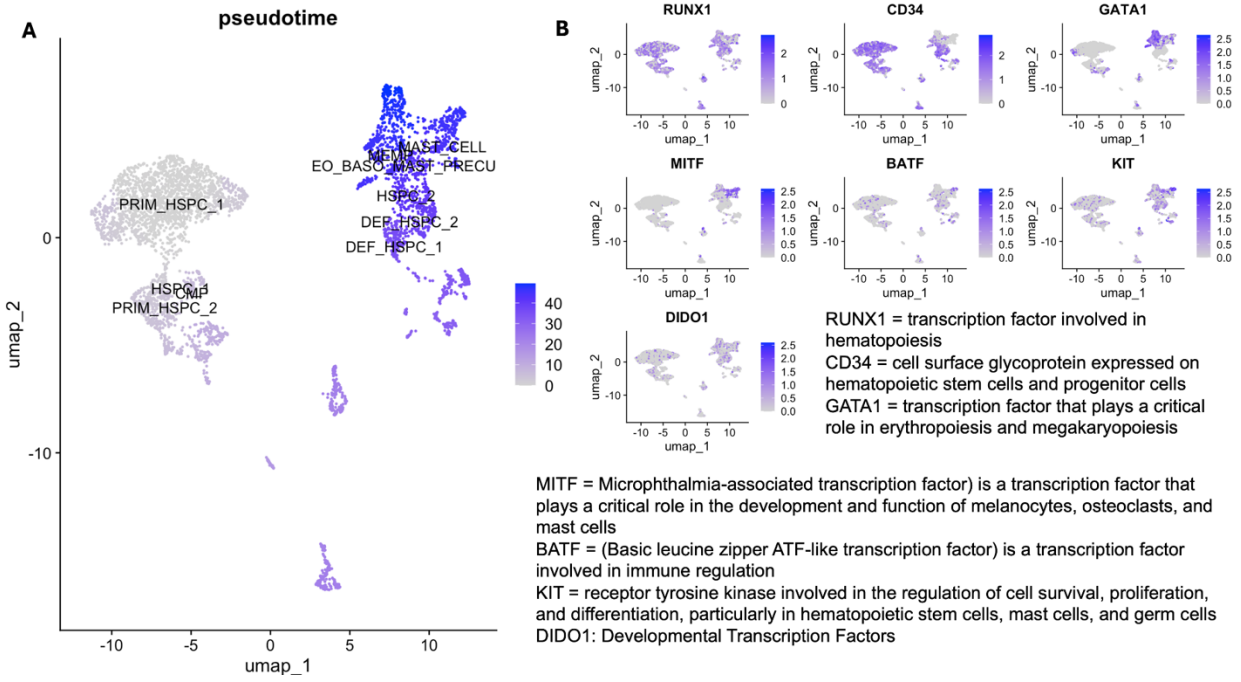


Figure 7: A) pseudotime trajectory with annotated cell types (no branching). B) Heat maps of cells important for hematopoiesis and mast cells and ones that came across as upregulated in the previous DEGs analysis.

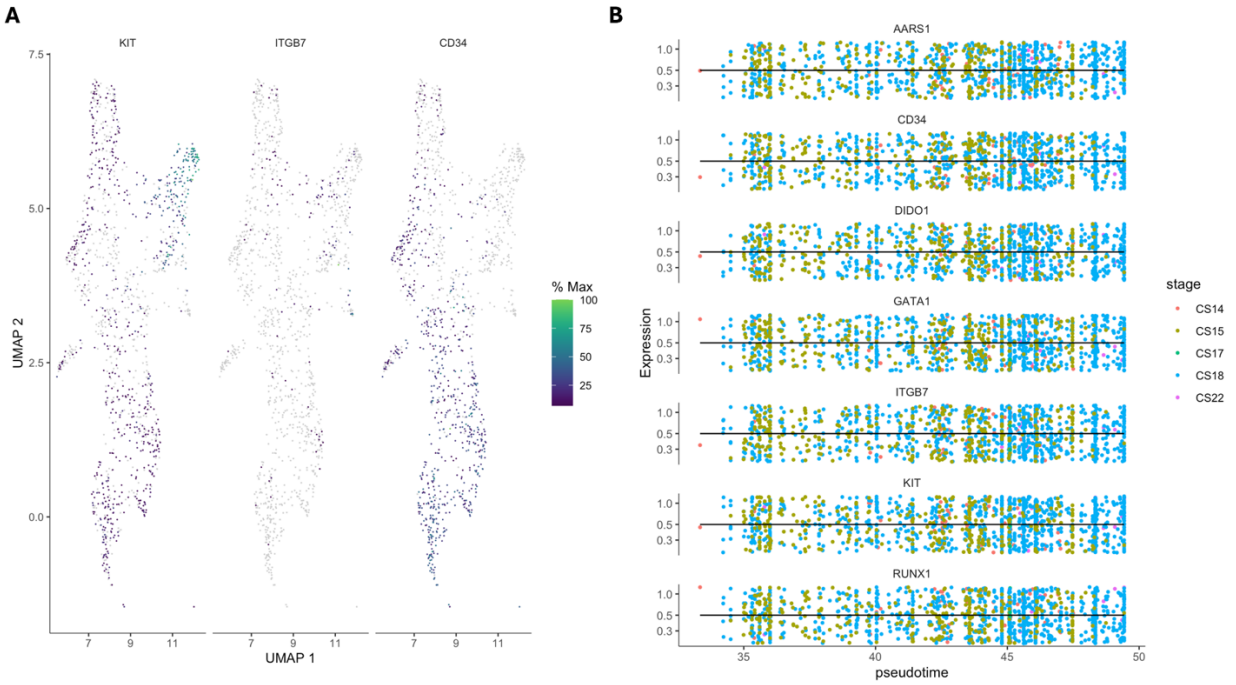


Figure 8: DEGs of the upper right branch of cells within the UMAP (cluster of cells containing mast cells, MEMPs, EBP, and definitive HSPCs. A) plot of DEGs for genes related to late, middle, and early Mast cell development. B) Plot of DEGs in subset based on pseudotime and stage of development.

5 Discussion

GSEA results We were able to find some enriched GO terms that were expected from mast cell progenitors such as "immune response" and "leukocyte differentiation", as well as some TF enrichment. We found that the GO term 'protein DNA complex organization' was enriched in both of the yolk sac clusters when compared to the rest of the clusters which suggested .Yolk sac cluster 1 was enriched for chromatin remodeling. An exciting next set of experiments would be to run ATAC-seq and see if those results line up with the RNA-seq data. Yolk sac cluster 0 was enriched for ZNF711, a TF involved in transcription factor binding activity.

Using Monocle 3 we were able to map hematopoiesis of Mast cells following the expected trajectory of HSPCs to EMP or CMP to Mast cells and look at the distribution of differentially expressed genes. When comparing pseudotime analysis between the data set that was not integrated or the one that was integrated with Seurat or Harmony we see some key differences mainly in the one integrated using Seurat Anchors. The trajectory analysis conducted using Seurat Anchors was not well defined and lacked key information. This may have been the better option for our dataset which was limited in sample size and varied due to batch effect.

Limitations We were able to find some enriched GO terms and TFs that we expected to see in the data. However, for some of the clusters (such as 2, 3, and 4) we were unable to identify any significantly enriched GO terms at all, and in the case of cluster 4, we were also unable to identify any enriched TF targets. Upon closer examination, the TF targets identified for clusters 2 and 3 correlated to an unusually large amount of unannotated proteins potentially explaining why we failed to identify significantly enriched GO terms for these clusters.

Limitations in the functional analysis came partly from the difficulty in correcting for batch effects between the different data sets. Additionally, there were a small number of mast cells in the yolk sac data, and after filtering there were even fewer. We attempted to run a pseudo bulk analysis, however the statistical power was too low to yield meaningful results due to limited sample size. For better results, a significantly larger amount of effort would need to be spent on QC and making sure that the data sets are integrated properly. Even with the extra effort however, this project would still be limited by the number of biological replicates as well as the number of cells available to analyze.

A possible limiting factor in our pseudotime analysis is the limited number of cells we had to work with. In our data set we only had 10 embryos, many with under 100 cells of the annotated populations (HSPCS, MEMP, EMP, CMP, or Mast Cells). There was also a very noticeable batch effect, with the seurat clusters lining up with the fetal IDs which could have also limited our analysis despite running through Seurat Anchors and Harmony.

Pseudotime trajectory analysis is a valuable tool to study cell fate and predict differentiation pathways but is limited due to its assumption of linear progression. Because biological processes follow complex branching, assuming linearity from initial to final state may be misleading and inaccurate. Our group did see this limitation very clearly when many of the primitive and definitive HSPCs were included in the trajectory analysis as they interrupted the expected biological trajectory when using Seurat Anchors to integrate. One possible way to overcome this limitation is to look instead at the RNA velocity of the cells which relies

on spliced and unspliced RNA reads to infer future states of individual cells. Since there is no trajectory to follow for RNA velocity, we would be better able to gauge the directionality of cell fate and capture dynamic changes in cellular processes that we are unable to with pseudotime trajectory analysis.

Pseudo bulk analysis was hindered due to the small number of biological replicates included in the yolk sac data set. A larger number of biological replicates would be needed to get statistical power.

6 Acknowledgement and Contributions

Authorship order was done by alphabetizing based on last name. Ashley Bielawski planned the group meetings, found DEGs, performed functional analysis, made the table summarizing GO analyses, and made the supplemental figure with the volcano plots for each cluster. She helped with discussion and interpretation of the results and writing of the methods and discussion section. Chase Lindeboom collaborated for the DEG and functional analysis, helped others troubleshoot at various stages of analysis, prepared the introduction for this report, contributed to the subsections of this report relating to differential gene expression and functional analysis, and helped edit this report. Christian Rizza was involved in processing the data (writing code, deciding appropriate QC metrics, discussion/interpretation of results, writing, and editing.) Paula Wu prepared the data for downstream analysis, did preliminary data preprocessing, created and helped others with some analysis and data organization. Mariana Sierra was focused on the pseudotime trajectory analysis of mast cells and myeloid progenitor cells. She conducted all of the integration and filtering of the the datasets with the myeloid progenitors and stem cells, translation of the seurat object for the yolk sac data set into a cell dataset object to run on monocle 3, and made all the images related to the pseudotime analysis. Mariana also wrote all the sections related to the pseudotime analysis, helped organize the data on github, and made a lot of the figures and tables not generated through R or Python code.

Key	Highest Contribution	High Contribution	High/ Moderate Contribution	Moderate Contribution	Moderate/ Little Contribution	Little Contribution	No Contribution

	Project Concept	Data Processing/ Filtering Reads	Integrating Datasets/ Clustering	Differential Gene Expression	Pseudotime Trajectory	Data Organization	Making Figures/ Tables	Writing	Editing
AB									
CL									
CR									
MS		*	*	*					
PW									

*Specifically related to YS data used for Pseudotime Analysis

Figure 9: Author Conrtibution Diagram

A Supplementary Material

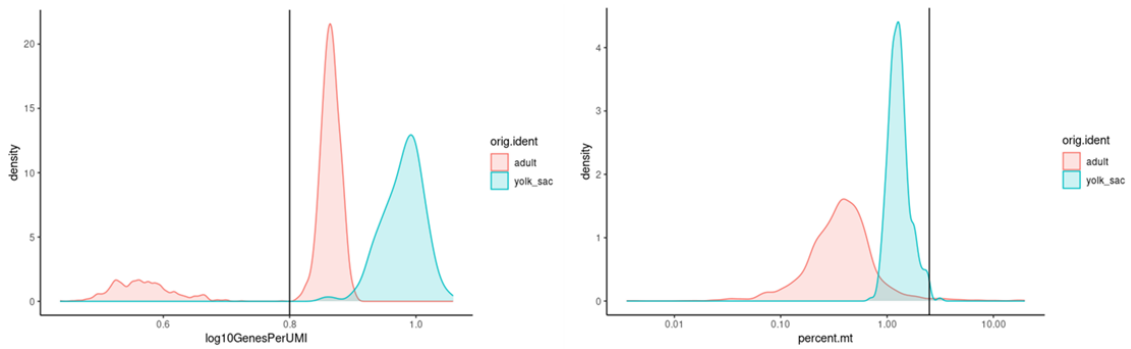


Figure S1: Library Complexities

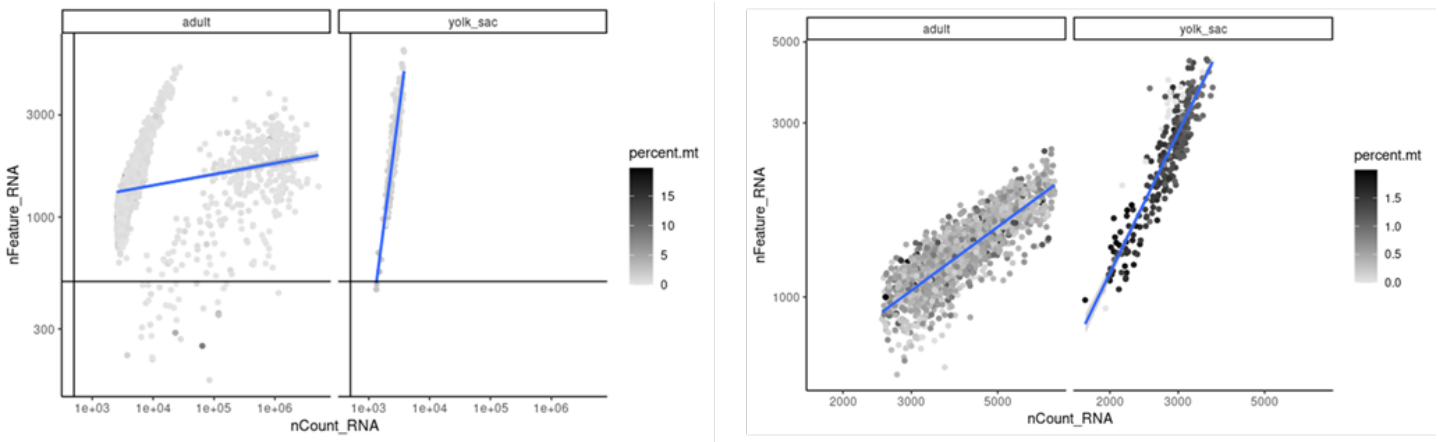


Figure S2: Mitochondrial DNA Content

	status	p_value	morans_test_statistic	morans_I	gene_short_name	q_value
AARS1	OK	0	55.36076	0.2712872	AARS1	0
ABCF2	OK	0	48.59860	0.2468723	ABCF2	0
ABCG2	OK	0	40.84116	0.2044630	ABCG2	0
ABRACL	OK	0	37.65973	0.1914447	ABRACL	0
ABRAXAS1	OK	0	71.79641	0.3651204	ABRAXAS1	0
AC004148.2	OK	0	64.93952	0.3303168	AC004148.2	0

Table S1: Genes that changes the most as a function of pseudotime

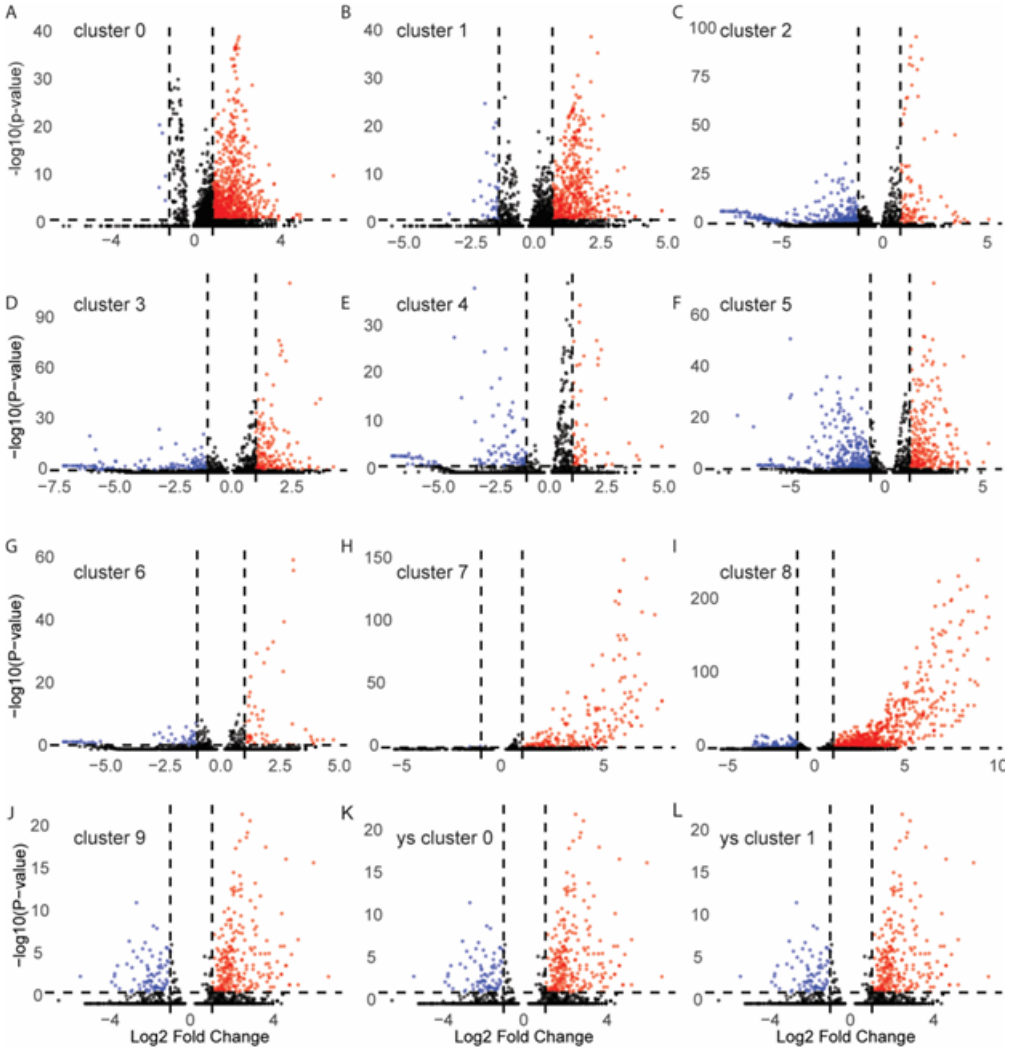


Figure S3: Vertical dotted lines are at -1 and 1 log2 fold change. Horizontal line is at the cut off p-value 0.05. Genes that are upregulated in a cluster based on the log fold change and p-value criteria are red, and downregulated genes are blue. **A-J**) Volcano plots showing the number of DEGs for each harmony cluster and **K-L**) each of the yolk sac clusters.

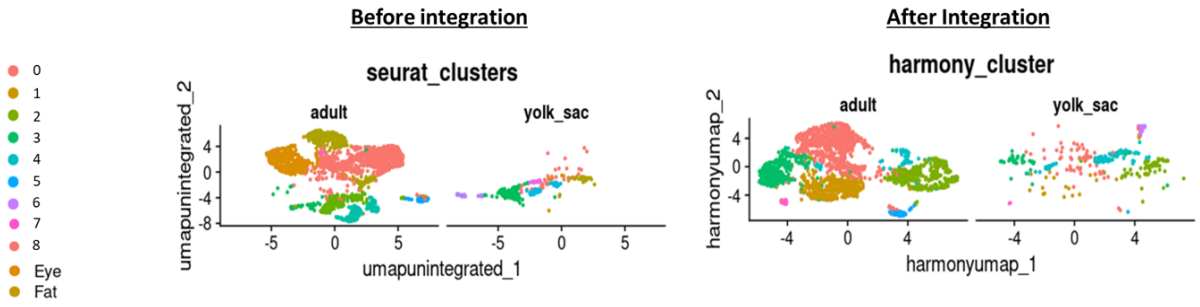


Figure S4: Comparison: Seurat & Harmony

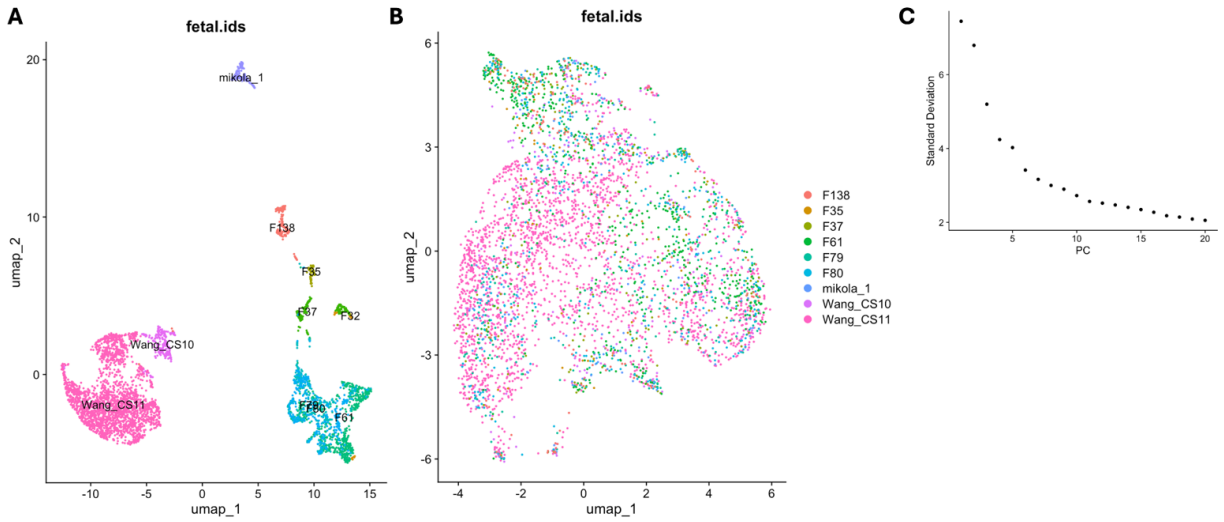


Figure S5: Clustering based on Seurat and associated elbow plot for optimal number of clusters.

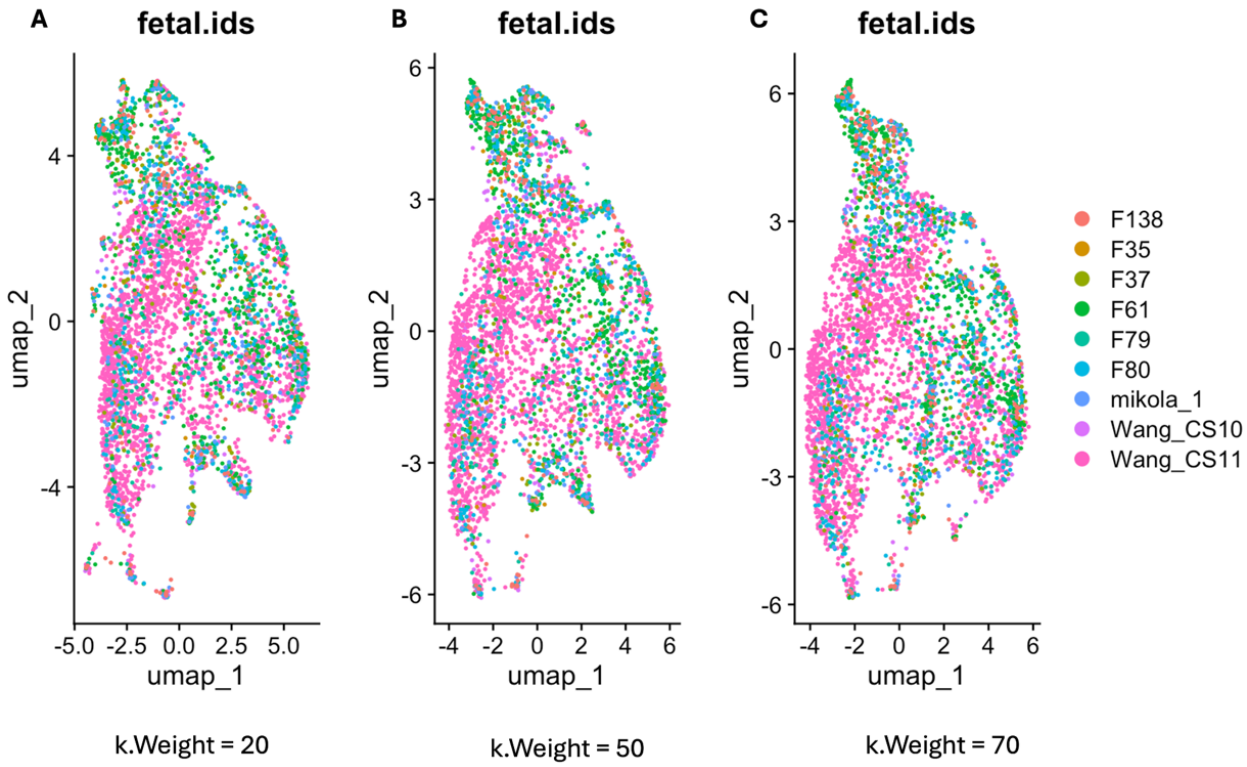


Figure S6: Seurat clustering with various k.Weights.

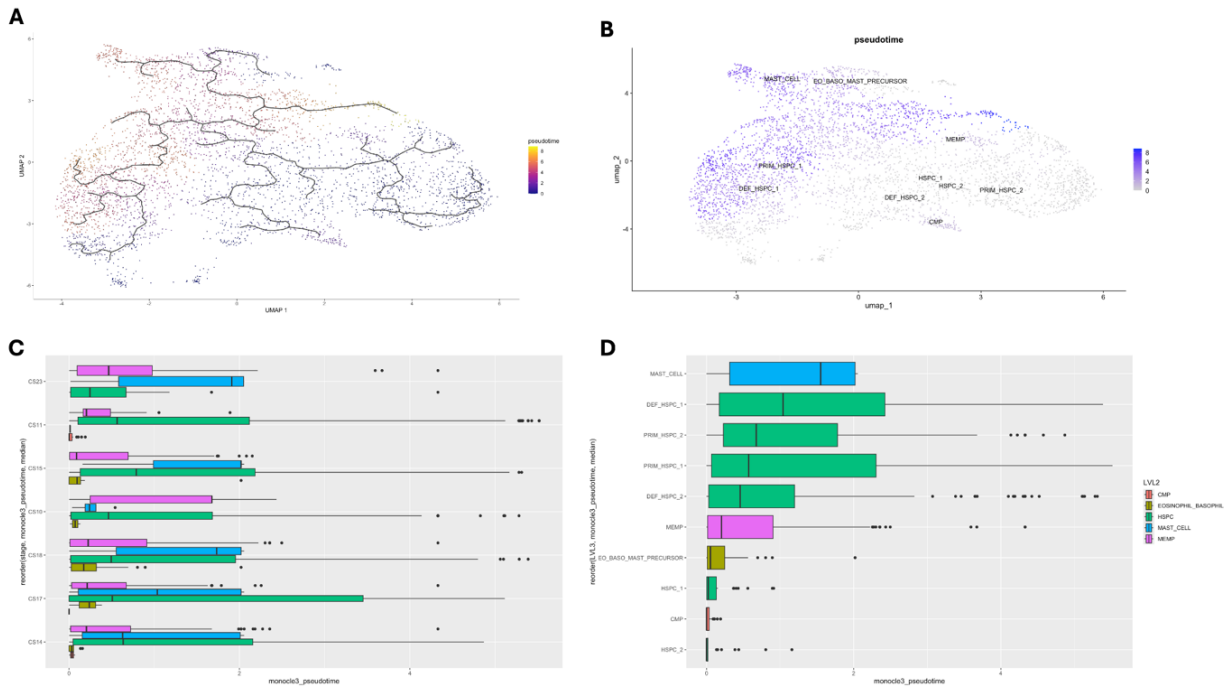


Figure S7: Pseudotime with Seurat clustering and associated barplots to show distribution of cells. A) Pseudotime Trajectory Plot B) Pseudotime Trajectory plot annotated C) Bar plots with Cell types based on developmental stage D) Bar plots with cells type to show how Monocle 3 annotated their pseudotime (cell fate).

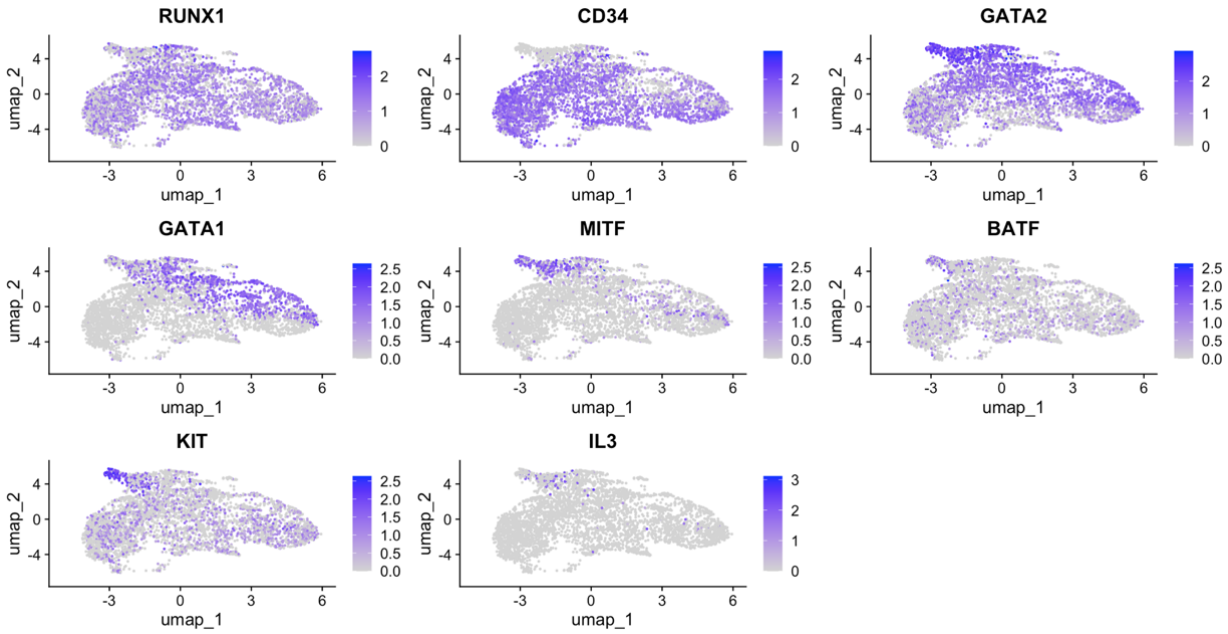


Figure S8: Key genes associated with Mast cell development and hematopoiesis.
Key:RUNX1 = transcription factor involved in hematopoiesis
CD34 = cell surface glycoprotein expressed on hematopoietic stem cells and progenitor cells
GATA1 = transcription factor that plays a critical role in erythropoiesis and megakaryopoiesis
GATA2 = Transcription factor involved in hematopoiesis, lymphangiogenesis, and vascular development
MITF = Microphthalmia-associated transcription factor) is a transcription factor that plays a critical role in the development and function of melanocytes, osteoclasts, and mast cells
BATF = (Basic leucine zipper ATF-like transcription factor) is a transcription factor involved in immune regulation
KIT = receptor tyrosine kinase involved in the regulation of cell survival, proliferation, and differentiation, particularly in hematopoietic stem cells, mast cells, and germ cells

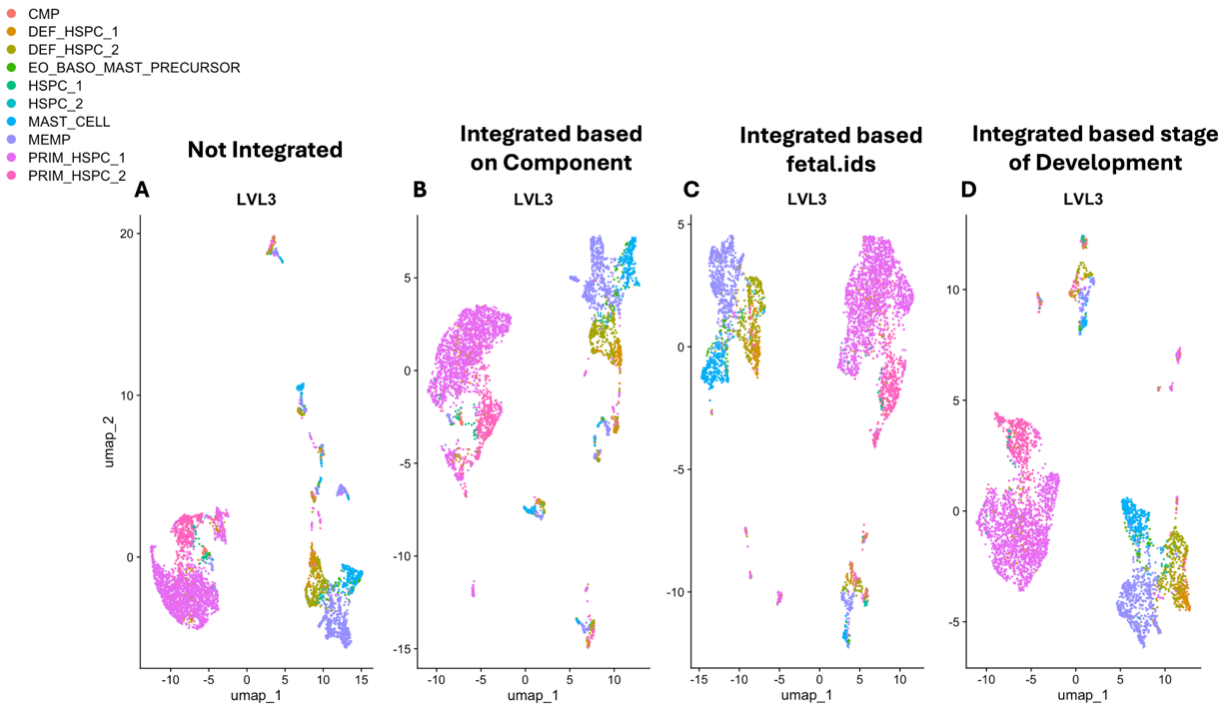


Figure S9: Integration based on Harmony and different set conditions.

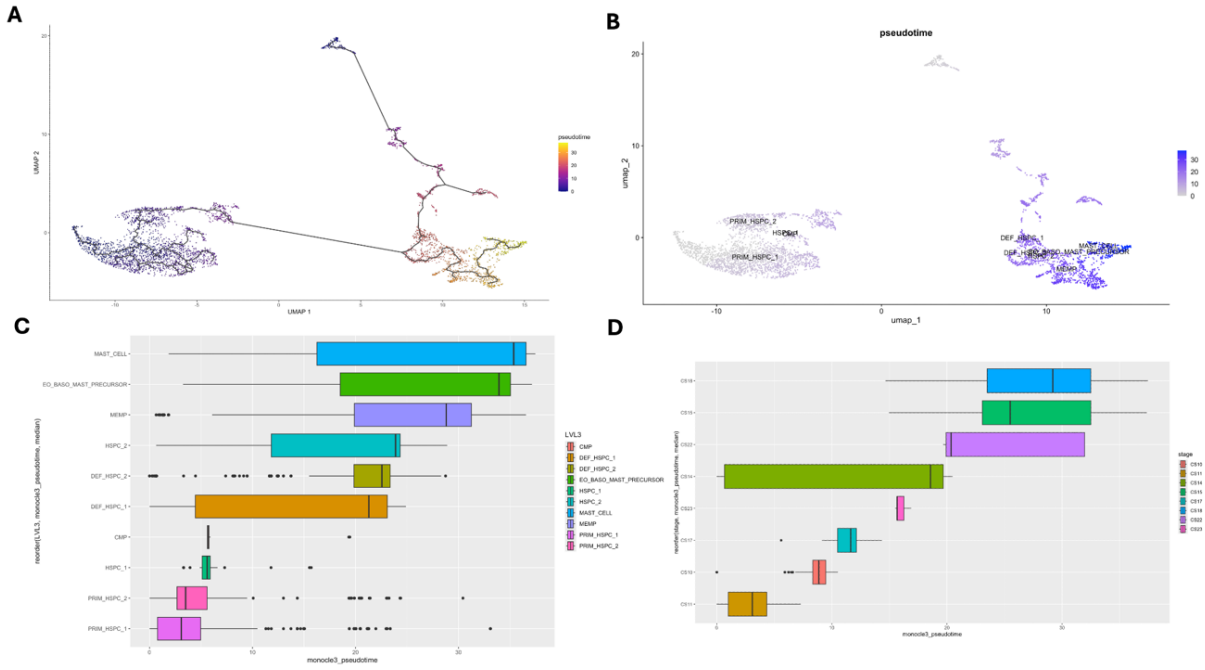


Figure S10: Unintegrated data sets Pseudotime (A-B) and plots showing distribution of cell types based on pseudotime (C) and Carnegie stage of development based on pseudotime (D).

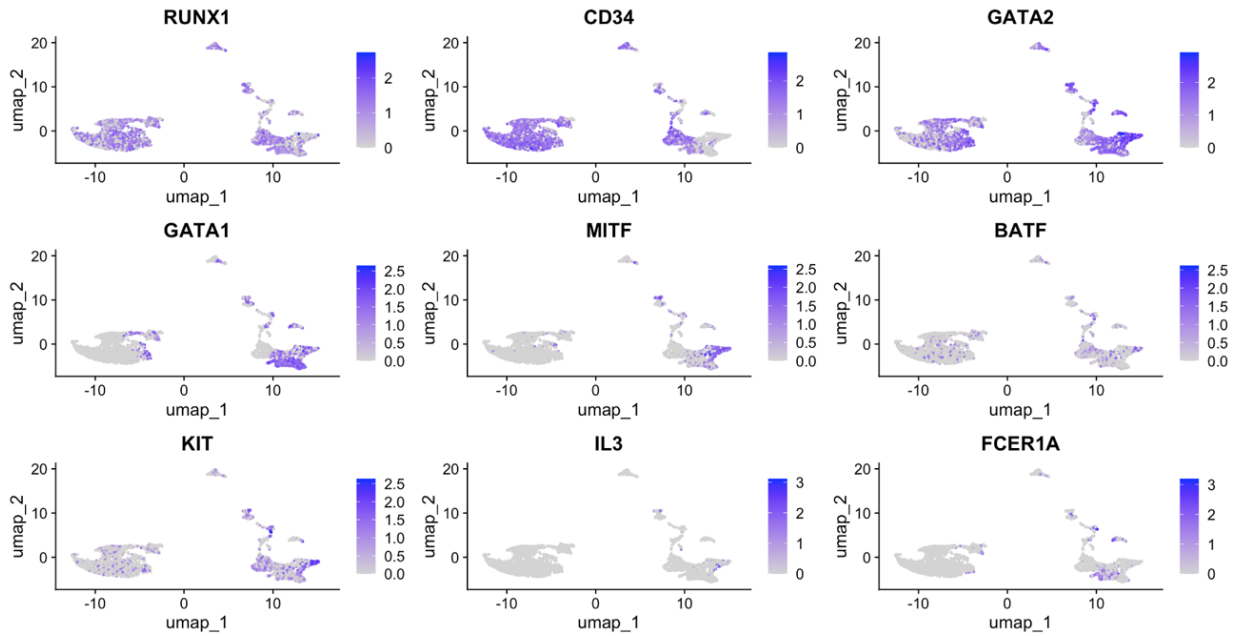


Figure S11: Key genes associated with Mast cell development and hematopoiesis.

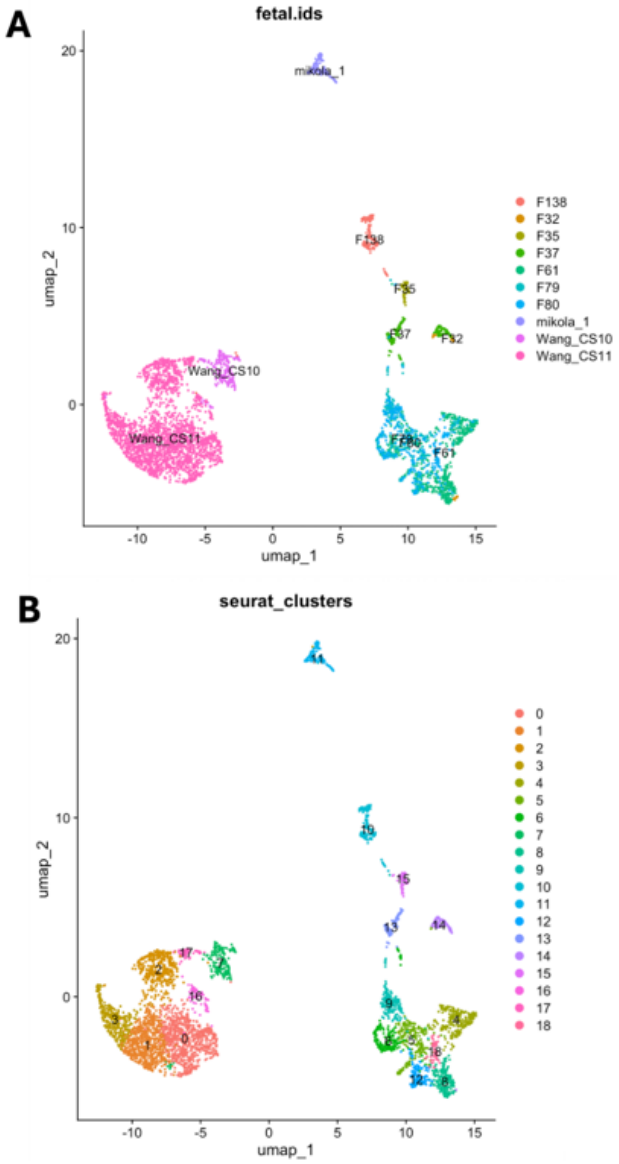


Figure S12: Mast cell clustering based on seurat object vs fetal ID before integration.

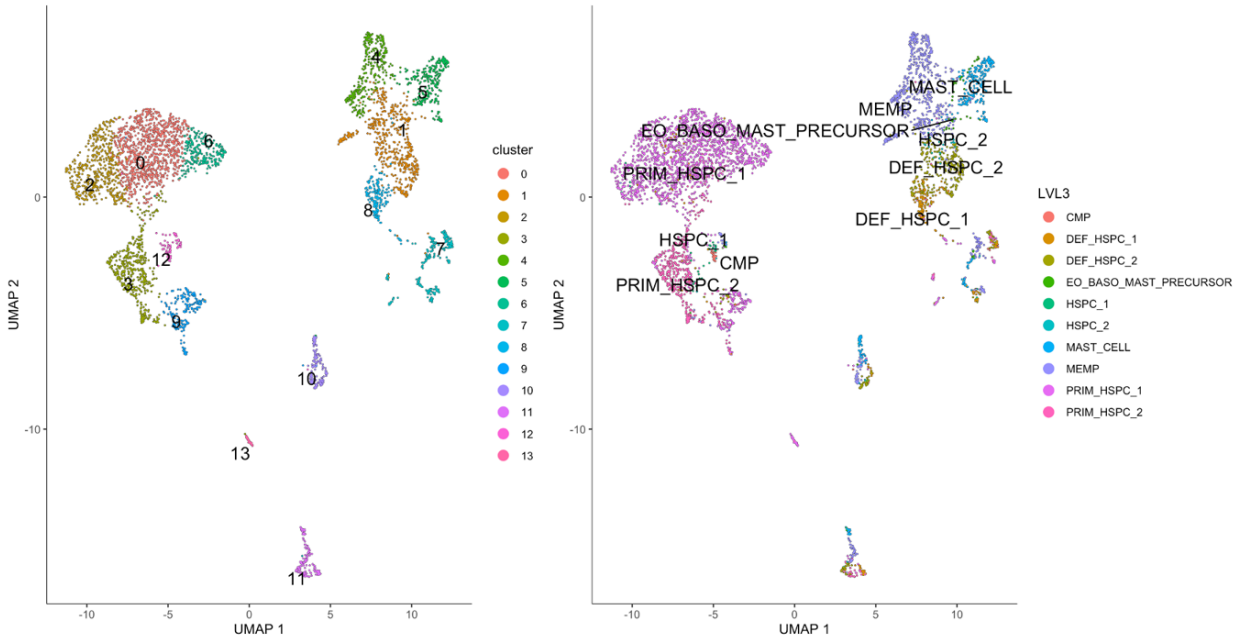


Figure S13: Clusters on Monocle 3 and annotation of cell types.

References

Boisset, J.-C., van Cappellen, W., Andrieu-Soler, C., Galjart, N., Dzierzak, E., and Robin, C. (2010). In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature*, 464(7285):116–120.

Chotinantakul, K. and Leeanansaksiri, W. (2012). Hematopoietic stem cell development, niches, and signaling pathways. *Bone marrow research*, 2012.

Crisp, A. J., Chapman, C. M., Kirkham, S. E., Schiller, A. L., and Krane, S. M. (1984). Articular mastocytosis in rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 27(8):845–851.

Frame, J. M., McGrath, K. E., and Palis, J. (2013). Erythro-myeloid progenitors:"definitive" hematopoiesis in the conceptus prior to the emergence of hematopoietic stem cells. *Blood Cells, Molecules, and Diseases*, 51(4):220–225.

Galli, S. J. and Tsai, M. (2012). Ige and mast cells in allergic disease. *Nature medicine*, 18(5):693–704.

Gentek, R., Ghigo, C., Hoeffel, G., Bulle, M. J., Msallam, R., Gautier, G., Launay, P., Chen, J., Ginhoux, F., and Bajénoff, M. (2018). Hemogenic endothelial fate mapping reveals dual developmental origin of mast cells. *Immunity*, 48(6):1160–1171.

Gilchrist, M., Henderson Jr, W. R., Morotti, A., Johnson, C. D., Nachman, A., Schmitz, F., Smith, K. D., and Aderem, A. (2010). A key role for atf3 in regulating mast cell survival and mediator release. *Blood, The Journal of the American Society of Hematology*, 115(23):4734–4741.

Gilfillan, A. M., Austin, S. J., and Metcalfe, D. D. (2011). Mast cell biology: introduction and overview. *Mast Cell Biology: Contemporary and Emerging Topics*, pages 2–12.

Goh, I., Botting, R. A., Rose, A., Webb, S., Engelbert, J., Gitton, Y., Stephenson, E., Quiroga Londoño, M., Mather, M., Mende, N., et al. (2023). Yolk sac cell atlas reveals multiorgan functions during human early development. *Science*, 381(6659):eadd7564.

Gomes, I., Sharma, T. T., Edassery, S., Fulton, N., Mar, B. G., and Westbrook, C. A. (2002). Novel transcription factors in human cd34 antigen–positive hematopoietic cells. *Blood, The Journal of the American Society of Hematology*, 100(1):107–119.

Jagannathan-Bogdan, M. and Zon, L. I. (2013). Hematopoiesis. *Development*, 140(12):2463–2467.

Kolkhir, P., Elieh-Ali-Komi, D., Metz, M., Siebenhaar, F., and Maurer, M. (2022). Understanding human mast cells: Lesson from therapies for allergic and non-allergic diseases. *Nature Reviews Immunology*, 22(5):294–308.

Lee, J. Y. and Hong, S.-H. (2020). Hematopoietic stem cells and their roles in tissue regeneration. *International journal of stem cells*, 13(1):1.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.

Lin, S., Zhao, R., Xiao, Y., and Li, P. (2015). Mechanisms determining the fate of hematopoietic stem cells. *Stem cell investigation*, 2.

Liu, Y., Kim, H., Liang, J., Lu, W., Ouyang, B., Liu, D., and Songyang, Z. (2014). The death-inducer obliterator 1 (dido1) gene regulates embryonic stem cell self-renewal. *Journal of Biological Chemistry*, 289(8):4778–4786.

Moore, M. A. and Metcalf, D. (1970). Ontogeny of the haemopoietic system: yolk sac origin of in vivo and in vitro colony forming cells in the developing mouse embryo. *British journal of haematology*, 18(3):279–296.

Sanbomics (2023). Convert h5ad anndata to a seurat single-cell r object. Online video.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502.

Sonoda, T., Hayashi, C., and Kitamura, Y. (1983). Presence of mast cell precursors in the yolk sac of mice. *Developmental biology*, 97(1):89–94.

St. John, A. L., Rathore, A. P., and Ginhoux, F. (2023). New perspectives on the origins and heterogeneity of mast cells. *Nature Reviews Immunology*, 23(1):55–68.

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

Tauber, M., Basso, L., Martin, J., Bostan, L., Pinto, M. M., Thierry, G. R., Houmadi, R., Serhan, N., Loste, A., Blériot, C., et al. (2023). Landscape of mast cell populations across organs in mice and humans. *Journal of Experimental Medicine*, 220(10):e20230570.

Tomar, S., Ganesan, V., Sharma, A., Zeng, C., Waggoner, L., Smith, A., Kim, C. H., Licona-Limón, P., Reinhardt, R. L., Flavell, R. A., et al. (2021). Il-4–batf signaling directly modulates il-9 producing mucosal mast cell (mmc9) function in experimental food allergy. *Journal of Allergy and Clinical Immunology*, 147(1):280–295.

Valent, P., Akin, C., and Metcalfe, D. D. (2017). Mastocytosis: 2016 updated who classification and novel emerging treatment concepts. *Blood, The Journal of the American Society of Hematology*, 129(11):1420–1427.