**Working title: Exploring cell type diversity in murine development through whole-embryo single-cell RNA sequencing**

Matthew B. Eastman1,2,3, Boyang Li1, Roddy M. Taing1, Eric M. Stas1, Ion I. Mandoiu4,5, Jason D. Gibson1,5, Craig E. Nelson1,5

1. University of Connecticut Department of Molecular and Cell Biology, Storrs, CT

2. University of Connecticut Department of Physiology and Neurobiology, Storrs, CT

3. Connecticut Institute for the Brain and Cognitive Sciences, Storrs, CT

4. University of Connecticut Department of Computer Science and Engineering, Storrs, CT

5. University of Connecticut Institute for Systems Genomics, Storrs, CT

Corresponding Author: craig.nelson@uconn.edu

**Figure outline:**

**Experimental workflow**

Might also consider: images of TS19 samples (head, hearts, etc), histograms of UMI and 3-NM QC, but these may be more suited for supplemental

**tSNE projections and example of IBS** - sample and one cluster level (3D needed?) - highlight global data structure

Scatter of sample and cluster (to show cell types present in all samples versus enriched in other samples)

(Can make a very large supplemental figure with proportions of cells from each sample per cluster or the concentric ring graph)

**Flow chart** of metrics used to assign biological function to clusters and branches

Grapes-and-raisins trees (cytoscape hierarchical layout?) featuring scores/significance of matches to GO:BP, LM:cell/organ/tissue, mousePheno (this is the rationale for assigning cell types at a particular level of splitting. Could be supplemental)

What are the distributions of scores across the levels of splitting? What are the max scores versus level of splitting?

Text comparison-based distance matrix for silhouette scores? (density in tSNE is “useless”)

**Global assignment of biological function** projected on IBS tree for GO:BP, LM:cell/tissue/organ, mousePheno

Colored arcs/wedges and/or branches

Specific biology and function of select clusters:

Sub-cluster projections of select clusters with GO:BP, LM:cell/tissue/organ, mousePheno (if not listed/redundant from above) and possibly other useful information (probably not lineage progression but something similar)

**Early lineage markers: Ectodermal/Mesodermal/Endodermal**

**Neuroectodermal lineages:**

Ectoderm IDed by IBS (concordance of projected marker lists (RD/LM) with IBS cluster assignment

Projections of Neural Progenitor, Retinal Progenitor, Neuronal lineage, Glial lineage

Specifics of: Ventral telencephalon MGE/LGE/CGE projections

GAD1/2, VGAT (overlap with neurogenic regions and mature markers)

**Mesodermal:** Similar to what’s done for neuroectodermal:

**Blood**

Roddy projecting ~15 clusters for JDG to check

**Cardiac**

Cardiac dumbbell

Angioblasts

Cardiac progenitors (not really a great signature, could leave out)

Supps:

3D -rotatable cluster tSNE in supplemental?

Just as a movie or guided walk-through. Embed plotly?

nGene/nUMI analysis

Supp: 8-panel figure of tSNE colored by IBS level

By sample: concentric ring graph, scatter plot, violin plot (show similar info)

**Discussion points:**

Large-scale, whole-embryo, transcriptome-wide analysis of development

Selection of embryonic timepoint and sampling of tissues and cells

Why don’t we recover many mature cardiomyocytes, even from the heart microdissection? Overwhelmingly blood

Even with longer time in enzymatic dissociation cocktail compared to whole-embryo and other samples, we still recover fewer cardiomyocytes than we would expect from our microdissection of the heart at this stage of development. We believe this is due to the structure of functionally differentiated heart muscle, which is composed of many branched and interconnected cardiomyocytes and may be more resilient to enzymatic dissociation, or make these cells less likely to survive the procedure.

To ensure adequate sampling of cardiac tissue for future investigations, different dissociation conditions will be required.

Application of single-linkage hierarchical clustering and iterative binary splitting using Ward’s hierarchical method. Iterative binary splitting performs as well as traditional clustering methods and offers a more intuitive method of tracking the implications of splitting on the biological signatures extracted from a given cluster by arranging each iteration of splitting as a binary pair of nodes, which can be easily compared.

Inspection of the 3D manifold of our tSNE projection led us to choose single-linkage hierarchical clustering to pull out well-separated clusters of cells. Comparison of cluster membership across these cells revealed good concordance with iterative binary splitting, suggesting that cluster assignment in these groups was robust to the clustering method used.

We did not detect a significant difference in either the maximum scores or the average scores of significant matches returned from Gene Analytics for cell cluster DGE/GSEA between Ward’s, k-means, agglomerative clustering, and iterative binary splitting (data not shown), indicating that all are, quantitatively, acceptable algorithms for clustering of our single-cell data. We therefore selected our iterative binary splitting approach for subsequent analysis due to the ease of keeping track of cell cluster identity for the different numbers of clusters generated at each iteration of splitting. This allowed us to “cut the branches” of our tree at different levels, depending on the strength of GO matches and the biological significance of the signatures present in cell clusters at each iteration of binary splitting.

Justification of our DGE method: Many tools for differential gene expression for single cell data rely on the assumption that most genes are not differentially expressed, and use the negative binomial distribution as a model to estimate gene-specific variation/dispersion.

We need to justify our choosing the “simple” equation for differential gene expression: do we discuss generalized linear models etc?

Looks like we will run a new DGE

Projection of early lineage markers onto tSNE space

Signatures of marker genes were coregistered with clusters identified through iterative binary splitting and single-linkage hierarchical clustering.

Ectodermal signatures and specific neuronal versus glial and neural progenitor signatures

Projection of LifeMap annotated gene expression for cell types, including medial ganglionic eminence and lateral ganglionic eminence.

Proximity of surface ectoderm and neuroectoderm in tSNE space

Nebulous “mesoderm” signature cluster

Recovery of a diverse array of hematopoietic system cell types, particularly the erythrocyte lineage.

**Results**

Our experimental workflow is detailed in **Figure 1**. Enzymatic dissociation with gentle trituration of E11.5 (Theiler stage 19) embryos for 30-40 minutes resulted in a suspension of viable single cells (**Table XXX**). Thousands of cells were recovered for our whole embryo sample and for each of our microdissected samples comprising TS19 head, trunk, and heart (**Table XXX**) for a total of 33,941 single cells passing CellRanger quality control thresholds.

Based on an examination of the distribution of the number of unique molecular identifiers (UMIs) detected for single cells in each sample, we decided to model this parameter in order to establish a quantitative method to exclude libraries that were vastly different in quality from the majority of our single cells (**figure SXa**). A mixture of three normal distributions (3-normal mixtures, JMP software) best fit our data from each sample. The 3-normal mixtures model was then applied to each sample, and cells whose number of UMIs fell outside the 95% confidence interval of our middle distribution were excluded from subsequent analysis (**figure XX**). This resulted in a final dataset consisting of 26,305 total mouse embryonic single-cell transcriptomes with an average of 3756 UMIs and 1604 genes detected per cell (summary statistics by sample are shown in **Table XX**). By excluding both high UMI and low UMI containing cells,we are able to minimize the potential effect of the technical artifact of library quality on our subsequent analysis.

During our application of 3-normal mixtures modeling to exclude both very low and high UMI-containing cells from the analysis, we noticed two apparent populations of cells based on their number of unique genes detected relative to total library complexity. These populations could be visualized by plotting cells by their number of genes detected (nGene) versus the number of UMIs detected (nUMI, **figure XXXA**). We determined that this feature of our data was not due to library complexity (i.e. number of UMIs detected per cell) because these cells spanned the entire range of nUMI (**Figure XXB**), but was instead driven by the number of genes detected (**figure XXXC,D**). We hypothesized that the population of cells with lower transcriptome complexity (nGene) at a given library complexity (nUMI), herrin referred to as nGene/nUMI LO cells, could represent the more terminally differentiated cell types we would expect to find at this stage of embryonic development, including erythrocytes, various white blood cell types, and cardiomyocytes. Analysis of these cell populations based on sample of origin revealed that nGene/nUMI LO cells were present in every sample, but were heavily enriched in the cardiac sample (TS19.WE: 6.1%, TS19.HD: 5.9%, TS19.TR: 4.1%, TS19.HT: 51%, **figure XXXE-I**) Differential gene expression and gene set enrichment analysis using Gene Analytics showed that cells in the nGene/nUMI LO population were significantly enriched for hematopoietic and erythrocyte associated signatures in the GO:Biological Process category, including heme biosynthetic processes, erythrocyte development and differentiation, and embryonic hematopoiesis (p>0.0001, **Figure XXXJ**). Analysis of LM:Cell Types matching entities corroborated this functional enrichment, returning hematopoietic cell types proerythroblasts and erythroblasts (**Figure XXXJ**).

We found that nGene/nUMI LO cells have a much lower average expression of numerous markers of pluripotency than the rest of the cell population (**Figure XXXK,L**; p<0.0001; pluripotency markers in **List 1**). They are absent for the pluripotency markers *Nanog* and *Oct3/4*, and have very low average expression of *Sox2*. They are also negative for *Fgf4*, which is associated with pluripotency and has been shown to be antagonistic to hematopoiesis (**REF Nakazawa 2006**).

Of the markers for potency we assessed, nGene/nUMI LO cells exhibited higher average expression of *Abcg2*, which has been associated with hematopoietic stem cells (**REF Tang 2010, Fatima 2012**), and *Lefty1*, which is associated with heart development (**REF Meno 1998**). Taken together, these results indicate that the group of nGene/nUMI LO cells consists of some of the most differentiated cells in our dataset, and that these cells have erythrocyte and other hematopoietic cell type identity. **Discussion: hematopoietic system; why did we not find other differentiated cell types? What are the outliers in panel L? Maybe we should subcluster nGene/nUMI LO**

In order to cluster and examine populations of cells within our entire dataset, we first performed principal component analysis and extracted the top 100 principal components (PCs), representing the most variability in our dataset, from the whole transcriptome of all single cells. We then used t-distributed stochastic nearest neighbor embedding (tSNE) using these top 100 PCs as our input to generate a 3D projection of our 26,305 cells in this reduced-dimensional space (**Figure 2**). We colored this tSNE projection by sample of origin (E11.5/TS19 whole embryo, head, trunk, and heart, **Figure 2A**). As expected, cells from the whole embryo sample populate the entire structure of the combined data cloud and coregister with cells from each dissected sample (**Figure 2B**), giving us confidence that cells aren’t clustering solely due to a technical artifact related to sample of origin.

Due to the large diversity of cell types we expect in our sampling of the entire E11.5/TS19 embryo, we developed an iterative binary splitting approach employing unsupervised hierarchical clustering using Ward’s method to group cells based on their transcriptional similarity (see Materials and Methods). This iterative method of splitting our population of cells allowed us to better keep track of parent-child and sister-sister cluster relationships, and to keep track of cell cluster membership across all iterations of splitting. This allowed us to compare enrichment scores for GO:Biological Process, LM:Organs and Tissues, and LM:Cell Types between clusters from multiple “levels” of splitting, a critical step in helping to determine where along each branch of our IBS tree to stop splitting any particular cluster of cells. A survey of our IBS tree landscape showed that distinct and informative biological processes emerge beginning at the fifth iteration of binary splitting. A tSNE projection of the fifth iteration of binary splitting (n=32 cell clusters) is shown in **Figure 3A**. Analysis of cell clusters at this level of splitting revealed that many clusters are composed of cells from multiple samples of the dissected embryo (**Figure 3B**), indicating that clustering of cells by transcriptional similarity is not solely due to a technical artifact or an unknown sample bias.

In order to empirically help us determine where along each branch of our IBS tree to stop splitting a cluster of cells, we overlaid the maximum score for GO:Biological Process onto our tree topology (**Figure 4A**). This resulted in a “grapes and raisins” style plot wherein each node represents a cell cluster and is sized proportionally to the score of the cluster’s top match for GO:Biological Process (large, significant nodes resembling “grapes”, and smaller nodes resembling “raisins”). This helped inform our decision about when to stop splitting a given cluster of cells by examining the effect each split would have on its daughter’s matches to GO:Biological Process and giving us a visual guide indicating where the strongest matches to GO:Biological Process are found along each branch of our IBS tree (**Figure 4A**, inset **4B**). A plot of the maximum score for GO:Biological Process for all cell clusters at each iteration of binary splitting is shown in **Figure 4C**. We created a similar overlay of the maximum scores of matches from LM:Organs and Tissues and LM:Cell Types to bolster our analysis.

As an additional metric to inform our decision of when to stop splitting a particular cluster, we created a similarity matrix of parent and child clusters based on term frequencies for GO:Biological Process, LM:Organs and Tissues, and LM:Cell Types. We then analyzed these scores along the branches of our hierarchical binary splitting tree to find sibling cell clusters which had nearly identical functional enrichments. This helped us determine whether a split of a given cluster of cells is informative, or uncovers new biological signatures that might justify the splitting of that cluster.

In order to visualize the biological complexity of the tree generated by our iterative binary splitting of cell clusters, we generated word clouds representing each cluster’s matches to GO:Biological Process, LM:Organs and Tissues, and LM:Cell Types. We then created a browsable version of our tree using a javascript library, and represented cell clusters as nodes with an overlaid word cloud, in which phrases represent matching entities and are sized based on strength of score (**Figure 5, url to tool**). This gave us an accessible way to begin to analyze the biological signatures of our cell clusters and how those signatures change in response to the splitting of a particular cluster. The root node represents our entire population of 26,305 single cells, and the first iteration of binary splitting gives rise to clusters ‘1’ and ‘2’. For subsequent splits, the parent cluster is referenced in the names of the child clusters—the second iteration of binary splitting gives rise to clusters ‘11’ and ‘12’ from cluster ‘1’, and ‘21’ and ‘22’ from cluster ‘2’. The number of cells in the cluster and the score of the top entity in that cluster’s word cloud is also annotated.

We then examined our word clouds to assign a biological function to clades on our hierarchical split tree and to assign these clades a putative tissue type identity (**Figure 6**). This first-approximation analysis takes into account GO:Biological Process, LM:Organs and Tissues, and LM:Cell Types, and provides a good overview of the biological signatures detected in cell clusters from our analysis. Our strongest matches for GO:Biological Process included hematopoietic, cardiac, and angiogenesis-related signatures (**Figure 6**). We also observed strong neural, immune, muscle, and ectodermal matches, as well as numerous matches relating to neural crest-derived tissues such as enteric nervous system development and craniofacial development (**Figure 6**). The ultimate decision of when to stop splitting a cluster of cells is a combinatorial effort based on quantitative analysis of enrichment scores for GO:Biological Process, LM:Cell Types, LM:Organs and Tissues, similarity to sister and parent clusters based on term frequency, and biological interpretation of the matching functions and entities.

**Early lineage signatures in the E11.5 embryo**

In order to investigate the biological signatures recovered by whole-embryo, whole-transcriptome single-cell RNA sequencing, we projected expression of marker gene lists associated with early ectoderm, mesoderm, and endoderm lineages (**refs**). We observed strong regional expression of ectodermal signature by cells in one region of our tSNE projection (**fig**), largely separate from the more diffuse mesodermal signature (**fig**) and endothelial signatures (**fig**).

We found that the expression of many of these marker genes were coregistered with clusters identified through our clustering method. Furthermore, the biological functions extracted from differential gene expression and gene set enrichment analysis between clusters was in good agreement with projections of marker gene lists for specific germ layers, structures, and cell types, giving us further confidence in our dimension reduction and clustering approach (**figs**).

**Mesodermal signatures recovered by RNAseq at E11.5**

We then examined more specific gene signatures for cell types for which we would expect to find robust biological process matches at E11.5. We detected a strong signature for cardiomyocyte marker genes (**ref, table**) in a compact, dumbbell-shaped region in our tSNE projection, corresponding to cluster 121122 identified at the 6th iteration of binary splitting. Differential gene expression and gene set enrichment analysis on the top 2% of differentially expressed genes indicate that this cluster has a strong cardiomyocyte signature (**fig**) and most likely contains functional cardiac cell types such as cardiomyocytes, in addition to heart tube cells (**fig**).

**Neurogenic signatures recovered in the E11.5 embryo**

Based on established anatomical anatomy at this stage of development (**ref**), we expected to recover strong neurogenesis-associated signatures from the region of our tSNE corresponding to ectodermal lineage marker expression (**fig**). Projections of marker genes for neural precursors (**ref**) colocalized with a subset of our ectodermal-signature cells (**fig**). Furthermore, projection of markers for committed neuronal lineage cells were largely distinct in this tSNE space from glial-lineage signatures (**fig**). The signature for neuronal lineage was more clearly distinct from the more general neural progenitors than the glial lineage signature cells, in agreement both with the established observations that neurogenesis largely precedes gliogenesis, and that neural progenitor cells have markers characteristic of glial cells (**ref**). Projection of markers for retinal progenitor cells largely colocalized with neural progenitor signature cells, but also overlapped with signatures of neuronal lineage-committed cells (**fig**), consistent with the expected increase in degree of functional specification of cell types found in this tissue (**ref**).

In addition to neural precursors migrating radially from the ventricular and subventricular zones into the developing cortex, there is extensive tangential migration of interneuron precursors to subcortical areas from progenitor cells located in the ganglionic eminences (**ref**). Projection of genes associated with ganglionic eminences highlighted a discrete group of cells in the neuronal-committed space of our tSNE (**fig**). This group of cells corresponded with cluster 221211, identified at iteration six of iterative binary splitting (**fig**). Analysis of the top differentially expressed genes showed an enrichment for GO:Biological Process terms including “nervous system development” and, specifically, “globus pallidus development”. A signature for “glutamate decarboxylation” is also present in this cluster, and projection of *Gad1* and *Gad2*, hallmark genes of the GABA synthesis machinery in interneurons,colocalizes with a subset of cells in this cluster (**fig**) and separate from a distinct group of cells co-localizing with neuronal lineage markers, indicating that some some of these cells may represent a transient cell state or population, or early activation of these genes in interneuron precursor cells. The strongest match to the LM:Cell Types database is for “medial ganglionic eminence progenitors”. Additionally, this cluster overlapped with expression of the gene *Foxg1* (**fig**), a marker for the ventral telencephalic progenitor cells which give rise to the medial, lateral, and caudal ganglionic eminences (**ref**).

A subsequent iteration of binary splitting divided this cluster into two groups, in which cluster 2212111 retained a strong signature for MGE progenitor cells, while cluster 2212112 had an increase in its strength of match to lateral and caudal ganglionic eminence progenitor cells (**fig**). Therefore, we conclude that this iteration of splitting begins to enrich for MGE versus LGE/CGE progenitor cells, respectively, and demonstrate that iterative binary splitting is a useful way to keep track of such relationships between cell clusters in a large dataset.

In order to investigate whether our approach could recover distinct subpopulations of neural progenitor cells, we projected expression of *Tbr2* (*Eomes*), a marker of intermediate progenitor cells which give rise to cortical neurons, onto our cells in tSNE space. *Tbr2*+ cells mapped largely to the neuronal portion of our ectodermal-signature cloud, and represented a population of the cells identified in our 8th iteration cluster 22121211 (**fig**). Functionally-enriched biological processes from DGE/GSEA were associated with neuronal function, including “chemical synaptic transmission”, “regulation of transmembrane ion transport”, and, to a lesser extent, “nervous system development”. The association of this cluster with functions of more mature neurons (instead of more general nervous system developmental processes) may be due to the fact that not all of the cells in this cluster are *Tbr2*+ (n=78 out of 178 cells) and that this cluster may contain postmitotic neurons, indicated by the expression of *Tbr1* by many (n=137) of the cells in this cluster (**ref**). Furthermore, these intermediate progenitor cells are largely a transient amplifying population producing new postmitotic neurons, in which case *Tbr1*/*Tbr2* double-positive cells would not be unexpected and were found to be present (n=57) in this cluster.

We investigated a region of our tSNE dataspace which colocalized well with early ectoderm lineage markers, but poorly with markers of neural progenitors, neuronal lineage, and glial lineage (**fig**). Cells in this region were clustered together at the 5th iteration of binary splitting in cluster 21122 (**fig**). This cluster overlapped with projection of marker genes associated with surface ectoderm (**ref, lifemap**, **fig**). Gene set enrichment analysis on the top differentially expressed genes in this cluster were functionally enriched for GO:Biological processes related to skin development, including “epidermis development”, “tight junction assembly”, and “keratinocyte differentiation”.

**Materials and Methods**

**Animals**

Six week old CD1 mice were purchased from Charles River, housed and bred in our vivarium. Animal procedures were approved by the Institutional Animal Care and Use Committee at the University of Connecticut and conform to the animal care guidelines of the National Institutes of Health and the Association for Assessment and Accreditation of Laboratory Animal Care.

**Preparation of single-cell suspensions for droplet-based single-cell RNA sequencing**

CD1 mice were mated, females were sacrificed on embryonic day 11.5 and uterine horns were collected. Embryos were dissected from decidua and extraembryonic membranes were removed under a dissection microscope in ice-cold Dulbecco’s modified Eagle’s medium (Life Technologies). Embryos were assessed and developmental stage was assigned using anatomical features described by Theiler staging criteria (**REFS**). Theiler stage 19 embryos were rinsed with PBS and dissociated into single-cell suspensions using AccuMax (StemPro) at room temperature with gentle rocking and trituration using a P1000 filter micropipet. Single-cell suspensions were filtered through a pre-wetted 40µm nylon mesh cell strainer (Corning) and resuspended in 0.4% bovine serum albumin in PBS. Cell concentration and viability was assessed using 0.4% Trypan Blue dye to label dead cells on a Countess II automated cell counter (Life Technologies), and single-cell suspensions were confirmed by visual inspection of cells by microscopy and by examination of Countess images for clumps of cells and debris. Sample characteristics are summarized in **table XXX (we may not need this table or it can be a supplement)**.

**Droplet-based sequencing using 10x Genomics Single-cell 3’ Assay and Chromium controller**

Samples were loaded into individual channels of a 10x Genomics Chromium Single-cell 3’ Assay microfluidic chip and processed according to the manufacturer’s protocol. After loading cells and lysis/reverse-transcription reagents using the Chromium controller, emulsion quality was assessed by light microscopy. Reverse-transcription and breaking of the emulsions was performed according to the manufacturer’s protocol. Samples were indexed and libraries were prepared for sequencing using the Nextera XT kit (Illumina) according to the manufacturer’s protocol. Reverse-transcription and subsequent library quality was assessed using a 2100 BioAnalyzer and high-sensitivity DNA assay (Agilent Technologies). Sequencing was performed on an Illumina NextSeq 500 (high-output mode, 75 cycle reagent kit, read 1: 56 - transcript, Index 1: 14 – cell barcode sequence, Index 2: 8 – sample index, Read 2: 10 – unique molecular identifier). Sequencing data was demultiplexed and processed using CellRanger software (version 1.0, 10x Genomics).

**Gene expression data quality control, normalization, and dimension reduction**

Reads were mapped to the UCSC mm10 genome assembly and filtered gene-barcode matrices were generated for each sample using CellRanger (version 1.0, 10x Genomics) software and imported into R (version 3.3.2) and JMP (version 13, SAS) statistical software. Histograms of the number of UMIs detected per cell for each sample were analyzed by fitting models to the distributions using JMP, and the 3-normal mixtures model was selected for subsequent exclusion of cells with low library quality (indicated by low UMIs) and very high UMI containing cells, which we found to dominate the structure of our reduced dimensional data and skew subsequent clustering and analysis (data not shown), or could indicate potential cell doublets in our droplet-based sequencing preparation (**REF needed?**). Therefore, only single-cell libraries of similar quality were used in our downstream analysis.

The data tables for individual samples were concatenated and processed using the Cell Ranger R kit (version 1.0, 10x Genomics). Briefly, expression tables were pooled, genes that were not detected in any cells were removed, and gene expression for each cell was normalized by scaling the sum of UMIs detected in that cell toward the grand median number of UMIs detected for all cells. Principal component (PC) analysis was performed on the log2-transformed expression matrix using the base ‘prcomp’ function in R, and the top 100 PCs were used as input for t-distributed stochastic nearest neighbor embedding (tSNE) using the “Rtsne” package in R (**REFS for tSNE and package**). Summary statistics of the cells included in our final dataset are shown in **Table XX**.

**Comparison of library complexity, number of detected genes, and pluripotency markers**

Unsupervised clustering was used to split single cells into two groups based on their ratio of genes detected versus number of UMIs detected (nGene/nUMI) using Ward’s hierarchical method. To examine the properties of these cells, plots of nGene, nUMI, nGene/nUMI, and 2D tSNE were generated and the proportion of cells from each sample was calculated using JMP (**Figure XX**).

A curated list of 57 genes associated with pluripotency and early lineage commitment (**List 1**) was used to subset our gene expression data table. The log2-transformed expression values for all pluripotency markers were summed to create “pluripotency score” for each cell, and the distribution of these scores for nGene/nUMI LO cells and nGene/nUMI HI cells were compared using the Mann-Whitney *U* test (Wilcoxon rank sum test with continuity correction) in R.

**Differential gene expression and gene set enrichment analysis**

Differential gene expression was calculated for each cluster by comparing the mean gene expression within a cluster expression to its global mean expression, and the top 400 differentially expressed genes were used as input to GeneAnalytics (<http://geneanalytics.genecards.org>) for gene set enrichment analysis. GeneAnalytics reports were generated, and scores for significant matches to Gene Ontology (GO):Biological process, LifeMap (LM):Organs and Tissue types, and Cell Types were extracted.

In order to correct for over or under-representation/annotation of terms in the gene ontology database for gene set enrichment analysis, we ran 1000 permutations of random gene lists through GeneAnalytics to generate “background” datasets for GO:Biological Process, LM:Organs and Tissues, and LM:Cell Type.

For GO:Biological process, the scores for each term in the background dataset were averaged, and this was subtracted from the corresponding enrichment score in the foreground dataset for each cluster to generate a “background-subtracted” score for that term.

Because all entities in the LM:Organs and Tissues and LM:Cell Types categories were represented in the background dataset, we were able to calculate a z-score for each significant match by first subtracting the average background dataset score for a given entity from its score in the foreground dataset (as for GO:Biological Process), then dividing by the standard deviation of the background dataset scores.

Word clouds were generated from the matched terms for each category using JMP with entities sized proportionally to their scores. The resultant dataset, with word clouds arranged in a hierarchical tree format, can be browsed at (<http://flora.mcb.uconn.edu/dgetool/treeView/multiView.php>).

**Unsupervised clustering of cells using Iterative Binary Splitting**

Cells from all samples were pooled and clustered simultaneously using an unsupervised “iterative binary splitting” approach based on Ward’s hierarchical clustering method. Ward’s method was run on three tSNE components to produce two clusters of cells (level 1 split). These clusters were then independently re-clustered, again using Ward’s hierarchical method, each into two daughter clusters (level 2 split). This process was repeated for eight iterations, resulting in 256 clusters at the eighth level of splitting and 510 total clusters populating our hierarchical tree structure.

**Assignment of functional identity to cell clusters**

GO:term enrichment and cell/organ/tissue type scores were evaluated for each cluster at each level of iterative binary splitting by comparing the maximum scores for each category along each branch of the tree. This allowed us to compare the effect of each split on enrichment scores between parent and daughter cell clusters and to inform our decision/determine at which level we should stop splitting the cells into separate clusters along a given branch.

Pairwise text-comparison was performed on the term frequency tables used to generate word clouds from cell cluster gene set enrichment analysis, and similarity matrices of GO:Biological Process, and Lifemap:Cell Type, and LifeMap:Organ and Tissue-matching terms were generated. By clustering these matrices using Ward’s hierarchical method, similarities and differences between parent and sibling cell clusters can be examined based on cluster distance. This was used in order to help determine whether a split of the cluster at a given iteration is informative based on similarity of terms between parent-children and siblings.

Analysis of the significance of enrichment scores, term similarity between parents and siblings for GO:Biological Process and Lifemap:Organs/Tissues/Cells, and examination of differentially expressed genes were used for biological interpretation and functional inference of cell cluster identity.

**Word Vomit: misc scratchpad for topics to hit in introduction, discussion, etc.**

**Introduction**

Understanding the development of a mammalian embryo from fertilization to birth is a monumental task, but is essential for the progression of research related to developmental diseases and translational stem cell therapies. While there is already significant interest in generating an atlas of adult human cell types (REF Human Cell Atlas and Chan-Zuckerberg Biohub) and intense interest in mapping the adult mouse brain (REF), we believe the creation of a single-cell atlas of development with lineage information in a model mammalian organism has enormous potential to revolutionize health-related experimental research and treatment of disease, much like the *C. elegans* lineage map revolutionized our understanding of cell fate commitment in the roundworm (ref).

The critical first step in generating a mammalian lineage map is to establish an atlas of the distinct cell types present at single snapshots of embryonic development. Herein we describe a global survey of cell types in the mouse embryo during organogenesis at Theiler stage 19, categorized based on single-cell gene expression profiles. This is the first attempt at a large-scale single-cell gene expression atlas in a developing mammalian embryo, and will provide a crucial framework for further investigation of embryonic cell typing, cell lineage progression, and changes in cell population hierarchy through multiple stages of development.

Profiling the cell types of the embryo requires balancing throughput/multiplexing with an assay that can adequately resolve each distinct cell type/closely related cell type present at each stage of development. Transcriptional profiling is an invaluable tool, and availability of next-generation technologies such as microfluidic or droplet-based single-cell isolation and chemistry now finally allow for cost-effective, large-scale single-cell RNA sequencing.

Transcriptional profiling/single-cell RNA-seq has been used successfully to identify/parse out distinct/differentiated cell types from the adult brain/fully differentiated organs (**REF Linnarsson, Tasic et. al., etc.**), as well as used to identify cell fate in the very early embryo (**REF Robson etc., Peng et. al. 2016,**) and identify new progenitors and subpopulations of specialized differentiated human white blood cells (**REF Villani**)

The development of computational methods/analysis tools for massively multiplexed single-cell data from highly diverse samples is lagging. Current methods underperform

Analysis of whole-embryo, whole-transcriptome single cell expression data during organogenesis presents many challenges due to the diversity of the cell population, which includes few fully differentiated cell types and many heterogeneous populations of transient progenitors that exist only during development. We have chosen E11.5 (Theiler stage 19) embryos for our initial analysis of cell types. At this stage of mouse organogenesis, we expect to identify several transient populations of progenitors, as well as a few differentiated cell types, based on gene expression profiles. We expect to identify transient cells of hematopoiesis as this process shifts from the primitive progenitors of the extraembryonic yolk sac to the definitive hematopoietic progenitors generated by the embryo proper, including those originating in the aorta-gonad-mesonephros (ref). We also expect to capture cells with unique neurogenic signatures, as spatially-constrained restriction of gene expression in the developing forebrain, which is beginning to regionalize, has been reported at this time (REF), and we expect to detect distinct gene signatures related to these specific populations of progenitor cells.

We propose a unique application of unsupervised hierarchical clustering of cells based on iterative, binary splits of our cell population into 2*n* clusters. For each iteration of binary splitting and reclustering, we are able to keep track of parent-child relationships between cell clusters and build a “hierarchy of splitting” represented by a binary tree. Using our entire population of cells as a root node, we can keep track of cell membership in clusters along each “branch” of our tree, from root to leaves (the *n*th iteration of binary splitting). By examining a combination of metrics related to gene set enrichment analysis on differentially expressed genes in clusters at each iteration of splitting, such as scores related to Gene Ontology:Biological Process, along with data-driven metrics such as the percentage of globally distinguishing differentially expressed genes shared between sister nodes, we can determine where a given split of cells into two clusters is not informative and choose that node to represent a “final” cluster of cells. In this manner, we can curate a list of biologically informative nodes which best represent clusters of similar “cell type.”

**Table XXX: Sample characteristics and summary of CellRanger output:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | Cells recovered | Percent viability | Number of cells | Number of reads | Mean reads per cell | Median genes per cell | Median UMI per cell | CDNA PCR duplication (%) | Percent reads in cell |
| TS19.WE | 4500000 | 66 | 8071 | 406452880 | 50359 | 1759 | 3767 | 68.3 | 63.8 |
| TS19.HD | 690000 | 80 | 11488 | 284803242 | 24791 | 1346 | 2418 | 49.7 | 68 |
| TS19.TR | 860000 | 71 | 7117 | 311229588 | 43730 | 1910 | 3937 | 60.4 | 67 |
| TS19.HT | 800000 | 80 | 7265 | 293577040 | 40409 | 1186 | 5384 | 63.3 | 91.2 |

**Table XX: Sample summary statistics after 3-normal mixtures UMI filtering of cells:**

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Number of cells | Median genes per cell | Median UMI per cell |
| TS19.WE | 6273 | 1788 | 3598 |
| TS19.HD | 9164 | 1391 | 2425 |
| TS19.TR | 5456 | 1942 | 3888 |
| TS19.HT | 5412 | 1120 | 5085 |

**List 1: List of pluripotency-associated genes** used in the comparison of nGene/nUMI LO versus HI cell populations:  
Tpbg  
Abcg2  
Acvr1b  
Acvr2b  
Alpl  
Cdh1  
Cbx2  
Cd9  
Tnfrsf8  
Kit  
Cdx2  
Chd1  
Cfc1  
Dnmt3b  
Dppa2  
Dppa4  
Dppa5  
Epcam  
Esrrb  
Gm7325  
Fbxo15  
Fgf4  
Fgf5  
Foxd3  
Gbx2  
Nr61a  
Gdf3  
Itga6  
Itgb1  
Klf4  
Klf5  
L1td1  
Lefty1  
Lefty2  
Lin28a  
Lin28b  
Trim71  
Maf  
Myc  
Nanog  
Pou5f1  
Podxl  
Zfp42  
Smad2  
Smad3  
Sox2  
Fut4  
Stat3  
Dppa3  
Suz12  
Tbx2  
Tbx3  
Tbx5  
Tex19.2  
Tex19.1  
Thap11