Material and methods draft

**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.

**Embryo harvest and single cell preparation**

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. ONE SENTENCE HERE ABOUT HOW 4 SAMPLES WERE DISSECTED (NUMBERS, PART, ETC.). 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).

**Primary data analysis from 10X cellranger goes here**

* Primary QC in cellranger
* Reads mapping and UMI counts statistics
* UMI matrix generation

**Secondary data analysis goes here**

**Model based quality control and normalization**

Besides primary quality control provided by 10X Genomics pipeline cellranger, we analyzed histograms of numbers of UMIs detected per cell from each sample. From all models tested, we selected the 3-normal mixtures model to exclude cells with low library quality. We believe the lower normal distribution derived from bad library indicated by low UMIs, while cells in the upper normal distribution could indicate potential doublets from sample preparation and skew subsequent analysis (data not shown). We removed all cells below lower 95% tail of lower distribution, and all cells above upper 95% tail of upper distribution. Therefore, only single-cell libraries of similar quality were used in our downstream analysis. We removed all non-expressed genes and pooled all samples by concatenating data tables for individual samples, which finally gave us a UMI matrix with dimension of 26305 cells by 20347 genes. To normalize UMI counts, we used median normalization by scaling the sum of UMIs detected in that cell toward the grand median number of UMIs detected for all cells, followed by a log transformation of log(1+UMI).

**Dimension reduction and clustering**

Principle Component (PC) analysis was applied to extract first 100 principle components using basic R function ‘prcomp’. STATISTICS ABOUT TOP 100 PCS(like percent variance explained). To visualize data, we further used R package ‘Rtsne’ (REF) to apply t-distributed stochastic neighbor embedding (tSNE) on first 100 PCs, extracting 3 tSNEs. Summary statistics of the cells included in our final dataset are shown in Table XX.

To cluster the cells, we used hierarchical single linkage clustering (HSLC) to extract 512 clusters in tSNE space. We removed all clusters with less than 50 cells (N cells in total removed), and finally yield 53 clusters.

NOT SURE IF WE INCLUDE IBS:

To study the substructure of big clusters from HSLC, we also used an unsupervised “iterative binary splitting” (IBS) 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. As a flexible algorithm, we were able to match our HSLC clusters to a group of IBS clusters with good concordance, which provides potential to study large HSLC clusters.

**Differential Gene Expression (DGE) analysis and Gene Set Enrichment Assay (GSEA)**

We used R package ‘Moncole’ to extract differentially expressed genes (REF). Each cluster was compared to whole population, genes with q-value >0.05 and fold change <2 were dropped. Top 400 genes passing the threshold were used in GSEA. If less than 400 genes extracted, then all those genes were used.

We used database from LifeMap tool GeneAnalytics (<http://geneanalytics.genecards.org>) for GSEA. For each cluster, scores of related matches for LM:Cell Types, GO:Biological Processes (GO:BP) and LM:Phenotypes 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:Phenotype, and LM:Cell Types. 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: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.

**Assignment of functional identity to cell clusters (NEED TO WORK ON THIS)**

Cell type annotation for all clusters were made based on a mixture of evidences. We analyzed high score matches from GSEA analysis on different categories including LM:Cell Types, GO: Biological Process and LM:Phenotypes. We used marker genes from LifeMap database and R&D database. We also used percent sample origin as a reference.

**LifeMap cell lineage mapping**

To map cell types to lineage map, we used lineage map from LifeMap. We simplify the lineage map by removing related anatomical compartment and tissue/organ information, which finally gave us 628 cell types. Cytoscape was used to generate network view of trimmed lineage map. For each possible cell type, we extracted related gene expression information from LifeMap database. We score our clusters for all cell types by calculating the sum normalized UMI of marker genes (probably explain LM marker gene system of 2 and 3). Then we normalize all scores by calculating z-scores. In this case, for each cluster we got z-scores for all cell types. We are able to project the scores on Cytoscape network to visualize dense region of high score cell types.