ClusterMap: Compare analysis across multiple Single Cell RNA-Seq profiling

Xin Gao, Deqing Hu, Hua Li

Abstract

Single cell RNA-Seq facilitates the characterization of cell type heterogeneity and developmental processes. Further study of single cell profiling across different conditions enable the understanding of biological processes and underlying mechanisms at sub-population level. However, proper methodology to compare multiple scRNA-Seq datasets remain missing. Here we developed ClusterMap, a systematic method and workflow, to facilitate the comparison of scRNA profiling across distinct biological contexts. With hierarchical clustering of the marker genes of each sub-groups, ClusterMap match the sub-types of cells across different samples. We introduce a new tree cut method that is designed to be felicitous for this matching problem. Utilized circos plot and regroup method visualize the results concisely. Furthermore, single value separability is defined to summarize the sub-population changes among all sample pairs. With three case studies, ClusterMap exhibits the ability to gain more insight into the different molecular mechanisms of the cellular sub-population across different conditions. ClusterMap is implemented in R and available at https://github.com/xgaoo/ClusterMap.

Keywords: sub-population, match, tree cut, separability, visualization

Introduction

Single cell RNA sequencing (scRNA-Seq) is an advanced technology that shed a light on the high-resolution heterogeneity and dynamics of transcriptome profiling. Following massive studies on the cell sub-types under static condition, researchers start to pursue the high resolution mechanisms of different biological processes. With more complicated experimental design, such as under different treatment conditions, different developmental time points, or different tissue types, researchers become interested in the heterogeneous changes of sub-population.

There are many existing methods and packages for cellular sub-types identification, developmental trajectory and differential expression analysis in single cell expression analysis (1–10). However, they are all restricted to one single cell dataset, except for batch effect correction and a couple of methods (11–14). There is an increasing demand for comparing multiple scRNA-Seq datasets across biological conditions, while the system analysis methodology is still missing. The current common strategy is combined analysis of all the datasets, which will loss the many critical information at each sub-type level. Here we designed a package ClusterMap to analyze and compare two or more single cell expression datasets and it provides the whole comparison analysis workflow to automatically address the following questions.

First, which pair of the sub-groups that should be compared? In different datasets, how the sub-groups in each dataset matches each other? The heuristic and common way is to check the expression of one or a few known marker genes to match the corresponding sub-groups (15–17). This manual step bias the matching towards several genes and might not be accurate. ClusterMap provides a method to match sub-groups automatically and unbiasedly, based on a list of marker genes detected in the analysis. It also makes the sub-group matching possible

even if there is no known maker gene in advance. In addition, the matching is not limited to be one-to-one match. With a certain cutoff, the sub-group can be a singleton without any matches, which is an extreme case of the sub-type change.

Second, after matching, how do the sub-group of cells change under different situations? The difference can be population size and properties. ClusterMap visualizes the matched pairs, the similarity of pairs and the cell percentage difference in a Circos plot. The t-Distributed Stochastic Neighbor Embedding (t-SNE) plots are re-colored to display the paired groups. The property change of the matched groups are estimated by separability. We define the single value separability to characterize the condition effect for each sub population. This single value measurement allowed quick and unbiased detection of most affected groups of cells among multiple groups and sample pairs.

Overall, ClusterMap provides easy-to-use and reliable methods and workflow to compare multiple single cell RNA-Seq datasets with all kinds of experimental design: across various treatments, across time points, as well as multiple tissues. We provide three case studies to explain the usage and advantage of ClusterMap. 1) Analyzed the scRNA-Seq of mammary glands epithelial cell across different estrus phase. 2) Compared the scRNA-Seq of immune stimulated PBMCs with the control. 3) Compared replicates of PBMCs datasets as a negative control. With ClusterMap, we coordinated the sub-groups in different conditions precisely, which allowed us to detect the difference and distinct sub-groups of cells in more accurate manner. This method helped us obtain a deeper understanding of the biological processes at the cellular sub-population level through the accurate matching and close comparison.

Methods

Workflow overview

ClusterMap focuses on the analysis of sub-group matching and comparison across single cell RNA-Seq samples. Thus ClusterMap is based on the pre-analysis of each single dataset. The summarized workflow of ClusterMap is shown in Figure1.

The first step of ClusterMap is taking the marker genes for each sub-group of each sample as the basic input to perform the matching. Matching is based on the existence (binary) of marker genes, which will overcome the batch effect. With binary hierarchical clustering and our special tree cut method, pre-defined sub-groups will be matched and grouped together with the most similar sub-groups in different samples. New group labels will be assigned according to the matching results. The similarity of the matched groups are extracted based on the clustering results, which measures the confidence of the matching. The marker genes can be generated by the Seurat package (1), Cell Ranger or other single cell analysis methods.

The second step is to visualize the matching results in multiple views. If the cell number of each sub-group is provided, ClusterMap can make Circos plot to summarize the results (Figure 3E). The cords link the matched sub-groups, while the cord color codes the different new groups and the transparency of the cord color indicates the similarity of the matched groups. The width of the black sectors represent the percentage of cell numbers in each sample. Thus, the sub-population size change is intuitive by comparing the sector size of linked

groups. Furthermore, if the t-SNE plot coordinates are provided for each cell, ClusterMap will re-color the plot to coordinate the colors for the matched groups (Figure 4A-C). This will facilitate the matching of sub-groups in the pre-analysis. ClusterMap can handle the Seurat object directly to perform the re-coloring.

The last step is to characterize the property changes across samples for each matched groups. We defined Separability to quantify the difference based on the distance of the K-nearest neighbors of inna- and inter- samples (Figure 3F). Separability is defined for a pair of samples. Pairwise separability will be measured if more than two samples. Separability enables the quick searching of the most affected cell sub-types among all the sample comparisons.

Following ClusterMap, differential expression analysis for most affected group might be the common choice to investigate the difference further. Since many methods, such as DESeq2, SCDE, BASiCS and so on (9, 18–20), have been well established, this part of analysis is not included in ClusterMap.

Binary hierarchical clustering

The sub-groups were matched by hierarchical clustering (Figure 3D). The existence (binary) of the marker genes in each group were used to measure the distance between the sub-groups across the samples. The union of all marker genes were considered together. Hierarchical clustering were performed using the average linkage and binary distance. The similarity of the matched groups is defined as 1 minus the height of the merging node of the matched groups in the dendrogram (Figure 3F). It measures the percentage of marker genes that overlap between the matched groups, number of intersections divided by the number of unions of marker genes.

Since the tree cut algorithm defined below will match a group with its most close group, groups with low similarity might be grouped together. With the similarity reported as part of the result, we keep all matched groups without filtering. It's reasonable to filter out the low similarity groups and treat them as unmatched groups for the downstream analysis. Note that, refine marker gene list might improve the similarity.

Purity Tree Cut

Our tree cut algorithm is designed to match the most similar sub-groups from different samples as much as possible, while avoid large cluster of sub-groups from the same sample. Thus, the algorithm decompose the dendrogram bottom-up based on both the distance of branches and the purity of the node. The traditional clustering assessment methods, such as Elbow method, Silhouette index, Dunn index or other indexes, are not optimal for this purpose. The algorithm decides to keep or cut a given node by checking three aspects. First, the purity of the node. Second, the edge length of the two branches. Third, the overlap of the sub-samples of two branches. Once all the nodes are checked, the sub-nodes of each kept node become a new group.

In a dendrogram, sample set of a node is a set of samples that all of its offspring nodes come from. If all the offspring nodes of a node are from the same sample, only one sample in its sample set, then the node is a pure node and considered the same as singleton. For example, in Figure 2D, the offspring of node n1 are S1_1 and S1_2. They are from sample S1 and S2. Thus the sample set of node n1 is {S1, S2} and n1 is not a pure node. For node n2.1.1, the

offspring S2_1 and S2_3 are from the same sample S2, thus n2.1.1 is a pure node and is considered as a singleton (Figure 2D).

Edge length in a dendrogram is the height difference between the upper node (such as N) and the lower node (such as n1 in Figure 2D). An edge length cutoff controls if two branches of a node should be merged as one group or not. If an edge is longer than the cutoff, then the branch is quite different from the other branch, it should not be merged. The heuristic cutoff is 0.1, 10% of the marker genes are different.

We search through all the nodes in the dendrogram from the bottom. For a given node N with two direct sub-nodes n1 and n2, and edges E1 and E2, N will be kept only under three conditions:

1. n1 and n2 are both singleton or pure (Figure 2B).
2. n1 is singleton or pure, AND the edge E2 is shorter than the edge cutoff, AND n2 is not cut based on the sub-nodes of n2 (Figure 2C).
3. Both edge E1 and E2 are shorter than the edge cutoff, AND the two sample sets SG1 and SG2 are not included by each other ($SG1 \not\subset SG2$ AND $SG2 \not\subset SG1$), AND none of n1 or n2 is cut (Figure 2D).

In all other conditions, node N will be cut, n1 and n2 are forming two groups. The algorithm is shown in Figure 2A. Take Figure 2D as an example, node n1 is under condition1, S1_1 and S2_2 are matched and form a new group. Node n2.1.1 is considered as singleton, thus node n2.1 is also under condition1, S2_1, S2_3 and S3_3 form one matched group. Node n2 is under condition2, but if the edge between n2 and n2.1 is longer than the cutoff, n2 will be cut, S1_2 can not be grouped together with the other branch. Node N violates conditon3, because SG1 is a subset of SG2, no matter if n2 is cut or not or the length of E1 and E2.

We generated a random tree with 4 samples and 10 sub-groups in each sample to test our tree cut algorithm. A simulated tree is shown in the Figure S1A. The results were as expectation (Figure S1B). If we loosed the edge cutoff, more sub-groups tended to merge into bigger groups but with lower similarity (Figure S1C). The following case studies also suggested that the tree cutting results matched the biological meaning.

Separability

For each new group, the Separability is defined as the median difference of inna- and inter- sample distance of each cell in the combined t-SNE coordinates (Figure S1D). The inna-sample distance is defined as the median distance of K-nearest cells of the same sample. The inter-sample distance is the same but cells of different sample. Using median instead of mean will reduce the variation due to the outliners. Increase K will improve accuracy but slow down the computation. Practically, K > 20 will improve the accuracy just slightly.

In another word, for a new group, given a cell $\mathbf{C_i^{(1)}}$ from sample1, $\mathbf{C_k^{(1)}}$ is one of the K nearest cells of $\mathbf{C_i^{(1)}}$ from sample1, $\mathbf{C_k^{(2)}}$ is one of the K nearest cells from another sample, sample2.

Inna-sample distance for cell x is

$$\mathbf{D_i^{Inna} = median_k \, || \, C_i^{(1)} - C_k^{(1)} \, ||}$$

Inter-sample distance for cell x is

$$\mathbf{D_i^{Inter} = median_k \| C_i^{(1)} - C_k^{(2)} \|}$$

The separability of sample1 vs sample2 in this new group is

$$\mathbf{SEP1 = median_i \ (D_i^{Inter} - D_i^{Inna})}$$

The same for cell $\mathbf{C_j^{(2)}}$ from sample2,

$$\mathbf{SEP2 = median_j \ (D_j^{Inter} - D_j^{Inna})}$$

Then the separability of sample1 vs sample2 in this new group is defined as

$$\mathbf{Separability = mean\ (SEP1, SEP2)}$$

Case study

Epithelial cells of different estrus cycle

To compare the sub-population change in two different biological phases, we apply ClusterMap to the epithelial datasets that were generated in the study of Pal et al. (15). Cells were collected from mammary glands of adult mice at different phases of the estrus cycle. By pooling the glands from two mice, scRNA-Seq of 2729 total epithelial cells in estrus and 2439 cells in diestrus were performed using the 10X Chromium platform.

The pre-analysis was performed before ClusterMap to define sub-groups of cells in the diestrus and esturs states separately (Figure 3A, 3B). We obtained 13 groups in diestrus and 10 groups in estrus. Marker genes for each of the 23 groups were called as part of the pre-analysis. To allow for the unbiased match of the sub-groups, we didn't label the cell type of each group by known markers at the beginning. To match the 13 groups to the 10 groups, ClusterMap first clustered the sub-groups based on all the marker genes (Figure 3C), and the clustering dendrogram was obtained as a result (Figure 3D). Then, the dendrogram needed to be decomposed by deciding which sub-groups should form a matched new group based on the similarity and the purity of the sub-groups (Methods). With the specific tree cutting algorithm of ClusterMap, the sub-groups were matched into 12 new groups, and the similarity of matched groups were reported as well (Figure 3F). Note that, new group 2 and 3 are one to multiple match. There is no matched group for the new group 7, 10, 11, 12 at current edge cutoff (Methods).

The results were summarized in the circos plot (Figure 3E). The 13 black sectors highlighted by the red sector represent 13 groups in diestrus as in Figure 3A, while the 10 sectors highlighted by the green sector represent 10 groups in estrus as in Figure 3B. The width of the black sectors represents the percentage of cells in each sample. Matched groups are linked by cord, and the transparency represents the similarity of the matched groups, the darker the similar. ClusterMap also recolored the t-SNE plots of each sample (Figure 4A, 4B) and combined sample with the new groups (Figure 4A, 4B, 4C). Matched groups are shown in the same color that we can recognize the matched sub-groups clearly. The same colored cells were clustered together in the combined sample, which confirmed our matching results. The cell percentage change of the new group 2, 3 and 9 were obvious in both circos plot and recolored t-SNE plot. The separability of each new group further emerged the most affected sub-types, such as group 6, 8 and 3, besides the no matched groups (Figure3F).

There are three main sub types of the mammary gland epithelial cells, including basal, mature luminal (ML) and luminal progenitors (LP). The known markers for basal cells are Krt14, Acta2, Myl9, Sparc, Mylk, and Cxcl14. Luminal cells express Elf5, Prlr, Areg, Ly6d, Stc2, Krt19, and mature luminal express Pgr, Prlr, Cited1, Esrrb, and Cxcl15 (15). Based on the known markers, we assigned group 2, 5, 6, 10 to mature luminal, group 1, 11, 12 to luminal progenitors, group 9 to luminal intermediate (Lum Int), and group 3, 4, 7, 8 to basal cells (Figure S2A, 4A-D).

In line with Pal's observation, we also noted that ML becomes two major sub-types and the Lum Int was substantially reduced in the diesturs phase. However, although group 2 and 10 are all mature luminal in diestrus based on the known markers, group 2 is much similar to the ML in estrus (group 2). Group 10 is a distinct ML sub-type from the ML in estrus. This is also obvious in the t-SNE plot of the combined analysis (Figure 4C, 4D). While group 2 is a mixture of both samples, group 10 is exclusively from the diestrus phase. Pal et. al. observed a hint as well that group 2 is "one of the clusters tightly associated with ML signature genes such as PgR". The same situation happened to the LP population. Group 1, 11 and 12 are all luminal progenitors, but group 11 and 12 are specific sub-types in diestrus (Figure 4A-D). This is different from Pal et al's conclusion that LP population was unaltered. The basal population was bigger in diestrus. Besides the change of the population size, there are much more details happened to each sub-type. Group 3 in basal was increased substantially in estrus, while group 7 is missing (Figure 4A-D).

Through the gene ontology and pathway analysis of the marker genes of the new groups (21), we found that group 7, 10 and 11 were extremely enriched for ribosome biogenesis, oxidative phosphorylation and metabolic process of ribonucleotides (Figure 4E-F, S2C-D). These findings reflect the increased level of progesterone in diestrus, which functions as a potent mitogen to stimulate expansion of mammary epithelia at this stage. It remains unclear about why only subpopulations of basal, ML and LP in diestrus stage could respond to progesterone to undergo rapid cell cycle progression but not other subtypes of the same identity, potentially suggesting the existence of differential regulation of mitogen-related cell signaling among sub-populations of the same cell type. Intriguingly, we noted Group 10 cells in diestrus are associated with the development of mammary gland alveoli, indicating some of cells within this group possess trans-differentiation potential for alveolar development (Figure 4F)(22). In addition, we observed a drastic increase in number of group 3 cells with characteristics of smooth muscle in estrus, indicating there may be a contractile switch of myoepithelial cells for lactation preparation at this period (Figure S2D).

PBMCs under immune stimulus

To compare the effect on the cellular populations across different treatment, we next applied ClusterMap on the datasets that were generated from the study of Kang HM, et al. (16). A pool of peripheral blood mononuclear cells (PBMCs) from eight lupus patients was studied. PBMCs from each patient were untreated as control or activated with recombinant interferon-beta IFN-β for 6 hours. 14,619 cells for control and 14,446 cells for stimulated were obtained from 10X Chromium instrument sequencing.

The pre-analysis is shown in Figure S3 to define sub-groups for each sample and find marker genes for each group. We obtained 11 and 13 sub-groups for the control and stimulated conditions. ClusterMap matched most sub-groups except the new group 10, 11 and 12 (Figure

5C). By checking the known marker genes (Figure S3C-D), the matching results were confirmed that the matched groups expressed same known marker genes. Based on the known marker genes, we recognized that group 1, 2, 3 are CD4+ T cells (IL7R), group 5 are B cells (MS4A1), group 6 are CD14+CD16– monocytes (CD14, LYZ), group 4 are CD14+CD16+ monocytes (FCGR3A, MS4A7), group 7, 8 are dendritic cells (FCER1A, CST3), and group 11 are erythroblasts. Group 9 is a mixture of natural killer cells (GNLY, NKG7) and CD8+ T cells (CD8A). After the IFN-β stimulation, these two groups of cells becomes more distinguishable (Figure 5D-E). For the group 10 and 12, although both groups express some megakaryocytes marker genes, such as PPBP, other marker genes did not match each other very well (Figure 5A arrows). Thus, ClusterMap didn't group them together. There is no obvious change of the cell percentage of each cell type after stimulation, which is consistent with Kang et al.'s conclusion. According to the separability of each pair, a global change was observed as Kang's observation. However, the global effect is not even across sub-groups. T cell group1 was less affected. The IFN-β stimulation affect monocytes and dendritic cells much more than the other immune cells (Figure 5C red dashed boxes). These are also shown in the combined analysis (Figure 5F-G).

PBMCs replicates

As a control of our method, we used ClusterMap to compare two replicate datasets, which supposed to match each other exactly with no difference observed. The 4K PBMCs and 8K PBMCs datasets were downloaded from the 10X Genomics support datasets. They are peripheral blood mononuclear cells (PBMCs) from the same healthy donor. The two datasets are replicates with different detected cell numbers, about 4000 and 8000 cells.

The group matching results demonstrate that the sub-types in 4K PBMCs matches perfectly with 8K PBMCs one-to-one (Figure 6). The similarity between matched groups are also much higher than other two datasets (Figure S4D). The cord color in the circos plot is much darker (Figure 6B). Majority of the cell percentage of sub-types are not changed. The separability demonstrates that there is no obvious difference between the two samples in any of the matched pairs (Figure S4D). These fulfill the expectation as 4K and 8K PBMCs are replicates.

Discussion/ Compare to other methods

Pre-analysis

ClusterMap rely on the pre-analysis results. The clustering analysis of each single sample and the marker genes of each sub-group will affect the comparison results. It's better to define meaningful sub-groups for each sample first before start comparison. The regroup step in ClusterMap will refine the clustering based on the matched results, such as merge some similar sub-groups in the same sample.

The clustering analysis may be performed in the combined sample as well. The resolution will be gained by increasing the cell numbers of pooled dataset. However, the new clustering results will be hard to match to the sub-groups in each single sample. The regroup results in ClusterMap keep the grouping information in the single samples. It's also making

sense to compare the match defined new groups and the clustering analysis defined new groups in the combined sample.

Binary distance

We use the binary value of marker genes instead of the expression level, because the expression of a gene is not comparable between different samples even after normalization. The relative expression level of a gene across cells or within a cell may variate a lot under different conditions, but a gene is recognized as a marker of the specific sub-group or not is more robust. This is the same rational as checking one or two known markers. The binary value also tolerate the global shift of transcriptome during matching. The prerequisite is the accurate identification of marker genes in the pre-analysis step. The binary distance as a consequence becomes more reasonable than the Euclidian distance.

Similarity and separability

Similarity is how well the pair match each other compare to the other groups based on the marker genes. Separability is how much the group property changes between the pair within a paired groups. They can both reflect the relationship of the paired groups, but at different aspect of view. Usually the more similar the less separate for a pair. However, it is possible that the pair have similar marker genes, but separate far apart (Figure5G, H, group 6). This is due to that we use binary value of marker genes to compare between groups, but use the expression level, transformed into TSNE coordinates, to compare cells within paired groups. Smaller groups with higher variance of marker genes tends to merge at lower similarity.

Batch effect correction

There are many study on comparison of multiple scRNAseq datasets from different batches, such as mutual nearest neighbors method (MNNs, (11)), distribution-matching residual networks method (12). The goal of these type of comparison is to remove the difference, while ClusterMap is to identify the difference.

For the sub-group matching step of ClusterMap, the batch effect will not affect the matching results. Because the marker genes are relative to the groups in the same sample, matching based on the existence (binary) of marker genes will overcome the batch effect. The batch effect only affects the separability of ClusterMap.

We didn’t remove batch effect in our case studies. One of the important reason is that batch effect and biological effect are always confounded. The Seurat package recently updated with canonical-correlation analysis (CCA), which can remove the common source of variation by decomposition on correlation matrix (13). They also applied the CCA to the same immune stimulated dataset (16). By aligning each cell, the samples are scaled to become similar to the most extend. Thus the global difference in the original datasets were treated as batch effect instead of biological. However, the IFN-β stimulus was expected to trigger a widespread immune response of PBMCs (16). It is hard to determine the effect is due to biology or technique. Beside, we observed uneven effect of IFN-β on different sub-groups, which is not obvious in the CCA aligned results.

To compare, we also performed CCA on the epithelial cell datasets (Figure S5). The two samples mixed more evenly under canonical correlation vector space. The two separated groups of mature luminal (ML, marked by Prlr) in destrus phase was not distinguishable unless

we highlighted them based on grouping in Figure 4 (Figure S5D). This indicated that CCA reduced the difference between sub-groups even in the same sample. CCA had difficulty to align sample specific cells. Butler et al. observed this effect as well for the rare population and suggested to use PCA for further analysis (13). However, the sub-group 10 in the diesturs phase (Figure 4A, 4C) was not a rare population. With CCA only, we may miss these important sub-groups. CCA and ClusterMap provide different view of comparing multiple scRNA-Seq datasets.

Another reason is that if samples are under good control, there is no obvious batch effect across different runs of the 10X Chromium, which is shown in the 4K and 8K PBMCs datasets. We also observed that in the epithelial cell datasets, the two samples in group 1 and 2 are overlapping much better than the other groups. This is a good internal control suggesting that the batch effect is not significant across the samples (Figure 4C, 4D). The batch effect might be more prominent across platforms. If necessary, the separability can be applied after the batch effect correction.

Compare to scmap

scmap (14) is a method mapping one scRNA-seq sample to the reference dataset to annotate the new dataset. scmap-cluster map each cell to the cluster centroid in the reference as a classification method. scmap-cell map each cell to the nearest cell in the reference. Both scmap and ClusterMap aim to match sub-types of cells from different samples. ClusterMap map at the cluster level while scmap map at the single cell level. Due to this different point of view, ClusterMap use marker genes of each sub-group for the mapping, and scmap use random selection, highly variable genes, genes with higher number of dropouts, different number of genes, and different similarity calculation to find the agreed mapping. Scmap emphasize on the similarity of one sample to the reference to annotate the new sample from the same origin, and ClusterMap focus on the difference across multiple samples under different treatments or conditions. ClusterMap can compare multiple samples at the same time, not limited to one to one map. Our method also provides the separability to quantify the changes across samples besides the cell percentage change shown in scmap.

Availability

The R package of ClusterMap with the tutorial are deposited at GitHub repository (https://github.com/xgaoo/ClusterMap). The code and results for the pre-analysis of the datasets in this study are also included.

The epithelial datasets and immune stimulated datasets were downloaded from Gene Expression Omnibus under accession number GSE103272 and GSE96583. The PBMCs replicates datasets were downloaded from the 10X Genomics support datasets (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k and https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k).

Figure Legends

Figure 1

Scheme of the workflow of ClusterMap.

Figure 2

Scheme of the algorithm of the dendrogram tree cut algorithm. A. Purity tree cut algorithm. B-D. Examples of the three conditions in the tree cut. B. Two singleton. C. One singleton. D. No singleton.

Figure 3

Cluster match of mammary gland epithelial cells at different phases of the estrus cycle. A. Pre-analysis of cells from diestrus phase. B. Pre-analysis of cells from estrus phase. C. Heat map of the hierarchical clustering of sub groups by the existence of marker genes. Each column is a marker gene for one of the sub groups. D. Dendrogram of the hierarchical clustering. E. Circos plot of the matched sub groups. The 13 black sectors highlighted by the red sector represent 13 groups in diestrus as in Figure 3A, while the 10 sectors highlighted by the green sector represent 10 groups in estrus as in Figure 3B. The width of the black sectors represents the percentage of cells in each sample. Matched groups are linked by cord, and the transparency represent the similarity of the matched groups, the darker the similar. F. ClusterMap results for the quantification of the sample comparisons.

Figure 4

Regroup of mammary gland epithelial cells. A. Re-colored t-SNE plot based on matching results for diestrus phase. B. Re-colored t-SNE plot based on matching results for estrus phase. C. Re-colored t-SNE plot based on matching results for the combined dataset. Matched groups were recolored in the same color and with the same label through all the three t-SNE plots. D. t-SNE plot with cells colored by sample. Based on the known markers (Figure S2A-B), we generally assigned group 2, 5, 6, 10 to mature luminal (ML), group 1, 11, 12 to luminal progenitors (LP), group 9 to luminal intermediate (Lum Int), and group 3, 4, 7, 8 to basal cells. E-F. Gene ontology and pathway analysis for the new marker genes of group 2 and 10 in combined sample (C) using Metascape.

Figure 5

ClusterMap analysis for the immune stimulated datasets. A-G. Output of ClusterMap, similar as in Figure 3 and 4, but for the immune stimulated datasets. Based on the known marker genes, we assigned group 1, 2, 3 to CD4+ T cells, group 5 to B cells, group 6 to CD14+CD16– monocytes, group 4 to CD14+CD16+ monocytes, group 7, 8 to dendritic cells, group 10, 12 to megakaryocytes, group 11 to erythroblast and group 9 is a mixture of natural killer cells and CD8+ T cells.

Figure 6

ClusterMap analysis for the PBMCs replicates datasets.

Bibliography

1. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.

2. Hart,Y., Sheftel,H., Hausser,J., Szekely,P., Ben-Moshe,N.B., Korem,Y., Tendler,A., Mayo,A.E. and Alon,U. (2015) Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods*, **12**, 233–5, 3 p following 235.

3. Pierson,E. and Yau,C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 241.

4. Haghverdi,L., Buettner,F. and Theis,F.J. (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, **31**, 2989–2998.

5. Angerer,P., Haghverdi,L., Büttner,M., Theis,F.J., Marr,C. and Buettner,F. (2016) destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, **32**, 1241–1243.

6. Ji,Z. and Ji,H. (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, **44**, e117.

7. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

8. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M., *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.

9. Kharchenko,P.V., Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

10. Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

11. Haghverdi,L., Lun,A.T.L., Morgan,M.D. and Marioni,J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 10.1038/nbt.4091.

12. Shaham,U., Stanton,K.P., Zhao,J., Li,H., Raddassi,K., Montgomery,R. and Kluger,Y. (2017) Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, **33**, 2539–2546.

13. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

14. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.

15. Pal,B., Chen,Y., Vaillant,F., Jamieson,P., Gordon,L., Rios,A.C., Wilcox,S., Fu,N., Liu,K.H., Jackling,F.C., *et al.* (2017) Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat Commun*, **8**, 1627.

16. Kang,H.M., Subramaniam,M., Targ,S., Nguyen,M., Maliskova,L., McCarthy,E., Wan,E., Wong,S., Byrnes,L., Lanata,C.M., *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.

17. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, **8**, 14049.

18. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and' ' dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

19. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.

20. Vallejos,C.A., Richardson,S. and Marioni,J.C. (2016) Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.*, **17**, 70.

21. Tripathi,S., Pohl,M.O., Zhou,Y., Rodriguez-Frandsen,A., Wang,G., Stein,D.A., Moulton,H.M., DeJesus,P., Che,J., Mulder,L.C.F., *et al.* (2015) Meta- and Orthogonal Integration of Influenza “OMICs” Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe*, **18**, 723–735.

22. Visvader,J.E. (2009) Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes Dev.*, **23**, 2563–2577.

Figure 1

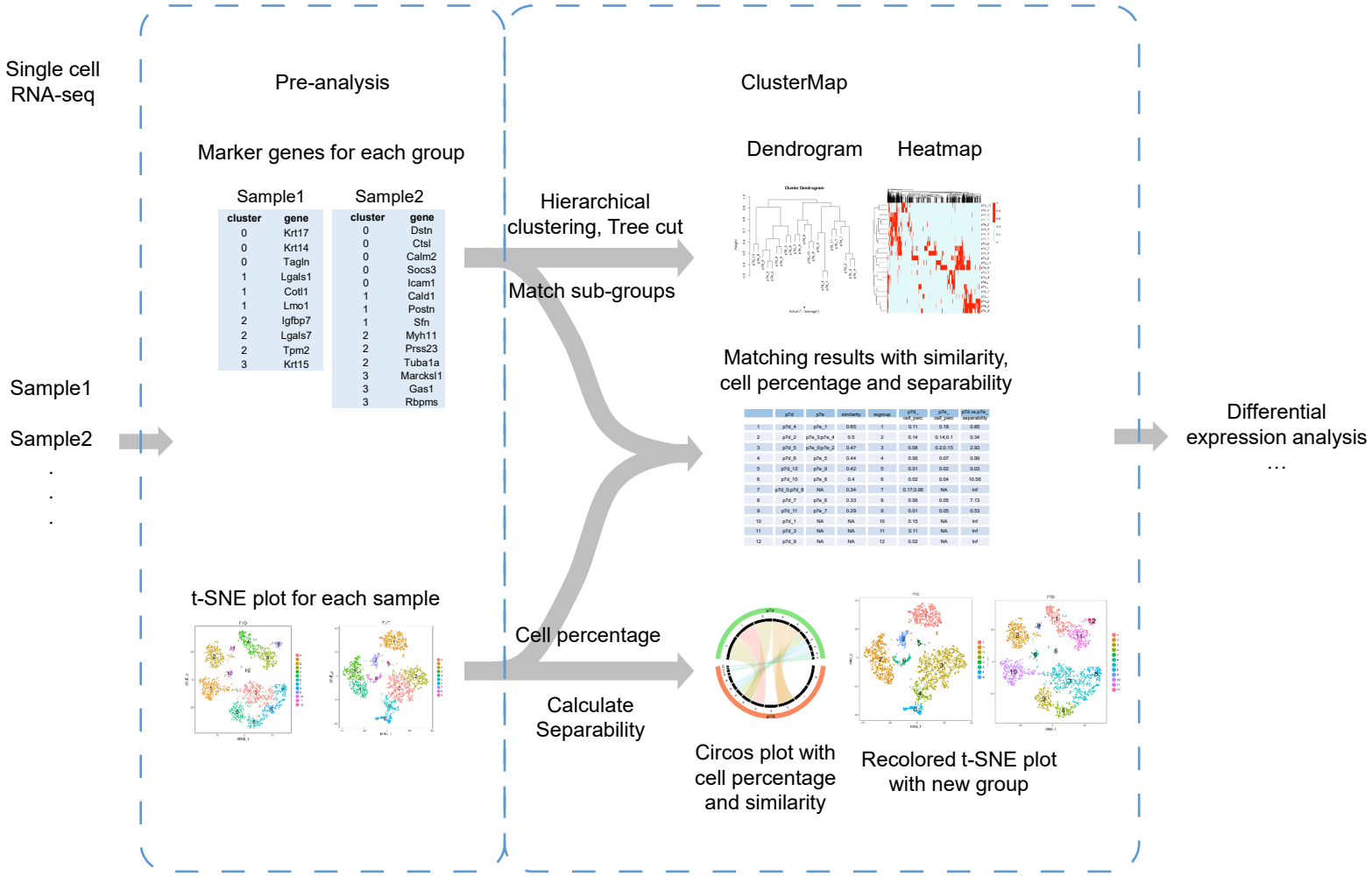
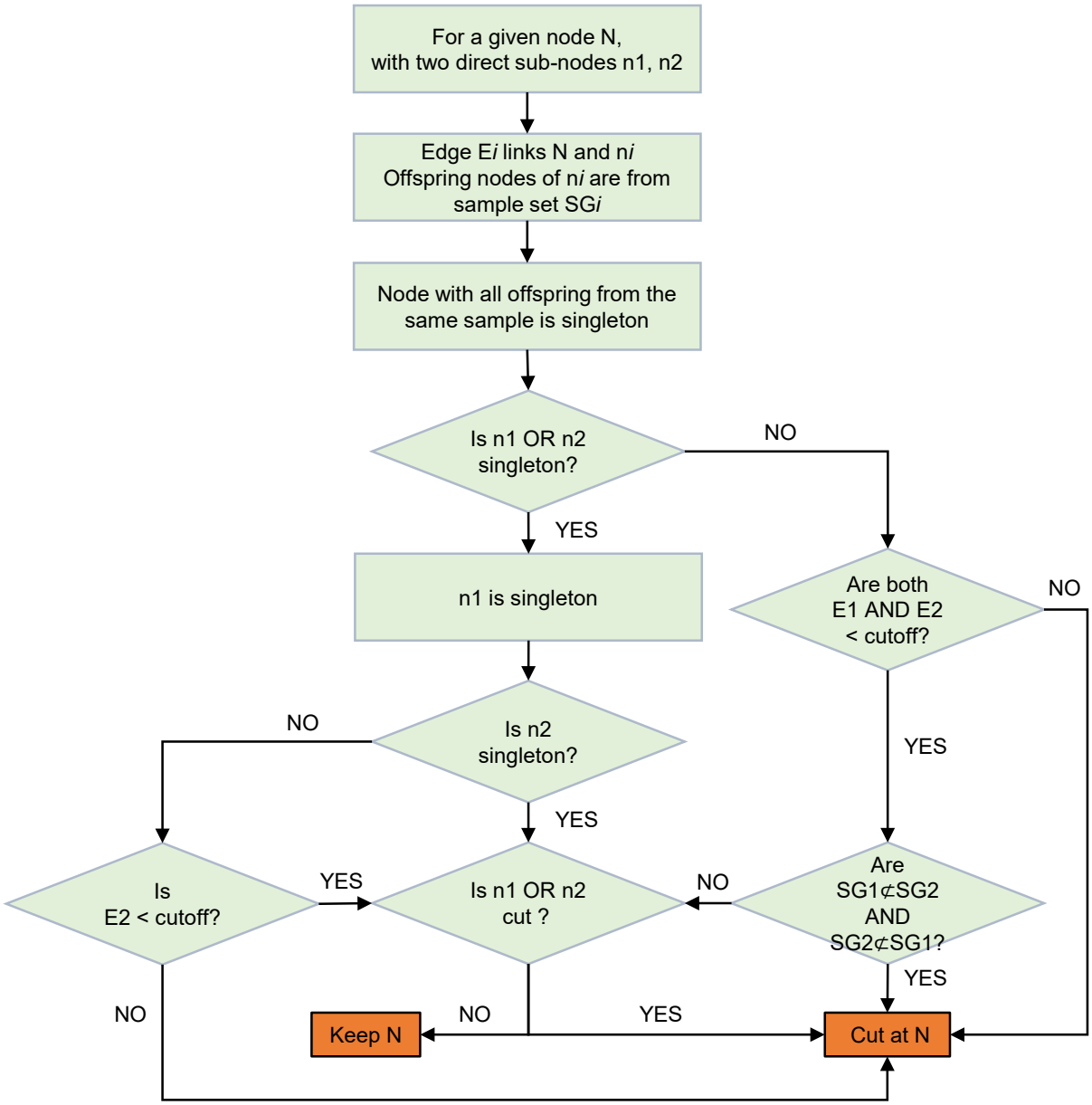
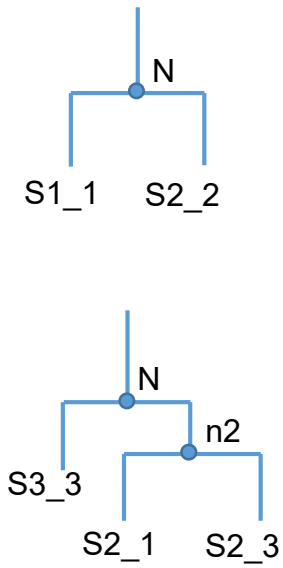

Single cell RNA-seq
Sample1
Sample2
.
.
.
Pre-analysis
Marker genes for each group
Sample1
Sample2
t-SNE plot for each sample
Hierarchical clustering, Tree cut
Match sub-groups
Cell percentage
Calculate Separability
ClusterMap
Dendrogram
Heatmap
Matching results with similarity, cell percentage and separability
Circos plot with cell percentage and similarity
Recolored t-SNE plot with new group
Differential expression analysis
…

Figure 2

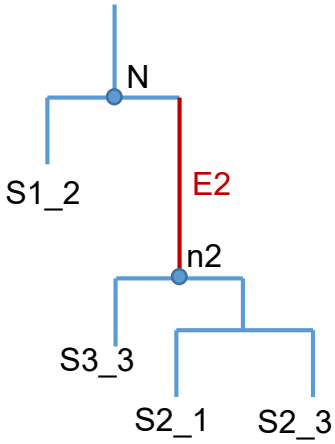
A



B Both singleton



C One singleton



D No singleton

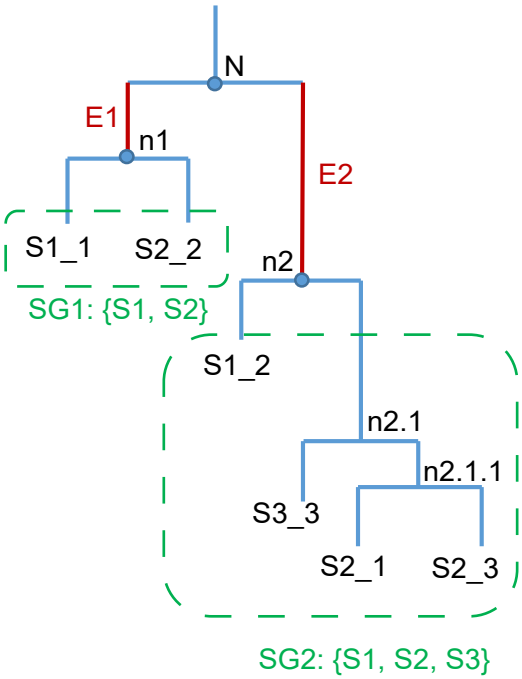
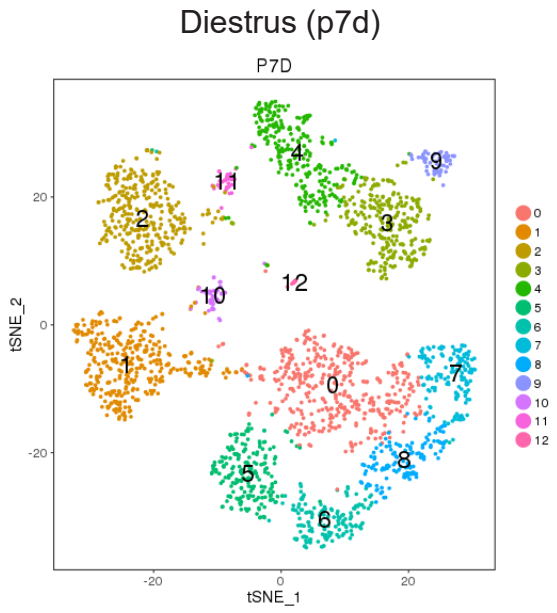
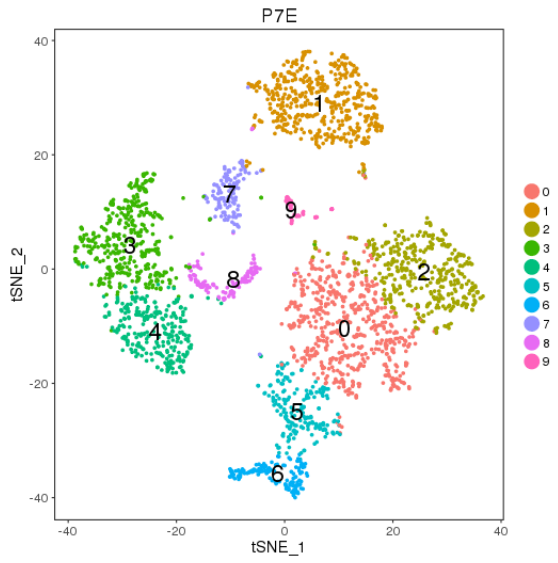


Figure 3

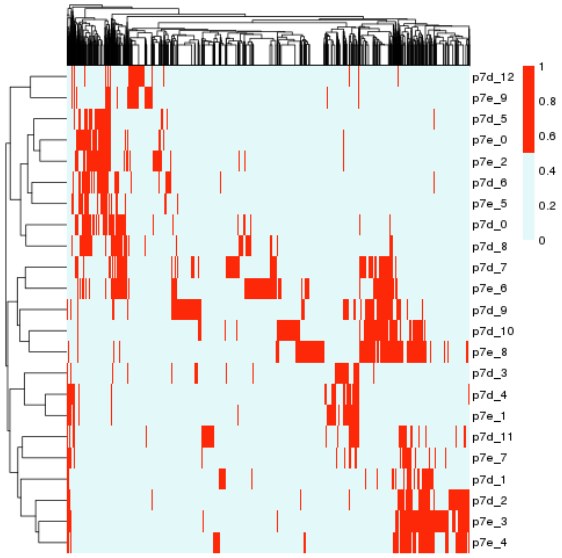
A



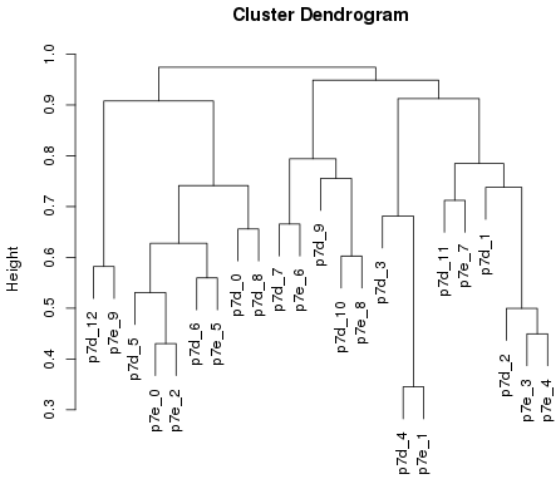
B



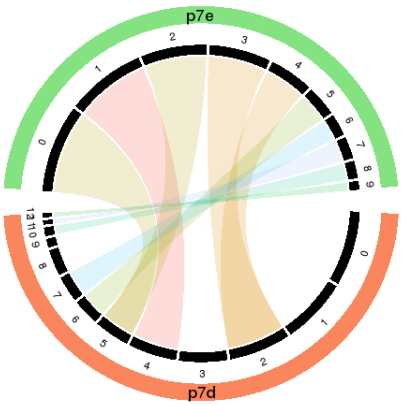
C



D



E



F

	p7d	p7e	similarity	regroup	p7d_ cell_perc	p7e_ cell_perc	p7d.vs.p7e_ separability
1	p7d_4	p7e_1	0.65	1	0.11	0.18	0.85
2	p7d_2	p7e_3;p7e_4	0.5	2	0.14	0.14;0.1	0.34
3	p7d_5	p7e_0;p7e_2	0.47	3	0.08	0.2;0.15	2.93
4	p7d_6	p7e_5	0.44	4	0.06	0.07	0.99
5	p7d_12	p7e_9	0.42	5	0.01	0.02	0.03
6	p7d_10	p7e_8	0.4	6	0.02	0.04	10.56
7	p7d_0;p7d_8	NA	0.34	7	0.17;0.06	NA	Inf
8	p7d_7	p7e_6	0.33	8	0.06	0.05	7.13
9	p7d_11	p7e_7	0.29	9	0.01	0.05	0.53
10	p7d_1	NA	NA	10	0.15	NA	Inf
11	p7d_3	NA	NA	11	0.11	NA	Inf
12	p7d_9	NA	NA	12	0.02	NA	Inf

Figure 4

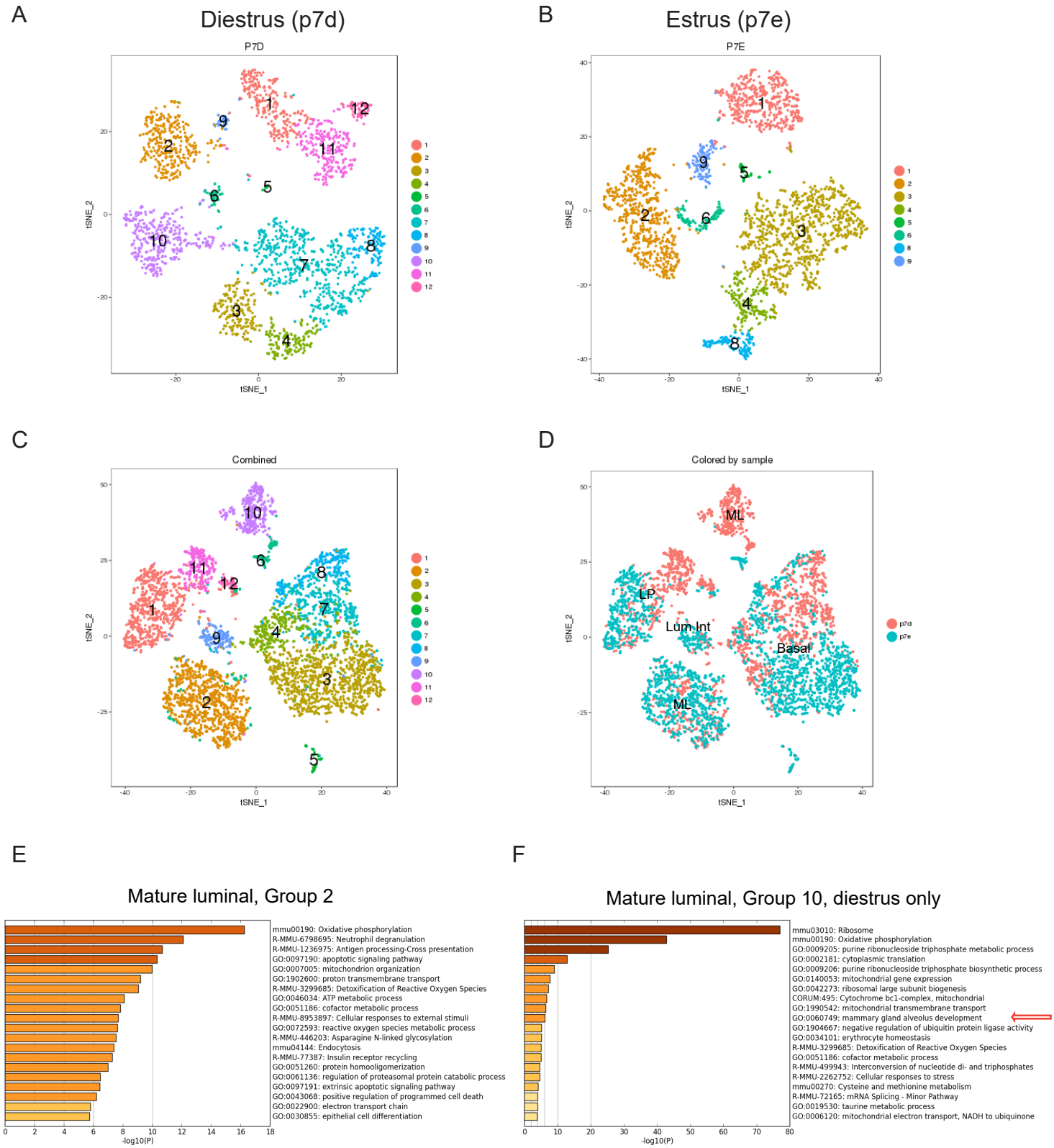
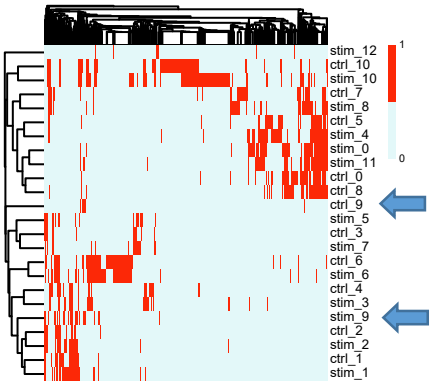
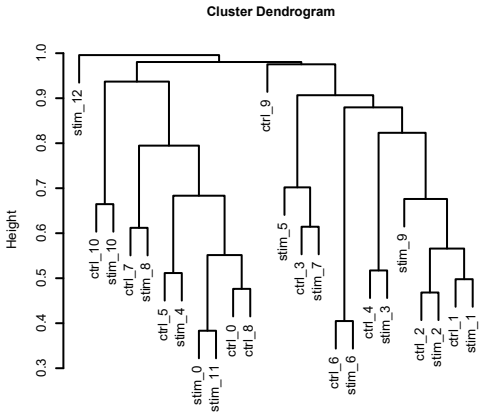

A
Diestrus (p7d)
P7D
tSNE_1
tSNE_2
B
Estrus (p7e)
P7E
tSNE_1
tSNE_2
C
Combined
tSNE_1
tSNE_2
D
Colored by sample
ML
LP
Lum.Int
Basal
ML
p7d
p7e
tSNE_1
tSNE_2
E
Mature luminal, Group 2
mmu00190: Oxidative phosphorylation
R-MMU-6798695: Neutrophil degranulation
R-MMU-1236975: Antigen processing-Cross presentation
GO:0097190: apoptotic signaling pathway
GO:0007005: mitochondrion organization
GO:1902600: proton transmembrane transport
R-MMU-3299685: Detoxification of Reactive Oxygen Species
GO:0046034: ATP metabolic process
GO:0051186: cofactor metabolic process
R-MMU-8953897: Cellular responses to external stimuli
GO:0072593: reactive oxygen species metabolic process
R-MMU-446203: Asparagine N-linked glycosylation
mmu04144: Endocytosis
R-MMU-77387: Insulin receptor recycling
GO:0051260: protein homooligomerization
GO:0061136: regulation of proteasomal protein catabolic process
GO:0097191: extrinsic apoptotic signaling pathway
GO:0043068: positive regulation of programmed cell death
GO:0022900: electron transport chain
GO:0030855: epithelial cell differentiation
-log10(P)
F
Mature luminal, Group 10, diestrus only
mmu03010: Ribosome
mmu00190: Oxidative phosphorylation
GO:0009205: purine ribonucleoside triphosphate metabolic process
GO:0002181: cytoplasmic translation
GO:0009206: purine ribonucleoside triphosphate biosynthetic process
GO:0140053: mitochondrial gene expression
GO:0042273: ribosomal large subunit biogenesis
CORUM:495: Cytochrome bc1-complex, mitochondrial
GO:1990542: mitochondrial transmembrane transport
GO:0060749: mammary gland alveolus development
GO:1904667: negative regulation of ubiquitin protein ligase activity
GO:0034101: erythrocyte homeostasis
R-MMU-3299685: Detoxification of Reactive Oxygen Species
GO:0051186: cofactor metabolic process
R-MMU-499943: Interconversion of nucleotide di- and triphosphates
R-MMU-2262752: Cellular responses to stress
mmu00270: Cysteine and methionine metabolism
R-MMU-72165: mRNA Splicing - Minor Pathway
GO:0019530: taurine metabolic process
GO:0006120: mitochondrial electron transport, NADH to ubiquinone
-log10(P)

Figure 5

A



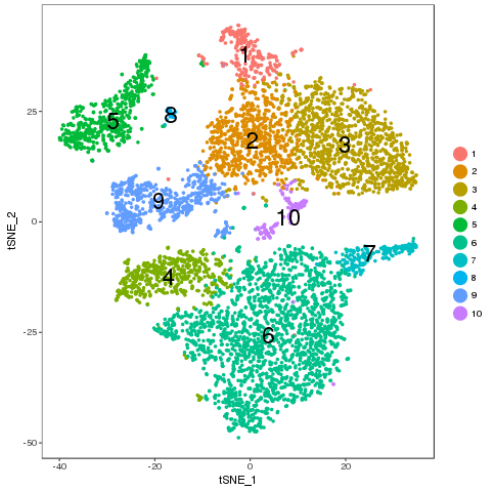
B



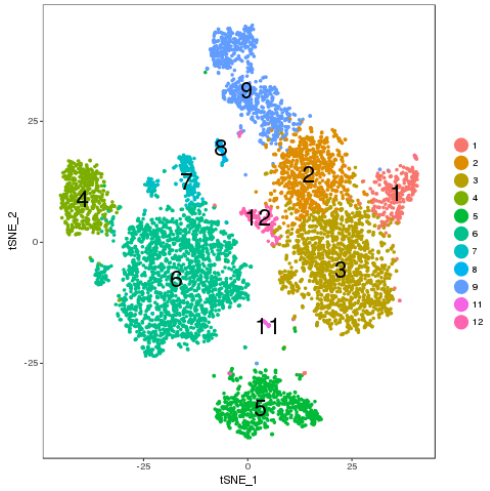
C

	ctrl	stim	similarity	regroup	ctrl_ cell_perc	stim_ cell_perc	ctrl.vs.stim_ separability
1	ctrl_6	stim_6	0.6	1	0.04	0.04	0.24
2	ctrl_2	stim_2	0.53	2	0.12	0.11	0.55
3	ctrl_1	stim_1	0.5	3	0.18	0.21	0.78
4	ctrl_5	stim_4	0.49	4	0.08	0.07	5.01
5	ctrl_4	stim_3	0.48	5	0.09	0.11	0.55
6	ctrl_0;ctrl_8	stim_0;stim_11	0.45	6	0.31;0.02	0.28;0.01	11.03
7	ctrl_7	stim_8	0.39	7	0.03	0.03	7.05
8	ctrl_10	stim_10	0.34	8	0.01	0.01	0.55
9	ctrl_3	stim_5;stim_7	0.3	9	0.09	0.07;0.04	0.37
10	ctrl_9	NA	NA	10	0.02	NA	Inf
11	NA	stim_12	NA	11	NA	0	Inf
12	NA	stim_9	NA	12	NA	0.02	Inf

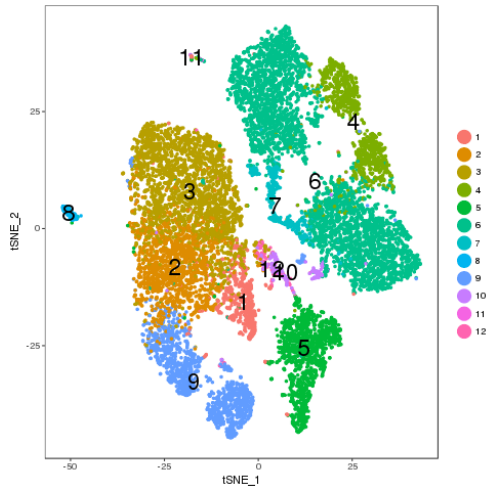
D



E



F



G

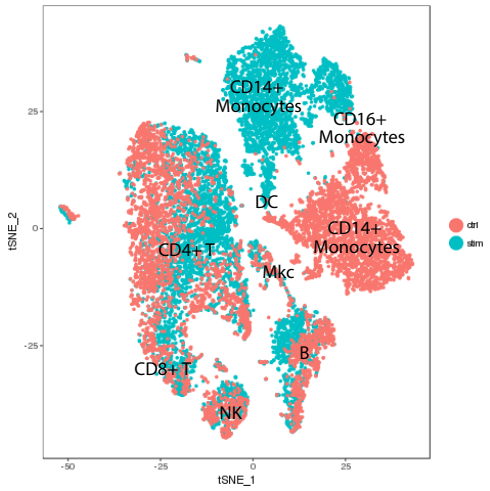
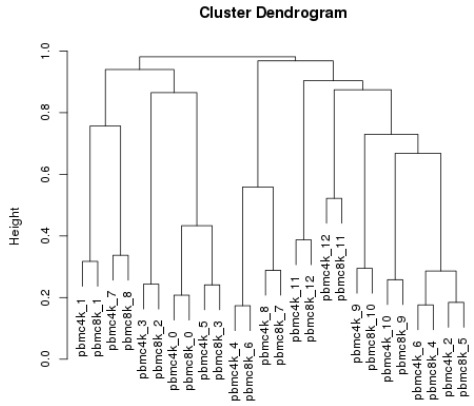
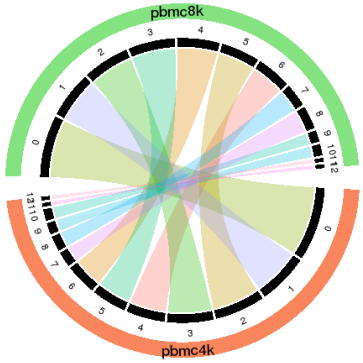


Figure 6

A

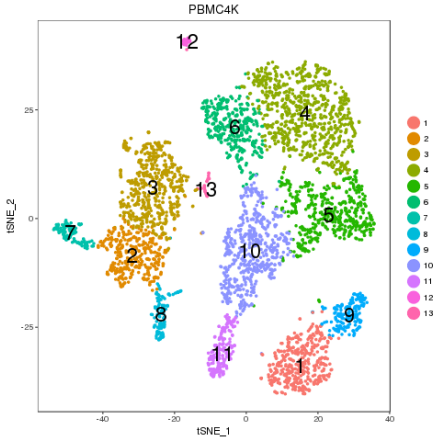


B



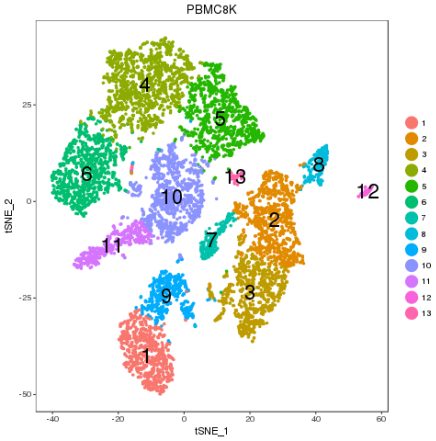
C

PBMC 4K

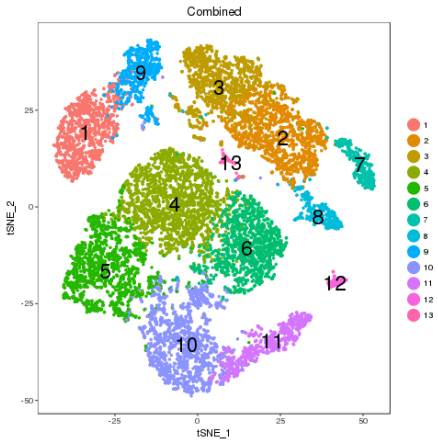
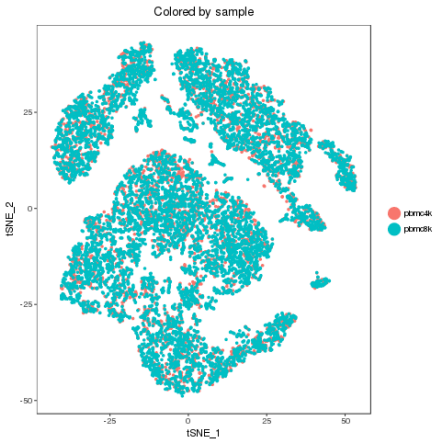


D

PBMC 8K



E



F