A case study on the detailed reproducibility of a human cell atlas project

(Supplementary Materials)

Kui Hua[1,2], …, Xuegong Zhang[1,2,3,*]

[1]MOE Key Laboratory of Bioinformatics Division and Center for Synthetic & System Biology, BNRIST, Beijing 100084, China

[2]Department of Automation, Tsinghua University, Beijing 100084, China

[3]School of Life Sciences, Tsinghua University, Beijing 100084, China

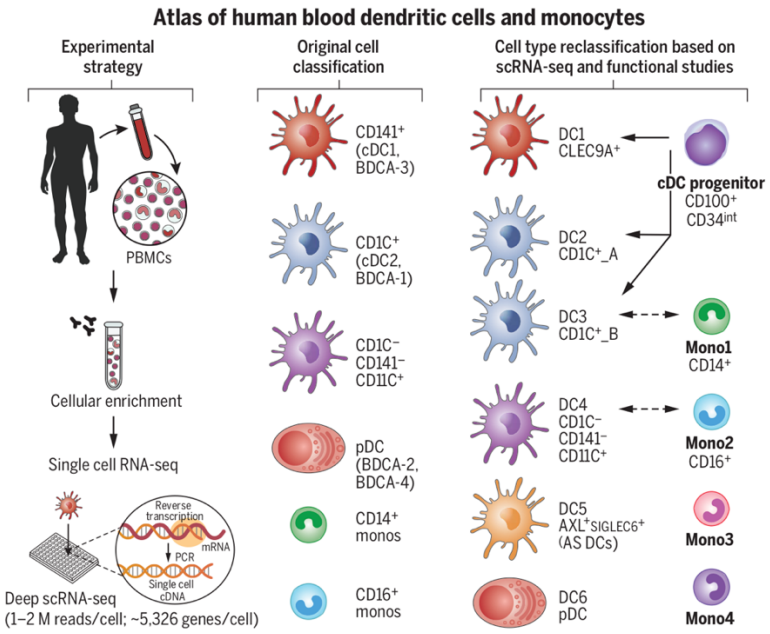
* Corresponding author. E-mail: zhangxg@tsinghua.edu.cn.

1. Introduction

This is the Supplementary Materials of the case study on the detailed reproduction of the work in the paper (Villani *et al.*, Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors, *Science,* vol. 356, no. 6335, 2017). This file contains records of technical procedures of the reproduction experiments. More details of the bioinformatics processing with live code are available at https://github.com/XuegongLab/HCA-reproducibility.

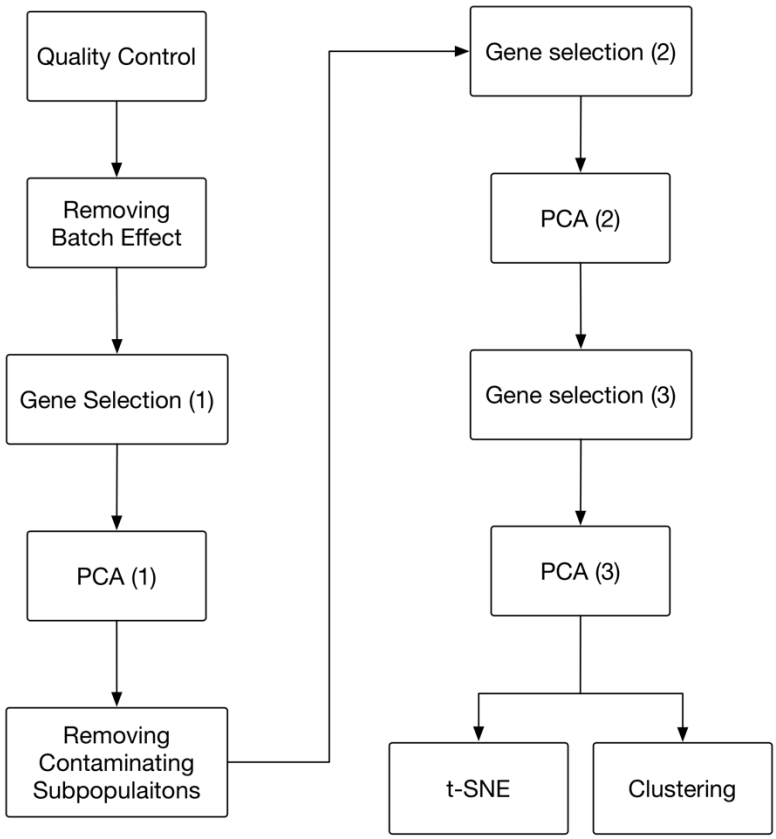
The selected paper revised the taxonomy of human blood dendritic cells (DCs) and monocytes by performing a single-cell study of ~2400 cells. By analyzing single-cell RNA-seq data of 768 DCs and 372 monocytes derived from a single healthy people (in total 1140 cells, referred to as ***discovery dataset*** in the following discussion), they identified a new subtype of DC, a new subdivision of a previous known subtype of DC, the existence of a conventional dendritic cell (cDC) progenitor and two additional subtypes of monocytes. The existence of these cell types is further confirmed in a subsequent study of additionally profiled ~1200 cells (referred to as ***validation dataset***). They also studied the function of these newly discovered cell types and revealed the relationship between some of them.

Here we focus on the analysis of dendritic cells (DCs) in the discovery dataset. Most of the analyses are performed using the R software package **Seurat** as the original paper did.



2. Workflow

Here is the workflow of the analyzing procedure that we extracted from the given code. We followed this workflow step by step in our reproduction.



3. Reproduction

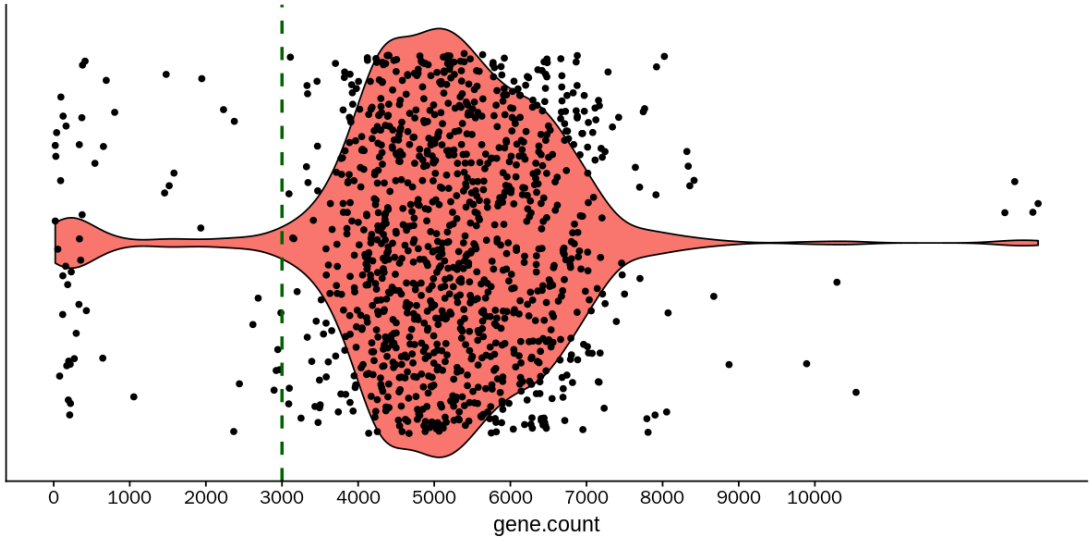
3.1 Data description

	Described	Provided
Number of cells	1140	1140
Number of genes	26593	26593
Cell preparation batch	8	8
Library preparation batch	2	NA
Traditional cell classification	768 DCs and 372 monocytes	768 DCs and 372 monocytes

3.2 Quality Control

The aim of the filtering step is to reduce the noise in the data by filtering out low-quality cells and low-frequency genes. First, a threshold of gene abundance (reads count, FPKM, TPM or other abundance estimation) is given to decide whether a gene is expressed. The 'quality' of a cell is then defined as the number of genes detected in the cell. Similarly, the frequency of a gene means the number of cells that express the gene. The cutoff of low-frequency genes is usually selected as 3 cells or some fraction of total cells (say 0.5%). The cutoff of low-quality cells, however, is usually data-specific, depending on the sequencing depths and other factors.

The original paper selected ***3000*** genes and ***3*** cells (it is described in the paper that this cutoff is set as ***0.5%*** of the total number of cells. The cutoff used in the given code is ***3***) as cutoffs to filter out low-quality cells and low-frequency genes. To see why they chose to set those parameters in the first place, we checked the distribution of gene numbers detected in each cell through violin plot. We observed a left 'tail' in the violin plot. Cells that distribute in the left tail were treated as low-quality ones. The boundary between the left tail and the main body, which was around ***3000*** in this case, was selected as the cutoff. It should be noted that 1. there is no exact boundary between the tail and the main body in the violin plot, researchers tend to select some multiple of hundred around the boundary as the cutoff. 2. in some studies, especially droplet-based sequencing studies, cells in the right tail are also treated as low-quality since more than one cells may be wrapped by one droplet.



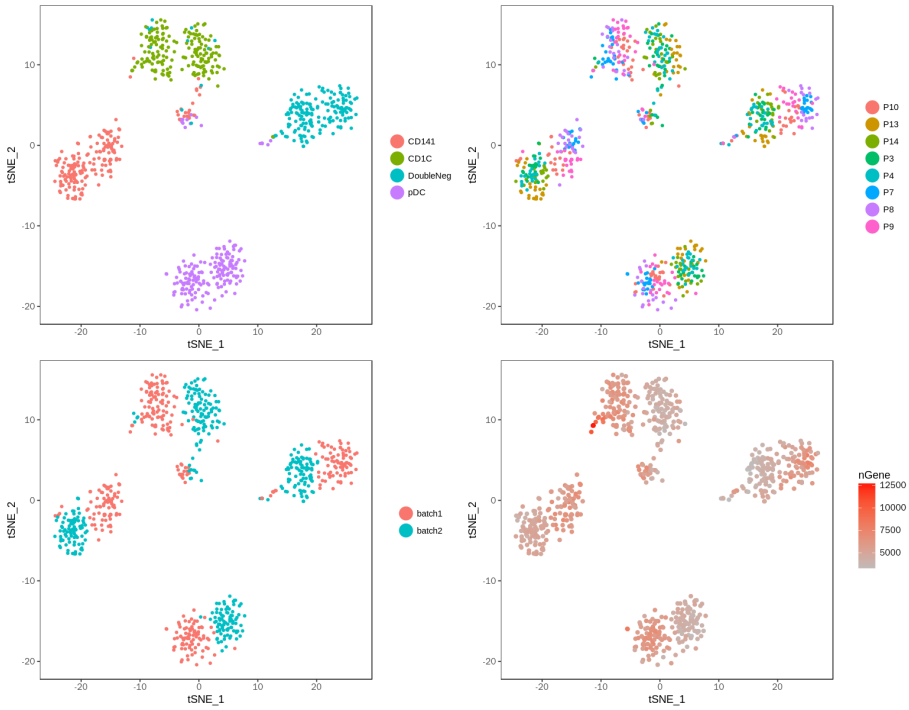
The filtering step resulting in ***19996*** genes and ***1085*** cells, including ***336*** Monocytes and ***749*** DCs. This is slightly different from the results in the paper (see the table below). The following analyses are only based on the data of the ***749*** DCs.

	Described	Provided
Number of DCs	NA	749
Number of monocytes	339	336
Number of genes	21581	19996

3.3 Removing the batch effect

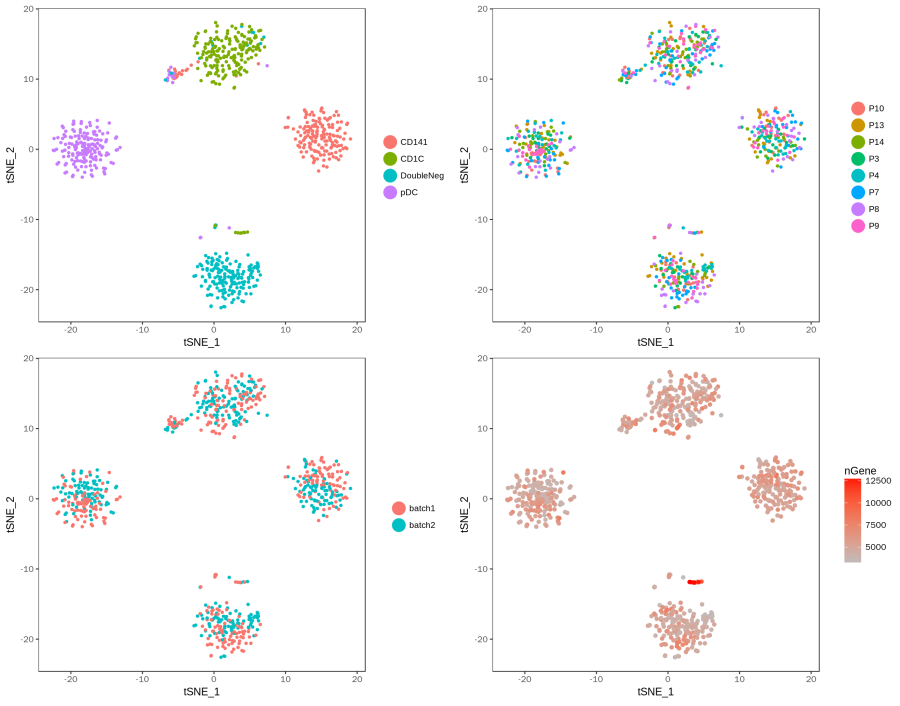
3.3.1 Check the batch effect

The DCs were profiled in 8 batches and the library preparations were carried out in 2 batches. To check if there is any batch effect in the data, we applied PCA to the whole data followed by t-SNE visualization. Below are the visualization results, colored by classical DC cell type, profiling batches, library preparation batches and the number of detected genes respectively (The library preparation batches information is not given in the paper, we inferred this based on the layout in the visualization of profiling batches). A weak batch effect were observed as the described in the paper. Besides, the layout of the visualization is also slightly biased by the number of genes detected in each cell.



3.3.2 Remove batch effect

To remove the batch effect and the effect of gene numbers, we regressed the gene expression using profiling batch indicator variable as well as the number of genes detected in the cell. The z-scored residuals are used as the adjusted gene expression for the downstream analyses. Below are the visualization for the adjusted data. Cells in different batches or with different detected genes numbers do not show distinct patterns, indicating that effects caused by those confounding factors have been removed.



3.4 Gene selection and PCA (1)

In single cells studies, the gene expression matrix usually contains tens of thousands of genes. As we know, dealing with high-dimension data is usually computationally heavy. Besides, since not all genes are informative for the interested questions and many genes are highly correlated in expression pattern, there is much redundancy in the high-dimension data. The aim of gene selection and dimension reduction is to remove redundancy in the data to improve the performance of downstream analysis.

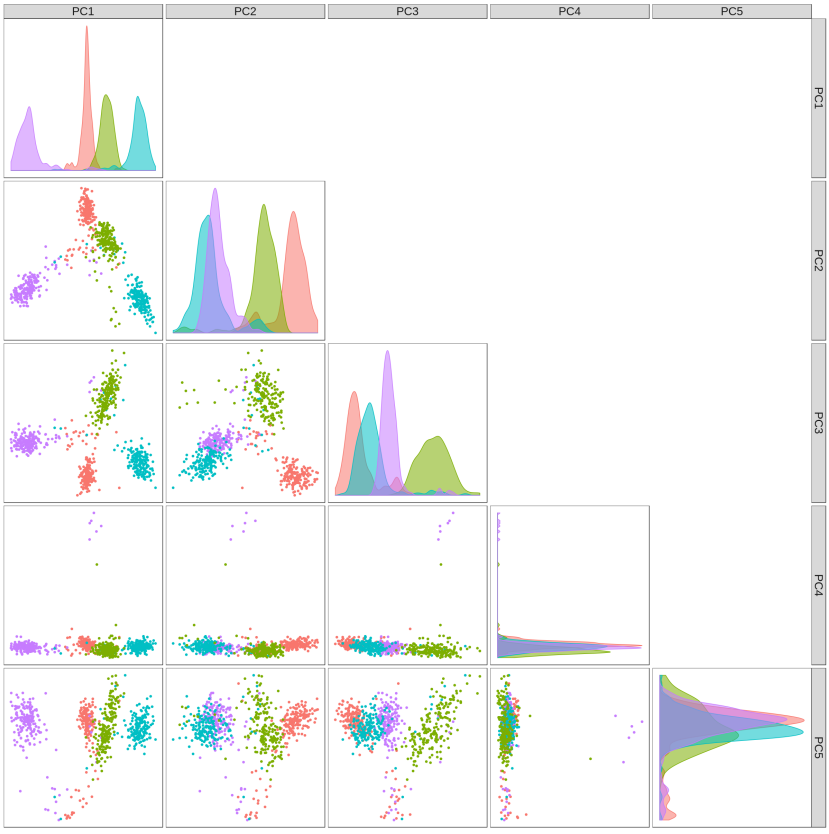
Gene selection removes redundancy in the data by selecting informative genes. One common strategy to measure the informativity of a gene is to calculate the variability in its gene expression pattern across cells. The variance-to-mean ratio or dispersion is usually involved in those calculations.

We did the first round of gene selection and PCA as the paper did. PCA was conducted on the ***563*** selected high variable genes. To grab an intuition of the dimension reduction results, we visualized PC1 to PC5 through both density plot and pair-wise scatter plot (colored by classical DC cell types). We got the following observations:

1. PC1 to PC3 has good discriminative power for the classical DC cell types.

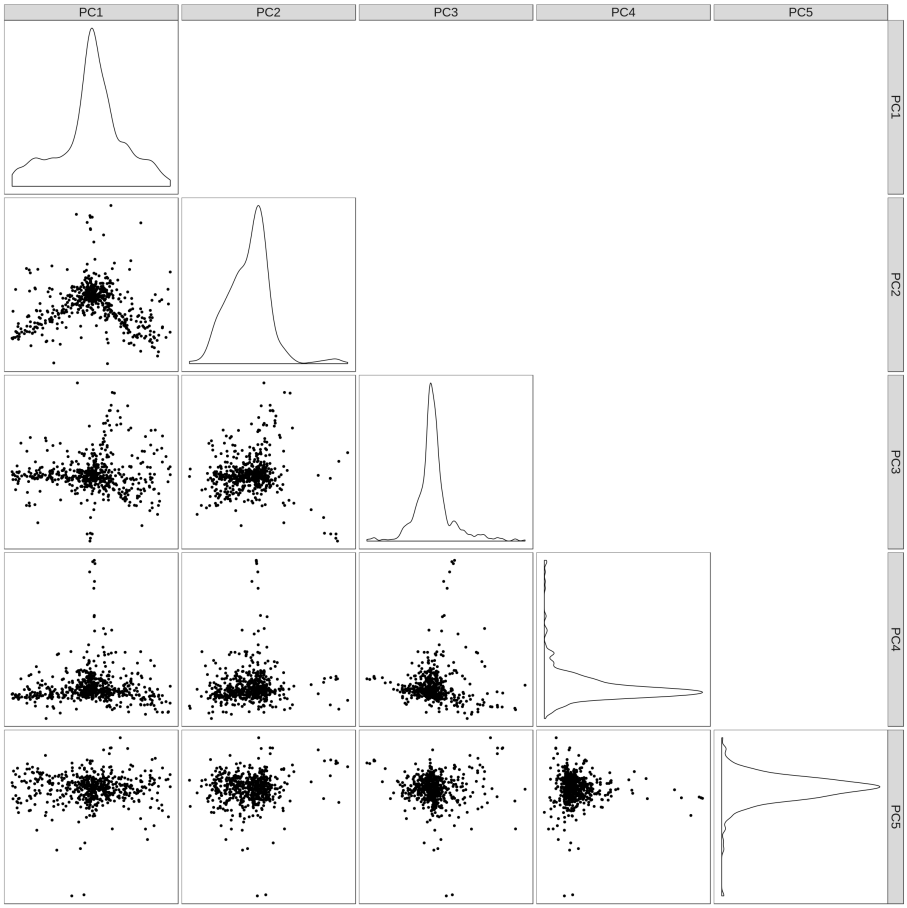
2. Seven cells appear to be outliers in PC4.

3. In PC5, other than the coincided main peaks for the classical DC cell type, there are some small peaks.

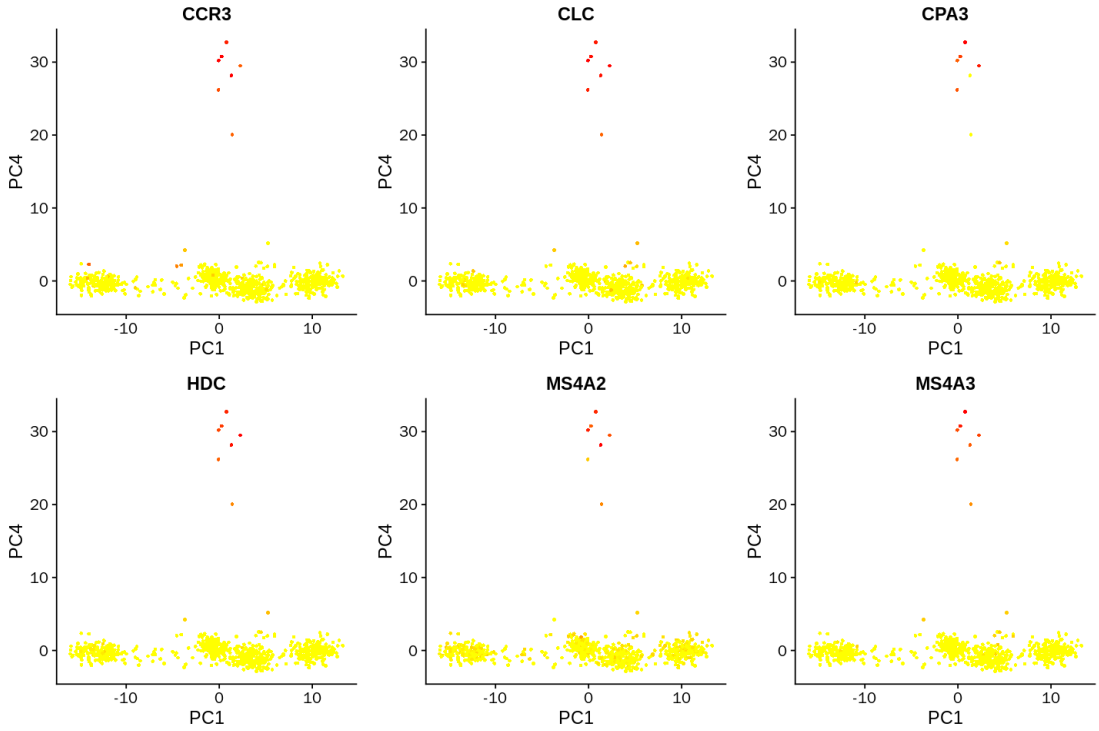


3.5 Remove contaminating subpopulations

In the original paper, the outliers in PC4 were removed as contaminating subpopulations. The paper also provides two genes as the marker gene of the contaminating subpopulations. Before doing so, how can we decide the outliers are contaminating subpopulations rather than a new cell type in the first place? A straightforward way is to do the differential gene expression detection between the outliers and other cells. The identity of the outliers can then be possibly decided by the detected marker genes. Another easier way is to check the gene loadings of the PCA results as follows.



In the gene loading plots, there appear some outlier genes in PC4, which apparently are the "markers" for the outlier cells in the original PCA plots. We illustrated this in the following figures. Each dot represents a cell, the y-axis represents the unnormalized cell loadings in PC4 and the color stands for the expression level of the selected gene in that cell. "CLC" and "MS4A2", two maker genes for the contaminating subpopulations given in the paper, are among those outlier genes. We removed those contaminating subpopulations according to cell loadings in PC4, leaving 742 DC cells for the downstream analysis.



3.5 Gene selection and PCA (2)

After removing the contaminating subpopulations, we redo the gene selection and PCA as in the paper, using the same method but with a different set of parameters. This step resulted in ***1531*** selected high variable genes which PCA was conducted on.

3.6 Gene selection and PCA (3)

The code given in the paper shows that another round of PCA and gene selection were conducted before t-SNE visualization and clustering. A randomization method called *jack straw* was used to identify *statistically significant* PCs and genes that contributed significantly to those selected PCs were used for the downstream dimension reduction and clustering analysis. This step results in ~400 selected genes (This step is stochastic so the number may change slightly between different runs).

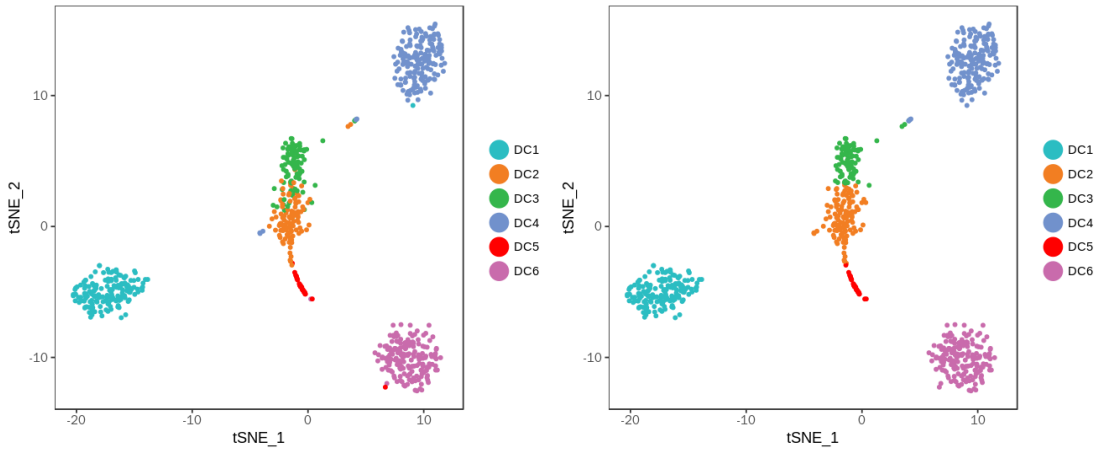
3.7 Clustering and t-SNE visualization

Graph-based clustering and t-SNE visualization were conducted after PCA. The original paper does not provide the exact code for clustering DCs but for the clustering of a combined analysis of DCs and monocytes. We use the first five PCs to construct the clustering as described in the paper and the parameter ***resolution*** was set ***1.4*** as in the combined analysis of DCs and monocytes. We conducted t-SNE on the first five PCs achieved in the previous steps with the perplexity 96 as in the paper.

3.7.1 Evaluation

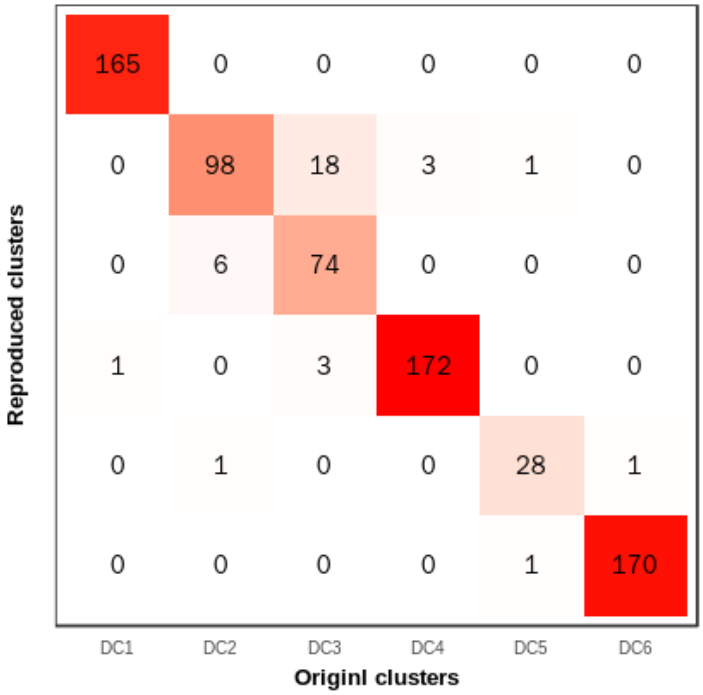
T-SNE visualization

The dots are colored by the original label in the paper (left) and the reproduced label (right) separately in the following figures.



Cell type assignment

Below is the comparison of cell type assignments. Cell types are aligned based on the Jaccard distance between original clusters and reproduced clusters.



Cell-type-specific marker genes

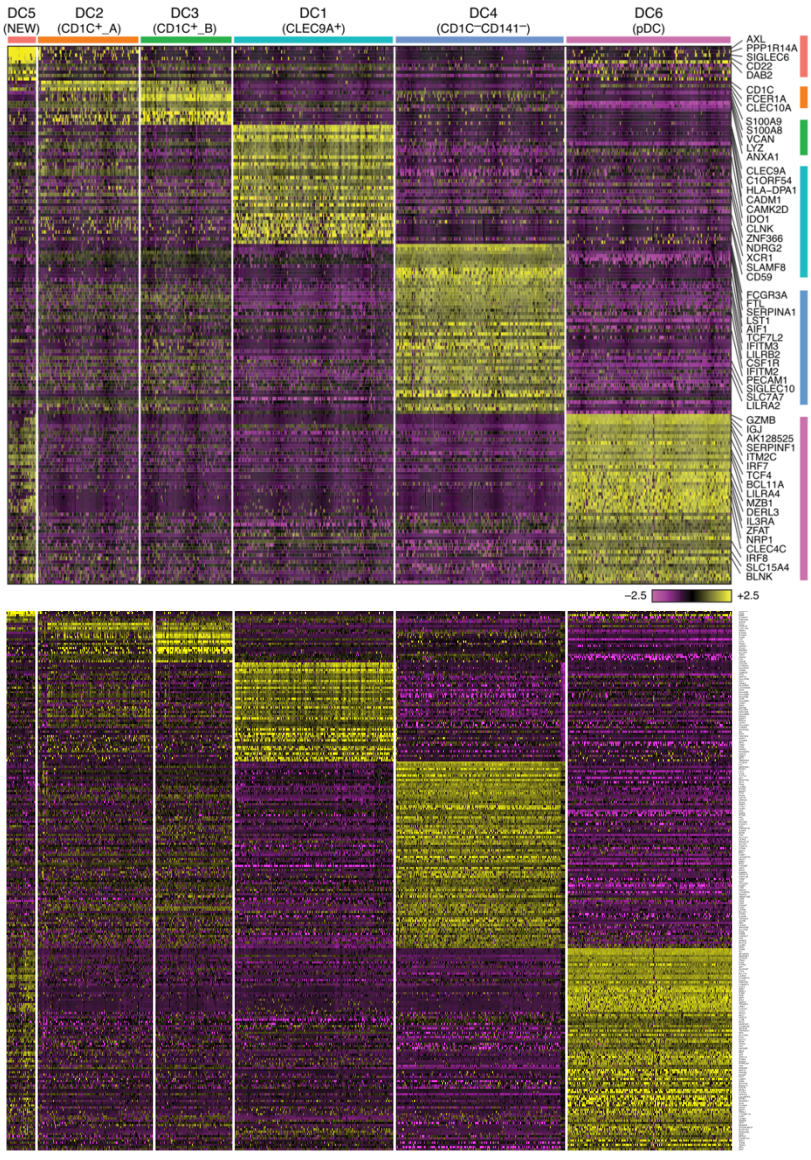
We first identified cell type-specific clusters for the reproduced cell types based on AUC scores as in the paper. Then we compared marker genes with AUC >= 0.85 between original results and reproduced results.

	DC1	DC2	DC3	DC4	DC5	DC6
Original paper	37	3	12	86	12	92
Reproduction	45	4	16	85	8	92
Overlap	36	3	11	82	8	90
Top 5 overlap	5	3	4	4	5	4

Heatmap of all the marker genes (AUC >= 0.85)

Original

Reproduced



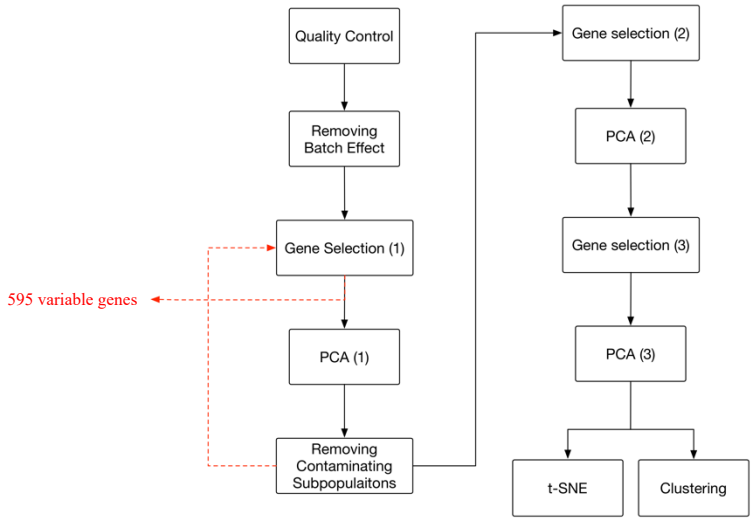
Expression patterns of marker genes

Expression patterns of cluster-specific marker genes can be found at https://github.com/XuegongLab/HCA-reproducibility.

4. Further exploration

4.1 The 595 variable genes in the paper

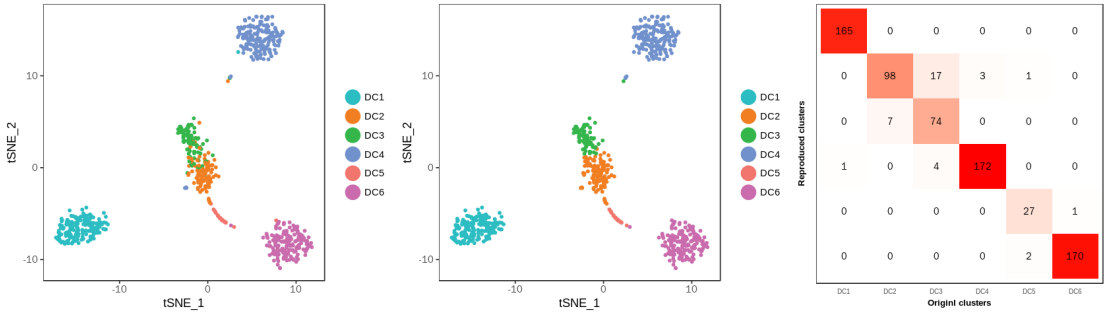
In scRNA-seq data analysis, it is common to first select variable genes as informative features for downstream analyses, and this feature selection step can sometimes affect the final results dramatically. For this work we tried to reproduce, it is described in the original paper that ***595*** variable genes were selected for doing PCA and other downstream analyses. However, the code shows that up to three rounds of PCA with three different selected gene lists were conducted and none of them contains *595* genes (the closest one is ***563*** genes in the first round of PCA). In our reproduction work, we figured it out how the number *595* came. We found that rerunning `FindVariableGenes()` after removing the contaminating subpopulations would results in ***595*** variable genes, as shown in the following flowchart.



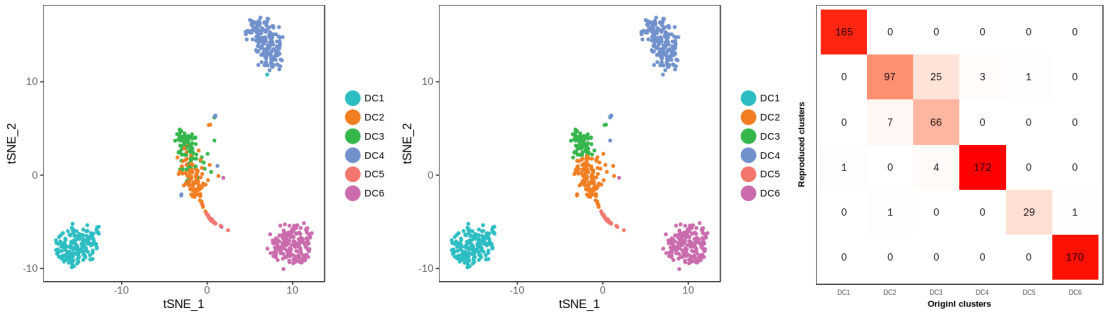
4.2 Effects of the selected variable gene list

The key step of the workflow is to conduct PCA on the selected variable genes and applying clustering and t-SNE on the achieved PCA space. So far we have ***4*** different gene lists in hand (***3*** used in each of the three rounds of PCA and the ***595*** genes we got in the last step). To check the effects of the selected gene list, we conducted the key steps on each of those 4 gene lists (the result given by gene list 3 is the 'reproduced result' we have discussed before and will be skipped here) and compared all the results with the original results in the paper.

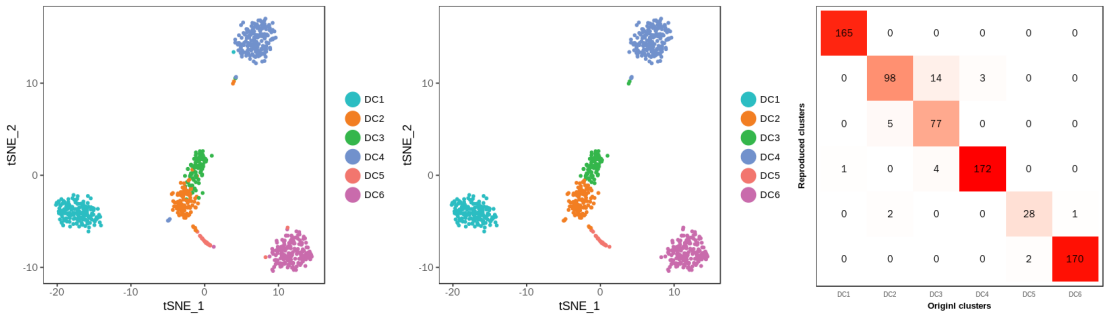
Gene list 1 (563 genes)



Gene list 2 (1531 genes)

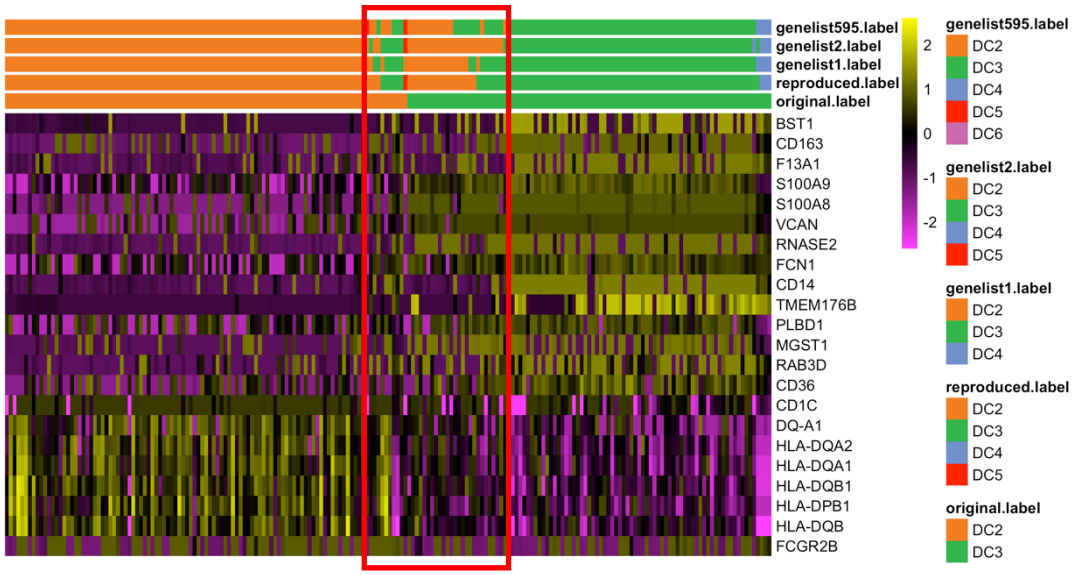


Gene list 4 (595 genes)



4.3 Blurry boundary of DC2 and DC3

Here is the heatmap of some selected marker genes (genes are selected according to Fig. 2A in the original paper). The boundary between DC2 and DC3 are not that clear. This blurry boundary is where the disagreement between different classification lies.

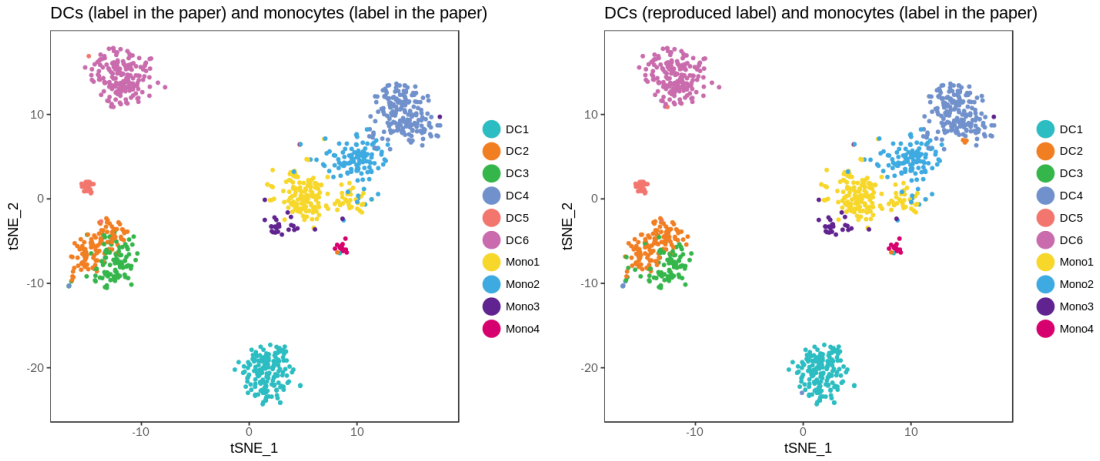


5. Reproduce some other figures in the original paper

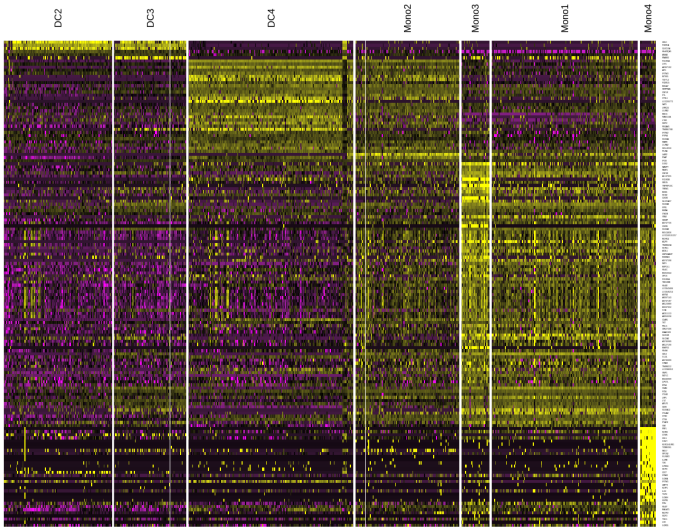
The details for generating most of the figures in the original paper is not provided. Here we try to reproduce some of them that related cell atlas.

5.1 Reproduce Figure 3B

The raw data matrix (TPM provided by the paper) of the ***742*** dendritic cells and the ***335*** monocytes (all these cells has a label available in the supplementary) were used in the analysis. The z-score normalization method was used on the log-transformed data (log(data+1)). Then we selected variable genes using the group-wise z-score method implemented in Seurat with the default parameters. This resulted in ***2629*** variable genes. The top ***2000*** variable genes were used for PCA. The first ***25*** PCs (decided based on the Elbow plot) were used for t-SNE with perplexity equals to ***60*** (this parameter was arbitrarily set after a few attempts). All monocytes were colored based on the label given in the paper. The DCs were colored according to the original paper and the reproduced label, separately.

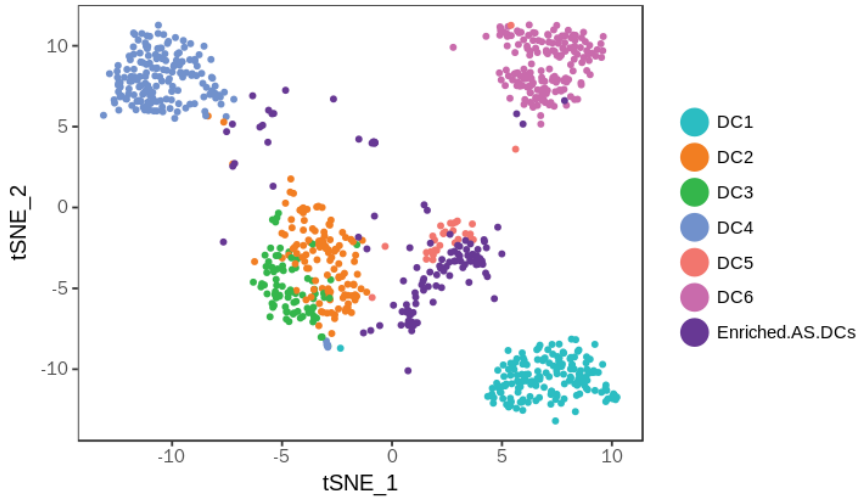


5.2 Reproduce Figure 3C



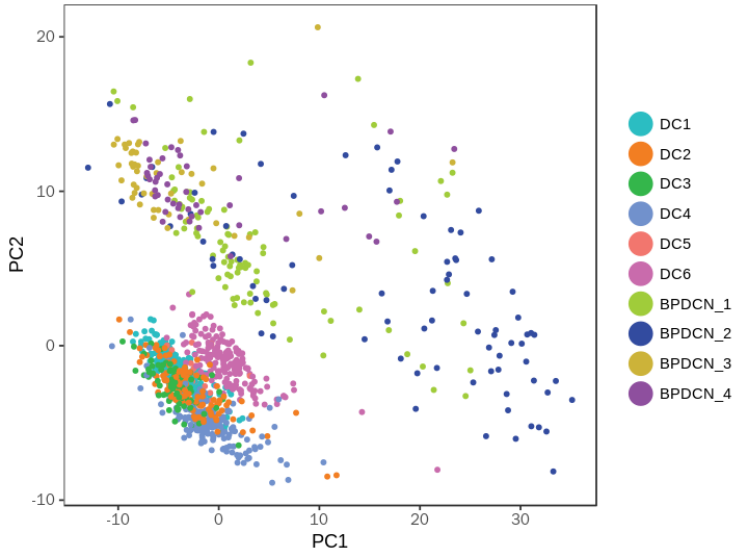
5.3 Reproduce Figure 4C

The reproduction of Fig.4C is quite straightforward. The raw data matrix of the 742 DCs and the 119 AS DCs achieved by cell sorting (in the paper, only 105 AS DCs were shown in Fig 4C. Since we don't how the cells were filtered, we used all the 119 cells here) were used in the analysis.



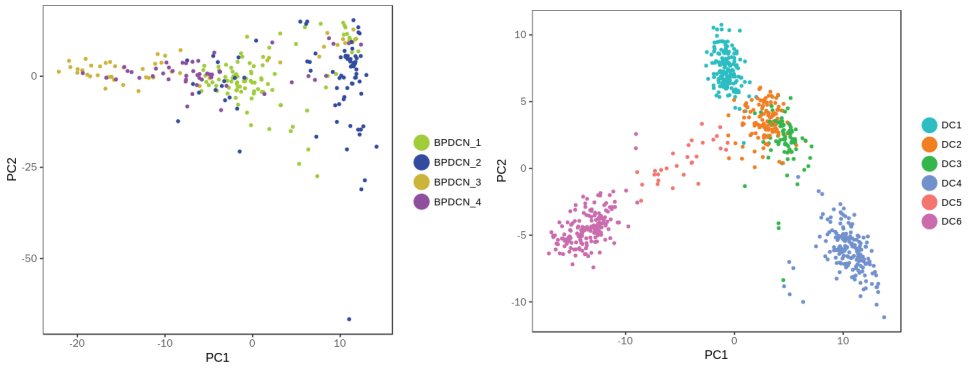
5.4 Reproduce Figure 6G

In the original paper, Figure 6G shows the result of mapping pathogenic cells from blastic plasmacytoid dendritic cell neoplasm (BPDCN) patients to the healthy DC atlas. In total, 269 BPDCN cells were obtained while only 174 of them were used to generated Figure 6G. The filtering detail was not provided, we therefore started the reproduction with all 269 cells.



We can see that cells from the 4 BPDCN patients are scattered aside the DCs in the map. They are close to DC6 but not mixed with it. Although the overall layout of the BPDCN cells is scattered, some of them do form a denser cluster which tends to be closer to DC6 than the more scattered ones. This may be the reason some of the cells were removed in the original

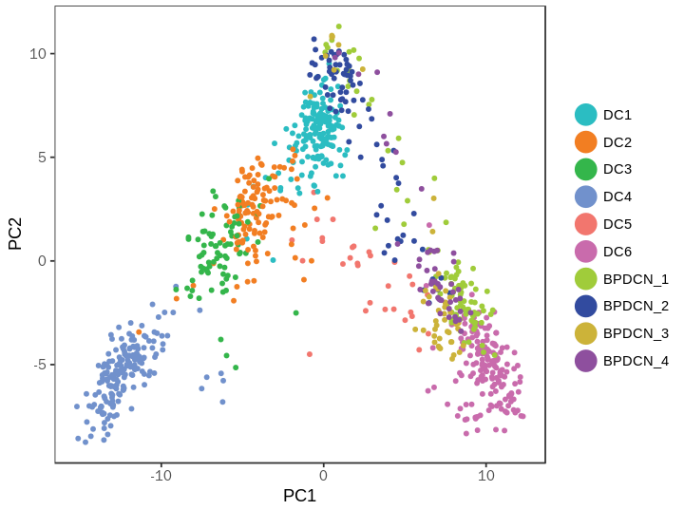
paper. To better reproduce the original result, we redid the analyses by first checking the BPDCN data before adding it to the DC data.



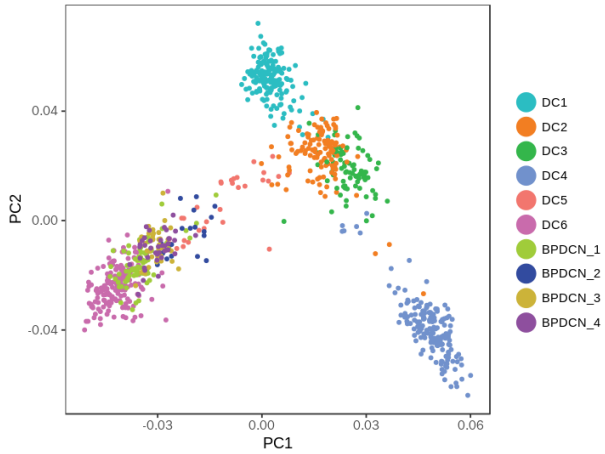
The major observations in the above plots are:

- Cells are scattered along PC1 for BPDCN data.
- The existence of an outlier along PC2 for BPDCN data.
- The layouts of DC1 -- DC6 are quite similar to those in the original Figure 6G.

To keep the layout of DC1 --DC6 as invariant as possible in the PCA visualization, we use the variable genes identified in analyzing DCs to do the PCA.



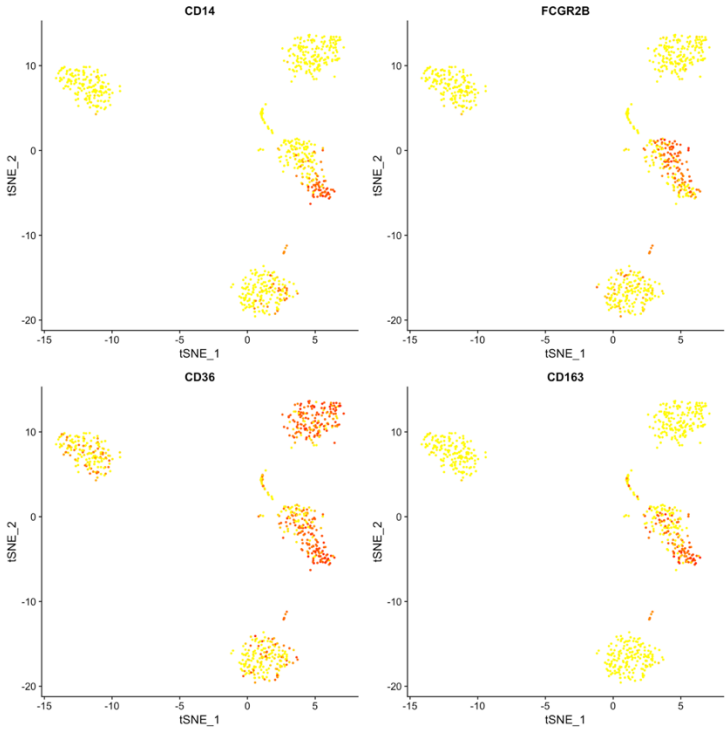
We can see that the layouts of DC1 - DC6 are similar to the original Figure 6G. BPDCN cells, however, present two major clusters (see the separated figures above). One is mixed with DC6 as described in the original paper while the other is close to DC1. To reproduce the original Figure 6G, we remove the cluster that is close to DC1 and redo PCA. This time we got a similar result to the original Figure 6G.



6. Other observations

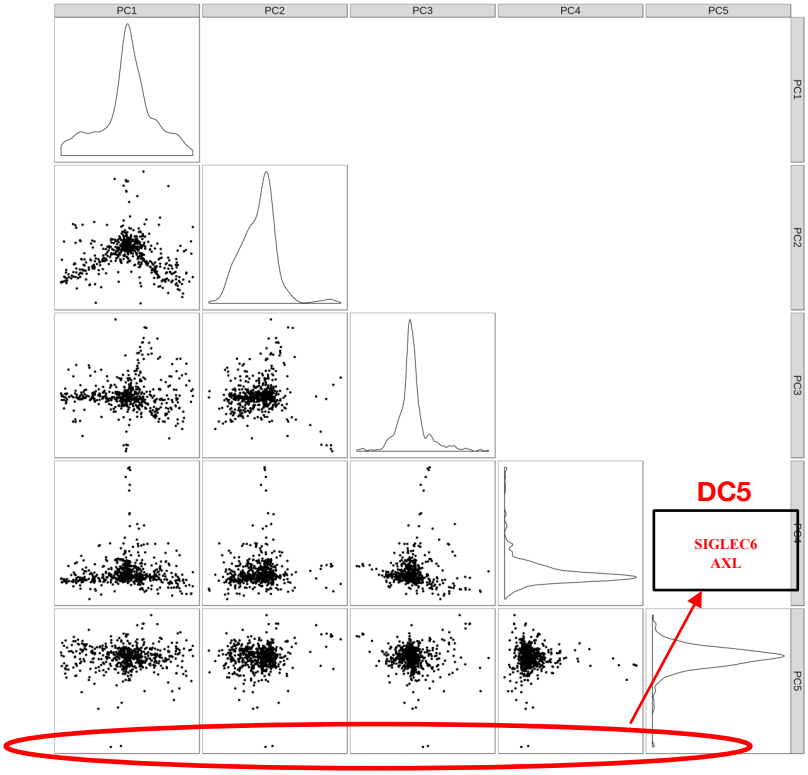
6.1 Expression patterns of some marker genes

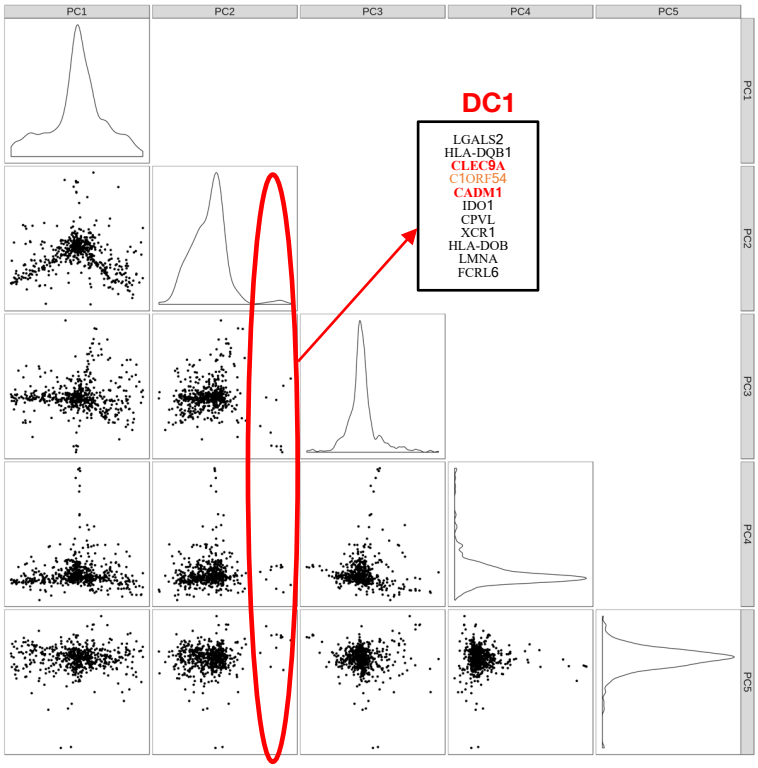
Here are the expression patterns of some marker genes used to sort DC2 and DC3 in the validation experiment.



6.2 Gene loading plot reflects some cell-type-specific marker genes

In the “removing contaminating subpopulation” step, we showed that the marker genes of the contaminating subpopulation cells appeared to be outliers in the gene loading plot. Likewise, we observed that some maker genes of identified cell types tend to be outliers in the gene loading plot as well. For example, SIGLEC6 and AXL, two markers of DC5, appear to be outliers in the PC5. Another two marker genes of DC1, CLEC9A and CADM1, also deviate from most genes in PC2.




PC1
PC2
PC3
PC4
PC5
DC1
LGALS2
HLA-DQB1
CLEC9A
C1ORF54
CADM1
IDO1
CPVL
XCR1
HLA-DOB
LMNA
FCRL6