Use of Single Cell RNA Sequencing Data to Understand the Process of T cell Maturation in Thymus

Master of Biomedical Research

Department of Surgery & Cancer

Faculty of Medicine

Imperial College London

2019-08-12

Principal supervisor: Prof. Matthias Merkenschlager

Day-to-day supervisor: Dr. Mahdi Karimi

Author: Xiaokai Cui (student of Data Science Stream)

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline. The experiments of biology were designed and implemented by the research group led by Prof. Matthias Merkenschlager at the Medical Research Council London Institute of Medical Sciences. The dataset used in this project was prepared by Dr. Mahdi Karimi who also suggested the multiplication equation for trajectory inference. The manually selected marker genes and the differentially expressed transcription factors were provided by Prof. Merkenschlager and Dr. Ya Guo respectively.

Acknowledgements

I would like to first acknowledge and thank my principal supervisor Prof. Matthias Merkenschlager and my day-to-day supervisor Dr. Mahdi Karimi. The instructions and the expectations from Prof. Merkenschlager are the powerful enhancers for me to improve my research work and performance. Dr. Karimi designed a one-month background learning scheme for me when I started the project, which is crucial for me to understand the essential background knowledge. In the Bioinformatics team, I also got valuable advice from Dr. Yifang Wang, Marian Dore and Sanjay Khadayate, who taught me a lot during my background learning. Finally, I appreciate the valuable advice from my supervisors for this thesis, their help means a lot to me and I cherish the experience and the opportunity of learning from and cooperating with them in this project.

Abbreviations

BXVP: binary-label XGBoost model with Venn-Abers Predictors

DE: differentially expressed

KSM: kinetic signalling model

MCCV: Monte Carlo cross-validation

MHC: major histocompatibility complex

NGS: next-generation sequencing

PC: principal component

PCA: principal component analysis

PEL: positive expression level

RF: random forest

RUE: ratio of undetected expression

scRNA-seq: single cell RNA sequencing

SVG: stochastic cell-to-cell variability of gene expression

T cell: T lymphocyte

TCR: T cell antigen receptor

DP: CD4 and CD8 coreceptor double positive

CD69-DP: CD69 protein negative, CD4 and CD8 coreceptor double positive

CD69+DP: CD69 protein positive, CD4 and CD8 coreceptor double positive

CD4+CD8low: CD4 coreceptor positive and CD8 coreceptor low

CD4SP: CD4 coreceptor single positive

CD8SP: CD8 coreceptor single positive

TCR+DP: TCR high CD4 coreceptor and CD8 coreceptor double positive

F-set: the set of DE genes detected by the FindMarkers function in Seurat

K-set: the set of DE genes detected by Kolmogorov-Smirnov test based on the PELs

M-set: the manually collected DE gene set based on domain knowledge

R-set: the set of randomly selected genes from the W-set

T-set: the set of DE transcription factors

W-set: the whole set of 9960 genes

Z-set: the set of DE genes detected by two proportion z-test based on the RUEs

Abstract

In this study, single cell RNA sequencing (scRNA-seq) data was used to better understand the dynamic process of T cell maturation in thymus. During the research, the ratio of undetected expression (RUE) of certain genes were found to relate to the biology of the system, rather than technical dropout events, and proved useful for distinguishing subpopulations of developing thymocytes. The use of RUEs described here may be useful beyond understanding T cell maturation and may be applied to other areas of research in bioinformatics. Application of a novel trajectory inference approach based on the binary-label classification model developed in my previous work was applied to the scRNA-seq data. Combined with differentially expressed genes detected based on RUEs as input, this approach was able to infer informative trajectories of T cell maturation in thymus than other existing methods. Thus, it enables the research on the continuous gene expression changing through the maturation trajectories.

1. Introduction

1.1 Background

Next-generation sequencing (NGS) refers to faster and cheaper sequencing technologies than the Sanger-based sequencing technologies. NGS-based technologies enable genome-wide research in genomics, transcriptomics and epigenomics and contribute to the long-standing challenge of understanding the relationship between genotypes and phenotypes. Bulk RNA sequencing technologies apply NGS to characterise transcriptomes cell populations. Bulk RNA sequencing ignores transcriptome variation between individual cells by analysing average gene expression for ensembles of cells. (1)

In 2009, Tang, et al. (2) first described an NGS-based experiment at the single-cell level to illustrate the diversity of transcriptome of individual mouse blastomere cells. In the following years, increasing number of studies showed that gene expression of individual cells in apparently homogenous cell types was heterogeneous, and that variation of single cell gene expression affects biological processes such as immune responses and cellular differentiation. (3, 4) ScRNA-seq technologies allow the identification of underlying cell types and stages based on the variation of gene expression at the cellular scale, thus, it is possible to discover rare populations and new regulators. (5) Moreover, with the aid of profiling of the gene expression of individual cells alongside the dynamic transition such as activation, differentiation and maturation, more accurate cellular trajectories in cell development can be inferred based on pseudotime analysis. (6)

Recently, scRNA-seq technologies have been applied to the research of T lymphocyte (T cell) development which contributes to the understanding of T cell responses to infection, autoantigens and vaccination, thereby promotes the development of pathogenic and therapeutic research. (7) For example, Proserpio et al. revealed the 3 major cell states during the differentiation of mature CD4 coreceptor positive T cells. (8) In 2019, Miragaia et al. compared the transcriptional features of regulatory T cells in different tissues based on scRNA-seq technologies. (9) These studies showed the promise of using scRNA-seq to illustrate the differentiation process of mature T cells. In this study, we aimed at analysing and modelling scRNA-seq data in depth to gain more useful information for the research of T cell maturation in the thymus.

1.2 T-cell Maturation in Thymus

The T cell precursors migrate from the bone marrow to the thymus, where they develop into mature T cells. Approximately 2% of developing T cells survive the process of maturation in thymus, while most fail either positive or negative selection and die by apoptosis. Positive selection allows only the T cell whose T cell antigen receptor (TCR) can bind major histocompatibility complex (MHC) molecules to pass the selection, which increases the probability that the surviving T cells can recognise foreign antigens. During positive selection, the TCR of a T cell can bind to the class I MHC molecule or class II MHC molecule with the aid of CD8 or CD4 coreceptor separately. The T cell whose TCR binds the class I MHC molecule will proceed along the CD8 single positive (CD8SP) developmental pathway, while the T cell whose TCR binds the class II MHC molecule will proceed along the CD4 single positive (CD4SP) developmental pathway. Negative selection allows only the T cell with low-affinity TCR for self-antigen presented by self-MHC molecules to pass the selection, which in turn ensures the surviving T cells are self-tolerant. Thus, T cells which survive in both positive and negative selection are enriched for recognition of foreign antigens in the context of self-MHC molecules. (10)

The CD4, CD8 and TCR combining with CD69 protein which appears on the surface of T cells only at the intermediate stages of T cell maturation are four biomarkers that identify subpopulations of T cells during their development in thymus. (11) In the beginning of T cell maturation in thymus, there are CD4 and CD8 double negative T cells who are the precursors of CD69 protein negative, CD4 and CD8 double positive (CD69-DP) T cells. For positive selection, the TCRs first appear in the CD69 protein positive, CD4 and CD8 double positive (CD69+DP) T cells which are developed from the CD69-DP T cells. (11) Since the kinetic signalling model (KSM) was introduced by Brugnera et al. in 2000 and contributed by Singer later on, it is widely accepted that the CD4 and CD8 lineage decision is committed in the CD4 positive and CD8 low (CD4+CD8low) T cells which are developed from CD69+DP cells. Then the mature CD4SP and CD8SP T cells are differentiated from the CD4+CD8low T cells during a reversible process. (Fig. 1) (12, 13, 14) Despite the relatively clear process of T cell maturation, the question how the CD4SP and CD8SP cells are continuously developed from DP cells still remains elusive. In this study, we integrated scRNA-seq data with machine learning approaches to infer the T cell maturation trajectories in order to advance the understanding of the process of T cell maturation in thymus.

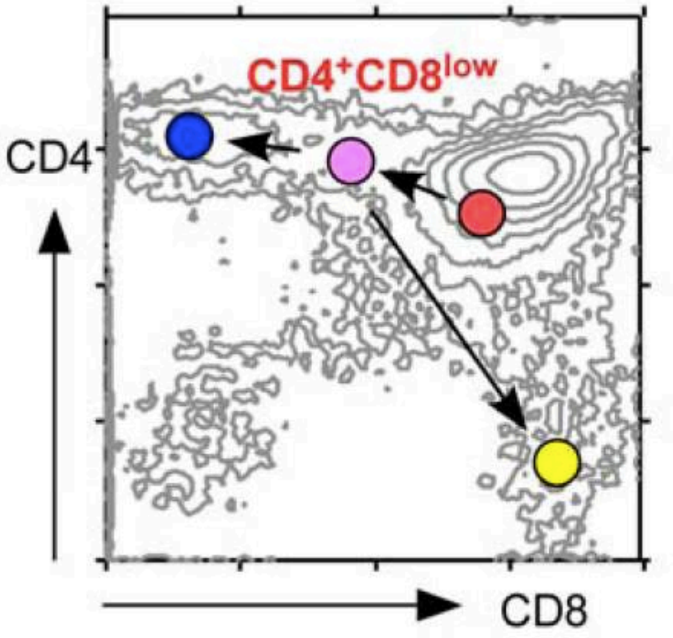


Figure 1: The intensity distribution of T cells based on the levels of its CD4 and CD8. The red, purple, blue and yellow circles represent the subpopulations of DP, CD4+CD8low, CD4SP and CD8SP respectively. The arrows point to the directions of T cell differentiation based on the KSM. (Adapted from (14))

There were six groups of T cells used in this study. The first five groups separately represent the five subpopulations (CD69-DP, CD69+DP, CD4+CD8low, CD4SP and CD8SP) used to reveal the developmental trajectories of T cells, while the last group of TCR+DP (TCR high CD4 and CD8 double positive) cells represent a subset that overlaps with CD69+DP cells (because TCR first appears in CD69+DP cells) and immature CD8SP cells. According to the KSM, a CD8-lineage-committed cell is differentiated from a CD4+CD8low cell, once the CD4+CD8low cell decides to go into the CD8SP lineage, its CD8 coreceptor will be produced again, then the CD4+CD8low cell will become a CD8SP through a transient DP stage. Considering that the TCR keeps high during the whole process, thus, a TCR+DP cell is possible to be at the immature stage of the differentiation from a CD4+CD8low cell to a CD8SP cell. This uncertainty of TCR+DP cells enables TCR+DP cells to be used as validation data for the trajectory inference by comparing their positions to the ones of CD69+DP and CD8SP cells.

1.3 Single Cell RNA Sequencing Analysis

In the majority of scRNA-seq experiments, cells are stained by several antibodies which are used for distinguishing cell populations following isolating cells from target tissues. Then the cells are separated based on light scatter and fluorescence parameters by fluorescence-activated cell sorting. After cell sorting, there are many protocols can be used to prepare libraries for sequencing, for example the Smart-seq2(15). The output of the scRNA-seq technologies is comprised by many FastQ files which store the nucleotide sequences and their corresponding read quality scores.

The FastQ files can be aligned to a reference genome by multiple tools such as TopHat2, STAR and Kallisto. (16, 17, 18) The bam files generated after the alignment can be used for accurate counting by the tools such as Rsubread and velocyto. (19, 20) The read counting matrix which is calculated by counting tools can be then normalised by log-transformation, DEseq2 or Seurat in order to remove systematic biases between cells. (21, 22) The normalised counting matrix generated by this way can be used for downstream analyses.

Although cells are labelled and sorted based on different fluorescent antibodies in scRNA-seq experiments, it is still meaningful to cluster cells using the gene expression data, for example discovering certain subpopulations which is impossible or difficult to be distinguished by cell staining. Hierarchical clustering and K-means clustering algorithms are two most common clustering methods used in scRNA-seq analysis, they can be very powerful by using the expressions of differentially expressed (DE) genes as the input data. (23, 24)

The principal component analysis (PCA) is widely used in scRNA-seq analysis for dimension reduction because of its good interpretability. A principal component (PC) calculated by PCA using gene expression data represents a common feature abstracted from the input genes. Considering that each PC is orthogonal to others, PCs can be used to reveal a number of distinct common features. For cell population research, the amount of useful information represented by a PC can be measured by the PC's capability for distinguishing cell populations. Because if a PC can separate cell populations well, it means that the common feature represented by the PC contains useful information for revealing the distinctive characteristics among different cell populations. Since the results of PCA varies depending on the input data, it is a good way of evaluating the useful information contained in the set of genes whose expression matrix is used as the input data of the PCA.

In the research of cell fate decision, studying the continuously changing cell status during cell development is significant for understanding the underlying mechanisms. ScRNA-seq technologies cannot monitor the gene expression of an individual cell over time because RNA extraction requires cell lysis. However, we can instead sample cells at distinct points in their development, then order the cells along the trajectory; this ordering is referred to as "pseudotime". (25) The dynverse R package includes more than 50 pseudotime analysis tools, in this study I used the Slingshot which was recommended by dynverse as the benchmark. (26, 27)

There are two classification algorithms which are binary-label XGBoost model with Venn-Abers Predictors (BXVP) and random forest (RF) used in this study, they were used with Monte Carlo cross-validation (MCCV) together in order to separate the training and testing datasets by randomly selecting cells from each subpopulation. (28, 29) The BXVP was developed based on a XGBoost model for binary classification and Venn-Abers Predictors for converting the prediction scores into more reliable probabilities. This algorithm has shown its powerful capability of classifying samples using numeric data in my previous work. (30) RF is a widely used algorithm for multi-label classification in biomedical research, in this study, it was used for measuring the PCs' capability for distinguishing cell populations.

1.4 The Detection of Differentially Expressed Genes

There are more than 20,000 protein-coding genes in mouse and human genome and scRNA-seq is capable of detecting many thousands of them in each cell. Nevertheless, in most of the cases, only a small portion of genes are differentially expressed between the biology conditions of interest. (29) Thus, in order to avoid noises and biases hiding in the scRNA-seq data, only the expression of DE genes is used in downstream analyses, such as cell clustering and pseudotime analysis. DE genes are detected by comparing their expression levels among cell populations. There are many ways for detecting DE genes, for example, Seurat is one of the most widely used tools in scRNA-seq analysis, which detects the DE genes using four parametric tests including t-test. (31) In addition, a non-parametric test called Kolmogorov-Smirnov test is also widely used to detect DE genes by comparing the distributions of gene expressions among cell populations.

When detecting the DE genes or analysing scRNA-seq data, the undetected expression of a gene among populations are either ignored or averaged with its corresponding positive expression levels (PELs) by most of the existing methods. For example, recent developed tools, such as scDesign and EnImpute, treat the undetected gene expression as a type of data bias which is imputed based on the cell-wise and gene-wise ratios of undetected expression (RUEs). (32, 33) The RUEs were not considered independently because it is generally believed that the RUEs and the PELs are basically correlated, in addition the PELs were considered more informative than the RUEs for scRNA-seq analysis because there are statistical reasons rather than biological reasons affecting the RUEs.

Besides the stochastic cell-to-cell variability of gene expression (SVG), a so-called “dropout event” describing a phenomenon that a gene expressed in one cell is not detected in other similar cells is generally used to explain the undetectable gene expression together with its causation. The dropout event is caused by the mRNA failing to be reversely transcribed in experiments due to its limited amount. (34, 35) It is clear that a mRNA missed the step of reverse transcription will then definitely miss the amplification step leading to the fact that it is impossible to be detected in the sequencing step. (More details of scRNA-seq work flow can be found in Sec. 2.1)

Although the dropout events and the SVG can be used to explain part of the reason why the expression of a gene cannot be detected in a cell where the gene should be expressed, there may be more reasons. Except the case that a gene is not expressed in a cell population at all, assuming that the dropout event and the SVG is the only reason for undetected gene expression, it can be deduced that the RUEs of a gene should be negatively correlated with its PELs. According to this deduction, the set of DE genes detected based on RUEs (Z-set) should be similar to the one detected based on PELs (K-set), in addition, Z-set should be at most equally informative comparing to K-set. Otherwise, it is reasonable to infer that, comparing to the PELs, the RUEs are governed by different unknown mechanisms which may provide more useful information of T cell maturation in thymus.

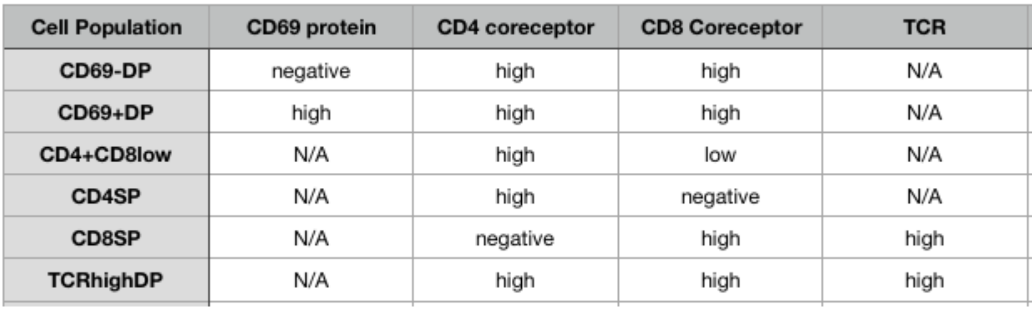
1.5 Aim and Hypothesis

In this study, we proposed the hypothesis that, despite many existing methods, there is potential for extending the useful information contained in scRNA-seq data to better understand the process of T cell maturation in thymus. This hypothesis was examined in two steps, in the first step, we aimed at illustrating that Z-set are more informative than K-set, which shows that the feature of RUEs is a potential key to discover more underlying biological mechanisms. In the second step, we aimed at developing a novel trajectory inference approach which is able to order the T cells in well-inferred maturation trajectories, which in turn allows the research on the continuously changing gene expressions of T cells in thymus during their maturation.

2. Methods

2.1 Prior Experiments and Data Preparation

In this study, the T cells were isolated from mouse thymus and then were stained by using 4 fluorescent antibodies to the cell surface biomarkers CD69 protein, CD4, CD8 and TCR respectively. Then those cells were sorted by a fluorescence-activated cell sorter into individual wells of 96-well plates depending on their fluorescent characteristics. The cell sorting step allowed us to select roughly 100 cells from each subpopulation despite the fact that the cells are unevenly distributed among subpopulations. The taxonomy of the T cells according to the staining features of the biomarkers is shown in Table 1 as below.



Cell Population	CD69 protein	CD4 coreceptor	CD8 Coreceptor	TCR
CD69-DP	negative	high	high	N/A
CD69+DP	high	high	high	N/A
CD4+CD8low	N/A	high	low	N/A
CD4SP	N/A	high	negative	N/A
CD8SP	N/A	negative	high	high
TCRhighDP	N/A	high	high	high

Table 1: The T cell labelling criteria based on the four staining features.

Following the cell sorting step, the RNAs of these cells were sequenced based on the Smart-seq2 protocol which is a more sensitive protocol and therefor suitable for sequencing low RNA-content cells such as T cells. (36) There are 5 main steps in the Smart-seq2 protocol to sequence RNA in single cells shown as below. (15)

Step 1. The sorted single cells are lysed in different wells.

Step 2. RNAs of single cells are converted to cDNAs by reverse transcription.

Step 3. Amplify the converted cDNA and exclude the by-product of primer dimers.

Step 4. Pool together libraries, each with a unique combination of adaptors (markers).

Step 5. cDNA sequencing and read data collection.

The RNA sequence data generated from the experiment was first aligned to the mm10 reference genome by TopHat2, then counted by velocyto using the same reference genome and normalised by the function of rlog in DEseq2. There are 9960 genes (the other genes were excluded by velocyto due to high RUEs) in the normalised gene expression matrix, for each subpopulation there are about 90 cells in the dataset and the size of the whole matrix is 551 by 9960. (16, 20, 21)

2.2 DE Gene Detection and Evaluation

There were 6 sets of DE genes detected by different approaches used in this study for comparison. One of them comprised by 89 DE genes was manually collected by Prof. Merkenschlager based on domain knowledge to describe the differentiation process, it is denoted as M-set. Another set provided by Dr. Guo was comprised by 303 transcriptional factor genes which were also DE genes among the cell subpopulations, it is denoted as T-set. There was also a set of 100 randomly selected genes from all of the 9960 genes, it is denoted as R-set. The other three sets of DE genes were generated by ranking DE genes based on the scores calculated by three computational approaches respectively. The ranking method is the same across the ways used for generating the three sets of DE genes.

The first scoring approach is based on the feature of RUEs, which uses the “Two-proportion z-test” (implemented in R) to test whether the RUE of a gene for a certain subpopulation is significantly different from the one for another subpopulation. (37) The second scoring approach is based on PELs, which uses the “Kolmogorov-Smirnov test” (implemented in the package of scipy) to test whether the distribution of PEL of a gene for a certain subpopulation is significantly different from the one for another subpopulation. (38) The last scoring approach is the function of “FindMarkers” in Seurat package, which also uses statistical tests to detect DE genes between subpopulations by using gene expression matrix which contains both undetected and positive expressions. (31) The p values calculated by these three approaches were used as the ranking scores in the following DE gene selection step.

In the selection step, the output data from each approach was separated into different groups, each group represents a comparison condition of a subpopulation versus another. As there were five subpopulations which could compose 10 distinct pairwise combinations, thus, 10 groups were formed in the selection step. Then around 100 distinct DE genes were collected by removing the duplications from the set of DE genes which were evenly selected from the 10 groups in the order of p values from low to high. The three DE gene sets generated this way are denoted as Z-set

(Two proportion z-test), K-set (Kolmogorov-Smirnov test) and F-set (FindMarkers in Seurat) separately according to the scoring method they used.

For evaluating a set of DE genes, the PCA transformation was applied to the expression matrix of those DE genes, then the top 3 PCs (ranked by the proportions of variance which the PCs explain) were kept in the new matrix because it was checked that the other PCs had very little capability of distinguishing cell populations in this study (Supplementary Fig. 3, 4, 5 and 6). Thus, a 431 by 3 matrix (The immature CD4SP (see Sec. 3.2) and TCR+DP cells were excluded from the matrix) was generated based on each set of DE genes which was used for PCA transformation, it was then used for predicting the subpopulations of cells after training the RF models. The prediction based on each matrix was repeated 20 times using MCCV with 80% data used for training, so that there were 20 accuracy ratios generated to quantitively measure the useful information contained in a set of DE genes. The average accuracy ratio measures the abundance of useful information contained in the corresponding DE gene set, which in turn describes the quality of the set. Furthermore, the quality of different DE gene sets can be compared by testing their corresponding average accuracy ratios using one-tail independent-samples t-tests by “t.test” function in R with the setting of “equal variances not assumed”. (37).

2.3 Mature CD4SP Cell Identification

In previous analysis, Prof. Merkenschlager found that there were immature T cells mixed in the CD4SP subpopulation which should be comprised by only mature T cells. Thus, it is significant to separate those immature cells from the mature ones in order to purify the CD4SP subpopulation. Considering that the immature and mature cells have different features in gene expression, thus, they can be distinguished by clustering algorithms using gene expression data. In this study, each cell was represented by a vector of gene expression, the original vector was transformed to a new vector whose L1-norm is equal to 1 by dividing the original vector by its L1-norm. After the transformation, Kullback–Leibler divergence (the calculation function is implemented by Scipy (38)) could be used to measure the distance between two cells based on their new vectors. (39) After finishing the calculation of all the distances between two cells, a cell-distance matrix was generated, which was then used as the input data of the hierarchical clustering (implemented in the package of scikit-learn (40)).

Hierarchical clustering was used to sort cells in order to let similar cells be close to each other. Then a heat map visualising the distances between cells was plotted based on the cell order calculated by hierarchical clustering (41). Ideally, one large cluster (comprised by more than 50% of CD4SP cells) can be manually selected, it is reasonable to assume that the larger cluster is more likely to be the mature cells considering the high accuracy of cell staining and sorting methods used in the experiments.

However, further tests are required to ensure that the mature CD4SP cells are properly separated from the immature ones. There are seven genes cautiously selected from different sources to verify the separation of mature CD4SP cell, the expressions of four of them are higher in the immature cells while the expressions of the other three of them are higher in the mature cells. The first two genes were ITM2A and STAT1 selected by Prof. Merkenschlager based on domain knowledge. In general, ITM2A expresses more in the immature cells while STAT1 expresses more in the mature cells. The expressions of CCR9 and CCR7 behave oppositely during the maturation of CD4SP cells as CCR9 expresses more in immature cells and CC7 expresses more in mature cells. (42) H2-Q7 coding QA-2 which is a marker of mature T cells can be used as a good marker for the mature cells while CD69 is a good marker for the immature cells due to its transient feature of expression (11). In addition, the cells expressing more CD24A (also known as HSA) are more likely to be the immature CD4SP cells, so CD24A is another good marker of the immature cells. (43)

2.4 Pseudotime analysis

Considering that there are only two lineages of the T cells in this study, thus, there is no need to apply dimension reduction methods which is widely used in many other pseudotime analysis methods to visualise the developmental trajectories of multiple lineages together. In this study, a specific coordinate system can be defined for visualising the T cell maturation trajectories in thymus. In details, its x axis indicates the maturation of a cell in the CD4SP lineage development and its y axis indicates the maturation of a cell in the CD8SP lineage development, in addition the origin represents the starting point of the development of both lineages, where the CD69-DP cells start to mature.

Another question following the definition of the coordinate system is how to calculate the x and y coordinates of a cell. A good measure used to quantify the maturation in the developmental process is the likelihood that a cell is a mature CD4SP or CD8SP cell. In this study, BXVP algorithm was used to calculate the probability of a T cell belonging to a certain subpopulation. To be more precise, three BXVP models were trained to predict the probabilities that a cell belongs to the CD69-DP, CD4SP and CD8SP subpopulations (denoted as P_0, P_4 and P_8 respectively) separately using the gene expression data. Thus, after predictions, each cell was assigned with three probabilities measuring its maturation in CD4SP and CD8SP lineages. If a cell has a lower probability of being a CD69-DP cell whilst it has a higher probability of being a CD4SP cell or being a CD8SP cell, it is more likely to be a mature CD4SP or CD8SP cell and vice versa. Thus, the $1 - P_0$ and P_4 measure the maturation of a cell in the CD4SP lineage together, while the $1 - P_0$ and P_8 measure the maturation of a cell in the CD8SP lineage together.

Similar to the dimension reduction methods which are used for visualising trajectories, two measures (coordinate x and y) are required to locate cells in the coordinate system. In this case, the two measures are required to describe the T cell maturation in lineage CD4 and CD8 separately by using $1 - P_0$, P_4 and P_8 (The measures of maturation are denoted as $\text{Maturation}_{\text{CD4SP}}$ and $\text{Maturation}_{\text{CD8SP}}$ respectively). The summation and the multiplication are two common ways of using the positively correlated independent variables to calculate the dependent variables. Thus, we designed four formula as Eq. 1, 2, 3 and 4, then comparing them using the corresponding trajectories generated based on the summation and multiplication equation respectively (Supplementary Fig. 2, Fig. 6). The two plots show that they have similar CD4 and CD8 lineage development trajectories, while the maturation trajectory starting from CD69-DP to CD69DP was better inferred by using the summation equation. According to this finding, the Eq. 1 and 2 were used to calculate the maturation levels of a cell in the CD4SP and the CD8SP lineage separately. Then the coordinates of x and y in the 2-dimensional coordinate system are calculated by scaling Maturation$_{\text{CD4SP}}$ and Maturation$_{\text{CD8SP}}$ by dividing by their maximal values respectively as shown in the Eq. 5 and 6.

$$\begin{cases} Maturation_{CD4SP} = P_4 + 1 - P_0 & \text{Equation 1} \\ Maturation_{CD8SP} = P_8 + 1 - P_0 & \text{Equation 2} \end{cases}$$

$$\begin{cases} Maturation_{CD4SP} = P_4 * (1 - P_0) & \text{Equation 3} \\ Maturation_{CD8SP} = P_8 * (1 - P_0) & \text{Equation 4} \end{cases}$$

$$\begin{cases} Coordinate_x = \frac{Maturation_{CD4SP}}{\max(Maturation_{CD4SP})} & \text{Equation 5} \\ Coordinate_y = \frac{Maturation_{CD8SP}}{\max(Maturation_{CD8SP})} & \text{Equation 6} \end{cases}$$

In the modeling step, the normalised gene expression matrix was used as the input data for training and predicting. For training the CD69-DP models, the cells of CD69-DP subpopulation were used as positive samples, while the other cells in the other four subpopulations were used as negative samples. It is similar to train the CD8SP and the CD4SP models, except that the CD8SP cells and the mature CD4SP cells were used instead of CD69-DP cells as positive samples respectively and their negative samples were accordingly changed as well. The TCR+DP and the immature CD4SP cells (see Sec. 3.2) did not participate in the training step, but they were predicted together with the other cells in the predicting step. The training dataset was formed by randomly selecting 80% cells in each of the five subpopulations. The predicting dataset was form by the other 20% cells in each of the five subpopulations combined with all of the immature CD4SP and TCR+DP cells.

Each of the three models was trained by a randomly generated training dataset and then was used to predict the cells in the corresponding predicting dataset. In order to ensure that each cell was predicted for at least 3 times, the training and predicting datasets were repeatedly generated and used for training and predicting for 60 times (the number of 60 was determined by multiple tests). Thus, after the whole process, each cell was assigned with three sets of probabilities describing the likelihoods that it belongs to the CD69-DP, CD4SP and CD8SP subpopulations, the size of the three sets was depending on how many times the cell was predicted by the models. Then, the average probabilities of the three sets calculated by the CD69-DP, CD4SP and CD8SP models were used as P_0, P_4 and P_8 to calculate the coordinates of the cell in the coordinate system.

3. Results

3.1 The Detection and Evaluation of DE Genes

The PCA transformation is significantly impacted by the genes used in the calculation. It is not wise to use as many genes as possible in PCA, similar to the clustering using whole set of 9960 genes (W-set) in Sec. 3.2, the disadvantages of using too many genes in PCA outweigh the advantages because of the noises brought by those non-DE genes. Moreover, sometimes even the DE genes can cause problems for PCA. For example, Z-set was able to be used to generate well interpretable PCs (Fig. 2), similarly, F-set can also generate similarly interpretable PCs. However, the combined set of DE genes of Z-set and F-set generated much less interpretable PCs (Fig. 3), because, comparing to top DE genes, the lower ranking DE genes may bring more noise than useful information due to the undetectable expression (a similar example can be found in Table 3). According multiple test, using about 100 DE gene could maximise the useful information in PCA, with which at most 3 interpretable PCs (PC1, PC2 and PC3) can be generated (More related information is provided in Supplementary Fig. 3, 4, 5 and 6).

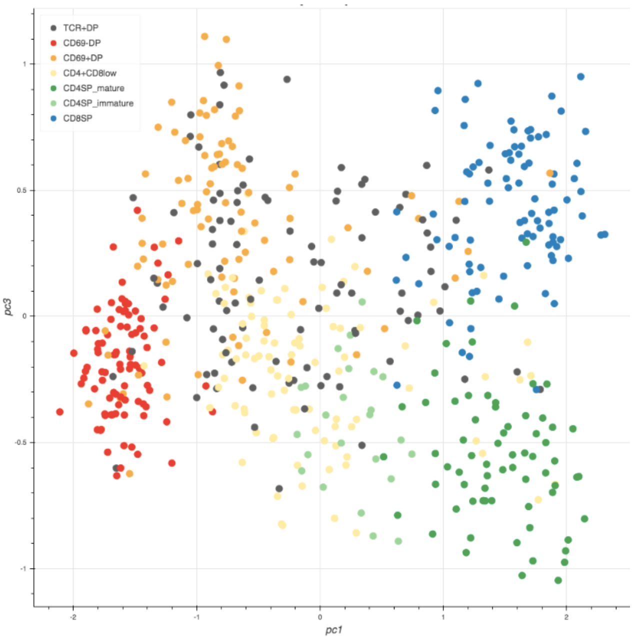


Figure 2: The PC1 versus PC3 of the PCA calculated based on Z-set.

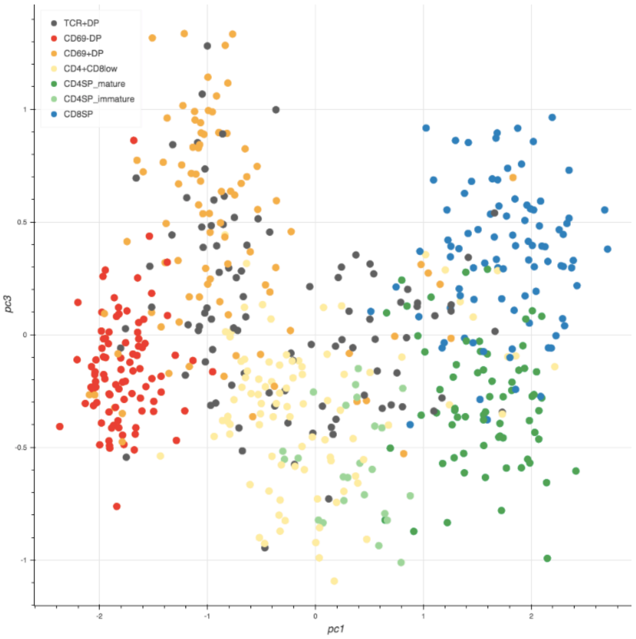


Figure 3: The PC1 versus PC3 of the PCA calculated based on the union of Z-set and F-set.

Z-set and K-set were generated using statistical tests based on the features of RUEs and PELs separately, in order to collect about 100 DE genes for each set, the top 21 and top 27 DE genes were selected from each group using the selection method introduced in Sec. 2.2. As a result, there are 102 and 99 distinctive DE genes in Z-set and K-set respectively shown in Table 2. Then their similarity was checked for verifying the hypothesis proposed in Sec. 1.4. It turned out that the Z-set and the K-set are quite different from each other, there are only 17 common DE genes in both sets while there are 85 and 82 distinct DE genes in Z-set and K-set separately. This finding conflicts with the hypothesis in Sec. 1.4, which shows that the RUEs and the PELs may be regulated by different mechanisms.

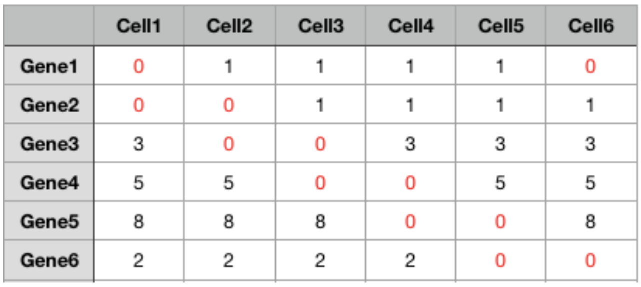
Set of DE Genes	Method	Number of Genes	Average Accuracy Ratio	Standard Deviation of Accuracy Ratios	P Value of the t-test
Z-set	Two proportion z-test based on RUEs	102	86.7%	0.027	N/A
K-set	Kolmogorov-Smirnov test based on PELs	99	77.6%	0.033	1.79E-11
F-set	FindMarkers in Seurat	102	84.9%	0.040	0.060
M-set	Manual selection based on domain knowledge	89	81.6%	0.040	2.50E-05
T-set	DE transcriptional factors	303	53.6%	0.041	< 2.2E-16
R-set	Random selection	100	34.9%	0.045	< 2.2E-16

Table 2: The evaluation of six DE gene sets using the method introduced in Sec. 2.2. The "method" column describes how the corresponding set was detected. The "number of genes" column shows the number of DE genes in each set. The "average accuracy ratio" column and the "standard deviation of accuracy ratios" column show the average accuracy ratio and the standard deviation of the 20 accuracy ratios of the classification tests based on the corresponding set. The "p value of the t-test" column shows the p values of the one-tail two independent samples t-test for comparing the average accuracy ratio of z-set to the one of the corresponding set.

The quality of Z-set and K-set was compared using the evaluation method introduced in Sec. 2.2. Their average accuracy ratios of classification tests were 86.7% and 77.6% respectively. In the corresponding one-tail t-test, a p-value of 1.79E-11 was calculated, this p-value shows the probability that the true difference of the average accuracy ratios between Z-set and K-set is not greater than 0. Thus, it is clear that their difference is significant showing that the Z-set is much more informative than K-set. The other sets of DE genes were also compared with Z-set in the same way for reference, except the F-set which has 72 common DE genes with the Z-set, the Z-set was significantly better than the other sets. Here, the hypothesis of this study proposed in Sec. 1.5 has been examined by the first step, which shows that the feature of RUEs is a potential key to discover more underlying biological mechanisms in the process of T cell maturation in thymus.

3.2 Mature and Immature CD4SP Cell Separation

As mentioned in Sec. 2.3, the immature and mature CD4SP cells should be separated prior to downstream analyses. The W-set and the Z-set whose DE genes were detected based on the feature of RUEs were used to cluster the cells of CD4SP, then two heat maps were plotted to show the distances between any pairs of cells according to the corresponding calculation.



	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6
Gene1	0	1	1	1	1	0
Gene2	0	0	1	1	1	1
Gene3	3	0	0	3	3	3
Gene4	5	5	0	0	5	5
Gene5	8	8	8	0	0	8
Gene6	2	2	2	2	0	0

Table 3: An artificial gene expression matrix. All of the six cells have the same PELs and RUEs for a certain gene. All of the gene have the same RUEs across the cells. However, different cells have different undetectable gene expression based on random selections. Ideally, these cells are identical, thus, the distances between them should be equal to 0. But in this case the Euclidean distance between two cells vary according to the undetected gene expression. This example clearly shows how undetected gene expression affects the cell similarity.

The Fig. 4 shows the heat map of the distances between cells calculated by using W-set, it is clear that, for any one of the cells, it is only itself which is closed to it (the green diagonal). This is probably because there are only a very small number of informative genes differentially expressed among the subpopulations, thus, the other less informative genes provide vast noise when calculating the distances between cells due to the undetected gene expression (Table 3). The Fig.

5 shows the heat map of the distances between cells calculated by using Z-set' which is detected based on the mixed CD4SP cells and other cells. Among those DE gene expressions, although there is still undetected gene expression which brings noise, the RUEs which contribute to the undetected gene expression can also provide useful information for the distance calculation, because the RUEs of those DE genes are probably related to the subpopulations rather than randomly distributed (Sec. 3.1). Thus, the heat map generated based on Z-set' shown in Fig. 5 shows a much better separation for the subpopulation of CD4SP. According to the deduction in Sec. 2.3, it is expected that the larger cluster framed by the blue square is the cluster of mature CD4SP cells.

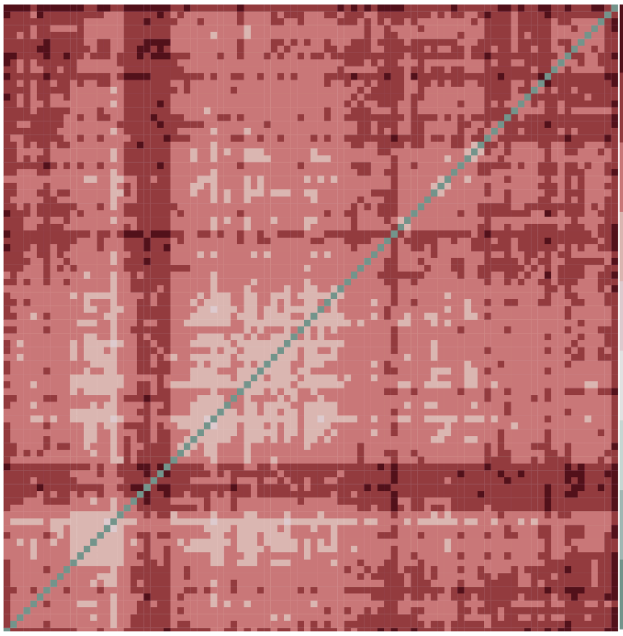


Figure 4: The distance heatmap of CD4SP cells based on W-set. The cells are ranked in the same order from left to right and from bottom to top. Each point represents the distance between 2 cells. The color from green to red corresponds to the distance from small to large.

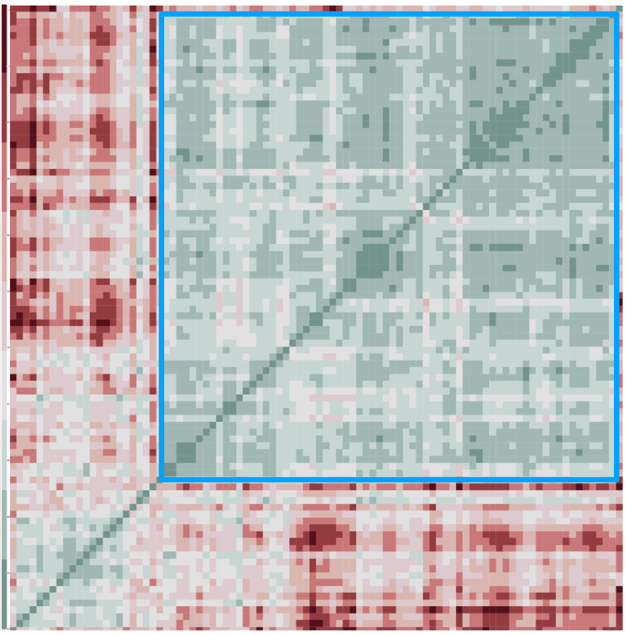
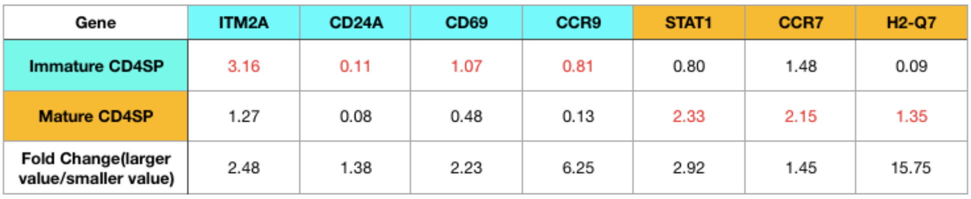


Figure 5: The distance heatmap of CD4SP cells based on Z-set'. The cells are ranked in the same order from left to right and from bottom to top. Each point represents the distance between 2 cells. The color from green to red corresponds to the distance from small to large. The blue square shows a cluster formed by mature CD4SP cells.

Then the seven marker genes were used to verify the separation, the data is shown in Table 4. It is clear that the separation well satisfies the domain knowledge of distinguishing immature and mature CD4SP cells. The immature cells express more ITM2A, CD24A, CD69 and CCR9 genes which are the marker genes for immature CD4SP cells, while the mature cells express more the other three genes which are the marker genes for mature CD4SP cells. In addition, except CD24A and CCR7, the immature and the mature cells have more than a 2-fold change for the other five marker genes. In conclusion, this separation is reliable, the Z-set' was then updated by using the mature CD4SP cells instead of the mixed CD4SP cells then. (The updated Z-set' is the Z-set)



Gene	ITM2A	CD24A	CD69	CCR9	STAT1	CCR7	H2-Q7
Immature CD4SP	3.16	0.11	1.07	0.81	0.80	1.48	0.09
Mature CD4SP	1.27	0.08	0.48	0.13	2.33	2.15	1.35
Fold Change(larger value/smaller value)	2.48	1.38	2.23	6.25	2.92	1.45	15.75

Table 4: The marker gene expression levels of immature CD4SP and mature CD4SP cells. The marker genes of immature CD4SP cells were filled in blue, the marker genes of mature CD4SP cells were filled in orange. The higher value in each gene was written in red.

3.3 Pseudotime analysis

The Fig. 6 shows the T cell differentiation trajectories inferred by BXVP models using the Z-set. (The trajectories inferred by using the other sets of genes can be found in the Supplementary Fig. 7-11) Generally, this plot can be divided into two parts, the left part shows the trajectory of maturation from the origin to the bifurcation in the middle of the plot. This trajectory is mainly formed by CD69-DP cells and a proportion of CD69+DP and TCR+DP cells, it is clear that the lineage decision has not been made in the trajectory. Another interesting finding is that the maturation trajectory seems to be well inferred as a nearly perfect straight line with a slope close to 1, which means that the maturation before the bifurcation does not show any lineage preference. The right part shows the trajectories of CD4SP and CD8SP differentiation towards different directions separately. The mature CD4SP cells were differentiated from the CD4+CD8low and immature CD4SP cells through the CD4SP maturation trajectory, while the CD8SP cells were differentiated from the TCR+DP cells through the CD8SP maturation trajectory. The TCR+DP and CD8SP cells plus some mature CD4SP cells were dispersed in the area (the area of “Both lineages” in Table 5) between CD4SP and CD8SP subpopulations. In addition, there are much less cells in the beginning of CD8 lineage trajectory than CD4 lineage trajectory, which indicates that T cells go to the CD4 lineage first as CD4+CD8low cells. Then some of them express CD8A and CD8B1 genes again and become CD8SP cells through DP cells (the TCR+DP cells in the top half area of Fig. 6). All of these findings are concordant with the KSM, thus, the cell ordering information inferred by BXVP models are useful for further analysis according to the KSM.

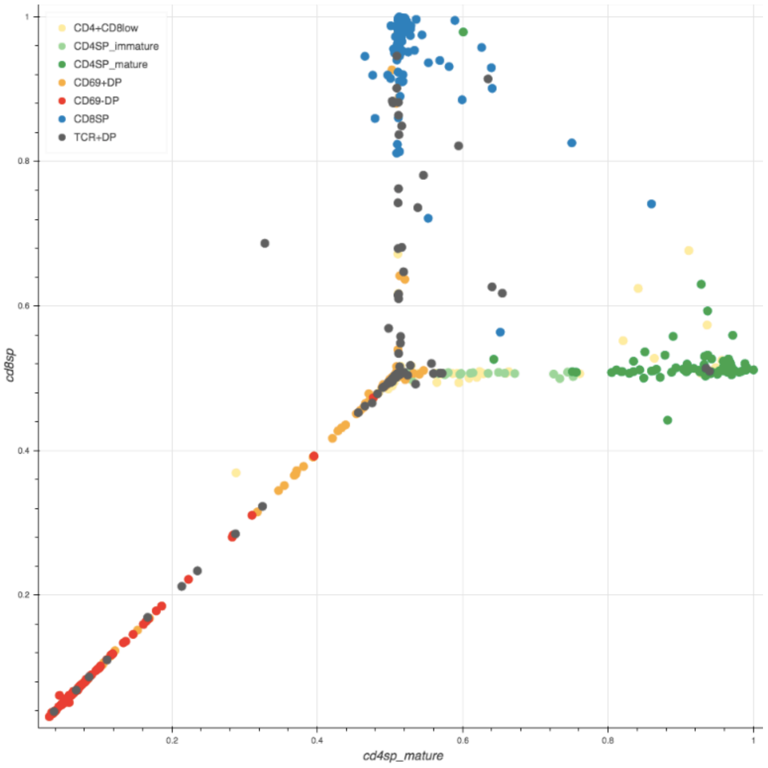
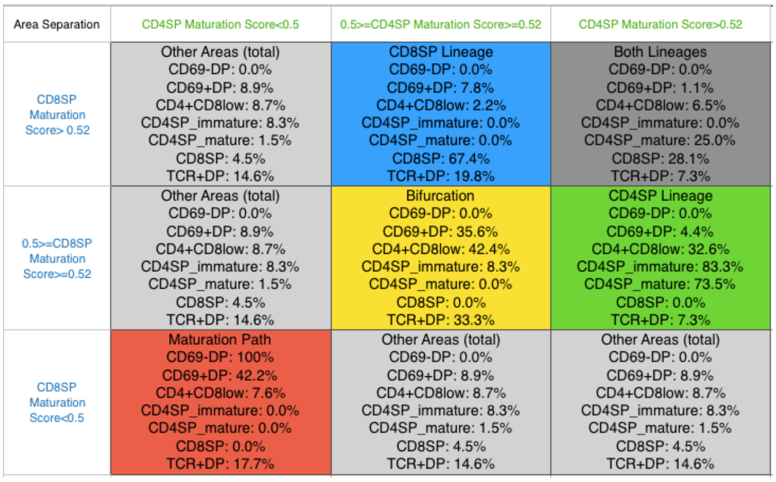


Figure 6: The inferred trajectories of T cell maturation in thymus by BXVP. The x axis indicates the level of maturation of CD4SP lineage while the y axis indicates the level of maturation of CD4SP lineage.

The Table 5 shows how the cell numbers of subpopulations distributed in different areas of Fig. 6. Most of the cells (93.3%) were distributed in the 5 important areas which are “Maturation Path”, “Bifurcation”, “Both Lineages”, “CD4SP lineage” and “CD8SP lineage”, although the they were account for only 50% area of the whole plot. Considering that the subpopulation information of CD69+DP, CD4+CD8low, immature CD4SP and TCR+DP cells had never been used during the BXVP model training, the models can still infer their position in the trajectories quite well according to the KSM, which validates this approach. The Fig. 7 shows how the “Bifurcation” area was identified. There are 19.1% cells mixed in the tiny area, in addition, those cells are in the four subpopulations whose labels were not used for training. Thus, the model may not learn enough useful information to distinguish them along the trajectory. However, the trajectories inferred by BXVP are still much better than the ones inferred by other existing methods.



Area Separation	CD4SP Maturation Score<0.5	0.5>=CD4SP Maturation Score>=0.52	CD4SP Maturation Score>0.52
CD8SP Maturation Score> 0.52	Other Areas (total) CD69-DP: 0.0% CD69+DP: 8.9% CD4+CD8low: 8.7% CD4SP_immature: 8.3% CD4SP_mature: 1.5% CD8SP: 4.5% TCR+DP: 14.6%	CD8SP Lineage CD69-DP: 0.0% CD69+DP: 7.8% CD4+CD8low: 2.2% CD4SP_immature: 0.0% CD4SP_mature: 0.0% CD8SP: 67.4% TCR+DP: 19.8%	Both Lineages CD69-DP: 0.0% CD69+DP: 1.1% CD4+CD8low: 6.5% CD4SP_immature: 0.0% CD4SP_mature: 25.0% CD8SP: 28.1% TCR+DP: 7.3%
0.5>=CD8SP Maturation Score>=0.52	Other Areas (total) CD69-DP: 0.0% CD69+DP: 8.9% CD4+CD8low: 8.7% CD4SP_immature: 8.3% CD4SP_mature: 1.5% CD8SP: 4.5% TCR+DP: 14.6%	Bifurcation CD69-DP: 0.0% CD69+DP: 35.6% CD4+CD8low: 42.4% CD4SP_immature: 8.3% CD4SP_mature: 0.0% CD8SP: 0.0% TCR+DP: 33.3%	CD4SP Lineage CD69-DP: 0.0% CD69+DP: 4.4% CD4+CD8low: 32.6% CD4SP_immature: 83.3% CD4SP_mature: 73.5% CD8SP: 0.0% TCR+DP: 7.3%
CD8SP Maturation Score<0.5	Maturation Path CD69-DP: 100% CD69+DP: 42.2% CD4+CD8low: 7.6% CD4SP_immature: 0.0% CD4SP_mature: 0.0% CD8SP: 0.0% TCR+DP: 17.7%	Other Areas (total) CD69-DP: 0.0% CD69+DP: 8.9% CD4+CD8low: 8.7% CD4SP_immature: 8.3% CD4SP_mature: 1.5% CD8SP: 4.5% TCR+DP: 14.6%	Other Areas (total) CD69-DP: 0.0% CD69+DP: 8.9% CD4+CD8low: 8.7% CD4SP_immature: 8.3% CD4SP_mature: 1.5% CD8SP: 4.5% TCR+DP: 14.6%

Table 5: The distributions of cells for the 7 cell groups in each area of the trajectory plot. The first row shows how x axis is partitioned while the first column shows how y axis is partitioned. The first row in each area is the name of the area. The percentage beside a group in one of the 9 areas represents how many per cents of cells in the given group appear in the given area.

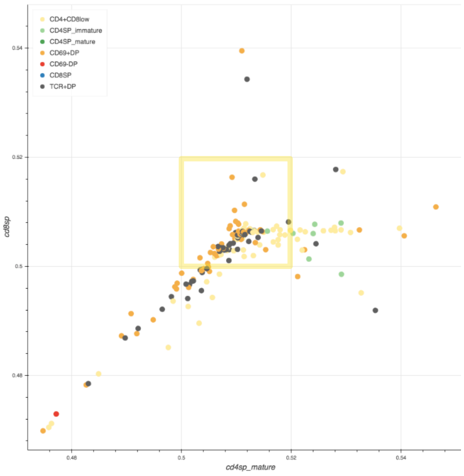


Figure 7: The yellow square shows the area of "Bifurcation" and how the cell distributed in the "Bifurcation" area.

By contrast, although Slingshot is the top-1 method recommended by dynverse from more than 50 trajectory inference methods based on the specific case in this study, it cannot infer equally reasonable trajectories comparing to BXVP, even if z-set were provided to Slingshot as an external enhancer (Fig. 8 and 9). In general, there are two common steps in most of the existing trajectory inference methods, the first step is dimension reduction which aims at projecting the complex gene expression data to a low-dimensional space for visualising the inferred trajectories. Despite the advantages, the dimension reduction step also limits those methods achieving a better performance. Because most of the dimension reduction methods project samples widely distributed in a low-dimensional plot rather than project them strictly ordered in a trajectory. Thus, it is very difficult for researchers to determine the explicit order of cells in an inferred trajectory. It is clear that BXVP models can handle the problems of two-lineage trajectory inference much better than those existing methods by ranking cells more explicitly along the inferred trajectory.

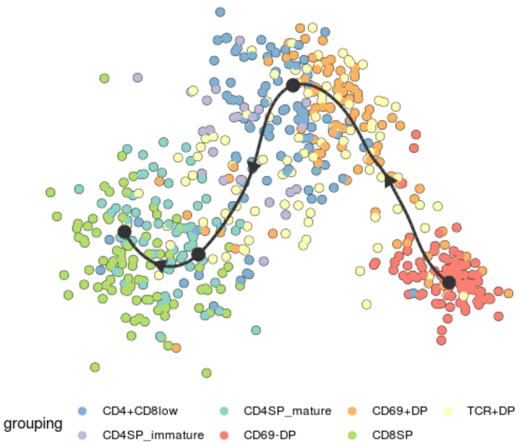


Figure 8: The trajectory of T cell differentiation inferred by Slingshot based on the DE genes detected by itself.

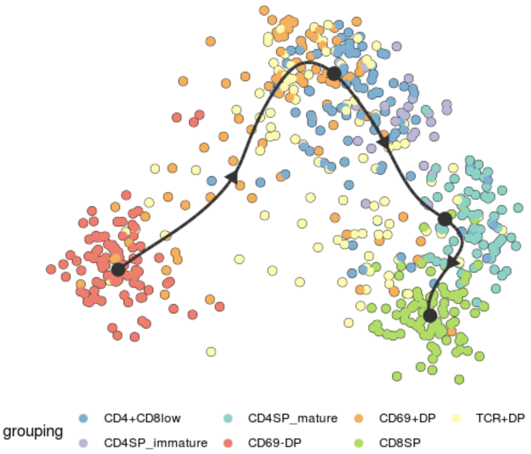
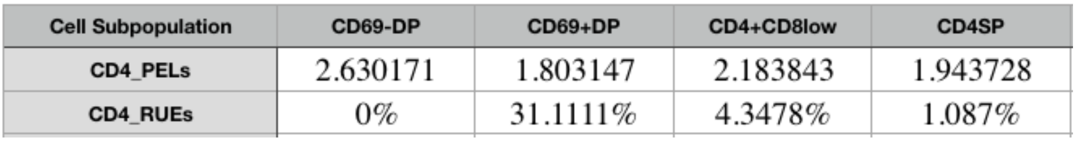


Figure 9: The trajectory of T cell differentiation inferred by Slingshot based on Z-set.

The second step in the majority of pseudotime analysis methods is the trajectory inference. In this step BXVP utilised the prior information to infer the trajectories, in details, CD69-DP cells were assigned as the first stage of the trajectories for model training, while CD4SP and CD8SP cells were assigned as the final stages of the trajectories for model training (This way is achieved by using the Eq. 1, 2, 3 and 4 to calculate the coordinates). Thus, comparing to Slingshot, BXVP requires more prior information to infer the trajectories, which narrows the applications of BXVP. Regardless of the advantages and the drawbacks of BXVP, in this study, it did provide additional useful information about T cell maturation and differentiation in thymus. Here, the second step of examining the hypothesis proposed in Sec. 1.5 has been achieved. In summary, the feature of RUEs was first illustrated to be a potential key to discover underlying biological mechanisms, then the novel trajectory inference method BXVP showed its capability of providing the detailed and reliable information of cell ordering along the well-inferred trajectories of T cell maturation in thymus. Thus, by the two steps of validation, the potential for extending the useful information contained in scRNA-seq data to better understand the process of T cell maturation in thymus has been confirmed.

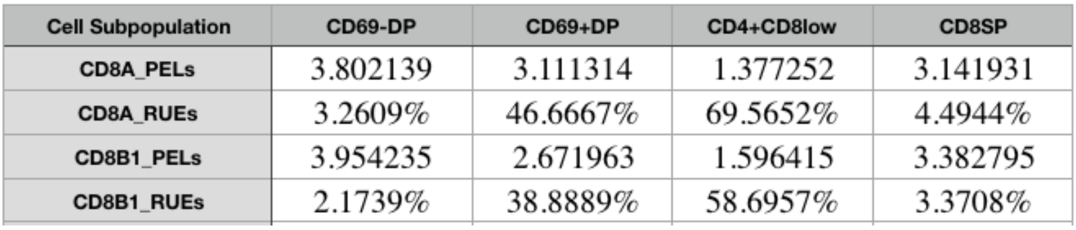
4. Discussion

4.1 The Conjecture about RUEs and PELs



Cell Subpopulation	CD69-DP	CD69+DP	CD4+CD8low	CD4SP
CD4_PELs	2.630171	1.803147	2.183843	1.943728
CD4_RUEs	0%	31.1111%	4.3478%	1.087%

Table 6: The CD4 PELs and RUEs for the four subpopulations in the maturation trajectory of CD4SP lineage.



Cell Subpopulation	CD69-DP	CD69+DP	CD4+CD8low	CD8SP
CD8A_PELs	3.802139	3.111314	1.377252	3.141931
CD8A_RUEs	3.2609%	46.6667%	69.5652%	4.4944%
CD8B1_PELs	3.954235	2.671963	1.596415	3.382795
CD8B1_RUEs	2.1739%	38.8889%	58.6957%	3.3708%

Table 7: The CD8A and CD8B1 PELs and RUEs for the four subpopulations in the maturation trajectory of CD8SP lineage.

The Table 6 and 7 shows the PELs and the RUEs of three representative genes during the maturation of CD4SP and CD8SP lineages. It is clear that, although having similar CD4 PELs, the

RUEs has a nearly 30-fold change between the CD69+DP and the mature CD4SP subpopulations in the CD4SP lineage. In addition, two highly correlated genes of CD8A and CD8B1 show similar patterns of RUEs related to the subpopulations during the differentiation. In details, it starts from a lower RUE in CD69-DP and goes up since CD69+DP until achieving the crest in CD4+CD8low, then the RUE ends up in CD8SP at another lower level similar to the one in CD69-DP. These findings illustrate that RUEs are not always correlated with PELs and it is reasonable to deduce that the RUEs of certain genes are related to the T cell subpopulations.

More diverse examples are shown in Fig. 10. In order to make the plot easier to read, 1 – RUE was used instead of RUE, which in turn allows the plot to show the positive correlation between 1 – RUE and PEL instead of negative correlation which may be more difficult to compare. ITM2A is a good example to show the correlation between RUEs and PELs. Although it is the only example showing the obvious correlation in the plot, there is a larger proportion of genes whose RUEs and PELs have similar correlation. The other four genes in Fig. 10 illustrate different types of relation between RUEs and PELs, which contradict with the correlation shown by ITM2A. Thus, it is reasonable to infer that there are more reasons, such as underlying biological mechanisms, for the occurrences of the undetected gene expression besides the dropout events and SVG.

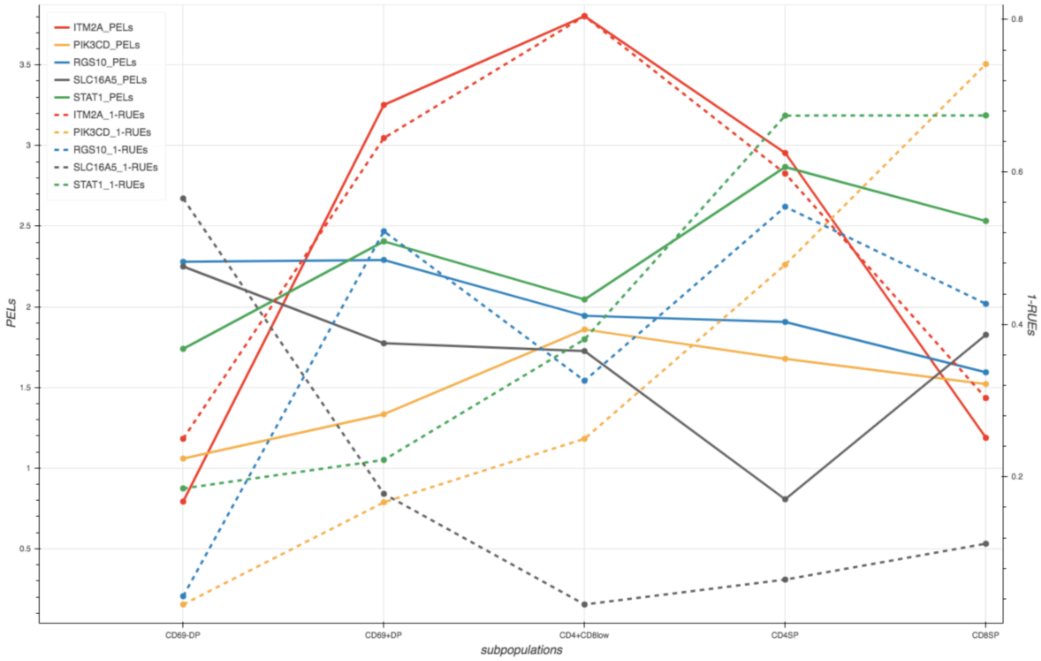


Figure 10: The PELs and 1-RUEs of ITM2A, PIK3CD, RGS10, SLC16A5 and STAT1. The PELs were indicated by solid lines based on the principal y axis on the left side while the 1-RUEs were indicated by dashed lines based on the secondary y axis on the right side. The five genes were distinguished by the colors. The PELs and 1-RUEs of ITM2A are positively correlated, while the PELs and the 1-RUEs of the other four genes were not always positively correlated. This fact shows that the PELs and the RUEs may be regulated by different mechanisms.

According to these findings, it is reasonable to conjecture that the gene expression is regulated by two distinct types of biological mechanisms. The first type of mechanisms takes in charge of turning the expression on for a period of time and then shut it down for another period of time for each gene separately, which in turn affects the RUEs. The process of the regulation is repeated regularly and steadily in different timelines in homogeneous cells according to the finding that cells in the same population simultaneously express different genes which are all generally expressed in their population. (For example, A and B are two genes expressed by a certain group of cells. At one time, cell X only expressed A while cell Y only expressed B, so that they are not in the same timeline) The other one controls the intensity of the expression for each gene when the corresponding gene expression is turned on by the first type of mechanisms, which in turn affect the distribution of PELs. Although these two types of mechanisms may interact with each other, they regulate the RUEs and PELs respectively, because there are so many examples showing the non-correlation between RUEs and PELs found in this study. In future studies, the process of T cell maturation in thymus can be illustrated more clearly by understanding how the gene expression is regulated from the aspects of time and intensity.

4.2 Future work

ScRNA-seq analysis and biological experiments are the good complement for each other, thus, the useful findings and insights discovered in this study deserve further research by biology experiments. An exciting insight in this study is the conjecture about RUEs in Sec. 4.1, which is not only a good feature to detect DE genes for scRNA-seq analysis, but also a potential key to discover more biological mechanisms. As the relation between RUEs and cell populations is discovered by this study, the following question why this relation is existent is needed to be researched by biological experiments. Furthermore, the conjecture that gene expression is simultaneously regulated by two types of mechanisms which control the activation and the intensity of gene expression respectively is also needed to be further verified. In addition, testing these findings by using other sources of scRNA-seq data may provide more evidence to researchers for further analysis.

The well-ordered cells in the trajectories inferred by BXVP allow the research on the continuous gene expression changing along the trajectories, which may contribute to the understanding of the process of T cell maturation in thymus significantly. Furthermore, another interesting finding is that the T cells did not show any lineage preference when maturing along the trajectory with slope close to 1, it is also worth to be delved deeper, which may let more related discoveries emerge.

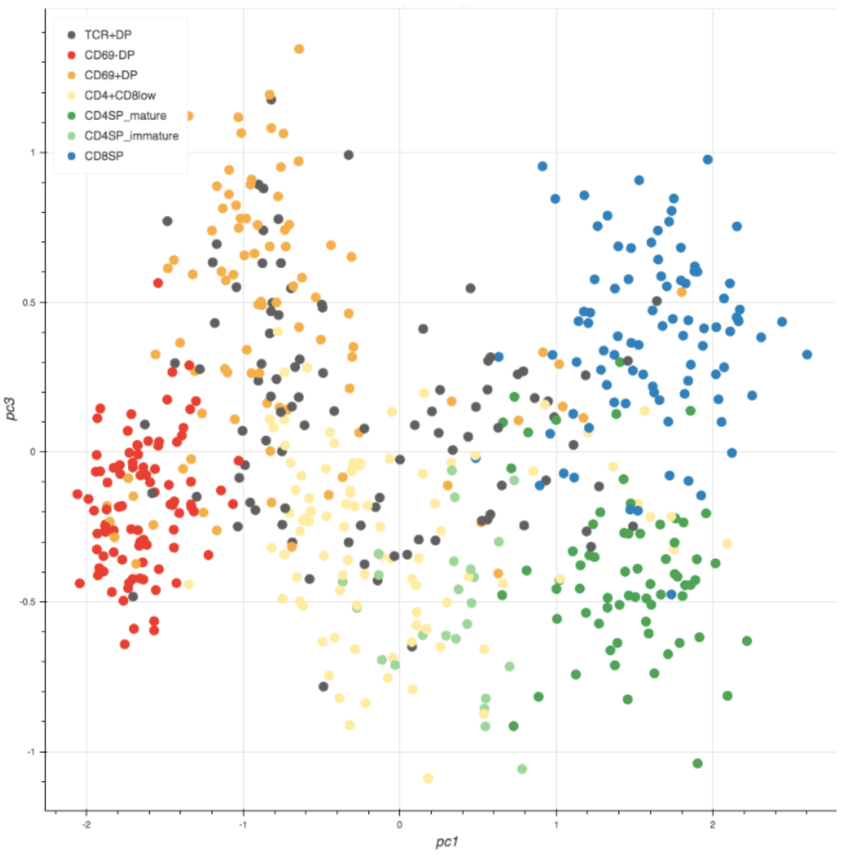
More work can be done to optimise the algorithm of DE gene detection by considering RUEs and PELs together when ranking DE genes. In addition, the manual selection step can be also developed by setting up more specific rules based on research purposes. (The whole analysis pipeline can be found in my Github repository at https://github.com/cxkai2008/DataScienceTools/)

References

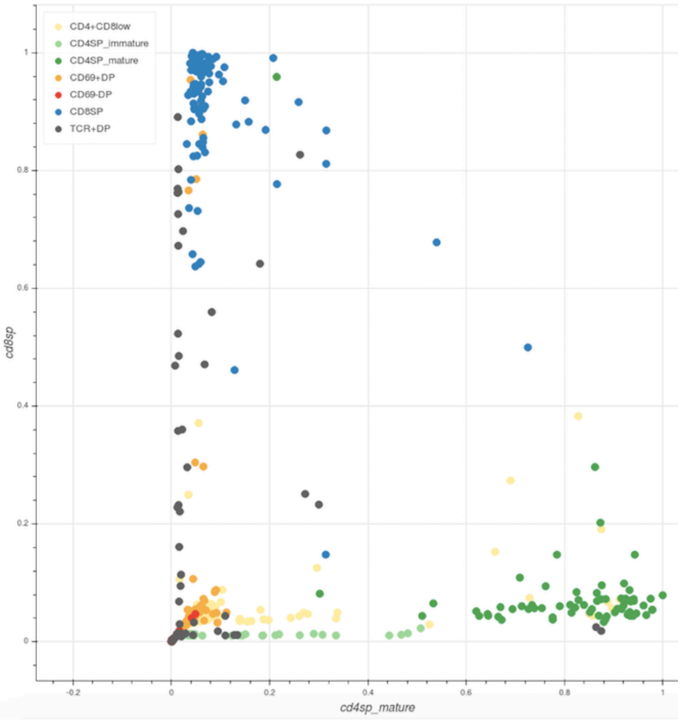
1. Hwang, B. et al. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. 50:96. doi: 10.1038/s12276-018-0071-8
2. Tang, F. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382 doi: 10.1038/nmeth.1315.
3. Shalek, AK. et al. (2014). Single cell RNA Seq reveals dynamic paracrine control of cellular variation. Nat. Jun 19;510(7505):363-9 doi: 10.1038/nature13437.
4. Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. Development 136: 3853-3862; doi: 10.1242/dev.035139
5. Linnarsson, S. et al. Single-cell genomics: coming of age. Genome Biol. May 10;17:97. (2016). doi: 10.1186/s13059-016-0960-x.
6. Cannoodt R. et al. (2016). Computational methods for trajectory inference from single-cell transcriptomics. Eur J Immunol. Nov;46(11):2496-2506. doi: 10.1002/eji.201646347.
7. Stubbington, MJT. et al. (2016). T cell fate and clonality inference from single cell transcriptomes. Nat Methods. Apr; 13(4):329-332. doi: 10.1038/nmeth.3800.
8. Proserpio, V. et al. (2016). Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. Genome Biol. May 12;17:103. doi: 10.1186/s13059-016-0957-5.
9. Miragaia, RJ. et al. (2019). Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation. Immunity. Feb 19;50(2):493-504.e7. doi: 10.1016/j.immuni.2019.01.001.
10. Owen, J. et al. (2013). Kuby Immunology 7th edition. W. H. Freeman; Stranford. ISBN-10: 1464137846
11. Yamashita, I. et al. (1993). CD69 cell surface expression identifies developing thymocytes which audition for T cell antigen receptor-mediated positive selection. Int Immunol. Sep;5(9):1139-50. doi: 10.1093/intimm/5.9.1139
12. Brugnera, E. et al. (2000). Coreceptor reversal in the thymus: signaled CD4+8+ thymocytes initially terminate CD8 transcription even when differentiating into CD8+ T cells. Immunity. Jul;13(1):59-71. doi: 10.1016/S1074-7613(00)00008-X
13. Singer A. (2002). New perspectives on a developmental dilemma: the kinetic signaling model and the importance of signal duration for the CD4/CD8 lineage decision. Curr Opin Immunol. Apr;14(2):207-15. doi: 10.1016/S0952-7915(02)00323-0
14. Singer, A. et al. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. Nat Rev Immunol. Oct;8(10):788-801. doi: 10.1038/nri2416.
15. Picelli, S. et al. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. Nov;10(11):1096-8. doi: 10.1038/nmeth.2639.
16. Kim, D. et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. Apr 25;14(4). R36. doi: 10.1186/gb-2013-14-4-r36
17. Dobin, A. et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics. Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635.
18. Bray, NL. et al. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. May;34(5):525-7. doi: 10.1038/nbt.3519.
19. Liao, Y. et al. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. May 7;47(8):e47. doi: 10.1093/nar/gkz114.
20. La Manno, G. et al. (2018). RNA velocity in single cells. Nature. Aug;560(7719):494-498. doi: 10.1038/s41586-018-0414-6.
21. Love, MI. et al. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
22. Stuart, T. et al. (2019). Comprehensive integration of single cell data. Cell. Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031.
23. Hastie, T. et al. (2009). "14.3.12 Hierarchical clustering". The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 520–528. ISBN 978-0-387-84857-0.

24. Lloyd, S. (1985). Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982 Mar;28(2):129-37. DOI: 10.1109/TIT.1982.1056489
25. Kiselev, V. et al. (2019). Analysis of single cell RNA-seq data. University of Cambridge Bioinformatics training unit. https://scrnaseq-course.cog.sanger.ac.uk/website/index.html
26. Saelens, W. et al. (2019). A comparison of single-cell trajectory inference methods. Nat Biotechnol. May;37(5):547-554. doi: 10.1038/s41587-019-0071-9
27. Street, K. et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. Jun 19;19(1):477. doi: 10.1186/s12864-018-4772-0.
28. Vovk, V. et al. (2014). Venn-Abers predictors. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence: AUAI Press. 829–838.
29. Chen, T. et al. (2016) XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: ACM. 785-794.
30. Cui, X. (2019) Using Machine Learning to Predict Genotoxic Mode of Action from High Content Imaging Data. Available at https://github.com/cxkai2008/DataScienceTools
31. Butler, A. et al. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 36;411–420. doi: 10.1101/164889
32. Li, W. et al. (2019). A statistical simulator scDesign for rational scRNA-seq experimental design. Bioinformatics. Jun 35;14:41–50 doi: 10.1093/bioinformatics/btz321
33. Zhang, XF. et al. (2019). EnImpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning. Bioinformatics. May 24. pii: btz435. doi: 10.1093/bioinformatics/btz435
34. Munsky, B. et al. (2012). Using gene expression noise to understand gene regulation. Science. Apr 13;336(6078):183-7. doi: 10.1126/science.1216379.
35. Kharchenko, PV. et al. (2014). Bayesian approach to single-cell differential expression analysis. Nat Methods. Jul;11(7):740-2. doi: 10.1038/nmeth.2967
36. Ziegenhain, C. et al. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. Feb 16;65(4):631-643.e4. doi: 10.1016/j.molcel.2017.01.023.
37. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/
38. Jones, E. et al. (2001), SciPy: Open Source Scientific Tools for Python, Available at http://www.scipy.org/
39. Kullback, S. et al. (1951). "On information and sufficiency". Annals of Mathematical Statistics. 22 (1): 79–86. doi:10.1214/aoms/1177729694.
40. Pedregosa, F. et al. (2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011. Available at https://scikit-learn.org/stable/
41. Bokeh Development Team (2018). Bokeh: Python library for interactive visualization Available at http://www.bokeh.pydata.org
42. Xu, X. et al. (2013). Maturation and Emigration of Single-Positive Thymocytes. Clin Dev Immunol.2013:282870. doi: 10.1155/2013/282870
43. Wilson, A. et al. (1988). Subpopulations of mature murine thymocytes: Properties of CD4−CD8+ and CD4+CD8− thymocytes lacking the heat-stable antigen. Cell Immunol. Dec;117(2):312-26. doi: 10.1016/0008-8749(88)90121-9

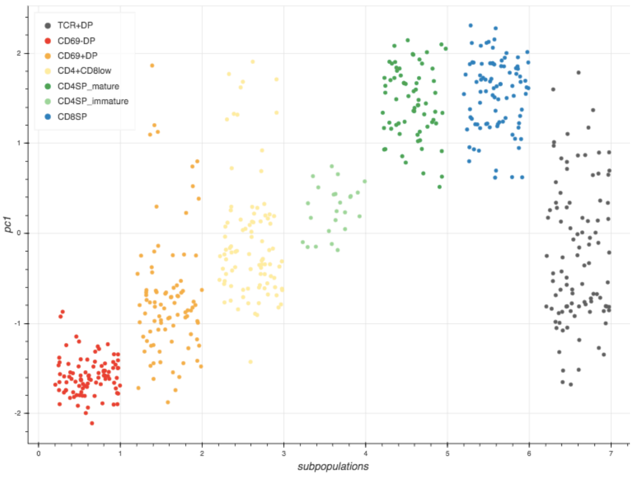
Supplements



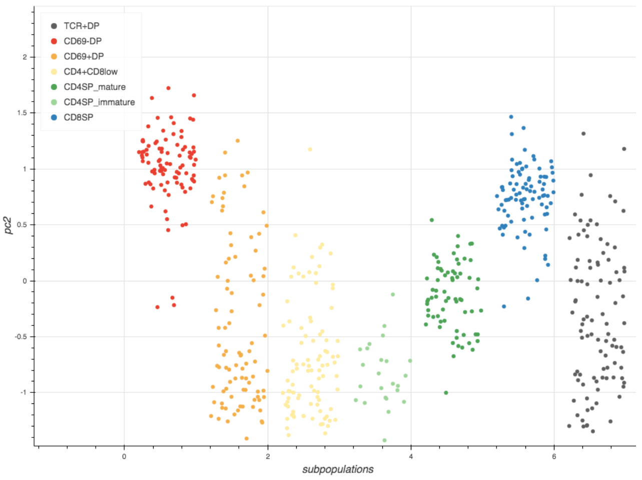
Supplementary Figure 1: The PC1 versus PC3 of the PCA calculated based on F-set.



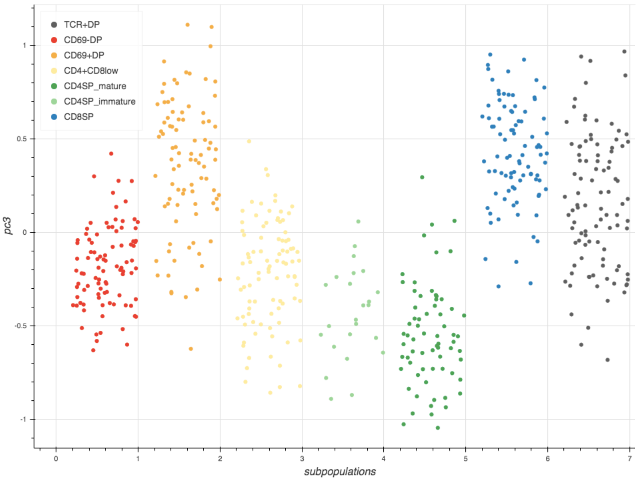
Supplementary Figure 2: The trajectory inferred by BXVP models based on Eq. 3 and 4. The trajectories of CD4 and CD8 lineages were inferred well, however the maturation trajectory from CD69-DP to CD69+DP cannot be revealed from this plot.



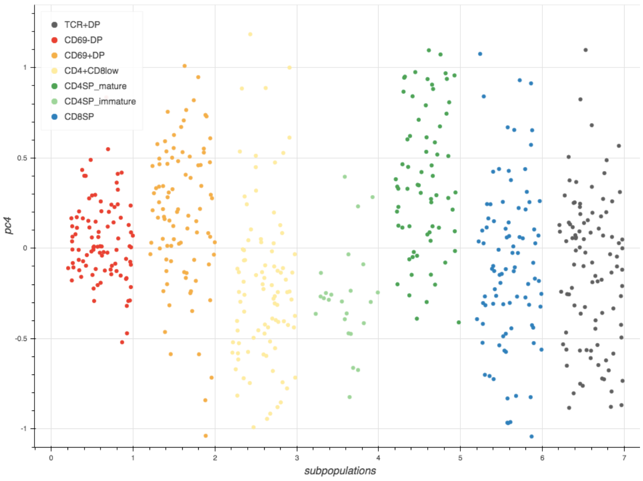
Supplementary Figure 3: The PC1 distributions of 7 groups of T cells calculated based on Z-set. For each group, the cells are randomly distributed in certain small range of x axis. The y axis represents the value of the PC1. It is clear that the PC1 generally goes up from CD69-DP cells through CD69+DP and CD4+CD8low cells to mature CD4SP and CD8SP cells. Thus, PC1 can be interpreted as the feature describing the maturation in the development of CD4SP and CD8SP lineages.



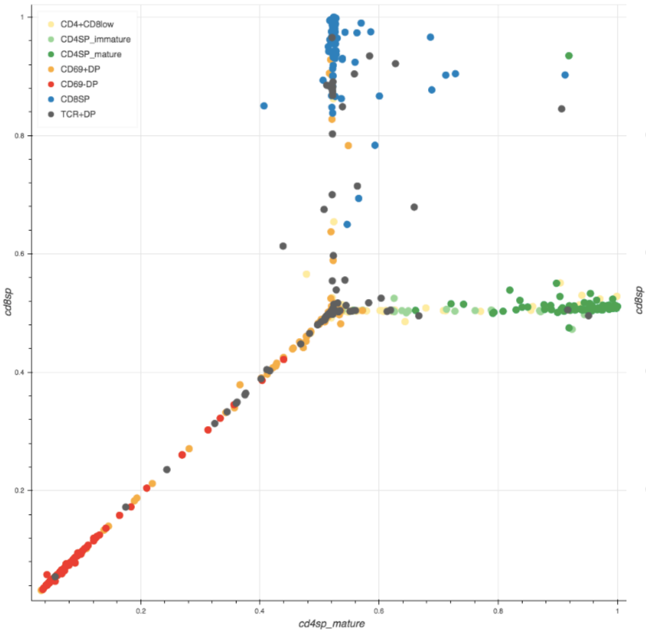
Supplementary Figure 4: The PC2 distributions of 7 groups of T cells calculated based on Z-set. For each group, the cells are randomly distributed in certain small range of x axis. The y axis represents the value of the PC2. It is clear that cells which are at either the starting or the end stage have relatively higher values of PC2, while, at the intermediate stages, cells have relatively lower values of PC2. Thus, PC2 can be interpreted as the feature describing the opposite side of transiently activation during the differentiation. However, PC2 is not as interpretable as PC1, because, unlike PC1, it does not treat the CD4SP and CD8SP subpopulations together as the final.



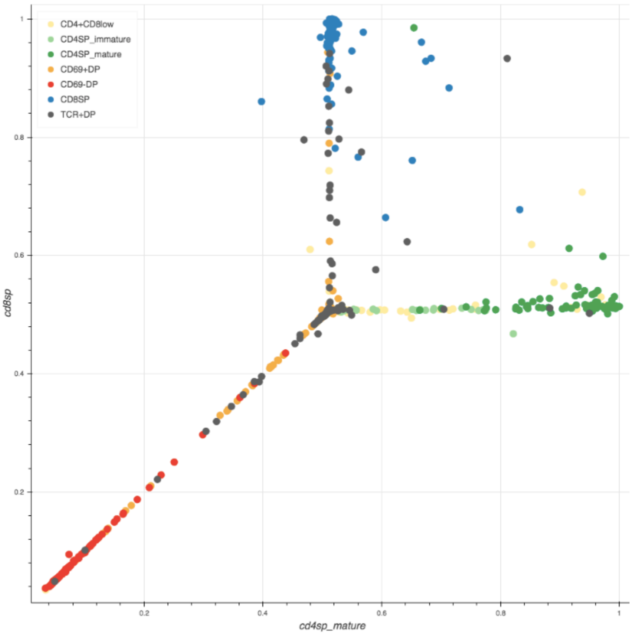
Supplementary Figure 5: The PC3 distributions of 7 groups of T cells calculated based on Z-set. For each group, the cells are randomly distributed in certain small range of x axis. The y axis represents the value of the PC3. It is clear that the mature CD4SP cells have relatively higher values in PC3 while CD8SP cells have relatively lower values in PC3. Although PC2 are also able to separate the majority of CD4SP cells from the CD8SP cells, PC3 shows a better capability of describing the lineage preference, because the subpopulations before the lineage decision failed to show obvious propensity in PC3.



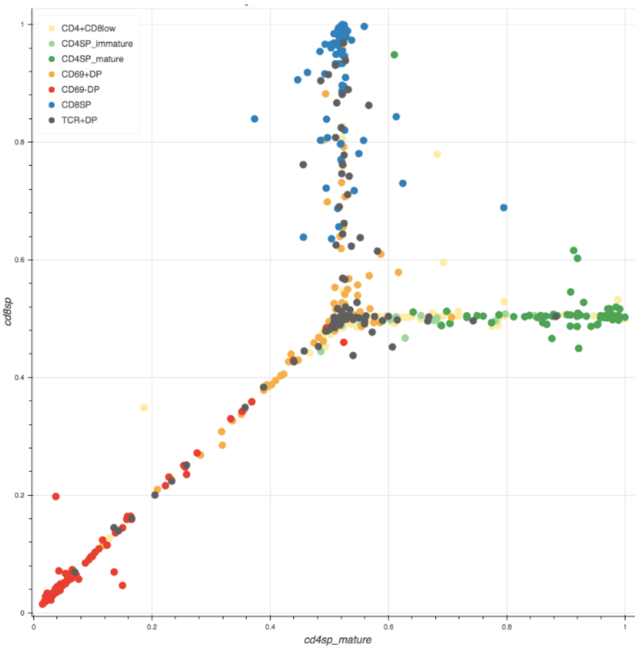
Supplementary Figure 6: The PC4 distributions of 7 groups of T cells calculated based on Z-set. Although it may have the ability of distinguishing some CD4+CD8low cells from CD69+DP and mature CD4SP cells, it is clear that the PC4 are not as informative as the PC1, PC2 or PC3. The situation was quite similar for the PCs calculated based on the other sets of genes, thus, the PC1, PC2 and PC3 were used to evaluate the gene sets.



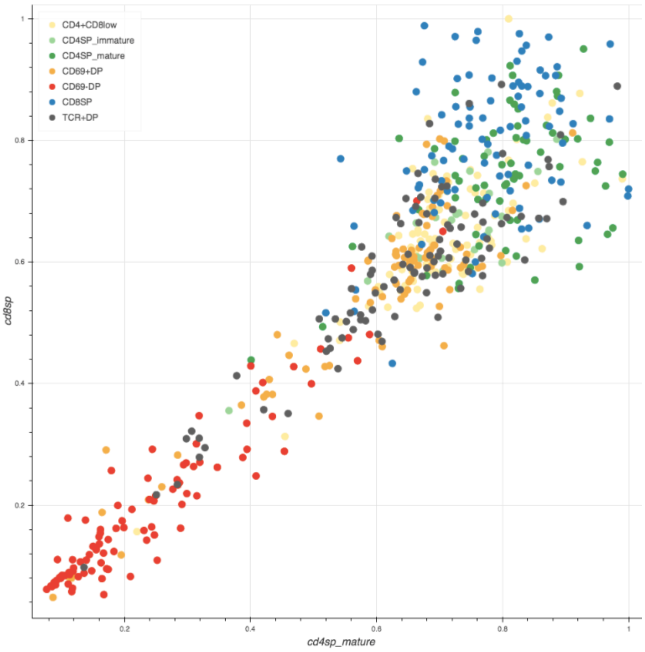
Supplementary Figure 7: The inferred trajectories of T cell differentiation by BXVP using K-set. The x axis represents the level of maturation of CD4SP lineage while the y axis represents the level of maturation of CD4SP lineage.



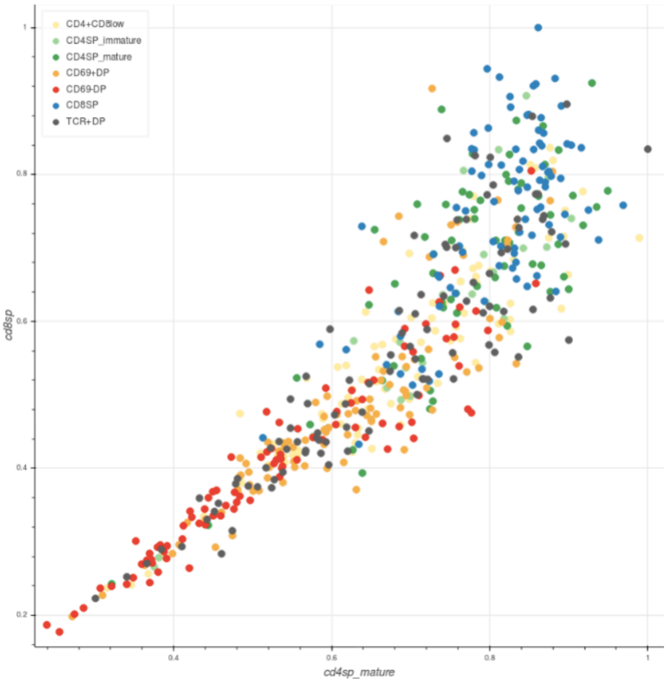
Supplementary Figure 8: The inferred trajectories of T cell differentiation by BXVP using F-set. The x axis represents the level of maturation of CD4SP lineage while the y axis represents the level of maturation of CD4SP lineage.



Supplementary Figure 9: The inferred trajectories of T cell differentiation by BXVP using M-set. The x axis represents the level of maturation of CD4SP lineage while the y axis represents the level of maturation of CD4SP lineage.



Supplementary Figure 10: The inferred trajectories of T cell differentiation by BXVP using T-set. The x axis represents the level of maturation of CD4SP lineage while the y axis represents the level of maturation of CD4SP lineage.



Supplementary Figure 11: The inferred trajectories of T cell differentiation by BXVP using R-set. The x axis represents the level of maturation of CD4SP lineage while the y axis represents the level of maturation of CD4SP lineage.