DOUBLETDETECTOR: A METHOD TO DETECT DOUBLET CELLS IN OPEN CHROMATIN REGIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Alper Eroğlu
September 2020

DoubletDetector: A Method to Detect Doublet cells in Open Chromatin Regions
By Alper Eroğlu
September 2020

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek(Advisor)

Can Alkan

Tolga Can

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

DOUBLETDETECTOR: A METHOD TO DETECT DOUBLET CELLS IN OPEN CHROMATIN REGIONS

Alper Eroğlu
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
September 2020

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a simple and effective technique in genomic studies that shows the chromatin accessibility of the genome. The open regions of the genome play an important role in DNA replication and transcription. It has many practical applications such as nucleosome mapping, identifying regulatory elements, cancer research and immune system aging. With the development of the technology used, this technique is now applied at single cell level in the form of single nucleus ATAC-seq (snATAC-seq). Single cell level resolution helps further the possible implications of ATAC-seq by helping in detection of rare cell types that play roles in the regulatory networks. Like other single cell technologies, snATAC-seq suffers from the existence of doublet cells that occur when multiple cells are simultaneously captured and sequenced which confounds downstream analyses. A unique property of snATAC-seq data is that at a given loci in the genome there can be at most two overlapping reads, one from the maternal and other from the paternal chromosome. When a loci has more than 2 reads this can be due to doublets or alignment/sequencing errors. We propose a count-based method, DoubletDetector, that makes use of this property to detect doublets. It identifies doublets by counting the number of loci within the cell that has more than 2 ATAC-seq reads. It also finds the types of the cells that formed the doublets, to further help understand their nature. DoubletDetector achieved high recall near 90% for detecting simulated doublets in human PBMC and islet snATAC-seq samples. Artificial doublets were then traced back to their cells of origin with near 78% recall using a marker peak-based algorithm. DoubletDetector is the first method to effectively identify both homotypic and heterotypic doublets from snATAC-seq.

Keywords: snATAC-seq, Doublet Cell, Cell Type Annotation.

ÖZET

DOUBLETDETECTOR: AÇIK KROMATİN BÖLGELERİNDE İKİLİ HÜCRE TESPİT ETME ARACI

Alper Eroğlu
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Eylül 2020

Transpozaz-Erişimli Kromatin Dizinlemesi ile Tahlili (ATAC-seq) kromotinin açıklığını ölçmeye yarayan bir genetik yöntemidir. Genomdaki açık bölgeler, DNA ikileşmesi ve transkripsiyonda rol oynamaktadır. Bu tahlilin nükleozom haritalaması, transkripsiyon düzenleyici unsurların tespiti, kanser ve bağışıklık sistemi yaşlanması alanlarında kullanılmaktadır. Teknolojik gelişmeler sayesinde bu teknik, tek hücre seviyesinde kullanılabilir ve buna tek çekirdek ATAC-seq (snATAC-seq) adı verilmektedir. Tek hücre seviyesine inebilen ATAC-seq sayesinde nadir rastlanan hücre tipleri keşfedilebilir ve onların düzenleme ağlarındaki rolleri belirlenebilir. Diğer tek hücre seviyesindeki teknikler gibi snATAC-seq de ikili hücrelerden zarar görmektedir. İkililer, birden fazla hücrenin aynı anda yakalanması ve dizinlenmesi ile meydana gelir ve daha sonra yapılacak çözümlemeleri kötü etkiler. snATAC-seq verisinin bir özelliği, herhangi bir genom bölgesinde sadece annenin ve babanın kromozomlarından gelen iki keşisen okuma olabilir. Eğer bir bölgede ikiden fazla kesişen okuma görülürse, bu; ikili hücreler veya bir dizinleme hatası nedeniyle meydana gelebilir. Biz de bu biyolojik özelliği kullanarak ikili hücreleri tespit etmeye yarayacak DoubletDetector adındaki aracı geliştirdik. Bu araç her hücre için kaç tane genom bölgesinde ikiden fazla kesişen okuma olduğunu saymakta ve bu durumun sık karşılaşıldığı hücereleri işaretlemektedir. Daha sonra bu ikiliyi meydana getiren hücreleri hangi hücre tiplerinden olduğunu da tespit etmektedir. Bu aracı tek çekirdekli çevresel kan hücreleri (PBMC) ve adacık hücreleri ile denedik ve simüle edilen ikili hücreleri %90 oranında doğru tespit edebildik. Bu yapay hücrelerin hangi hücre tiplerinden geldiklerini ise %78 oranında doğruladık. DoubletDetector, başarılı şekilde hem çok-kaynaklı hem de tek-kaynaklı ikili hücreleri tespit edebilen ilk araçtır.

Anahtar sözcükler: snATAC-seq, İkili hücre, Hücre Tipi Belirleme.

Acknowledgement

I want to start by thanking my supervisor A. Ercüment Çiçek who put his trust in me and gave me a position in his lab. I was more than lucky to work closely with an instructor that passed me his passion of teaching during my undergraduate years. He kept on pushing me in my ups and downs, even when I was a closed box. Last year he gave me an opportunity that changed my life. I am forever in his dept.

I also want to thank Duygu Uçar for accepting me in her lab in The Jackson Laboratory and much more. I wouldn't have survived across the pond without her trust, guidance and support. I especially want to thank Asa Thibodeau from Uçar Lab, who played an important role in developing the counting overlaps part of this work.

I want to thank Asst. Prof. Can Alkan and Prof. Tolga Can who were kind enough to be in the thesis committee for my defense. Thank you for going through my thesis and for the valuable feedback.

During my study, I was lucky enough to call some of the brightest people I know, as colleagues and friends. Its always better to go on a journey together rather than alone, so thank you Onur for being a roommate, friend and brother. Cihan and Gizem, I was lucky to be able to call you friends and hope we can make more trips to MADO. Çağlar, FMA, Miray and Ömer, thank you for all the conversations over a cup of coffee. Lets do it again as "Yüksek Mühendis"s.

Finally, I want to thank my family. No form of writing is enough to describe the unconditional support and love my parents gave me. Thank you for making me the person I am today. Nothing makes me happier than making you proud. I also want to thank my sister, who was my muse and gave me inspiration to work harder for the future. Dedem, my hero, I will always cherish and honour you memory.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a widely used method in genomics for measuring chromatin accessibility. In this method, the open regions of the chromatin are sequenced [1]. It is used in research areas such as nucleosome mapping, epigenomics and identifying regulatory elements [2, 3, 4, 5, 6]. There are also applications into cancer research [7]. Recently, ATAC-seq along with other methods such as RNA sequencing and flow cytometry has been used to find the changes in the human immune system with age [8].

Single nucleus ATAC-seq (snATAC-seq) is a far recent improvement that allows mappings at single cell level, instead of the bulk nature of ATAC-seq [9, 10]. This allows for a more-in-depth analysis of scarce cell types and their epigenomics. In this method, each cell is captured inside a droplet and sequenced on its own. A common problem in this method is the formation of doublets or multiplets, where multiple cells are gathered inside the same droplet [11]. This phenomenon causes issues in downstream analysis as the cells in the droplet are sequenced together. There are two type of doublets; homotypic and heterotypic [12]. Homotypic doublets are cases when doublets from the same cell type are caught together in the same droplet. Heterotypic doublets occur when cells from different cell types or lineages are inside the same droplet.

In this thesis, we propose a two-part method called DoubletDetector, that is able to detect heterotypic and homotypic doublets in snATAC-seq data and predict the cell types that formed the heterotypic ones. Figure 1.1 shows an outline of the steps taken by the method. DoubletDetector uses a basic biological fact that no more than 2 ATAC-seq reads should overlap a locus from a single cell. These overlaps are expected to be only from the maternal and paternal chromosomes. If there are more than 2 overlapping reads on a locus, we assume for our method that this is due to;

- the cell being a doublet
- repetitive elements or copy number variations within the locus
- error during sequencing or read alignment

DoubletDetector efficiently counts unique ATAC-seq reads overlapping a locus in a single cell and enumerates loci that are spanned by greater than 2 unique reads. If a cell has many loci that are spanned with greater than 2 unique reads, this is a potential doublet cell. We use the Poisson cumulative distribution function to identify potential doublet cells, applying the Benjamini-Hochberg procedure to Poisson probabilities to correct for multiple hypothesis testing [13].

To understand the nature of these doublets, we apply a second step to find the origin cell types of the cells that formed them. To this end, we cluster the single cells in our data to get populations of different cell types. We then find the peaks or open chromatin regions that are unique to each cell type. Finally, we annotate doublets based on the distribution of their reads on these marker peaks. Shannon entropy of this distribution is used to differentiate between homotypic and heterotypic doublets, as heterotypic doublets are expected to have higher information gain.

We ran DoubletDetector on snATAC-seq data from human peripheral blood mononuclear cell (PBMC) (n=2) and human pancreatic islet (n=2) samples. To evaluate the method, we simulated artificial doublets by combining reads from multiple cells. We created heterotypic doublets by combining cells from different cell type clusters. To create homotypic doublets, we paired cells within the same cell type cluster. We were able to correctly identify these artificial cells as doublets with recall values around 95% for the PBMC samples and around 90% for the islet samples. The performance was highly effected by the read depth of the cells. 5-10% of the real cells in the data were detected as doublets. We tested the annotation pipeline with artificial cells as well. We compared predictions made by the pipeline with the actual origin cell types of cells used to create the artificial cells. We had a recall around 77% for the PBMC samples when the top prediction was used and an average recall of 80% for the islet samples. The majority of the real cells detected were homotypic for both PBMCs and islets.

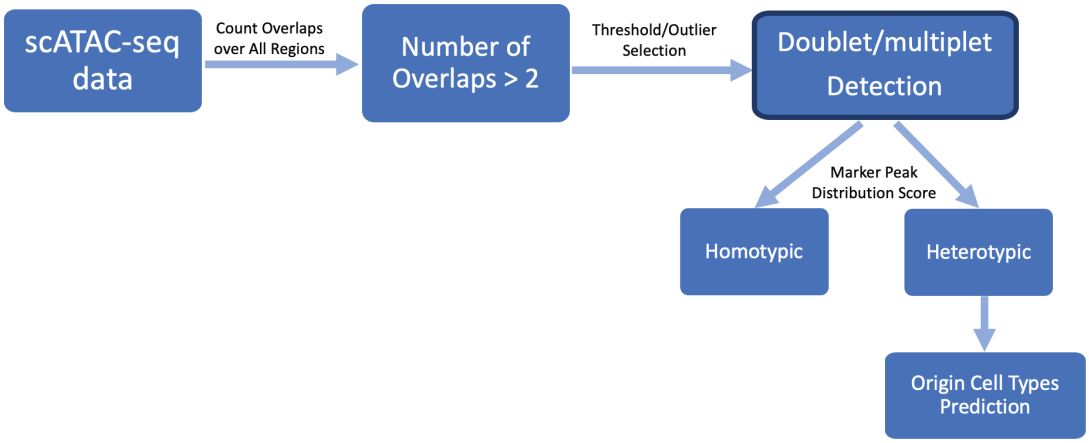


Figure 1.1: The figure shows the outline of DoubletDetector. First the loci with read overlaps greater than 2 are identified. Then Poisson Probability Mass Function is applied with Benjamini-Hochberg correction to detect outlier cells. To annotate the doublet cells, we find marker peaks for each cell type and look at the distribution of cells on these marker peaks. Finally, we predict if a doublet is heterotypic or homotypic and if heterotypic, which cells contributed in its creation.

Chapter 2

Related Work

2.1 ATAC-seq and snATAC-seq

Buenrostro et al. describe ATAC-seq as a method were the Tn5 transposase splits and sequences the open regions of the chromosome [1]. ATAC-seq is described as a method requiring less number of cells and time for library preparation when compared with other chromatin mapping techniques [4]. The open chromatin regions have multiple uses in research. *Schep et al.* have used ATAC-seq data to map the positions of the nucleosome, to better understand how the transcription is regulated [2]. Realtive positions of the nucleosomes are decided with the help of open regions near them detected by ATAC-seq. *Corces et al.* have used ATAC-seq with tumor samples across multiple cancer types [7]. They identified new regions in the non-coding DNA that regulate transcription and extended the network for regulation. *Marquez et al.* used ATAC-seq data along with RNA-seq and flow cytometry to analyze the difference between immune system of aging men and women [8]. They showed that some cell types were acting contrary in men and women and that men aged faster in terms of the immune system.

The single cell level of this method is snATAC-seq. In this method, the cells are captured individually, instead of inside a large bulk, and labeled with a barcode

inside a droplet [14]. The nucleus of the cells inside the droplet are subjected to Tn5 and library preparation is started. Figure 2.1 visualizes this process. *Rai et al.* have showed that using snATAC-seq they can identify rare cell types that play a regulatory role in the Type-II diabetes [15].

2.2 Doublets in single cell analysis

Doublets are a common occurrence in other single cell methods such as scRNA-seq and appear in a similar way as they do in snATAC-seq. Doublets occur when multiple cells are captured inside the same droplet and this occurs when a large quantity of cells are loaded [12, 16]. Large number of cells are loaded in order to have a high throughput of cells that can be further analyzed and to save time. Costly methods such as cell hashing are used for scRNA-seq in which the cells are additionally marked with unique tags that later help point out droplets with multiple tags [17]. Methods like demuxlet are used to differentiate doublets formed when cells from different individuals are loaded together by utilizing the unique genotype properties of each individual by statistical modelling [18]. In terms of detecting doublets from different cell types with the same sample, three frequently used methods exist: DoubletFinder, DoubletDecon and Scrublet [12, 19, 20]. In all of these scRNA-seq methods artificial doublets are simulated and used for classification. DoubletFinder and Scrublet are very similar methods that were published around the same time that make decisions based on the closeness of the cells to these artificial doublets. The number of artificial doublets inside a cell's k-nearest-neighborhood is used in both methods. For DoubletDecon, the approach is slightly different [12]. The gene expression of cells and artificial cells are deconvoluted with each singlet cell cluster as reference. This gives a profile that shows how much each cell cluster contributed to that cell's gene expression. The synthetic doublets are expected to have a mixed profile with influences from multiple clusters. The cells which have similar profiles to the synthetic doublet profiles are labeled as doublets.

In their preprint *Granja et al.* propose a pipeline called ArchR for analyzing snATAC-seq data that also detects doublets inspired by methods used for scRNA-seq cells, especially by DoubletFinder and Scrublet [21]. Again, synthetic doublets are created in the snATAC-seq data by combining the open chromatin reads of two cells. The cells at the proximity of these synthetic cells are labeled as doublets. This method is able to detect heterotypic doublets to some extent but lack detection of homotypic doublets. As an alternative in snATAC-seq research expert knowledge is used for doublet detection. For their paper analyzing human pancreatic islets with *Rai et al.* by integrating their snATAC-seq data with scRNA-seq [15]. They can detect cell type specific markers and find the corresponding genes with that peak. Finally, they label cells that give mixed signatures in different cell type marker peaks as doublets.

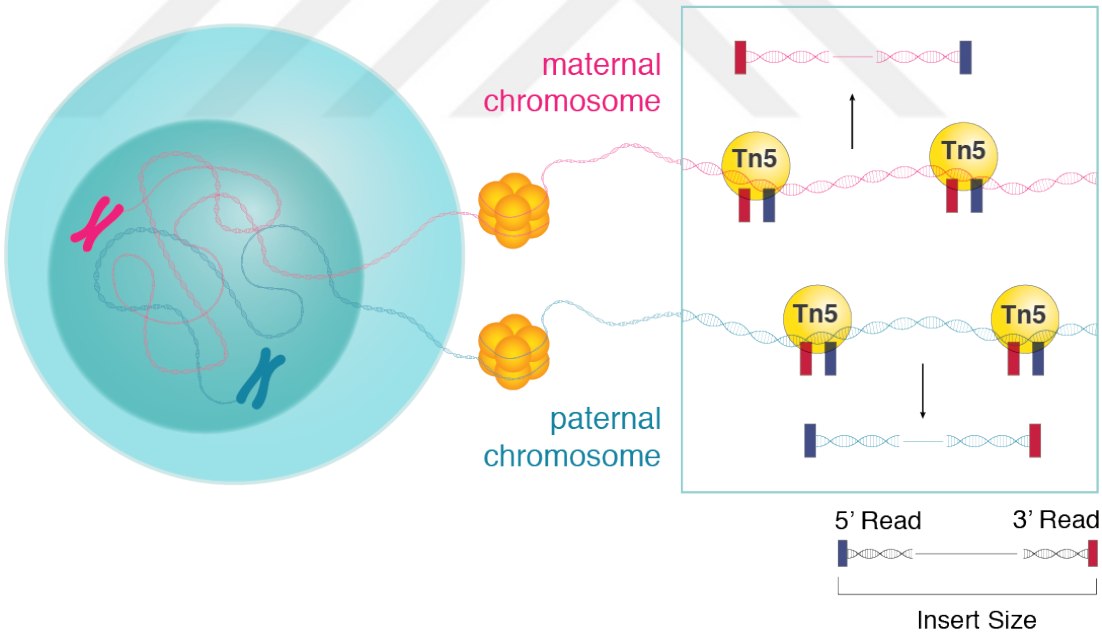


Figure 2.1: The figure shows the sequencing of reads from the maternal and paternal chromosome in snATAC-seq.

Chapter 3

Methods

3.1 Methodological background

3.1.1 Poisson Probability Mass Function

In order to model the number of loci that have overlaps greater than two, we used the Poisson Distribution:

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.1}$$

Where X is a discrete random variable, λ is the expected value of X and k is the desired number of observations. For our context, k is the number of overlaps greater than two. We use this function to determine the probability that the observed number of times an overlap is greater than 2.

3.1.2 Benjamini-Hochberg Procedure

After getting the p-values from our probability function, we use the Benjamini-Hochberg procedure to correct and control the false discovery rate [13]. With this method, we can refrain from incorrectly classifying singlet cells as doublets.

In this procedure, the p-values are sorted in increasing order and each is given a rank based on their position from 1 to the number of values. For every p-value a critical value is calculated based on the rank and number of values:

$$criticalvalue = (i/m)Q \tag{3.2}$$

Where i is the rank of the p-value, m is the number of values and Q is the chosen false discovery rate (FDR). The largest p-value that is smaller than its corresponding critical value is found. All values ranked below this particular p-value are considered significant. For our analysis, we tried two FDR values 0.01 and 0.001 with detecting artificial cells and chose the best one according to the recall.

3.1.3 UMAP

Uniform manifold approximation and projection for dimension reduction (UMAP) is a common technique used for visualization of single cell data [22]. The main goal of the method is to retain the population structure in lower dimensions. For this a weighted graph is formed and the weights are given according to a cell's nearest neighbors. With the application of this tool in single cell analysis, we are able to see clear and distinct cell type populations in the form of clusters.

3.1.4 Marker peak detection in single cell

In single cell analysis a common approach is to find unique features that help separate groups. In scRNA-seq these features are the genes. The unique genes separating a group within the scRNA-seq data are found with the expression of groups for the genes. For snATAC-seq data the features are called peaks. Peaks are the genomic regions that are frequently accessible and the reads are piled up at these locations because of this. In snATAC-seq research, the marker peaks are found similarly to scRNA-seq data. We find the marker peaks in our data with

the FindMarkers feature of the Seurat package, which is commonly used in this domain [23]. Through this method, each feature is given a p-value that shows its significance for being a unique peak. We used the logistic regression test for this procedure described by *Ntranos et al.* [24]. For this test, a logistic regression model is fit to the reads on the peaks. With this model the cluster membership of each cell is predicted. Finally, a likelihood test is performed by comparing this model with a null model. This helps identify peaks that are unique to each cell type cluster.

3.1.5 Shannon entropy

Entropy is the average information inherent in a random variable [25]. It is denoted by the following formula:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log(P(x_i)) \tag{3.3}$$

Where P(x_i) is the probability of outcome x_i of random variable X among n such outcomes. A uniform probability is said to have higher uncertainty and higher entropy. We use this value to assess the how distributed the reads are on the marker peaks. If for a doublet cell, the reads are spread out onto marker peaks of different cell types, it will have a high entropy will be a heterotypic doublet that is combination of multiple cell types. If the reads are piled up on the marker peaks from a single cell type, we will have lower entropy and most likely a homotypic doublet.

3.2 Artificial doublet cell simulation

Both heterotypic and homotypic doublets were simulated by combining the reads of two different cells. Figure 3.1 shows the outline of this process. To create heterotypic doublets; for every cluster combination of two, cells with high read counts were chosen by random sampling from the clusters and paired together for

artificial cell simulation. In default, for each pair of clusters, 100 artificial cells are simulated. To create homotypic doublets; for every cluster, cells with high read count were chosen from that cluster and split into two separate sets. The cells in these sets are paired to create homotypic artificial cells for that cluster. Again, 100 artificial cells are simulated for each of the clusters in default. For both cases of artificial doublet simulation the number of cells simulated is changed according to the number of cells in the parent clusters used.

The barcodes for these pairs of cells were mapped to the same artificial identity so that DoubletDetector can process them as a single cell. First, the number of overlaps greater than 2 are found for these artificial cells. They are later tested in two different approaches, first each one of them is separately added to the existing set of single cells and doublet predictions are made on this updated set. Secondly, to simulate a population of doublets, a group of these artificial cells are added to the existing set of single cells. The group size is 2.5% of the original number of cells and again the number of artificial cells correctly identified in this group is reported.

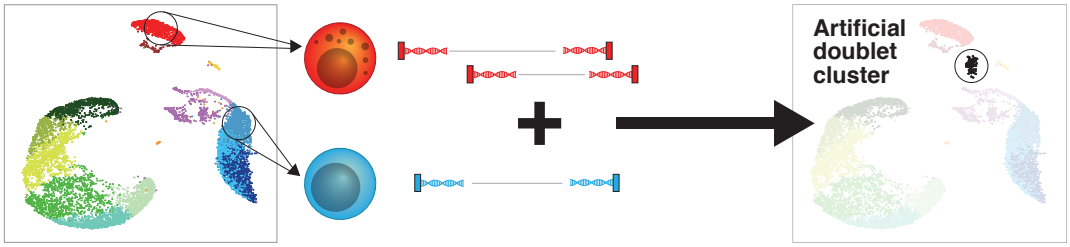


Figure 3.1: The figure shows the creation of the artificial cells. Cells from different clusters are chosen and assigned a combined barcode represents the created doublet cell.

3.3 Count based doublet detection

In this section, we describe the approach to detect doublets.

3.3.1 Identifying snATAC-seq read overlaps

Position sorted paired-end read alignments from snATAC-seq data were compared to detect all genomic regions where greater than 2 reads overlap within a single cell. To reduce overlaps due to alignment errors, we removed all reads with;

1. Mapping quality scores less than or equal to 30
2. Insert sizes (i.e., the end to end distance between 5' and 3' read positions) greater than 900bp (6 nucleosomes) in length

Additionally, we discarded read pairs that aligned to different chromosomes. Reads for each cell and chromosome were compared from start to end efficiently in linear time to identify all instances where reads overlapped greater than a specified threshold (i.e., 2).

To identify instances of greater than 2 reads overlapping, we first identified all genomic region intervals, for which an overlap was observed for at least two paired-end read alignments. Reads defining each interval were then compared to one another to identify all sub-intervals that exceed the specified overlap threshold. To efficiently identify these sub-intervals, for each subset, interval breakpoints were defined at the start and end positions of each paired end read. For each interval break-point, an integer value of 1 was assigned to all breakpoints originating from a start position and -1 to all breakpoints originating from an end position. Interval breakpoints are then visited in order by chromosome position to generate a cumulative sum based on the assigned values at each break-point. The cumulative sum indicates the total number of overlaps between two interval breakpoints and efficiently identifies all sub-intervals with a number of overlaps greater than the specified threshold.

Once all sub-intervals satisfying the threshold are identified for a subset of reads, the algorithm repeats this process for the remaining paired-end read subsets. Each step is performed using a linear time algorithm (i.e., $O(n)$, n is the number of total reads), with an additional $O(\log m)$ (m equals the number of cells) overhead for each read to identify their respective cell origin, resulting in $O(n \log m)$ run time. This algorithm assumes that reads are sorted beforehand and is otherwise superseded by time it takes to sort reads by their chromosome and position (i.e., $O(n \log n)$).

3.3.2 Detecting significant doublets from overlap counts

Genomic regions with greater than 2 reads overlapping from all cells were merged together, combining any common genomic regions that overlapped by at least one base pair. Using this unified list of genomic regions, a binary matrix was generated where columns in the matrix represent the genomic regions with greater than 2 reads overlapping for at least one cell, and the rows represent the individual cells within the sample. Values within the matrix were assigned to 1 if the cell and genomic region combination observed greater than 2 reads overlapping, and 0 otherwise. From this matrix, doublets are detected using rows sums (i.e., the total number of greater than 2 read overlap instances for a specific cell).

The events of observing greater than 2 reads overlapping within the same region for multiple cells or across multiple regions within the same cell can be modeled using the Poisson distribution. Occurrences of these events are independent, counted within set intervals, are either present or not within these intervals, and have a constant average rate of occurring, satisfying all assumptions associated with the Poisson distribution. We therefore detected significant doublets using the Poisson cumulative distribution function, assigning a probability based on the row sums. The means of row sum counts were used as the average rates/expected values to calculate the respective doublet Poisson probabilities. Finally, Benjamini-Hochberg correction was applied to doublet probabilities to

adjust for multiple hypothesis testing. Doublets were predicted by selecting regions or cells with adjusted Poisson probabilities less than 0.01.

3.4 Doublet annotation pipeline

Identified doublets are annotated with regard to the clusters identified in the samples. Figure 3.2 shows the outline of this pipeline. To make sound annotations we re-group the clusters found previously into broader clusters that represent cell types. This helps avoid a possible over-clustering of the data due to batch effects. For the PBMC samples we are able to identify these cell types as naive T, memory CD4 T, memory CD8 T, NK, monocytes and B cells. For the islet samples, these large cell type groups were alpha, beta, ductal and delta cells. Marker peaks are found for these cell type clusters by using the FindMarkers function of Seurat, with the logistic regression test setting. The marker peaks were only found for clusters that have at least 150 cells. For the sake of unison, 100 marker peaks were identified for each of the clusters.

Due to the sparsity of the snATAC-seq data, aggregate reads are calculated for each of the cells that are identified as doublets. This cell is aggregated by taking the median of reads from that cell's 15 nearest neighbors in the singular value decomposition (SVD) reduced dimensions. The procedure of using SVD dimensions is a common method in snATAC-seq data [14]. Cumulative distribution function is used to find the abundance of reads on the cluster's marker peaks for each of the doublet cells. The *ecdf* function of *R* is used for this calculation. The resulting abundance scores sum up to 1. In order to distinguish homotypic and heterotypic doublets, the Shannon Entropy is calculated on these abundance scores. The abundance scores with low entropy hold less information and the reads for that doublet cell are concentrated on one single cluster's marker peaks. These doublet cells are classified as homotypic doublets. The cells that have high entropy are identified as heterotypic doublets. The high/low value for entropy is selected by calculating the entropy score for every cell in the dataset

and the doublets that have a score higher than the median are considered heterotypic. For each heterotypic doublet cell, the clusters are paired by summing their corresponding abundance scores for that particular cell. The pairing that yields the highest total abundance score is reported as the parent clusters that stuck together to form that doublet.

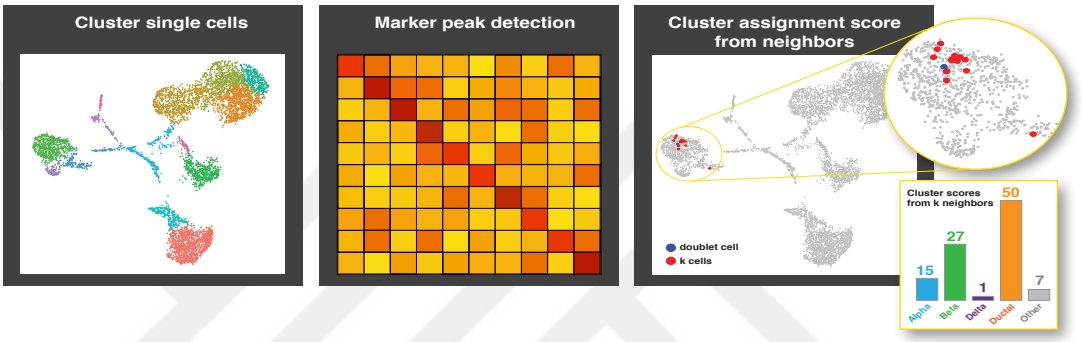


Figure 3.2: The figure shows the annotation pipeline for detecting the origins of the doublets. First the cells are clustered and their cell types are annotated. Second, the marker peaks for these cell types are found. Finally, the nearest neighbor aggregated read count distribution of the cells are used to find if the doublet is homotypic or heterotypic and if heterotypic from which cell types did the cells come from.

Chapter 4

Results

4.1 Datasets

To test the efficacy of DoubletDetector for doublet detection, we generated snATAC-seq maps from 2 human peripheral blood mononuclear cells (PBMCs) and 2 human pancreatic islet samples. These libraries were generated using 10X protocol as described in *Satpathy et al.* [14]. These datasets were pre-processed using CellRanger ATAC software, where the fragments of reads were processed into a computer-readable format that associates the fragments with the corresponding single cell. To identify the cell type specific clusters, the single cells in each sample were separately clustered using a pipeline inspired by *Satpathy et al.* in which the clustering is performed in two steps [14]. The snATAC peaks are called for these clusters using the MACS2 software [26].

4.1.1 PBMC Samples

As a result of the library generation, we obtained 6,143 and 4,974 single cells respectively for each of the PBMC samples. The clustering pipeline resulted in 16 clusters for one of the samples and 15 clusters for the other one. Figure 4.1

shows the clusters identified from the single cell in UMAP space. As a result of peak calling, 174,800 and 188,951 peaks were found by MACS2 respectively. On average, 12,400 reads were detected on the peaks for one of the samples and an average of 18,500 reads were detected on the other sample.

To find the cell types of the clusters, the aggregated snATAC profiles of the clusters were correlated with bulk ATAC profiles of samples that are of a specific cell type. Bulk sample that highly correlated with the cluster was assigned as the cell type of that cluster. Figure 4.2 shows the correlation matrix used for cell type annotations. The clusters are confined in groups of cell type specific samples for annotation.

4.1.2 Islet Samples

The pancreatic islet samples are made up of 6,623 and 5,722 single cells. For the islet samples, 14 clusters were identified for one and for the other 12 clusters were identified. The cell type clusters identified for the first islet sample is shown in Figure 4.3. Upon peak calling 211,767 and 204,566 peaks were obtained. The read counts were on average 8,600 and 8,000 respectively for each of the samples.

Expert knowledge was used to assign the cell types for the clusters in the islet samples.

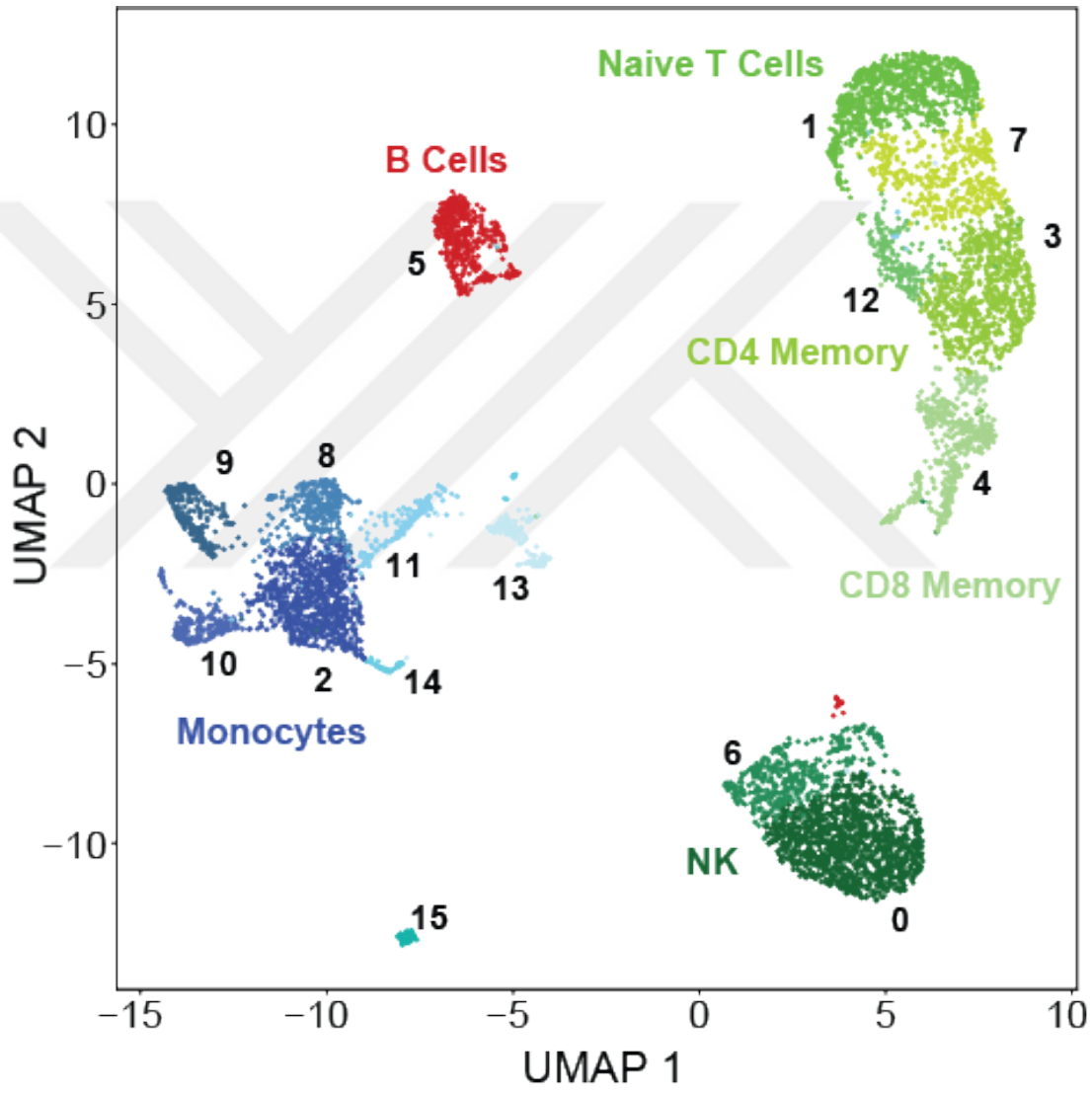


Figure 4.1: The figure shows the single cell clusters that are obtained after the two-step pipeline for one of the PBMC samples. Clustering on the first PBMC sample resulted in 16 clusters that represent different cell type populations. The clusters 0 & 6 are NK cell type clusters. Cluster 1 is Naive T cell cluster. 2, 8-11, 13 & 14 are Monocytes. 3, 4, 7 & 12 are Memory CD4 cells. 5 is a B cell cluster.

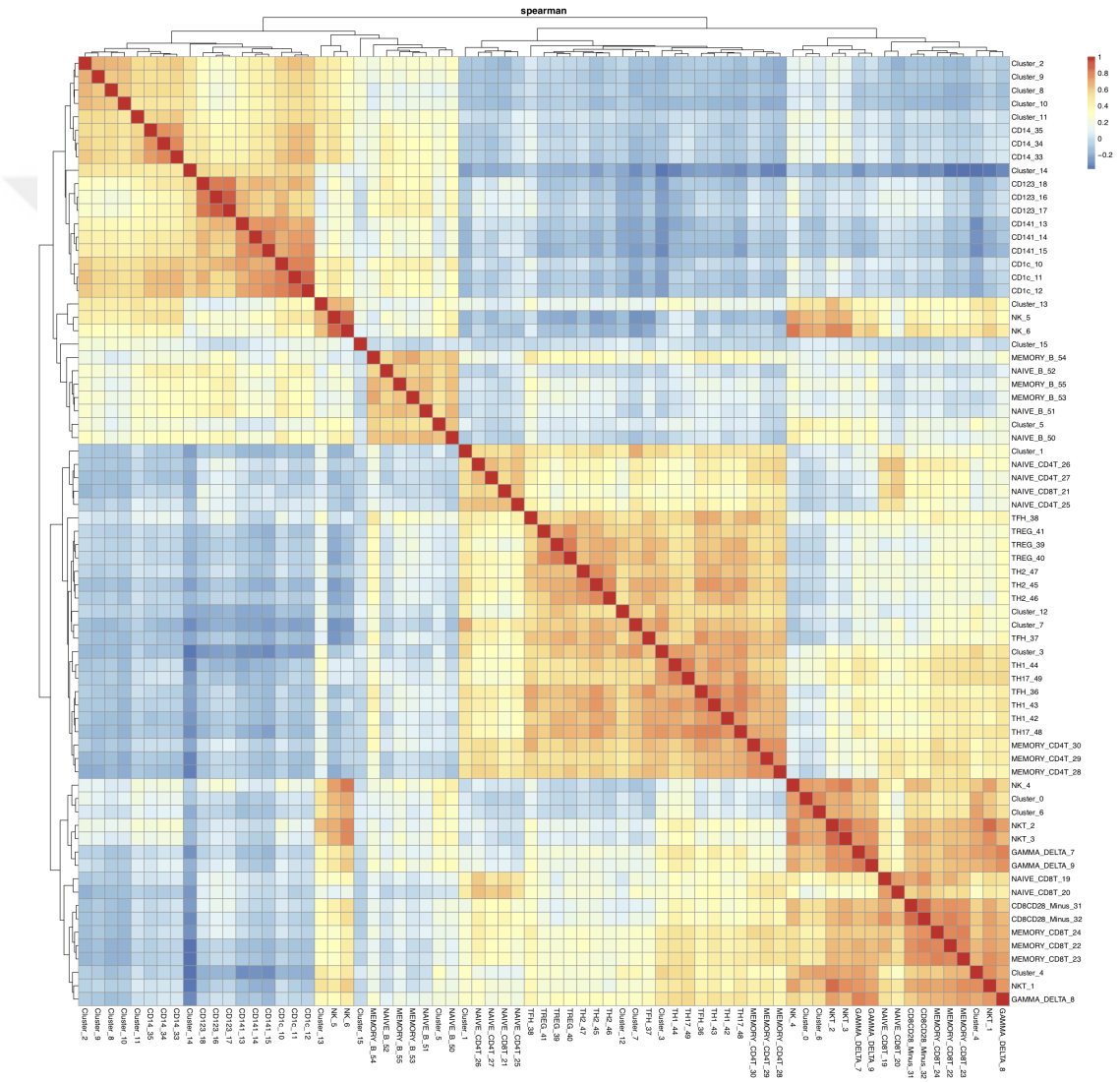


Figure 4.2: The figure shows correlation matrix for the cell type annotation of clusters in PBMC sample 1. The aggregate read count profiles of clusters are correlated with bulk ATAC-seq samples that are sorted. These bulk samples are sorted and only contain cells from a single cell type.

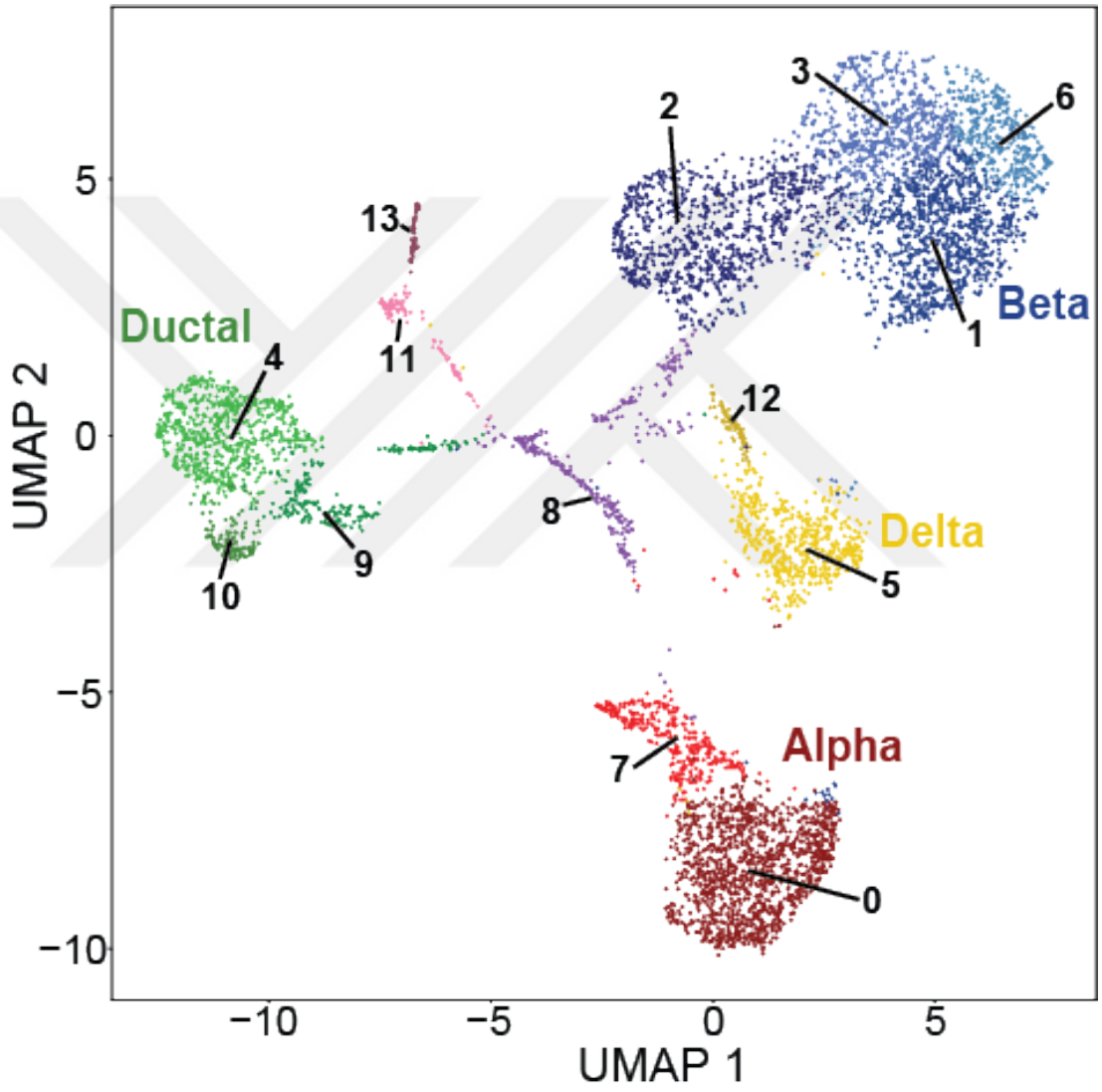


Figure 4.3: The figure shows the single cell clusters that are obtained after the two-step pipeline for one of the islet samples. Clustering on the first islet sample resulted in 14 clusters that represent different cell type populations. Clusters 0 & 7 are Alpha cell clusters. 1-3 & 6 are Beta cell clusters. 5 & 12 are Delta cell clusters. 4, 9 & 10 are Ductal cell clusters.

4.2 Artificial doublet detection

Artificial doublet cells were introduced to evaluate the performance of the method in correctly detecting multiplets. We introduced two different types of doublets for this evaluation step; heterotypic and homotypic. Heterotypic doublets were introduced by combining cells from different clusters and assigning an artificial barcode to these combined transcripts from the cells. Figure 4.4 shows the 100 artificial cells simulated by CD4 Memory and Monocyte cell combinations for PBMC sample 1. These artificial cells were clustered around an island cell cluster between larger cell type clusters. Similarly, Figure 4.5 shows the 100 artificial heterotypic doublets that were added by combining cells from Alpha and Beta cell type clusters for Islet sample 1 in UMAP space.

We first tested the performance of DoubletDetector by evaluating each artificial cell individually, together with the existing cells in the datasets. The recall for detecting the artificial cells of PBMC cells was over 95% for both samples as seen in Figure 4.6 when each cell is tested individually. 16,122 and 13,299 artificial cells are created in this step respectively for each PBMC sample from different clusters. For the Islet samples, more than 91% of the simulated doublet cells were correctly identified as doublets when they were individually classified along with the existing cells as seen in Figure 4.6. 8,291 and 5,536 artificial heterotypic cells are created and evaluated for this analysis.

To create a more realistic setup, we added artificial cells to account for 2.5% of the total population. For the PBMC samples, the recall value was more than 94% for the artificial population analysed along with the actual cells as seen in Figure 4.7. On the other hand, more than 85% of the simulated cells were still correctly classified as doublets out of this artificial population for the Islet samples as shown in Figure 4.7.

Homotypic artificial cells were created by combining cells within the same cluster. More than 87% of the homotypic doublets classified individually with the actual cells were correctly identified by DoubletDetector as doublets for the

Islet samples as seen in Figure 4.8. For this analysis, 650 and 554 artificial cells were created from within the same cluster. As seen in Figure 4.8 recall values were improved for the PBMC homotypic artificial cells, as more than 96% of them were correctly identified. For these samples, 754 and 671 homotypic artificial cells were analyzed individually along with the existing cells.

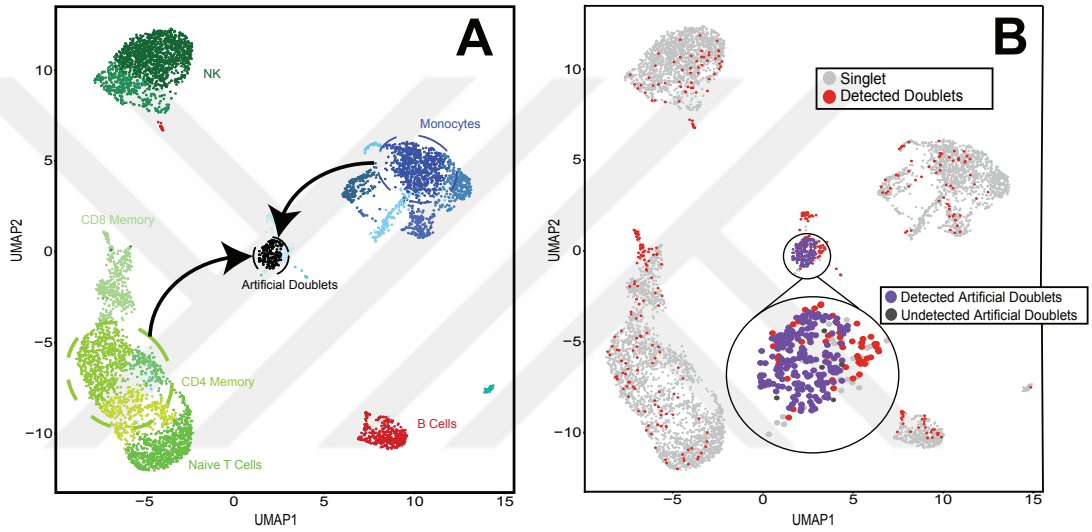


Figure 4.4: The figure shows (A) the 100 artificial cells that are added by the combination of cells from CD4 memory and Monocyte cell clusters. (B) The original and artificial cells that are detected by DoubletDetector for PBMC sample 1.

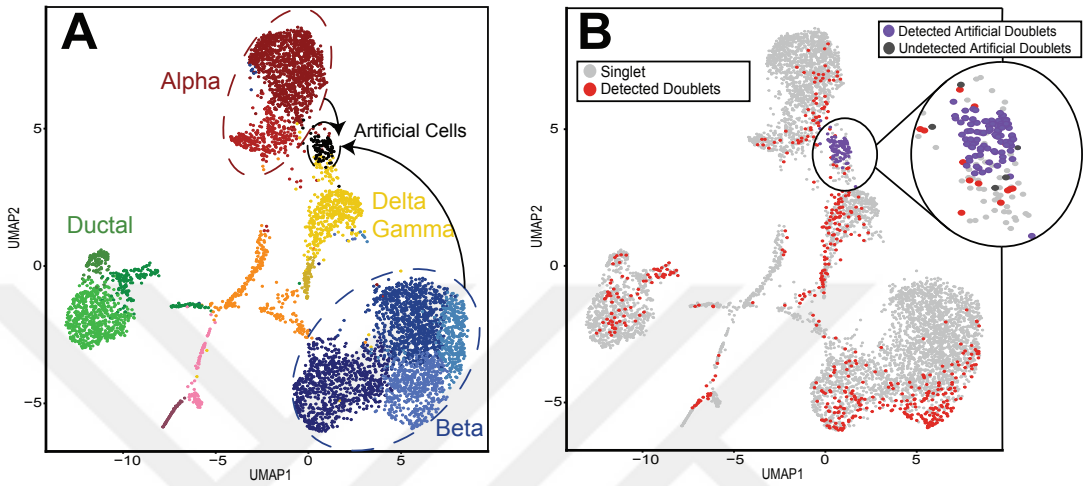


Figure 4.5: The figure shows (A) the 100 artificial cells that are added by the combination of cells from Alpha and Beta cell clusters. (B) The original and artificial cells that are detected by DoubletDetector for islet sample 1.

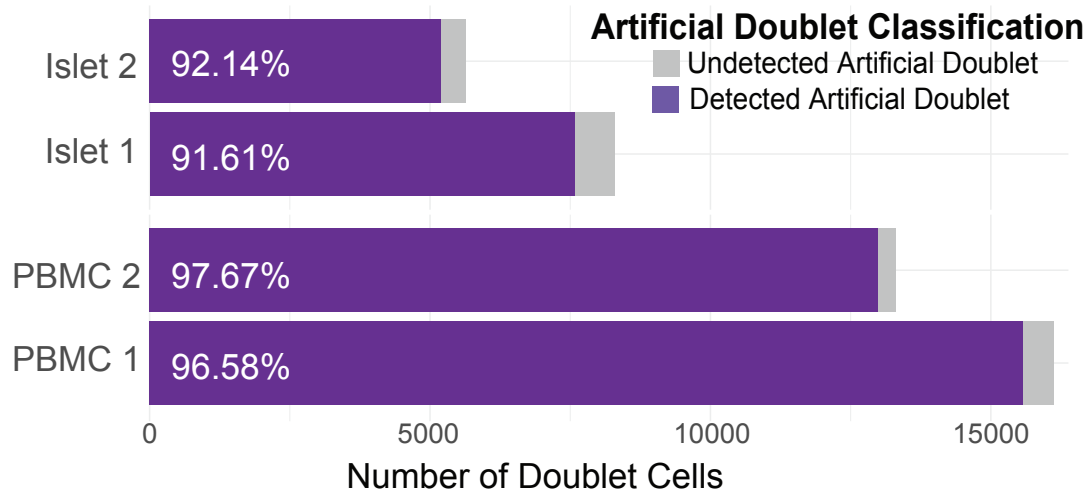


Figure 4.6: The figure shows the percentage of heterotypic artificial cells that were correctly detected when each doublet was analyzed individually. In this analysis for each artificial cell the count based method is ran by adding it to set of real cells.

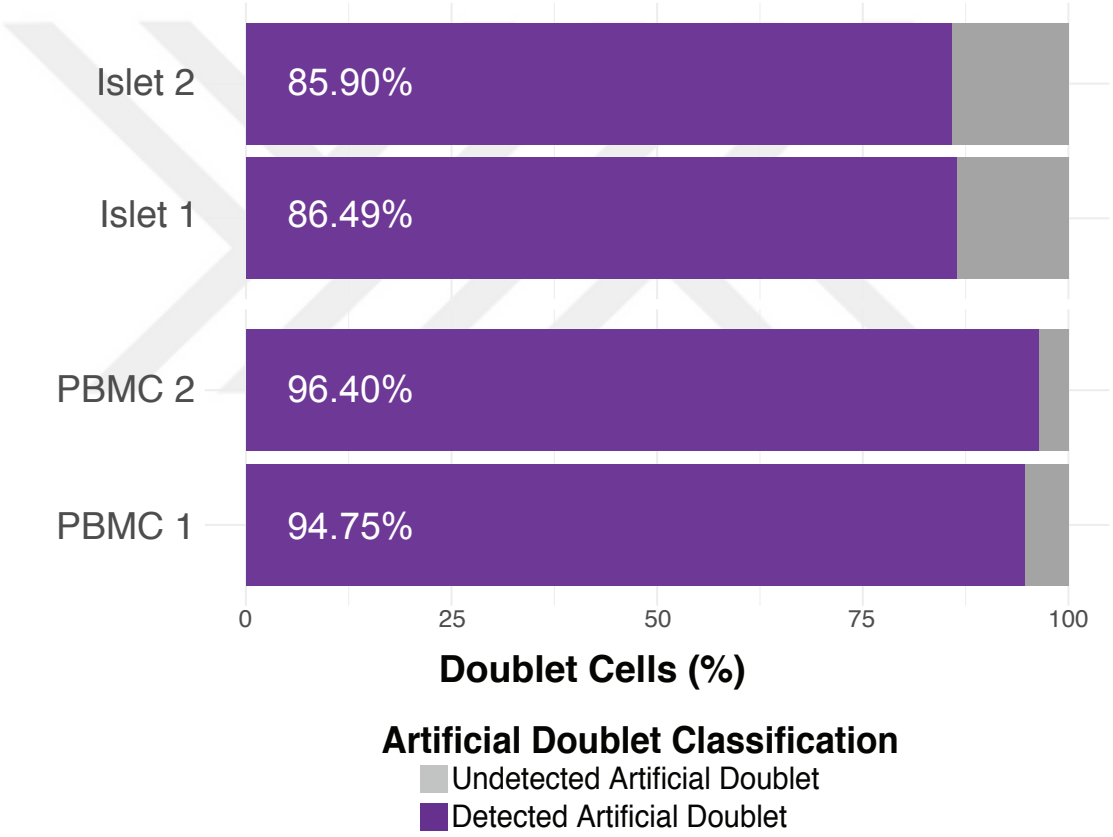


Figure 4.7: The figure shows the percentage of heterotypic artificial cells that were correctly detected when they were added to account for the 2.5% of the existing population.

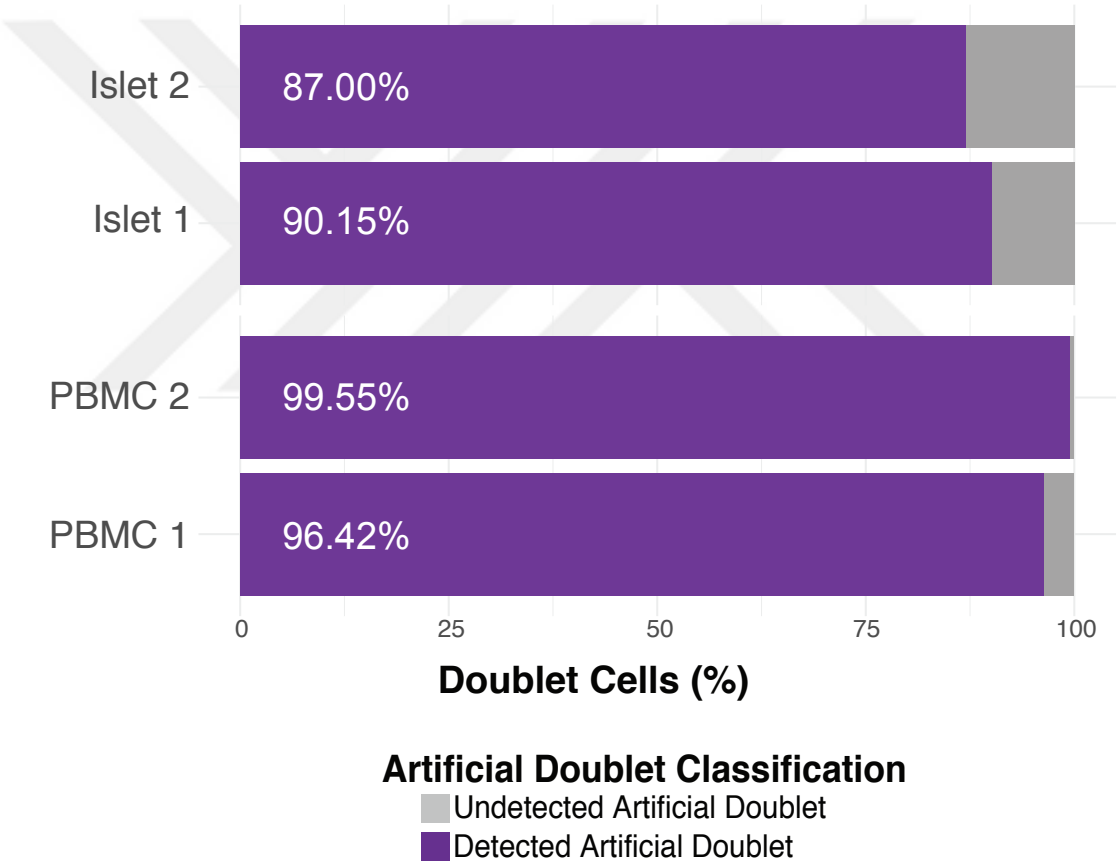


Figure 4.8: The figure shows the percentage of homotypic artificial cells that were correctly detected when each doublet was analyzed individually.

4.3 Doublet cell identification

After the promising results in detecting artificial doublets, we ran DoubletDetector on the unfiltered cells in our datasets. The results when the method was ran on the cells for each sample can be seen in Figure 4.9. DoubletDetector detected 330 of the 6,623 cells in Islet sample 1 as doublets. 285 cells out of 5,722 of Islet sample 2 were identified as doublets by DoubletDetector. PBMC sample 1 had 429 doublets in 6,143 cells, whereas 538 cells were identified as doublets out of 4,974 for PBMC sample 2. The cells that are detected as doublets along with the correctly identified heterotypic artificial cells are shown in Figure 4.4 for the PBMC sample 1 and in Figure 4.5 for the islet sample 1.

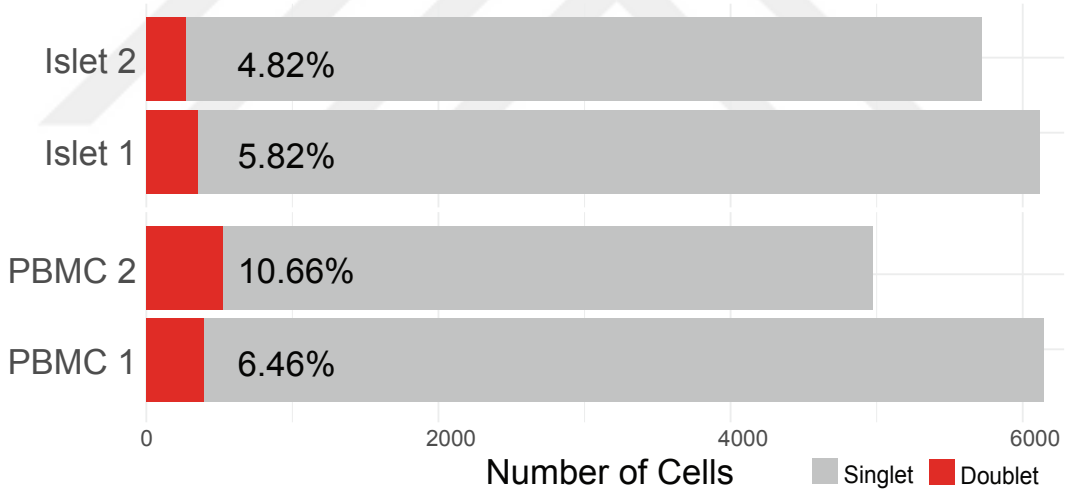


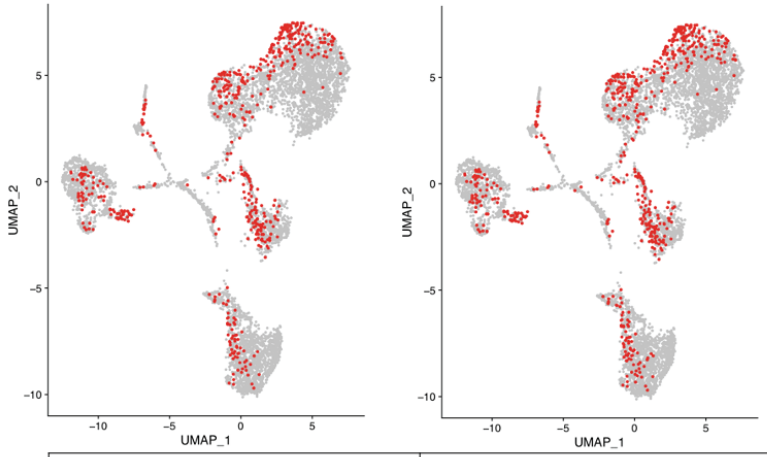
Figure 4.9: The figure shows the percentage of doublets that were detected in each sample.

4.4 Benjamini-Hochberg false discovery rate (FDR) selection

For the doublet detection part we tried two different values for the false discovery rate(FDR) of Benjamini-Hochberg testing; 1% and 0.1%. Figure 4.10 shows how the different FDR values worked for islet sample 1. The percentage of doublets detected were similar for both cases at 5% for 1% FDR and 6% for 0.1% FDR. The detection of artificial doublets were hence an important aspect in deciding the right FDR value. As we got a higher recall with 1% FDR and not very much difference in actual cell detection, we went ahead with 1% FDR value for DoubletDetector.

FDR Cut-Off	0.001	0.01
Islet Tissue Cells Identified As Doublets (6623 cells)	317 (5%)	389 (6%)

Doublets Identified in Original Cells



Artificial Doublet Detection (8291 doublet cells added)		Actual		Actual	
		1	0	1	0
Predicted	1	7595	*NA*	7711	*NA*
	0	691	*NA*	580	*NA*

Figure 4.10: The figure shows the impact of changing the false discovery rate in Benjamini-Hochberg correction on the doublet detection performance for islet sample 1.

4.5 Library depth and doublet detection

During the evaluation of DoubletDetector, one of the important observations we made was the impact of library depth's of the cells chosen to create the artificial cells. While creating the artificial cells for the analysis above, we chose cells with high number of reads on the peaks as source for the combination. Figure 4.11 shows how DoubletDetector's artificial heterotypic cell detection is affected when cells with different library depth are used for evaluation. When the library depths of the cells used as sources for the artificial cells are below a certain limit, the number of artificial cells detected decreases. Although majority of doublets formed are expected to have high library depth, this is still a limitation for the method. This is supported by the Figure 4.12 and Figure 4.13 which show the difference between read depths of doublet and singlet cells as classified by DoubletDetector. In all samples, the doublet cells have higher number of reads when compared with the singlet cells.

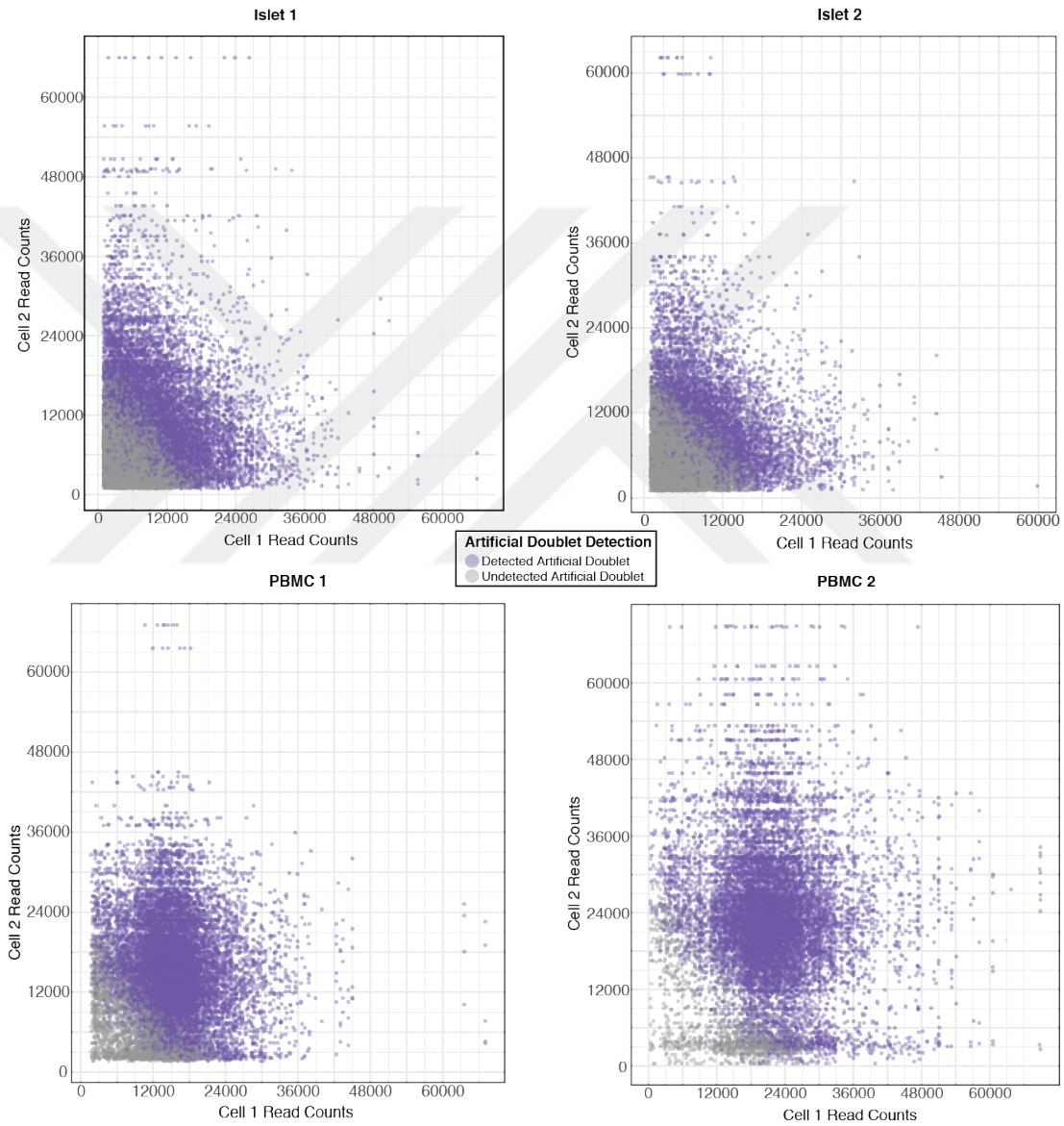


Figure 4.11: The figure shows the impact of the library depth of cells chosen for creating artificial cells. For each sample, under a certain threshold, less number of artificial cells are correctly detected as doublets. The gray points are the cells that are undetected by DoubletDetector.

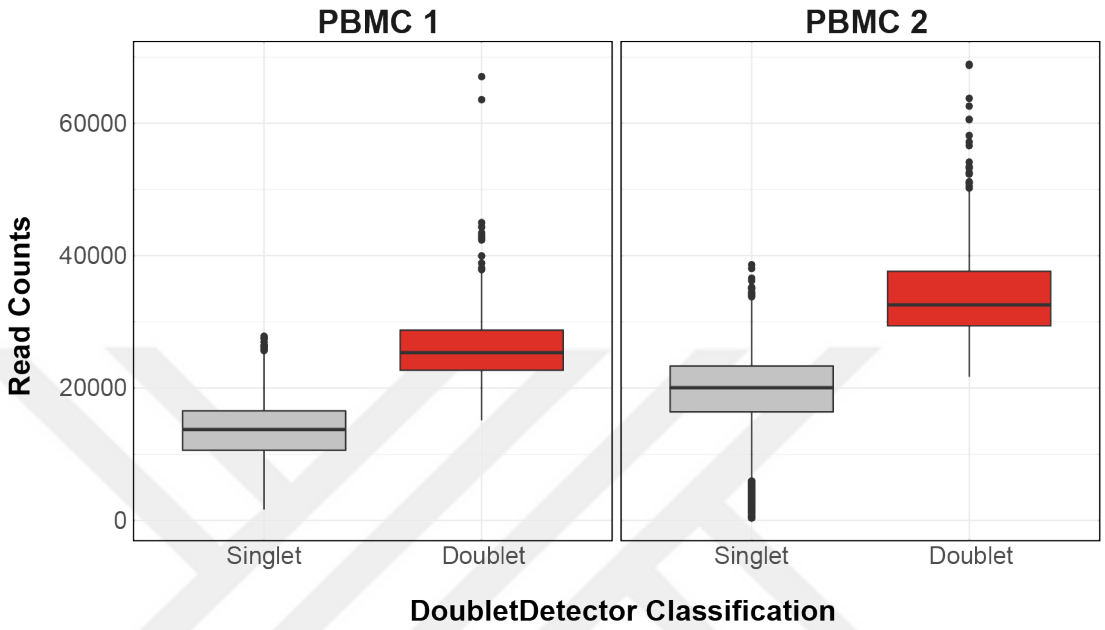


Figure 4.12: The boxplots plots show the difference between read counts of cells classified as doublets by DoubletDetector vs unclassified singlet cells for PBMC sample 1.

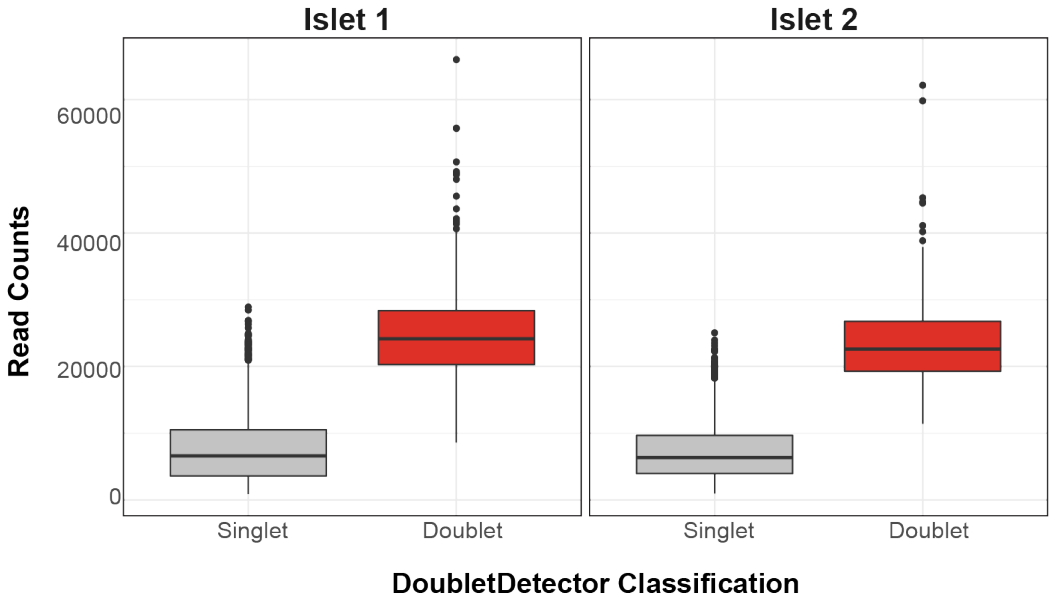


Figure 4.13: The boxplots plots show the difference between read counts of cells classified as doublets by DoubletDetector vs unclassified singlet cells for islet sample 1.

4.6 Doublet origin annotation

After identifying the cells which are doublets, we moved onto finding the cell types of the cells that formed them. The clustering and cell type definitions of the cells were already made and found as described in Section 4.1. According to these cell type clusters, we moved onto finding their unique marker peaks as described in Section 3.4. These marker peaks are the genomic regions were the cells of that certain cell type cluster have significantly higher number of reads when compared with the other cell types in the dataset. Figure 4.14 shows the distributions of the aggregate reads for clusters 5 and 13 on the marker peaks for PBMC sample 1. The aggregate reads are found by taking the median of reads of the cells in that cluster. As cluster 5 is made up of mostly singlet B cells, most of the reads are concentrated on the marker peaks of B cells as expected. However, cluster 13 is similar to an island and mostly composed of doublet cells. Hence, the read distributions are spread onto multiple cell type's marker peaks and not concentrated on a single cell type. Although not as clearly, Figure 4.15 shows a similar situation for the islet sample 1. As an alternative the read counts are separated according to clusters instead of cell types, but the colors indicate the cell types. As fitting, the reads of cluster 0, which is an Alpha cell cluster, are mostly concentrated on the marker peaks of Alpha cell clusters; 0 and 7. Although cluster 5 is a Delta cell type cluster, some of the cells on this cluster is doublets as seen in Figure 4.3. This causes the spread of reads onto the marker peaks of other cell types when compared with the profile of cluster 0. The difference in this representation between PBMC and islet samples comes from the unique property of the cell types in these tissue types. As PBMC cell types such as CD4 and Monocytes are far different when compared with islet cell types such as Alpha and Beta, the profiles tend to be more separate on the unique cell clusters.

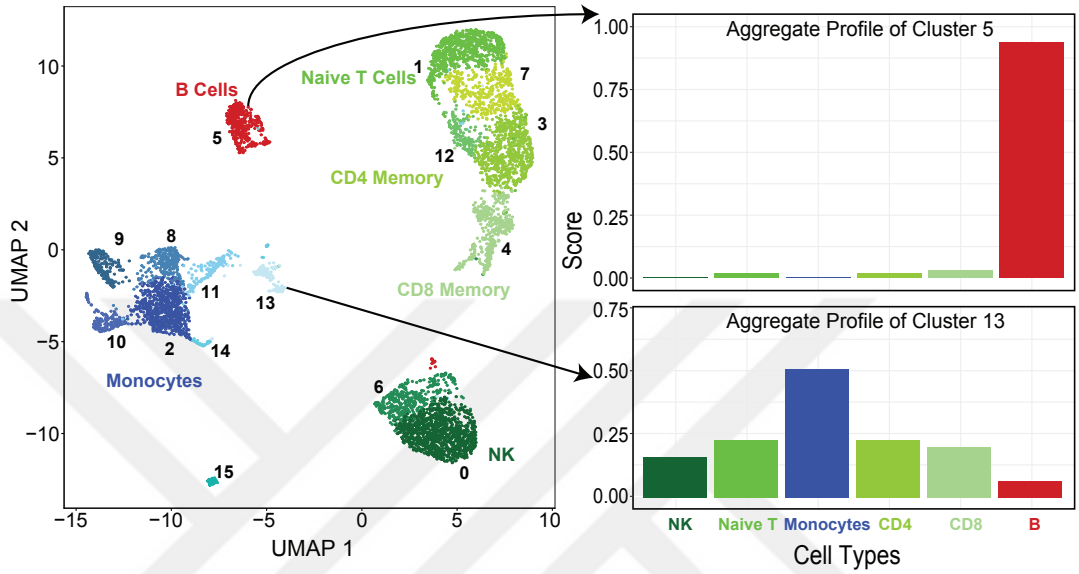


Figure 4.14: The figure shows aggregate read count distributions of clusters 5 (B cell cluster) and 13 (possible doublet cluster) on the cell type specific marker peaks for PBMC sample 1.

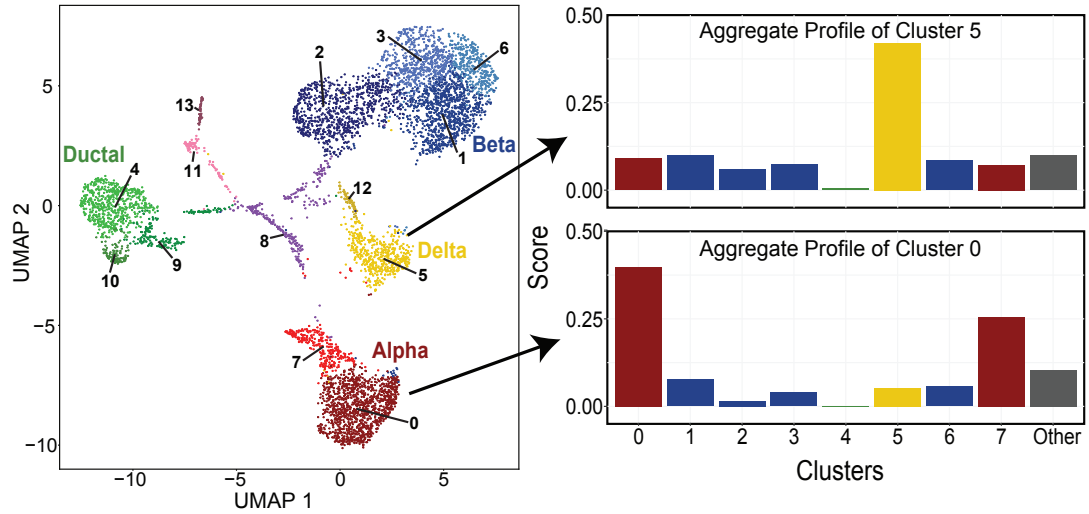


Figure 4.15: The figure shows aggregate read count distributions of clusters 0(Alpha cell cluster) and 5(Delta cell cluster with doublets) on the cluster specific marker peaks for islet sample 1.

4.6.1 Artificial cell annotation

Similar to the evaluation of doublet detection, artificial cells are again used for evaluating the annotation pipeline. Again, the artificial cells are created by combining cells from different cell type cluster, in this case only heterotypic doublets are used. The annotation pipeline was ran on these in order to see if the source clusters that formed the artificial cell can be correctly predicted. For every pair of cell types in the sample, 100 artificial cells are simulated and analyzed.

Table 4.1 shows the results of this evaluation for PBMC sample 1. Pairings between distinct cell types such as Monocyte and B cells are correctly identified with recall values around 90% as the top prediction. However, as the T cell subtypes such as NK, Naive T, CD4 memory and CD8 memory are very similar, their pairings with the other cell types give mixed results in top prediction. But if the top 2 hits of the pipeline is considered, the recall value improves for T cell sub-type pairings with B and Mononcyte cell types. However, the pipeline has an improved recall value for inner T cell pairings such as NK-Naive T and Naive T-CD memory that are slightly more distinct sub-types. The results for PBMC sample 2 are shown in Table 4.2.

The performance of the annotation pipeline on islet sample 1 is shown in Table 4.3. The recall values are improved for the islet sample when compared with PBMC sample 1. Pairing of the two major cell types, Alpha and Beta, are predicted with high recall. As the Delta cell type is fairly more similar with every other cell type, its pairings do not have such a high recall value in the top prediction. However, it is somewhat improved when the top 2 predictions of the pipeline is used for evaluation. The results for the other islet sample are shown in Table 4.4.

Table 4.1: This table shows the performance of the annotation pipeline on the artificial cells from PBMC sample 1. Top predicted annotation is achieved when the most probable pairing from the pipeline is the same as the clusters used to form the artificial cell. Correct annotation in top 2 shows the percentage of cases where the actual artificial cell pairing was within the top 2 spots of the pipeline's predictions.

Cell 1 Type	**Cell 2 Type**	**Top Predicted Annotation (%)**	**Correct Annotation in Top 2(%)**
NK	Naive T	89	99
NK	Monocytes	90	91
NK	CD4 Memory	56	79
NK	CD8 Memory	85	97
NK	B	95	100
Naive T	Monocytes	90	90
Naive T	CD4 Memory	98	100
Naive T	CD8 Memory	33	56
Naive T	B	97	100
Monocytes	CD4 Memory	73	85
Monocytes	CD8 Memory	49	59
Monocytes	B	97	99
CD4 Memory	CD8 Memory	57	75
CD4 Memory	B	76	97
CD8 Memory	B	49	63
Combined		**76%**	**86%**

Table 4.2: This table shows the performance of the annotation pipeline on the artificial cells from PBMC sample 2. Top predicted annotation is achieved when the most probable pairing from the pipeline is the same as the clusters used to form the artificial cell. Correct annotation in top 2 shows the percentage of cases where the actual artificial cell pairing was within the top 2 spots of the pipeline's predictions.

Cell 1 Type	**Cell 2 Type**	**Top Predicted Annotation (%)**	**Correct Annotation in Top 2(%)**
B	CD4 Memory	80	91
B	Monocytes	89	97
B	CD8 Memory	75	90
B	NK	62	93
CD4 Memory	Monocytes	80	97
CD4 Memory	CD8 Memory	83	99
CD4 Memory	NK	32	94
Monocytes	CD8 Memory	91	99
Monocytes	NK	93	99
CD8 Memory	NK	100	100
Combined		**79%**	**64%**

Table 4.3: This table shows the performance of the annotation pipeline on the artificial cells from islet sample 1. Top predicted annotation is achieved when the most probable pairing from the pipeline is the same as the clusters used to form the artificial cell. Correct annotation in top 2 shows the percentage of cases where the actual artificial cell pairing was within the top 2 spots of the pipeline's predictions.

Cell 1 Type	**Cell 2 Type**	**Top Predicted Annotation (%)**	**Correct Annotation in Top 2(%)**
Alpha	Beta	96	100
Alpha	Ductal	100	100
Alpha	Delta	78	88
Beta	Ductal	98	100
Beta	Delta	60	76
Ductal	Delta	70	78
Combined		**84%**	**90%**

Table 4.4: This table shows the performance of the annotation pipeline on the artificial cells from islet sample 2. Top predicted annotation is achieved when the most probable pairing from the pipeline is the same as the clusters used to form the artificial cell. Correct annotation in top 2 shows the percentage of cases where the actual artificial cell pairing was within the top 2 spots of the pipeline's predictions.

Cell 1 Type	**Cell 2 Type**	**Top Predicted Annotation (%)**	**Correct Annotation in Top 2(%)**
Beta	Alpha	87	96
Beta	Delta	76	99
Beta	Ductal	98	100
Alpha	Delta	50	87
Alpha	Ductal	97	100
Delta	Ductal	47	62
Combined		**76%**	**91%**

4.6.2 Real doublet cell annotation

The doublet cells identified and reported in Section 4.3 are ran through the marker peak distribution based annotation pipeline. Initially, the cells are separated into heterotypic and homotypic based on the information gain from their marker peak distribution. For this part, we ran the pipeline on all cells and found a threshold for separation based on entropy scores. Doublet cells that have entropy above the median of combined entropy values were classified as heterotypic and others as homotypic. The median values were 1 and 1.25 for PBMC samples respectively, 1 and 0.8 for the islet samples. Figure 4.16 shows the homotypic and heterotypic doublet cells in the UMAP space for PBMC sample 1. The doublets that are formed by the pairings of Y cell sub-types are seen inside the large mixed T cell cluster. As expected the island cluster that had a mixed marker peak profile is mostly made up of heterotypic doublets. The homotypic doublets are found inside the cluster body and are seen as ordinary cells. The doublets of islet sample 1 are visualized in Figure 4.17. Heterotypic doublets of this sample are grouped around the Delta cell cluster that again had a comparably mixed profile. Homotypic doublets of the Alpha and Beta cells are mostly found at the outer edges of the clusters. Figure 4.18 shows the difference of entropy values between singlet and doublet cells in islet sample 1. For the heterotypic and homotypic threshold, a value near 1 can be established to capture the doublets that have high entropy and are heterotypic as singlet cells will be expected to have similar entropy values with homotypic doublets.

Figure 4.19 shows in detail the pairings that formed the doublets in the PBMC samples. For both samples around 60% of the doublet cells are homotypic. Most prominent heterotypic pairings are the T cell sub-type pairings such as NK-CD8 memory and CD4 memory-CD8 memory. The barplots below show the distribution of the homotypic doublets inside the cell types. For both samples most of the homotypic doublets were inside NK, Monocytes and B cell clusters.

Figure 4.20 shows the annotation pipeline results for the islet samples. Percentage of homotypic doublets are increased when compared with PBMC samples

and are 75% for one sample and around 90% for the other. The majority of heterotypic doublets are pairings with the Delta cell cluster and other cell types. Majority of the homotypic doublets are found in the Beta cell cluster for both of the samples which can have further implications about the doublet forming nature of this cell type.

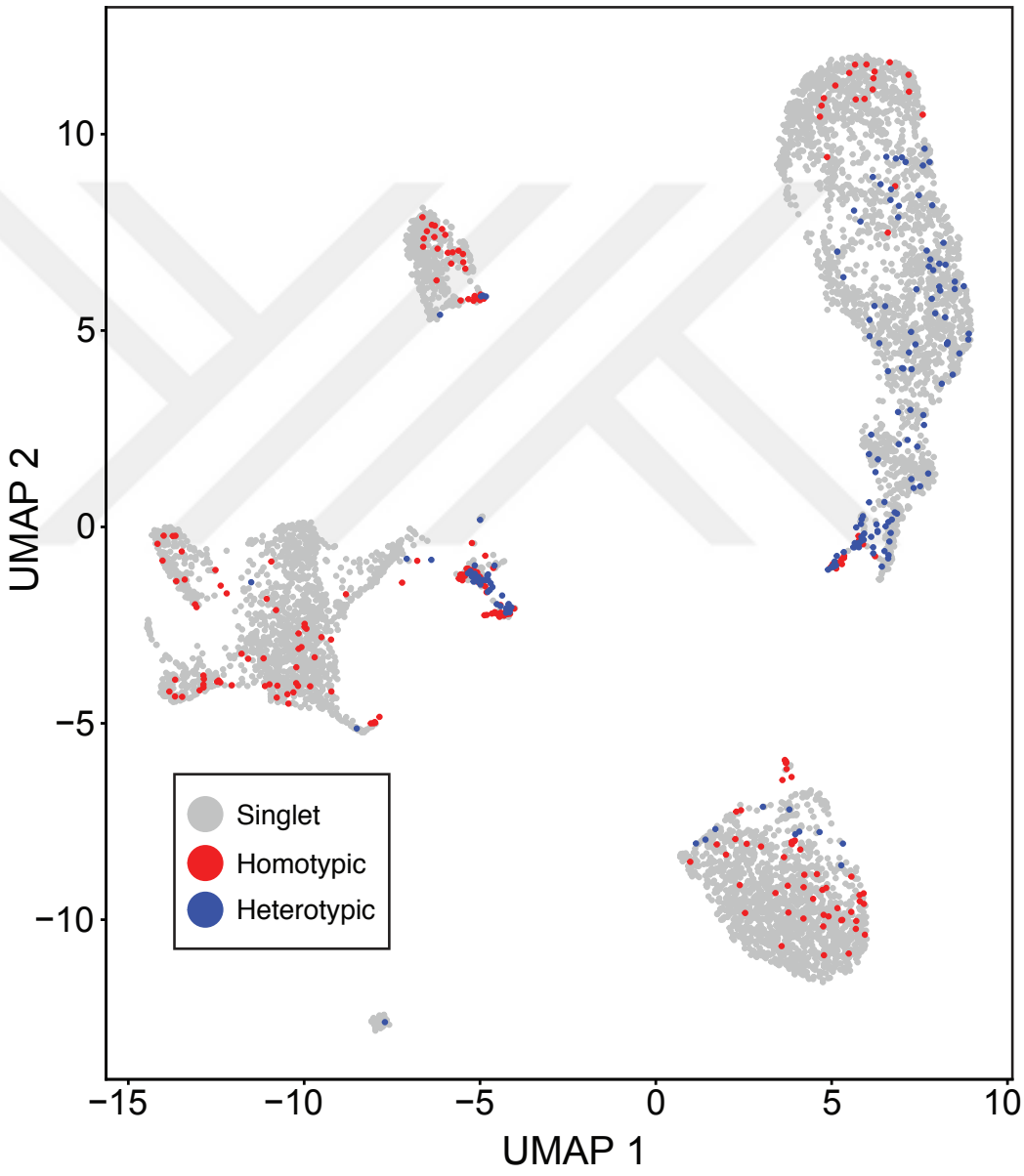


Figure 4.16: The figure shows the homotypic and heterotypic doublets that were identified by the annotation pipeline for PBMC sample 1.

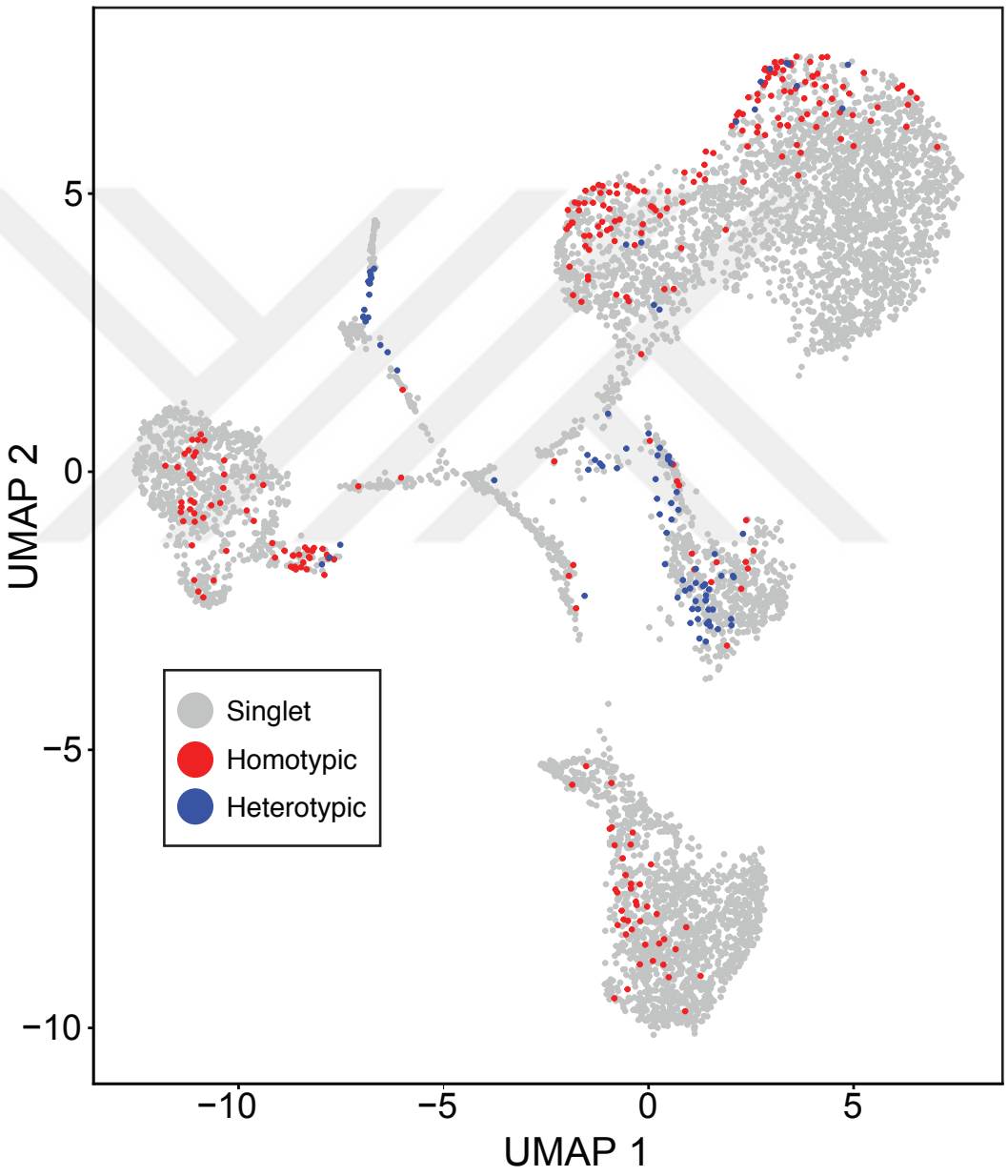


Figure 4.17: The figure shows the homotypic and heterotypic doublets that were identified by the annotation pipeline for islet sample 1.

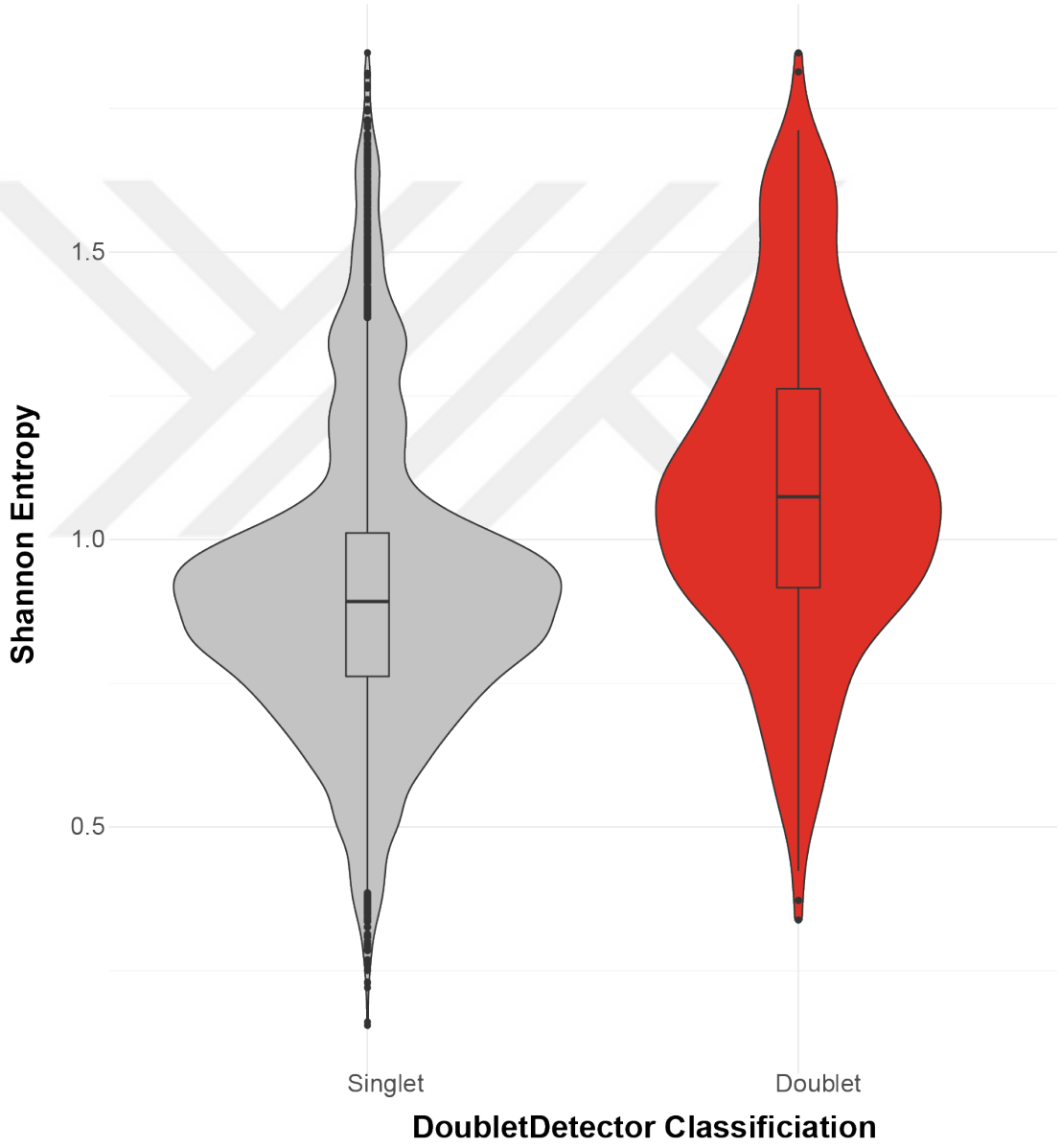


Figure 4.18: This violin plot shows the difference between the entropy values of singlet and doublet cells as classified by DoubletDetector in islet sample 1.

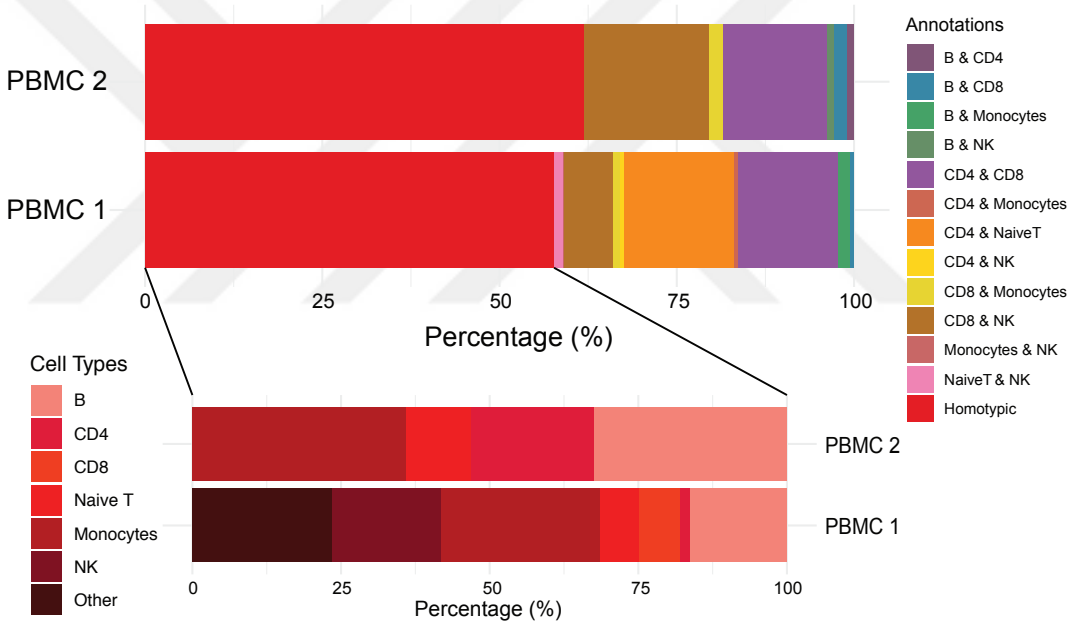


Figure 4.19: This barplot shows the distribution of the doublet types and origins for the PBMC samples. The first set of barplots show the origin pairings that are seen in the samples. The second set show the distribution of the homotypic doublets in the cell type clusters.

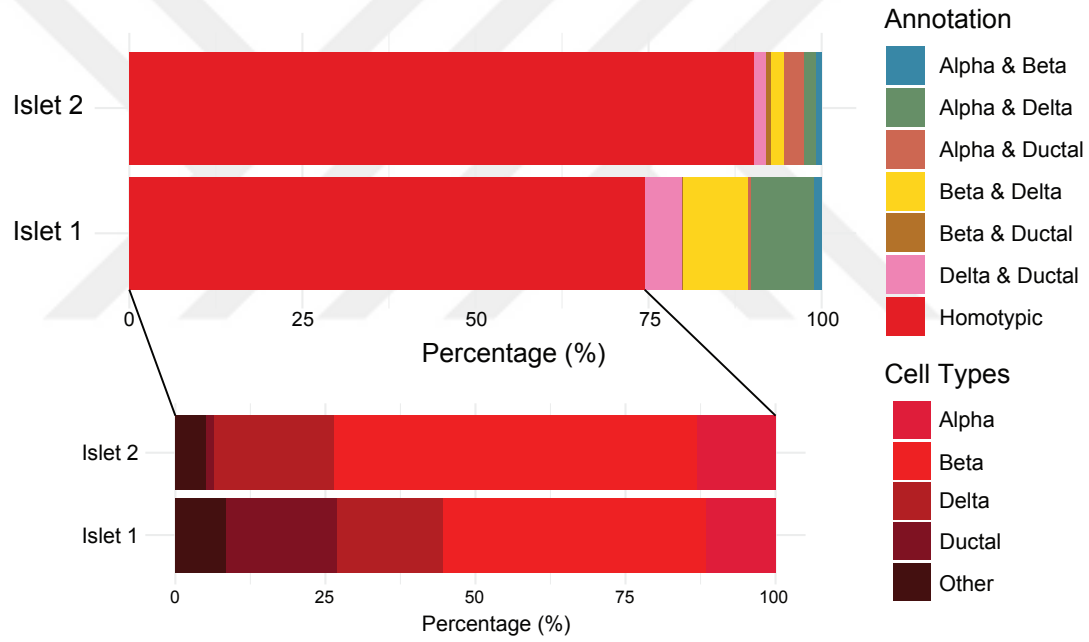


Figure 4.20: This barplot shows the distribution of the doublet types and origins for the islet samples. The first set of barplots show the origin pairings that are seen in the samples. The second set show the distribution of the homotypic doublets in the cell type clusters.

Chapter 5

Conclusion

Doublet cells are one of the prominent issues in single cell analysis and they must be identified beforehand for sound downstream analyses. Doublet detection methods exist for single cell RNA sequencing data that are widely used in research. However, there is no method that utilizes the single cell property that DoubletDetector makes use of doublet detection in single nucleus ATAC-seq domain. We propose a method that detects doublets in snATAC-seq samples and annotate the type of these doublets to further understand the nature of multiplets. To show that DoubletDetector works for different tissue types, we demonstrated with both PBMC and pancreatic islet samples.

One drawback of DoubletDetector is the impact of read depth. As the read depth of a doublet decreases, the detection rate of that cell is decreased. In the future, a hybrid approach combining both DoubletDetector and an approach using nearest neighbors on accessibility profiles maybe able to overcome this limitation. DoubletDetector can be used to first identify the initial set of doublets and nearest neighbors can then be used to identify additional low read depth doublet cells that cluster more closely to the initial doublets identified.

Bibliography

[1] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature Methods*, vol. 10, pp. 1213–1218, Oct. 2013.

[2] A. N. Schep, J. D. Buenrostro, S. K. Denny, K. Schwartz, G. Sherlock, and W. J. Greenleaf, "Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions," *Genome Research*, vol. 25, pp. 1757–1770, Aug. 2015.

[3] Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa, "Identification of transcription factor binding sites using ATAC-seq," *Genome Biology*, vol. 20, Feb. 2019.

[4] C. Liu, M. Wang, X. Wei, L. Wu, J. Xu, X. Dai, J. Xia, M. Cheng, Y. Yuan, P. Zhang, J. Li, T. Feng, A. Chen, W. Zhang, F. Chen, Z. Shang, X. Zhang, B. A. Peters, and L. Liu, "An ATAC-seq atlas of chromatin accessibility in mouse tissues," *Scientific Data*, vol. 6, May 2019.

[5] J. F. Fullard, M. E. Hauberg, J. Bendl, G. Egervari, M.-D. Cirnaru, S. M. Reach, J. Motl, M. E. Ehrlich, Y. L. Hurd, and P. Roussos, "An atlas of chromatin accessibility in the adult human brain," *Genome Research*, vol. 28, pp. 1243–1252, June 2018.

[6] D. A. Cusanovich, A. J. Hill, D. Aghamirzaie, R. M. Daza, H. A. Pliner, J. B. Berletch, G. N. Filippova, X. Huang, L. Christiansen, W. S. DeWitt,

C. Lee, S. G. Regalado, D. F. Read, F. J. Steemers, C. M. Disteche, C. Trapnell, and J. Shendure, "A single-cell atlas of in vivo mammalian chromatin accessibility," *Cell*, vol. 174, pp. 1309–1324.e18, Aug. 2018.

[7] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, W. J. Greenleaf, and H. Y. C. and, "The chromatin accessibility landscape of primary human cancers," *Science*, vol. 362, p. eaav1898, Oct. 2018.

[8] E. J. Márquez, C. han Chung, R. Marches, R. J. Rossi, D. Nehar-Belaid, A. Eroglu, D. J. Mellert, G. A. Kuchel, J. Banchereau, and D. Ucar, "Sexual-dimorphism in human immune system aging," *Nature Communications*, vol. 11, Feb. 2020.

[9] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, pp. 486–490, June 2015.

[10] D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure, "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing," *Science*, vol. 348, pp. 910–914, May 2015.

[11] C. A. Lareau, S. Ma, F. M. Duarte, and J. D. Buenrostro, "Inference and effects of barcode multiplets in droplet-based single-cell assays," *Nature Communications*, vol. 11, Feb. 2020.

[12] E. A. DePasquale, D. J. Schnell, P.-J. V. Camp, Í. Valiente-Alandí, B. C. Blaxall, H. L. Grimes, H. Singh, and N. Salomonis, "DoubletDecon: Deconvoluting doublets from single-cell RNA-sequencing data," *Cell Reports*, vol. 29, pp. 1718–1727.e8, Nov. 2019.

[13] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal*

Statistical Society: Series B (Methodological), vol. 57, pp. 289–300, Jan. 1995.

[14] A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, P. Shah, J. C. Bell, D. Jhutty, C. M. Nemec, J. Wang, L. Wang, Y. Yin, P. G. Giresi, A. L. S. Chang, G. X. Y. Zheng, W. J. Greenleaf, and H. Y. Chang, "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion," *Nature Biotechnology*, vol. 37, pp. 925–936, Aug. 2019.

[15] V. Rai, D. X. Quang, M. R. Erdos, D. A. Cusanovich, R. M. Daza, N. Narisu, L. S. Zou, J. P. Didion, Y. Guan, J. Shendure, S. C. Parker, and F. S. Collins, "Single-cell ATAC-seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures," *Molecular Metabolism*, vol. 32, pp. 109–121, Feb. 2020.

[16] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[17] M. Stoeckius, S. Zheng, B. Houck-Loomis, S. Hao, B. Z. Yeung, W. M. Mauck, P. Smibert, and R. Satija, "Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics," *Genome Biology*, vol. 19, Dec. 2018.

[18] H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell, and C. J. Ye, "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation," *Nature Biotechnology*, vol. 36, pp. 89–94, Dec. 2017.

[19] S. L. Wolock, R. Lopez, and A. M. Klein, "Scrublet: Computational identification of cell doublets in single-cell transcriptomic data," *Cell Systems*, vol. 8, pp. 281–291.e9, Apr. 2019.

[20] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, "DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors," *Cell Systems*, vol. 8, pp. 329–337.e4, Apr. 2019.

[21] J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf, "ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis," Apr. 2020.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018.

[23] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. M. III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, pp. 1888–1902, 2019.

[24] V. Ntranos, L. Yi, P. Melsted, and L. Pachter, "Identification of transcriptional signatures for cell types from single-cell RNA-seq," Feb. 2018.

[25] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[26] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-seq (MACS)," *Genome Biology*, vol. 9, no. 9, p. R137, 2008.