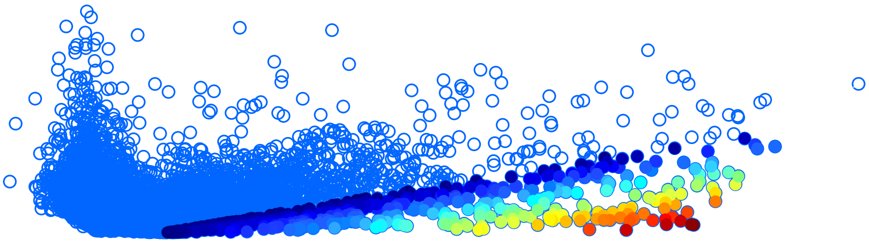
Association Plots visualize cluster-specific genes from high-dimensional transcriptomics data

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Elżbieta Gralińska



Berlin, 2022

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachterin: Dr. Ewa Szczurek

Tag der Disputation: 4. Juli 2022

Selbstständigkeitserklärung

Name: Gralińska
Vorname: Elżbieta

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Berlin, den 07. April 2022

Elżbieta Gralińska

Acknowledgements

First and foremost, I would like to thank my supervisor, Martin Vingron, for his guidance through each stage of my doctoral studies. I am grateful for providing such an inspiring and unforgettable research atmosphere in our department and for his invaluable advice. I would also like to express my gratitude to Stefan Haas for being a member of my TAC committee, for all biological insights, as well as for his invaluable support.

Special thanks go to all members of our lab: Gözde Kibar, Aybuge Altay, Lam-Ha Ly, Tris Rapakoulia, Maryam Ghareghani, Prabhav Kalaghatgi, Hossein Moeinzadeh, Robert Schöpflin, Alena van Bömmel, Yan Zhao, Daniel Rosebrock, Clemens Kohl, Florian Klimm, Emel Comak, Nico Alavi, Persia Akbari Omgba, Ekin Deniz Aksu, as well as all current and previous members of the Vingron Department, for being absolutely wonderful colleagues. I wish to extend my special thanks to Philipp Benner for being a wonderful office mate. I would also like to thank Kirsten Kelleher for her assistance along all the steps of the PhD, and Martina Lorse for creating such a friendly atmosphere in our department.

Additionally, I would like to thank my colleagues who were very much involved in the development of the software implementing the concept of the Association Plots. First of all, Clemens Kohl for his contribution to the R package, as well as Bita Sokhandan Fadakar for her contribution to the Shiny App.

I would also like to express my gratitude to Ewa Szczurek for her time in reviewing this thesis and her insightful comments on the project during our TAC meetings.

Finally, I want to thank Laura Cifuentes Fontanals and Melissa Bothe for critical reading of parts of the thesis, Lam-Ha Ly for helping me with the abstract and for the valuable comments on the figures, as well as Clay Alter for critical reading of the entire thesis.

Lastly, I want to thank my brother, Artur, and my parents for their continuous support along the way. Without them this would have not been possible.

Contents

CHAPTER 1

Introduction

A re-occurring question in transcriptomics data analysis is which genes are characteristically highly expressed in a given cluster of conditions, i.e. are associated to this cluster. Approaches to this question occur in many forms, be it as biclustering (Tanay *and others*, 2002; Pontes *and others*, 2015) or in the search for so-called marker genes. While there are tools available today, such as differential expression testing methods, exploration of genes which characterize a cluster still require tedious sifting through long lists of program output.

The term 'cluster-specific genes', i.e 'marker genes', refers to genes with expression profiles characteristic for a cluster of conditions. In the case of single-cell data, a cluster of conditions can refer to a cell cluster representing a cell type or a cell identity, whereas, in the case of bulk data it can refer to a group of samples consisting of multiple cells. Although the identification of marker genes is commonly of interest in genomics studies, due to biological variability of cell types as well as current technological limitations there is no catalogue of marker genes covering existing cell types or various biological conditions. Often times, a condition is defined based on the expression of 'historical' marker genes, i.e. genes that have been found over many years of research under given conditions. Yet, the expression of certain genes might vary depending on the conditions they are under. This might be the case for genes less studied in the past, and thus, less present in the literature than other, more frequently studied genes. As a consequence, relying on marker genes derived from the literature might lead to underestimating the heterogeneity within a cluster, followed by a wrong biological interpretation of a data set.

While the identification of marker genes is fairly easy for small data sets, for data sets with a higher number of samples or cells, such as single-cell RNA-seq, the situation is more complex and poses a significant challenge to analysis and visualization methods currently available. Although methods such as principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933) have been successfully employed for many years, they have serious limitations when applied to large and complex data sets. For such data the first two or three principal components may explain only a small fraction of the total variance. This renders the low-dimensional representation, obtained after projection, effectively pointless due to the massive loss of information. Due to this we are faced with the challenge of visualizing information from complex data sets from their higher dimensions.

This thesis seeks to address this problem by presenting Association Plots, a novel method for determining and visualizing cluster-specific gene sets from high-dimensional transcriptomics data. Association Plots are derived from correspondence analysis (CA) (Greenacre, 1984), itself a data projection technique closely related to PCA. However, when applied to transcriptomics data, CA enables the joint embedding of genes and conditions in one space. Such a simultaneous representation of genes and conditions in one space conveys information as to the association between them. We take advantage of this CA property for our definition of Association Plots. By measuring the distances between a given cluster of conditions and the genes in the multidimensional CA space we are able to plot the genes in a two-dimensional coordinate system. We call this representation the 'Association Plot'.

The horizontal axis of the Association Plot (Fig. 1.1) indicates how strongly

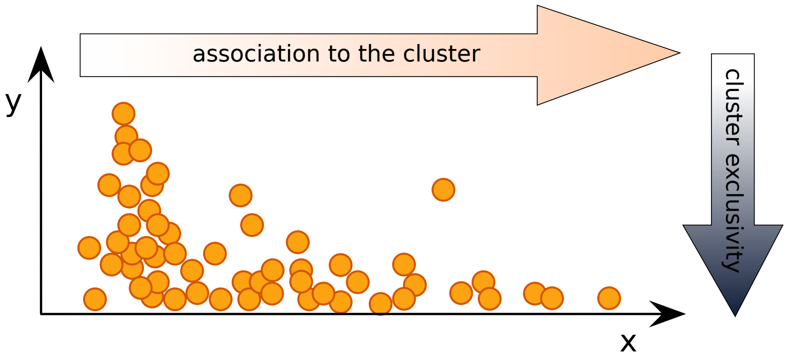


Figure 1.1: Association Plot scheme. The horizontal axis indicates the strength of the gene-cluster association, while the vertical axis indicates whether other clusters also show expression of a given gene. Orange circles represent genes.

the gene is associated with a cluster, with genes furthest to the right having the strongest association. The vertical axis indicates whether other clusters also show expression of this gene, such that the most characteristic genes for a cluster can be found near the x-axis far to the right. Due to this, Association Plots can be seen as a visualization method for prediction of cluster-specific genes even from high-dimensional data, in a manner independent of data size.

Applying Association Plots to transcriptomics data offers a wide range of applications. First of all, it facilitates exploration of the data and thereby enables the user to gain a deeper understanding of the data. Second, generating Association Plots for clusters from the data allows the identification of marker genes characterizing different biological samples, cell types, or cell identities. Moreover, Association Plots can also be applied for cluster annotation purposes. This is effected by comparing literature-derived marker genes, e.g. for different cell types, with the marker genes derived from the Association Plots generated for cell clusters from investigated data.

To implement the concept of Association Plots we developed *APL*, an R package for the identification and visualization of cluster-specific genes from both bulk- and single-cell transcriptomics data. After providing input data, either a single-cell RNA-seq data set with precomputed clusters of cells or bulk RNA-seq data with biological sample information, *APL* generates Association Plots. The computed plots can then be interactively queried by the user, which gives further insights into the data. Furthermore, a ranked list of genes characteristic for a selected cluster can be extracted. To facilitate the analysis of single-cell data, *APL* has been developed in way which allows its integration into single-cell analysis pipelines.

This thesis is organized as follows. First, Chapter 2 provides an overall overview of bulk- and single-cell RNA-seq methods, and gives an introduction to the topic of marker gene identification. Chapter 3 focuses on a mathematical overview of PCA, singular value decomposition (SVD), and CA, and explains the link between them. Chapter 4 presents the newly developed Association Plots. Within this chapter, first the formalism behind Association Plots is introduced and then two demonstration example data sets are provided, one simulated and one from economics, which serve to illustrate the new visualization method. Lastly, SVD-based denoising is introduced and an issue of scoring genes according to their cluster-specificity is addressed. Chapter 5 focuses on the application of Association Plots to an example bulk RNA-seq data set. Here, it is demonstrated how to identify tissue marker genes and investigate cluster similarities using Association Plots. Chapter 6 presents the usage of Association Plots for the visualization of cluster-specific genes in high-dimensional single-cell RNA-seq data, identification of novel marker genes, and cluster annotation purposes. This is demonstrated by applying Association Plots to two example single-cell RNA-seq data sets. This chapter explores also the relationships between Association Plot-derived cluster-specific genes and those obtained by other computational approaches. Finally, in Chapter 7 presents our newly developed R package *APL* and provides examples of Gene Ontology enrichment analysis in the framework of Association Plots.

CHAPTER 2

Biological Background

2.1 Introduction to RNA-seq

Transcriptomics data, which is the focus of this work, contains information on the complete set of RNA molecules present in a cell or in a population of cells. Although various methods exist for measuring the amount of RNA transcripts, the most common one nowadays is RNA-sequencing (RNA-seq). RNA-seq was invented approximately 15 years ago, shortly after the invention of next-generation sequencing (NGS). Owing to its advantages over the so-called hybridization-based approaches it became a powerful tool that largely contributed to the field of molecular biology (Wang *and others*, 2009; Weber, 2015).

Hybridization-based approaches such as DNA microarrays were predecessors to RNA-seq method (Schena *and others*, 1995). DNA microarrays, also called DNA chips, were developed in 1988 by four independent groups of scientists, and consist of a plastic, glass or nylon slide covered with multiple spots (Pevzner, 2000). Each spot contains a small amount of a particular DNA sequence, which is used as a probe for quantification of gene expression from two biological samples, e.g. reference and experimental samples. Thus, to conduct a two-color microarray experiment, mRNA molecules from both samples need to be first converted into cDNA and labeled with one of two fluorescent dyes (e.g. red and green), depending on the sample they originate from. Subsequently, equal amounts of cDNA molecules are mixed together and hybridized to the slide. The expression of a gene is then quantified by measuring the intensity of a fluorescent signal in a given spot, which results from cDNA molecules binding to a probe from a given spot. For example, if a gene is equally expressed in both samples, the microarray spot will appear yellow, while in the case of the overexpression of a gene in one of the two samples the spot will appear either green or red.

Although DNA microarrays were a relatively affordable tool for a high-throughput gene expression quantification, they had several limitations. First of all, microarrays require a prior knowledge on DNA sequences, whereas RNA-seq allows for detection of unknown DNA sequences or gene variants (Simoneau *and others*, 2021). Second, microarrays are oftentimes characterized by high background signal due to cross-hybridization events (Casneuf *and others*, 2007). This, together with signal saturation issues, leads to difficulties in detection of genes with either very high or very low expression, which is not the case for RNA-seq. Finally, factors such as manufacturing errors, sample preparation, or array processing lead to a high technical variation across arrays, which in turn makes the comparison of the detected expression values across multiple samples particularly challenging (Parmigiani *and others*, 2003; Wang *and others*, 2009).

The advantages of RNA-seq over a hybridization-based approach led to the replacement of microarrays by RNA-seq. Today, RNA-seq is a commonly-used method for differential expression analyses, as well as several other applications. RNA-seq experiments contributed to new findings in research areas such as alternative splicing, non-coding RNAs, discovery of new variants, and enhancer studies. Moreover, in addition to the scale up in throughput, the wide use of RNA-seq experiments resulted in a variety of methods originating from a standard RNA-seq protocol (Stark *and others*, 2019), such as single-cell RNA-seq. Since this thesis describes the application of Association Plots to bulk- and single-cell (sc) RNA-seq data, we will now focus on these two types of data.

2.2 Bulk RNA-seq data

The term 'bulk RNA-seq' refers to sequencing of RNA transcripts from a population of cells. A basic bulk RNA-seq workflow (Fig. 2.1) begins with the extraction of RNA molecules from a biological sample. Additionally, depending on the topic of interest, an RNA selection or depletion step can be conducted. Afterwards, RNA molecules are reverse-transcribed to cDNA molecules, fragmented, and ligated with adapter sequences. The resulting library is then amplified and sequenced using a high-throughput platform (Stark *and others*, 2019).

Once the sequencing is done, the next step is the mapping of sequencing reads. In most cases the reads are mapped to a reference genome using existing mapping tools such as STAR (Dobin *and others*, 2013) or TopHat (Trapnell *and others*, 2009). Alternatively, when no high-quality annotation is available, the transcriptome is reconstructed by *de novo* assembly of the reads, and the reads are subsequently re-mapped to the reconstructed transcriptome.

Once the reads are mapped, the gene expression can be quantified. At this stage, raw reads for each gene are counted e.g. using HTSeq (Anders *and others*, 2015). Since raw read counts are affected by both the sequencing depth and gene length, they need to be normalized before comparing them across samples. Therefore, gene expression measures such as RPKMs (reads per kilobase of transcript per million reads mapped), FPKM (fragments per kilobase of transcript per million reads mapped) or TPMs (transcripts per million) need to be calculated.

Often, the aim of genomics analyses is the identification of genes differentially expressed between conditions in an experiment. This step is called differential expression analysis, and allows for revealing genes up- or down-regulated across varying conditions. Therefore, multiple bioinformatics tools can be used, e.g. DESeq2 (Love *and others*, 2014), edgeR (Robinson *and others*, 2010), limma (Ritchie *and others*, 2015).

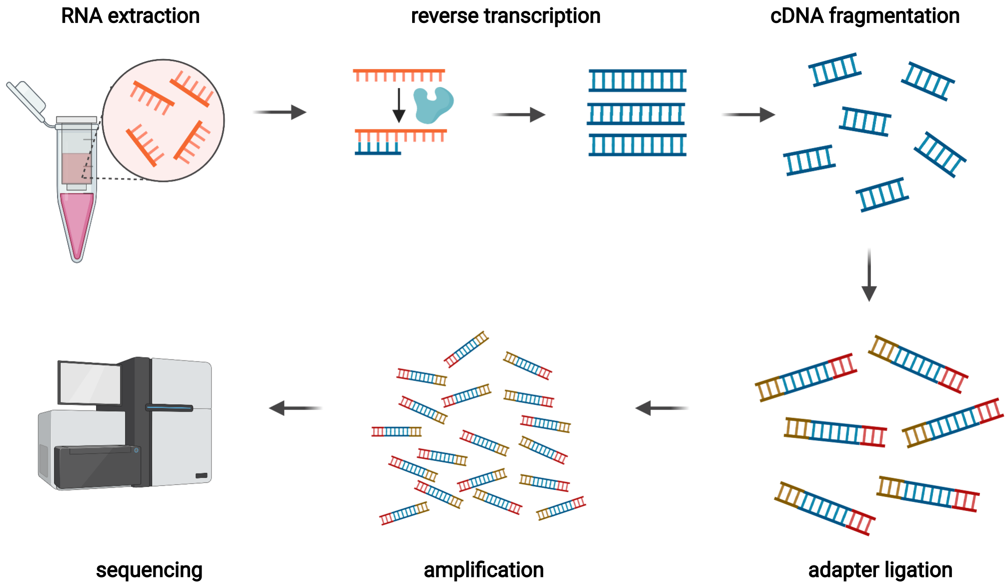


Figure 2.1: Overview of RNA-seq workflow. Created with BioRender.com.

2.3 Single-cell RNA-seq data

Although bulk RNA-seq is a convenient approach for measuring gene expression levels in a biological sample, it does not reveal information on heterogeneity of its cell population. Oftentimes, a cell population consists of cells belonging to various cell types and captured in different cell states. Due to this, the bulk RNA-seq method also fails to reveal spatial information from the cells. These limitations of bulk RNA-seq led to the development of single-cell RNA-seq, a method published for the first time by Tang *and others* (2009) which allows for quantifying the amount of transcripts in a single cell (Stark *and others*, 2019).

A typical sc RNA-seq protocol begins with the isolation of single cells followed by their lysis. Afterwards, mRNA from the lysed cells is reverse transcribed into cDNA and then the cDNA gets amplified. The amplification process can be done either by PCR or *in vitro* transcription. Subsequently, the library is generated and sequenced. An important step in the protocol is barcoding of samples. The barcoding can be conducted either during the reverse transcription, or during library preparation. The main goal of barcoding is the labeling of transcripts from a given cell, so that it is possible to compare the gene expression levels across individual cells (Kolodziejczyk *and others*, 2015; Haque *and others*, 2017).

Although the workflow of scRNA-seq experiments is largely similar to the bulk RNA-seq workflow, there are two main experimental challenges which are not present in bulk experiments and which have an influence on the data processing and analysis steps. First of all, due to the small size, capturing of single cells is challenging (Haque *and others*, 2017; Kolodziejczyk *and others*, 2015). Even though there exist multiple methods such as laser capture microdissection, micropipetting, microdroplets, fluorescent-activate cell sorting (FACS), and microfluidics, each of these method has its limitations, e.g. perturbation of the transcriptional profiles, capturing neighboring cells. The second challenging aspect of the single-cell experiments is the amplification step, mainly due to minute amounts of mRNA present in a single cell.

The challenges mentioned above, that occur while conducting scRNA-seq experiments, can lead to the sparsity of generated data, i.e. the high amount of zero values in the data (Lähnemann *and others*, 2020; Haque *and others*, 2017). This phenomenon is often described as a 'dropout' event. However, zero values in the data are not only related to technical problems, they can also result from a lack of gene expression, which makes handling dropout events even more challenging. In addition to dropout events, scRNA-seq data is more variable than bulk RNA-seq data, due to biological variation between single cells caused by events such as transcriptional bursting (Suter *and others*, 2011) or cell size variation, which influence the amount of transcripts in a cell and which are not directly observed when sequencing a pooled population of cells (Haque *and others*, 2017).

2.4 Marker genes

As was mentioned in the introduction to this thesis, Association Plots can be used for extraction of cluster-specific genes or marker genes. Although we often use these two terms interchangeably, these terms, together with terms such as 'biomarkers', 'genetic markers', or 'differentially expressed genes', can lead to confusion and their meaning needs to be defined.

We start by examining the term biomarker. Although the definition of this term is under discussion and there exist multiple definitions of it, it is often used in clinical research in reference to a molecular characteristic for a medical state, which can be measured in an accurate and reproducible way, e.g. blood pressure or body temperature (Strimbu and Tavel, 2010). Since the development of omics methods (genomics, transcriptomics, proteomics, metabolomics) the definition of biomarkers has become very broad. Multiple molecular factors such as mutations in mitochondrial DNA, copy number variations, single-nucleotide polymorphisms (SNPs), circulating metabolites or RNA levels, are now studied to predict the risk of various medical conditions, e.g. stroke (Montaner *and others*, 2020), Parkinson disease (Lempriere, 2021; Deng *and others*, 2022), various cancers (Engebraaten *and others*, 2021; Wang *and others*, 2019*b*), and many more. The term genetic markers refer only to a subset of such biomarkers.

Genetic markers, also known as genomic biomarkers, can be defined as DNA or RNA characteristics which are indicators of various medical states (Novelli *and others*, 2008). These are mainly mutations in DNA and RNA levels.

This thesis focuses on the identification of marker genes from transcriptomics data. As marker genes we describe genes upregulated in a given set of biological conditions. Since the term 'marker genes' sounds similar to 'genetic markers', the term 'cluster-specific genes' is more precise as this indicates the genes whose expression is specific to a given cluster of conditions, be it a cell type, a tissue, or any selected group of biological samples from the experiment.

In Chapter 5 we will compare the set of cluster-specific genes obtained using Association Plots to the list of genes identified using different differential expression testing tools. Such tools aim to identify genes over- or under-represented in a group of conditions. Thus, differentially expressed genes which are over-represented in a group of conditions are similar to cluster-specific genes which are the focus of this thesis.

CHAPTER 3

Mathematical Background

This chapter provides an overview of three closely related data dimension reduction methods: principal component analysis (PCA), singular value decomposition (SVD), and correspondence analysis (CA). For the purpose of this chapter we define data as a set of vectors $m_i \in \mathbb{R}^C,\ i = 1, 2, ..., G$. We assume $G \geq C$. The vectors can be arranged in form of a matrix M, where the i^{th} row is represented by the i^{th} vector from the set. The resulting matrix consists of G rows (observations) and C columns (variables).

In terms of transcriptomics data, an observation from M corresponds to a gene's expression values, while a variable corresponds to the biological sample or a cell in which the measurement was performed. We use matrix M throughout this chapter as a starting point for the explanation of each of the three methods.

The content of this chapter is mainly based on the books of Jolliffe (2002) and Greenacre (1984, 2017). The content of Section 3.5 is an adapted part of a manuscript posted on bioRxiv (Gralinska and Vingron, 2020), and which is under review as of March 2022, as well as of a manuscript Gralinska *and others* (2022).

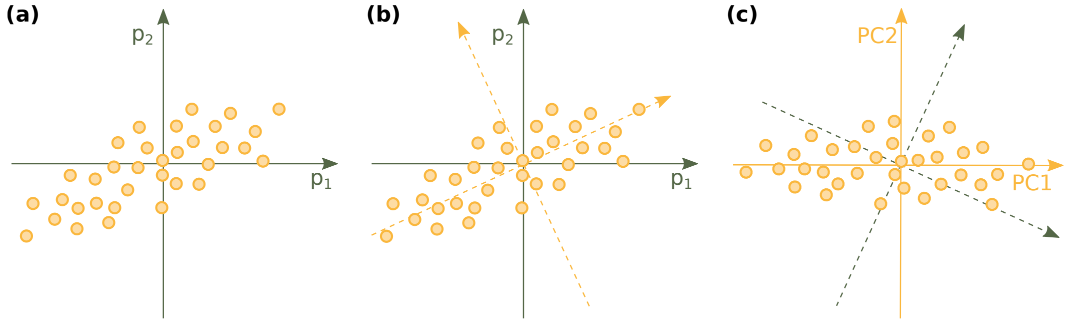


Figure 3.1: Transforming the original coordinate system into a new one using PCA. (a) Projection of observations (yellow circles) from a data matrix onto the first two variables, p_1 and p_2. (b) The variance of the data is the highest along the direction represented by one of two yellow, dotted lines. (c) By rotating the original coordinates a new coordinate system is created.

3.1 Principal component analysis (PCA)

Principal component analysis (Pearson, 1901; Hotelling, 1933) is a method for reducing the dimensionality of a data while preserving most of its variation. Its main concept assumes that a set of variables from the data matrix M is transformed into a smaller set of summary variables called principal components, which facilitate the visualization and statistical modelling of complex data.

To illustrate this, M needs to be first mean-centered, i.e. each variable from M needs to be zero-centered by subtracting the average of its observations from these observations. Each observation from M can be then represented as a linear combination of all variables, and represented in the space with a coordinate system defined by these variables (Fig. 3.1a). In such coordinate system, the first coordinate of an observation is equal to the length of the projection of this observation onto the first axis, the second coordinate of an observation is equal to the length of the projection of this observation onto the second axis, and so forth.

However, in the context of data dimensionality reduction, a more convenient way of representing M would be by using a new coordinate system, in which the axes refer to the amount of variation from the data. To illustrate this, in the example presented in Fig. 3.1b the direction along which the variance of the data is the highest is represented by one of two yellow dotted lines, and not by the first or second variable. The direction defined by this yellow line is called the first principal component direction. Thus, by rotating the original coordinates we obtain a new coordinate system (Fig. 3.1c), in which the first axis is defined by a direction along which the variation of the data is the highest, the second axis is orthogonal to the first one and is defined by a direction along which the variation of the data is the second highest, and so forth. By projecting the data matrix onto these new directions we obtain the so-called principal components (PCs).

Now, an important question is how to find such principal component directions. It turns out that the new coordinate system can be calculated as eigenvectors of a covariance matrix. Therefore, we first calculate a covariance matrix $\sum$:

$$\sum = \frac{1}{n} M^T M.$$

In this matrix, the entry in the j^{th} row and i^{th} column represents the covariance between j^{th} and i^{th} variables if $j \neq i$, and the variance of variable j^{th} if $j = i$. The covariance matrix is symmetric and positive semi-definite, hence it is diagonalizable:

$$\sum = A \Lambda A^T.$$

As shown in Fig. 3.2, Λ is a diagonal matrix and its diagonal element λ_l, where $l = 1, 2, ..., C$, represents the l^{th} eigenvalue of $\sum$. A is an orthogonal matrix and its l^{th} column represents the l^{th} eigenvector of $\sum$. The obtained eigenvectors form a new coordinate system, in which the first eigenvector is the first PC, the second eigenvector is the second PC, and so on. More details on the derivation of the PCs and its mathematical proof can be found in Jolliffe (2002) in Sections 1.1 and 2.1.

Knowing how to transform the original coordinate system into a new one, we can calculate new coordinates of observations from M. Let x be an observation from M (Fig. 3.3a). The problem of calculating its first coordinate in the new coordinate system, y, can be reduced to the calculation of the scalar product of

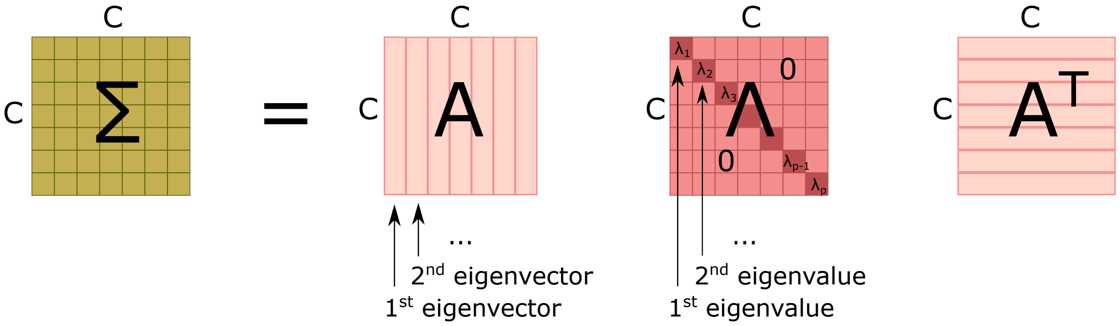


Figure 3.2: Diagonalization of the covariance matrix $\sum$. A is an orthogonal matrix and its columns represent eigenvectors of $\sum$. Λ is a diagonal matrix with diagonal elements representing eigenvalues of $\sum$.

two vectors. As shown in Fig. 3.3b, the orthogonal projection of a vector $\vec{b}$ onto a vector $\vec{a}$ is equals:

$$|\vec{b}| \cos \alpha,$$

where α is the angle between these vectors.

Given the fact that the inner product of two vectors, $\vec{a}$ and $\vec{b}$, is defined as:

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos \alpha,$$

we can represent the scalar projection of $\vec{b}$ onto $\vec{a}$ as

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}|}.$$

In the case of PCA we can use this formula to compute the first coordinate of x in the new coordinate system. $\vec{a}$ corresponds then to the first eigenvector, and $\vec{b}$ to the point x. Additionally, we can think of eigenvectors as normalized to the length of 1 ($|\vec{a}| = 1$). As a result, the first coordinate of x is equals the inner product of the first eigenvector and x. The further coordinates of x are calculated then as inner products of x and the corresponding eigenvectors.

To compute simultaneously all coordinates of x a vector with old coordinates

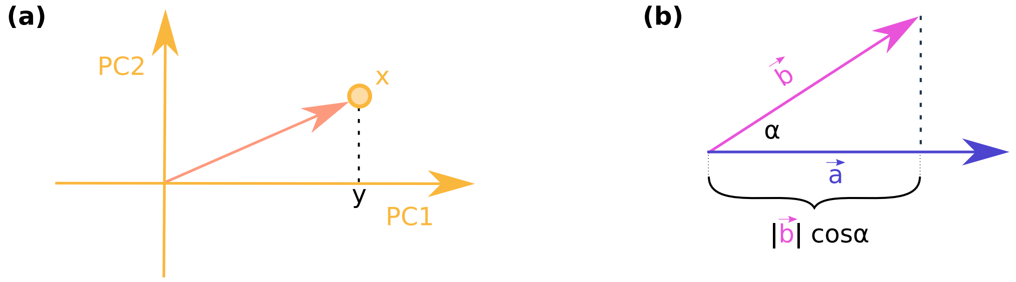


Figure 3.3: Coordinates of an observation in the new coordinate system. (a) y, the first coordinate of the observation x is obtained by projecting x onto the first PC. Therefore, the scalar product of the first eigenvector and x can be used. (b) Orthogonal projection of a vector $\vec{b}$ onto a vector $\vec{a}$.

of x needs to be multiplied by the matrix of eigenvectors (Fig. 3.4a). To compute coordinates simultaneously for all observations from M, the matrix with all old coordinates for all observations needs to be multiplied by the matrix of eigenvectors (Fig. 3.4b).

As mentioned at the beginning of this section, PCA is a dimensionality reduction method. To reduce the dimensionality of the data, the subset of coordinates for all observations from M can be computed by multiplying the matrix of all old coordinates by a matrix containing only a subset of selected eigenvectors to the leading eigenvalues (Fig. 3.5).

3.2 Singular value decomposition (SVD)

Singular value decomposition is a matrix decomposition method which allows to represent M as a product of three matrices (Fig. 3.6):

$$M = USV^T.$$

S is a diagonal matrix of size $G \times C$ and its diagonal elements are known as singular values of M. Furthermore, the matrices U (with elements u_{gc}) and V contain left- and right singular vectors, respectively, which are represented by the

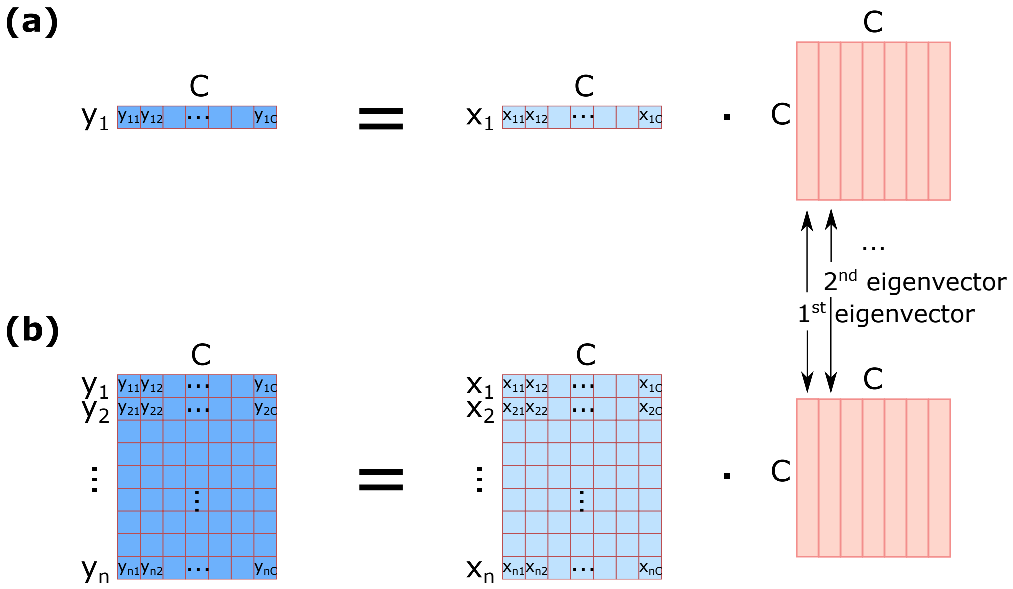


Figure 3.4: Computation of new coordinates in PCA. (a) To compute new coordinates $(y_{11}, y_{12}, ..., y_{1C})$ of an observation x_1, the vector of original coordinates $(x_{11}, x_{12}, ..., x_{1C})$ needs to be multiplied by a matrix of eigenvectors. (b) To compute new coordinates for a vector of all observations $(x_1, x_2, ..., x_C)$, the matrix with original coordinates for all observations needs to be multiplied by a matrix of eigenvectors.

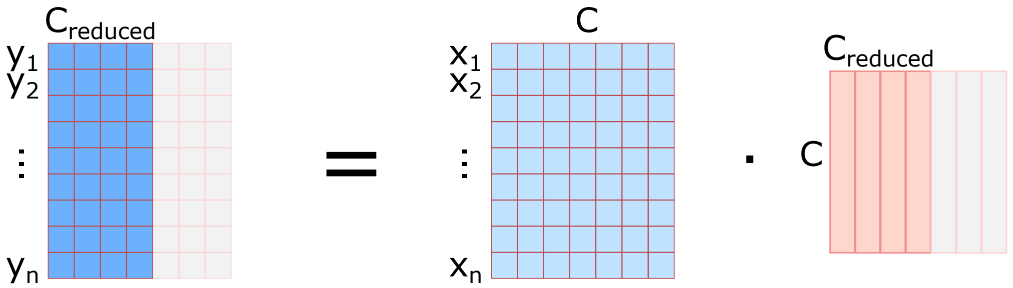


Figure 3.5: Dimensionality reduction in PCA. To reduce the dimensionality of a data matrix from C to C_reduced a matrix of original coordinates for all observations needs to be multiplied by a matrix with C_reduced eigenvectors.

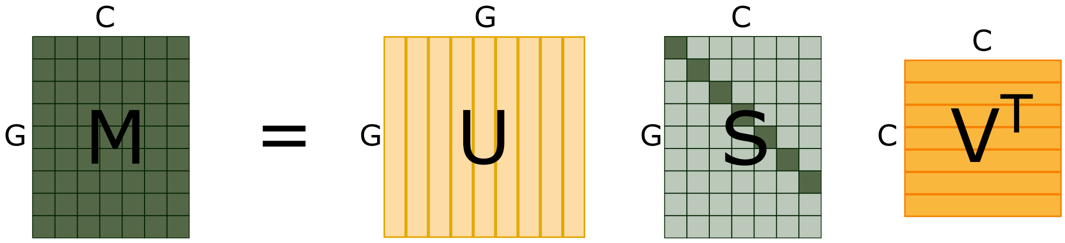


Figure 3.6: Generalized form of SVD. Matrix M of size $G \times C$ is submitted to SVD and gets decomposed into a product of three matrices: U $(G \times G)$, S $(G \times C)$, and transposed V $(C \times C)$.

columns. The matrix U is of size $G \times G$ and its columns are orthonormal, i.e. $U^T U = I_G$, where I_G is an identity matrix of rank G. Similarly, the matrix V is of size $C \times C$ and its columns are orthonormal, i.e. $V^T V = I_C$, where I_C is an identity matrix of rank C (see Fig. 3.7).

Importantly, the columns of U are eigenvectors of MM^T, the columns of V are the eigenvectors of $M^T M$, and the diagonal elements of S are square roots of eigenvalues of MM^T and $M^T M$.

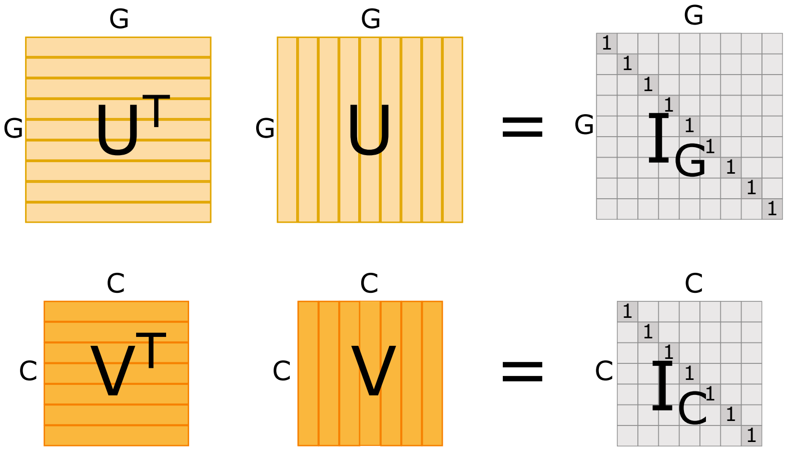


Figure 3.7: U and V are orthogonal matrices. U is a matrix of left singular vectors of M and V is a matrix of right singular vectors of M.

3.3 Link between SVD and PCA

Instead of eigenvalue decomposition, PCA can also be conducted by applying SVD to the covariance matrix $\sum$. As shown in Section 3.1, the covariance matrix of the matrix M is calculated using the formula:

$$\sum = \frac{1}{n} M^T M.$$

After applying SVD to the matrix M, which results in:

$$M = USV^T,$$

we can represent the matrix $\sum$ as:

$$\begin{aligned} \sum = \frac{1}{n}(USV^T)^T(USV^T) = \\ = \frac{1}{n} VS^T U^T USV^T. \end{aligned}$$

Since the columns of U are orthonormal and S is a diagonal matrix, we obtain:

$$\sum = \frac{1}{n} VS^2V^T,$$

which resembles the diagonalization formula of $\sum$ (Fig. 3.2). Thus, the columns of V represent the eigenvectors of $M^T M$, while the diagonal elements of S represent the square roots of eigenvalues, the so-called singular values. In other words, the right singular vectors of matrix M are equivalent to the eigenvectors of $M^T M$. Thus, the new PCA coordinates of M can be computed by multiplying M by V, where V is the matrix with the right singular vectors of M. Finally, this results in:

$$MV = USV^T V = US,$$

thus the matrix of new coordinates can be computed by multiplying the matrix of left singular vectors of M by the singular values of M. This demonstrates that SVD can be seen as an alternative way of performing PCA. Nowadays, owing to its shorter computation time, the approach involving SVD is more common for computing PCs than the one involving eigenvalue-decomposition.

3.4 Correspondence analysis (CA) - intuitive explanation

CA is a data projection method which allows for a simultaneous embedding of both variables and observations from a data matrix in one real-valued space. For example, by applying CA to single-cell transcriptomics data the resulting space would contain both genes and cells. We call the resulting space with points for cells and for genes *CA-space*. In this space, CA places cells with similar transcript profiles near each other, and, likewise, arranges genes with similar distribution over cells near each other. Most importantly for our application, in CA-space a cluster of similar cells defines a direction from the origin to that cluster, and genes which are highly expressed in this cluster but not elsewhere lie in that very direction. We will use this feature of CA as a starting point for generating Association Plots.

Similar to other data projection methods, points from the CA-space are traditionally projected down into two or three dimensions. While in terms of visualization purposes such low-dimensional data projections usually allow for a basic understanding of the data, we refrain from doing so. Instead, we keep a large number of dimensions of the original data so as to reduce noise while maintaining the defining information. We discuss the practical aspects of dimension reduction and the choice of number of dimensions to keep in Section 4.4.

3.5 CA - mathematical explanation

Correspondence analysis (Benzécri, 1973; Greenacre, 1984, 2017) is a projection method for visually representing a data matrix in a high-dimensional space. While it was originally intended for analysis of contingency tables, Gower *and others*, 2011, p. 290, state that "the elements of $\mathbf{X}$ [the data matrix] may contain any nonnegative values". Unlike PCA, CA does not submit the data matrix M itself to a singular value decomposition, but the "object of interest" is the matrix of Pearson residuals derived from M.

CA is computed in the following steps. By m_{gc} we denote a value in the gth row and cth column of M, and by m_{++} the grand total of M. One calculates an observed proportion matrix P with elements $p_{gc} = m_{gc}/m_{++}$ and uses this for calculating row and column masses. The gth row mass, p_{g+}, is defined as the sum across the g^{th} row, and the cth column mass, p_{+c} , as the sum of all values in the cth column of P. With the expected proportions $e_{gc} = p_{g+} * p_{+c}$ Pearson residuals $f_{gc} = (p_{gc}–e_{gc})/\sqrt{e}_{gc}$ are computed. For comparison to contingency tables, note that the sum of the squares of all Pearson residuals, multiplied by the grand total m_{++} (Greenacre and Blasius, 1994, (3.2.12)) would form the χ^2 statistic.

In analogy to PCA, it is now this matrix $F = (f_{gc})$ that gets submitted to singular value decomposition, factoring it into the product of three matrices: $F = USV^T$. S is a diagonal matrix and its diagonal elements, s_{cc}, are known as singular values of F. Furthermore, the matrices U (with elements u_{gc}) and V (with elements v_{gc}) contain left- and right singular vectors, respectively, which are represented by the columns. From the left and right singular vectors coordinates of points for rows and columns in an high-dimensional space are computed. The coordinates of $\nu^{(g)}$ depicting the g^{th} row are defined as $\nu_n^{(g)} = u_{gn}/\sqrt{p}_{g+} * s_{nn}$, for $n = 1, \ldots, m$, where $m = \min(G, C) - 1$ (Greenacre, 2017, (A.8)). The coordinates of $\omega^{(c)}$ representing the c^{th} column are given by $\omega_n^{(c)} = v_{cn}/\sqrt{p}_{+c}$ (Greenacre, 2017, (A.7)). In the literature (Greenacre, 2017) this choice of scaling is called "asymmetric", with the rows represented in "principal coordinates" and the columns in "standard coordinates".

It is a key feature of correspondence analysis that one can interpret the joint map of points for rows, the ν's, and columns, the ω's, in the same space. The (full) dimension of this space will be $\min(G, C) - 1$, which above we called m, (Greenacre, 2017, p. 203) and both sets of points can be thought of as elements of $\mathbb{R}^m$. We will refer to this space as the CA-space. Traditionally, one uses only the first two or three dimensions and calls this a biplot. In the examples in Section 4.3 we will also depict the two-dimensional biplots, although only for illustration purposes. The focus of our exposition, however, is on large data sets where too much information would be lost upon projection into two dimensions. Instead of "explained variance" that is used in PCA, CA speaks of "inertia". Inertia is the sum of the squares of the elements of the matrix F, or in other words, the χ^2 statistic divided by the grand total m_{++} (Greenacre, 2017, p. 28). Just like the explained variance in PCA, inertia gets approximated increasingly better with increasing number of dimensions used.

A particular question which can be answered using CA pertains to the associations between rows and columns, or between rows and a cluster of columns. For a mathematical definition of association we follow the logic of contingency tables, where the likelihood ratio for the entry in cell (i, j) (the entry in the i^{th} row and j^{th} column) to be a product of chance would be $\frac{p_{ij}}{e_{ij}}$, the observed frequency in the cell divided by the expected frequency. When this ratio is close to 1, we have little reason to believe that there is an association, whereas a large ratio hints at an association between the row and the column. Subtracting 1 from the ratio we can write this as $\frac{p_{ij}-e_{ij}}{e_{ij}}$, which will be near 0 in the absence of an association. We call this quantity the "association ratio" and abbreviate it as $a(i, j)$.

Since we are also interested in the association between observation and a cluster of variables, we proceed to add up the association ratio for the respective cells of the matrix. Let $\mathcal{C}, |\mathcal{C}| = K$ be a cluster of K variables $j_1, \ldots, j_K$ which behave similarly. We extend the notation to clusters by defining

$$a(i, \mathcal{C}) := \frac{1}{K} \sum_{l=1}^{K} a(i, j_l)$$

and call this the association ratio of the gene with the cluster $\mathcal{C}$.

The geometry of the Association Plots which will be defined in the next chapter rests on two key features of CA (Greenacre, 2017). First of all, a column-point can be expressed as a weighted sum of row-points, where the weights are related to the contribution of the rows to the particular column in P. This is Greenacre's transition equation (Greenacre, 2017, (A.16)). It is the reason for the clustering of similar observations and similar variables, respectively, in CA space.

Second, what is even more important for our application is that the inner product of i^{th} row- and j^{th} column-vector approximates the respective association ratio, i.e.

$$a(i,j) = \langle \nu^{(i)}, \omega^{(j)} \rangle + \epsilon$$

where in the m-dimensional CA-space the error term $\epsilon = 0$. This is Greenacre's reconstitution formula (Greenacre, 2017, formula (13.5) or (A.14)). It pertains directly to our goal of describing association in geometrical terms: Due to the inner product, the observations that are associated to a variable lie in the direction of that variable; the more aligned the two are, and the longer the vectors, the higher the association. When a low-dimensional projection is permissible, this feature is usually clearly visible.

Note that the reconstitution formula also allows for generalization to clusters: observations that are associated to a cluster of variables lie in the direction of these variables. Since the inner product is bi-linear, it also means that the association ratio sums nicely over groups of observations or clusters of variables. This will be utilized in Section 4.2.

CHAPTER 4

Association Plots

In this chapter we present Association Plots, a tool for identification of cluster-specific genes from high-dimensional transcriptomics data. The content of this chapter was published in bioRxiv (Gralinska and Vingron, 2020) and the manuscript is under review, as of March 2022. The parts of this chapter come also from Gralinska *and others* (2022).

4.1 Intuitive explanation of Association Plots

Association Plots are primarily a visualization tool for genes associated to a cluster of cells or samples. To give an intuitive overview of Association Plots we will present them in the context of transcriptomics data. When applied to such data, they enable to depict genes associated to a cluster of conditions (e.g. cells from single-cell data). This capability of Association Plots is derived from a feature of CA, which was introduced in Sections 3.4 and 3.5 and which is summarized in Fig. 4.1.

As shown in Fig. 4.1a, in CA-space cells with similar transcript profiles are located near each other and create a cluster. To identify genes specific for a selected cluster using an Association Plot we need to focus on the direction defined by the vector from the origin to the centroid (grey dot with orange border) of the given cluster (orange dots). Genes that are associated to the cluster (black dots) lie in this direction in space. Note that this would be the same geometry even in a space of much higher dimension. The stronger an association between a gene and a cell cluster, the farther out towards this cluster a gene will be located. Therefore, the length of the orthogonal projection of the gene-point onto the vector towards the centroid is an indicator of the strength of the association. As shown in Fig. 4.1b, we use this length $(d * cos(\gamma))$ as the x-axis for the gene-point in the Association Plot. The perpendicular distance from a gene to this vector $(d * sin(\gamma))$ constitutes the y-axis of the gene-point in the Association Plot. This distance will be short when the gene is very specific for the cluster. When the gene is also expressed in other clusters, then it will be farther away from the direction to the centroid, and therefore have a larger y-coordinate in the Association Plot.

The Association Plot for a given cell cluster depicts all genes with these two coordinates in a two-dimensional space. Genes positively associated with the selected cluster will be located in the right bottom part of the plot. This is due to the fact that in the CA space such genes align with the direction towards this cluster and are located in close proximity to the cluster centroid. On the other hand, genes which do not show any association with the cluster will be located close to the Association Plot's origin. In the CA space such genes do not align with the direction towards the selected cluster and are located closer towards other clusters.

Due to their features, Association Plots can be seen as a tool for identification and visualization of cluster-specific genes. Importantly, despite being planar Association Plots are independent of the dimension of the CA-space thanks to which they capture information from the high-dimensional space without discarding dimensions.

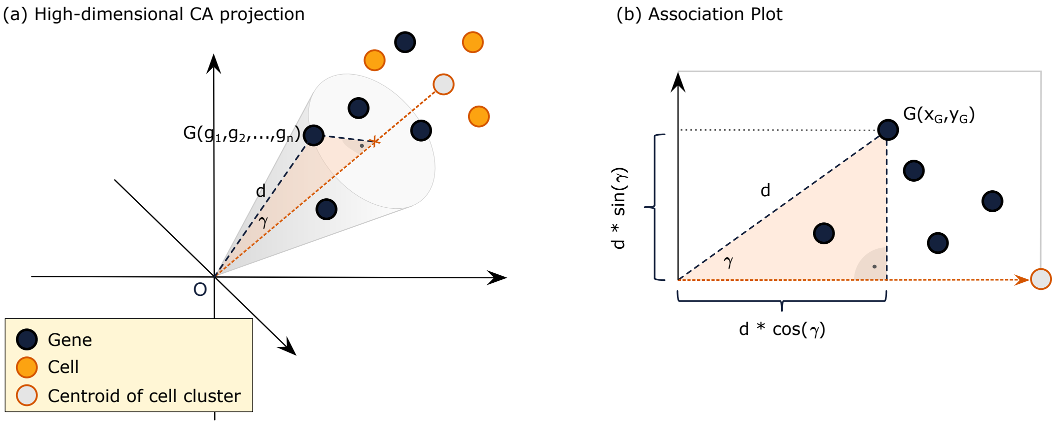


Figure 4.1: Association Plots delineate cluster-specific genes. (a) In a high-dimensional CA space a cluster of cells (orange dots) defines a direction, here represented by the orange line pointing from the origin to the centroid of the cell cluster. The genes (black dots) associated to this cluster of cells are located close to this line along its direction. **(b)** For the Association Plot we only use the length from the origin to the genes‘s projection onto the orange line (d*cos(γ)) as the first coordinate of the gene in the Association Plot. The length of the perpendicular distance from the gene to the line (d*sin(γ)) is the second coordinate. Thus, the x-axis of the Association Plot corresponds to the line pointing towards the cluster centroid shown at the end of the vector.

4.2 Mathematical explanation of Association Plots

A cluster of variables $\mathcal{C} = j_1, \ldots, j_K$ can be represented by the centroid of its variable vectors $\omega^{(j_l)}$, $l = 1 \ldots K$ in CA-space. We call this centroid $\vec{X}$:

$$\vec{X} := \frac{1}{K} \sum_{l=1}^{K} \omega^{(j_l)}$$

Due to linearity, we can see the average association ratio also in the inner product between an observation and the vector from the origin to this centroid. Let $\vec{r}$ be an observation-vector in CA-space representing row r of the data. Then we can express the association between r and the cluster $\mathcal{C}$ as:

$$a(r, \mathcal{C}) = \frac{1}{K} \sum_{l=1}^{K} \langle r, \omega^{(j_l)} \rangle + \epsilon = \langle \vec{r}, \vec{X} \rangle + \epsilon$$

This is a trivial consequence of the reconstitution formula and the definitions from above.

This observation forms the basis for a simple 2-dimensional visualization of the high-dimensional information. The inner product is determined by the length of $\vec{r}$, $|\vec{r}|$, the length of $\vec{X}$, $|\vec{X}|$, and the angle between the two vectors. We call the angle $\phi(\vec{r})$, or just ϕ where the context is clear. In this notation, the inner product from above can be written as

$$\langle \vec{r}, \vec{X} \rangle = |\vec{r}| \; |\vec{X}| \; \cos(\phi(\vec{r}))$$

Therefore it makes sense to introduce a 2-dimensional representation where the x-axis corresponds to the direction of the centroid vector, and we represent $\vec{r}$ by the following x- and y-coordinates:

$$x(\vec{r}) := |\vec{r}| \; \cos(\phi(\vec{r}))$$

$$y(\vec{r}) := |\vec{r}| \; \sin(\phi(\vec{r}))$$

Clearly, $|\vec{r}|\cos(\phi(\vec{r}))$ is the length of the projection of $\vec{r}$ onto $\vec{X}$, or $\langle\vec{r}, \frac{\vec{X}}{|\vec{X}|}\rangle$, and $|\vec{r}|\sin(\phi(\vec{r}))$ is the length of the orthogonal distance of $\vec{r}$ to $\vec{X}$, or $|\vec{r} - \frac{x(\vec{r})}{|\vec{X}|}\vec{X}|$. Also, $|\vec{r}|$ is equal to the length of the vector $(x(\vec{r}), y(\vec{r}))^T$. We define the Association Plot for cluster $\mathcal{C}$ as the 2-dimensional plot where each observation-vector $\vec{r}$ in CA-space is represented by these 2-dimensional points $(x(\vec{r}), y(\vec{r}))$.

Introducing $\tilde{X}$ as the 2-dimensional vector

$$\begin{pmatrix} |\vec{X}| \\ 0 \end{pmatrix} =: \tilde{X}$$

we can ascertain the conservation of the inner products, and with it the association ratio, between CA-space and Association Plot:

$$a(r, \mathcal{C}) = \langle\vec{r}, \vec{X}\rangle + \epsilon = |\vec{r}|\ cos(\phi)\ |\vec{X}| + \epsilon = \langle\begin{pmatrix} |\vec{r}|\ cos(\phi) \\ y(\vec{r}) \end{pmatrix}, \begin{pmatrix} |\vec{X}| \\ 0 \end{pmatrix}\rangle + \epsilon = \langle\begin{pmatrix} x(\vec{r}) \\ y(\vec{r}) \end{pmatrix}, \tilde{X}\rangle + \epsilon$$

This demonstrates that the association ratio can be seen both as an inner product in the m-dimensional CA-space as well as an inner product in the 2-dimensional Association Plot. When the full m dimensions are used for the vectors $\vec{r}$ and $\vec{X}$, the error term ϵ will be 0. In Section 4.4 we will propose to use fewer dimensions than m, actually relying on the approximation.

From this simple line of algebra above one also notes that $y(\vec{r})$ gets multiplied with 0. This means that the inner product, and with it the association ratio between observation and cluster, is constant along vertical lines in the Association Plot. Intuitively, this is due to the fact that by definition the association ratio $a(r, \mathcal{C})$ calculated for a given cluster of variables is not influenced by other variables or clusters in the data, which might also attract an observation. The angle ϕ contributes information because the less competition there is for an observation from other clusters, the smaller will be ϕ, whereas when an observation is shared also by other clusters, this will be reflected in a larger ϕ. This will be visible in the example below, and will be studied further in the section on significance of the visual patterns (Section 4.5).

To aid interpretation one can also embed samples into the Association Plot using the same coordinate system of projection length onto the centroid vector vs. orthogonal distance to it. The examples to follow will illustrate that closely clustered sample points in an Association Plot indicate a coherent cluster, while widely spread out points indicate a heterogeneous cluster. When one suspects that two clusters are close to each other, one can also display their respective positions in the Association Plot to check on the proximity between clusters. This will be done for the GTEx data in Section 5.4.

4.3 Simple data examples

While the ideas laid out in this chapter are really meant for the analysis of large data sets, we want to first show in two simple examples how Association Plots work. We will start with a simulated data set and then proceed to employment data from the International Labor Organization.

4.3.1 Simulated data

To demonstrate the concept of Association Plots, we apply Association Plots to a simulated data set of 100 observations (row categories) and 15 variables (column categories), where the 15 variables fall into 5 clusters of 3 variables each. This data was generated using the function *make_blobs* from the Python library *scikit-learn* (Pedregosa *and others*, 2011) which can be used for generating isotropic Gaussian point clouds for clustering. The clusters of variables are clearly visible upon projection into a two-dimensional correspondence analysis biplot (Fig. 4.2), in which the observations (light blue circles) span the space between the five clusters of variables (crosses).

We aim at delineating the cluster-specific observations characterizing each cluster. As described in Section 3.5, in CA-space the observations characteristic for a given cluster of variables will be located in the direction of this cluster. Therefore, in the two-dimensional CA projection of the simulated data the observations which are located in the direction of cluster 1 (red crosses) are expected to be specific

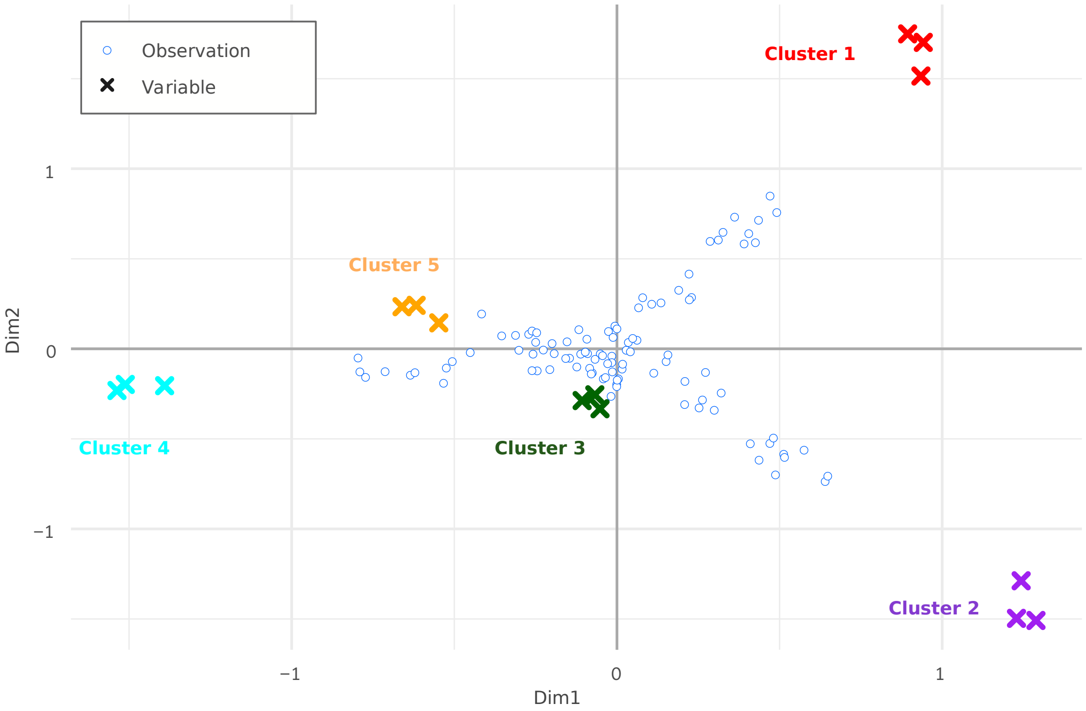


Figure 4.2: Two-dimensional CA projection of the simulated data. The projection arranges the cluster members in proximity to each other. Observations pointing toward clusters 1, 2, and 4 are recognizable, while a direction toward cluster 3 or 5 is not visible in this two-dimensional projection.

for this cluster. In the CA biplot the observations located in the direction towards cluster 1 are easy to notice, as the planar projection happens to nicely resolve this.

Alternatively, the cluster-specific observations can be visualized in the Association Plot for cluster 1 (Fig. 4.3a). In this plot one can observe that all three variables belonging to cluster 1 are clearly separated from the other variables, and the observations associated to cluster 1 are located in the right part of the plot. This selection of observations is confirmed in Fig. 4.3b where one can see that the 10 observations with the highest specificity for cluster 1 are indeed highly over-represented in cluster 1 in comparison to the remaining clusters.

The challenge, however, lies in delineating the cluster-specific observations also for clusters which are not as nicely visible in the low-dimensional biplot as cluster 1. For example, identification of observations specific for cluster 3 (dark green

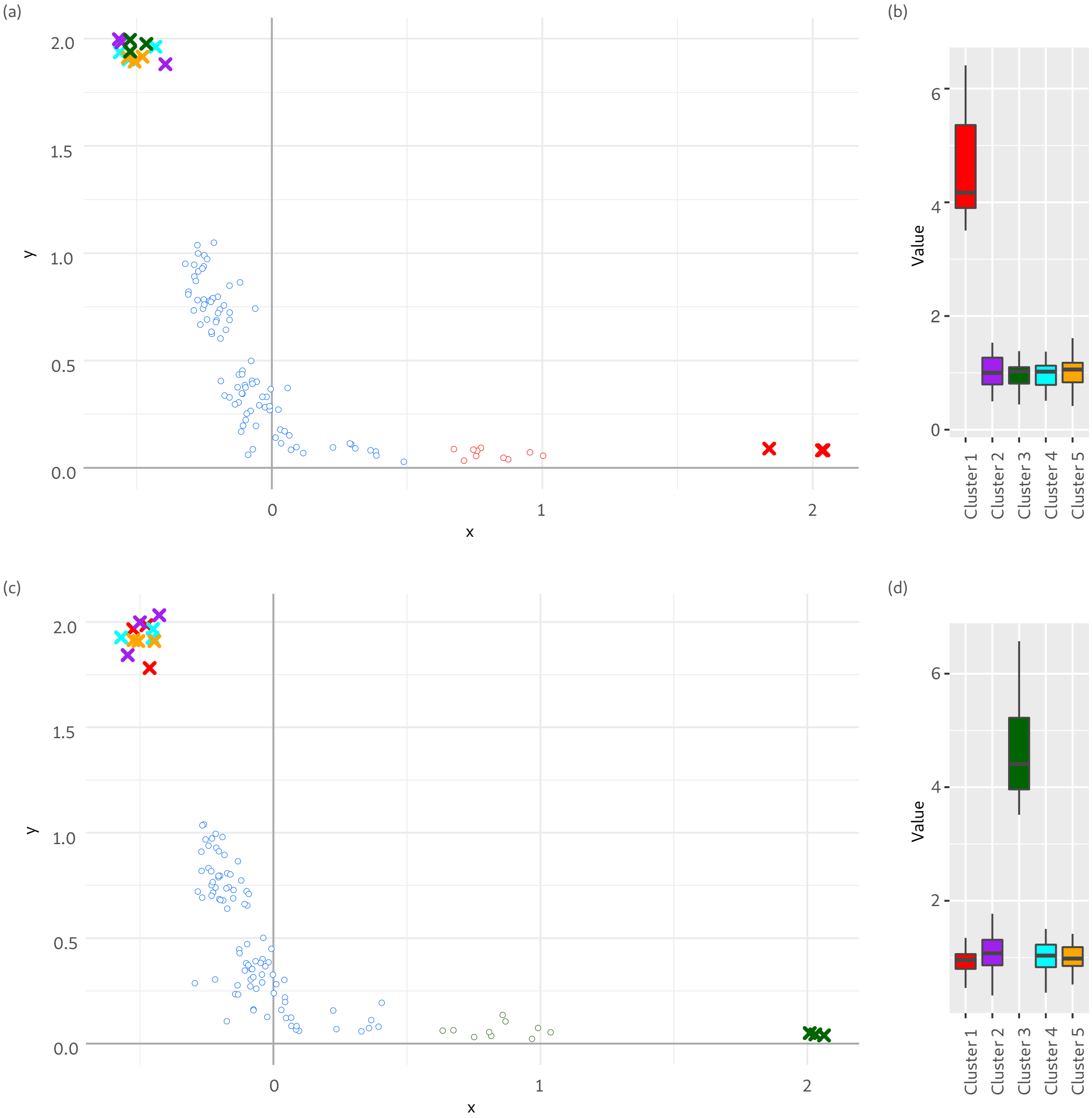


Figure 4.3: Identification of cluster-specific observations from the simulated data using Association Plots. (a, c) Association Plots were generated for (a) cluster 1 and (c) cluster 3 of variables using four CA dimensions. 10 observations with the highest specificity for given cluster (according the generated Association Plot) are highlighted in (a) red or (c) green. **(b, d)** Over-representation of 10 detected cluster-specific observations in (b) cluster 1 or in (d) cluster 3.

crosses) from the simulated data is not possible based on the two-dimensional data projection. The Association Plot generated for cluster 3 (Fig. 4.3c) reveals a set of observations characteristic for this cluster, which can again be confirmed in a box plot of the observations (Fig. 4.3d).

4.3.2 Employment data

Our second illustrative example is a real data set and therefore not as "clean" as the simulated data. It comes from the International Labour Organization and describes people's sector of employment in 233 countries (International Labour Organization). Employment sectors are agricultural sector, industry sector, services sector, or unemployed. Data are further divided according to gender (m/f) and year (2000, 2005, 2010, 2015). This results in a matrix with 233 rows for the countries and 32 columns for employment category, gender, and year. For example, the column "industry, female, 2005" would represent the percentage of female labor force of a given country that was employed in industry in 2005. Details on the data set can be found in Appendix C.

Data categories are "agriculture", "industry", "services", and "unemployed", further refined by gender. These form homogeneous groups in the data which is easily confirmed by visual inspection of the two-dimensional CA projection (Fig. 4.4). The year from which the data comes seems to have a lesser influence - at least in the 2D projection.

The blue dots in the biplot represent the countries, and their location within the plot provides a clue on the employment profile of that country in 2000-2015. For example, countries in which a high percentage of the labor force was employed in the agricultural sector are located towards the agricultural sector cluster. Based on this it is possible to identify countries which were the employment leaders in the agricultural sector. This is a simple application of the interpretation of the directions as discussed with the reconstitution formula. However, the clusters for industry and services lie very close to each other in the two-dimensional projection, and it is hard to discern whether there are countries specifically associated to either of the two categories.

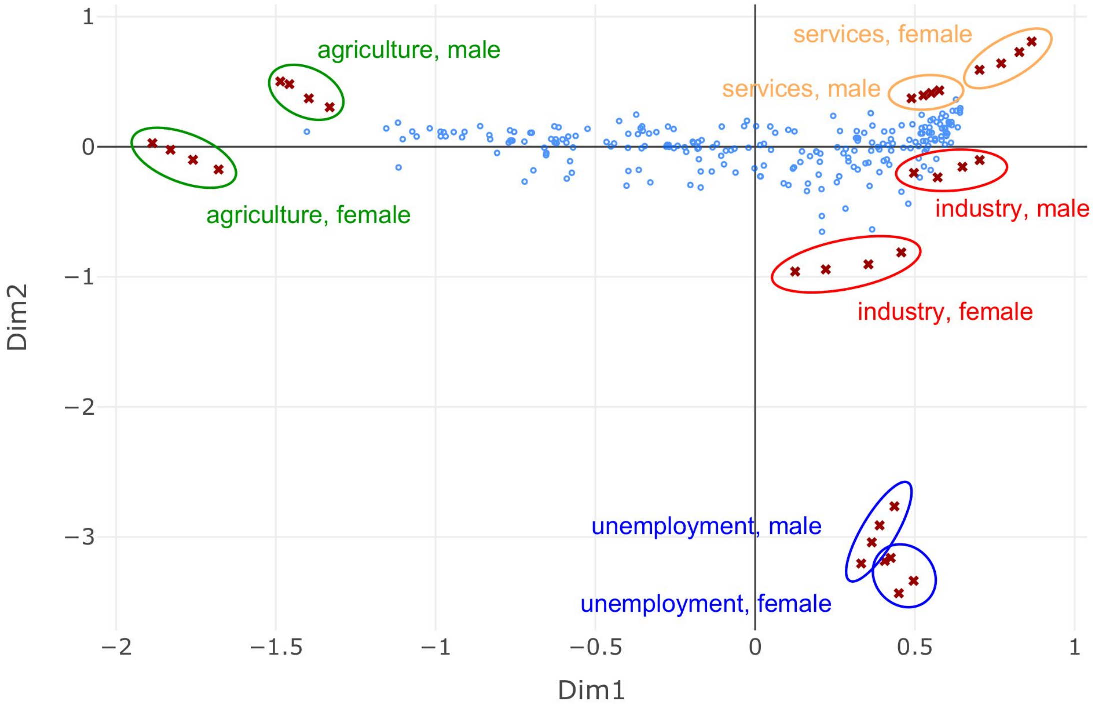


Figure 4.4: Two-dimensional CA projection of the employment data. The projection allows discerning four main clusters: "agriculture sector", "services sector", "industry sector", and "unemployed". Each of these clusters contains data points from four years (2000, 2005, 2010, and 2015). The location of the countries, represented in the biplot by blue dots, implies their employment profiles in the years 2000-2015.

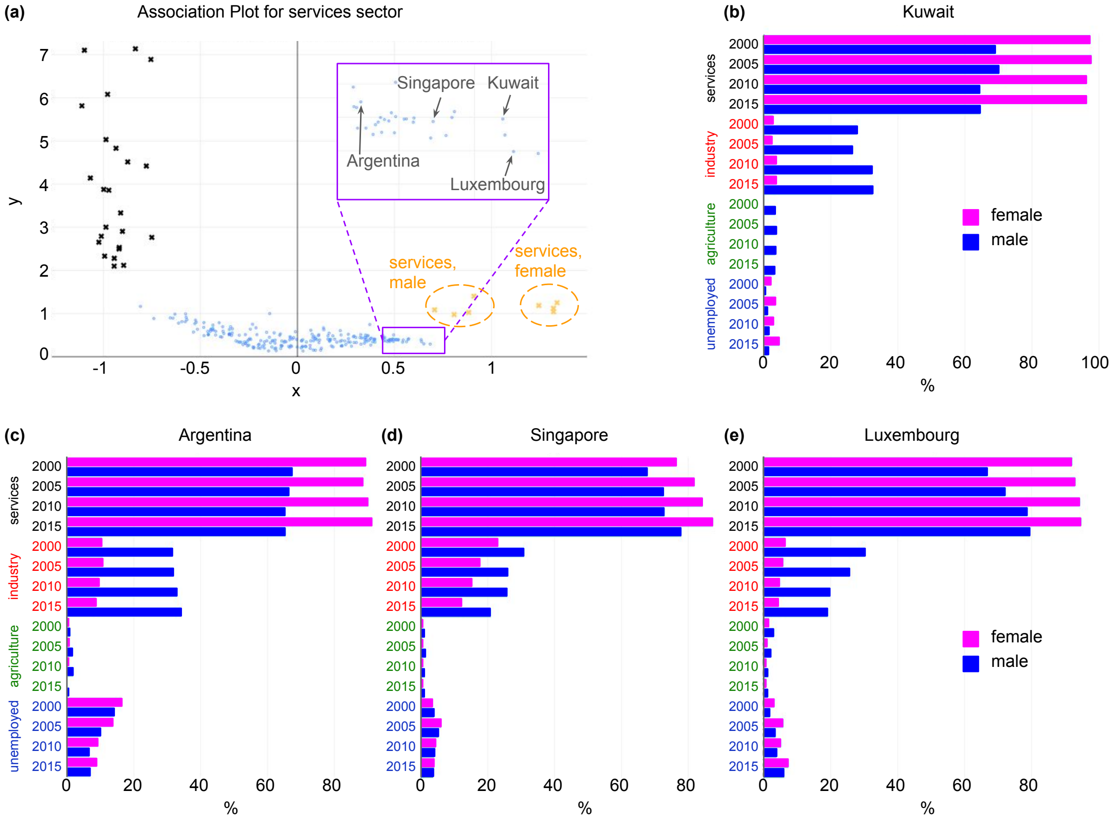


Figure 4.5: Association Plot for services sector. (**a**) The Association Plot was generated from the employment data using eight CA dimensions. The countries with the highest percentage of population employed in the services sector (yellow crosses) are located towards the right part of the plot. The names of four example leader countries are shown. The countries with the lowest proportion of population employed in services are located towards the left part of the plot, in the direction of other employment sectors (black crosses). (**b**-**e**) Employment profiles of the example leaders in services sector: (b) Kuwait, (c) Argentina, (d) Singapore and (e) Luxembourg. The presented barplots illustrate the percentages of female and male population in a country employed in different sectors in the years 2000-2015.

Finding out which countries are the leaders in, for example, the services sector is made possible by the respective Association Plot (Fig. 4.5a). In the figure one can clearly see that the services category (represented by yellow crosses) is separated from the other employment clusters (black crosses). The yellow crosses are in turn divided into two groups, which correspond to the male and female data. The countries with the highest percentage of labor force employed in the services are located towards the right part of the plot. Starting from right, the leaders are: Hong Kong, Luxembourg, Guam, and Kuwait, followed by Macao, Saudi Arabia, Brunei Darussalam, Singapore, Netherlands and the United Kingdom. Representative barplots for some of these countries are given in Fig. 4.5b-e. The countries with the lowest percentage of the labor force employed in services are on the opposite side of the Association Plot (Burundi, Ruanda, Central African Republic, Niger and Rwanda). The Association Plot visualizes this information while it would be invisible in the low-dimensional projection. The Association Plot for the industry cluster (Fig. 4.6) actually indicates a lack of countries exclusively associated to industry. This is represented by a clear distance between industry categories and countries, as well as by the lack of a tail of points extending toward the cluster members in the Association Plot generated for the services sector.

4.4 Selecting the dimension number

Traditionally, both PCA and CA are performed with the goal of depicting the information either in the plane or, possibly, in three dimensions. While for small data sets such low-dimensional data representation is usually satisfying, for large data sets it is often associated with a large loss of information. In such the case, the SVD can still play an essential role and provide for noise reduction by canceling the dimensions that belong to small singular values. Although employing this mechanism implies a projection into a still high-dimensional subspace which cannot be visualized, it allows for maintaining only the relevant information in the data. We adopt this procedure for our purposes and use it for noise reduction in the CA space, followed by computation of Association Plots.

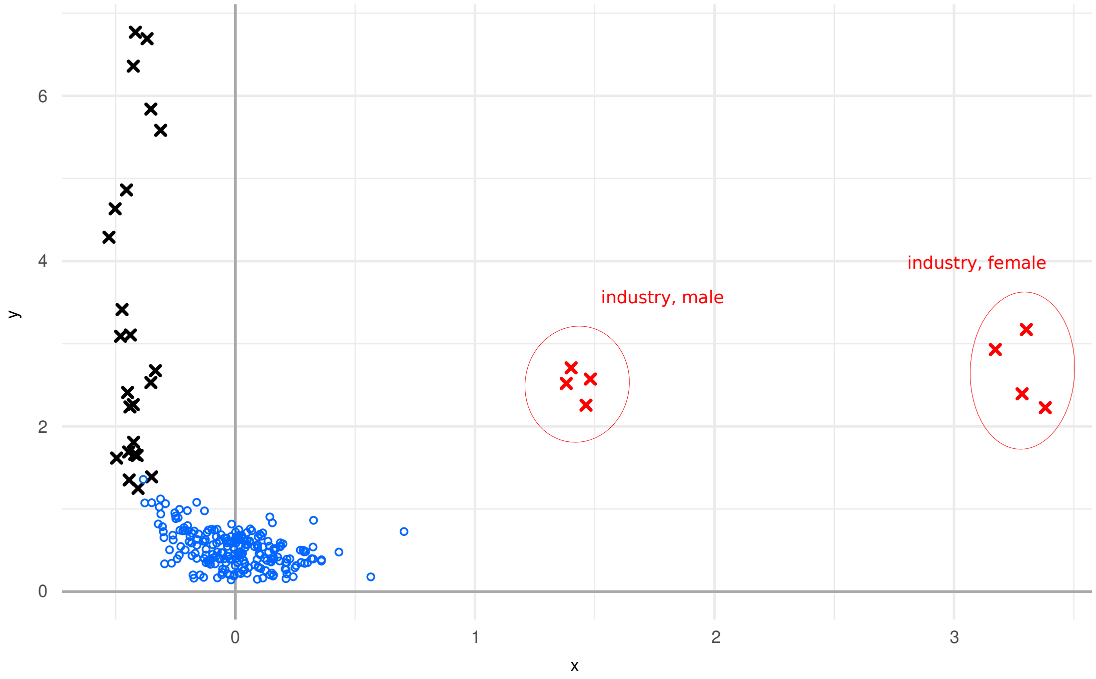


Figure 4.6: Association Plot for the industry sector. The Association Plot was generated from the employment data using eight CA dimensions. Each blue dot represents one country, each cross represents one employment category (services sector, industry sector, agriculture sector, unemployment) for males or females in years 2000, 2005, 2010, or 2015. Clear distance between industry categories and countries, as well as the lack of a tail of points extending toward the cluster members in the Association Plot indicates a lack of countries exclusively associated to industry.

An important consideration when conducting noise reduction is how many dimensions should be retained in the analysis. However, there is no clear guideline (Greenacre and Blasius, 1994) and many papers have been written on the problem of selection of the right number of singular values (see the literature on spectral clustering, e.g., Zelnik-Manor and Perona, 2005; Von Luxburg, 2007). For instance, when data happen to fall nicely into clusters, a common approach is to retain the number of dimensions which roughly reflects the number of clusters in the data. Therefore, the Association Plot in the employment example in Section 4.3.2 was computed for eight dimensions.

The other commonly used methods are computational and estimate the number of dimensions by analyzing the scree plot, i.e. the plot showing the sorted singular values from largest to smallest. One of such approaches, the “elbow rule” (Ciampi *and others*, 2005), is based on a scree plot for original and randomized data. A randomized data matrix is generated by permuting the values for each variable separately. By applying correspondence analysis to it a vector of sorted randomized singular values is calculated. After repeating these steps multiple times we compute an average of the first randomized singular values, the second randomized singular values, etc. The number of dimensions to retain is then chosen as the intersection point between the actual singular values and the average of the randomized singular values. See Fig. 4.7 for an example.

The two further approaches also use scree plots. The first one retains the minimum number of dimensions which in total account for more than 80% of total inertia (Greenacre and Blasius, 1994). We call this approach the ”80% rule“. The second one, which we call the ”average rule“, retains those dimensions that explain more inertia than one dimension on average (Greenacre and Blasius, 1994).

All three presented approaches for estimating the number of CA dimensions to retain for generating Association Plots are also implemented in our software described in Chapter 7. Additionally, in Section 5.3 we will study the influence of the dimension number retained in the analysis on the results and structure of Association Plots, and we will discuss the real-life meaning of this choice.

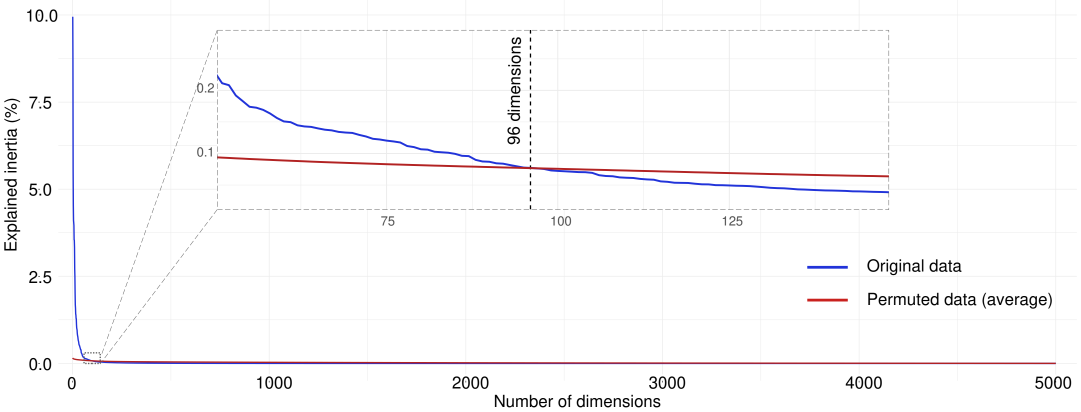


Figure 4.7: Elbow plot generated for the heart samples from GTEx data. Blue line represents the percentage of explained inertia from the original data. Red line represents the average percentage of explained inertia from 1,000 permutations of the data. The number of 96 dimensions results from the intersection point between the original data and the average of the permuted data.

4.5 S_α score

To obtain a better understanding of the visual patterns we observe in the Association Plot we study random data. For CA alone, randomized data would lead to a dense, roughly ellipsoidal cloud of gene-points around the origin. For a random Association Plot, however, it does not suffice to randomize the data, but one also needs to define a random cluster of conditions to orient the Association Plot in the direction of the centroid. Taken together, we first randomize a given data set by permuting the rows of the data matrix following Tusher *and others*, 2001. Next we select a number of conditions to form a random cluster. The centroid of this random cluster will determine an arbitrary direction in space. Additionally, we have observed that the cardinality of the cluster also influences the appearance of a random Association Plot. This seems to be due to the angle between the centroid and the cluster members depending on the size of a cluster.

Fig. 4.8a shows such an Association Plot generated from randomized data, where a "pseudo"-cluster of 600 conditions was selected for plotting the Association Plot. In fact, the randomized data in this example come from the example data set (GTEx) discussed in Section 5.1, and the number of cluster members corresponds

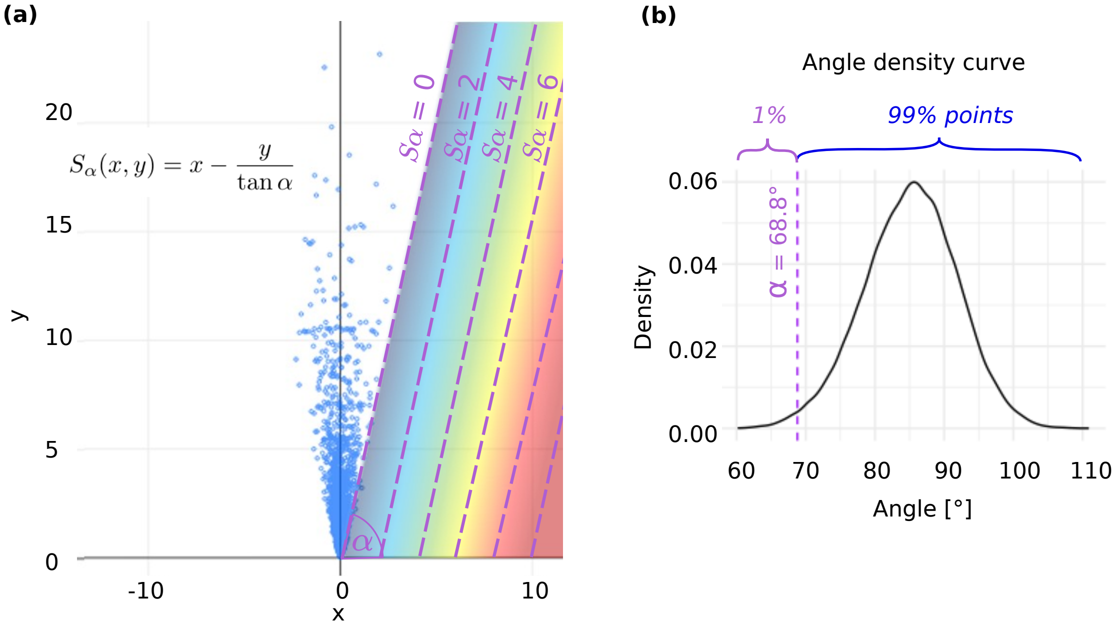


Figure 4.8: Scoring system of candidate genes. (a) Association Plot generated from randomized GTEx data for a "pseudo"-cluster comprising 600 conditions. For each point the angle between a given point and the x-axis was calculated. 1% of points with the smallest angle determines the α threshold, which will be further used for calculating the gene scores S_α in the original data. **(b)** Distribution of the angles between points from (a) and the x-axis. In this example the threshold of 1% resulted in $\alpha = 68.8°$.

to the cardinality of the heart cluster. One observes a V-shaped cloud of points, which is the Association Plot's view on the dense cloud of points around the origin in CA-space. The width of the V provides information as to the area occupied by chance and to the right of which relevant information may start. Thus, we aim to determine the angle of the ray delineating the V towards the right. The empirical distribution of the number of points falling to the right of the delineating line is shown in Fig. 4.8b. It allows for the choice of a threshold on the angle α. We have chosen the line with degree 68.8° which delineates 1% of the points to the right of the V.

Based on the angle thus determined we propose the following heuristic choice of a scoring function $S_\alpha(x, y)$ for an individual point (x, y) in the Association Plot:

$$S_\alpha(x, y) = x - \frac{y}{\tan \alpha}.$$

S_α is 0 along the delineating line of degree α which the simulation yielded and which in Fig. 4.8a is annotated with "$S_\alpha = 0$". The scoring function S_α is designed such that parallels to this line constitute level lines of increasing S_α as they shift to the right (see Fig. 4.8a). This serves the purpose of giving higher scores to points further towards the right, while at the same time decreasing the scores as one moves upward. With this choice one adds additional power of distinguishing among genes which otherwise would have the same association value with respect to a cluster. The scoring function S_α can, e.g., be used to rank the genes based on an Association Plot. The data examples in the next two chapters will provide examples illustrating and supporting this choice of the scoring function.

CHAPTER 5

Association Plots Applied To Bulk RNA-Seq Data

This chapter presents the application of Association Plots to an example bulk RNA-seq data and demonstrates how they aid in investigating similarities between clusters from the input data. The content of this chapter is an adapted and partially extended part of a manuscript posted on bioRxiv (Gralinska and Vingron, 2020) and which is under review, as of March 2022.

5.1 Introduction to GTEx data

To demonstrate the utility of the Association Plots in finding cluster-specific genes we present an application of our method to a large biological data set: the Genotype-Tissue Expression ("GTEx") data (Carithers *and others*, 2015). This data set comprises 11,688 tissue samples generated by RNA-seq from 30 human tissues collected from different donors (see Appendix C for GTEx data processing). Each column of the data input matrix contains one of many replicates for each tissue, whereas the rows of the matrix correspond to the genes whose expression levels were measured. In total, we use the 5,000 genes with the largest variance across columns of the chi-square component matrix. The tissue information provides a natural clustering of the columns, and we interpret each tissue as one cluster. Altogether, the matrix used for further analysis has a size of 5000 x 11,688. Several aspects of the data have been analyzed in the publications of the GTEx consortium (see, e.g, Carithers *and others*, 2015) or in conjunction with specialized methods development (see, e.g., Dey *and others*, 2017). The question we address is "Which genes are associated to a certain tissue?", which corresponds to the search for marker genes for a tissue.

5.2 Identification of tissue marker genes using Association Plots

We first conducted correspondence analysis and projected the GTEx data into a three-dimensional subspace. The three dimensions of this projection explain together only ca. 24% of the inertia in this data set. The plot is clearly organized around the three directions for the tissues: pancreas, blood, and pituitary gland (Fig. 5.1a). Genes that are located in one of these three directions are known to be specifically expressed in the tissue represented by the given direction. Alternatively, to delineate all genes specific for the given tissue one can generate an Association Plot. In Fig. 5.1b we present the Association Plot computed for the cluster of pancreas samples. Pancreas is the tissue which is clearly separated from other tissues in the three-dimensional projection of the data, and thus, the Association Plot for pancreas (Fig. 5.1b) contains no surprise: Many genes point to the right, in the direction of the pancreas centroid. We provide a zoom into the right tail of the plot. This clearly shows a set of pancreas-specific genes, of which we colored in red the known pancreas marker genes as determined by the Human Protein Atlas (Uhlén *and others*, 2015).

The real challenge, however, lies in the invisibility of the remaining 27 tissues in the three-dimensional projection. These tissues cannot be distinguished from each other since they form a dense cloud around the origin of the coordinate system. Consequently, it is also impossible in the three-dimensional projection to identify

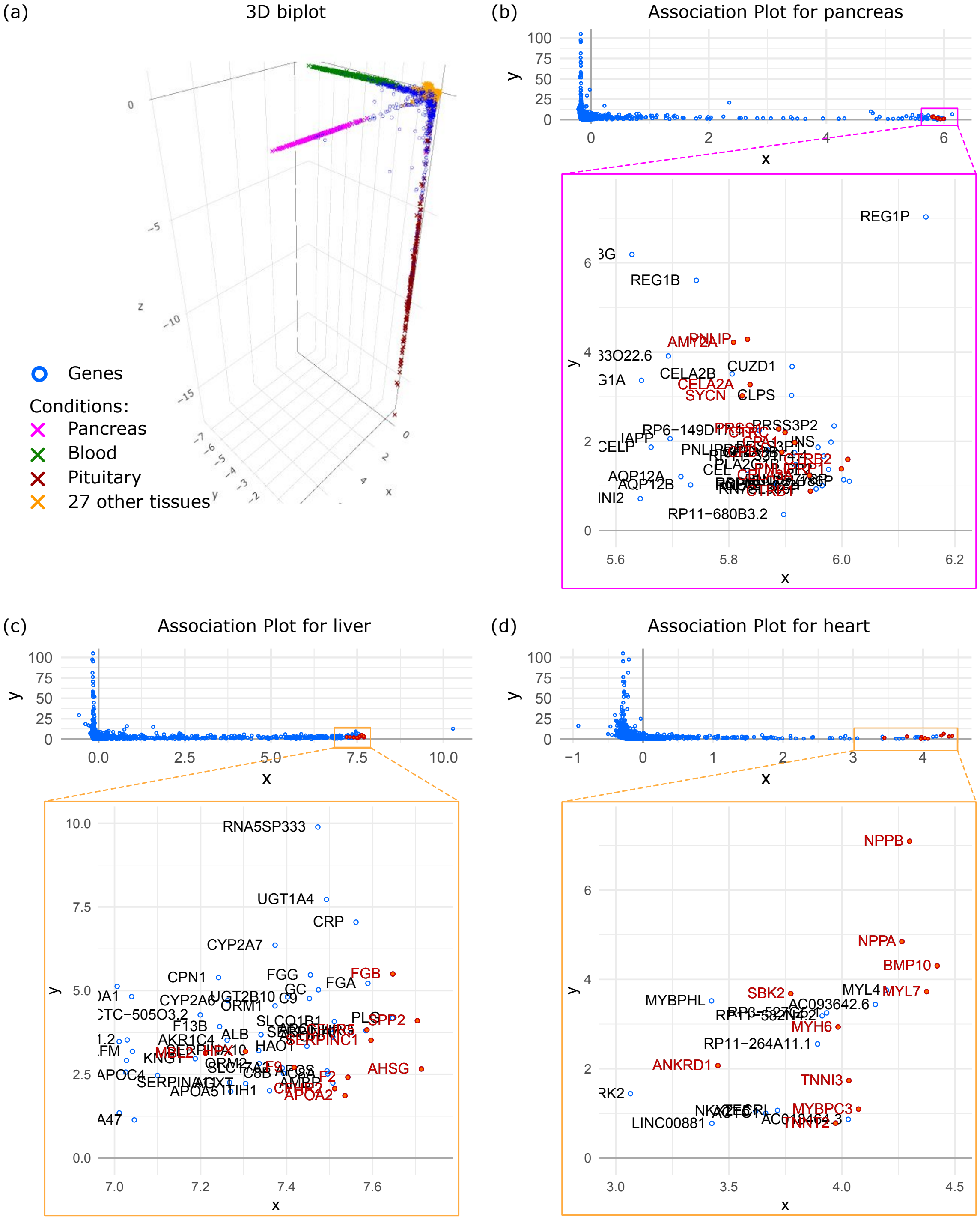


Figure 5.1: Applying CA to GTEx data. **(a)**, Classical CA projection into a three-dimensional subspace, which allows discerning three different tissues (pancreas, blood, pituitary). Other tissues remain lumped in the centre of the observed structure. Association Plots enable obtaining further tissue-specific information. **b-d,** The Association Plots can be used for delineating pancreas- **(b)**, liver- **(c)** and heart-specific **(d)** genes. Such genes are located in the right bottom part of the plot. Red color indicates marker genes from Human Protein Atlas for a given tissue. The presented Association Plots serve as examples. Each tissue can be inspected separately and respective marker genes can be visualized. Each Association Plot is shown above a square plot which contains zoom into the right tail of the Association Plot.

marker genes for these tissues. Yet, any of these “invisible” tissues can be analyzed using an Association Plot to extract further tissue-specific information. Fig. 5.1c depicts the Association Plot generated for the liver samples. Again, a zoom into the right tail of the plot shows numerous genes, a subset of which are known as liver-specific marker genes as given by the Human Protein Atlas (Uhlén *and others*, 2015). One can in principle go through the tissues in this manner and for each of them generate one Association Plot which will highlight genes located in the direction of the given tissue cluster in the CA space, and thus, whose expression is characteristic for this tissue. As another representative example heart-specific genes are shown in the Association Plot generated for the heart cluster (Fig. 5.1d).

When studying any Association Plot, one also faces the question which points to call associated to the cluster. The score S_α allows to further distinguish among genes of equal $x(r)$ but at different distance to the x-axis. Fig. 5.2a shows the heart-specific genes colored by the S_α value. The angle α is taken from the simulated data, which are exactly the ones shown in Fig. 4.8. The ranking of genes provided by the color code is a rather intuitive ranking. Some examples of the actual distribution of gene expression values over the tissues are shown in Fig. 5.2b and provide an intuition for the associations depicted in these ranked Association Plots.

5.3 Influence of dimension number on Association Plots

The Association Plots generated for the GTEx data, which were presented in the previous section, were computed based on the first 96 dimensions of the CA-space as determined by the elbow rule (see Fig. 4.7 for the elbow plot). However, visual inspection shows that Association Plots are fairly robust with respect to the precise choice of dimension for the computation.

To illustrate this, Fig. 5.3 presents Association Plots generated for three example tissues from the GTEx data: pancreas, liver, and heart. For each tissue three Association Plots were generated, using 37, 96, or 225 dimensions kept from the CA. These numbers were obtained using one of the three approaches for selecting a

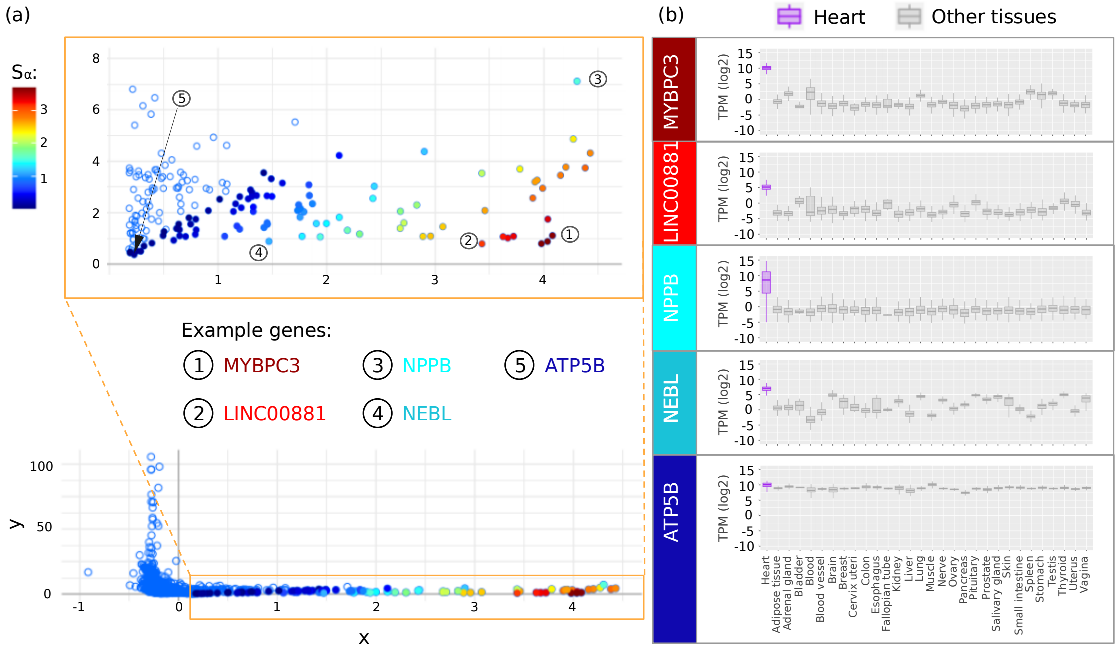


Figure 5.2: Heart-specific genes colored by their S_α values. **(a)** Association Plot was generated for heart samples from GTEx data using 96 CA dimensions. The color of a gene (circle) refers to its S_α value. The higher the S_α value, the higher the specificity of the given gene for heart samples. The Association Plot is shown below a square plot which contains zoom into the right tail of the Association Plot. **(b)** Boxplots illustrating the expression of five example genes (MYBPC3, LINC00881, NPPB, NEBL, and ATP5B) across 30 tissues. The boxplots generated for genes with high S_α values clearly show their upregulation in the heart tissue in comparison to other tissues. When comparing the expression profiles of the five highlighted genes, MYBPC3 shows the highest upregulation in the heart tissue, whereas ATP5B the lowest.

number of dimensions presented in Section 4.4: 80% rule, elbow rule, and average rule, respectively.

Inspection of the cluster-specific gene sets shows large overlaps across these choices. To demonstrate this, for each tissue 12 genes with the highest levels of enriched expressions in a given tissue were obtained from the Human Protein Atlas (Uhlén *and others*, 2015) and highlighted in the corresponding Association Plots. Importantly, independent of the dimension number used for the analysis all these genes are located in the right bottom part of the Association Plots among tissue-specific genes, which demonstrates similarities between the results obtained using different number of CA dimensions.

5.4 Investigating cluster similarities using Association Plots

Until this point we have only focused on the associations between genes and a cluster of samples revealed by Association Plots. However, Association Plots can also provide information on the similarities between clusters. Although in the GTEx data we can distinguish 30 clusters representing 30 distinct tissues, between some of these clusters we expect to find gene expression similarities. For instance, due to their biological identities, tissue pairs such as nerve and brain, or muscle and heart are expected to share the characteristic expression of a higher amount of genes than with the remaining tissues.

The similarity between muscle and heart is visible in the Association Plot for muscle (Fig. 5.4a). There, many heart samples (red crosses) point towards the muscle samples (orange crosses), which indicates an overlap between the gene expression profiles of these two tissues. The remaining 28 tissues are located in the left part of the plot, and are well separated. We also highlighted the lung samples (green crosses) which are not expected to show any similarity to muscle. Indeed, lung samples are isolated from the muscle samples and are located in the left part of the plot. Similarly, nerve and brain share many genes, and in the Association Plot for nerve samples (Fig. 5.4b, light pink crosses) the brain samples (dark pink

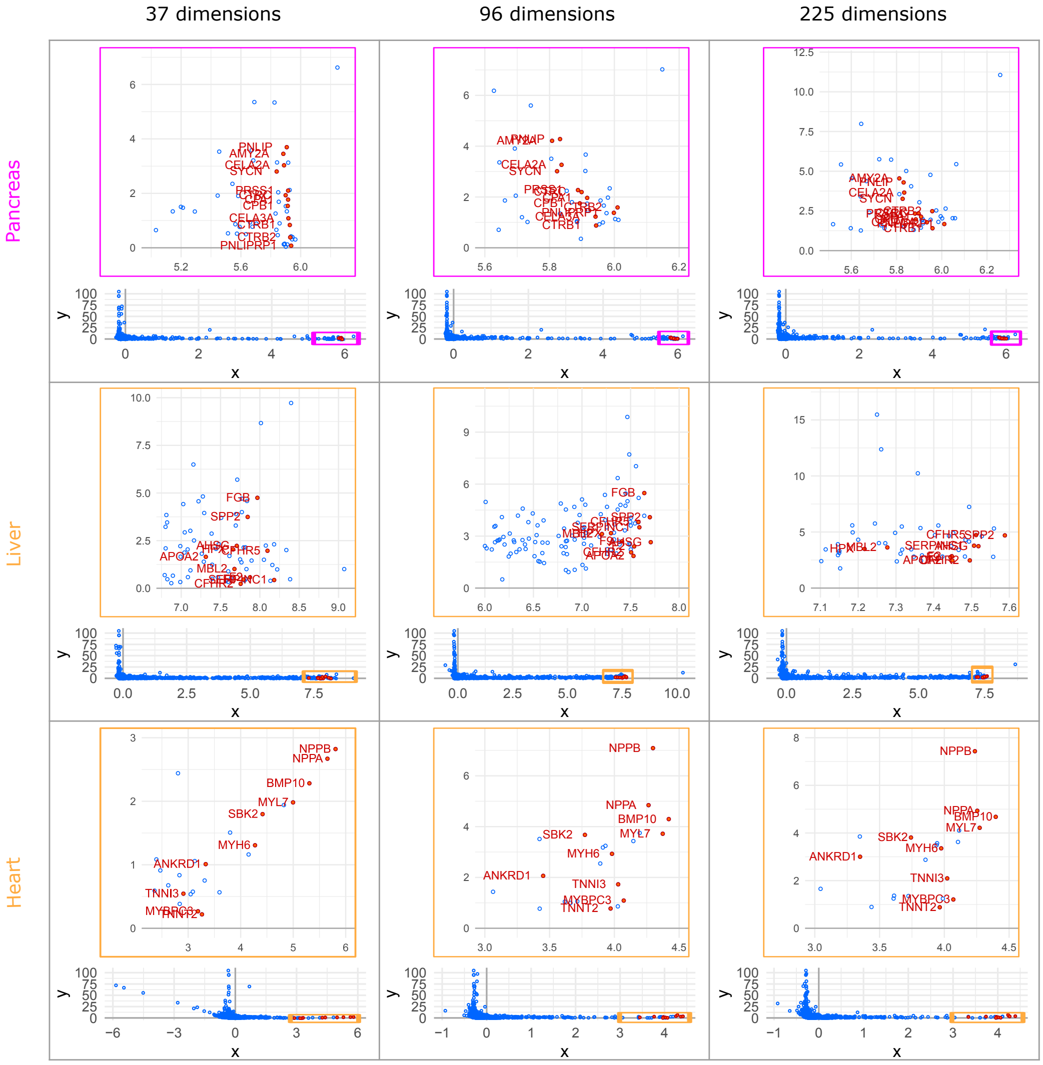
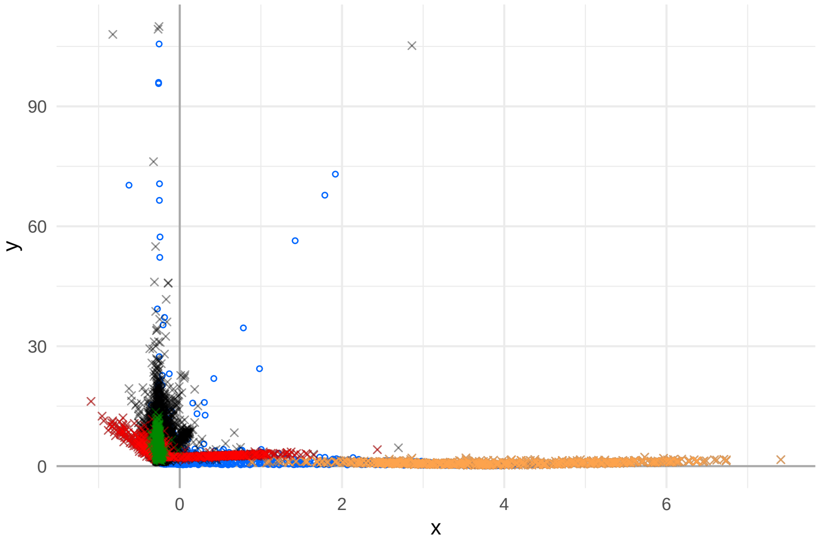
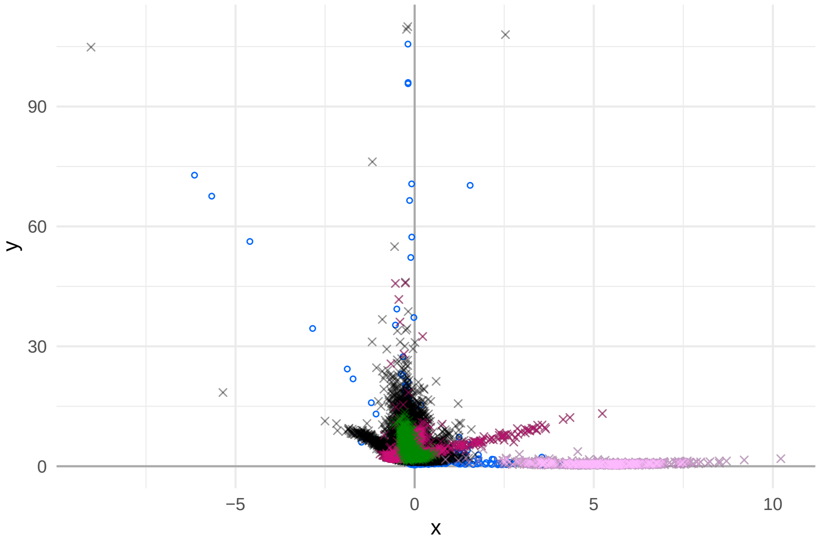


Figure 5.3: Comparison of Association Plots generated for three tissues (pancreas, liver, heart) using three different numbers of dimensions. The numbers of dimensions (37, 96 and 225) were calculated using three different approaches (keeping the minimum number of dimensions that account for more than 80% of inertia in total; elbow rule (Ciampi *and others*, 2005); keeping only those dimensions that explain more inertia than one dimension on average). For each tissue the cluster-specific gene sets obtained using the three different approaches show large overlaps. Genes with red labels – genes with the highest enriched expression levels in a given tissue obtained from the Human Protein Atlas (Uhlén *and others*, 2015). Each time the Association Plot is shown below a square plot which contains the zoom into the right tale of the Association Plot.

crosses) also point towards the right. This visualizes that in CA space there is a small angle between the brain samples and the nerve samples, and thus they are closer to each other than to the remaining samples.



(a) Association Plot for muscle samples. (Orange: muscle samples, red: heart samples, green: lung samples.)



(b) Association Plot for nerve samples. (Light pink: nerve samples, dark pink: brain samples, green: lung samples.)

Figure 5.4: Information on the cluster similarities revealed by Association Plots. (a) Association Plot for muscle reveals similarities between heart and muscle samples. Colorful crosses represent highlighted samples: muscle - orange, heart - red, lung - green, samples from the remaining 27 tissues - black. **(b)** Association Plot for nerve reveals similarity between brain and nerve samples. Colorful crosses represent highlighted samples: nerve - light pink, brain - dark pink, lung - green, samples from the remaining 27 tissues - black. The presented Association Plots were generated based on GTEx data using 37 dimensions. Blue dots represent genes.

CHAPTER 6

Association Plots Applied To Single-Cell RNA-Seq Data

A remarkable amount of transcriptomics data produced nowadays is data obtained using the single-cell RNA-sequencing. In this chapter we present application of Association Plots to single-cell RNA-seq data on two example data sets. The content of this chapter is an adapted and extended version of Gralinska *and others* (2022).

6.1 3k PBMC data

6.1.1 Introduction

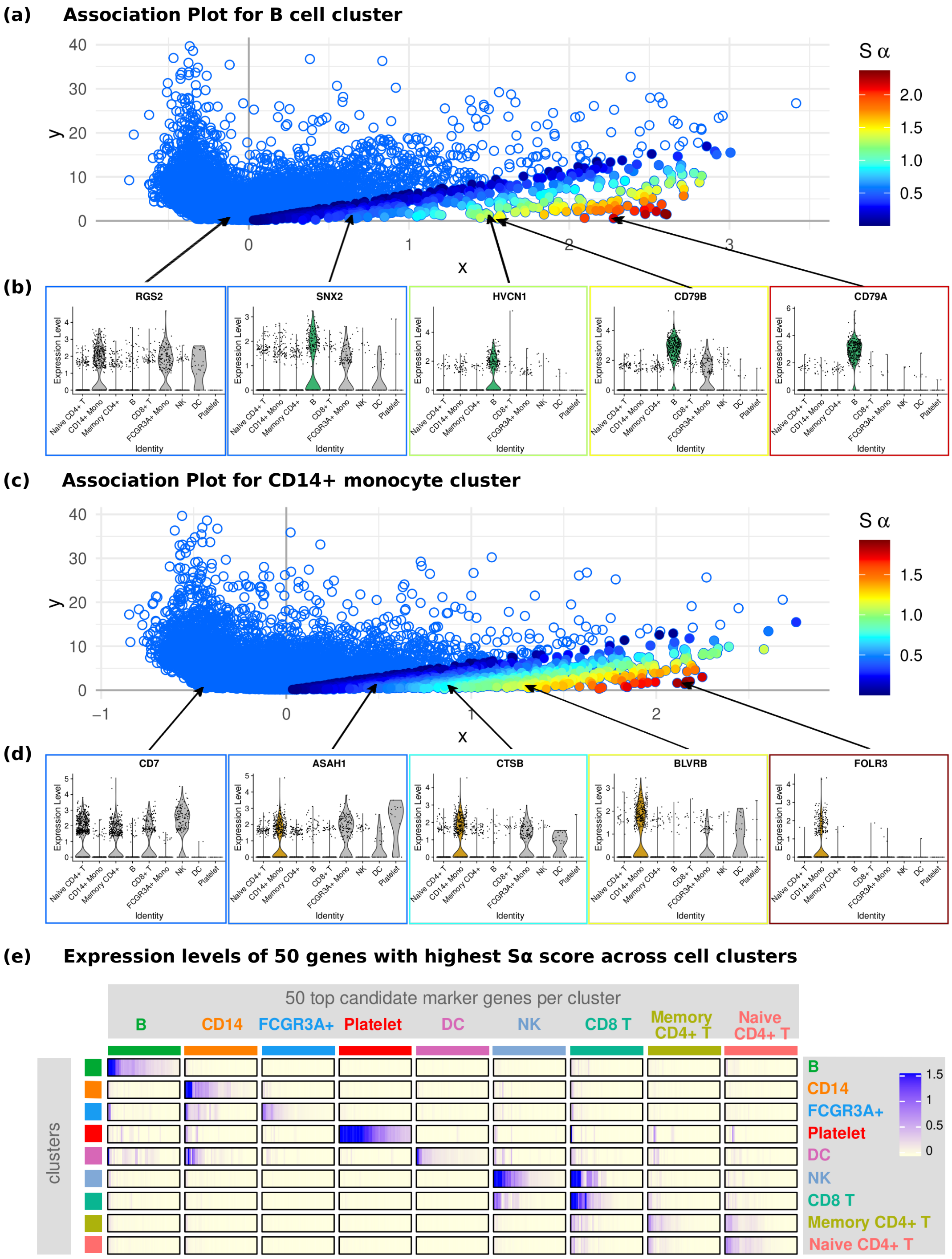
To demonstrate how to use Association Plots for studying gene-cluster associations and for identification of novel marker genes characterizing a cell cluster from single-cell data, we applied our method to the 3k Peripheral Blood Mononuclear Cell (PBMC) single-cell RNA-seq data (Zheng *and others*, 2017; 10x Genomics, 2016). PBMC data was generated by 10X Genomics and allows for studying the immune populations within PBMCs from a healthy donor.

When applying Association Plots to a single-cell transcriptomics data, clustering of cells will typically be part of data pre-processing. For this we followed the vignette from the Seurat package (Hao *and others*, 2020) (more details in Appendix C). Using Seurat the clusters were annotated to different cell types based on the expression of canonical marker genes for immune cells (Fig. A.1 in Appendix A) (Hao *and others*, 2020). This annotation allows us to address clusters by their cell type rather than by a number. The following cell types were identified: B cells, naive CD4+ T cells, memory CD4+ T cells, CD8+ T cells, FCGR3A+ monocytes, natural killer (NK) cells, CD14+ monocytes, dendritic cells (DC), and platelets.

6.1.2 Identification of cluster-specific genes using Association Plots

To identify cluster-specific genes from 3k PBMC data we generated Association Plots for each of the cell clusters. Fig. 6.1 presents two example Association Plots for the B cell- (Fig. 6.1a) and CD14+ monocyte clusters (Fig. 6.1c). Genes with the positive S_α score are highlighted in color according to the color map given. To illustrate the linkage between S_α score and gene expression patterns across cell clusters we focus on 10 example genes with different S_α values. In Fig. 6.1b we present the expression levels of five random genes from the B-cell cluster Association Plot across nine clusters for comparison: RGS2, SNX2, HVCN1, CD79B, and CD79A. As shown in the violin plots, with the increasing S_α score the over-expression of a given gene in the B cell cluster gets more pronounced. For instance, while in the case of CD79A, a gene with the highest S_α score equal 2.18, we observe a clear over-expression in the B cell cluster, in the case of SNX2, a gene with the S_α score of 0.51, the over-expression signal in the B cell cluster is almost not visible. On the other hand, for RGS2, a gene with a negative S_α score located in the left bottom part of the Association Plot, we observe the over-expression in three other cell clusters and not in the B cell cluster.

Fig. 6.1d demonstrates analogous observations as in Fig. 6.1b, this time for five genes from the CD14+ monocyte Association Plot: CD7, ASAH1, CTSB, BLVRB, and FOLR3. For CD7, a gene located in the bottom left part of the



(a) Association Plot for B cell cluster

(b)

(c) Association Plot for CD14+ monocyte cluster

(d)

(e) Expression levels of 50 genes with highest Sα score across cell clusters

Figure 6.1: Cluster-specific genes from 3k PBMC data. (**a**) Association Plot generated for the B cell cluster. Each circle represents one gene from the input data. Genes with the positive S_α score are highlighted in color according to the color map given. (**b**) Comparison of the expression levels of five example genes from (a) across nine cell types. (**c**) Association Plot generated for the CD14+ monocyte cluster. (**d**) Comparison of the expression levels of five example genes from (c) across nine cell types. (**e**) Average expression levels of 50 candidate marker genes per cell type, identified using the Association Plots, across all nine cell types.

Association Plot, no over-expression signal in the CD14+ monocyte cluster is observed. Instead, it is over-expressed in four other cell clusters: naive CD4+ T, memory CD+, CD8+ T, and NK cells. On the other hand, a gene with the highest S_α value of 1.93, FOLR3, shows a strong over-expression in the CD14+ monocyte cluster. This, together with the three remaining plots generated for ASAH1, CTSB and BLVRB confirms the link between S_α score and gene expression patterns across cell clusters.

Next, we employ Gene Set Enrichment Analysis (GSEA) (Subramanian *and others*, 2005; Mootha *and others*, 2003; Liberzon *and others*, 2015) to show that genes with high S_α values are associated to the cell type for which the respective Association Plot was computed. GSEA was performed on the 100 genes with the highest S_α values. The results for two cell clusters, B cell- and CD14+ monocytes cluster, are presented in Appendix B in Table B.1 and Table B.2, respectively. For the B cell cluster nine out of the top 10 enriched gene sets are linked to the B cell population. For the CD14+ monocyte cluster four out of the top 10 enriched gene sets are directly related to the monocyte population. Two further gene sets are related to the myeloid cell population, which reflects the monocyte-specificity of the gene markers from the CD14+ monocyte cluster. Thus, the genes associated to a cell type based on high S_α are indeed characteristic for the respective cell types.

Fig. 6.1e provides an overview of uniqueness vs sharing of cluster-specific genes for the cell-type clusters in the PBMC data. From each of the nine Association Plots we extracted the 50 genes with the highest S_α score. Each of the nine rectangles on the main diagonal of the matrix represents those 50 genes from the respective Association Plot each as a little heatmap: The genes' within-cluster average expression strength is encoded in color, with genes sorted by expression from left to right. The rectangles in the same column contain the genes from the main, diagonal rectangle in the same order, and with the color commensurate to the average expression level in the other cluster.

While marker genes obtained from well-separated clusters such as B cells, platelet or dendritic cells show a strong over-expression in only one cluster, the identified marker genes from clusters located in close proximity to another cluster in correspondence analysis space are also partially up-regulated in the neighboring clusters. This is observed for natural killer cell- and CD8+ T cell cluster, as well as for the CD8+ T cell cluster, the memory CD4+ T cell cluster, and the naive CD4+ T cell cluster. Cell clusters located in close proximity to each other share similar gene expression profiles, which results in a low number of genes characteristic for only one of these clusters.

6.1.3 Extraction of novel marker genes

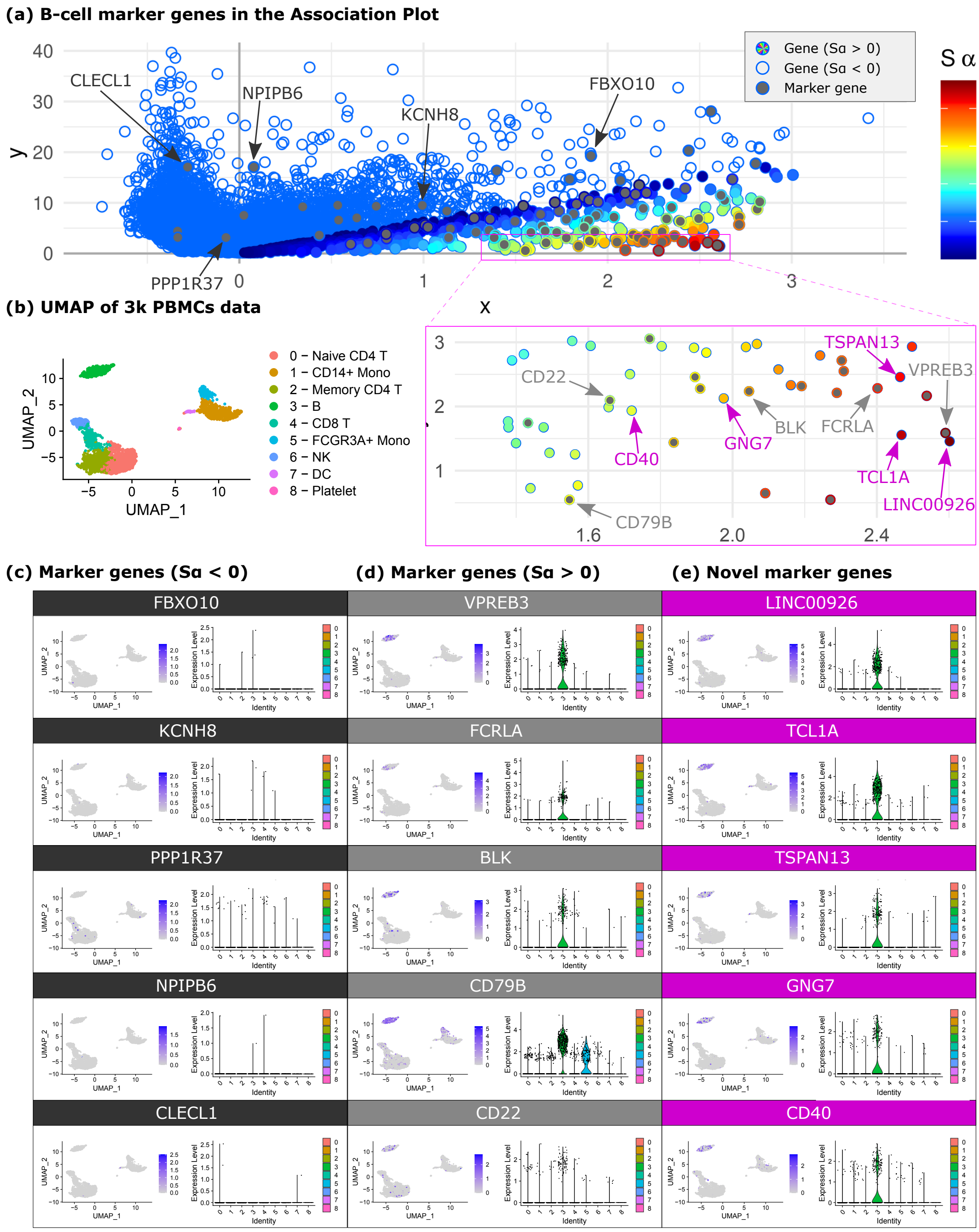
When a set of established marker genes for a cell type is given, an Association Plot for a cell cluster corresponding to that cell type may serve to support the identification of novel marker genes. This is an important task in the context of less well characterized cell types. We demonstrate how to proceed on the example of the Association Plot generated for the B cell cluster from the 3k PBMC data.

Fig. 6.2a presents the Association Plot for the B cell cluster, with 242 B-cell enriched genes highlighted using grey filling. This set of genes was obtained from the Human Blood Atlas (Uhlén *and others*, 2015), a collection of information on the human protein-coding genes across distinct human blood cell types. In the Human Blood Atlas all genes with at least four times higher normalized expression values in B cells than in any other cell type are qualified as B-cell enriched genes (Uhlen *and others*, 2019). In the Association Plot these genes obtained statistically higher S_α scores than the remaining genes (Wilcoxon test, p-value 1.505e-35). This is also visible in the Association Plot with the majority of the B-cell enriched genes located within the area of positive S_α values, which confirms a substantial overlap between the Association Plot results and the marker gene set from the Human Blood Atlas. In addition to this, single marker genes are located outside of the positive S_α area. This might be caused by the differences in data sets used for computing Association Plots and for extracting cell type enriched genes in the Human Blood Atlas.

Novel marker genes will be located among genes with high S_α scores, and which at the same time are not annotated yet as marker genes for a given cell type. In the Association Plot generated for the B cell cluster we highlighted five example genes which can be considered as marker gene candidates: LINC00926, TCL1A, TSPAN13, GNG7, and CD40 (Fig. 6.2a). As presented in Fig. 6.2e, each of these genes is over-expressed in the B cell cluster. For comparison, in the generated Association Plot we also highlighted 10 further example genes listed among the B-cell enriched genes in the Human Blood Atlas. As expected, the first five genes (FBXO10, KCNH8, PPP1R37, NPIPB6, CLECL1) located outside of the positive S_α area do not show any over-expression signal in the B cell cluster (Fig. 6.2c), while the further five genes (VPREB3, FCRLA, BLK, CD79B, CD22) are characterized by positive S_α values and show over-expression in the B cell cluster (Fig. 6.2d). Moreover, their expression profiles resemble the profiles of the newly detected candidate marker genes.

The first detected gene is LINC00926, a long non-coding RNA over-represented in the B cell cluster. Even though LINC00926 has not been well-characterized yet, its abnormal expression was observed in several cancer types (Wang *and others*, 2021). For instance, the up-regulation of LINC00926 in B cells in lung adenocarcinoma patients was observed to improve their overall survival (Li *and others*, 2021). Due to this, LINC00926 was suggested to be a B-cell specific marker gene protecting against lung adenocarcinoma. Moreover, it was also observed to suppress breast cancer growth by down-regulating the expression of phosphoglycerate kinase 1 (PGK1) (Chu *and others*, 2021). LINC00926 was also described in the context of acute myeloid leukemia (Wang *and others*, 2018) and Hodgkin lymphoma (Liang *and others*, 2020). To our knowledge, there is only one publication available which describes LINC00926 as a B cell marker (Sellers *and others*, 2021), together with another gene, TCL1A, the second marker gene candidate identified using the Association Plot.

TCL1A, T-cell leukemia/lymphoma protein 1A, is a gene involved in the regulation and differentiation of B cells. Over-expression of this gene is linked to the T- and B-cell lymphomas (Brinas *and others*, 2021; Aggarwal *and others*, 2009). Although TCL1A is not present in the B-cell enriched gene list from the Human Blood Atlas, it is classified there as a “cell lineage group enriched gene”,



(a) B-cell marker genes in the Association Plot

(b) UMAP of 3k PBMCs data

(c) Marker genes (Sα < 0)

(d) Marker genes (Sα > 0)

(e) Novel marker genes

Figure 6.2: Identification of novel marker genes using Association Plots. (**a**) Association Plot for B cell cluster from the 3k PBMC data. Genes known to be enriched in B cells according to the Human Blood Atlas are highlighted in grey. (**b**) Cell clusters in 3k PBMC data. (**c-d**) Expression levels of example genes enriched in B cells according to the Human Blood Atlas with (c) negative and (d) positive S_α score. (**e**) Expression levels of B-cell specific genes detected using the Association Plot, which are not listed among B-cell enriched genes according to the Human Blood Atlas.

and thus, its up-regulation is simultaneously observed in B cells and plasmacytoid DCs. However, in the 3k PBMC data TCL1A is up-regulated only in some cells from the DC cluster and, thus, it still scores high in the Association Plot for B cells.

Further example genes from the Association Plot showing B cell cluster specificity are GNG7, CD40, or TSPAN13. Similar to TCL1A, GNG7 is classified as a “cell lineage group enriched gene” in the Human Blood Atlas and its up-regulation is observed both in B cells and plasmacytoid DCs. However, according to the 3k PBMC data this gene is partially up-regulated only in the B cells. The fourth detected gene, CD40, according to the Human Blood Atlas is only enhanced in the B cells, while in the 3k PBMC data it is visibly over-expressed in the B cell cluster. TSPAN13, the last gene identified from the Association Plot, is classified as enhanced both in the naive B cells and plasmacytoid DCs according to the Human Blood Atlas, while in the 3k PBMC data it is up-regulated only in the B cell cluster.

6.2 Human cell atlas of fetal gene expression

6.2.1 Introduction

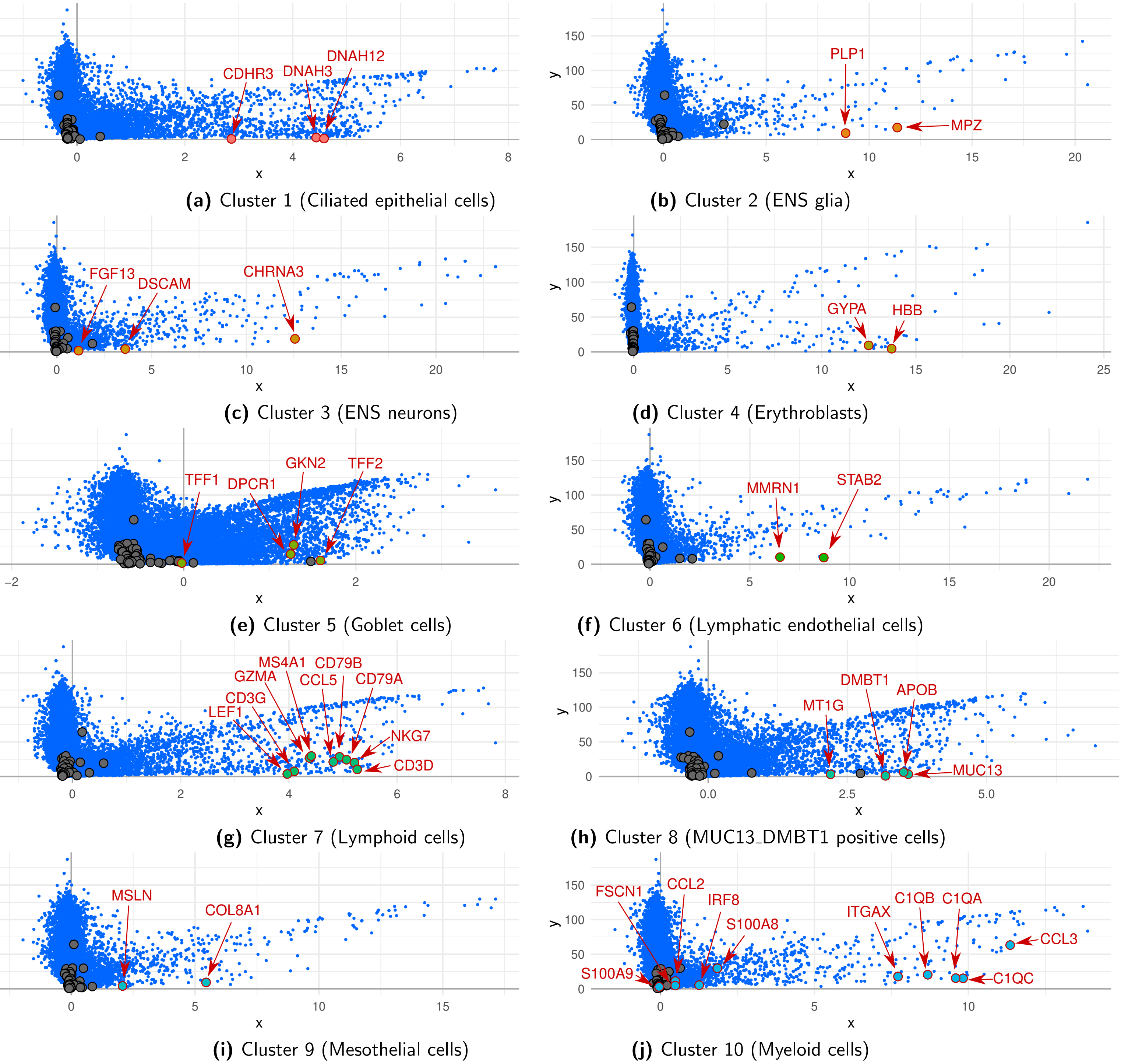
Another typical task in single-cell data analysis is the annotation of clusters of cells to known cell types. We proceed to demonstrate this on the example of stomach single-cell data from the human cell atlas of fetal gene expression (Cao *and others*, 2020). The human cell atlas of fetal gene expression is a reference atlas generated by profiling almost four million cells across 15 human organs. In our analysis we focus on the cells from the stomach organ, which cluster into 16 stomach cell types. More information on the data extraction can be found in Appendix C.

6.2.2 Cluster annotation using Association Plots and within-tissue marker genes

Among the existing methods for cluster annotation one can distinguish two main types. The first group encompasses methods that rely on a reference database. In this case expression profiles of cells from a given cluster are compared against expression profiles of various cell types from a reference database. Alternatively, cluster annotation can be conducted using a literature-derived list of marker genes for various cell types, where the expression analysis of such markers allows then for matching a given cell cluster to a cell type. We proceed according to this second paradigm and work with given lists of marker genes for different cell types from the stomach samples, obtained from the human cell atlas of fetal gene expression.

The single-cell data from the stomach consists of 16 subclusters according to the subcluster analysis conducted by the authors of the original study (see Methods Section ”Subclustering analysis” in Cao *and others* (2020)). For each subcluster, we generated its Association Plots, yielding 16 Association Plots depicted in Fig. 6.3a-6.3p. To annotate the generated Association Plots to 16 cell types we use a set of within-tissue marker genes from stomach provided by the authors of the original study. Altogether, among the sets of within-tissue marker genes reported by them, there were 64 marker genes with subsets characteristic of individual cell types from stomach. Thus, in each generated plot we highlighted the complete set of 64 marker genes for all 16 stomach cell types, leading to the images of Fig. 6.3a-6.3p.

In each of the generated plots the majority of the highlighted genes are located on the left side of the plot, which indicates no association between them and the depicted cell subcluster. However, in each plot a few genes are located on the right hand side. These are the marker genes for that respective subcluster. Therefore, in the last step of the analysis we focused only on these genes and used them to match each Association Plot to one of the 16 stomach cell types based on the provided list of within-tissue marker genes. This allows to easily assign the identity of all 16 stomach clusters from the data as shown in a UMAP (Fig. 6.4). To illustrate the results, in Fig. 6.3a-6.3p we show all the 16 Association Plots together with their



(a) Cluster 1 (Ciliated epithelial cells)

(b) Cluster 2 (ENS glia)

(c) Cluster 3 (ENS neurons)

(d) Cluster 4 (Erythroblasts)

(e) Cluster 5 (Goblet cells)

(f) Cluster 6 (Lymphatic endothelial cells)

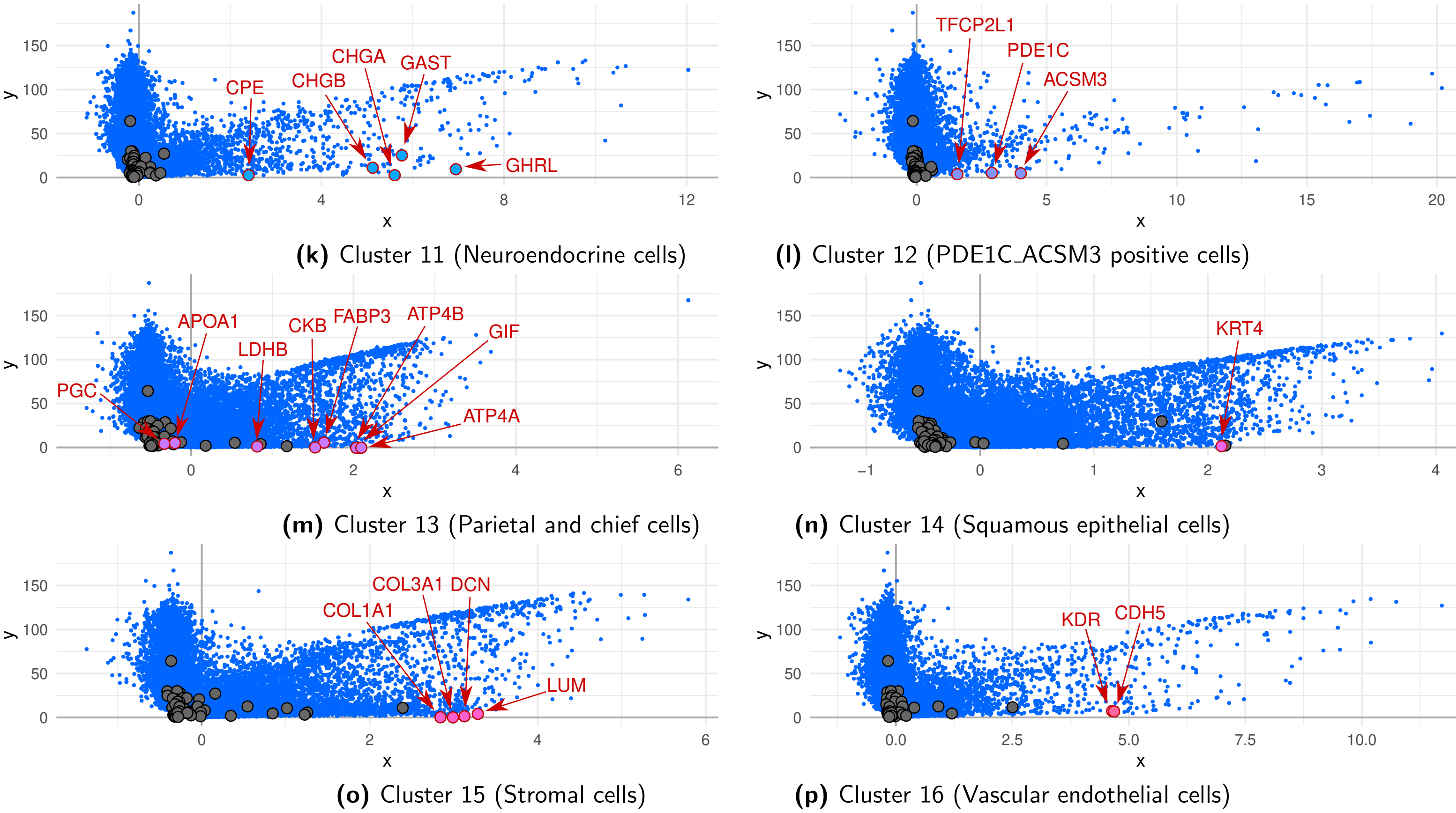
(g) Cluster 7 (Lymphoid cells)

(h) Cluster 8 (MUC13_DMBT1 positive cells)

(i) Cluster 9 (Mesothelial cells)

(j) Cluster 10 (Myeloid cells)

Figure 6.3: (continuation on the next page)



(k) Cluster 11 (Neuroendocrine cells)

(l) Cluster 12 (PDE1C_ACSM3 positive cells)

(m) Cluster 13 (Parietal and chief cells)

(n) Cluster 14 (Squamous epithelial cells)

(o) Cluster 15 (Stromal cells)

(p) Cluster 16 (Vascular endothelial cells)

Figure 6.3: Annotation of cell clusters from the human cell atlas of fetal gene expression using known marker genes. (**a**-**p**) Association Plots for each of the 16 stomach subclusters with a set of 64 within-tissue marker genes. Marker genes specific for a given cell type are highlighted in color. The colors correspond to the UMAP color scheme (Fig. 6.4).

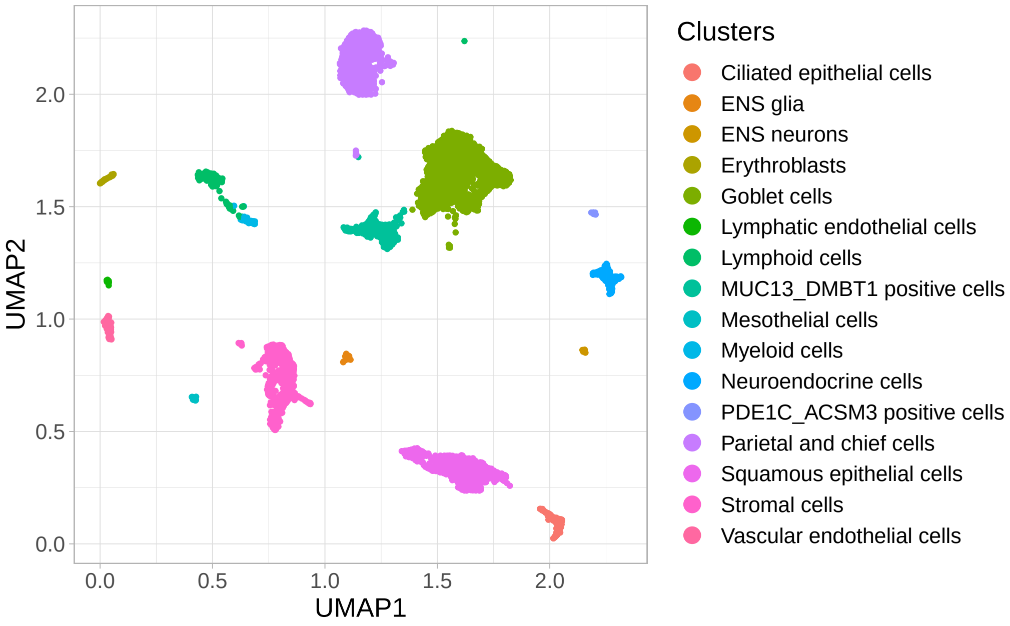


Figure 6.4: UMAP visualization of stomach cells from human cell atlas of fetal gene expression. The cells were annotated to cell types using Association Plots from Fig. 6.3 and the provided list of within-tissue marker genes for each of the 16 cell types. The plot was generated using the UMAP coordinates of stomach cells obtained from the processed data from the original publication Cao *and others* (2020).

real (as given by the original authors) cell type, and the corresponding marker genes for a cell type highlighted in color.

We have used this example to demonstrate how easy it is to obtain cell-type assignments based on Association Plots. There is no need to search for the right set of marker genes for a cell cluster, but mapping the union of all marker genes into the Association Plots yields an easy to interpret visualization from which the identity of the cluster can be inferred. Our results were in agreement with the cluster information from the original data, which demonstrates that Association Plots can be applied for annotating the cell clusters to known cell types based on the predefined list of marker genes.

6.3 Comparison to differential expression testing tools

As a tool for visualizing genes that are characteristic for a cell cluster, Association Plots can also be seen as a way of determining genes that are differentially expressed between the cells in the cluster vs all other cells. Thus, we need to answer the question how comparable are the results obtained using Association Plots to the results from commonly used differential expression testing tools.

To address this question we need to choose a set of differential expression (DE) tools for single-cell RNA-seq data to include in a comparison. There is no consensus in the community on which of the existing differential expression testing methods is the best one for single-cell RNA-seq data. Comparative studies of various existing tools revealed an unsatisfying agreement among them (Van den Berge *and others*, 2018; Wang *and others*, 2019*a*). Even though there also exist tools developed especially for single-cell RNA-seq data, it was recently suggested that the standard tools for bulk RNA-seq data do not perform worse than the specialized single-cell RNA-seq tools (Van den Berge *and others*, 2018). Therefore, we decided to follow recent recommendations (Luecken and Theis, 2019) and focused on two differential expression testing tools for bulk RNA-seq data: DESeq2 (Love *and others*, 2014) and edgeR (Robinson *and others*, 2010; McCarthy *and others*, 2012), combined with ZINB-WaVE weight estimation method (Risso *and others*, 2018). In addition to those we include in the comparison the *FindAllMarkers* function from Seurat (Hao *and others*, 2020), which is specifically designed for delineating marker genes from single-cell data.

To investigate the agreement among results obtained with DESeq2, edgeR, Seurat, and Association Plots, we applied them to the 3k PBMC data. First of all, sets of 1000 most up-regulated or cell-type specific genes were extracted for each cluster and tool (see Appendix C for details), and the overlaps between them were investigated. For Association Plots the top 1000 genes refer to the ranking by S_α score. Fig. 6.5 shows the overlap between the results of the four approaches for all nine cell types of the 3k PBMC data. For all cell types, results obtained using Association Plots agree the most with Seurat results. The overall lighter color of the matrix for dendritic cells (DC) indicates that in this cell type the different methods agree the least. In DC, Association Plots share 547 out of 1000 genes with Seurat, whereas in natural killer cells (NK) 833 genes are shared between these two methods. DESeq2 and edgeR, in turn agree more with each other than with either Seurat or Association Plots, as can be seen for eight out of nine cell types.

We proceed to demonstrate how mapping the sets of differentially expressed genes into an Association Plot allows to visualize and study the differences or agreements. We focus on the case of the dendritic cells since there the agreement between methods was smallest. Fig. 6.6 shows three times the same Association Plot for the DC cluster, overlaid respectively with the 250 most differentially expressed genes from Seurat, edgeR, and DESeq2. The genes are chosen according to thresholds selected appropriately for the individual method (see Appendix C).

Comparison of the highlighted differential gene sets in the three subfigures shows general agreement with interesting particular differences. For example, Seurat classifies a few genes very far to the right as differential. Those genes are characterized by high expression of the respective gene albeit only in a subset of cells in the cluster. As a consequence, a method like DESeq2 may assign a high fold-change but a non-significant p-value. This is also illustrated in Fig. 6.7

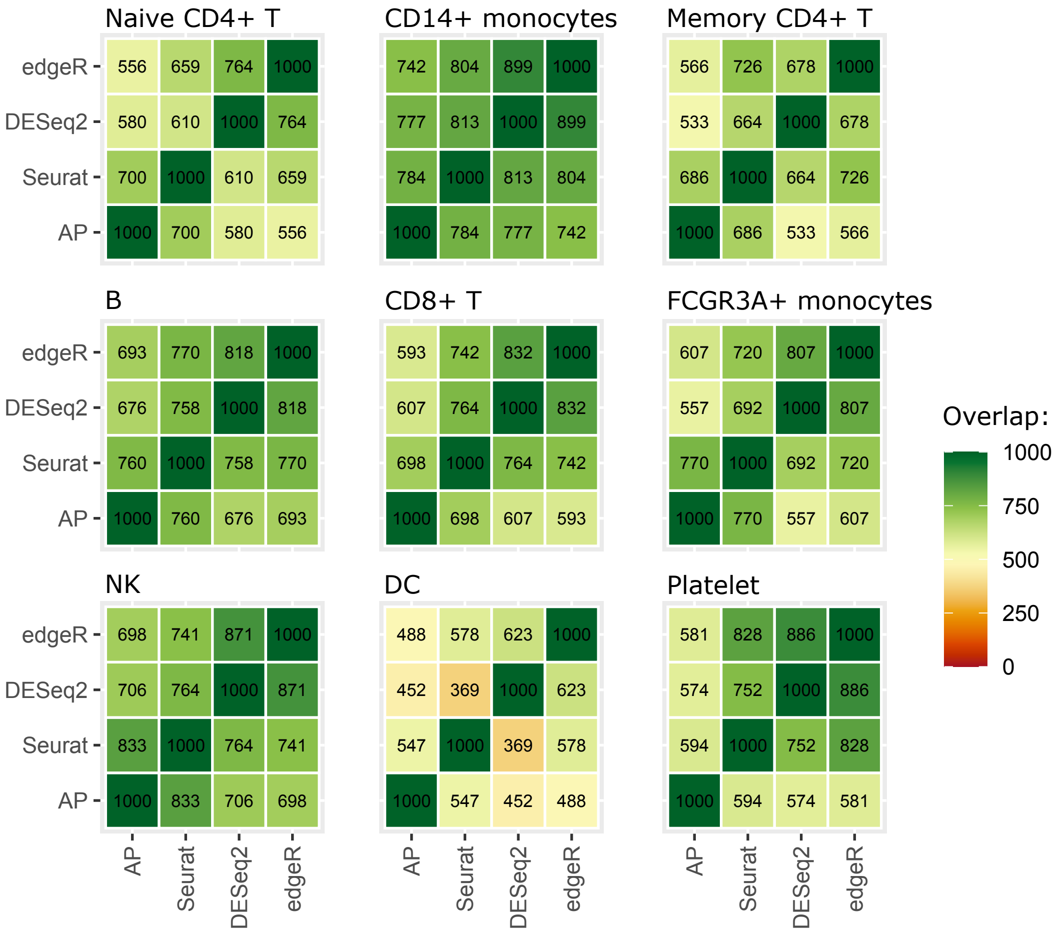


Figure 6.5: Agreement among the results obtained with DESeq2, edgeR, Seurat, and Association Plots for the 3k PBMC data. For each cell cluster and tool 1000 most up-regulated or cell-type specific genes were extracted. The overlaps between them are shown in the heatmaps. AP, Association Plot.

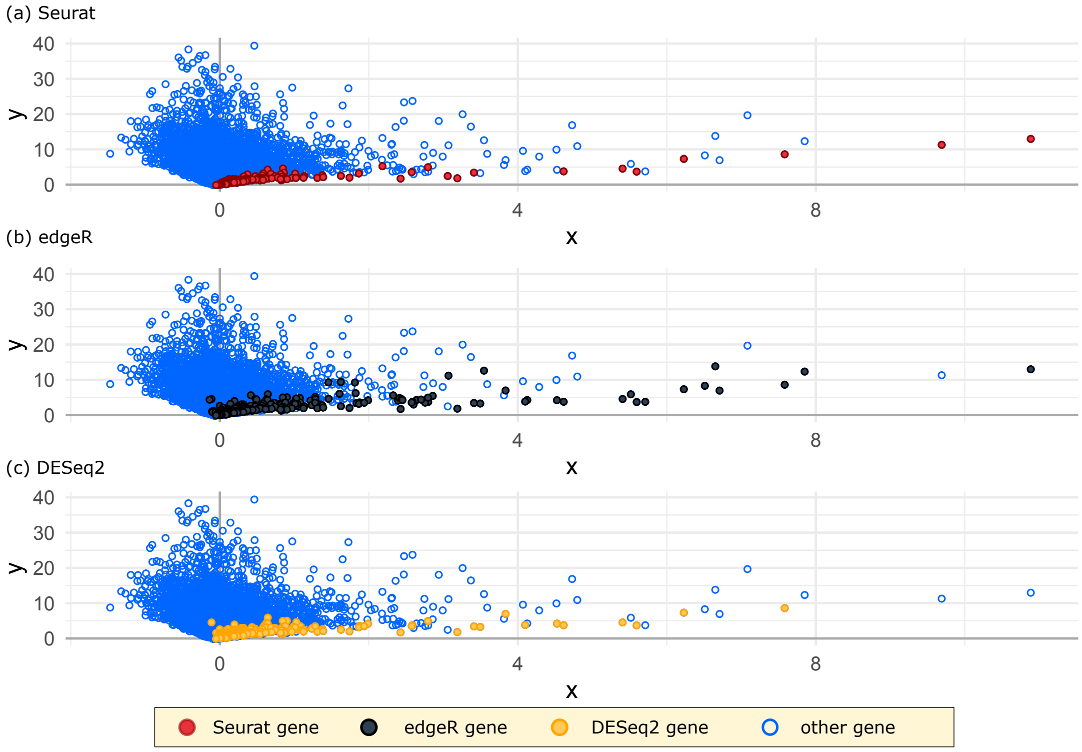


Figure 6.6: Association Plot for the DC cluster from 3k PBMC data, overlaid with 250 most differentially expressed genes from (**a**) Seurat, (**b**) edgeR, and (**c**) DESeq2.

in the dot plot generated using the Seurat package (Hao *and others*, 2020). The dot plot demonstrates the average expression of selected 30 differentially expressed genes across all nine cell clusters as well as the percentage of cells from a cluster in which a given gene is expressed. Per method only five top and five bottom genes out of 250 genes from Fig. 6.6 are shown. Similar to the previous observations, the top differentially expressed genes detected using Seurat show expression signal only in the subset of cells from the DC cluster, while they are not expressed in other clusters. For edgeR and DESeq2 we observe an opposite pattern, and thus, the most differentially expressed genes detected using these two methods are also expressed in other clusters, however their average expression is the highest in the DC cluster. Genes with the expression signal only in a subset of cells from the DC cluster are assigned higher p-values and are not present in the top part of the list.

To summarize, for most of the genes that one would judge as differential from the visual impression provided by the Association Plot, there is at least one of the other, differential expression testing methods that would also identify that gene. Thus, the Association Plot serves well as a summary of the relevant genes to be explored further.

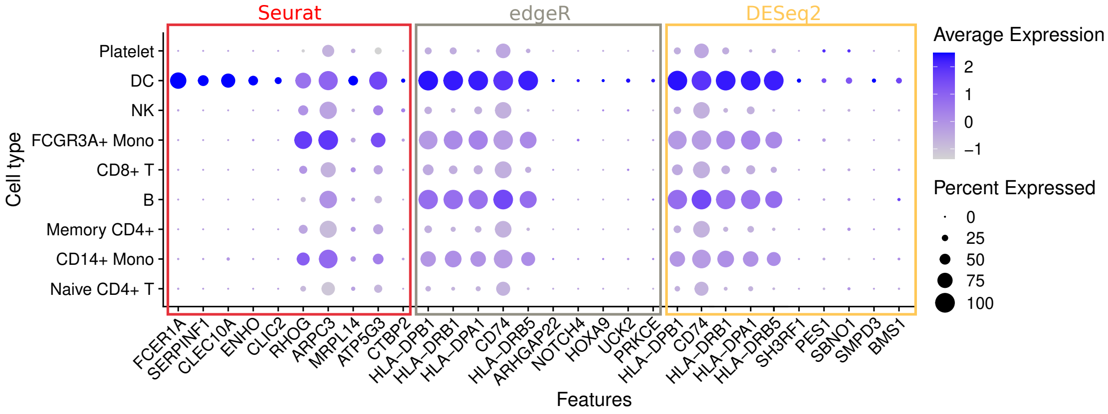


Figure 6.7: Dot plot illustrating the expression of selected 30 genes which are differentially expressed in the DC cluster from 3k PBMC data. This gene set was identified using Seurat, edgeR, and DESeq2. Per method we show five top and five bottom genes out of 250 most differentially expressed genes revealed by this method. The dot size illustrates the percentage of cells in which a gene is expressed, while the color shows the AverageExpression value from Seurat across cells belonging to a given cluster.

CHAPTER 7

Software

This chapter presents the software implementing the concept of Association Plots. First of all, I will introduce a Bioconductor R package *APL* developed together with Clemens Kohl. Subsequently, I will present a Shiny App developed together with Bita Sokhandan Fadakar. The parts of this chapter come from Gralinska *and others* (2022).

7.1 R package

7.1.1 Introduction to the package

APL is a freely available R package for identification of cluster-specific genes using Association Plots. The package was developed in a way that allows for applying it to single-cell transcriptomics data and extracting a list of genes specific for any selected cell cluster.

Further information about the package, installation, usage details and examples can be found below and in the vignette provided with the package.

7.1.2 Download and installation

The R package *APL* is available from Bioconductor at `https://bioconductor.org/packages/release/bioc/html/APL.html`, as well as from the GitHub repository at `https://github.com/VingronLab/APL`. The package requires the R program, which is freely available from CRAN at http://cran.r-project.org. For using *APL* we highly recommend installing *pytorch* (Paszke *and others*, 2019) because it provides a fast implementation of the singular value decomposition. For more details on package installation please refer to the vignette available on Bioconductor or GitHub.

7.1.3 Usage

When working with single-cell transcriptomics data we recommend providing the input data as a *Seurat* or *SingleCellExperiment* object. Alternatively, the input data can be provided in form of a normalized count matrix, with rows representing genes and the columns representing cells. In addition to this, *APL* can also be applied to any type of the data represented in form of a matrix with non-negative entries. To run the analysis, the input data should be specified as the *obj* parameter in the function *cacomp*.

Association Plots are computed using a function *apl_coords*. To run this function the user needs to specify the *group* parameter indicating for which cells from the input data the Association Plot should be computed. Therefore, the user should use the indices or names of cells belonging to a cluster of interest, e.g. according to the clustering information provided beforehand. We additionally implemented a wrapper function *runAPL*, which automates the above-described steps. Finally, to display the computed Association Plot a function *apl* should be called.

S_α scores for ranking genes are computed with the function ***apl_score***. The S_α scores are then stored in the ***APL_score*** attribute of a *ca* object.

To investigate the expression of a gene that was identified as interesting in an Association Plot across the clusters of the single-cell transcriptomics data, external plotting functions such as *VlnPlot*, for generating a violin plot, and *FeaturePlot*, for generating a feature plot, from the *Seurat* package can be used.

By default the computation of Association Plots is done using 5,000 genes with the highest variance across cells. This number can be changed using the parameter *top* in the functions *cacomp* or *runAPL*. The wrapper function *runAPL* uses by default the number of CA dimensions computed using the elbow rule. When using the *cacomp* function the user should specify the number of CA dimensions using the parameter *dims*. We implemented three methods for selecting a dimension number which are also described in Section 4.4: elbow rule (`elbow_rule`), 80% rule (`maj_inertia`), and average rule (`avg_inertia`). The user can also estimate the number of dimension using a scree plot, which can be generated using the function `scree_plot`.

Association Plots are computed based on the geometry of correspondence analysis. Therefore it is possible to plot the two- or three-dimensional input data projection of the correspondence analysis space. This is done by the functions `ca_biplot` and `ca_3Dplot`, respectively.

The *APL* package can be integrated into existing pipelines, and Association Plot results can be used as an input for functions from other packages. For instance, to conduct GO enrichment analysis of cluster-specific genes identified using Association Plots we developed a function ***apl_topGO***, which allows for conducting a GO enrichment analysis using the *topGO* package.

7.1.4 Gene Ontology Enrichment Analysis

To interpret the biological meaning of a cluster-specific gene set identified using Association Plots the *APL* package allows for conducting and visualizing Gene Ontology (GO) enrichment analysis using the R package *topGO* (Alexa and Rahnenfuhrer, 2021). We demonstrate this on the example of the lymphoid cell cluster from stomach, as obtained from the human cell atlas of fetal gene expression (introduced in Section 6.2).

To conduct the analysis we applied the function ***apl_topGO*** to 358 genes from the Association Plot for the lymphoid cells with S_α score above 1. In Fig. 7.1

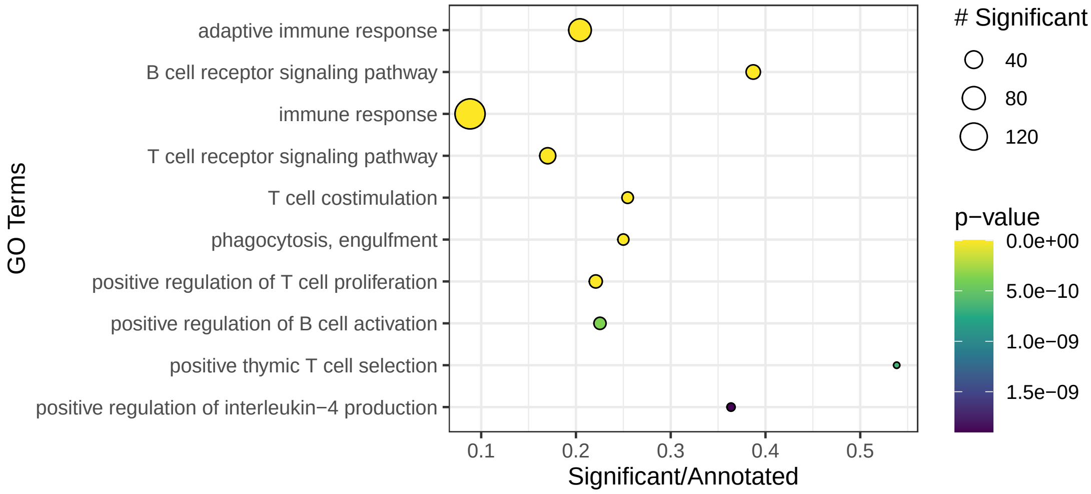


Figure 7.1: GO enrichment analysis of lymphoid cells from the stomach cluster obtained from human cell atlas of fetal gene expression. The figure presents our visualization of the *topGO* results, implemented in the *APL* package. Only 10 most significantly enriched GO terms are shown.

the 10 most significantly enriched GO terms are shown. A prominent role of GO terms related to T cells and immunity is apparent in the figure. According to the authors of the data set (Cao *and others*, 2020), the majority of lymphoid cells in stomach are T cells which is in line with most of the terms being related to T cells.

These results can be made more intuitive by mapping the information into the Association Plot. Genes belonging to a given GO term can be highlighted in the Association Plot. As an example, in the Association Plot for lymphoid cells we highlighted all genes annotated to a GO term 'GO:0050853 B cell receptor signaling pathway' (Fig. 7.2). Most of these genes are significantly enriched (located in rainbow area). Other genes, which are not in the region where $S_\alpha > 1$ are visibly still close to this region. In particular, one can recognize those genes of a GO category which are strongly associated to the cluster as opposed to others which are apparently shared with other clusters or are lowly expressed in the cluster.

Although in the presented example we demonstrated the results for lymphoid cells obtained using the *topGO* package, the results from Association Plots can be smoothly integrated with various R packages.

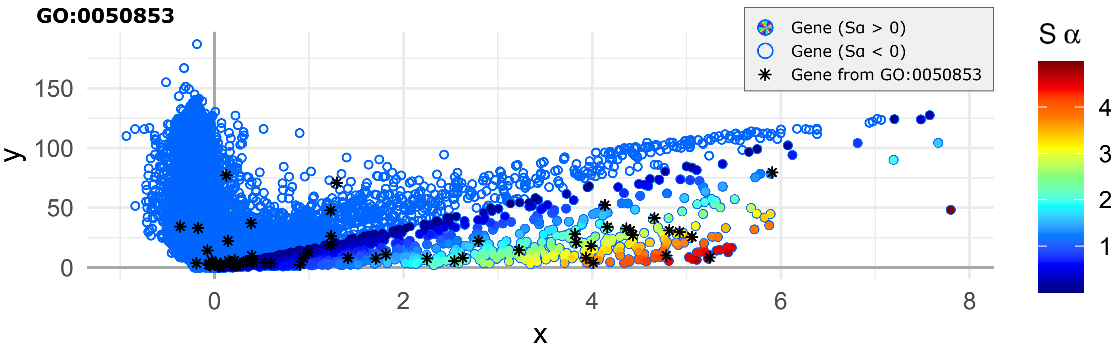


Figure 7.2: Location of genes annotated to the GO term 'GO:0050853 B cell receptor signaling pathway' in the Association Plot for the lymphoid cells. Genes belonging to this GO category are marked using black stars.

7.2 Shiny app

7.2.1 Introduction to the Shiny app

The Shiny app *APL* is an interactive application for exploration of input data and identification of condition-specific genes using Association Plots. In contrast to the R package *APL*, the Shiny app does not require any programming skills and offers an interactive, user-friendly interface. The software not only displays Association Plots but also allows for interactive exploration of the resulting plots.

7.2.2 Download and installation

The app was generated using 'shiny' (Chang *and others*, 2020), a web application framework for R, and is available from the GitHub repository `https://github.com/elagralinska/APL`. Only the SVD routine (torch.svd) is taken from the Python3 torch package (Paszke *and others*, 2019) and gets called from the R code. The Shiny app requires the R program, which is freely available from CRAN at http://cran.r-project.org. More details on the app installation and further requirements can be found in the tutorial available on GitHub.

7.2.3 Usage

After opening the graphical interface of the app, the input data needs to be provided using the file input widget called *Gene expression data input*. The app accepts any matrix with non-negative entries saved in form of a tab-separated .txt file. Importantly, the gene expression input matrix needs to contain column names and the first column needs to contain gene names.

After a successful data upload a notification with two numbers will be displayed in the app: the total number of genes in the data matrix, and the number of genes with a non-zero variance. The number of genes with the highest variance across all conditions, which will be used for further analysis, can be specified by the user in the field *Number of genes*. By default, the computation is done using 1,000 genes. Next, the user should specify the number of CA dimensions using the widget *Number of dimensions*. Therefore, in the tab *Suggested number of dimensions* three dimension numbers, computed using elbow rule, 80% rule, and average rule (see Section 4.4), are suggested by the app. Finally, to run the program the button *Start calculations* should be used.

Association Plots generated by the program can be seen in the tab *Association Plot*. The cluster of conditions, for which the plot should be generated, can be changed using the widget *Which conditions*. This can be done either by manually providing column IDs of conditions of interest, or by uploading a cluster annotation file when working with single cell data. The generated Association Plots are interactive, and by clicking the mouse over a gene a barplot illustrating the expression of a selected gene across all conditions from the input data will be shown. To highlight selected genes in the Association Plot the user should provide the gene names in the widget *Genes to highlight* located in the left panel. The selected genes will then appear in the Association Plot in red. To download the list of condition-specific genes obtained using the Association Plots, the button *Download gene ranking* located below the plot should be used.

Besides Association Plots, the program allows also for conducting a Gene Ontology enrichment analysis of the condition-specific genes obtained from the Association Plot. Therefore, the program uses the R package topGO (Alexa and

Rahnenfuhrer, 2021) and the annotation file needs to be specified by the user by using the buttons *Annotation file* in the left panel. The list of available annotation files is provided in the tutorial on GitHub. The output of the enrichment analysis is provided both in the graphical and tabular form.

Finally, the *APL* shiny app allows for generating a two- and three-dimensional representation of the data. The resulting plots will be shown in the tabs *2D plot* and *3D plot*. Both types of plots are interactive and facilitate the exploration of the input data. Further details on the usage of the tool are provided in the tutorial available in the GitHub repository.

CHAPTER 8

Discussion and conclusions

Motivated by frequently occurring questions about cluster-specific genes in transcriptomics studies we developed Association Plots, a novel method for visualization and analysis of bulk- and single-cell data. Today, given the size and complexity of such data, it is challenging to extract information from data visualization or data queries. As a potential solution to this we use Association Plots for the identification of genes associated to individual clusters of conditions in the data, the so-called marker genes or cluster-specific genes.

At the heart of Association Plots lies the geometry of the correspondence analysis biplot. As described in Section 3.5, in the CA space the direction from the origin towards a cluster of conditions allows for an interpretation of gene-condition associations within the data. This allows the genes to be represented by their orthogonal distance to the cluster direction vector. The smaller is this distance and the distance between a gene's projection onto this vector and a cluster centroid, the higher is the cluster-specificity of a given gene.

The geometry of correspondence analysis is a necessary prerequisite for Association Plots. As described in Chapter 4, Association Plots depict the gene-condition associations in a planar coordinate system, independent of the original dimension of the data. This is particularly important when working with single-cell RNA-seq, due to the large amount of data contained in these data sets. A traditional exploratory visualization method like, e.g., PCA may provide little help, and projection of such data into the plane would lead to a huge loss of information. Association Plots tackle this problem and offer the capability to visualize and interact with the sets of genes that are associated to a cluster of conditions, in a manner only dependent on the clusters to be studied but independent of the size of the data set. However, unlike with, e.g., PCA, this is not achieved by simple projection. Rather, Association Plots constitute a non-linear mapping of the high-dimensional image into a plane, while preserving the gene-cluster association features.

To demonstrate the application of Association Plots to bulk and single-cell transcriptomics data we focused on three example data sets. First of all, in Chapter 5 we applied Association Plots to the Genotype-Tissue Expression data to present how to use Association Plots for visualization of associations of genes to groups of samples in bulk data. Then, in Chapter 6, we applied Association Plots to two example single-cell data sets: the 3k PBMC data set containing information on gene expression in peripheral blood mononuclear cells, and the human cell atlas of fetal gene expression containing information on *in vivo* gene expression across diverse organs and cell types. These last two examples demonstrate that Association Plots can be used for identification of novel marker genes as well as cluster annotation purposes in single-cell data.

To implement the concept of Association Plots we developed an R package and a shiny app *APL*, which is freely available on Bioconductor and GitHub and allows for interactively querying various aspects of the data. Using *APL*, one can extract sets of genes specific to a given cluster of samples or cells, map marker genes into an Association Plot for the purpose of cell type annotation, or visualize gene set enrichment in a cell cluster. Additionally, *APL* is integrated with Gene Ontology enrichment to further support the annotation process. The developed software is described in Chapter 7.

An important consideration when generating Association Plots is choosing the number of dimensions to retain. When defining CA space we apply dimension reduction merely for the purpose of noise reduction, and not to a degree where the visualization would cancel significant amounts of information from the data. Given large data resides in thousands of CA dimensions, applying singular value decomposition typically results in many dimensions representing noise. These are associated to small singular values. Thus, we employ methods such as elbow rule, described in Section 4.4, for estimating the number of dimensions to keep in our representation of the data in CA space. This will typically be way more than three dimensions, but also way less than the full dimensionality of the data. A positive side effect of reducing the number of dimensions in this way is that computations become faster than when done on the original data. Therefore, in the *APL* package we implemented three alternative methods for computing the number of dimensions to keep.

In the example presented in this thesis we focused on the dimension numbers calculated using three common approaches, which resulted in fairly robust Association Plots. However, a too high or too low number of CA dimensions can significantly influence the structure of an Association Plot. For instance, for a selected cluster of conditions from an input data a too high number of dimensions retained in the analysis would result in a plot in which the characteristic tail of genes does not align with the x-axis. Instead, a positive angle between the tail of genes and the x-axis could be observed. This is due to the projection length of a cluster-specific gene onto a vector from origin towards a cluster centroid, which increases with every further CA dimension retained in a CA space. On the other hand, retaining a too low number of CA dimensions, lower than a number of condition clusters from input data, poses a significant problem as it hinders the identification of cluster-specific genes. This is caused by the fact that in a too low-dimensional CA space some clusters from an input data cannot be clearly distinguished from each other as the differences between them are too small.

Another important aspect, often overlooked in data analysis methods, is whether a method will indicate that its assumptions are not met. Thus, Association Plots can reveal that a given cluster of conditions from the data is in fact not a coherent cluster. In this case the typical structure of genes pointing in the direction of the cluster centroid in the CA space will be dispersed, and, as a consequence, the right tail in the Association Plot generated for this cluster will be short and not clearly visible anymore. Flagging the violation of a clustering assumption is a desirable feature of our method.

Finally, a marker gene is meant to distinguish a particular condition from other conditions. In the context of transcriptomics data, a marker gene can highlight one cluster over the other clusters in the data set. The marker genes delineated by an Association Plot thus need to be understood as 'relative' marker genes dependent on the given data set, and the composition of the set of marker genes may vary depending on the set of conditions present in the data. Clearly, this is no different from other commonly-used tools for differential expression testing, and thus, searching for novel marker genes should always be accompanied by the appropriate experimental design.

We believe that Association Plots offer a new viewpoint on transcriptomics data analysis which opens up many further interesting questions. For now, the *APL* package requires a pre-clustered data, and the cluster information needs to be provided. Therefore, focusing on the connection of Association Plots to biclustering (Tanay *and others*, 2002; Pontes *and others*, 2015) as well as to spectral clustering (Zelnik-Manor and Perona, 2005; Von Luxburg, 2007) is most worthwhile.

Bibliography

10x Genomics. ((2016)). 3k PBMCs from a Healthy Donor, Single Cell Gene Expression Dataset by Cell Ranger 1.1.0.

Aggarwal, Mohit, Villuendas, Raquel, Gomez, Gonzalo, Rodriguez-Pinilla, Socorro M, Sanchez-Beato, Margarita, Alvarez, David, Martinez, Nerea, Rodriguez, Antonia, Castillo, Maria E, Camacho, Francisca I *and others*. (2009). Tcl1a expression delineates biological and clinical variability in b-cell lymphoma. *Modern Pathology* **22**(2), 206–215.

Alexa, Adrian and Rahnenfuhrer, Jorg. (2021). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.46.0.

Anders, Simon, Pyl, Paul Theodor and Huber, Wolfgang. (2015). Ht-seq—a python framework to work with high-throughput sequencing data. *bioinformatics* **31**(2), 166–169.

Benzécri, Jean-Paul. (1973). *L'analyse des correspondances*. Dunod, Paris.

Brinas, François, Danger, Richard and Brouard, Sophie. (2021). Tcl1a, b cell regulation and tolerance in renal transplantation. *Cells* **10**(6), 1367.

Cao, Junyue, O'Day, Diana R, Pliner, Hannah A, Kingsley, Paul D, Deng, Mei, Daza, Riza M, Zager, Michael A, Aldinger, Kimberly A, Blecher-Gonen, Ronnie, Zhang, Fan *and others*. (2020). A human cell atlas of fetal gene expression. *Science* **370**(6518).

Carithers, Latarsha J, Ardlie, Kristin, Barcus, Mary, Branton, Philip A, Britton, Angela, Buia, Stephen A, Compton, Carolyn C, DeLuca, David S, Peter-Demchok, Joanne, Gelfand, Ellen T *and others*. (2015). A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking* **13**(5), 311–319.

Casneuf, Tineke, Van de Peer, Yves and Huber, Wolfgang. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC bioinformatics* **8**(1), 1–13.

Chang, Winston, Cheng, Joe, Allaire, JJ, Xie, Yihui and McPherson, Jonathan. (2020). *shiny: Web Application Framework for R*. R package version 1.4.0.2.

Chu, Zhong, Huo, Nan, Zhu, Xiang, Liu, Hanxiao, Cong, Rui, Ma, Luyuan, Kang, Xiaofeng, Xue, Chunyuan, Li, Jingtong, Li, Qihong *and others*. (2021). Foxo3a-induced linc00926 suppresses breast tumor growth and metastasis through inhibition of pgk1-mediated warburg effect. *Molecular Therapy*.

Ciampi, Antonio, González Marcos, Ana and Castejón Limas, Manuel. (2005). Correspondence analysis and 2-way clustering. *SORT* **29**(1).

Deng, Kaiwen, Li, Yueming, Zhang, Hanrui, Wang, Jian, Albin, Roger L and Guan, Yuanfang. (2022). Heterogeneous digital biomarker integration out-performs patient self-reports in predicting parkinson's disease. *Communications Biology* **5**(1), 1–10.

Dey, Kushal K, Hsiao, Chiaowen Joyce and Stephens, Matthew.

(2017). Visualizing the structure of rna-seq expression data using grade of membership models. *PLoS genetics* **13**(3), e1006599.

Dobin, Alexander, Davis, Carrie A, Schlesinger, Felix, Drenkow, Jorg, Zaleski, Chris, Jha, Sonali, Batut, Philippe, Chaisson, Mark and Gingeras, Thomas R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**(1), 15–21.

Engebraaten, Olav, Yau, Christina, Berg, Kristian, Borgen, Elin, Garred, Øystein, Berstad, Maria EB, Fremstedal, Ane SV, DeMichele, Angela, Veer, Laura van't, Esserman, Laura *and others*. (2021). Rab5a expression is a predictive biomarker for trastuzumab emtansine in breast cancer. *Nature communications* **12**(1), 1–11.

Gower, John C, Lubbe, Sugnet Gardner and Le Roux, Niel J. (2011). *Understanding biplots*. John Wiley & Sons.

Gralinska, Elzbieta, Kohl, Clemens, Sokhandan Fadakar, Bita and Vingron, Martin. (2022). Visualizing cluster-specific genes from single-cell transcriptomics data using association plots. *Journal of Molecular Biology* **434**(11), 167525.

Gralinska, Elzbieta and Vingron, Martin. (2020). Association plots: Visualizing associations in high-dimensional correspondence analysis biplots. *bioRxiv*.

Greenacre, Michael. (2017). *Correspondence analysis in practice*. Chapman and Hall/CRC, Boca Raton (FL).

Greenacre, Michael J. (1984). *Theory and applications of correspondence analysis*. Academic Press, London.

Greenacre, Michael J and Blasius, Jörg. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. Academic Press, London.

Hao, Yuhan, Hao, Stephanie, Andersen-Nissen, Erica, Mauck, William M., Zheng, Shiwei, Butler, Andrew, Lee, Maddie J., Wilk, Aaron J., Darby, Charlotte, Zagar, Michael, Hoffman, Paul, Stoeckius, Marlon, Papalexi, Efthymia, Mimitou, Eleni P., Jain, Jaison, Srivastava, Avi, Stuart, Tim, Fleming, Lamar B., Yeung, Bertrand, Rogers, Angela J., McElrath, Juliana M., Blish, Catherine A., Gottardo, Raphael, Smibert, Peter *and others*. (2020). Integrated analysis of multimodal single-cell data. *bioRxiv*.

Haque, Ashraful, Engel, Jessica, Teichmann, Sarah A and Lönnberg, Tapio. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine* **9**(1), 1–12.

Hotelling, Harold. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24**(6), 417.

International Labour Organization.

Jolliffe, IT. (2002). Isbn 978-0-387-95442-4,“. *Principal component analysis, series: Springer series in statistics* **29**(487), 28.

Kolodziejczyk, Aleksandra A, Kim, Jong Kyoung, Svensson, Valentine, Marioni, John C and Teichmann, Sarah A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell* **58**(4), 610–620.

Lähnemann, David, Köster, Johannes, Szczurek, Ewa, McCarthy, Davis J, Hicks, Stephanie C, Robinson, Mark D, Vallejos, Catalina A, Campbell, Kieran R, Beerenwinkel, Niko, Mahfouz, Ahmed *and others*. (2020). Eleven grand challenges in single-cell data science. *Genome biology* **21**(1), 1–35.

Lempriere, Sarah. (2021). Exosomal microrna is promising biomarker in pd. *Nature Reviews Neurology*, 1–1.

Li, J, Guo, H, Ma, Y, Chen, H and Qiu, M. (2021). 11p linc00926 is a b cell-specific long non-coding rna in lung adenocarcinoma and is associated with the prognosis of patients with this disease. *Journal of Thoracic Oncology* **16**(4), S703.

Liang, Yuexiong, Zhu, Haifeng, Chen, Jing, Lin, Wei, Li, Bing and Guo, Yusheng. (2020). Construction of relapse-related lncrna-mediated cerna networks in hodgkin lymphoma. *Archives of Medical Science: AMS* **16**(6), 1411.

Liberzon, Arthur, Birger, Chet, Thorvaldsdóttir, Helga, Ghandi, Mahmoud, Mesirov, Jill P and Tamayo, Pablo. (2015). The molecular signatures database hallmark gene set collection. *Cell systems* **1**(6), 417–425.

Love, Michael I, Huber, Wolfgang and Anders, Simon. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**(12), 1–21.

Luecken, Malte D and Theis, Fabian J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology* **15**(6), e8746.

McCarthy, Davis J, Chen, Yunshun and Smyth, Gordon K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research* **40**(10), 4288–4297.

Montaner, Joan, Ramiro, Laura, Simats, Alba, Tiedt, Steffen, Makris, Konstantinos, Jickling, Glen C, Debette, Stephanie, Sanchez, Jean-Charles and Bustamante, Alejandro. (2020). Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nature Reviews Neurology* **16**(5), 247–264.

Mootha, Vamsi K, Lindgren, Cecilia M, Eriksson, Karl-Fredrik, Subramanian, Aravind, Sihag, Smita, Lehar, Joseph, Puigserver, Pere, Carlsson, Emma, Ridderstråle, Martin, Laurila, Esa *and others*. (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**(3), 267–273.

Novelli, Giuseppe, Ciccacci, Cinzia, Borgiani, Paola, Amati, Marisa Papaluca and Abadie, Eric. (2008). Genetic tests and genomic biomarkers: regulation, qualification and validation. *Clinical cases in mineral and bone metabolism* **5**(2), 149.

Parmigiani, Giovanni, Garrett, Elizabeth S, Irizarry, Rafael A and Zeger, Scott L. (2003). The analysis of gene expression data: an overview of methods and software. *The analysis of gene expression data*, 1–45.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca *and others*. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026–8037.

Pearson, Karl. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2**(11), 559–572.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent *and others*. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830.

Pevzner, Pavel. (2000). *Computational molecular biology: an algorithmic approach*. MIT press.

Pontes, Beatriz, Giráldez, Raúl and Aguilar-Ruiz, Jesús S. (2015). Biclustering on expression data: A review. *Journal of biomedical informatics* **57**, 163–180.

Risso, Davide, Perraudeau, Fanny, Gribkova, Svetlana, Dudoit, Sandrine and Vert, Jean-Philippe. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications* **9**(1), 1–17.

Ritchie, Matthew E, Phipson, Belinda, Wu, DI, Hu, Yifang, Law, Charity W, Shi, Wei and Smyth, Gordon K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**(7), e47–e47.

Robinson, Mark D, McCarthy, Davis J and Smyth, Gordon K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140.

Schena, Mark, Shalon, Dari, Davis, Ronald W and Brown, Patrick O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* **270**(5235), 467–470.

Sellers, Subhashini A, Fischer, William A, Heise, Mark T and Schughart, Klaus. (2021). Highly dampened blood transcriptome response in hiv patients after respiratory infection. *Scientific reports* **11**(1), 1–7.

Simoneau, Joël, Dumontier, Simon, Gosselin, Ryan and Scott, Michelle S. (2021). Current rna-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics* **22**(1), 140–145.

Stark, Rory, Grzelak, Marta and Hadfield, James. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics* **20**(11), 631–656.

Strimbu, Kyle and Tavel, Jorge A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS* **5**(6), 463.

Subramanian, Aravind, Tamayo, Pablo, Mootha, Vamsi K, Mukherjee, Sayan, Ebert, Benjamin L, Gillette, Michael A, Paulovich, Amanda, Pomeroy, Scott L, Golub, Todd R, Lander, Eric S *and others*. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550.

Suter, David M, Molina, Nacho, Gatfield, David, Schneider, Kim, Schibler, Ueli and Naef, Felix. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *science* **332**(6028), 472–474.

Tanay, Amos, Sharan, Roded and Shamir, Ron. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18**(suppl_1), S136–S144.

Tang, Fuchou, Barbacioru, Catalin, Wang, Yangzhou, Nordman, Ellen, Lee, Clarence, Xu, Nanlan, Wang, Xiaohui, Bodeau, John, Tuch, Brian B, Siddiqui, Asim *and others*. (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature methods* **6**(5), 377–382.

Trapnell, Cole, Pachter, Lior and Salzberg, Steven L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics* **25**(9), 1105–1111.

Tusher, Virginia Goss, Tibshirani, Robert and Chu, Gilbert. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**(9), 5116–5121.

Uhlén, Mathias, Fagerberg, Linn, Hallström, Björn M, Lindskog, Cecilia, Oksvold, Per, Mardinoglu, Adil, Sivertsson, Åsa, Kampf, Caroline, Sjöstedt, Evelina, Asplund, Anna *and others*. (2015). Tissue-based map of the human proteome. *Science* **347**(6220), 1260419.

Uhlen, Mathias, Karlsson, Max J, Zhong, Wen, Tebani, Abdellah, Pou, Christian, Mikes, Jaromir, Lakshmikanth, Tadepally, Forsström, Björn, Edfors, Fredrik, Odeberg, Jacob *and others*. (2019). A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**(6472).

Van den Berge, Koen, Perraudeau, Fanny, Soneson, Charlotte, Love, Michael I, Risso, Davide, Vert, Jean-Philippe, Robinson, Mark D, Dudoit, Sandrine and Clement, Lieven. (2018). Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome biology* **19**(1), 1–17.

Von Luxburg, Ulrike. (2007). A tutorial on spectral clustering. *Statistics and computing* **17**(4), 395–416.

Wang, Fangce, Tian, Xiaoxue, Zhou, Jie, Wang, Guangming, Yu, Wenlei, Li, Zheng, Fan, Zhuoyi, Zhang, Wenjun and Liang, Aibin. (2018). A three-lncrna signature for prognosis prediction of acute myeloid leukemia in patients. *Molecular medicine reports* **18**(2), 1473–1484.

Wang, Tianyu, Li, Boyang, Nelson, Craig E and Nabavi, Sheida. (2019*a*). Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics* **20**(1), 1–16.

Wang, Yejinpeng, Chen, Liang, Ju, Lingao, Qian, Kaiyu, Liu, Xuefeng, Wang, Xinghuan and Xiao, Yu. (2019*b*). Novel biomarkers associated with progression and prognosis of bladder cancer identified by co-expression analysis. *Frontiers in oncology* **9**, 1030.

Wang, Yanbo, Liu, Jing, Ren, Fenghai, Chu, Yanjie and Cui, Binbin. (2021). Identification and validation of a four-long non-coding rna signature associated with immune infiltration and prognosis in colon cancer. *Frontiers in Genetics* **12**.

Wang, Zhong, Gerstein, Mark and Snyder, Michael. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**(1), 57–63.

Weber, Andreas PM. (2015). Discovering new biology through sequencing of rna. *Plant physiology* **169**(3), 1524–1531.

Zelnik-Manor, Lihi and Perona, Pietro. (2005). Self-tuning spectral clustering. In: *Advances in neural information processing systems*. pp. 1601–1608.

Zheng, Grace XY, Terry, Jessica M, Belgrader, Phillip, Ryvkin, Paul, Bent, Zachary W, Wilson, Ryan, Ziraldo, Solongo B, Wheeler, Tobias D, McDermott, Geoff P, Zhu, Junjie *and others*. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049.

List of Figures

List of Tables

APPENDIX A

Supplementary Figures

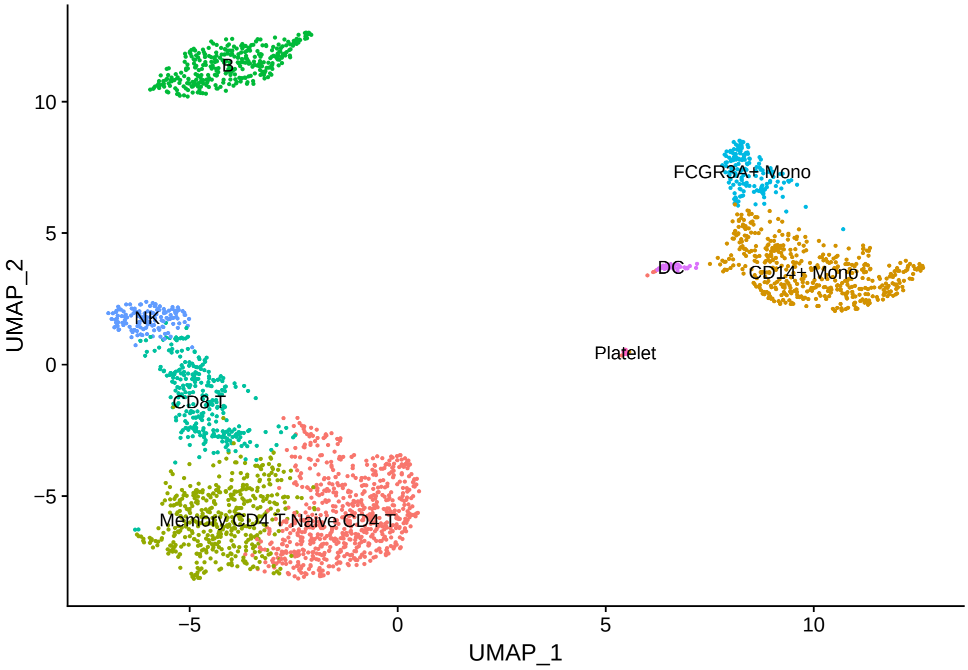


Figure A.1: Cell type subpopulations in 3k PBMC data. Colors of the cells in the UMAP refer to the subpopulations of PBMCs. NK, natural killer cells; DC, dendritic cells.

APPENDIX B

Supplementary Tables

Table B.1: GSEA results of 100 top genes from B cells.

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HAY_BONE_MARROW_FOLLICULAR_B_CELL	142	—	48	0.338	1.01E-95	2.07E-91
GSE10325_CD4_TCELL_VS_BCELL_DN	194	Genes down-regulated in comparison of healthy CD4 [GeneID=920] T cells versus healthy CD19 [GeneID=920] B cells.	51	0.2629	1.33E-95	2.07E-91
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_BCELL_DN	195	Genes down-regulated in comparison of systemic lupus erythematosus CD4 [GeneID=920] T cells versus systemic lupus erythematosus B cells.	47	0.241	1.65E-85	1.71E-81
GSE4984_UNTREATED_VS_GALECTIN1_TREATED_DC_DN	191	Genes down-regulated in monocyte-derived dendritic cells: control versus treated with LGALS1 [GeneID=3956].	40	0.2094	2.63E-69	2.04E-65
GSE29618_BCELL_VS_MONO-CYTE_DAY7_FLU_VACCINE_UP	195	Genes up-regulated in comparison of B cells from influenza vaccinee at day 7 versus monocytes from influenza vaccinee at day 7.	39	0.2	1.27E-66	7.92E-63
GSE29618_BCELL_VS_MONOCYTE_UP	194	Genes up-regulated in comparison of B cells versus monocytes.	38	0.1959	1.88E-64	9.75E-61
GSE29618_BCELL_VS_MDC_DAY7_FLU_ VAC-CINE_UP	192	Genes up-regulated in comparison of B cells from influenza vaccinee at day 7 post-vaccination versus myeloid dendritic cells (mDC) at day 7 post-vaccination.	37	0.1927	2.16E-62	9.6E-59
GSE10325_BCELL_VS_MYELOID_UP	196	Genes up-regulated in comparison of healthy B cells versus healthy myeloid cells.	35	0.1786	1.28E-57	5E-54
GSE3982_MEMORY_CD4_TCELL_VS_BCELL_DN	197	Genes down-regulated in comparison of memory CD4 [GeneID=920] T cells versus B cells.	34	0.1726	2.3E-55	7.95E-52
GSE22886_TCELL_VS_BCELL_NAIVE_DN	198	Genes down-regulated in comparison of naive CD4 [GeneID=920] CD8 T cells versus naive B cells.	34	0.1717	2.77E-55	8.63E-52

Table B.2: GSEA results of 100 top genes from CD14+ monocytes.

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HAY_BONE_MARROW_NEUTROPHIL	450	—	56	0.1244	3.91E-85	1.22E-80
GSE11057_PBMC_VS_MEM_CD4_TCELL_UP	197	Genes up-regulated in comparison of peripheral mononuclear blood cells (PBMC) versus memory T cells.	30	0.1523	1.38E-46	2.15E-42
GSE29618_MONOCYTE_VS_MDC_DAY7_FLU_VACCINE_UP	200	Genes up-regulated in comparison of monocytes from influenza vaccinee at day 7 post-vaccination versus myeloid dendritic cells at day 7 post-vaccination.	30	0.15	2.24E-46	2.33E-42
GSE29618_MONOCYTE_VS_PDC_UP	199	Genes up-regulated in comparison of monocytes versus plasmacytoid dendritic cells (pDC).	29	0.1457	2.06E-44	1.25E-40
GSE29618_MONOCYTE_VS_MDC_UP	200	Genes up-regulated in comparison of monocytes versus myeloid dendritic cells (mDC).	29	0.145	2.41E-44	1.25E-40
GSE29618_MONOCYTE_VS_PDC_DAY7_FLU_VACCINE_UP	200	Genes up-regulated in comparison of monocytes from influenza vaccinee at day 7 post-vaccination versus plasmacytoid dendritic cells (mDC) at day 7 post-vaccination.	29	0.145	2.41E-44	1.25E-40
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_MYELOID_DN	200	Genes down-regulated in comparison of systemic lupus erythematosus CD4 [GeneID=920] T cells versus systemic lupus erythematosus myeloid cells.	27	0.135	2.36E-40	9.33E-37
GSE6269_HEALTHY_VS_STAPH_PNEUMO_INF_PBMC_DN	170	Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) from healthy donors versus PBMC from patients with acute S. pneumoniae infection.	26	0.1529	2.4E-40	9.33E-37
DURANTE_ADULT_OLFACTORY_ NEUROEPITHELIUM_ DENDRITIC_CELLS	117	—	21	0.1795	3.55E-34	1.23E-30
GSE10325_BCELL_VS_MYELOID_DN	200	Genes down-regulated in comparison of healthy B cells versus healthy myeloid cells.	23	0.115	1.13E-32	3.2E-29

APPENDIX C

Supplementary Methods

Employment data

The employment data set was downloaded on 06.11.2020 from Ilostat database (International Labour Organization). The data describes sectors of employment in 233 countries and groups of countries. In the example presented in this study we focused on the data from years 2000, 2005, 2010, 2015, from the eight following category types: "employment in agriculture, female", "employment in agriculture, male", "employment in services, female", "employment in services, male", "employment in industry, female", "employment in industry, male", "unemployment, female", "unemployment, male". Additionally, to obtain the percentages of total female or male labor force employed in three different sectors the data from the employment categories were multiplied by the factor $1 - x/100$, where x describes the percentage of the labor force for given year and gender, which was unemployed (data read from categories "unemployment, female" or "unemployment, male". At the end, the resulting matrix of 233 rows and 32 columns was submitted to correspondence analysis and the analysis was done using eight CA dimensions.

GTEx data

RNA-seq data of postmortem non-disease human tissues was retrieved from GTEx Portal (?). The files ”GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz” containing gene TPM values and ”GTEx_v7_Annotations_SampleAttributesDS.txt” containing sample annotations (both files available from https://gtexportal.org/home/datasets) were downloaded on 03.07.2019. A detailed description of data processing procedures is available from https://gtexportal.org/home/documentationPage.

Correspondence analysis was computed using a subset of 5,000 genes with the highest variance across 11,688 columns of the chi-square component matrix, which was calculated based on the input gene expression matrix. For generating the Association Plots in Fig. 5.1 the first 96 CA dimensions (number obtained using the elbow rule) were considered.

3k PBMC data

The UMI count matrix (”Gene/cellmatrix(filtered)”) of peripheral blood mononuclear cells (PBMCs) from a healthy donor Zheng *and others* (2017); 10x Genomics (2016) was downloaded on 02.08.2021 from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k and analyzed according to the Seurat Guided Clustering Tutorial available from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html. Below we present the steps of the conducted analysis.

Features detected in less than three cells were removed from the data. Additionally, the cells for which less than 200 or more than 2,500 features were detected, as well as cells with expression of mitochondrial genes higher than 5% of total counts, were also removed from the data. The filtered data was then normalized using a method ”LogNormalize“ from Seurat 4.0 package Hao *and others* (2020) and linearly transformed using its ”ScaleData“ function. Subsequently, PCA was performed on the matrix of the 2,000 most variable genes and the first 10 PCs were selected for constructing K-nearest neighbor graph. Next, cells were clustered using the Louvain algorithm (resolution parameter of 0.2) and the UMAP visualization was generated using the first 10 PCs. To match the clusters to known cell types the expression of canonical markers was investigated (Supplementary Fig. A.1). Finally, correspondence analysis was applied to the normalized UMI count matrix using all 13,713 genes. For generating the Association Plots, the first 223 CA dimensions (number obtained using the ”elbow rule”) were considered.

Gene set enrichment analysis

The gene set enrichment analysis of the PBMC cell clusters was conducted using the Molecular Signatures Database (MSigDB) v7.2 Liberzon *and others* (2015) together with the GSEA method Subramanian *and others* (2005); Mootha *and others* (2003) available from `http://www.gsea-msigdb.org/gsea/msigdb/annotate.jsp`. The analysis was run using the default parameters and all available MSigDB gene sets. The results were sorted according to the size of the overlap between the input gene set and the gene sets from the MsigDB collection.

Human cell atlas of fetal gene expression data

A processed data set with normalized counts from all cells (”`Human\_RNA\_processed.loom''`) was downloaded on 01.15.2021 from `https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/` (Cao *and others*, 2020). The normalized counts were obtained using the protocol described in the publication (Cao *and others*, 2020).

Correspondence analysis was applied to the normalized count matrix using all genes across 12,106 cells from the stomach clusters. The Association Plots for each of the 16 stomach cell types were generated using the first 4047 CA dimensions.

Differential expression testing tools

To investigate the agreement among results obtained with Association Plots, DESeq2 Love *and others* (2014), edgeR Robinson *and others* (2010); McCarthy *and others* (2012), and Seurat 4.0 Hao *and others* (2020), we applied them to each cell type from the 3k PBMC data. The gene rankings from each tool were computed in the following way.

For DESeq2 Love *and others* (2014), the analysis was performed by combining DESeq2 package with the `zinbwave` function from ZINB-WaVE package Risso *and others* (2018), as recommended in the DESeq2 vignette from 10/27/2020 for single-cell analysis. For this purpose we applied DESeq2 to zimbwave-weighted count matrix using `test="LRT"` for significance testing, and the following `DESeq` arguments: `useT=TRUE`, `minmu=1e-6`, and `minReplicatesForReplace=Inf`. Finally, the genes were sorted by p-value in increasing order and the genes with non-positive log2 fold-change values were removed from the analysis.

For edgeR Robinson *and others* (2010); McCarthy *and others* (2012), the differential expression analysis for each cluster was also performed using the observational weights computed by zinbwave function. We followed the zinbwave vignette from 10/28/2020. The genes were then sorted by p-values in an increasing order, and genes with non-positive log2 fold-change values were discarded.

For Seurat Hao *and others* (2020), the gene rankings were computed for each cluster separately using the `FindAllMarkers` function with the parameters: `only.pos=FALSE`, `min.pct=0`, `logfc.threshold=0`, `return.thresh=1.01`. The genes were then sorted by p-value in increasing order and the genes with non-negative log2 fold-change values were removed from the analysis.

For Association Plots, genes were ranked by S_α in decreasing order.

To generate the heatmaps from Fig. 6.5 1,000 top genes from each gene ranking were extracted, and the size of the gene overlaps between the rankings was computed.

To generate the Association Plots for the dendritic cells from Fig. 6.6, from each gene ranking we selected 250 genes with the lowest p-values, which passed a log2 fold-change threshold of: 1.9 (DESeq2), 0.5 (Seurat), 1 (edgeR), and highlighted them in the Association Plots.

APPENDIX D

Zusammenfassung

Ein immer wiederkehrendes Problem bei der Analyse von Transkriptomdaten ist die Suche nach Assoziationen zwischen Clustern von Bedingungen und den dazugehörigen hochexprimierten Genen. Ansätze für dieses Problem gibt es in vielen Formen, wie zum Beispiel bei dem Biclustering oder bei der Suche nach Markergenen. Während die Identifizierung von Markergenen bei kleinen Datensätzen relativ einfach ist, stellt sie bei komplexen Datensätzen wie Einzelzell RNA-seq Experimenten eine erhebliche Herausforderung für die derzeit verfügbaren Analyse- und Visualisierungsmethoden dar. Insbesondere, Methoden zur Darstellung niedrigdimensionaler Daten wie die Hauptkomponentenanalyse (PCA) führen zu Informationsverlusten, da sie die in den höheren Dimensionen enthaltenen Informationen nicht anzeigen.

In dieser Arbeit wird dieses Problem durch die Einführung von Association Plots (APs), einer neuartigen Methode zur Bestimmung und Visualisierung von clusterspezifischen Genen in hochdimensionalen Daten, angegangen. APs werden von der Korrespondenzanalyse (CA) abgeleitet, einer Projektionsmethode ähnlich der PCA, die jedoch eine gemeinsame Einbettung von Genen und Bedingungen ermöglicht. Bei einer solchen Einbettung liegen Gene, die mit einem Cluster von Bedingungen assoziiert sind, zusammen in einer bestimmten Richtung im hochdimensionalen Raum. Die Messung der Abstände zwischen Genen und den dazugehörigen Bedingungen führt zu APs, die unabhängig von der Dimensionalität der Daten sind und bei der Identifikation von Markergenen helfen können.

Wir präsentieren die Anwendung von APs auf populationsbasierenden- und Einzelzell RNA-seq-Daten. Zunächst wird die Identifizierung von Markergenen mithilfe von APs am Beispiel von Genotype Tissue Expression (GTEx) und 3k Peripheral Blood Mononuclear Cell (PBMC)-Daten vorgestellt. Als Nächstes zeigen wir, wie APs bei der Annotation von Zellclustern zu bereits bekannten Zellidentitäten helfen, indem wir eine vordefinierte Liste von Markergenen am Beispiel des Zellatlases menschlicher fetaler Genexpressionsdaten verwenden. Gleichzeitig demonstrieren wir, wie APs zur Untersuchung von Ähnlichkeiten zwischen Clustern aus den Daten eingesetzt werden können und vergleichen die Ergebnisse von APs mit den Ergebnissen bestehender Tools zur differenziellen Genexpressionsanalyse. Abschließend demonstrieren wir *APL*, das entwickelte Bioconductor R-Paket und die Shiny App. *APL* implementiert das Konzept der APs und ist mit einer Überprüfung von Gene Ontology Begriffen ausgestattet.

APPENDIX E

Summary

A re-occurring problem in transcriptomics data analysis is the search for associations between clusters of conditions and the highly expressed genes these conditions share. Approaches to solve this problem occur in many forms, for instance biclustering or the search for marker genes. While for small data sets identification of marker genes is fairly easy, for complex data sets such as single-cell RNA-seq it poses a significant challenge to analysis and visualization methods currently available. In particular, low-dimensional data representation methods such as principal component analysis (PCA) lead to information loss, as they do not show information contained in higher dimensions.

In this thesis, we address this problem by presenting Association Plots (APs), a novel method for determining and visualizing cluster-specific genes in high-dimensional data. APs are derived from correspondence analysis (CA), a projection method similar to PCA, which however enables the joint embedding of genes and conditions. In such an embedding, genes associated to a cluster of conditions lie in a particular direction in high-dimensional space. Measuring distances between genes and conditions leads to APs which are independent of the data dimensionality and can aid in delineating marker genes.

We present the application of APs to bulk- and single-cell RNA-seq data through several examples. First, we show the identification of marker genes using APs on Genotype Tissue Expression (GTEx) and 3k Peripheral Blood Mononuclear Cell (PBMC) data. Next, we present how APs aid in cell cluster annotation using a predefined list of marker genes on human cell atlas of fetal gene expression data. Simultaneously, we also demonstrate how to apply APs for studying similarities between clusters from the data, and we compare results from APs to results from existing differential expression testing tools. Finally, we demonstrate *APL*, the developed Bioconductor R package and shiny app. *APL* implements the concept of APs and is integrated with the Gene Ontology enrichment tool.