Network inference from sparse single-cell transcriptomics data

Exploring, exploiting, and evaluating the single-cell toolbox

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Master of Science *Lisa Maria Steinheuer*

geboren am 04. März 1995 in Bonn

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Jörg Hackermüller (UFZ Leipzig, Deutschland)
2. Professor Dr. Wolfgang Enard (LMU München, Deutschland)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 09.03.2022 mit dem Gesamtprädikat *magna cum laude*.

Contents

Abstract **vii**

Acknowledgments **xi**

List of Abbreviations **xv**

1 Biological Introduction **1**

1.1 Introduction into Transcriptomics 1

1.1.1 Basis of transcriptomics 2

1.1.2 Workflow and challenges of next-generation sequencing 4

1.2 The Emergence of the Single-Cell World 6

1.2.1 Experimental workflow 7

1.2.2 Single-cell tools and applications 8

1.3 From Counts to Co-regulation - Gene Networks 11

1.3.1 Gene correlation networks 11

1.3.2 Gene networks in single-cell transcriptomics 11

1.4 Example use case: Study of age-related macular degeneration using an organoid system 13

1.4.1 Characteristics of AMD 14

1.4.2 Composition of the human retina 14

1.4.3 Disease progression of AMD 15

1.5 The Research Gap 17

2 Methodology 19

2.1 Single-cell data Imputation 19

2.2 Low-dimensional embedding and community detection. 22

2.3 Cell annotation via transfer learning 23

2.4 Pseudotime calculation 23

2.4.1 Inference of gene networks via `WGCNA` 25

3 Effects of dropout and data imputation on single-cell correlation networks 29

3.1 Workflow 30

3.1.1 Data Sets 32

3.1.2 Data imputation 33

3.1.3 Network Analysis using `WGNCA` 34

3.1.4 Cell cluster embedding and annotation 35

3.2 Generation of single-cell like reference data sets 38

3.3 Impacts of data imputation on network inference 43

3.4 Cell cluster annotation in human retina organoid data 47

3.5 Discussion 51

4 Imputation of cell type-specific gene regulatory networks in human retina organoids 55

4.1 Workflow 55

4.1.1 Data preparation 57

4.1.2 Network inference by means of `WGCNA` 57

4.1.3 Network characterization 58

4.2 Network inference after `DCA` imputation 60

4.2.1 `DCA` imputation allows to generate scale-free topology networks. 60

4.2.2 Specific modules for rods and cones can be identified in cell type-specific networks 63

4.3 Discussion 74

5 Characterization of human retinal organoids 79

5.1 Workflow 80

5.1.1 HRO characterization 80

5.1.2 Correlating HRO system to foveal and peripheral data sets 84

5.1.3 Approach developmental dynamics via RNA velocity 85

5.2 Annotation of neonatal human retina organoids 87

5.2.1 Comparison of manual and machine learning annotations 97

5.2.2 Comparison of HRO sample variance 99

5.3 Ambiguous correlation results of HRO towards fovea and periphery 101

5.4 RNA velocity analysis revealed dynamics within retina organoid systems . . . 106

5.5 Discussion 110

6 Conclusion 119

7 Supporting material 127

7.1 Data and Code availability 128

7.2 Chapter 3 128

7.3 Chapter 4 135

7.4 Chapter 5 137

List of Figures i

List of Tables iii

Bibliography v

Curriculum Scientiae xvii

Publications xix

Abstract

Large-scale transcriptomics data studies revolutionised the fields of systems biology and medicine, allowing to generate deeper mechanistic insights into biological pathways and molecular functions. However, conventional bulk RNA-sequencing results in the analysis of an averaged signal of many input cells, which are homogenised during the experimental procedure. Hence, those insights represent only a coarse-grained picture, potentially missing information from rare or unidentified cell types. Allowing for an unprecedented level of resolution, single-cell transcriptomics may help to identify and characterise new cell types, unravel developmental trajectories, and facilitate inference of cell type-specific networks. Besides all these tempting promises, there is one main limitation that currently hampers many downstream tasks: single-cell RNA-sequencing data is characterised by a high degree of sparsity. Due to this limitation, no reliable network inference tools allowed to disentangle the hidden information in the single-cell data.

Single-cell correlation networks likely hold previously masked information and could allow inferring new insights into cell type-specific networks. To harness the potential of single-cell transcriptomics data, this dissertation sought to evaluate the influence of data dropout on network inference and how this might be alleviated. However, two premisses must be met to fulfil the promise of cell type-specific networks: (I) cell type annotation and (II) reliable network inference. Since any experimentally generated scRNA-seq data is associated with an unknown degree of dropout, a benchmarking framework was set up using a synthetic gold data set, which was subsequently affected with different defined degrees of dropout. Aiming to *desparsify* the dropout-afflicted data, the influence of various imputations tools on the network structure was further evaluated. The results highlighted that for moderate dropout levels, a deep count autoencoder (`DCA`) was able to outperform the other tools and the unimputed data. To fulfil the premiss of cell type annotation, the impact of data imputation on cell-cell correlations was investigated using a human retina organoid data set. The results highlighted

that no imputation tool intervened with cell cluster annotation.

Based on the encouraging results of the benchmarking analysis, a *window of opportunity* was identified, which allowed for meaningful network inference from imputed single-cell RNA-seq data. Therefore, the inference of cell type-specific networks subsequent to `DCA`-imputation was evaluated in a human retina organoid data set. To understand the differences and commonalities of cell type-specific networks, those were analysed for cones and rods, two closely related photoreceptor cell types of the retina. Comparing the importance of marker genes for rods and cones between their respective cell type-specific networks exhibited that these genes were of high importance, i.e. had hub-gene-like properties in one module of the corresponding network but were of less importance in the opposing network. Furthermore, it was analysed how many hub genes in general preserved their status across cell type-specific networks and whether they associate with similar or diverging sub-networks. While a set of preserved hub genes was identified, a few were linked to completely different network structures. One candidate was EIF4EBP1, a eukaryotic translation initiation factor binding protein, which is associated with a retinal pathology called age-related macular degeneration (AMD). These results suggest that given very defined prerequisites, data imputation via `DCA` can indeed facilitate cell type-specific network inference, delivering promising biological insights.

Referring back to AMD, a major cause for the loss of central vision in patients older than 65, neither the defined mechanisms of pathogenesis nor treatment options are at hand. However, light can be shed on this disease through the employment of organoid model systems since they resemble the *in vivo* organ composition while reducing its complexity and ethical concerns. Therefore, a recently developed human retina organoid system (HRO) was investigated using the single-cell toolbox to evaluate whether it provides a useful base to study the defined effects on the onset and progression of AMD in the future. In particular, different workflows for a robust and in-depth annotation of cell types were used, including literature-based and transfer learning approaches. These allowed to state that the organoid system may reproduce hallmarks of a more central retina, which is an important determinant of AMD pathogenesis. Also, using trajectory analysis, it could be detected that the organoids in part reproduce major developmental hallmarks of the retina, but that different HRO samples exhibited developmental differences that point at different degrees of maturation. Altogether, this analysis allowed to deeply characterise a human retinal organoid system, which revealed *in vivo*-like outcomes and features as pinpointing discrepancies. These results could be used to refine culture conditions during the organoid differentiation to optimise its utility as a disease model.

In summary, this dissertation describes a workflow that, in contrast to the current state of the art in the literature enables the inference of cell type-specific gene regulatory networks. The thesis illustrated that such networks indeed differ even between closely related cells. Thus, single-cell transcriptomics can yield unprecedented insights into so far not understood cell regulatory principles, particularly rare cell types that are so far hardly reflected in bulk-derived RNA-seq data.

Acknowledgements

First, I want to thank my supervisors Peter F. Stadler, Jörg Hackermüller and Sebastian Canzler for their input, constant support, lively discussions, and the opportunity to work on this upcoming, diverse research field.

Next, I would like to thank all my colleagues and friends at the 'Young Investigators Group Bioinformatics & Transcriptomics', for their scientific advice and support. Especially to Stefan Krämer, with whom I worked since our starting days of the PhD in 2018. Special thanks also go to Jana Schor for her efforts in teaching me the importance of meaningful and easy data visualizations, and early morning swim sessions. Likewise, I want to express my gratitude again to Sebastian, without whom this dissertation would not have been successful. Furthermore, to Stephan Schreiber, Matthias Bernt, and Ali Yazbeck.

It was a pleasure being a member of this group.

Additionally, I would like to thank my friends, especially Constanze and Christine who cheered me up and fueled me when needed. During my PhD in Leipzig, I got the chance to meet many people, from whom I can truly call some friends. Thanks to Leo, Daniel and Kathleen for nice weekend trips, the overuse of unnecessary meme quotes, unforgettable cooking sessions and weekend strolls through the city.

Finally, I am extremely thankful to my parents, my brother, and the rest of my family, who support me from the very beginning of my (scientific) career and accompanied every single step of this journey.

Thank you for always believing in me!

This thesis is based on the following publications:

Steinheuer, L. M., Canzler, S. & Hackermüller (2021).
Benchmarking scRNA-seq imputation tools with respect to network inference highlights deficits in performance at high levels of sparsity. In preparation. doi: https://doi.org/10.1101/2021.04.02.438193 .

Völkner, M. , Wagner, F. , Steinheuer, L. M., Carido, M. , Kurth, T. , Yazbeck, A. , Schor J. , Wieneke, S. , Ebner, L. , Del Toro Runzer, C. , Taborsky , D. , Zoschke, K. , Vogt, M. , Canzler, S. , Hermann, A., Khattak, S. , Hackermüller, J. & Karl, M. O. (2021).
HBEGF-TNF induces a complex outer retinal atrophy with macular degeneration hallmarks in human organoids. In revision.

Glossary

`CaSTLe` Classification of single cells by transfer learning. 72

`WGCNA` Weighted Gene Correlation Network Analysis. 13, 33

AE Autoencoder. 12

AMD Age-related macular degeneration. 20, 71

cDNA complementary DNA. 4

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats. 20

DNA Deoxyribonucleic acid. 3

E Expectation. 11

FBS Fetal bovine serum. 20

GOIs Genes-of-interest. 72

HRO human retina organoid. 71

iPSC induced-pluripotent stem cells. 20

LASSO Least absolute shrinkage and selection operator. 13

log2-FC log-2 Fold Change. 38

M Maximazation. 11

MCC Matthew's Correlation Coefficient. 42

MG Müller Glia. 44

MI Mutual information. 17

miRNA micro RNA. 4, 6

MM Module membership. 16

mRNA messenger RNA. 4

MSC Mean Silhouette Coefficient. 44

ncRNA non-coding RNA. 4

PCA Principal components analysis. 10

piRNA Piwi-interacting RNA. 6

RNA Ribonucleic acid. 3

RNA-seq ribonucleic acid sequencing. 6

RNN Recurrent neural network. 12

RPE retinal pigment epithelium. 21

scRNA-seq single-cell ribonucleic acid sequencing. 8

SD Standard Deviation. 42

siRNA short interfering RNA. 6

snRNA small nuclear RNA. 4

SSL Semi-supervised learning. 12

t-SNE t- distributed Stochastic Neighbor Embedding. 10, 44

TOM Topological Overlap Matrix. 16, 42

tRNA transfer RNA. 4

UMAP Uniform Manifold Approximation and Projection. 10, 72

ZINB Zero-inflated negative binomial. 12

Chapter 1

Biological Introduction

1.1 Introduction into Transcriptomics

This chapter will provide and present the fundamental knowledge and theories used for this research. Starting with the essential processes in gene expression and its regulation, which are quantified and analysed in transcriptomic studies, the development to the most recent technological breakthrough of single-cell analysis will be described. Spanning a range of promising analysis tools and methods, finally, the theory of gene co-regulation networks will be introduced. Apart from technical and theoretical aspects, the disease pattern of age-related macular degeneration will be characterized. Since the defined pathophysiological cascades leading to a loss of central vision remain unknown, single-cell transcriptomic approaches might shed a light on this.

The thirst for knowledge has driven human discoveries since the earliest times. Not only Gregor Johann Mendel has proven that the field of biology represents no exception to this phenomenon. While the monk Mendel focussed on the inheritance of phenotypic features in the midst of the 19th century[Miko 2008], it was not until a whole century later that scientists were able to extract genetic information from the cell's nucleus [Dairawan & Shetty 2020]. With the help of radiation, the deoxyribonucleic acid (DNA) was firstly 'photographed' in 1952 by Rosalind Franklin [Franklin & Gosling 1953]. A final deciphering of the double-stranded DNA was achieved in 1953 by Francis Crick and James Watson [Watson & Crick 1953], which was awarded the Nobel Prize of medicine in 1962. Alongside, another genetic sequence was discovered and characterized, which was the single-stranded messenger ribonucleic acid (RNA)[Cobb 2015]. In 1902, scientists already discussed that proteins consist of a sequence of amino acids that are connected via peptide bonds [Pevsner 2015].

However, the clear connection between DNA, RNA, and proteins was not discussed and stated until 1970 the *central dogma of molecular biology* was published by Crick [1970]. It describes the flow of genetic information from DNA over RNA to finally protein (see Figure 1.1).

As the first step of protein-biosynthesis, the cell transcribes certain DNA sequences into RNA molecules. Generally, two different classes of RNA's are present: messenger RNA (mRNA) and non-coding RNA (ncRNA). As the name implies, ncRNA's do not contain information about coding sequences, nevertheless, they fulfil important roles in splicing, gene expression regulation and translation [Mattick & Makunin 2006]. Common subtypes of the ncRNA family are for example micro RNA (miRNA), small nuclear RNA (snRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Both later types, tRNA and rRNA are crucial players during the translation process [Cooper 2000]. Whereby two types of rRNA represent one ribosome, tRNAs contain different triples of so-called anticodons. Depending on the anticodon sequence, one amino acid is loaded per tRNA.

The mRNA is used as an intermediate product to transport coding information from the DNA to the ribosomes. After a set of post-transcriptional modifications including polyadenylation at the 3' end, the mRNA serves as an input for translation.

During this process, complementary mRNA-tRNA triplets (called codons and anti-codons) are matched, and the respective amino acids are attached via peptide bonds to finally assemble the protein.

This well-oiled machinery is subjected to alterations in distinct developmental conditions, tissues, or cell types, meaning that certain sets or proteins are translated at different amounts and frequencies. However, this cascade can also be disturbed by external stimuli such as environmental changes or exposure to toxins. The quantitative differences of this alteration can be approximated via the amount of mRNA at a specific time point. Depending on the relative change of the amount of mRNA of one gene, the transcription of this gene is either up- or downregulated in comparison to normal conditions.

1.1.1 Basis of transcriptomics

Based on the aforementioned biological processes, a whole discipline in biology has evolved, focussing on the analysis of these transcriptomic adaptions. All transcripts from one condition or experiment, called the transcriptome, are quantified and analysed to gain a complete view. This information can help to uncover complex developmental processes as well as disease progressions.

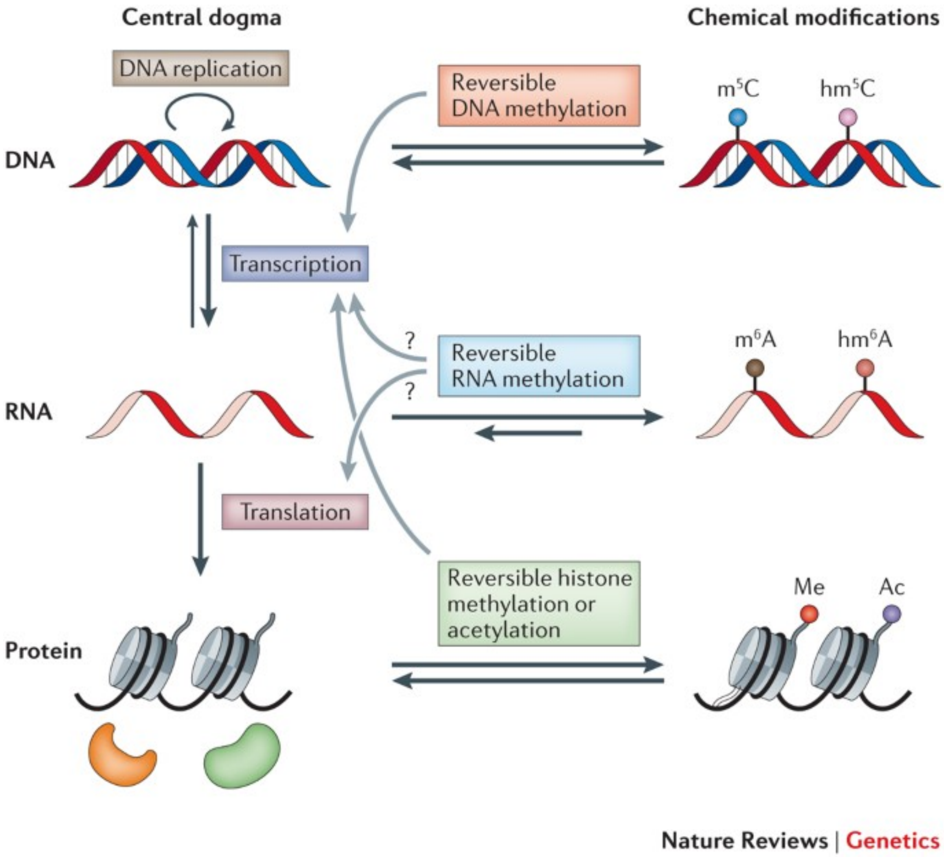

Central dogma
Chemical modifications
DNA replication
Reversible DNA methylation
m⁵C
hm⁵C
DNA
Transcription
?
Reversible RNA methylation
?
m⁶A
hm⁶A
RNA
Translation
Reversible histone methylation or acetylation
Me
Ac
Protein
Nature Reviews | Genetics

Figure 1.1. *Central Dogma of Molecular Biology, taken from Fu et al.[Fu* et al. *2014]. As stated by Francis Crick in 1970, the central dogma of molecular biology describes the flow of genetic information from DNA over RNA to proteins. Once reaching the protein level, the information cannot be re-translated into RNA.*

The strength of a transcriptomic analysis was first demonstrated in the 1970s. There, the full transcriptome of an MS2 bacteriophage was sequenced by Walter Fiers and his colleagues [Fiers *et al.* 1976]. However, since RNA is single-stranded and a variety of RNases are present, it is delicate to handle. With the discovery of reverse transcription*, unstable RNA could be converted to more stable complementary DNA, called cDNA [Perevozchikov *et al.* 1973]. Based on this rationale, different experimental set-ups were established over the past decades to facilitate a broad, quick, and cost-efficient transcriptome analysis. Generally, two major approaches are currently utilised being either hybridization- or sequence-based.

---

*The principle of reverse transcription describes the transformation from the single-stranded RNA molecule back to a double-stranded DNA molecule. Retroviruses use this approach to hijack the host cell's transcription machinery[JM & H 2016].

Hybridization-based techniques aim at quantifying mRNA by incubating labelled cDNA on microarrays. This approach has many advantages such as the high throughput, cost efficiency, and the ability to customize the arrays. However, the inferred results were biased to the limited 'search-space' and sensitivity [Shendure 2008]. Next to this, high noise background levels due to cross-hybridization represent another limitation.

Instead of looking for a predefined, limited set of known genes, sequence-based methods infer the cDNA sequence directly. Initially, the sequencing was done via the cost-intensive and low throughput Sanger chemistry [Sanger *et al.* 1977, Wang *et al.* 2009]. But it was not until the development of high throughput DNA sequencing techniques, for example by Illumina or 454, and the availability of appropriate computational methods that RNA sequencing (RNA-seq) became more frequently used [Ozsolak & Milos 2011]. Being able to quantify RNA transcripts in an un-targeted and high throughput manner fueled the comprehensive and systematic analysis of gene expression patterns through, e.g., the identification of single nucleotide polymorphisms (SNPs), the detection of unknown splice variants, or the characterization of whole non-model organism transcriptomes. While short-read sequencing techniques (such as Illuminas NovaSeq, HiSeq, NextSeq, and MiSeq instruments) generally generate sequences of up to 600 bases, long-read sequencing techniques reach more than 10 kb [Amarasinghe *et al.* 2020]. Depending on the sequencing platform, different quantities of reads can be generated per sequencing experiment. Illumina, for example, can currently reach between 4 million and 1.1 billion reads per run[*], depending on the sequencing device and budget.

1.1.2 Workflow and challenges of next-generation sequencing

In brief, the workflow of a next-generation sequencing experiment looks as follows: the source material is a mixture of different RNA molecules that are extracted from a population of lyzed cells. To increase the sensitivity of the RNA-seq experiments, mRNA molecules can be enriched by selecting their polyadenylated tails [Lowe *et al.* 2017]. To allow for DNA sequencing in a high throughput manner, fragile mRNA molecules are translated into more stable cDNA in a reverse transcription reaction. Usually, larger cDNA strands are fragmented into smaller ones by DNase digestion or sonication. Finally, these cDNA fragments will be ligated to sequencing adaptor sequences which, for example, contain unique barcodes allowing to discriminate different samples, before being amplified by PCR [Parekh *et al.* 2016].

Before sequencing the cDNA libraries, the input must be submitted to several 'manipulation'

---

[*] `https://emea.illumina.com/systems/sequencing-platforms.html`

steps. Whereas smaller RNA molecules, such as microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs) and short interfering RNAs (siRNAs) can be directly sequenced, other larger RNA strands must be fragmented [Wang *et al.* 2009]. Performing this step at either the RNA or cDNA level can introduce small to stronger bias because cDNA fragmentation strongly influences the identification of the transcript's 3' end. Another problem arises from the tendency to over-amplify short cDNA sequences prior to sequencing, which can be mistaken for highly abundant RNA.

Apart from the experimental difficulties, some challenges are faced in the bioinformatic context. Since generally larger data sets are generated during RNA sequencing, it requires efficient ways to access, retrieve, analyze, and store the data. To sort these numerous RNA sequences, they must be either mapped to a reference sequence or can be assembled *de novo*. Both tasks are especially challenging for very small reads. Other problems can arise from sequencing errors or repetitive sequences. However, those problems can be avoided or reduced by using deeper coverage and better reference sequences, respectively. Defining the optimal coverage for a sequencing experiment is a frequent problem and demands to specify a trade-off between the insights that want to be gained and the financial costs it will raise. Being able to detect lowly-expressed genes at a stable level or to analyse larger and more complex transcriptomes generally requires a higher coverage, and hence result in higher sequencing costs.

While high throughput RNA sequencing *principally* allowed for a deeper look into the transcriptional dynamics under certain conditions, it though faces certain limitations on projecting these results onto fine-grained contexts, like highly heterogeneous tissues or disease-related analysis. As explained before, to be able to extract a sufficient amount of RNA, a population of cells has to be used as an input. But this bears the risk that solely the signal from a mixed population can be inferred, and hence rare cell types or even mutated cells are eventually overlooked or *flattened out*. In general, this approach is referred to as bulk RNA-seq.

These limitations can nowadays be overcome with the possibility to extract and analyse the transcriptional information from individual cells.

1.2 The Emergence of the Single-Cell World

With the emergence of the single-cell sequencing methodology, the previously described limitations can be overcome. Instead of using a mixture of cells, solely the genetic information of a single cell is isolated and analysed. By this experimental setup, cell(type) specific biological insights can be gained. While being initially limited in cell throughput, the emergence of droplet-based approaches opened the door for extensive single-cell characterization of transcriptomes. Through clustering these cells by their gene expression pattern, groups of similar cells can be identified and annotated to a specific cell type, using for example gene-of-interest (GOI) lists or automatic pipelines [Lieberman *et al.* 2018, Tan & Cahan 2019]. Due to the increased resolution, rare as well as new cell types can be identified and characterized [Plassschaert *et al.* 2018, Keren-Shaul *et al.* 2017]. An extensive explanation of the methods used in this dissertation will be provided in Chapter two.

Though single-cell transcriptomics offers solutions to several previously identified issues in bulk RNA-seq, there are also several challenges and limitations within the method itself. Similar to bulk RNA sequencing, large data sets are produced during the sequencing procedure, however, the dimensionality on the cell level is massively increased. Since some high-throughput platforms allow the analysis of hundreds of thousands of cells, huge data sets are being generated, and thus raising the bars for an efficient and optimized processing, storage, and analysis even further. Though the resolution on cell level is increased in single-cell RNA sequencing (scRNA-seq), only a very limited biological source material is available which results in large amounts of zero entries in the expression matrix. Generally, two types of zero values are encountered in scRNA-seq datasets [Lähnemann *et al.* 2020]. They can either represent a truly absent count or be introduced methodologically where a transcript was expressed but not measured. This second type of zero, referred to as sparsity, can therefore arise from two different scenarios: either systematically, such as mRNA degradation during cell lysis, or by chance where a barely expressed transcript is measured in one cell but not in the other though present in both [Kharchenko *et al.* 2014]. The latter phenomenon is also denoted as 'dropout'. Therefore, single-cell RNA sequencing (scRNA-seq) data is big but sparse by definition. In addition, single-cell data owns higher degrees of variability and technical noise compared to bulk RNA-seq data [Chen & Mar 2018].

Another major challenge in data analysis lies in the recency of this research field. Apart from missing 'gold standard' workflows and extensive method development, certain domains such as gene-correlation network inference are only about to become investigated [Luecken & Theis 2019].

1.2.1 Experimental workflow

The central part in data generation represents single-cell isolation and (sufficient) mRNA extraction. Therefore, many efforts in method development were dedicated to these aspects. While relying on manual single-cell isolation and generation steps, only a magnitude of ten to a hundred cells were analysable. As discussed earlier, despite the low throughput, the data quality was generally high and less affected by dropout. Labelling individual cells and subsequent isolation of mRNA, promoted the development of other (closed) systems [Aldridge & Teichmann 2020]. An overview of the different methodologies can be seen in Figure 1.2.

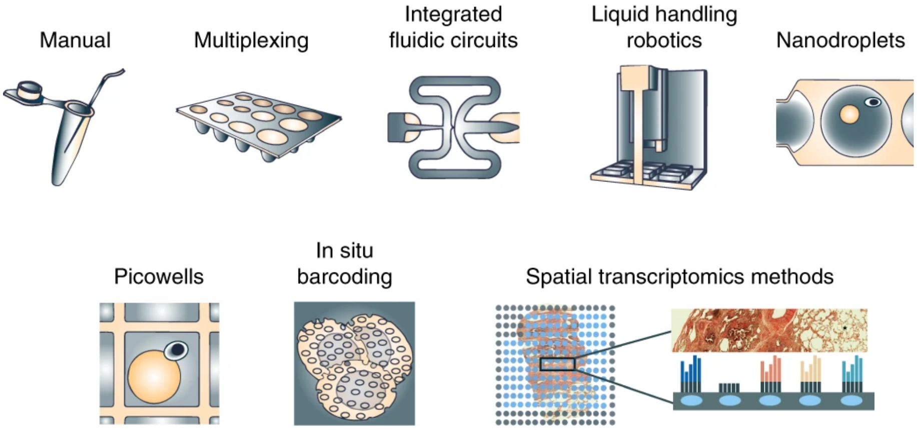


Figure 1.2. *Overview of single-cell preparation methods, taken from Aldridge & Teichmann [2020].*
While single-cells were initially extracted via manual workflows, the upcoming high throughput systems, such as fluidic circuits or robotics, increased the throughput drastically. The most recent boost in throughput was initialised by the employment of nanodroplets, as they are used in 10X single-cell sequencing [Zheng et al. *2017]. Other approaches such as picowells and in situ barcoding do not focus on throughput but quality of the expression data (less sparsity). The latter and most recent approach allows adding spatial information to the expression data.*

It was not until the employment of integrated fluidic circuits and liquid handling robotics that the throughput on cell level was increased to several thousand. Systems such as 10X [Zheng *et al.* 2017] or Seq-Well [Gierahn *et al.* 2017] rely on nanodroplets or picowells, respectively, increasing the throughput further. Very recent approaches such as *in situ* barcoding or spatial transcriptomic methods allow analysing even more cells or retrieve additional information about the spatial location, next to the transcriptome, respectively.

Over the last decade, improving data generation methods and techniques facilitated the generation of enormous datasets, including thousands of individual cells. However, data dropout and the sheer size highlight major deficits on the computational side.

1.2.2 Single-cell tools and applications

Pipelines. single-cell transcriptomics data offers a novel level of resolution and hence promoting the development of appropriate analysis tools. By the end of 2020, 23 different tool categories were listed on the `scrna-tools`[*] website, ranging from mainly data visualization and cell clustering to imputation and simulation. Combining more than one tool or task, complete data analysis pipelines, such as `Seurat` [Butler *et al.* 2018], `Scater` [McCarthy *et al.* 2017], and `scanpy` [Wolf *et al.* 2018], were developed to enable an end-to-end solution from data preprocessing to analysis.

The first analysis pipeline, `Scater`is implemented in `R` and was originally published in 2017. Both, `Seurat` and `scanpy`, were developed and released in 2018 while building onto the `Scater`framework. Whereby `Seurat` is also implemented in `R`, storing the data as a `Seurat`-object, `scanpy` is up to now the only pipeline implemented in `python`. This mainly allows for a speed-up in data loading, processing and visualization [Luecken & Theis 2019]. Common steps which can be found across platforms are for example quality control, expression data normalization, identification of highly variable genes, data scaling, dimensionality reduction and an initial data visualization. For the last two steps, usually, a principal component analysis (PCA) is conducted and the cells are either embedded in a t-distributed Stochastic Neighbor Embedding (t-SNE) or a Uniform Manifold Approximation and Projection (UMAP). Whereby both tools reliably capture the structures of the data, UMAP runs faster which is especially useful for growing single-cell data sets [Mcinnes *et al.* 2020]. Though comparably younger than t-SNE, UMAP proofed its suitability over the past years.

In the following section, an excerpt of common applications and tools in the field of single-cell transcriptomics is introduced.

Cell clustering and annotation. To analyse the single-cell data, usually, the cells are projected into a lower dimensional space using UMAP or t-SNE. The following step, cell clusters can be defined based on the *unique* gene expression pattern of the cells. Using *a priori* knowledge, e.g., GOI lists or marker genes, the cell clusters can be annotated to specific cell types [Kim *et al.* 2019, Cowan *et al.* 2020]. Since this approach involves a series of subjective,

---

[*] `https://www.scrna-tools.org/`

unreproducible decisions, other tools try to utilise the results from previously annotated, reference data sets. These *transfer learning* tools, try to learn patterns from a reference data set and use these to classify the test data. `SingleR`[Aran *et al.* 2019] for example allows using bulk information as a reference dataset. Another transfer learning tool used in this thesis is `CaSTLe` which uses a random forest-based classification algorithm [Lieberman *et al.* 2018].

Pseudotemporal ordering. Another goal is to resolve developmental pathways, which can elucidate when and where progenitor cells commit to defined, mature cell types [Wolf *et al.* 2019, Herring *et al.* 2018]. Since cells are disrupted during the sequencing procedure, direct tracking over a specific time period is impossible. However, various *snapshots* are collected, which differ slightly. Inferring these differences allows ordering the sequenced cells along a pseudotime axis ranging from the earliest time point to the most mature captured in the experiment.

One task which is related to this field is the inference and quantification of RNA velocity [La Manno *et al.* 2018, Bergen *et al.* 2020]. Briefly, using the dynamics between spliced and unspliced transcripts allows calculating the transition probability between two cells.

To allow for meaningful visualization of these cell developmental trajectories, many efforts were undertaken in implementing novel tools. One of them is the partition-based graph abstraction algorithm called `PAGA`, which provides a graph-like map of the single-cell data manifold [Wolf *et al.* 2019]. Without prior information about the developmental tree[*], `PAGA` calculates a coarse-grained single-cell embedding with connections between the cell types. Incorporating the information from the RNA velocity analysis, also the directionality of these connections can be stated.

Imputations. The huge fraction of data sparsity also affects the aforementioned applications on single-cell data. One potential direction to improve the efficacy of single-cell analysis tools, however, is to leverage the data sparsity through imputation approaches intending to 'interpolate' the previously missed expression. In the past years, many single-cell specific imputation tools have already been published [Huang *et al.* 2018, Gong *et al.* 2018, Peng *et al.* 2019a] and their evaluation showed that denoising sparse simulated data can, for example, help to reobtain original cell clusters and time-course patterns [Eraslan *et al.* 2019]. These tools aim at inferring the zero or NA entries by using different mathematical assumptions or paradigms, such as repeated clustering or the use of autoencoder networks.

---

[*] The developmental tree refers to the general transitions of cell types. Based on previous experiments, this layout is often known and can be used as input for trajectory inference.

Because of the rapid increase of published imputation methods, several review articles [Lähnemann *et al.* 2020, Chen *et al.* 2019] and benchmarking analysis [Zhang & Zhang 2018, Patruno *et al.* 2020] also investigated specific fields within the downstream analysis realm, such as differential gene expression [Hou *et al.* 2020].

While Hou *et al.* focussed on the influence of imputations on differential gene expression, unsupervised clustering and pseudotime trajectory analysis, their research highlighted that results of imputation approaches should be regarded with caution, since the majority did not outperform the unimputed data regarding downstream tasks. Similar results were stated by Chen and Mar for network inference tools on sparse scRNA-seq data [Chen *et al.* 2019].

1.3 From Counts to Co-regulation - Gene Networks

Using potent methods such as (single) RNA sequencing, which allows quantifying transcripts in a high throughput manner across different conditions, developmental stages and genetic backgrounds, can help in deciphering complex gene regulation systems. Since genes do not act *independently*, it is more useful to identify larger gene regulatory networks. Pinpointing these orchestrated expression programs or responses is a major discipline in transcriptomics. These networks might help in deepening the understanding of gene function, biological processes and complex disease mechanisms.

1.3.1 Gene correlation networks

One way to proxy this co-regulation is via the concept of gene expression correlation. A schematic overview of the different steps in gene network construction is provided in Figure 1.3. Here, the main tasks include network reconstruction, gene module identification and the assessment of biological functions or tasks. This includes for example the functional enrichment of biological functions per module, differential network analysis and finding hub genes. The gathered information could be used to pinpoint potential disease genes with a module. In this dissertation, the Weighted Gene Correlation Network Analysis tool (`WGCNA`) was used which will be explained in detail in Section 2.4.1 [Langfelder & Horvath 2008].

Information-theoretic models also rely on the pairwise correlation coefficients. However, they use a generalization which is termed mutual information (MI) [Steuer *et al.* 2002, Bansal *et al.* 2007]. The MI measures the statistical dependency between two variables. Some representative tools using information theory are for example `ARACNE` [Margolin *et al.* 2006], `RelNet` [Butte & Kohane 2000] and `CLR` [Faith *et al.* 2007]

These system-level assignments of genes via correlation networks can be useful, since they can, for example, aid in predicting the function of unknown genes or RNA sequences.

1.3.2 Gene networks in single-cell transcriptomics

Network inference approaches, such as `WGCNA` already proofed their worth in various RNA-sequencing studies [Yang *et al.* 2014, Kogelman *et al.* 2014, Lee *et al.* 2011]. With the ability to infer gene expression data from single cells, also cell type-specific gene correlation networks can be inferred, as indicated in Figure 1.4. Early `WGCNA`-based approaches also allowed new insights

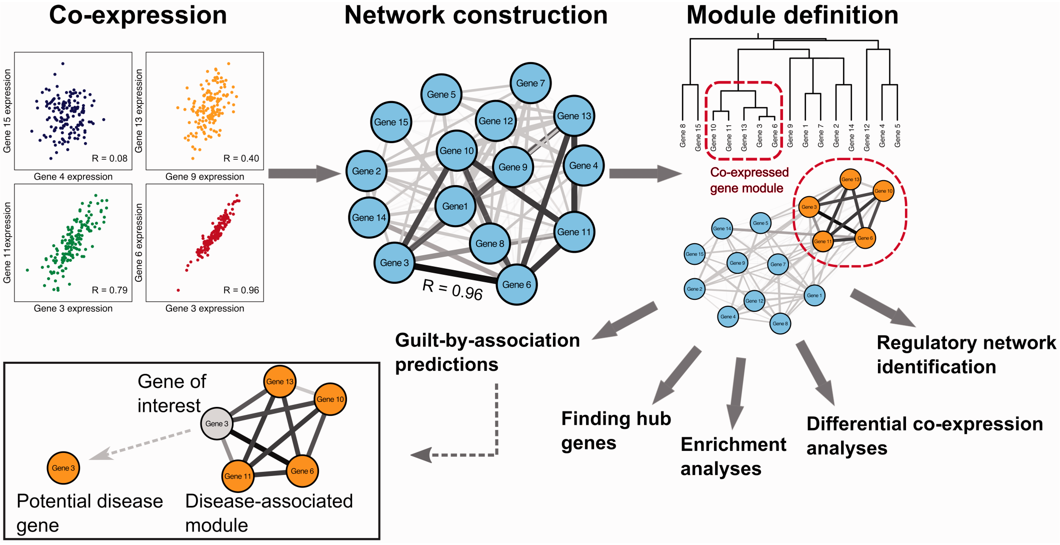


Figure 1.3. *Overview of individual tasks during gene network inference, taken from van Dam* et al. *[2018].*

Using the gene co-expression data, pairwise gene correlation values are calculated. Transforming these values into interconnectedness measures, a (weighted) network can be extracted and groups of densely connected genes, called modules, can be identified. Various downstream tasks can be applied to this higher level assignment, such as hub gene detection and gene set enrichment.

into genetic programs of low throughput single-cell transcriptomic approaches [Luo *et al.* 2015, Xue *et al.* 2013]. With the emergence of higher throughput methods, these bulk-derived, non-single-cell specific tools showed only mediocre results [Chen & Mar 2018]. Subsequent developments of highly specific single-cell tools, however, were also not able to transform the unique characteristics of this data into the desired outcome [Chen & Mar 2018]. More recent work benchmarked single-cell specific tools using curated reference models [Pratapa *et al.* 2020, Nguyen *et al.* 2020]. Including 12 algorithms as well as three different sources of reference data sets, `BEELINE`, for example, offers an evaluation framework to facilitate the development of novel tools [Pratapa *et al.* 2020]. While testing conditions were improved, network inference remains a challenging problem. Although algorithms such as `SCENIC` or `CellOracle` [Aibar *et al.* 2017, Kamimoto *et al.* 2020] have been proven to be meaningful, they calculate the network inference on a restricted search space by subsetting the number of investigated genes, and thus still neglect an overall and unbiased picture.

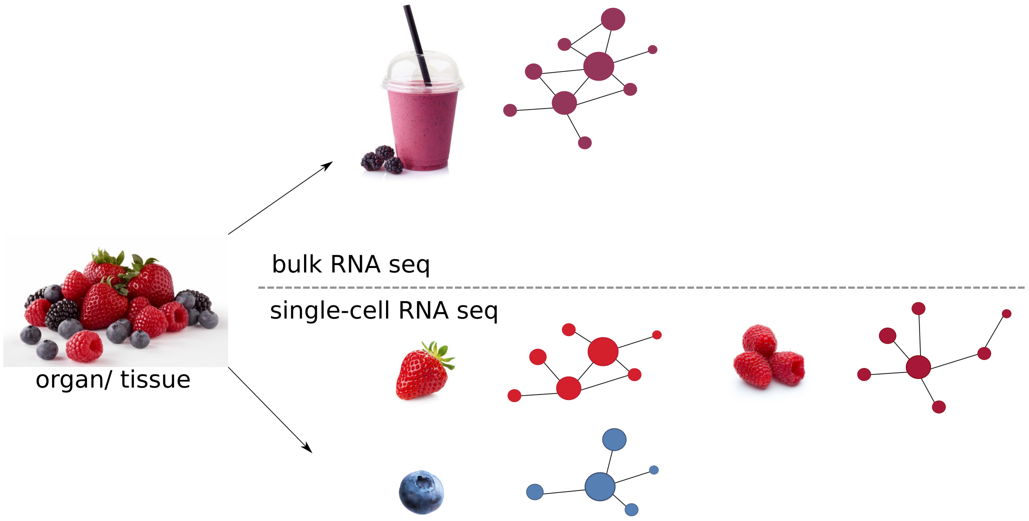


Figure 1.4. *Differences in resolution from bulk and single-cell RNA-seq data on the level of gene networks.*
Illustrating the input of sequencing experiments as various fruits (left) would result in a homogenized, mixed source material for bulk RNA-seq data, much like a fruit smoothie (top). Inferred networks would therefore reflect an average of all the signals detected in the mixture of, potentially different, cells. Single-cell RNA-seq on the other hand would preserve the information of the individual fruits, hence different cell types, and would allow for cell type-specific gene correlation networks (bottom). Thus, similarities, as well as differences in networks between cell types, could be carved out that would otherwise be hidden or covered in the bulk RNA-seq derived network.

1.4 Example use case: Study of age-related macular degeneration using an organoid system

In the previous chapters, a variety of single-cell transcriptomic applications such as cell annotation and trajectory inference were introduced. Here, these potential tools were applied in a real-world scenario, more specifically a novel human retinal organoid. The following section will provide background information about the disease pattern of age-related macular degeneration (AMD) as well as the history of model systems.

1.4.1 Characteristics of AMD

Being able to receive and process visual stimuli in light and colour is one central feature of the human eye. While visual impairments (near and farsightedness) are becoming more common, the same trend can be stated for complete loss of vision. Apart from the formation of cataracts or diabetic retinopathy, AMD is one major cause for the loss of central vision [Al-Zamil & Yassin 2017]. Currently, AMD affects 10% of people older than sixty-five years, and 25% older than seventy-five. It is estimated that by 2040, more than 288 million people will be affected by AMD [Wong *et al.* 2014]. Apart from age, other major risk factors are family history and other genetic factors [Al-Zamil & Yassin 2017, Klein *et al.* 2004]. Caucasians, Hispanics, and Asians are known to have the highest risk for developing AMD, while African Americans are the least affected. Minor risk factors are lifestyle, diet, and nutrition.

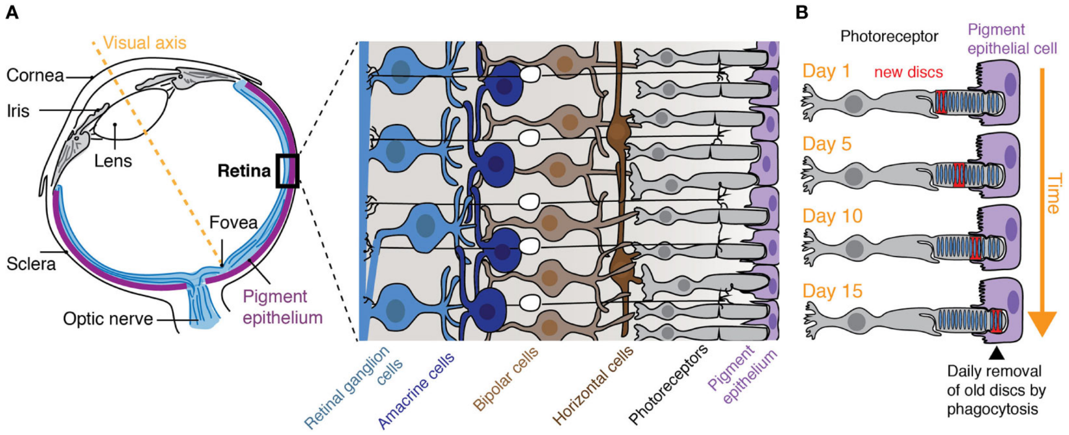


Figure 1.5. *Composition of the human retina, Figure taken from Kwon & Freeman [2020]. (A) A schematic overview of the cross-section of a human eye (left hand side) and the different cell types from inner to the outer layer (right hand side). While the retinal ganglion cells are the first ones receiving the light, it is the most inner photoreceptors which can process its information. (B) The maintenance of those photoreceptors is operated by the retinal pigment epithelium through continuous removal of old discs. These discs contain phototoxically damaged opsins.*

1.4.2 Composition of the human retina

The human retina is a complex system allowing for the reception of light stimuli, which are processed into colourful pictures of our surroundings [Al-Zamil & Yassin 2017]. As depicted in

Figure 1.5(A), many cell types are represented in the retina. While the actual light stimulus is processed in the photoreceptors in the outer layer, the retinal neuron cells (horizontal, amacrine and bipolar cells) are adjacent to them. In total, two different photoreceptor cell types are represented, denoted as rods and cones, that can detect changes in light intensity [Molday & Moritz 2015]. Whereby cones are responsible for the perception of fine-grained details and colour, rods process information about brightness. This is due to the different opsins contained in the photoreceptors. Whereas rods rely on rhodopsins, cones contain cone opsins [Terakita 2005]. Both subtypes show repeated compartments, so-called discs, which contains the cells metabolic machinery as well as the previously mentioned opsins [Besharse & Pfenninger 1980]. The photoreceptors are embedded in the retinal pigment epithelium (RPE), which is a non-dividing cell type. Its main function is the maintenance of the photoreceptor cells, which is mainly facilitated by continuous disc removal since opsins are susceptible to phototoxic damage [Kwon & Freeman 2020] (see Figure 1.5(B)).

1.4.3 Disease progression of AMD

With increasing age, an accumulation of focal, yellow, extracellular, and polymorphous material occurs, which is called drusen. Under normal conditions, these cells would be eliminated by the choriocapillaris. The choriocapillaris is a tissue positioned under the retina near Bruch's membrane. Bruch's membrane is located close to the RPE and serves as a molecular sieve [Booij *et al.* 2010]. However, a dysfunction leads to changes in the permeability of Bruch's membrane. The presence of drusen describes a hallmark of AMD along with hypo- and hyper-pigmentation. It is furthermore associated with a thickening of collagenous layers in Bruch's membrane, degeneration of elastin, and calcification. With more accumulation of drusen, the RPE is stepwise lost, which ultimately leads to photoreceptor death and the loss of central vision. This progression is considered a *dry* AMD. A visual representation is shown in Figure 1.6(b). Oppositely, a *wet* AMD involves upregulation of VEGF, the vascular endothelial growth factor, which promotes the abnormal growth of choroidal vessels underneath the RPE (see Figure 1.6(c)). Upon blood vessels bursting, disc-like scars are formed, leading to a permanent loss of vision.

Though the disease progression of AMD is roughly known, the general and individual causes and as well as pathophysiology remains masked. Moreover, due to the human-specific characteristics of AMD, animal models cannot be employed. Using a sufficiently complex yet interpretable system, 3D cell culture models may shed a light on the mechanisms occurring in AMD. In this approach, human cells can be used, providing a more organ-like scenario to

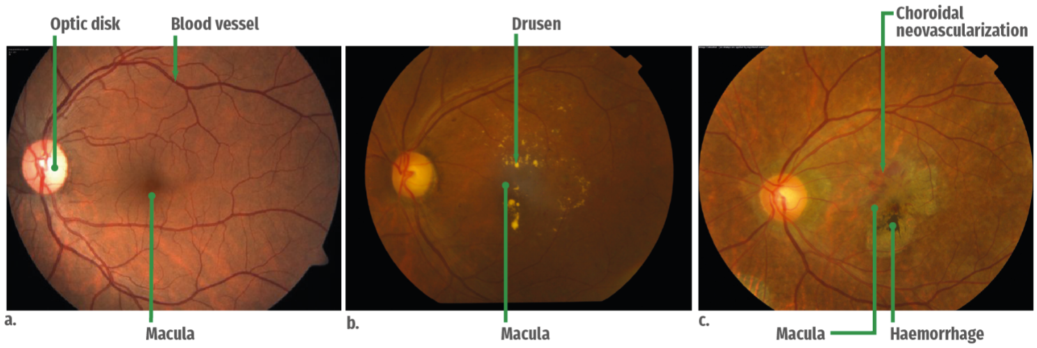


Figure 1.6. *Differences of a healthy eye and different AMD subtypes, taken from Acharya* et al. *[2016].*

(a) In a healthy eye, the macula can be identified as a dark spot in the middle of the eye, with a slight shift to the optical disk. (b) In dry AMD, the formation of drusen, here shown as bright yellow spots on top of the macula, resulting in a loss of central vision due to a photoreceptor loss. (c) Though sharing the same disease outcome, wet AMD is characterised by abnormal choroidal neovascularization and subsequent vessel bursting. These haemorrhages ultimately lead to a loss of vision.

increase the expressiveness and scalability of the results. Alternatively, induced pluripotent stem cells (iPSC)[Ye *et al.* 2013] can be used to allow organ development from scratch. Due to the organ-like structure, the resulting cell collection is called an organoid.

In the past, these organoids proved to be useful, for example, as tubular organoids were used to model kidney disease [Cruz *et al.* 2017]. With the emergence of scRNA-seq, a fine-grained picture of these organoids systems can now be drawn [S *et al.* 2019].

Retinal organoids are thought to help to decipher these questions and, in the recent past, many efforts have already been undertaken to generate explorable and meaningful organoid systems [Kim *et al.* 2019, Cowan *et al.* 2020, Völkner *et al.* 2021]. Despite these compelling prospects, it currently remains unclear whether these organoids resemble an *in vivo* organoid in cell type composition and development and how much variance can be expected between organoids.

1.5 The Research Gap

Summing up the individual introductory chapters, the general prospects of transcriptomic studies were introduced. The inference of gene regulatory networks from conventional bulk RNA-seq approaches has led to important biological insights, for example, in recurrence-associated genes of colon cancer [Zhai *et al.* 2017] or gene architectures associated with Alzheimer's disease [Acquaah-Mensah *et al.* 2015]. With the breakthrough of single-cell genomics, it should now be feasible to analyse an even higher resolution level and infer cell type-specific gene correlation networks. However, limitations due to size and high data sparsity are currently still hampering this task. To alleviate the latter, data imputation methods aim to replace missing entries utilising various modelling strategies.

Therefore, two main questions arise: (I) How do increasing levels of sparsity impact gene correlation network structures, and (II) can data imputation of the sparse data aid in network inference? To furthermore be able to infer cell type-specific gene correlation networks, two premisses must be met such as i) the identification of cell types and (ii) the possibility to infer gene networks.

These questions are tackled in chapter three, using a benchmarking framework, including six different sparsity levels as well as six published imputation tools. As a proof-of-concept, the gained insights were used on an experimentally derived data set. Chapter four evaluated the possibility to derive meaningful cell type-specific gene correlation networks.

Besides network inference, various other tools can be used to generate valuable insights into single-cell data sets. Such knowledge could be used, for example, to describe organoid systems and compare them to the corresponding *in vivo* tissues. This dissertation investigated if a human retina organoid can be characterised via the single-cell toolbox. Ultimately, these organoid systems could aid in deciphering the pathophysiology of age-related macular degeneration, the major cause for loss of central vision. Here, the main questions concern the stability of cell annotation results, the similarity of two different organoid samples, and the correspondence to the *in vivo* organ development.

Chapter five will highlight the workflow as well as the results, aiming to answer the previously mentioned aspects.

Chapter 2

Methodology

From the previous sections, an introduction to the underlying theories and applications was provided, which will become eminent during this dissertation. While giving a rather general overview of *state-of-the-art* methods and tools, the research questions and working assumptions were postulated. In the following section, the employed algorithms and their implementations will be explained in more detail to provide a basis for the upcoming chapters. Following the order of the introductory sections, single cell-specific tools and methods, such as data imputation and pseudotime inference, will be explained at the beginning. The second part will then deal with network inference and different measures to compare network structures.

2.1 Single-cell data Imputation

As demonstrated in Section 1.2.2, the single-cell toolbox allows for a diverse set of applications. Though the potentially high resolution within the data, high amounts of sparsity limit, for example, the inference of (cell type-specific) gene correlation networks. Therefore, the question arises if data imputation could alleviate the sparsity. To perform a systematic evaluation, well-established and representative tools were used in accordance with the four classes defined by Lähnemann *et al.* [2020]: (1) deep-learning-based (`DCA`), (2) smoothing-based (`DrImpute`), (3) model-based (`SAVER`), and (4) low-rank matrix-based (`ENHANCE`)*. Furthermore, an additional class (5) of tools that utilise gene networks (`scNPF`) was created. In the following, six different, single cell-specific imputation tools will be introduced, and their underlying mathematical formulation discussed.

---

*`https://github.com/yanailab/enhance-R`

`DCA` . Since denoising autoencoders (AE) inherently suit the problem of separating true signal from noise, it is not surprising that they are frequently used to impute sparse scRNA-seq data. Here, the deep count autoencoder (`DCA`) is introduced [Eraslan *et al.* 2019]. While using the expression matrix as an input (x), a zero-inflated negative binomial (ZINB) distribution is used to infer the dropout (π), dispersion (Θ) and the mean (μ) (see equation 2.1). Generally, the autoencoder is based on an encoder, a bottleneck and a decoding part. The initial encoder part condenses the input data via reduction of the number of nodes over multiple connected layers, aiming to extract the relevant information from the noisy data. After the bottleneck layer, which represents the narrowest part in the neural network, the decoder part is used to reobtain the original dimensions of the input layer.

$$ZINB(x; \pi, \mu, \Theta) = \pi\delta(x) + (1 - \mu) * NB(x; \mu, \Theta) \tag{2.1}$$

`DISC` . Though imputation via neural networks is generally an unsupervised learning task, another tool called `DISC` combines a deep learning approach with semi-supervised learning (SSL) [He *et al.* 2020]. Model parameters are trained by combining an AE, a recurrent neural network (RNN) and the SSL approach. The AE and the RNN are thereby used to extract a low dimensional latent representation of the cell expression profiles. This compressed matrix z^t represents the basis for predicting the cell expression profile, the predictor matrix (y^t). Simultaneously, the latent space is reconstructed by the decoder layer of the AE to derive the reconstructor ($\hat{y}^t$). The imputation result is inferred via a weighted average of the predictor y_t. Likewise, a weighted average of the reconstructor $\hat{y}^t$ was calculated. Both were used to support the SSL model, which learns the parameter specifications itself in `DISC`. Being able to learn from positive- as well as zero-count genes, `DISC` can search for the best expression structures to preserve the latent data manifold produced by the AE initially.

`DrImpute` . Apart from neural networks, other methods can be used to infer dropout entries. `DrImpute` for example, uses the result from a repeated clustering $C_1, ..., C_H$ and averaging [Gong *et al.* 2018]. As depicted in Equation 2.2, the mean value of the (i,j)th component from expression matrix X can be approached per clustering result C_h. Using the log-transformed expression data, a similarity matrix among the cells is calculated using Pearson and Spearman correlations. Extracting the first 5% of the principal components of this similarity matrix, k-means clustering is performed using between ten and fifteen clusters. In total, twelve different clustering results build the basis for the final imputation step, the average of the individual clustering results.

$$E(_{ij}) = mean(E(x_{ij}|C)) = \frac{1}{H}\sum_{h=1}^{H} E(x_{ij}|C_h) \tag{2.2}$$

`ENHANCE` . Another imputation tool that relies on PCA is `ENHANCE` [Wagner *et al.* 2019]. Using the PCA on variance-stabilised data, the authors hope to separate true biological signals from noise. To reduce the noise level in the source data prior to PCA, a k-nearest neighbour aggregation step is included. The rationale behind this is that technical noise is more likely to be included in higher PCs. Generally, `ENHANCE` can be divided into three phases. In the first phase, the number of significant PCs is determined by calculating the background noise level from a simulated data set owning the same dimension of the analysed data. After that, the second phase aims in generating an aggregated expression matrix. Thereby, the subsequent PCA steps are less biased towards highly expressed genes. Lastly, a PCA is applied to the aggregated expression matrix from the previous step, and the signal from the first PCs is extracted. To match the signal *intensity* of the input data, the expression profiles are scaled.

`SAVER` . Using an empirical Bayes-like approach, `SAVER` models the dropped out expression values via a Poisson LASSO regression of other, predictive genes [Huang *et al.* 2018]. As the first step, `SAVER` uses the count data and models the gene g in cell c (Y_{gc}) using a Poisson-gamma mixture. To estimate the posterior gamma distribution for λ_{gc} given the observed counts Y_{gc}, the authors adopted an empirical Bayes-like approach to estimate the prior mean (μ_{gc}) and variance (ν_{gc}). The number of predictive genes used to infer the missing data is reduced employing a LASSO regression. LASSO adds a penalty parameter λ to the likelihood, controlling for predictors that have nonzero coefficients. The variance is approached assuming a constant noise model, called dispersion (ϕ_g). Additional parameters which were used in the noise model are a constant shape parameter α, a constant Fano factor ϕ_{gc}^{F}, and a constant rate parameter β_{gc}. After approaching $\hat{\mu}_{gc}$ and $\hat{\nu}_{gc}$, the posterior distribution can be stated as seen in equation 2.3. So finally, the posterior mean (see equation 2.4) is used as the recovered, denoised expression values.

$$\lambda_{gc}|Y_{gc}, \hat{\alpha}_{gc}, \hat{\beta}_{gc} \sim Gamma(Y_{gc} + \hat{\alpha}_{gc}, s_c + \hat{\beta}_{gc}) \tag{2.3}$$

$$\hat{\lambda}_{gc} = \frac{Y_{gc} + \hat{\alpha}_{gc}}{s_c + \hat{\beta}_{gc}} = \frac{s_c}{s_c + \hat{\beta}_{gc}}\frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c + \hat{\beta}_{gc}}\hat{\mu}_{gc} \tag{2.4}$$

`scNPF`. Lastly, `scNPF` imputes missing entries based on a network propagation process using a random walk with restart (RWR) [Ye *et al.* 2019]. Using the RWR on a gene-interaction network allows inferring the importance of a gene-based on connectivity measures. This tool can be run using two different modes: using *a priori* knowledge or extracting networks directly from the data. While many external databases contain experimentally proven gene-interaction data, they can be used as an input for the RWR. Alternatively, the weighted gene correlation network analysis tool (`WGCNA`) is used to calculate gene networks.

In order to smooth the expression measurement across the network, an RWR is used as described in equation 2.5. The restart vector P_0 records the initial expression levels, and W represents the degree-normalised adjacency matrix of the network. r describes the trade-off between prior knowledge and network diffusion. This propagation function is run until convergence is achieved.

$$P_{t+1} = rP_0 + (1 - r) * P_t W \tag{2.5}$$

2.2 Low-dimensional embedding and community detection.

In the past years, many efforts were undertaken in generating comprehensive analysis pipelines to read-in, clean up, and visualise single-cell transcriptomics data. Though mainly three different preprocessing pipelines exist, the general aim is shared across them, which is the generation of a low-dimension data embedding. Based on this embedding, groups of similar cell expression profiles should be identified, which can help to assign a biological function to them.

In this dissertation, a principal components analysis (PCA) on the most variable genes was used as a fundament to calculate the neighbourhood graphs. A PCA is an unsupervised dimensionality reduction technique that aims at capturing most of the variance in the data by linear combinations. Based on a fixed set of principal components, the neighbourhood is calculated and embedded in a two-dimensional space. In the python-based pipeline `scanpy`, the Uniform Manifold Approximation and Projection (UMAP) embedding is highly recommended [Wolf *et al.* 2018]. The UMAP algorithm is based on three assumptions: (I) The data is uniformly distributed on the Riemannian manifold, (II) the Riemannian metric is locally constant, and (III) the manifold is locally connected [*]. Generally, these Riemannian manifolds can be considered as extensions of the euclidean space [Vaugon 2006].

Alternatively, also t-Distributed Stochastic Neighbor Embedding (t-SNE) can be used. t-SNE is again an unsupervised data reduction technique that operates non-linearly [Van Der

---

[*] `https://umap-learn.readthedocs.io/en/latest/`

Maaten & Hinton 2008]. In contrast to PCA, t-SNE mainly conserves smaller pairwise distances between cells. Briefly, the t-SNE aims at defining an embedding for the data in a high- and low-dimensional space, while maximising their similarity. This is achieved via the minimization of the Kullback-Liebler divergence (KL) cost function.

Based on this low-dimensional embedding, cell clusters can be inferred, grouping cells with similar expression patterns, which might represent cells of one biological cell type. In this dissertation, the Louvain algorithm was used [Blondel *et al.* 2008]. This algorithm operates in two different phases. Generally, it aims at optimising the modularity, by maximising the differences between the detected number of edges in a community versus the expected. While in the first step, the Louvain algorithm locally moves the nodes, it aggregates them in the second step. This procedure is repeated until the quality measure cannot be improved any further.

2.3 Cell annotation via transfer learning

Based on the low-dimensional cell clustering, biologically relevant cell types could be annotated. However, it is also possible to re-use the cell annotation from other reference data sets, extract their features and classify new data. In this dissertation, the transfer learning tool `CaSTLe` was used, which operates according to the previously mentioned workflow [Lieberman *et al.* 2018]. As a first step, `CaSTLe` selects a set of features, in this case, genes, which will have a high expression in the reference and target data and own high mutual information only in the reference. Then, correlated genes within this subset will be removed and the entries are binned. Finally, a pre-tuned XGBoost classifier is trained on 80% of the randomly selected reference data and evaluated on the remaining 20%. XGBoost, short for *eXtreme Gradient Boosting*, is a supervised machine learning tool, that builds on a collection of regression trees [Chen & Guestrin 2016]. This random-forest based approach aims at minimizing the prediction error and the model complexity. If the classification performance was sufficient, this model can be used to annotate an unseen target data set, cell by cell.

2.4 Pseudotime calculation

Besides data manipulation and low dimensional embedding, other tools also allow deciphering development aspects within the single-cell transcriptomics data. In this dissertation, RNA velocity was used to add a pseudotime variable to the expression data, using the `scVelo` package [Bergen *et al.* 2020]. Generally, RNA velocity is approached via the ratio between unspliced (new) and spliced (mature) mRNA transcripts for all genes with a splicing variant.

In equation 2.6, the splicing dynamics of the unspliced transcript $u(t)$ with state-dependent rate $\alpha^{(k)}$ into mature mRNA $s(t)$ with rate β is shown. Finally, the mature mRNA is degraded via the parameter γ.

$$\phi \xrightarrow{\alpha^{(k)})} u(t) \xrightarrow{\beta} s(t) \xrightarrow{\gamma} \phi \tag{2.6}$$

To learn the transcription dynamics more adaptively, `scVelo` aims at inferring the variables via an Expectation-Maximation approach. The expectation step (E) estimates the unspliced/spliced trajectory and assigns a latent time to the observed mRNA values and transcriptional states. `scVelo` assumes three different states which are on, off, and a steady-state. During the maximization step (M), the likelihoods are updated based on the parameters inferred in the E-step. This procedure is repeated until convergence is reached.

Finally, an RNA velocity can be calculated as shown in equation 2.7. Based on these velocities, cell transition probabilities can be calculated, which allow visualizing dynamics within the low dimensional cell embedding.

$$v(t) = \beta u(t) - \gamma s(t) \tag{2.7}$$

Furthermore, using this information to construct developmental associations between cell-types, `PAGA`, a partition-based graph abstraction, allowed to transform this information. `PAGA` uses the cell cluster annotation and expression pattern to generate a connected graph between cells. In the implementation symmetrized kNN-like graphs were used to conduct the nearest neighbour search within the low-dimensional embedding.

In a second step, the graph is partitioned based on the number of in- and outgoing edges based on the velocity graph that originated from the previously described RNA velocity analysis.

Subsequently, the pseudotime variable can be estimated using the extended version of diffusion pseudotime (DPT). The basic implementation measures a progression through branching lineages via a random-walk-based distance in the diffusion map space [Haghverdi *et al.* 2016]. In `PAGA`, the extension accounts for disconnected Eigen-subspaces in the graph adjacency.

Using this `PAGA`-embedding, the low-dimensional UMAP can be recalculated to fit the `PAGA`-layout, facilitating improved interpretability.

2.4.1 Inference of gene networks via `WGCNA`

While RNA velocity can infer the relationship between cells and cell types, gene network inference can identify the relationship between genes.

There are various mathematical approaches at hand helping to identify these groups of 'connected' genes. One option is to proxy co-regulation through the concept of correlated gene expressions. Genes that are similarly expressed will own a correlation coefficient close to one; or minus one if they are anti-correlated. This can help to identify groups of correlated genes instead of focusing on individual candidates.

In the following section, the concept of gene correlation will be explained using the tool `WGCNA` since it is heavily used in this dissertation [Langfelder & Horvath 2008]. The 'Weighted Gene Correlation Network Analysis' package is implemented in `R`.

Network construction. To state the similarity of genes or groups of genes, a co-expression similarity matrix is calculated. This can be done by common correlation measures like Pearson correlation, but here a more robust* method is applied with the biweight midcorrelation.

As stated by Barabási & Albert [1999], biological networks own a scale-free topology, meaning that only a few nodes will have a high connectivity whereas the majority of nodes reveal a low connectivity. These highly connected nodes represent so-called 'hub genes' which take a central role in their respective gene group (gene module). To achieve this network architecture, correlation values are transformed by increasing them to the power of a β value:

$$a_{ij} = \left(\frac{bicor(x_i, x_j) + 1}{2} \right)^{\beta} \tag{2.8}$$

where x_i and x_j are gene-nodes i and j, respectively. The range of i and j is equal to the number of genes included in the data set, whereby the size of x_i and x_j corresponds to the number of replicates in bulk RNA-seq or the number of cells in scRNA-seq. The optimal β value was determined using the following criteria: First, the scale-free topology model fit across a range of β values was plotted. Here the model fit must be at least 0.8. If that criterion was met, the β value was chosen that lies in the 'elbow-phase' shortly before reaching the plateau phase. If no model fit above 0.8 was reached, the β value, which resulted in the smallest median connectivity above one hundred was used† In cases where two networks were compared to each other, two identical β values have been selected such that both criteria were

---

* The biweight mid-correlation is more robust with respect to outliers.

† `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html`

met as good as possible.

The similarity matrix a, with size $|G|x|G|$ and G being the number of genes, that is based on correlation across all genes raised to the power β is used as an adjacency matrix to construct a so-called weighted gene correlation network.

Gene Module detection. After network construction, groups of densely interconnected genes are inferred. As a basis for that, the row sum of the adjacency matrix is firstly calculated (see equation 2.9). In an unweighted network, the gene connectivity k_i is reflected by the number of direct neighbours, whereby in weighted networks the sum of connections strengths to other nodes is indicated. Network interconnectedness in the `WGCNA` implementation is approached by the topological overlap measure (TOM) [Yip & Horvath 2007] as depicted in Equation 2.10.

Finally, these TOM values are grouped using hierarchical clustering. Depending on the implementation, some post-processing steps such as tree cutting and module merging can be applied.

$$k_i = \sum_{j \in G} a_{ij} \tag{2.9}$$

$$TOM(i,j) = \frac{\sum_{u \in \ G} a_{iu} a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2.10}$$

Relate modules to external information. In a final step, the inferred modules can be associated to certain biological functions or clinical outcomes. If external data such as trait associations are available, correlation to the identified modules can be calculated. Otherwise, gene set enrichments or over-representation analysis tools can be used to identify biological functions. Alternatively, the measure of module membership (MM) can be used to identify genes that are important for the module. The MM of a gene i is determined based on the correlation value of the node x_i to the Eigengene vector E of the module q, which is the first principal component of the module:

$$MM(q) = K_{cor,i}^{(q)} := cor(x_i, E^{(q)}), for\ i \in \{1, ..., |G|\} \tag{2.11}$$

Module preservation. Apart from inferring the gene correlation networks, some applications also seek to compare module structures across data sets, for example from different genders, or validate the reproducibility of modules. In `WGCNA`, this can be achieved via a composite measure Z_{summary}that combines different network connectivity and density measures Z statistic values that result from the permutation test (Equation 2.12) [Langfelder *et al.* 2011].

$$Z_{summary} = \frac{Z_{density} + Z_{connectivity}}{2} \tag{2.12}$$

For this approach, a reference data set is required which allows inferring how many of its modules are preserved within a test set. Within the density measures $Z_{density}$, information about the mean correlation, adjacency, module membership, and the proportion of variance explained (PVE) is reflected (see Equation 2.13). The PVE can be considered as the mean squared module membership, aiming to define how well an eigengene represents the whole module.

The connectivity measure $Z_{connectivity}$ include the correlation values of the intramodular connectivity, the module membership, and the correlation values itself between the reference and test set, as indicated in Equation 2.14.

$$\begin{aligned} Z_{density} =& median(Z_{mean\ correlation}, Z_{mean\ adjacency}, \\ & Z_{proportion\ of\ variance\ explained}, Z_{mean\ module\ membership}) \end{aligned} \tag{2.13}$$

$$\begin{aligned} Z_{connectivity} =& median(Z_{cor\ intramodular\ connectivity}, Z_{cor\ module\ membership}, \\ & Z_{cor\ correlation\ values}) \end{aligned} \tag{2.14}$$

Finally, modules can be assigned to be either non-preserved, moderately preserved or strongly preserved compared to the reference [Langfelder *et al.* 2011].

Chapter 3

Effects of dropout and data imputation on single-cell correlation networks

Single-cell transcriptomics data potentially offers a high level of resolution, however many applications are currently hampered by the level of sparsity, or more specifically by the level of dropout. Depending i.a. on the experimental throughput, different scRNA-seq platforms produce different levels of dropout. In this chapter, the influence of increasing level of dropout on gene correlation networks is therefore systematically investigated. Alleviating the limitations of data sparsity, imputation approaches may help to interpolate the missed expression values. However, it remains unclear to which extent true signals are rescued or noise being introduced.

The inter-cellular heterogeneity in single-cell transcriptomics data furthermore promises to uncover gene regulatory networks that are specific to a cell type or cluster of cells that so far remain buried. However, to fulfil this promise, two things must be achieved using scRNA-seq data: (i) the identification of cell types, and (ii) the possibility to infer gene correlation networks. Therefore, the influence of data imputation on marker genes used to annotate a human retina data set was additionally analysed.

In the beginning, the workflow to answer this question will be described, while the main findings will be presented subsequently. The last part will be dedicated to discussing the impact of the results on the scientific community.

3.1 Workflow

Single-cell transcriptomics may open up hidden gene correlation networks within cell subtypes. However, due to the high level of dropout, which highly depends on the experimental setup and sequencing technique, direct network inference remains challenging. To answer the question raised above on how much noise is introduced and to what extent the hidden signals can be recovered, a synthetic data set is required that i) allows for gene correlation network inference based on its inherent correlation structure and ii) allows for precise and systematic discrimination between signal and noise. When both criteria are met, the direct influence of data imputations on the sparse data can be inferred (illustrated in Figure 3.1).

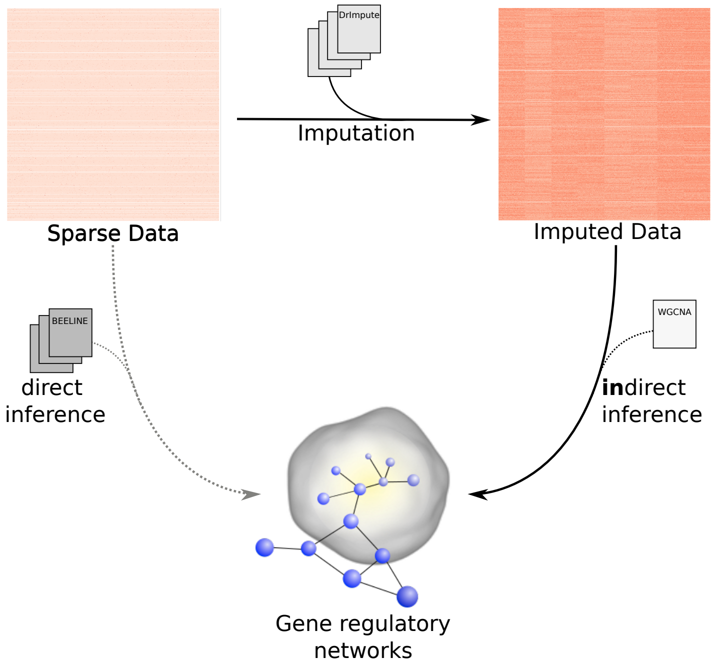


Figure 3.1. *Overcoming sparsity via imputation approaches to allow for network inference in single-cell transcriptomics data.*
While more and more tools are designed to allow for direct network inference in sparse, single-cell transcriptomics data, the approaches still lack robustness. One potential idea to circumvent this problem is the application of imputation approaches, trying to de-sparsify the data to enhance the masked signals. Finally, well-known, established network inference tools such as `WGCNA` *may be used to infer robust, meaningful gene correlation networks.*

A visual summary of the main parts of this analysis is given in Figure 3.2: Starting with the generation of the synthetic gold data, artificial dropout data sets were derived and used as input for different imputation tools. Based on these *recovered* data sets, gene correlation networks have been calculated using `WGCNA`. To measure the influence of imputation approaches on the network level, the preservation of modules and the edge recovery was quantified.

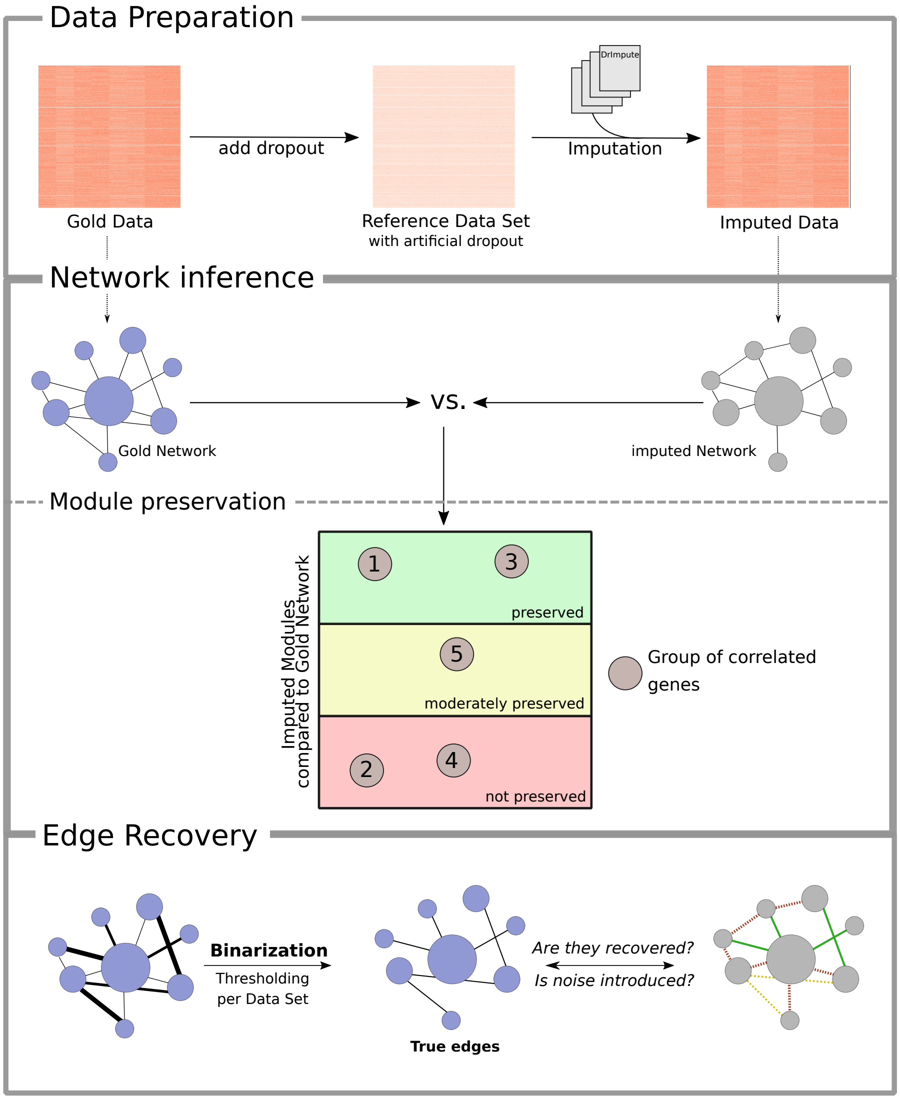


Figure 3.2. *Workflow to investigate the impact of imputation of gene regulatory networks (GRN) in single-cell transcriptomic data.*
By downsampling a bulk RNA seq dataset, a dropout-free, single-cell like, gold dataset was generated. Introducing increasing levels of dropout, six artificial sparse datasets were generated, which severed as the input for six imputation tools. The ability to preserve correlated gene groups from the gold data was inferred using the quantitative measure of module preservation from the `WGCNA` package. In a second step, the ability to recover true edges was evaluated by binarizing the weighted networks.

3.1.1 Data Sets

To state the influence of data imputation on single-cell transcriptomics data, two different datasets were used in this chapter. While a synthetic dataset served as a ground-truth reference, a human retina dataset was used to infer the impact of data imputation on cell cluster annotatability.

Generation of synthetic reference datasets. A dropout-free gold dataset preserving an appropriate correlation structure needs to be generated to evaluate the ability to reconstruct gene regulatory networks from single-cell RNA-seq data. As shown by Peng *et al.* [2019a], the existing correlation structure from bulk RNA-seq datasets can be used to create single-cell-like data. The bulk dataset of mice hair follicles was taken from Wang *et al.* [2017]* and contained 48795 genes across 48 conditions. Single-cell data generation was done in accordance with the workflow by [Peng *et al.* 2019a] and is available on their Github page. Briefly, eight conditions and 5000 genes were randomly selected during the downsampling procedure. Each of these conditions was used to simulate a certain cell type. After replicating each cell type 100 times, an 800 cell data set was generated. To resemble the single-cell-like gold data, (five times) the standard deviation of each gene in each condition was used to introduce noise via a random normal distribution with mean zero and one hundred observations. A dropout rate was modelled via the λ parameter (range from 0 to 1). The higher the λ, the smaller the resulting sparsity. This dropout rate followed an exponential function $e^{-\lambda * \text{mean expression}^2}$. Zero values representing the explicitly dropped out gene expressions, were at the end introduced via a Bernoulli distribution defined by the dropout rate. In total, six different λ values were used with 0.01,0.09, 0.21, 0.42, 0.7, and 0.99 to generate six different sparse datasets ranging from 84 % to 40% dropout, respectively. All data sets were filtered for genes and samples with missing values using the `WGCNA GoodSamplesGenes` function and transformed to count values.

Retina organoid data set. To tackle the second important aspect of the preservation of cell cluster annotatability, a human retina organoid dataset from Kim *et al.* [2019](GEO:GSE119343) was used. In their original study, retina organoids were clustered and annotated according to the major cell types in the retina: rods, cones, and Müller Glia cells. Specific marker genes were extracted to describe each retinal cell type. This dataset contained 19426 genes across 1346 cells and included 85% zeros.

---

[*] `https://github.com/software-github/scrabble_paper`

3.1.2 Data imputation

Prior to the imputation, genes not being expressed in at least two cells were removed using the `preprocessSC` function from the `DrImpute` package [Gong *et al.* 2018]. Data input was performed as summarized in Table 3.1 and described in the section 2.1.

Table 3.1. *Overview of published imputation approaches used in this thesis.*

Imputation	Publication	Class	Data input	Output	Code
`DrImpute` (v1.0)	[Gong *et al.* 2018]	(2)	Log10(X+1) data	Logged data	`R`
`SAVER` (v1.1.2)	[Huang *et al.* 2018]	(3)	Count data	Count data	`R`
`DCA` (v1.3.1)	[Eraslan *et al.* 2019]	(1)	Count data	Count data	`python`
`ENHANCE`	[Wagner *et al.* 2019]	(4)	Count data	Both possible (Logged used)	`R`
`scNPF` (v0.1.0)	[Ye *et al.* 2019]	(5)	Count data + TOM data	Count data	`R`
`DISC` (v1.1.2)	[He *et al.* 2020]	(1)	Count data	Count data	`python`

All imputation tools were run according to their recommended settings, which are described in R-markdown files on the Github repository[*]. Solely `ENHANCE` was applied with a fixed knn-parameter of eight for the synthetic dataset.

The `scNPF` tool made use of a gene correlation network to guide the imputation process. In its implementation, two different types of networks can be imported which are either directly derived from the sparse data or extracted from a reference database. For the synthetic dataset, `WGCNA` (v1.69) was used for both scenarios: a network based on the sparse datasets to reflect the first case and a network of the gold data to resemble the latter case. In either way, topological overlap measure (TOM) matrices were used as `scNPF` input. For the human retina dataset, a `WGCNA`-derived, sparse network as well as reference data by the STRING database, was used.

`DISC` required the input data in a `LoomPy`[†] format which was accomplished in accordance with the tutorial[‡]. While succeeding on the synthetic data, `DISC` failed to operate on the larger human retina dataset.

---

[*] `https://github.com/lisbeth-dot-95/Dissertation`

[†] `http://loompy.org/`

[‡] `https://nbviewer.jupyter.org/github/iyhaoo/DISC/blob/master/reproducibility/Data\%20Preparation\%2C\%20Imputation\%20and\%20Computational\%20Resource\%20Evaluation/Data\%20Pre-processing/MELANOMA.ipynb`

3.1.3 Network Analysis using `WGNCA`

Gene correlation networks were detected using the Weighted Gene Network Correlation Analysis (`WGCNA`) R-package [Langfelder & Horvath 2008] (version 1.69). The general steps were performed according to the tutorial stated on the website *.

Module detection. The first step when inferring a network with `WGCNA` was to remove genes and cells containing zero values utilizing the `GoodSamplesGenes` function. Subsequently, the scale-free model fit was calculated. The criteria on how to choose the optimal β value were presented in section 2.4.1.

For subsequent module detection, a minimal cluster size of 20 was used in the case of the synthetic data set, while the remaining parameters were left with their default values.

Gene module detection in the synthetic dataset was only performed on the gold dataset which does not contain artificial dropout. Neither in the sparse nor the imputed data a module detection was necessary.

Investigate Module preservation statistics. The module preservation was calculated using the provided `modulePreservation` function. A detailed explanation of this functionality is also provided in Section 2.4.1 . To reduce the random reduction of modules, the maximal module size was set to the number of genes included in the dataset. The number of permutations was set to 100, the `dataIsExpr` parameter was left on default, and all other arguments were chosen according to the tutorial. To make the results more tangible, log2 fold changes (log2-FC) of the module-specific Z_{summary} scores were calculated between the gold and the test datasets. Gold data Z_{summary} values were obtained by calculating the module preservation against itself. A negative value indicates a lower Z_{summary} value in the test than in the gold dataset. A value of zero corresponds to equal values.

Evaluation of edge detection. Next to module preservation, other measures were included to assess whether true gene correlation can be inferred after imputation. Therefore, the weighted `WGCNA` network was transformed to an unweighted, binarized network.

Gene modules were detected in `WGCNA` using the topological overlap matrix (TOM) measure. By setting certain thresholds to the TOM values, the presence of an edge was binarized. Since the distribution of TOM values was different before and after imputation, a distribution-depended threshold was applied, using the following formula: $e * SD + mean; e = \{1, 2\}$. To

---

[*] `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/`

avoid bias in the model evaluation measures, the self-correlated gene diagonal was removed from the subsequent analysis.

Here, the gold network was used as a *true* reference and the imputed networks as *predictions.* Three different measures were applied to analyse the problem from different perspectives. On the one hand, precision and recall were used to approximate the recovery of true gene-correlations. On the other hand, *Matthew's correlation coefficient* (MCC) was applied to infer the performance of the whole edge classification task. The MCC is particularly useful since these datasets are heavily unbalanced towards true negative entries (non-correlated genes). The measures are defined as follows:

Precision:

$$\frac{True\ Positives(TP)}{True\ Positives(TP) + False\ Positives(FP)} \tag{3.1}$$

Recall:

$$\frac{True\ Positives(TP)}{True\ Positives(TP) + False\ Negatives(FN)} \tag{3.2}$$

Matthews Correlation Coefficient:

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3}$$

3.1.4 Cell cluster embedding and annotation

Using a human retina organoid dataset, the influence of data imputation on cell cluster annotatability is investigated. The influence on cell clustering, or more specifically on cluster separability is also analysed. Therefore, the analysis pipeline `scanpy` is used (version 1.6.0)[Wolf *et al.* 2018] . In the original Kim *et al.* [2019] publication, the retina organoid data was processed using the `Seurat`(version 2.3.4) workflow. To exclude misinterpretations based on the different analysis workflows, a comparison of `scanpy` and `Seurat` has been conducted.

The cell embedding and subsequent cluster annotation with `scanpy` were run as follows: Sparse and imputed data were filtered for low-quality cells and genes upon the initial data import into `scanpy`. As described in the original publication, cells with less than 600 genes and genes which do not occur in at least 5% of cells were discarded. Here, 5% account for 67 cells. This parameter was changed for `ENHANCE` imputed data, where genes without any count were removed. Otherwise, the subsequent downstream tasks would not have been conductible after imputation.

In the following step, cells expressing an aberrantly high number of genes will be detected and removed. Since some imputation tools introduced a uniform number of expressed genes across cells and to ensure comparability across imputation approaches, this threshold was set individually per dataset to avoid excessive filtering. Those individual thresholds are given in Table 3.2.

Using the data normalization procedure described in the `scanpy` tutorial, followed by a log10(x+1) transformation, highly variable genes were detected and kept using default parameters. Subsequently, unwanted variation from the total gene counts was regressed out and data was scaled according to parameters given in the paper.

Finally, a dimensionality reduction using principal component analysis (PCA) prior to the t-SNE was done. The neighbour graph was calculated using ten neighbours and 40 principal components. Cluster detection was done via the Louvain algorithm using standard parameters.

Table 3.2. *Overview of filtering thresholds used for data preprocessing. The table shows the individual gene number upon which cells were discarded in case they exceed this cutoff of expressed genes.*

Imputation	**Number of genes**
`Sparse`	4000
`DrImpute`	7900
`SAVER`	13369
`DCA`	14845
`ENHANCE`	14000
`scNPF`	13369
`scNPF` String-reference	13369

Clustering performance evaluation. To infer the performance of the Louvain clustering before and after the imputation, the mean silhouette coefficient from the `sk.learn` package (version 0.23.2) was used [Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel *et al.* 2011]. This score helps to infer clustering performance by investigating the tightness and separation of each cluster [Rousseeuw 1987]. Therein, the expression data and the Louvain cluster assignment was used as input. The Euclidean distance was applied as a distance metric.

Automatic cell cluster Annotation. Based on marker genes expression, cell clusters could be annotated to their corresponding cell type. To infer whether this was still possible after the imputation, an automated cluster annotation pipeline was implemented.

Using the originally published marker genes for cones, rods, and Müller Glia cells [Kim *et al.* 2019], the mean expression per gene and Louvain cluster was calculated. To allow for a fair comparison, the data was scaled per Louvain cluster with a min-max-normalization. In the next step, the scaled data was binarized using a threshold of 0.5. A Louvain cluster was then annotated if more than 75% of the marker genes were expressed above this threshold. Three classes of annotation results were inferred such as pure, mixed and unassigned clusters. Pure clusters showed an assignment for exactly one cell type and mixed clusters exhibited at least a double assignment. The third class included Louvain clusters with no successful annotation. Their relative numbers were compared across imputation tools.

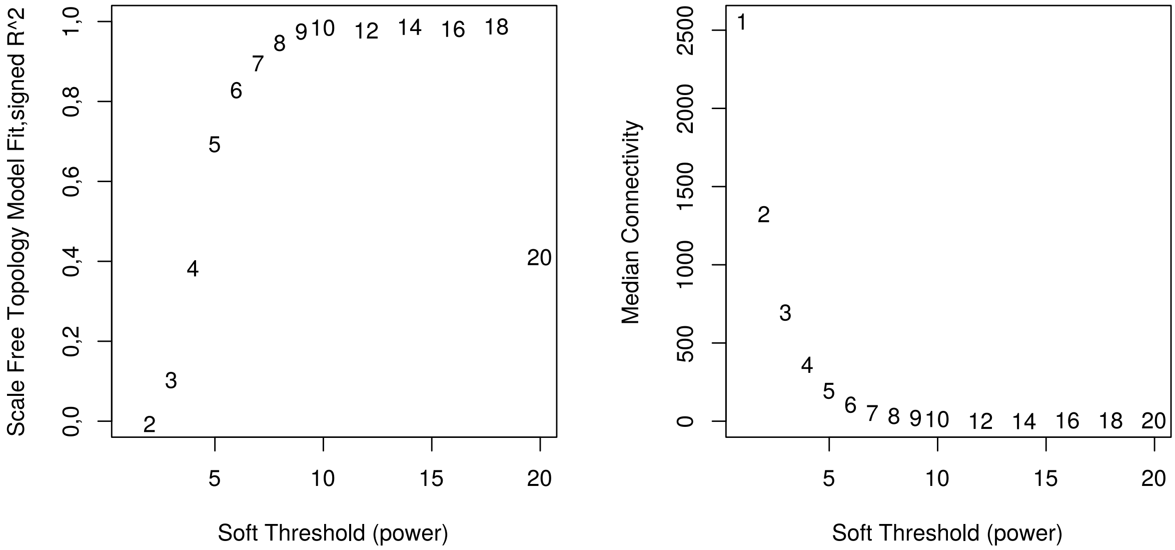
3.2 Generation of single-cell like reference data sets

While scRNA-seq allows for unprecedented biological insights, technical as well as biological noise produce an extremely sparse data matrix hindering current network inference methods [Pratapa *et al.* 2020, Chen & Mar 2018, Nguyen *et al.* 2020]. Tackling how data imputation influences gene correlation network inference, a data set is required which only holds ’true signals’. A true signal in this chapter refers to truly measured expression values and more specifically non-dropped out entries. Since, however, all experimentally inferred data sets suffer from some degree of dropout, only a synthetic scenario was appropriate.

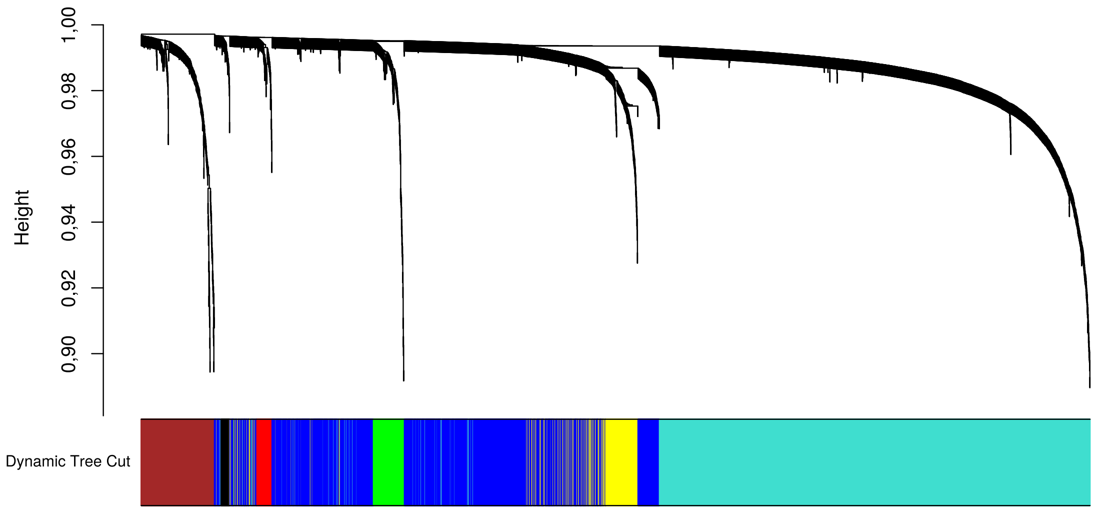
Gold data. When generating synthetic data sets, coping with reproducibility, scalability, documentability and suitability for the particular aim is of major importance. Here, the latter point mainly refers to a biologically meaningful correlation structure. Using a workflow proposed by Peng *et al.* [2019a], an existing mouse hair follicle bulk data set by Wang *et al.* [2017] was used as an input to generate a single-cell-like data set. It contained 48,795 genes across 48 conditions (over 20 cell types). For the downsampling procedure, eight hair follicle cell types were chosen randomly such that the synthetic gold data set owned 5000 genes across 800 cells. Since only true zeros were included in this data, it will be referred to as *gold data.*

After a quality filtering to remove barely expressed genes, a gold data set with 4960 genes across 800 cells was created. In total, this data set contained 15% of true zeros. As a proof-of-concept, `WGCNA` [Langfelder & Horvath 2008] was applied to the gold data set to ensure that the natural correlation structure of the bulk RNA-seq data set was retained, see Figure 3.3. Figure 3.3a shows the R^2 of a scale-free topology model fit, which represents a hallmark of many networks, over different soft threshold β values.

Based on this plot, a β value was chosen to generate a network with approximately scale-free topology and sufficiently high median connectivity. For the gold data, a plateau for β values larger than eight was reached. Other simulation tools, like `splatter` [Zappia *et al.* 2017] or `GeneNetWeaver` [Schaffter *et al.* 2011] failed in producing data sets that allowed reconstructing approximately scale-free networks. This may be due to an insufficient ability of these tools to correspond to real biological processes that have been previously reported [Pratapa *et al.* 2020]. By choosing a β of nine for the gene network inference, the gene dendrogram showed a clear hierarchical structure within the genes (Figure 3.3b). In total, seven different gene modules were detected, whereby the turquoise module was the largest with 2143 genes and the black module was the smallest with only 21 genes.



(a) *Scale-free model fit and median connectivity of the gold data.*



(b) *Cluster dendrogram of synthetic Gold data.*

Figure 3.3. *Characteristics of the synthetic gold data.*

(a) Development of scale-freeness and median node connectivity over twenty β values. A plateau of R^2 values was reached for β values larger than eight. (b) Cluster Dendrogram of gold data using the `WGCNA` *package. The gene module assignment is shown in the colour bar below. The hierarchy of the genes is indicated on the y-axis. This result is based on a β of nine. In total seven gene modules were defined.*

Reference data sets. After defining the *ground-truth* data, different but defined levels of dropout had to be added to *mask* the expression signals. As guidance, conventionally achieved sparsity levels (here percentage of zeros) from different single-cell workbenches were used. Based on the gold data set, six additional data sets were generated with increasing degrees of dropout, ranging from 40% to 84%. Accounting for the 15% true zeros in the gold data, a total of 55% to 99% zeros were included. These data sets together with the gold data will be referred to as *reference data sets*. The characteristics of these reference data sets are summarized in Figures 3.4 and 3.5. A glimpse into the influence of the dropout on the gene expression data is given in Figure 3.4. Here one hundred genes from the gold data, as well as one representative for the mild (40%), moderate (66%) and high (84%) dropout are shown.

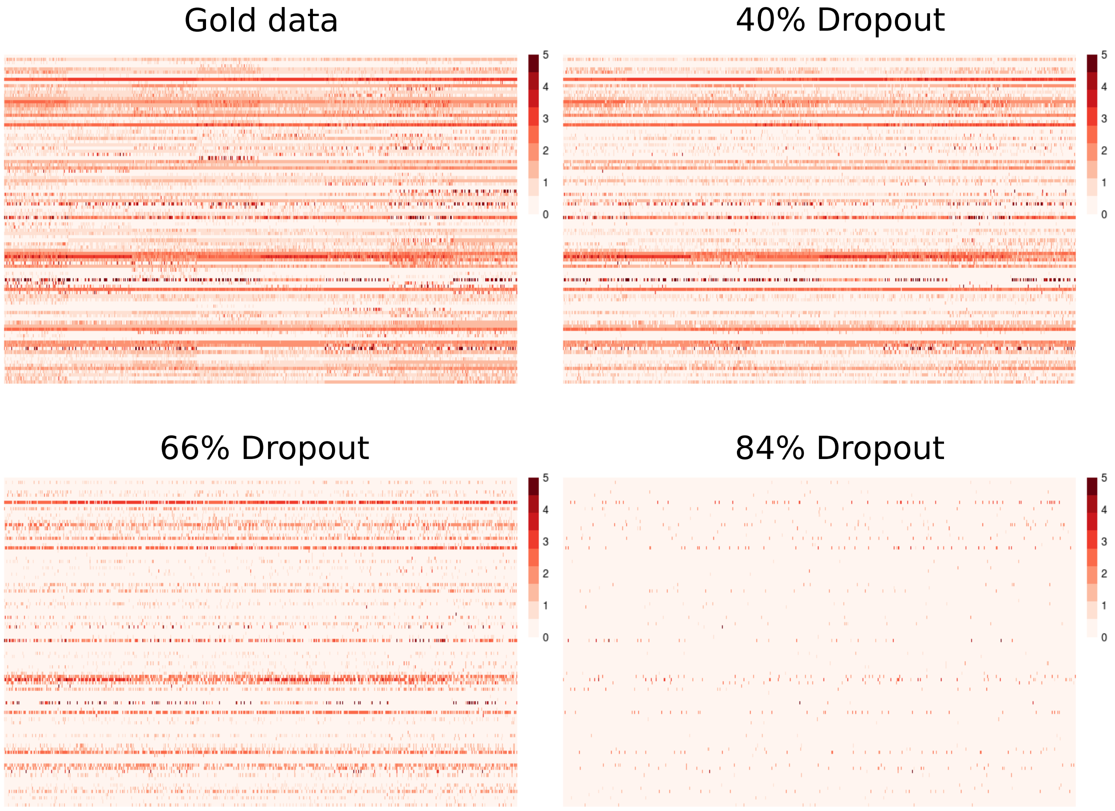


Figure 3.4. *Heatmaps of selected genes in the gold and three sparse data sets.*
Heatmaps showing log10(X+1) expression of first 100 genes to demonstrate impact of dropout in selected data sets. While the lowly expressed genes mainly dropped out in 40 and 66% dropout data sets, only a sparse gene expression is left in the 84% dropout data set.

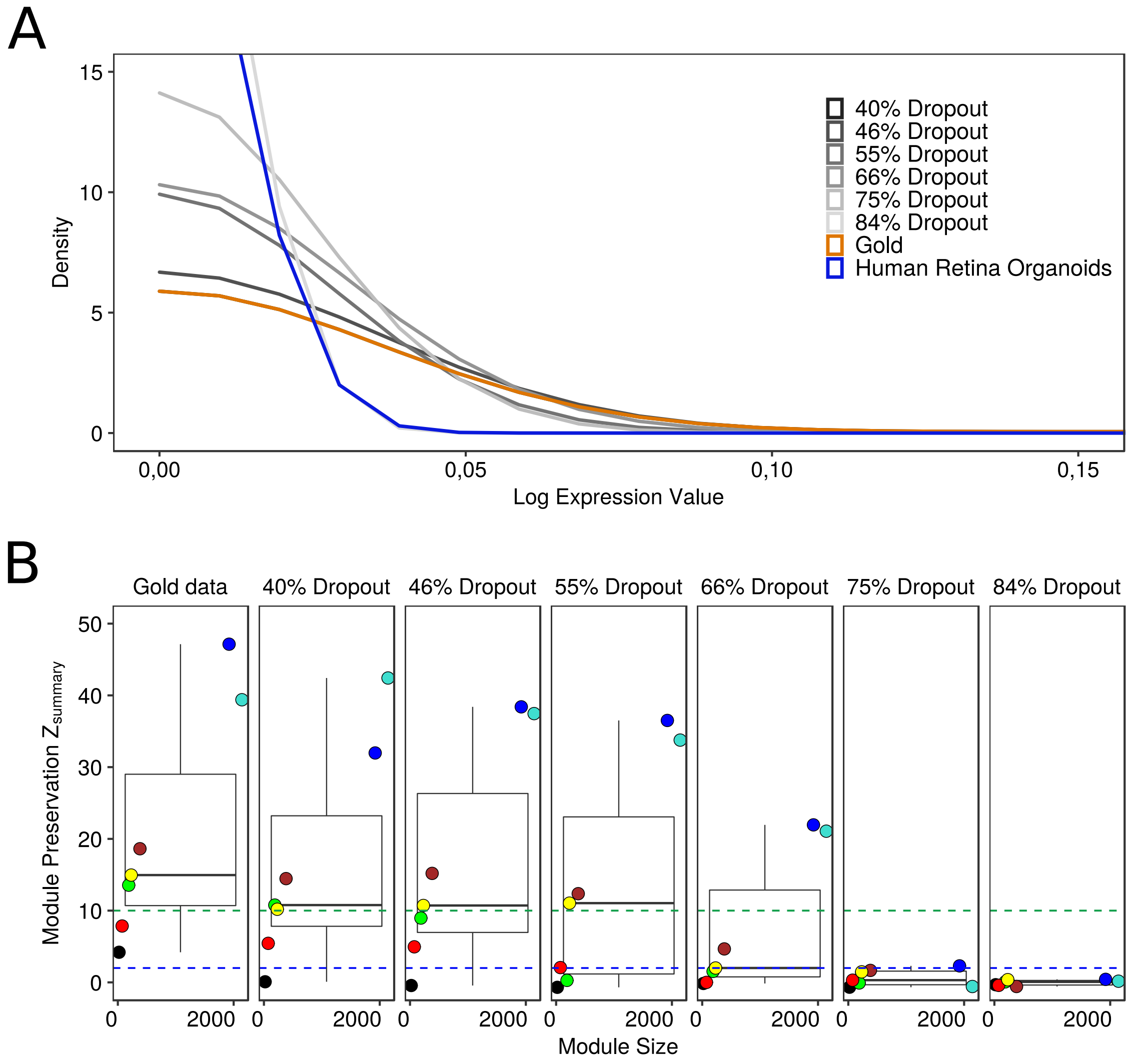


Figure 3.5. *Reference data characteristics.*

The term dropout refers to the amount of artificially introduced non-true zeros in each of the reference data sets.(A) Distribution of logged expression values of all reference data sets and the human retina data set. The density of expression values across eight data sets is shown to contrast the impact of different dropout rates. The gold data is plotted in orange, all six dropout reference data sets are shown with a grey gradient and a biological data set is plotted in blue. Here, only an excerpt of the x-axis between 0.0 and 0.2 is shown. (B) Boxplots showing the behaviour of module preservation across different dropout levels. The $Z_{summary}$ *measure implemented in* `WGCNA` *is a composite, permutation-based metric of various network density and connectivity measures. The blue and the green line indicate the threshold towards moderate and strong module preservation, respectively. Colouration of the dots corresponds to the individual modules. An increase in dropout is associated with a decrease in module preservation.*

While mainly lowly expressed genes were dropped out in the mild and moderate scenario, only a minority remained detectable in the high dropout data set. With an increasing degree of dropout, the distribution of gene expression values indicated the expected shift towards larger proportions of lowly expressed or missing genes (see a clipped version in Figure 3.5A and the complete graph in Supplemental Figure S-1. The gold data showed the smallest fraction of lowly expressed genes, while the 84% dropout data set showed the opposite behaviour. A scRNA-seq data set of human retina organoid cells [Kim *et al.* 2019], generated via the 10X Genomics procedure was included and plotted for comparison, exhibited a very similar distribution as the 84% dropout data set.

The Z_{summary} statistic from `WGCNA` quantifies module preservation [Langfelder *et al.* 2011] and therefore helps to investigate to what extent hallmarks of network structure can be recovered or are lost with increasing degrees of dropout (Figure 3.5B). This measure indicates how well a group of correlated genes (modules) from a reference data set is preserved in a test data set. Here, all artificially dropped out data sets were compared to the gold data. Since the Z_{summary} is dependent on module size, the gold data was compared to itself as a reference. As introduced by Langfelder and colleagues, two different thresholds define the preservation of gene modules [Langfelder *et al.* 2011]: modules with a Z_{summary} value below two are considered to be not preserved at all, whereby values larger than ten indicate strong preservation. Modules in between both thresholds are considered to be moderately preserved. Within the gold data, modules showed a median Z_{summary} value above the strong preservation threshold as also seen for slightly increased dropout levels (40% - 55%). Z_{summary} values close to zero were detected for both high sparsity data sets containing 75% and 84% dropouts. Moreover, no single module remained even moderately preserved in the 84% dropout reference data set. In general, it can be stated that an increase in dropout correlates with a decrease in module preservation. Nonetheless, correlation networks appear fairly robust for low to intermediate levels of dropout. Up to a level of 55 % dropout, a strong preservation for the majority of modules could be observed.

Based on those findings, a meaningful reference data sets were generated that (1) resembled true biological data sets in their distribution of expression values, (2) retained the natural correlation structure of its original bulk RNA-seq ancestor, and (3) were shown to have decreasing network preservation with higher levels of dropout.

3.3 Impacts of data imputation on network inference

Since dropout seems to compromise network inference, next it was investigated, whether imputation methods alleviate the situation.

Five imputation tools representing different types of imputation methodologies were selected, spanning a variety of underlying mathematical assumptions as well as implementation languages (see Table 3.1).

Module Preservation. To compare the imputation tools based on Z_{summary} values, a module-wise Z_{summary} log-2 fold change was calculated (log2-FC) that reflects the difference in module preservation to the gold data set before and after imputation, see Figure 3.6. A negative fold change, therefore, represents a module preservation smaller than the gold data, and thus a loss of network information.

The unimputed, dropout-affected reference data sets (shown in black) showed lower negative log2-FC values with an increase in dropout, reflecting the loss of module preservation as was described above. All log2-FC in the 40% dropout data set were still close to zero, while it decreased drastically in the 84% dropout data set. If imputation restored the correlation structure, values closer to zero would be obtained. However, a rather diverse picture was observed: For low levels of dropout, where network inference still worked robustly without imputation, only `DrImpute` and `DCA` yielded a log2-FC comparable to the unimputed data. All other tools performed considerably worse. At intermediate levels of dropout (55-66%), where network inference without imputation was increasingly affected but still feasible, `DrImpute` and `DCA` improved the log2-FC. `SAVER`, `ENHANCE` and `DISC` resulted in similar log2-FC distributions as the unimputed data, while both variants of `scNPF` even diminished the network information.

For the high dropout levels, none of the imputation tools enabled network inference concerning the overall small log2-FC. While `DCA` and `ENHANCE` for 75% dropout and `scNPF` for 84% of dropout performed substantially better than the unimputed data most of the network structure was still lost. An overview of the individual module preservation values is presented in Supplemental Figure S-2.

In summary, from a module preservation perspective, imputation using `DrImpute` and `DCA` facilitated network inference for data sets with low to intermediate levels of dropout, whereas several other tools severely compromised the inference at those dropout levels. For high-level dropout data sets, however, no imputation tool showed convincing and promising results.

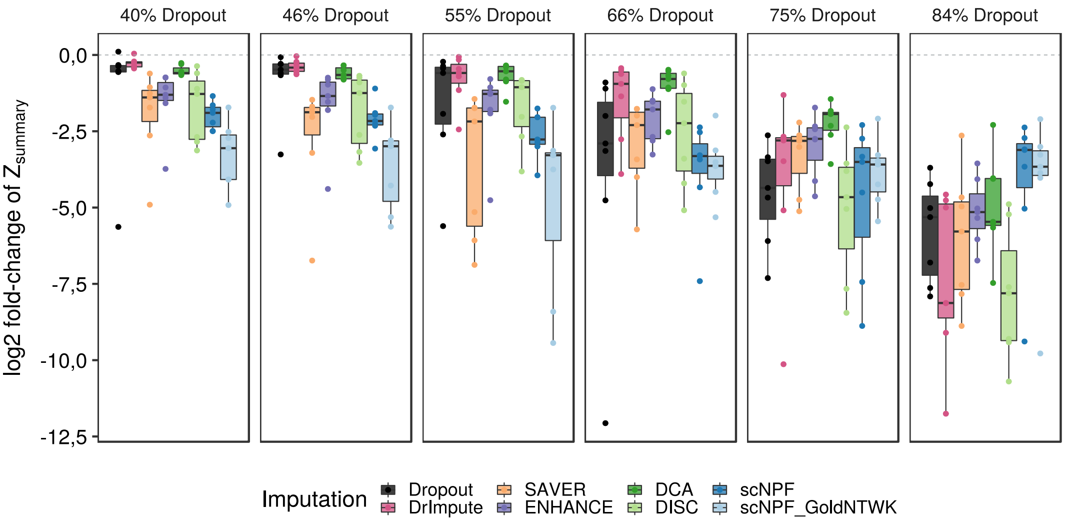


Figure 3.6. *$Z_{summary}$ before and after imputation of reference data sets.* *Boxplot of the $Z_{summary}$ log2 fold change (log2-FC) of all reference data sets compared to gold data before and after imputation. $Z_{summary}$ was computed for gold data compared to itself and for any reference data set, with and without imputation, compared to gold data, respectively. Subsequently, a log2 fold change was computed between the gold versus gold and any reference versus gold $Z_{summary}$ to illustrate how well gene modules from the gold data were preserved in the dropped out and imputed data. The dashed line indicates the threshold for completely recovered modules. Dots represent the values of individual gene modules. The dropout data is depicted in black. Implementation tools that employ related methodological approaches were grouped by similar colours.*

Edge recovery. By calculating module preservation scores, a similarity measure to infer the impacts of imputation on network inference was employed. By switching the perspective, the ability of imputation methods to recover *true* gold data gene-gene interactions was analysed.

Therefore, the network with continuous edge weights was transformed to an unweighted network either stating that an edge is present or not. This transformation allowed the computation of measures like precision, recall, and Matthews correlation coefficient (MCC) to quantify the potential of each imputation technique to recover true edges. To account for different data distributions and ranges, the binarization threshold was derived individually per data set. See the workflow section for details. Initially, the precision was evaluated, which measures the fraction of true edges over all detected edges, see Figure 3.7. The higher the precision, the fewer false positive edges were detected. Before imputation, the reference data

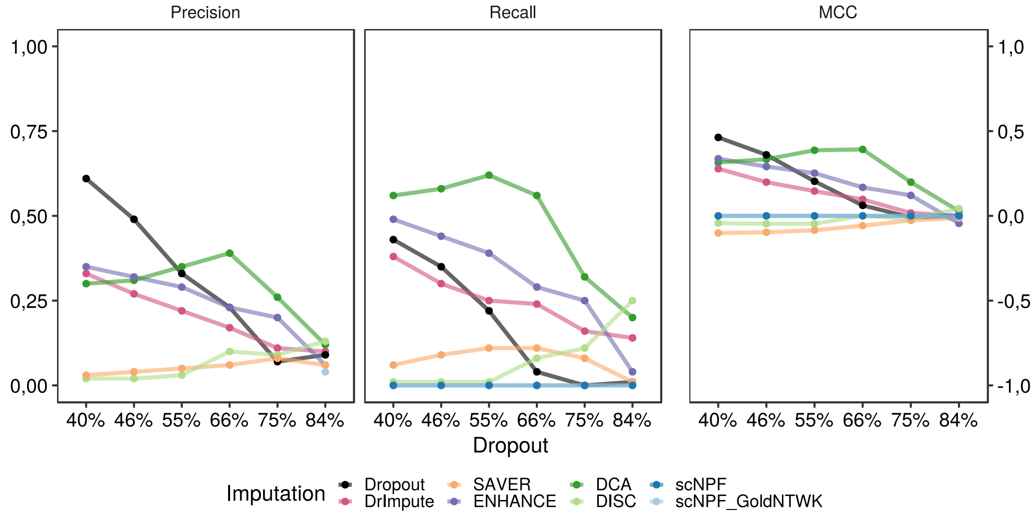


Figure 3.7. *Edge recovery trend before and after imputation.*

Trends of Precision, Recall, and Matthews correlation coefficient (MCC) over all dropout levels indicating the ability to recover true edges after binarizing the gold data network. The mean and standard deviation (SD) of the topological overlap matrix (TOM) per data set were computed. Edges with a TOM greater than 1 SD + mean of the gold and all dropped out and imputed data sets were retained. Retained edges of gold data were considered as true edges for computing Precision, Recall, and MCC, respectively.

sets revealed moderate precision in the 40% and 46% dropout data sets. With increasing dropout, the precision continuously decreased towards zero.

Imputation of low sparsity data sets (40-55% dropout) did not improve the precision. Whereby some tools such as `DCA`, `DrImpute`, and `ENHANCE` reached precision values half as high as the dropped out data, `SAVER`, and `DISC` revealed values close to zero. It was not possible to calculate precision values for both `scNPF` approaches. Based on a high mean and a high standard deviation in the TOM values, no edge was retained. As soon as the dropout level exceeded 55 %, `DCA` retrieved a higher precision than the unimputed data with a peak of performance at the 66% dropout data set. For both high dropout data sets, all approaches showed low precision values. These results suggest, that independent of the method used, imputation methods applied to high dropout data sets tend to inflict the correlation network with large amounts of false edges.

The recall measures the fraction of true edges in the gold network that was classified as edges in the dropout or imputed data sets. Similar to the precision, the dropout data sets revealed a stepwise decrease in recall with increasing dropout levels. Overall, this trend also persisted after data imputation. However, `DCA` showed increasing recall rates until a dropout level of 55% was reached. `DCA`, `ENHANCE`, and `DrImpute` consistently outperformed the unimputed data sets. `DISC` showed a steady increase in recall starting from dropout levels of 55%, resulting in the highest recall value at 84%. Both `scNPF` resulted in recall rates of zero based on the fact that no edge was detected in those data sets.

Similar to scale-free biological networks with only a few highly connected edges, the data set was highly unbalanced towards true negative edges. The MCC represents a more robust evaluation of the classification for such imbalanced data sets since it makes use of all entries in the confusion matrix. Perfectly predicted values reflect an MCC of one, whereby a random assignment would result in values close to zero. Wrong predictions would reveal negative values ranging up to minus one. Here, trends similar to the precision were observed. While the dropout data sets exhibited the highest MCC values at 40 and 46% dropout, `DCA` and later `ENHANCE` outperformed the unimputed data exceeding 46% dropout. For the highest dropout data set, neither the sparse nor any imputation tool reached good MCC values. Both `scNPF` approaches and `SAVER` consistently showed values close to or below zero.

Gaining a broader picture of the effect of imputation prior to network inference, the edge recovery analysis revealed that good recall values in data imputation were mainly caused by inflating the data with more correlation, i.e. artificial edges, compared to the gold data as highlighted by the moderate precision values (Supplementary Figure S-3). However, as already indicated by the module preservation analysis, overall good edge recovery values were obtained for the unimputed data with low dropout levels. `DCA` revealed a moderate performance on data sets affected with intermediate levels of dropout.

3.4 Cell cluster annotation in human retina organoid data

Before identifying cell type-specific network properties, the respective cell type must be annotated. Therefore, also the cell correlation before and after imputation was analysed. For that reason, a human retina organoid data set that was published by Kim *et al.* [2019] was used. This data set is afflicted by dropout (total percentage of zeros: 85%) and will be denoted hereafter as sparse data set. In their study, Kim *et. al* clustered the cells and annotated them as either rod, cone, or Müller Glia (MG) cells based on marker genes extracted from their low-dimensional t- distributed stochastic neighbour embedding (t-SNE).

As many tools improve cell clustering [Hou *et al.* 2020], the impact of data imputation on low-dimensional cell embedding was analysed. Here, t-SNE, as well as Louvain cluster embedding, was used, as already performed in the original Kim *et al.* [2019] publication. Due to the higher robustness of the implementation, here the python-based `scanpy` workflow instead of the original R-based `Seurat` pipeline was used. A comparison between the two results is shown in Figure 3.8. As depicted in (B) and (D), the shown t-SNE representations appear different, however, the marker genes expression shown in (A) and (C) reveal similar patterns. While six clusters were detected in the `Seurat` workflow, eight clusters were identified using `scanpy`. Generally, the clusters from the `scanpy` workflow appear less distinct than the ones obtained by `Seurat`. Nevertheless, distinct expression patterns of the marker genes were observable across both pipelines. Especially focussing on the marker genes associated with rods, cones, and MG cells, unique expression patterns per cell cluster were detected. Also, the relative expression *strength* (here the colour of the heatmap) corresponds across pipelines.

Starting from this sparse data set, the impact of the previously introduced imputation methods concerning cell clustering and a subsequent marker gene-based annotation of cell types was investigated.

The Louvain cluster detection within the t-SNE of the original sparse data set revealed eight cell clusters, see Figure and 3.9(a). Most imputation tools produced comparable numbers of eight to eleven clusters, Figure 3.9(b-f). Solely the imputation by `ENHANCE` resulted in a total of twenty-one clusters being detected (Figure 3.9(g)). Cluster sizes were comparable between the sparse data and most imputation methods. The largest Louvain cluster contained around 300 cells in the original organoid data, as well as for both `scNPF` variants, `DCA`, and `SAVER`. Explicitly smaller sizes were obtained after `ENHANCE` and `DrImpute` imputation with 128 and 195, respectively. The smallest Louvain cluster counted between 39 to 46 cells, which was found in the sparse data set and after both `scNPF`, `SAVER`, and `DrImpute`. Deviations were detected after `ENHANCE` and `DCA` imputation with 19 and 62, respectively.

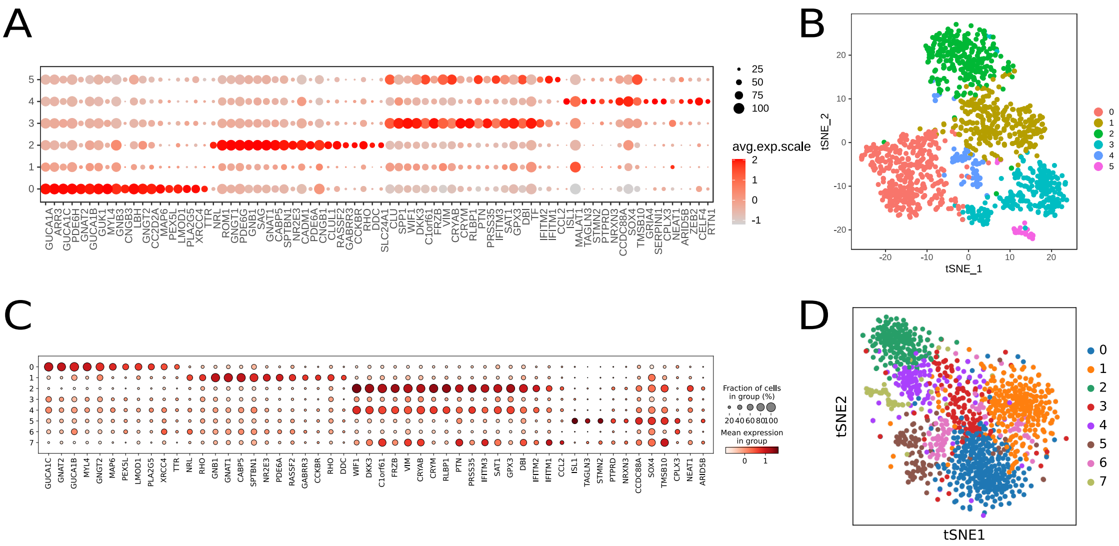
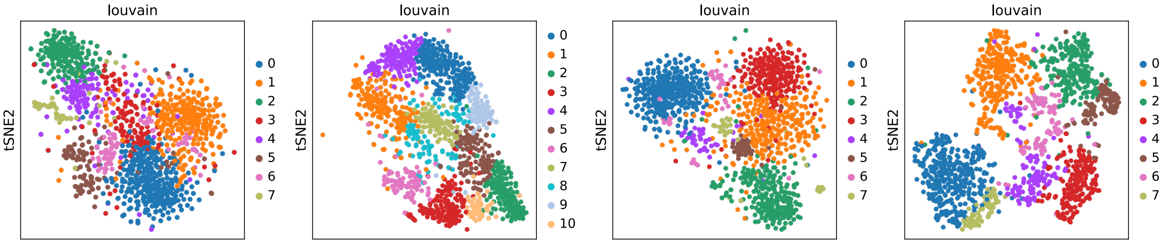


Figure 3.8. *Comparison of* `Seurat` *and* `scanpy` *preprocessing results.*

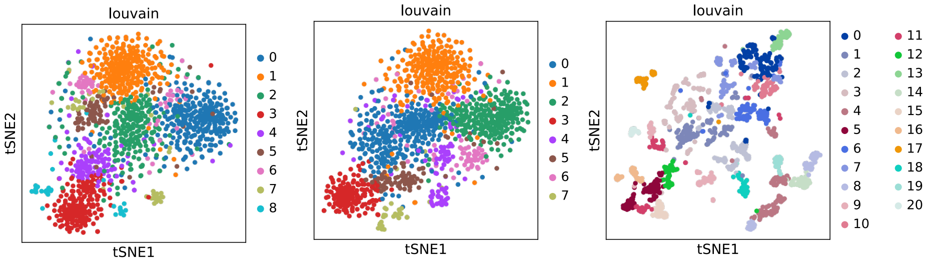
(A) Dot plot showing the expression of marker genes across Louvain clusters using the `Seurat` *pipeline. Colouring corresponds to the mean gene expression and the dot size to the percentage of cells per cluster expressing the respective gene. (B) t-SNE of Louvain clusters using* `Seurat`*. (C) Dot plot showing the expression of marker genes across Louvain clusters using the* `scanpy` *pipeline. The expression was scaled per Louvain cluster. (D) t-SNE of Louvain clusters using* `scanpy`*. Colours between t-SNE plots do not correspond.*

By mere visual inspection of the cluster density for the sparse data, both `scNPF` alternatives and `SAVER` showed clusters with intertwining boundaries. Imputation through `DrImpute` and `DCA` resulted in slightly more dense clusters while `ENHANCE` led to a completely different cluster layout with much more fine-grained clusters. To back these findings up with a statistical measure, the mean silhouette coefficient (MSC) was calculated that allowed to quantify the density of clusters. Clusters revealing an MSC closely to one reflect well separated, dense cell clusters, whereas values close to zero hint towards overlapping clusters. Negative values (lowest:-1) indicate that a cell was most likely misclassified. Before and after imputation overall MSC values were found to be close to zero. Slightly positive values were obtained by `DCA` and `ENHANCE`. The weakest performance was found for `DrImpute`. These results support the visual observation of overlapping and not well-separated clusters. Thus, data imputation did not improve the cluster density substantially compared to the sparse data.

Based on their original clustering, Kim *et al.*, derived unique marker genes for the retina-specific cell types: rods, cones, and MG cells. In the following, the reproducibility and



(a) sparse (b) DrImpute (c) SAVER (d) DCA



(e) scNPF (f) scNPF String (g) ENHANCE

Figure 3.9. *Louvain clusters detected in human retina organoid data before and after imputation.*

The sparse and imputed data sets were subjected to a standard preprocessing procedure in `scanpy` *and projected into a low dimensional t-SNE. The embeddings were coloured according to the Louvain cluster detection. Colours are specific for the individual plot and do not correspond between data sets. Eight to eleven Louvain clusters were generally detected. Solely* `ENHANCE` *imputation resulted in twenty-one clusters. In most cases, cell clusters were found to be overlapping on the cluster boundaries. Visually more separated clusters were obtained after* `DCA` *and* `DrImpute`*.*

consistency of these marker genes before and after imputation was investigated by comparing quantities of annotated cell types. Therefore, an automated cluster annotation pipeline was set up based on the expression patterns of these marker gene lists (see section 3.1.4). All dot plots, showing the expression patterns of these marker genes before and after imputation are provided in Supplemental Figure S-4. Briefly, a cluster was annotated as rods, cones, or MG cells when more than 75% of the cells in this cluster expressed the respective marker genes. A summary of this analysis is depicted in Figure 3.10(C).

To compare the results on a cluster level, three classes were generated: (i) *pure clusters* represented the percentage of clusters where all cells had the same annotation either as rods, cones, or MG cells, (ii) *mixed annotation* summarized clusters with at least two different annotations, while (iii) *not assignable* contained clusters where a unique annotation was not

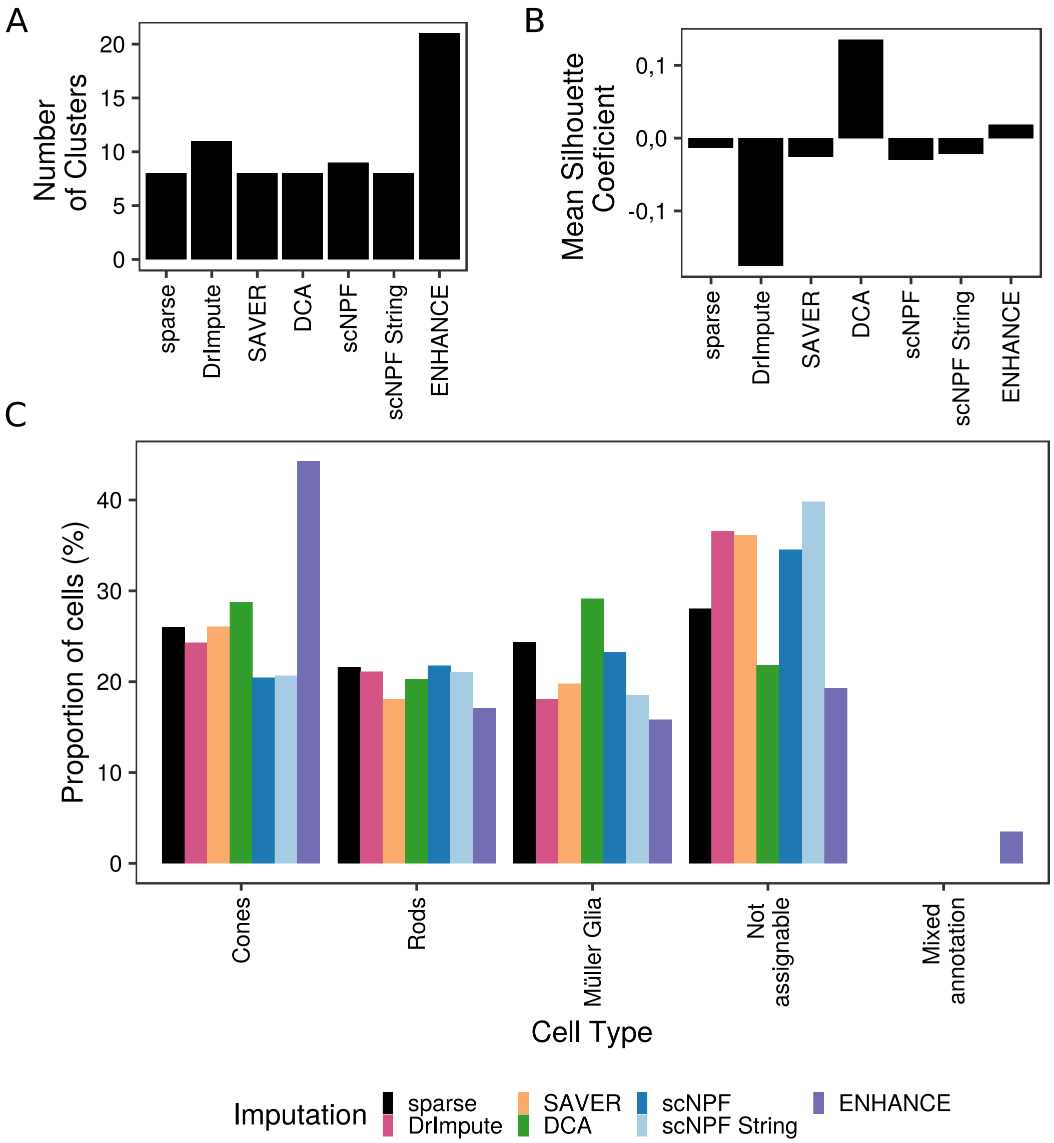


Figure 3.10. *Effect of imputation on cell cluster annotation in human retina organoid data. (A) The barplot shows the number of Louvain clusters for sparse and imputed data, obtained as described in Figure 3.9. (B) The bar plot shows the mean silhouette coefficient indicating the overall cluster density. (C) Result of the automated cell cluster annotation procedure based on the expression of known marker genes. Expression values per gene and cluster were scaled between zero and one, and binarized (threshold: 0.5). When 75% of the marker genes representing a certain cell type were expressed in a cluster, this cluster was annotated to the respective cell type. The barplot visualises the percentage of cells annotated as rods, cones, or Müller Glia (MG) cells as well as the percentage of clusters with a mixed, or without a cell type-specific annotation.*

feasible.

Generally, across all imputation tools, cells were assigned to all included retina-specific cell types (cones, rods and MG cells). While in the sparse data 26, 22, and 24% of the cells were annotated as cones, rods, and MG cells, respectively, most imputation tools achieved comparable quantities. Similar ratios were detected after `scNPF` and `scNPF`-String, though lower relative values were calculated. However, `ENHANCE` produced the most diverging results concerning both the cell type ratio, and the relative abundances. Here, percentages of 44% cones, 17% rods and 15% MG cells were calculated.

A higher percentage of MG cells, compared to the sparse data was stated after `DCA`-imputation with 29% (compared to 24%). Overall, similar quantities of rods were detected across all imputation tools ranging from 17% (`ENHANCE`) to 22%(`scNPF`). Concerning the other photoreceptor cell type, the lowest cone abundance was stated after both `scNPF` approaches with around 20%. `SAVER` achieved an equal percentage of cones, compared to the sparse results.

After stating that `ENHANCE` produced the most divergent cell type annotation compared to the sparse data, `ENHANCE` was also the only tool that led to cell clusters with *mixed annotations*. Here 3.5% of the imputed human retina cells were assigned to this class.

The amount of *not assignable* clusters was quite different between all imputed and the unimputed data. Regarding this last class, `ENHANCE` resulted in the lowest (19%) and `scNPF String` in the highest percentages (40%). In the sparse data 28 % of the cells were not annotatable. `DCA` also reached a low percentage of non-annotatable cells with 22%.

Summarising the above-described results, cluster quantities, clustering performance as well as cluster annotatability behaved comparably before and after imputation. Again, `DCA` improved the results marginally regarding cluster density and the amount of annotated retina cell types. `ENHANCE`, on the other hand, produced a different ratio in cones towards rods and MG cells.

3.5 Discussion

Single-cell omics approaches may provide a unique opportunity to gain unprecedented insights into cell type-specific regulatory programs via the inference of cell type-specific networks and the comparative analysis of these networks between cell types or states. Here, it was investigated to what extent the sparsity or dropout observed in scRNA-seq data interferes with correlation network inference and the identification of cell types using marker genes and whether imputation approaches improve these tasks.

Investigating the effect of dropout on network inference in single-cell data requires reference data sets with defined levels of dropout. Since all experimentally generated data sets are afflicted with certain degrees of dropout, this goal can only be achieved with a synthetic data set. Though several single-cell data simulation tools were available at the beginning of this thesis, none of those were specifically designed to deliver data with a proper gene correlation structure, which is an essential requirement of correlation network inference.

Therefore, a downsampling approach was applied to a bulk RNA-seq data set [Peng *et al.* 2019a] to generate a non-sparse single-cell like gold data set. Based on the results, synthetic data proofed to have suitable gene-gene correlation properties that enabled the inference of networks with an approximately scale-free topology. This data set contains per definition only true zeros, and hence it was feasible to subsequently produce data sets with defined increasing degrees of dropout.

These reference data sets enabled a thorough investigation of the impact of data dropout on vanishing network information. The module preservation measure, although dependent on module sizes, clearly indicated a negative correlation between dropout levels and the preservation of modules. A similar trend was stated by Zhang & Zhang [2018] where the sum of squared errors increased while Pearson's correlation coefficient decreased with higher ratios of dropout.

It was demonstrated that retrieving reliable and meaningful biological gene networks from low dropout scRNA-seq data is still maintainable (up to 55% dropout). data sets with dropout levels beyond 75% or even 84%, which resemble, e.g., typical up-to-date 10X Genomics data output, are fairly inappropriate for network inference analysis. An option to overcome this situation could be to manipulate the data prior to the network inference, for example, through data imputation. This might help to potentially lift the aforementioned restrictions to finally take advantage of the higher resolution of the source data.

Diving deeper into the potentials of imputation approaches, seven different methods were utilised to preprocess the six dropout-afflicted data sets. By calculating the log2-FC of the $Z_{summary}$ values, it was able to investigate the question if those methods allowed to regain at least parts of the buried gene correlation structure, and hence enabled the usage of scRNA-seq data for (sub)network inference.

In general, it was observable that the impact of imputation was highly dependent on the applied method and even more so on the dropout level of the source data sets. In addition, the results suggest that most algorithms alter the complete gene correlation structure instead of restoring previously hidden information (see Supplementary Figure S-3).

Most imputation tools did not improve module preservation especially for low dropout levels (up to 46%) with the exceptions `DCA` and `DrImpute` which achieved comparable results. However, for such low dropout levels, even `DCA` and `DrImpute` fell short of unimputed data in terms of precision of edge recovery suggesting that low dropout data sets barely benefit from imputation. Platforms such as Smart-seq2 are only moderately afflicted by dropout [Ziegenhain *et al.* 2017]. Hence it is proposed to stick to the original data instead of applying data imputation for such low-dropout data. For intermediate levels of dropout (up to 66%), imputed data sets revealed better module preservation compared to the sparse data. While still `DCA` and `DrImpute` appear to preserve modules best, the highest precision and recall rates of `DCA` compared to all competitors highlighted that `DCA` might be the most suitable option to recover hidden gene correlations in moderately sparse data sets. Beyond 75% of dropout, none of the imputation tools was able to approximately restore the true gene correlation structure.

Taking into consideration the results of module preservation and edge recovery, it is proposed to infer networks directly from low dropout single-cell transcriptomics data and use `DCA`-imputed data for intermediate levels of dropout. Although Andrews & Hemberg [2019] declare `SAVER` as the 'safest' option for data imputation, this statement cannot be supported based on these findings. The results of the edge recovery analysis, for example, indicated fairly low precision and recall values for `SAVER`. In the case of high dropout levels, the value and benefit of network inference is simply not given either with the original sparse data or with current imputation methods.

Next to reliable network inference, identification of specific cellular populations via cell clustering and annotation of clusters is the other important requirement for studying cell type-specific networks. One frequently employed approach for cluster annotation is relying on unique expression patterns of known marker genes. After finding that imputation methods tend to induce false positive signals into (high dropout) gene-gene correlation data, additionally, the effect on the correlation between cells was investigated. Intrinsically, the unique expression profile should prevent marker genes from being too prone to dropout, which makes an imputation not necessarily required prior to cell type annotation. However, in case imputation tools have been applied, it is of major importance that the imputation methods have no negative effect on the overall expression profiles of marker genes. Andrews and Hemberg already pointed out that the reproducibility of marker genes tends to be reduced upon data imputation when extracting and comparing marker genes before and after imputation [Andrews & Hemberg 2019]. Here, the impact of imputation on known marker genes was investigated on a biological data set.

Clusters without clear distinctions between one another were obtained both for sparse data as well as most imputation tools, except for minor improvements for `DCA`. These findings suggest that imputation did not strongly improve cell type separation in low dimensional space. This is in contrast to previous findings by [Eraslan *et al.* 2019], who used a synthetic data set with lower levels of dropout, which may explain the discrepancy.

Based on the tSNE and Louvain cluster detection, an automated annotation pipeline was implemented to compare quantities of annotated retina-specific cell types before and after imputation. All over, annotation of pure retina clusters was only marginally improved using `DCA`. All remaining tools resulted in fewer retina-specific cell types compared to the unimputed data, raising concerns about their applicability and usefulness in this respect. Looking at this analysis part alone, the results suggest that `DCA`, as well as `ENHANCE`, allow to annotate marginally more retina cells and can thus improve, for example, cell type quantification. However, `ENHANCE` produced a distorted low dimensional cluster embedding as well as cell type ratio compared to the other data sets.

In general, it was demonstrated that the applied imputation tools maintain the usability of marker genes leading to comparable quantities of annotated cell clusters. Based on this analysis, again `DCA` was able to slightly optimize both, cell clustering and annotatability compared to the sparse situation. Here, no tool led to an eradication of the marker gene profiles, but they have not been found to add certain definiteness to the problem.

Summing up all results, using a benchmarking framework based on a downsampled bulk data set, the effects of data imputation prior to network inference was investigated over various dropout levels. While major network structures were still preserved in the low dropout data sets, none of the included imputation tools helped to infer gene correlation networks from high dropout data sets. `DCA` outperformed the unimputed data with respect to network preservation and edge recovery precision on both moderate dropout data sets.

These results suggest that in data sets owning a moderate range of dropout, indeed `DCA` may allow to infer cell type-specific gene correlation networks.

Chapter 4

Imputation of cell type-specific gene regulatory networks in human retina organoids

In this chapter, the encouraging results of the previous analysis were further investigated. The systematic evaluation of six different imputation approaches and their effect on gene correlation networks highlighted that `DCA` allowed inferring meaningful network structures from moderately dropout affected data sets. This might represent a *window of opportunity* for inferring cell type-specific gene correlation networks. `DCA` was therefore applied to a human retinal data set [Kim *et al.* 2019] to subsequently infer gene correlation networks, and the differences between cell type specific networks were analysed and quantified. In a final step, selected gene networks between both photoreceptor cell types (rods and cones) were created and compared.

4.1 Workflow

To investigate how `DCA` influences the inference of (cell type-specific) gene correlation networks, different analysis steps has to be considered. While still focussing on the whole network topology, it was analysed in a first step, if a scale-freeness was achievable in the unimputed and `DCA`-imputed data set. Diving deeper towards cell type-specific networks, the `DCA`-imputed cells were annotated via the marker genes provided in the original publication by Kim *et al.* [2019]. The inferred cell types were then in a final step used to derive cluster-specific networks which were biologically characterized and compared to another. An overview of this workflow is depicted in Figure 4.1.

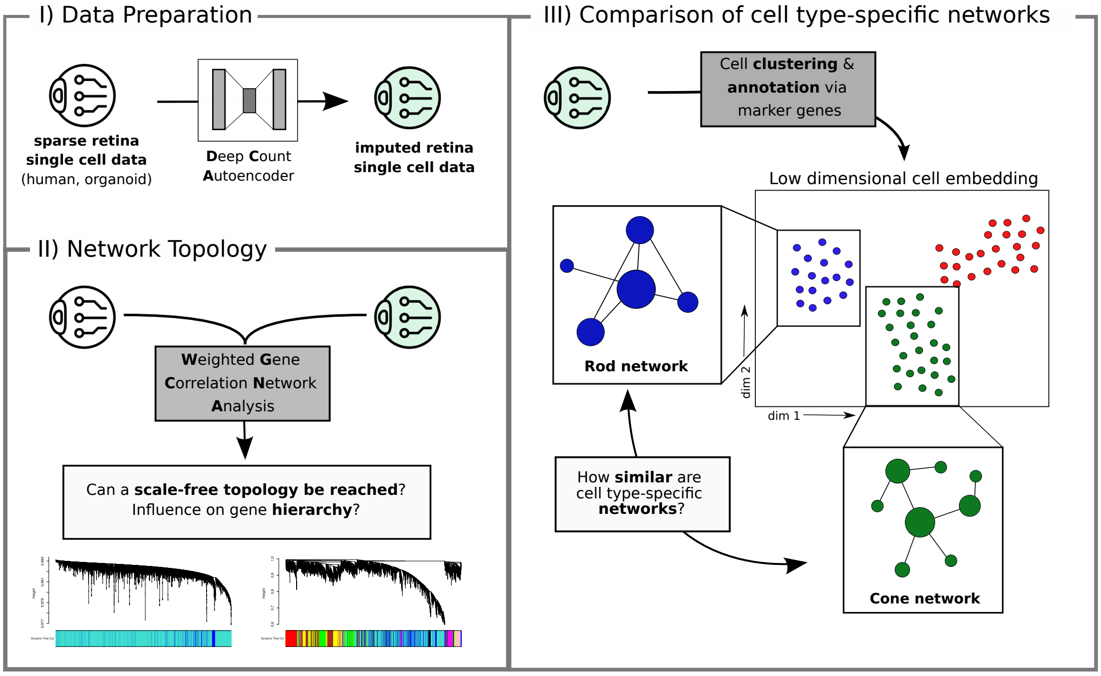

I) Data Preparation
sparse retina single cell data (human, organoid)
Deep Count Autoencoder
imputed retina single cell data
II) Network Topology
Weighted Gene Correlation Network Analysis
Can a scale-free topology be reached? Influence on gene hierarchy?
III) Comparison of cell type-specific networks
Cell clustering & annotation via marker genes
Low dimensional cell embedding
Rod network
dim 2
dim 1
How similar are cell type-specific networks?
Cone network

Figure 4.1. *Evaluating the influence of `DCA` on inferring gene correlation networks from a human retina organoid data set [Kim* et al. *2019].*
After an initial data preparation and imputation step (I), the ability to infer scale-free topology networks was investigated for both the original and `DCA`-imputed organoid data set utilizing `WGCNA` (II). Subsequently, the cone- and rod-specific clusters were annotated, allowing to infer and compare the cell type-specific correlation networks (III).

4.1.1 Data preparation

As described in the workflow Figure 4.1, a human retina organoid data set by Kim *et al.* [2019] was used. This data set contained 19426 genes across 1346 cells and included 85% zeros.

As already described in section 3.1.2, `DCA` operated on the sparse count data directly. Again, version 1.3.1 of the tool was used, employing the standard conditions of the implementation. The `DCA`-imputed and sparse data were preprocessed and annotated subsequently.

In their original study, Kim *et al.* clustered and annotated the cells w.r.t. the major cell types in the retina: rods, cones, and Müller Glia cells. This was achieved via a set of unique marker genes, which were provided in the original publication. As neither an annotation vector nor a script was provided by Kim *et al.*, these marker genes were used to perform a cell cluster annotation according to the method section described in their paper.

`Seurat`-preprocessing. Briefly, the retina single-cell data sets were preprocessed as indicated in the original publication using the `Seurat` environment (version 3.1.5.). Upon data import, cells with less than 600 expressed genes, and genes that were not detected in at least 67 cells were discarded. These thresholds were derived from the publication. Subsequently, the count data was normalized, log-transformed, and variable genes were extracted, using default parameters as indicated in the tutorial*. Prior to running the principal component analysis (PCA), the logged-expression data was scaled. Cells were subsequently clustered using the Louvain algorithm on the t-distributed stochastic neighbour embedding (t-SNE). Finally, the expression of the cone-, rod-, and Müller Glia specific marker genes was analysed across the detected Louvain clusters. The final cell cluster assignment was based on the expression patterns of the marker genes within the Louvain clusters. An overview of the annotation result is depicted in Supplemental Figure S-5. All annotation results can be extracted from the uploaded `Seurat`-objects. More details will be provided in Section 7.1. Here, cells were either annotated to *pure* retina cell types (rods, cones, MG cells), *mixed* retinal signals or *unassignable* clusters.

4.1.2 Network inference by means of `WGCNA`

In this chapter, again the network inference tool `WGCNA` was used. Generally, networks were inferred for, both the complete data sets and for the respective cone and rod subclusters. Re-

---

[*] `https://satijalab.org/seurat/articles/pbmc3k_tutorial.html`

garding the cell type-specific networks, only clusters with an unambiguous cluster assignment were used. Unambiguity here refers to pure retinal cell types.

After defining the optimal β value, gene modules were detected using a minimal cluster size of 100, while all other parameters were left on default. The defined criteria on how the variable β should be approached were discussed in detail in section 2.4.1 of the previous chapter. Based on the cluster annotation, cell type-specific networks were inferred using the same `WGCNA`-workflow.

4.1.3 Network characterization

After inferring cell type-specific networks from the `DCA`-imputed data, their biological information was extracted and evaluated. While comparing developmentally and functionally very similar cell types, it was analysed if cone- as well as rod-specific modules could be identified. Furthermore, it was investigated if similarity important genes across both cell type-specific networks were associated to different correlation networks.

GOI-module detection. In order to identify modules that were highly specific for the cone or rod networks, the GOI-genes provided in the Kim *et al.* publication were used. Approaching the importance of each gene inside the modules, the module membership (MM) per subnetwork was calculated according to the `WGCNA`-tutorial. An explanation of the MM can be found in Section 2.4.1.

Subsequently, the genes per module were ranked by the absolute MM values, such that the highest MM value got assigned the highest rank. Finally, the resulting distribution of these GOI-ranks across all modules was used to select cone- and rod-specific modules.

Hub gene preservation. As mentioned previously, subnetworks derived from functionally similar cell types were compared. Thus, it is expected that many regulatory components are shared across networks. To investigate, how many signals were shared, the preservation of hub genes was investigated. In this context, genes with the highest MM-values within a module were referred to as hub genes. Generally, two different approaches were used. In total, 220 hub genes (20 hub genes × 11 modules) were analysed for both the rod and cone network. To generally analyse their preservation as key regulators in the opposing network, their maximal MM value score across all modules was derived and compared.

Differential network analysis. Using another approach, not only preserved hub genes were evaluated but the networks they regulate. For this approach, the top 30 highly connected genes per module per photoreceptor subnetwork were extracted. The number of top hub genes was increased to enlarge the number of candidate genes. Intersecting genes were considered as preserved. For each preserved hub gene (target), two correlation subnetworks were extracted - one for each retinal cell type. To do so, the top 300 genes with the highest adjacency value towards the target gene were extracted. The target gene itself was added manually.

For target-specific networks between both photoreceptor classes, the differences in connectivity and shared genes were calculated. Using the `intramodularConnectivity`-function by `WGCNA` the connectivity k within the respective module was calculated per target across both networks. The Tanimoto similarity was used to quantify the network similarity across cell types (see Equation 4.1). This measure calculates the ratio of common genes (intersect) and the union of all genes in both networks.

Tanimoto similarity:

$$\frac{cone\ genes \bigcap rod\ genes}{cone\ genes \bigcup rod\ genes} \tag{4.1}$$

Network visualization. Selected gene correlation networks were also visualised in this dissertation. To do so, a smaller target subnetwork was calculated as described earlier, with the exception that only 30 genes plus target were included. The network information was visualised via the `Cytoscape` (version 3.8.2) tool. There, the gene layout was generated via the *Edge-weighted Spring Embedded Layout* algorithm, whereby the weight represented the absolute adjacency values. Furthermore, the spring strength and rest length were adjusted to 100 and 150, respectively. All other layout setting were left on default.

4.2 Network inference after `DCA` imputation

Benchmarking six different imputation algorithms highlighted, that the imputation quality is highly dependent on the tool itself, and the level of dropout. Among the included tools, `DCA` showed the overall best performance on moderately sparse synthetic data sets. Therefore, `DCA` was applied on an experimentally derived retinal organoid data set by Kim *et al.* to infer gene correlation networks.

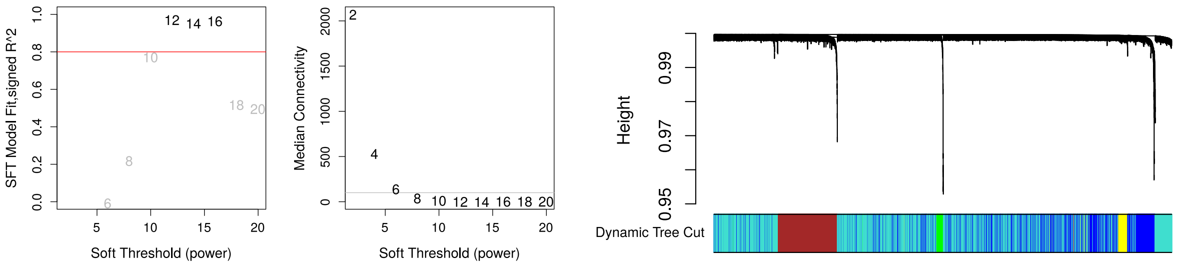
4.2.1 `DCA` imputation allows to generate scale-free topology networks.

The development of the SFT R^2 model fit over twenty β values in the whole unimputed Kim data set is shown in Figure 4.2a. Section 2.4.1 described the workflow on how the β value was selected. As it can be seen in 4.2a, a SFT R^2 model fit over 0.8 was reached with a β value of at least twelve. The median connectivity development is shown next to it.

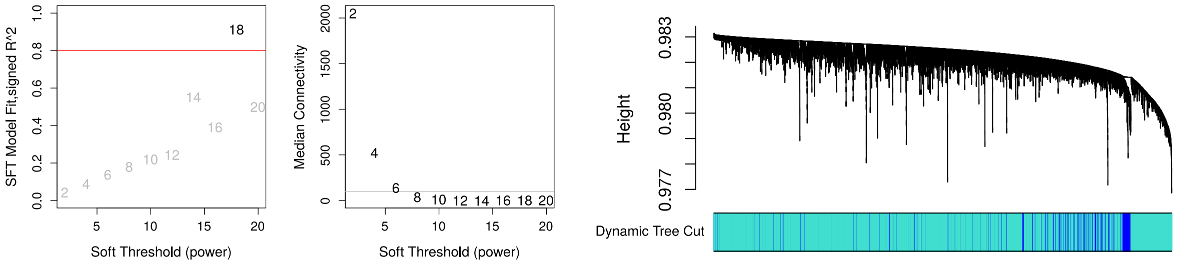
Here, a β of twelve was used to construct a gene dendrogram. As depicted in Figure 4.2b, a dendrogram with a shallow hierarchy with solely five modules was detected. Scooping the full potential of the single-cell transcriptomics data, cell type-specific gene correlation networks were inferred. Therefore, the marker genes derived by Kim *et al.* were used to annotate the sequenced cells. In the following step, it was investigated if an SFT-model fit can be achieved in these subnetworks.

The results are depicted in Figure 4.2c and Figure 4.2e for cones and rods, respectively. In both cases, the overall SFT model fit proofed to be low with only a peak in the rod data at a β of 18. Since both data sets should be compared to each other, the same β value was used. As stated earlier, if a model fit below 0.8 was revealed, a median connectivity above one hundred should be approached. Therefore, a β of six was used to infer both cone and rod subnetworks. As it can be seen in Figure 4.2d as well as 4.2f, again a shallow gene hierarchy with only two modules was detected.

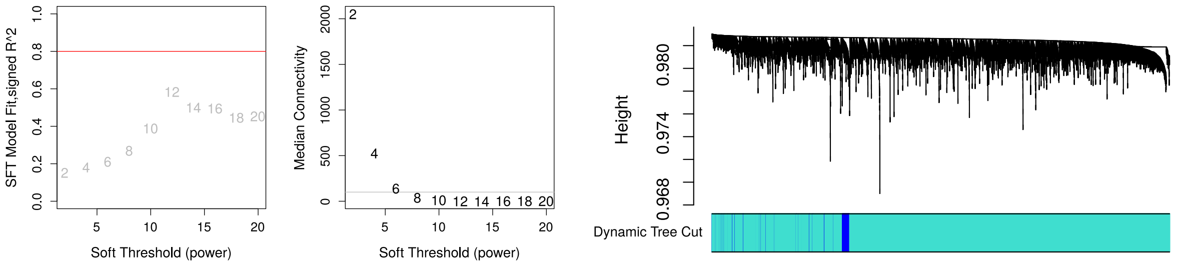
After investigating the unimputed organoid data, it was analysed if and how `DCA` imputation influences network inference. Therefore, the whole data set was used as an input for the `DCA` pipeline. Again, it was analysed in a first step if an SFT R^2 model fit can be achieved for the whole as well as the cell type-specific networks.



(a) *SFT-model fit, whole unimputed* **(b)** *Cluster dendrogram, whole unimputed*

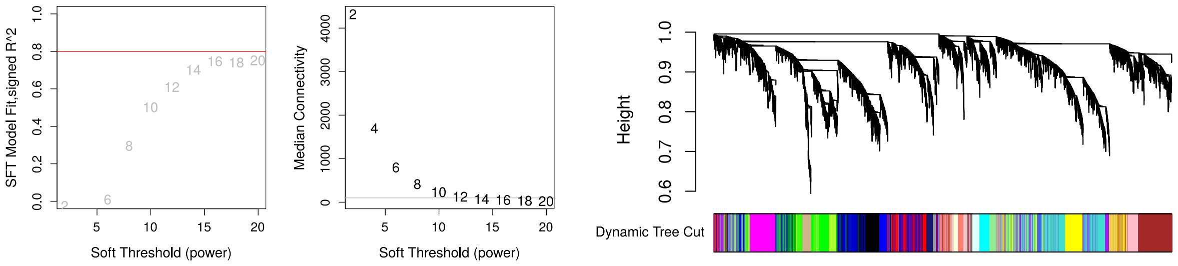


(c) *SFT-model fit, cone-subdata unimputed* **(d)** *Cluster dendrogram, cone-subdata unimputed*



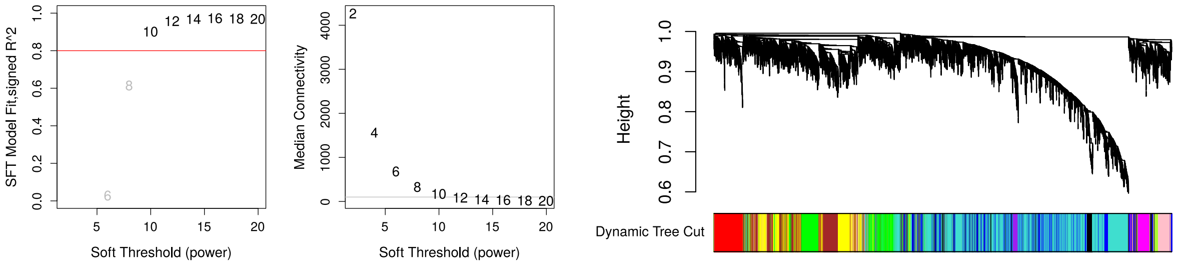
(e) *SFT-model fit, rod-subdata unimputed* **(f)** *Cluster dendrogram, rod-subdata unimputed*

Figure 4.2. *Scale-free topology estimation and gene cluster dendrograms for weighted gene correlation network analysis in the sparse organoid retina data.*
(a,c,e)Scale-free topology model fit and median node connectivity over twenty β values. (b,e,f)Gene cluster dendrograms. Data visualised corresponds to the complete data (a,b) as well as the cone(c,d) and rod(e,f) subsets. No SFT R^2 model fit value larger than 0.8 was determined for both cell type-specific networks. Besides, all resulting gene cluster dendrograms exhibited a shallow hierarchy with only a small quantity of detected modules.



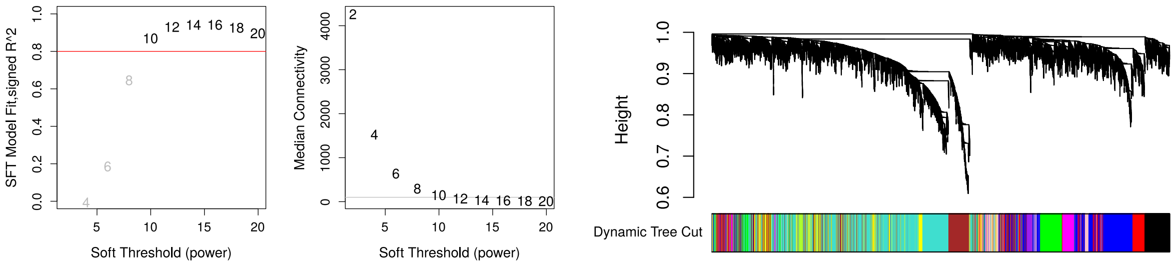
(a) *SFT-model fit, whole* `DCA`*-imputed*

(b) *Cluster dendrogram, whole* `DCA`*-imputed*



(c) *SFT-model fit, cone-subdata* `DCA`*-imputed*

(d) *Cluster dendrogram, cone-subdata* `DCA`*-imputed*



(e) *SFT-model fit, rods-subdata* `DCA`*-imputed*

(f) *Cluster dendrogram, rod-subdata* `DCA`*-imputed*

Figure 4.3. *Scale-free topology estimation and gene cluster dendrograms for weighted gene correlation network analysis in* `DCA`*-imputed organoid retina data.*

(a,c,e)Scale-free topology model fit and median node connectivity over twenty β values. (b,e,f)Gene cluster dendrograms. Data visualised corresponds to the complete data (a,b) as well as the cone(c,d) and rod(e,f) subsets. While no SFT R^2 model fit larger than 0.8 was determined in the complete `DCA`*-imputed retina organoid data set, still a deep hierarchy was detected in the corresponding cluster dendrogram. The cone- as well as the rod-specific gene correlation networks highlight a SFT R^2 model fit above 0.8. Similar to the complete network data, a deep hierarchy was detected.*

As it can be seen in Figure 4.3a, similar to the unimputed data, none of the included β values reached a model fit above 0.8. Given the β-selection procedure, a median connectivity above one hundred was aimed for. Here, a β value of 12 was selected to construct a gene cluster dendrogram for the whole, `DCA` imputed data set.

Figure 4.3b illustrates the gene hierarchy within the whole `DCA` imputed organoid data set. In contrast to Figure 4.2b, a deep gene structure could be uncovered, giving rise to twenty modules with sizes ranging from 124 to 1993 genes.

This trend was even more pronounced in the cell type-specific subnetworks. Based on the annotation via marker genes, a cone-specific subset of the imputed organoid data was generated. Figure 4.3c shows the SFT model fit of this subpopulation. Here, a β value of twelve resulted in a sufficiently high R^2 value, which was close to the plateau phase.

Based on this β value, a gene dendrogram was calculated (Figure 4.3d). As already seen for the whole organoid data set, a deep hierarchical structure was achieved. In total eleven modules were detected with 220 genes in the smallest and 5176 genes in the biggest module.

Consistent and comparable results could be achieved for the rod-specific subpopulation. Next to an optimal model fit for a β value of twelve (see Figure 4.3e), a meaningful gene dendrogram could also be constructed (Figure 4.3f). In this rod-specific subnetwork, eleven modules were detected containing a minimum of 161 and a maximum of 4590 genes. Though the same number of modules was detected as already highlighted in the cone subnetwork, there is no biological connection between the module namings.

Summing this part up, it can be seen that prior to imputation, generally no optimal, consensus SFT fit could be determined. When calculating the gene dendrograms, no hierarchy was detected. This result was found for the whole as well as for the cell type-specific subnetworks. After `DCA` imputation of the whole organoid data set, the SFT fit was better approachable. Although no SFT model fit did exceed a value of 0.8 in the overall network, the typical plateau-phase was observable, resulting in a deep hierarchical structure within cluster dendrogram. Both cell type-specific subnetworks highlighted an optimal SFT fit with a β of twelve, as well as hierarchically structured gene dendrograms revealing eleven modules. Allover no grey module was detected which usually contains the lowly correlated genes.

4.2.2 Specific modules for rods and cones can be identified in cell type-specific networks

After stating the influence of `DCA` on network topology, the following analysis steps will evaluate whether cell type-specific networks differences were identifiable. In the first step, rod-as well as cone-specific modules were identified in the respective cell type-specific networks, again using the GOI lists that have been used previously to annotate the cell clusters.

For each gene, a module membership (MM) score was calculated, by correlating the individual genes to the module eigengenes. MM-values range from minus one to plus one, with values on both ends of the spectrum representing highly (anti-) correlated genes, and hence indicate their importance for the module. It is assumed that GOI genes, which were used to annotate the rod- and cone-specific cell clusters play a regulatory role in the respective cell type. Therefore, these genes should have some hub gene-like properties, which would be associated with low ranks in MM. Here, the absolute MM-values were used to calculate a ranking of the importance of each gene in each module.

Figure 4.4 shows the distribution of ranks across all detected modules for the cone- and rod-specific GOIs within their respective cell type-specific networks.

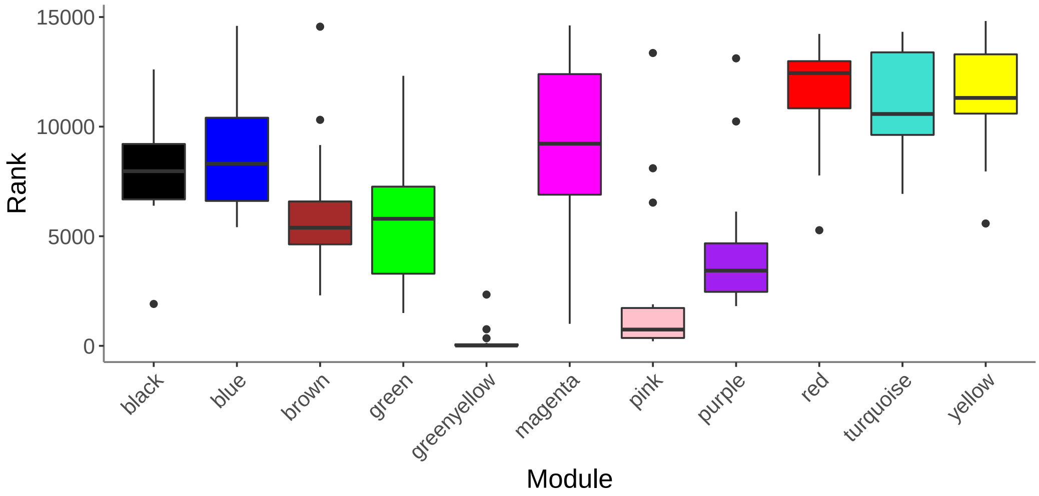
For the rod network (Figure 4.4(A)) as well as for the cone network (Figure 4.4(B)), one module could be identified that accumulated GOIs with an overall lower median rank. Rod GOIs were found to play an important role in the greenyellow module in the rod network, while cone GOIs were found to be important in the purple module in the cone network. Although other modules such as the pink (in the rod network) or the red (in the cone network) owned generally low ranks as well, the variance was larger compared to the aforementioned modules.

Assuming that the greenyellow module from the rod network contained mainly rod-specific network configurations, no similar network configuration should be detected in the cone-specific network. The same holds true for the other direction meaning that cone-specific genes from the purple module, should not accumulate high ranks in modules in the rod network.

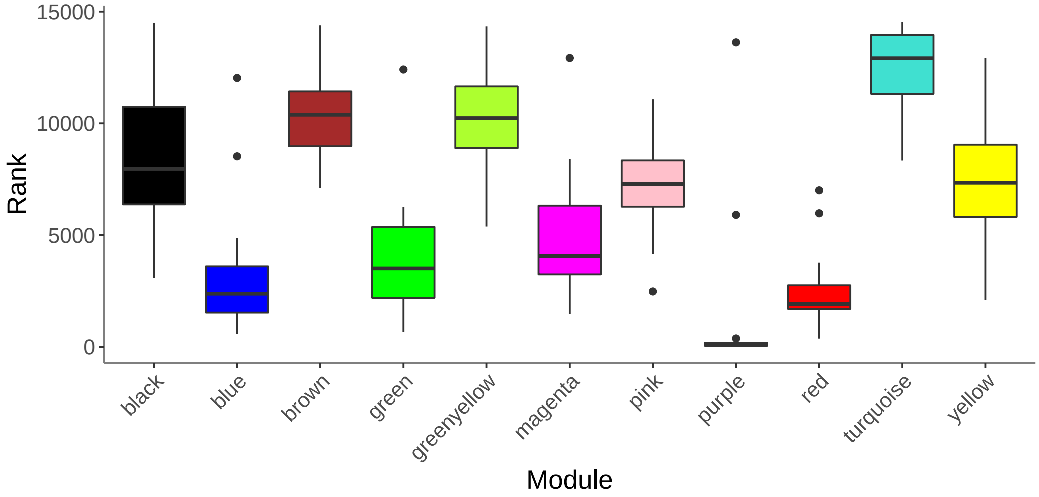
To examine the degree of truth behind this hypothesis, the MM-values of the top 20 genes from the cell type-specific modules were compared across all modules of the opposing network. The result is depicted in Figure 4.5. In cases where these hub genes would retain their importance in the respective other cell type's network, they would again accumulate high MM-values in some modules.

Regarding the greenyellow hub genes, originating from the rod network, most MM-values were found to be close to zero across all modules in the cone network. Solely the cone-magenta and cone-red module showed slightly higher median MM values. However, no module was identified that owned consistently high MM-values.

For the purple hub genes, originating from the cone network, the results were slightly different in the rod network (see Figure 4.6b). Here, more modules revealed elevated median MM-



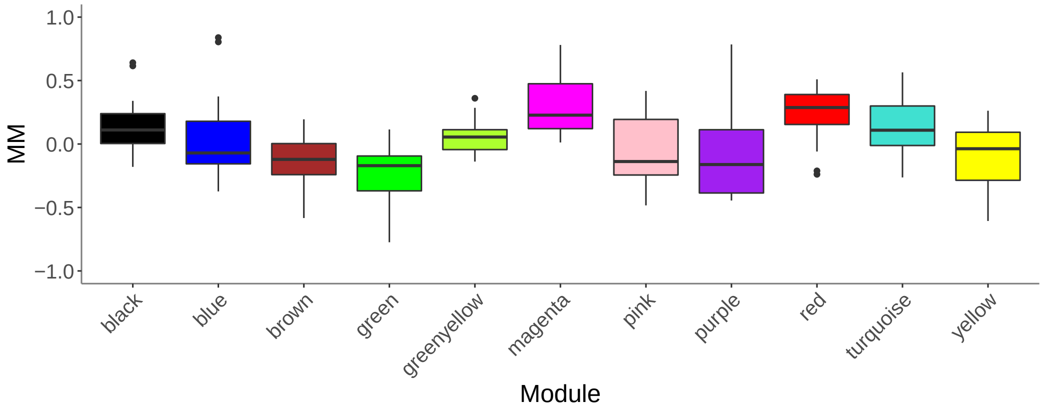
(a) *Rod network*



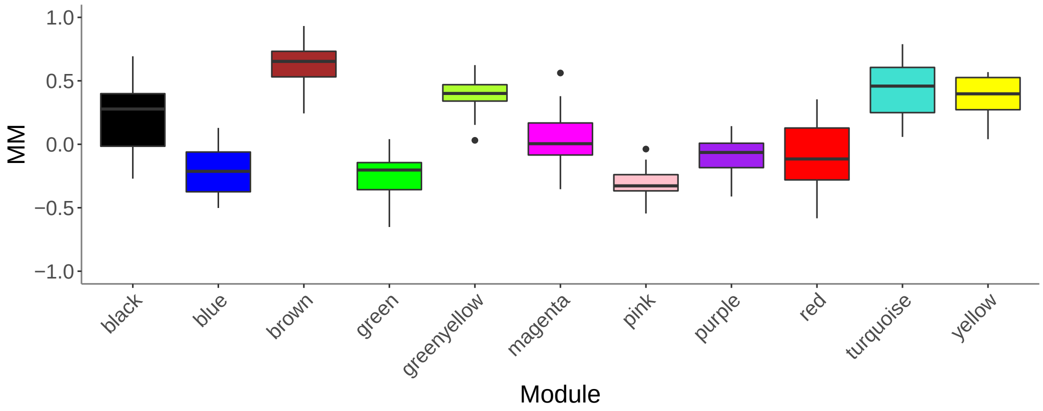
(b) *Cone network*

Figure 4.4. *Distribution of GOI ranks in the `DCA`-imputed retina sub-cell type networks. Using the correlation of genes towards the first principal component of the modules, called module membership (MM), the importance of the GOI can be stated across modules. Therefore the absolute MM values were translated into ranks. Lower ranks hereby correspond to higher importance or relevance. For the rod (a) as well as cone (b) sub-networks, clearly one module can be highlighted. Whereby the greenyellow module contains the lowest median ranks for the rod network, similar results were found for the purple module in the cone network.*

values. Though the rod-brown module even showed a median value of 0.65, it remained lower than the median MM-values of inside the original cone-purple module (0.89).



(a) *Rod-module hub genes in cone network*



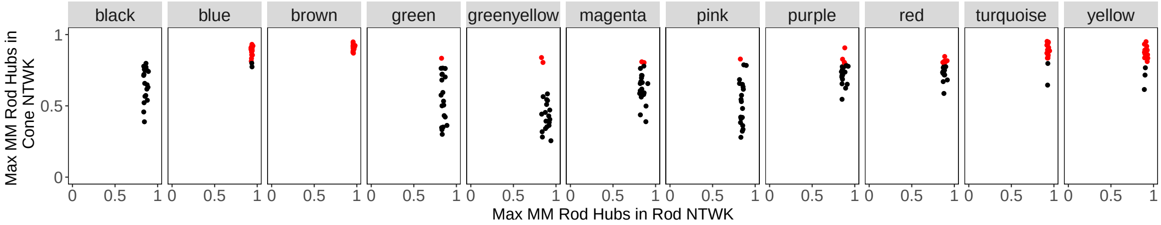
(b) *Cone-module hub genes in rod network*

Figure 4.5. *Distribution of cone and rod specific module membership values in complementary network.*
After identifying the cone and rod specific modules in the respective network, the uniqueness of the hub genes in these modules is analysed. Therefore,the top 20 MM-genes per module are extracted and the distribution of their MM-values is compared across all other modules in the opposite network. Both, the MM values of the rod-greenyellow hub genes in the cone network (a), as well as the cone-purple hub genes in the rod network (b) show no module with high MM-values but low variance.

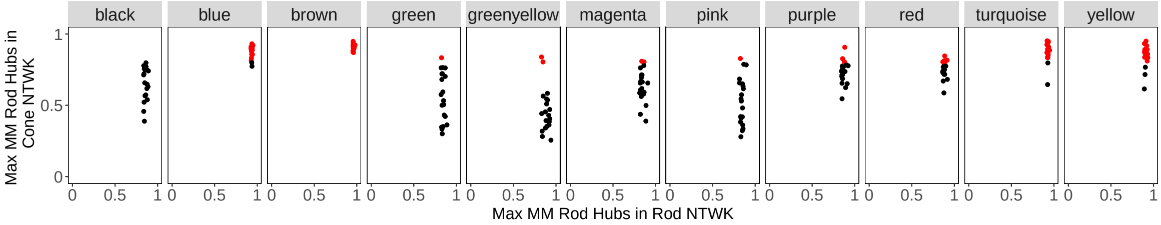
In summary, cell type-specific networks from the rod- and cone subpopulation highlighted one module, where the respective GOIs contained a high importance. These results were unique for the respective network, as the hub genes of the cone-purple and the rod-greenyellow module were not preserved in their status in the opposite network.

After stating that cell type-specific configurations of the rod- and cone networks remain unique, it was investigated how many hub genes preserve their hub status at all. Similar to the previous analysis, hub genes which remain important should retain a high MM value in any module of the opposite network. Therefore, not only the hub genes of the GOI-modules were analysed in the opposite network, but the hub genes across all modules.

To highlight the maximal importance these hub genes can reach in the opposite network, the highest scoring MM-value across all modules was extracted. The results are indicated in Figure 4.6. In total 220 hub genes of each cell type-specific network were extracted. Generally, whereby the hub genes of the blue, brown, turquoise, and yellow module from the rod network appeared to score high MM values in the cone data, black and greenyellow hub genes showed the opposite result. In the cone network, however, hub genes from the black, blue, greenyellow, and turquoise modules revealed many MM values above 0.8 in the rod subnetwork. Green and magenta hubs however indicated lower max-MM-values. In concordance to the analysis before, the rod- and cone-specific modules indicated lower MM values across networks.



(a) *Rod-hub genes in Cone network*



(b) *Cone-hub genes in Rod network*

Figure 4.6. *Maximal module membership values of hub genes in the complementary networks. The distribution of module membership(MM) values from 20 hub genes per rod (A) and cone (B) network in the opposite network. Red dots indicate a MM value above 0.8 in the opposite network. While some hub genes were preserved across network, such as rod-brown and cone-black, other hub genes were less preserved. Especially the cone-green and cone-magenta hubs were more unique for the cone network.*

After stating that some hub genes conserve their hub gene status across networks by scoring high MM-values, it was investigated if the preserved hub genes were associated to similar gene correlation networks. Therefore, the top 30 hub genes of each module in both photoreceptor networks were extracted. Genes occurring in both networks were considered as preserved. In a following step, the gene correlation networks around these hub genes were identified in each subnetwork and compared using the Tanimoto similarity. A Tanimoto similarity close to one indicates similar gene sets, whereby values closer to zero state the opposite. An overview of the results is provided in Table 4.1.

In total, 35 hub genes were found to be preserved between the cone and rod subnetworks. However, the Tanimoto similarity differed widely across the surrounding correlation networks. RRN3P2-, ALLC-, ADAMTSL3-, and RHOF-specific networks, for example, owned a higher Tanimoto similarity of above 0.6, indicating that these hub genes are densely connected to the same set of genes in both networks. other gene networks TBC1D10A, PPL, and EIF4EBP1 revealed the opposite trend, since only small similarity values of 0.14, 0.09 and 0.08 were calculated, respectively.

Moreover, larger differences within the intramodular connectivity (K_{IM}) values of the target genes were detectable. The K_{IM} values were calculated to quantify the compare the hub gene-status of the target genes across networks, since a fixed number of high-connected genes was extracted. Generally, the highest K_{IM} values were identified in the cone network for LANCL3 with 669. In the rod-specific network, MUC16 reached the highest K value with 414.4. While 23 genes revealed K values below 100 in the cone-specific data, only 8 were detected in the rod data.

It was found that the activation of the ErbB signaling pathway was associated to an eye disease called age-related macular degeneration [Sheu *et al.* 2019]. EIF4EBP1 exhibited to be an interesting candidate for a detailed investigation, as it is a member of the ErbB-signaling pathway, and is part of different subnetworks in rods and cones. Therefore, smaller subnetworks of solely 30 genes plus the target EIF4EBP1 were generated, to ensure a visual inspection of the network structure. Again, only a small intersect of three genes was detected with a Tanimoto similarity of 0.051 within this smaller subnetworks. The resulting networks are visualised in Figure 4.7 and Figure 4.8.

As already indicated by the small Tanimoto similarity, both EIF4EBP1-subnetworks reveal very different layouts. Though RAD23A and ZNF146 were found in both networks, their roles and connectivity changed exceedingly. Whereby both genes were highly connected to the target in the cone network, they illustrate less strong interactions in the rod data. While the target gene EIF4EBP1 pointed out a stronger relative correlation towards a group of highly

Table 4.1. *Overview of preserved hub genes.*

From the cone and rods specific network, the top 30 hub genes were extracted and the intersect was further investigated. Building subnetworks around these target genes (here 300 genes) from the cone and rod data, the intersect and Tanimoto similarity was calculated. Furthermore, the intramodular connectivity (K) of the target gene from both data sets was defined.

Gene name	K_{IM} Target Gene (Cones)	K_{IM} Target Gene (Rods)	Intersect	Tanimoto Similarity
EIF4EBP1	88.90	245.60	47	0.08
PPL	129.20	277.50	52	0.09
TBC1D10A	189.40	250.40	75	0.14
SOX13	132.10	252.60	109	0.22
MIR601	639.80	106.40	124	0.26
KCNA2	28.50	16.60	124	0.26
SHANK1	31.30	187.90	127	0.27
ADRA2B	29.40	197.00	128	0.27
IL17C	17.60	181.00	134	0.29
C1QTNF7	22.60	173.20	134	0.29
TMC5	30.60	16.70	135	0.29
MUC16	658.60	414.40	138	0.30
LANCL3	669.00	392.70	139	0.30
MOGAT3	39.60	18.20	139	0.30
C6orf141	18.50	183.90	139	0.30
SLC1A6	23.60	171.80	141	0.31
TCF7L1	124.80	248.10	141	0.31
NR4A1	29.00	173.00	141	0.31
CARNS1	97.60	48.30	143	0.31
KCNJ16	636.60	370.80	144	0.31
QPCT	86.90	31.60	146	0.32
TLR9	22.10	183.20	146	0.32
CNTN6	22.20	185.40	154	0.34
ASB5	27.10	183.70	156	0.35
CPT1B	36.50	16.00	170	0.39
QRFP	28.20	183.40	175	0.41
HTN1	113.40	58.60	182	0.43
GSTA7P	654.60	382.80	193	0.47
UGT8	50.10	19.70	197	0.49
HOXC11	27.60	188.20	199	0.49
RHOF	129.60	250.10	228	0.61
ADAMTSL3	21.20	171.90	235	0.64
UPB1	25.80	181.10	237	0.65
ALLC	129.70	261.10	238	0.65
RRN3P2	25.00	175.00	241	0.67

connected genes (RAD23A-EIF4EBP1-ILVBL-mir210HG-FKBP1A) in the cone subnetwork, it appeared less connected in the rod data. There, two bigger cliques of genes (OSGEP-SAP130-HAGHL and RGS5-PARP16-ZNF600-PSAT1-FAM162A) were detected. Generally, many zinc-finger proteins (ZNF) were included in both networks.

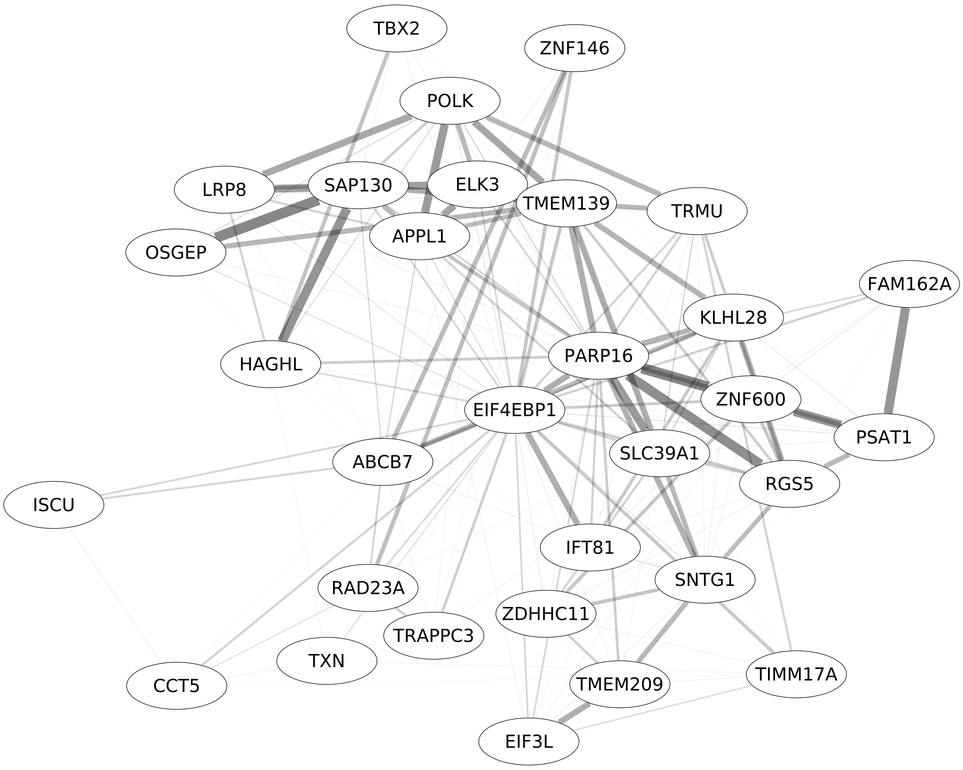


Figure 4.7. *Rod-specific subnetworks of EIF4EBP1.*

Based on the correlation value, the top 30 genes around EIF4EBP1 (target) were extracted and plotted. The width and transparency of the edges corresponds to the relative correlation strength. All included genes were coding genes.

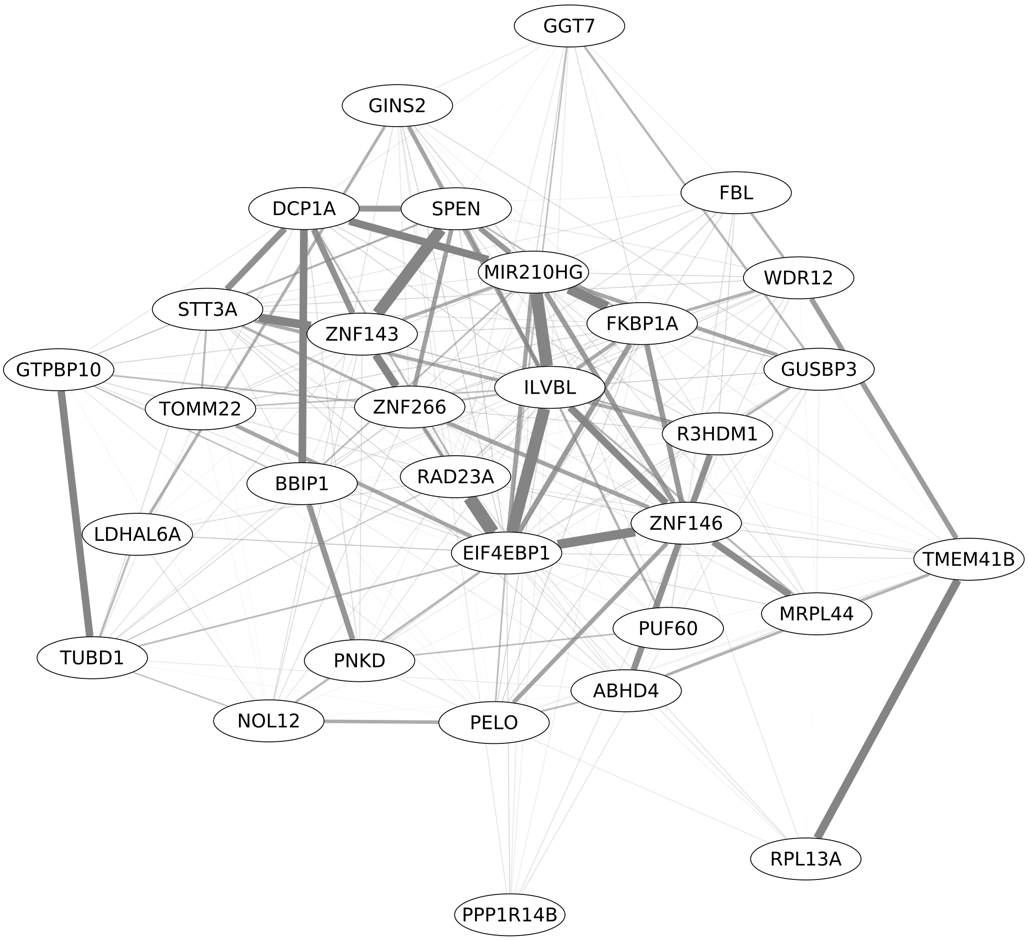


Figure 4.8. *Cone-specific subnetworks of EIF4EBP1.*

Based on the correlation value, the top 30 genes around EIF4EBP1 (target) were extracted and plotted. The width and transparency of the edges corresponds to the relative correlation strength. All included genes were coding genes.

In order to put these networks into a biological context, the `ARCHS`^{4} was search for predicted human phenotypes and GO-terms of retina-related results. An overview is provided in Table 4.2.

For the cone-specific subnetwork of EIF4EBP1, ILVBL and mir210HG were strongly connected for example. Both genes were predicted to hemorrhage associated terms, since mir210HG revealed connections to angiogenesis. Alongside, also FKBP1A was connected to mir210HG revealing an association to the TGF-β signaling pathway.

In the rod-EIF4EBP1 network, PARP16 and ZNF600 for example were predicted to be involved in either retinal dysplasia and abnormal rod and cone electroretinograms. Another stronger correlation was found between the target and IFT81, which was predicted to influence retinal rod cell development. Astrocyte associated prediction were found for EIF3L and TMEM209, which revealed a high correlation in the rod-EIF4EBP1 network.

Summing up the previous section, constructing target gene-specific networks from conserved hub genes across rod- and cone-specific subnetworks highlighted the following: While some target-specific networks were consistent across networks, others differed severely. One target of the later scenario was EIF4EBP1, which is known to be associated to the ErbB signaling cascade. By constructing smaller networks across cell types, the structure was visualised and analysed. Indeed, many predicted human phenotypes associated to retinal abnormalities were identified for various members of both networks.

Coming back to the initial aim of comparing rod- and cone-specific gene correlation networks on a biological level, it can be stated that using the included analysis steps, a groundwork was generated.

Table 4.2. *Overview of gene associations detected in the EIF4EBP1 subnetworks. Predicted human phenotypes and GO-terms of EIF4EBP1-network genes. The results were obtained from* `https://maayanlab.cloud/archs4`*, except for entries found in 'other'. All results were filtered for retina-associated terms. ERG refers to electroretinogram, a diagnostic test used to measure the electrical potential of the retina.*

Gene name	Network origin	Predicted human phenotype	Predicted GO-term	Other
ILVBL	Cone	Hemorrhage of the eye	-	-
mir210HG	Cone	-	-	Association to angiogenesis
FKBP1A	Cone	Retinal dysplasia	-	Association to TGF-β signaling pathway
RPL13A	Cone	Absent rod-and cone-mediated responses on ERG	-	-
TUBD1	Cone	Retinal dysplasia	-	-
PARP16	Rods	Retinal dysplasia	retinal cone cell development	-
ZNF600	Rods	Abnormal rod and cone electroretinograms	-	-
RGS5	Rods	Chorioretinal atrophy	retina vasculature morphogenesis in camera-type eye	-
HAGHL	Rods	Attenuation of retinal blood vessels, Retinal dysplasia	-	-
SAP130	Rods	Abnormality of the astrocytes	-	-
OSGEP	Rods	Abnormal rod and cone electroretinograms, Attenuation of retinal blood vessels, Absent rod-and cone-mediated responses on ERG, Severe visual impairment	-	-
ELK3	Rods	-	retina vasculature morphogenesis in camera-type eye	-
POLK	Rods	Attenuation of retinal blood vessels, Abnormal rod and cone electroretinograms	positive regulation of astrocyte differentiation	-
TMEM139	Rods	Pigmentary retinal degeneration, Hemorrhage of the eye	-	-
EIF3L	Rods	Abnormality of the astrocytes	-	-
TMEM209	Rods	Abnormality of the astrocytes	-	-
IFT81	Rods	Retinal dysplasia	retinal rod cell development	-

4.3 Discussion

Based on the encouraging results that have been discussed in the preceding chapter, it became necessary to verify them with respect to their real world application and their biological relevance. In this chapter it was therefore investigated whether the proposed workflow allowed for the inference of cell type-specific gene correlation networks from real scRNA-seq data.

`DCA` imputation enables network inference. As the first criterion, the ability to infer scale-free topology gene correlation networks was investigated on the complete and celltype-specific as well as sparse and `DCA`-imputed retina data. After stating that indeed `DCA` allowed for scale-freeness in the cell type-specific networks, gene modules were identified. The resulting cluster dendrograms revealed a deep hierarchy of genes, which was not found in any sparse data set.

From previous studies it was stated that `DCA` introduces spurious correlation signals into the data [Breda *et al.* 2021, Andrews & Hemberg 2019]. Here, it can be noted that no grey module was detected after `DCA`-imputation. In `WGCNA`, the grey module usually summarizes all lowly and uncorrelated genes within the data set. Concerning the sparse retina organoid data, also no grey module was detected, but also no scale-free topology was determinable. As already stated in the previous chapter, the general increase fo gene correlation values by `DCA` would explain why no lowly- or uncorrelated genes were identified. However, in contrast to conventional bulk RNA-seq analysis in `WGCNA`, here a highly filtered data set was used. A large grey module could moreover hint towards an insufficient data cleaning prior to network inference. In general, the presence or absence of a grey module does not *per se* define the quality of the inferred gene correlation networks. But taking also the increased overall correlation structure into consideration, it becomes mandatory to analyse the resulting gene correlation networks with care and caution.

Unique celltype-specific signals can be found. After stating that `DCA` indeed, allowed inferring a scale-free topology and deeply structured celltype-specific network, the resulting gene modules were furthermore analysed.

By employing the set of marker genes for cones and rods, provided by Kim *et al.* [2019], it was investigated if unique network configurations in certain modules could be identified. The importance of these marker genes was therefore determined via a ranking approach. Thus, one module per network was identified in which the celltype specific marker genes revealed overall high low. In order to verify the uniqueness of these celltype-specific modules, the importance of these module hub genes was checked in the opposite network based on their module membership, which corresponds to the intramodular connectivity.

The overall importance was very low for the rod-specific genes that were assigned to the greenyellow module in the cone network. This outcome indicates that the rod-signatures were very specific for the rod-network, which cannot be detected in the cone-network. Regarding the importance of the cone-purple hub genes in the rod network, a different result was seen. Though no module was identified, which owned similarly high MM-values than in the cone-purple module, the rod-brown module revealed higher MM-values. This altogether indicates, that the cone-specific hub genes own a relatively high importance in the rod-specific network. When investigating the cell type embedding in Supplemental Figure S-5b, it can be seen that indeed a few cone cells were detected in the rod cluster.

While celltype-specific networks mainly rely on the cell (cluster) annotation result, the *purity* of the celltypes dictate the quality of the correlation network. The celltype-specific networks could be strongly biased by impure clusters, potentially masking true or imposing false signals. Here, marker genes provided by Kim *et al.* were used to examine the expression pattern of the previously clustered, `DCA`-imputed data. These marker genes were, however, identified in a data-driven approach, based on the cell clustering of the sparse data as well as external knowledge. Therefore, these expression markers might produce diverging cluster annotation results in the `DCA`-imputed data compared to the unimputed retina organoid.

Another major challenge may arise from the fact that two different celltypes were compared which were closely related to each other. Thus, the question arises, if these marker genes are selective enough to sufficiently separate these two related celltypes. Rods and cones both belong to the class of photoreceptors, which are responsible for visual photo-transduction. Perhaps, contrasting photoreceptors and MG cells would derive more distinct results.

Apart from the marker genes, also the annotation procedure itself influences the cluster purity. Depending on the annotation pipeline as well as the prior parameter choices during data preprocessing and cell cluster detection, the annotation result can change. Annotation procedures via marker genes often rely on an initial cell clustering, so that ultimately whole clusters will be assigned to a respective celltype to increase the robustness of the results. Also in this dissertation, whole cell clusters were annotated to either rods, cones , MG cells or non-retinal groups, while only unambiguous clusters were used for network inference. Still the choice of the low dimensional embedding, the cluster detection algorithm as well as the resolution parameters influence the clustering and ultimately the annotation. For the retina organoid data set, neither an annotation script nor the annotation result (in form of a vector) was provided. Therefore, the pipeline was recreated using the sparse information from the publication. To increase reproducibility, the actual Kim-annotation results could have been directly compared to the `DCA`-imputed and clustered data. A more in-depth discussion regarding the reproducibility of annotation results will be presented in the end of this thesis.

Still, the acquired results indicate that cone as well as rod specific modules were detectable in the `DCA`-imputed data using the set of marker genes provided by Kim *et al.*, which were unique for the respective network.

Inferred networks revealed biological meaningful results. Using marker genes to identify rod- and cone-specific modules in the rod- and cone-specific networks already highlighted differences within the celltype-specific networks. Based on this analysis, conserved hub genes were systematically analysed for the respective gene correlation networks across celltypes. EIF4EBP1, a translation initiation factor binding protein was identified as a conserved hub genes, regulating different networks in cones and rods. Predicted human phenotypes as well as GO-terms were extracted for the top 30 surrounding nodes. Though the results were filtered for retina-associated terms, often pathology-related results were gathered. When furthermore considering the connectivity between nodes, homogenous results were identified.

Given this data-driven approach on a `DCA`-imputed single cell transcriptomics data set, celltype-specific gene correlation networks were inferable. The resulting biological analyses of one network highlighted that indeed, retina associated predictions were detected. While only three genes, including the target, were encountered in both celltype-specific networks, the incorporation was different. These results suggest that EIF4EBP1 plays different roles in both photoreceptor celltype-specific networks.

While the target gene itself was less connected in the cone-specific EIF4EBP1 subnetwork compared to the rod network, it revealed some stronger correlations towards ILVBL and miR210HG, which were associated to hemorrhage-like terms. Additionally, FKBP1A was also highly correlated to mir210HG in the cone-subnetwork. This gene revealed some associations to the TGF-β signaling pathway, which itself proofed to take a significant role in so-called *wet AMD* [Wang *et al.* 2019]. Furthermore, was miR210 described to regulate CFB, a Complement Factor B, which is associated to AMD via a promotion of drusen accumulation in RPE cells [Ghanbari *et al.* 2017, Chen *et al.* 2008]. In the rod-specific subnetwork, more general rod (IFT81) and cone (PARP16) cell developmental terms showed high correlations towards the target EIF4EBP1. As already seen for the cone-specific subnetwork, three other hemorrhage-associated genes (TMEM139, ELK3, and OSGEP) were detected. However, all of these genes were mainly connected via PARP16 and not EIF4EBP1. Summing this part up, though different gene correlation subnetworks were associated to EIF4EBP1 in rods and cones, similar biological terms and functions were discovered. In both cell type-specific subnetworks many hemorrhage-like terms were highly correlated to the target EIF4EBP1 directly or via a direct neighbor. As for example neo-vascularization represents a central characteristic of wet AMD, these modules represent compelling candidates, which should be analysed under (wet) AMD conditions.

Since the field of network inference is still up-and-coming in the domain of single cell transcriptomics, it will demand further efforts. Though this analysis indicated a possible *window of opportunity* for using data imputation prior to network inference, further evaluation is still necessary. However, given the results presented in this chapter, `DCA` allowed inferring scale-free, hierarchically structured, celltype-specific networks which revealed distinct biological characteristics.

Chapter 5

Characterization of human retinal organoids

Age-related macular degeneration (AMD) is the major cause of loss of vision in industrialised countries. However, the mechanisms of pathogenesis remains widely unknown to the current day which still hinders an optimal therapeutic approach. With the combination of single-cell transcriptomics and its higher resolution, it might be possible to shed some light on the molecular characteristics of this disease. Though animal models often facilitated investigation and deciphering disease progressions in the human system [Kim *et al.* 2020], they are proofed to be unsuitable in AMD studies since the architecture of the eye severely differ across 'domains'. An overview of the retina as well as AMD was already provided in Section 1.4 of the Introduction.

Here, a novel, neo-natal, human retina organoid (HRO) system should be characterized to infer its suitability. Therefore, two, untreated HRO organoids were subjected to a scRNA-seq analysis. From now on, these two organoid samples will be denominated as HRO-2 for sample two, and HRO-3 for sample three.

As of the current knowledge, AMD is understood to start in the foveal-parafoveal region of the retina [Curcio 2001]. Therefore, the question arises whether the cellular composition of these organoids is sufficiently close to the foveal region. Using two different approaches to annotate the single-cell transcriptomics data of two untreated organoids, the overall cellular composition is described. In a latter step, these organoids are further characterized by comparing the data to other (human tissue) reference data set. All previous parts aimed in characterizing the HRO system, trying to state the suitability of the HRO system based on cell type distribution and expression pattern correlation.

Apart from the cellular composition, another important question concerns the developmental maturity of the organoid systems. Therefore, it was investigated to what extend known developmental processes can be reproduced by the organoid model. Using RNA velocity, developmental processes within the single-cell data can be visualised, potentially providing a new analysis depth.

5.1 Workflow

Human retina organoids were established to study the defined molecular and cellular changes upon AMD onset (Völkner *et al.*, in revision). The general workflow of this chapter is depicted in Figure 5.1. To characterize the HRO system, two different annotation approaches were used. Further steps analysed if the HRO system owns expression patterns that are closer correlated to the inner retinal region (fovea) or periphery. Using the ratio between unspliced and spliced transcripts, a developmental trajectory was inferred to order the cells based on a pseudotime variable.

5.1.1 HRO characterization

HRO single-cell transcriptomics data preprocessing, visualization, and most of the downstream analysis were conducted in `scanpy` (version 1.3.1). Two untreated HRO samples were used for the characterization and annotation procedure.

Data preprocessing. All steps were adapted from the `scanpy`-tutorial webpage. Cells with less than 200 expressed genes and genes which were not expressed in at least three cells were discarded immediately. Based on the distribution of gene counts and the percentage of mitochondrial genes, cells with less than 2500 genes and four per cent mitochondrial genes were kept. The count data was normalised according to the tutorial, highly variable genes were detected using the standard parameters and kept for downstream analysis. Finally, the data was log-transformed, variation factors were regressed out (number of genes and percentage of mitochondrial genes) and scaled in accordance to the `scanpy` tutorial.

Cluster detection. Dimensionality reduction was done using principal component analysis (PCA). Therefore, a neighbourhood graph was calculated using ten neighbours and 40 principal components. Following a Uniform Manifold Approximation and Projection (UMAP) embedding, cell clusters were detected based on the Louvain clustering implementation of `scanpy` using a resolution parameter of two, to detect smaller clusters.

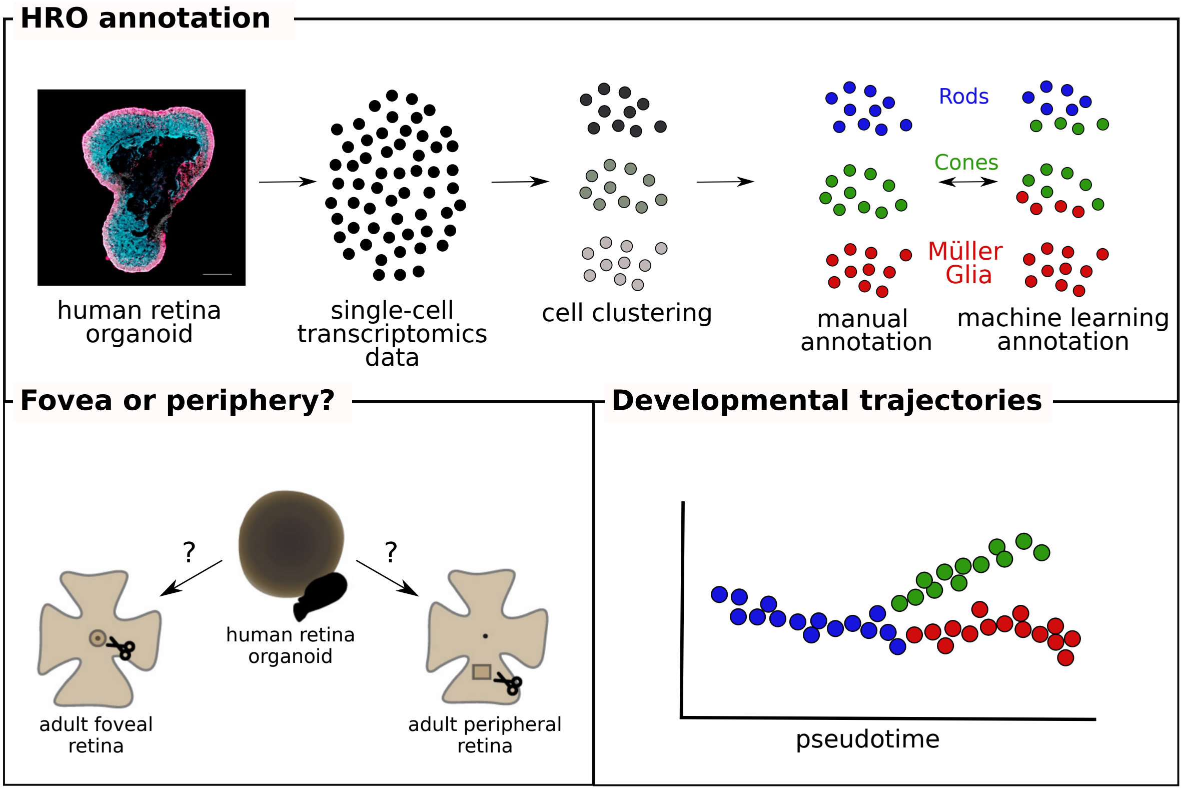


Figure 5.1. *Analysing and characterising the HRO systems.*

In a first step, the single-cell HRO transcription data was analysed and annotated using two approaches: (1) a manual annotation using unique expression pattern of marker genes and (2) a machine learning approach, which uses transfer learning. After identifying retinal cell populations, the correlation towards foveal or peripheral tissue single-cell data was analysed. Finally, developmental trajectories within the HRO system were calculated and investigated via RNA velocity.

Cluster annotation

In order to annotate the HRO cells, two different approaches were used. While one approach requires external gene of interest (GOI) lists, that are based on published knowledge, and compares the expression pattern of predefined clusters, the other approach uses a transfer learning algorithm, which depends on an annotated reference data set to learn from.

Manual annotation. GOI lists were assembled of known marker genes aiming to discriminate the following cell types: cones, rods, Müller Glia (MG) cells, bipolar cells and Amacrine-Horizontal-Ganglion (AHG) cells. These genes were derived using expert knowledge via personal communication with Mike Karl. The overlap of genes between GOI lists and single-cell data was taken as input for a dot plot representation in `scanpy` where the expression of the selected marker genes across all Louvain clusters was analysed. All dot plots were generated using `scanpy` version 1.4.1. Different `scanpy` versions were used, since the initial manual cluster annotation was performed using the older version. To avoid differences in the cell cluster embedding and therefore a re-evaluation of the GOI expression patterns, preprocessing was performed using version 1.3.1. The local expression pattern of these markers was additionally plotted onto the UMAP embedding. Generally, if the majority of the cell type-specific marker genes were uniquely expressed in one Louvain cluster, this cluster was assigned the respective cell type. Finally, all the Louvain clusters were manually annotated to one of the five previously mentioned cell types adding one premature photoreceptor cluster.

Transfer Learning using `CaSTLe`. In opposite to manual annotation, a machine learning tool called `CaSTLe` was applied which learns expression profiles from a reference data set and transfers that knowledge to an un-annotated data set [Lieberman *et al.* 2018].

Here, three different reference data sets were used: a fully developed human retina organoid, an adult periphery data set, and an adult fovea data set (all taken by Cowan *et al.* [2020]). While 37 different retinal sub-cell types across 44 000 cells were detected in the Cowan developmental organoid data set, 41 and 53 cell types were annotated in the adult foveal and peripheral data, respectively. While the adult foveal data set includes 20 000 cells, the peripheral counterpart contains 35 000 cells collected from three donors. An overview of the condensed cell counts is represented in Table 5.1. All three Cowan *et al.* [2020] data sets were retrieved from the iob-webpage*.

Though a high resolution of retinal sub-cell types was provided by all Cowan *et al.* [2020] data sets, the sub-cell types were compressed into major groups to increase the training performance. Initially, the Cowan data sets contained a rather fine-grained cell type classification. For example, MC (Müller Glia) cells were split into three distinct subtypes with only a few dozen annotated cells in one of those subtypes.

The annotation procedure was run in accordance with the tutorial of the `CaSTLe` github page†. `CaSTLe` requires both, the reference as well as the unannotated data set to be in a `SingleCellExperiment` format. From here on those data sets will be referred to as *source* and *target*. After identifying the set of common genes between source and target, both data sets were subsetted on those. Then, 100 genes with the highest mean expression in the source

---

*`https://data.iob.ch/`

†`https://github.com/yuvallb/CaSTLe/blob/master/CaSTLeMultiClass.R`

Table 5.1. *Overview of cell types and amounts of the Cowan* et al. *[2020] data sets. In the original data, more numerous cell types were found. To increase classification performance, the annotation data by Cowan was condensed. An overview of the whole data is depicted in Supplementary Table S-2 and Table S-3. Legend: AC - Amacrines, Ast - Astrocytes, CdBC/ChBC - Bipolars, CM - Choroidal melanocyte, END - Endothelial cells, FB - Fibroblasts, GC - Ganglions, HC - Horizontals, MC - Müller Glia, PER - Pericytes, RBC - Rod bipolar cell, RPE - Retinal pigment epithelium, uG - Mircoglia*

Cell type	Cellcount		
	Organoid	Foveal	Peripheral
AC_B	1230	192	417
AC_Y	81	123	330
Ast	-	149	172
CdBC	1700	2058	2418
ChBC	658	398	1182
CM	-	-	157
cones	12973	1375	1202
END	-	368	208
FB_02	-	252	1355
GC	-	6086	35
HC	1762	1037	844
MC	10542	3886	8207
PER	-	75	85
RBC	461	1016	4191
rod	13913	1894	13029
RPE	130	84	186
uG	-	172	171

and target were identified. Additionally, the top 100 genes with the highest mutual information from the source data were extracted. The union of these gene sets were furthermore preprocessed by removing high inter-feature correlations. Finally, the logged expression entries were converted into four ordinal bins $[0], (0, 1], (1, 6], (6, \infty)$, as described in the tutorial and original publication [Lieberman *et al.* 2018]. Prior to training the classification model, genes were removed for which all values fell in the same bin. The cell type annotation of the source data was transformed into numerical entries, starting at zero. In `CaSTLe`, an xgboost classifier was trained on 80% of the source data which was randomly sub-sampled. All model parameters were elected following the original publication. The model performance was evaluated on the remaining 20% of the source data, using sensitivity (Equation 5.1) and specificity (Equation 5.2).

$$Sensitivity = \frac{True\ positives}{False\ negatives + True\ positives} \tag{5.1}$$

$$Specificity = \frac{True\ negatives}{False\ positives + True\ negatives} \tag{5.2}$$

Afterwards, the target data was fed to the classification model. In this dissertation, the multiclass implementation was used. Finally, the `CaSTLe`-classified HRO data was re-translated from numerical entries into the respective cell type levels.

Merging both HRO data sets. After annotating both untreated HRO systems via the manual marker gene expression and machine learning workflow, both data sets were merged on top of each other. Here, the low dimensional UMAP embedding of HRO-2 was used as a basis. In a first step, both data sets were filtered on the set of common genes. Then, a PCA, neighbor detection, and UMAP embedding was perfomed on the filtered HRO-2. The `ingest` function from the `scanpy` tool was used for mapping labels from HRO-2 to HRO-3. Subsequently, both data sets were concatenated. To infer the consistency of rods, cones and MG cells, a PCA was performed on those three cell types.

5.1.2 Correlating HRO system to foveal and peripheral data sets

In the previous approach, the HRO cells were annotated manually or via a transfer learning tool. To infer whether HRO photoreceptors and MG cells are closer to a foveal or peripheral expression pattern, the correlation between three reference single-cell data sets and HRO photoreceptors and MG cells were calculated.

Here, three different adult reference data sets were used: two human data sets by Cowan *et al.* [2020] and Voigt *et al.* [2019], as well as a macaque experiment by Peng *et al.* [2019b]. All human data sets contained information from three different donors. Each reference data set as well as the HRO single-cell data was subsetted to rods, cones, and MG cells, respectively, to compare their expression patterns. For the HRO samples, the `CaSTLe` cell annotation was used. For each specific cell type, the gene intersect between the reference and the HRO data was identified. Afterwards, a mean gene expression vector of the reference rods, cones and MG cells was calculated using the log10(x+1) transformed count data. In cases where more than one donor was present, one mean vector per donor was calculated. The HRO single-cell data was also log10(x+1) transformed to match the reference data.

Finally, the Pearson correlation was calculated for each selected HRO cell against all previously derived mean expression vectors of the reference data sets.

5.1.3 Approach developmental dynamics via RNA velocity

During the experimental workflow, cells are disrupted to extract the RNA. Therefore, any developmental information is theoretically lost. However, the ratio between unspliced and spliced transcripts can be used to infer a pseudotemporal variable, called RNA velocity. Inferring developmental dynamics in form of the RNA velocity required a few preprocessing steps, which are stated on the github page*. First, `samtools`(version 1.6.) was used to sort the input BAM files by position. Second, `velocyto`(version 0.17.17) was used to calculate the RNA velocity [La Manno *et al.* 2018] on both HRO samples. Therefore, the sorted BAM-files, the cell-barcodes, and the genome annotation file was required. Finally, the tool `scVelo` (version 0.2.3)[Bergen *et al.* 2020] was used to translate the RNA velocity into cellular dynamics. `scVelo` is embedded in the `scanpy` framework.

The RNA velocity analysis pipeline was taken from the example analysis of the endocrine pancreas, provided on the github page†. Briefly, the filtered, normalised, scaled and annotated expression data and the `scVelo` output file were merged. For this analysis, the `CaSTLe`-annotation was used. Before the velocity was calculated, the first- and second-order moments ‡ were computed using the PCA space. Afterwards, the velocity was calculated by fitting the ratios between unspliced and spliced mRNA abundances. Additionally, `scVelo` allows calculating the pseudotime, as well as root and endpoints of the input data. Local dynamics identified by this analysis can then be plotted on top of the UMAP embedding.

Though this high resolution is very useful for some analysis aspects, it does not allow to compare general developmental trajectories. Therefore, a partition-based graph abstracting

---

[*] `https://scvelo.readthedocs.io/getting_started/`

[†] `https://scvelo.readthedocs.io/Pancreas/`

[‡] According to the tutorial, first and second order moments correspond to the mean and uncentered variance computed among nearest neighbors in the PCA space.

(`PAGA`) tool was used which produces a coarse-grained representation of the single-cell cluster data. Similar to `scVelo`, `PAGA` itself is embedded in the `scanpy`-environment (version 1.7.1). Starting from the preprocessed and `CaSTLe`-annotated data, the `PAGA`-graph was calculated using the grouping from the `CaSTLe`-annotation under model *v1.0*, and the calculated RNA-velocity. For visualization of the graph, the *Fruchterman-Reingold* (fr) layout, as well as a threshold of 0.15, was used. Based on this cell type graph, the UMAP embedding was recalculated to allow for better interpretation. Finally, the RNA velocity stream embeddings were projected onto this new embedding. Using other internal functions, the pseudotime, and the root and endpoints were highlighted in a heatmap.

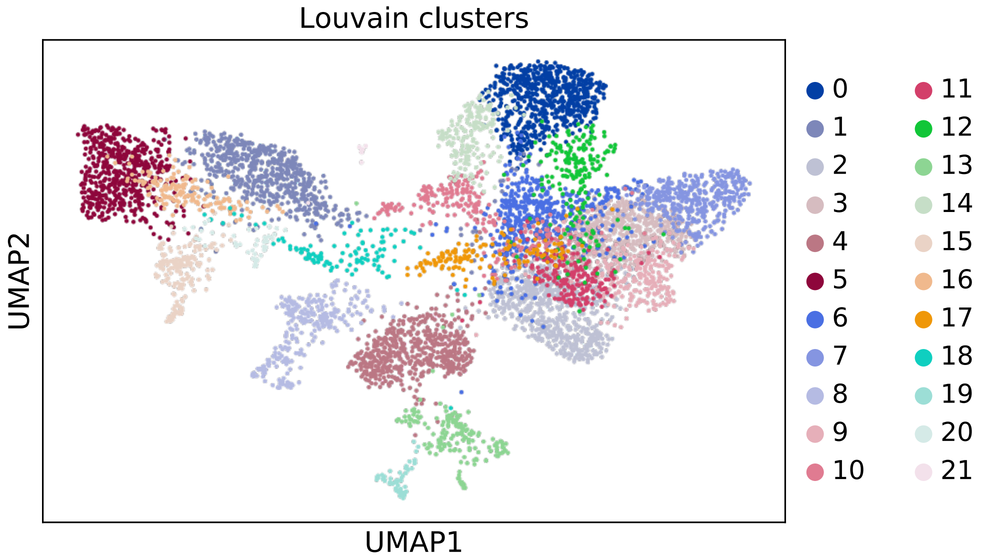
5.2 Annotation of neonatal human retina organoids

As previously described, different approaches should be utilised to investigate the composition and characteristics of a human retina organoid. In a first attempt to characterize the organoids, individual cells in each of the two organoid samples must be annotated to a specific retinal cell type, such as cones, rods, Müller-Glia, or bipolar cells. Here, two different approaches were used with a manual annotation relying on genes-of-interest (GOI) lists and a machine learning approach using transfer learning. The detailed workflows are described in section 5.1.

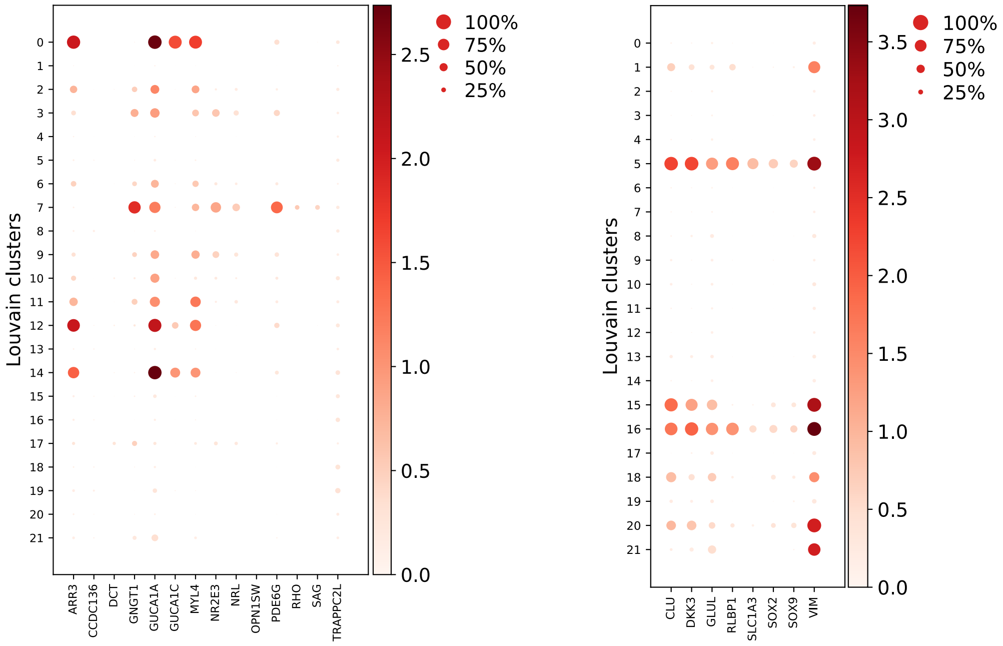
Both HRO single-cell transcriptomics data sets were initially subjected to the `scanpy` preprocessing pipeline, including the removal of lowly expressed and non-variable genes, as well as a data transformation to scaled, normalised log-values. In total, 4031 genes across 6665 cells remained after filtering in HRO-2, while 3771 genes across 5370 cells were kept for HRO-3. Subsequently, the dimensionality reduction and UMAP embedding were calculated for both samples and similar cell clusters (based on expression values) were detected via the Louvain algorithm. As indicated in Figures 5.2a and 5.3a, 22 and 20 clusters were detected in HRO-2 and HRO-3, respectively. Using Louvain clustering, cluster '0' always represents the largest cluster with 665 (HRO-2) and 556 (HRO-3) cells. The smallest clusters were cluster '21' with 10 cells (HRO-2) and cluster '19' with 13 cells (HRO-3). After identifying groups of cells owning similar gene expression patterns, these clusters can be annotated. But assigning cell clusters as certain cell types allows inferring for example the cell type distribution of the data set or developmental processes.

Manual annotation. One possibility to annotate specific cell types is to use their expression 'behaviour' across genes-of-interest (GOIs) sets. This collection of marker genes are based on literature and expert knowledge. In Figures 5.2 and 5.3, the expression pattern of selected marker genes that are specific for photoreceptors (b) and MG cells (c) is shown for HRO-2 and HRO-3, respectively. Through a comparison of those specific expression patterns across all Louvain clusters in each HRO sample, it becomes feasible to distinguish different cell types manually.

In the given examples, Louvain clusters 0, 2, 3, 6, 7, 9, 11, 12, and 14 showed expression for photoreceptor-specific genes in HRO-2 (Figure 5.2b), clusters 1, 5, 15, 16, 18, 20, and 21 revealed MG cell-specific expression patterns (Figure 5.2c). In HRO-3, photoreceptor genes were uniquely expressed in clusters 1, 2, 5, 7, 9, 13, 15, and 16 (Figure 5.3b), while MG-cell specific genes were specific for clusters 0, 3, 17, and 18, see Figure 5.3c.



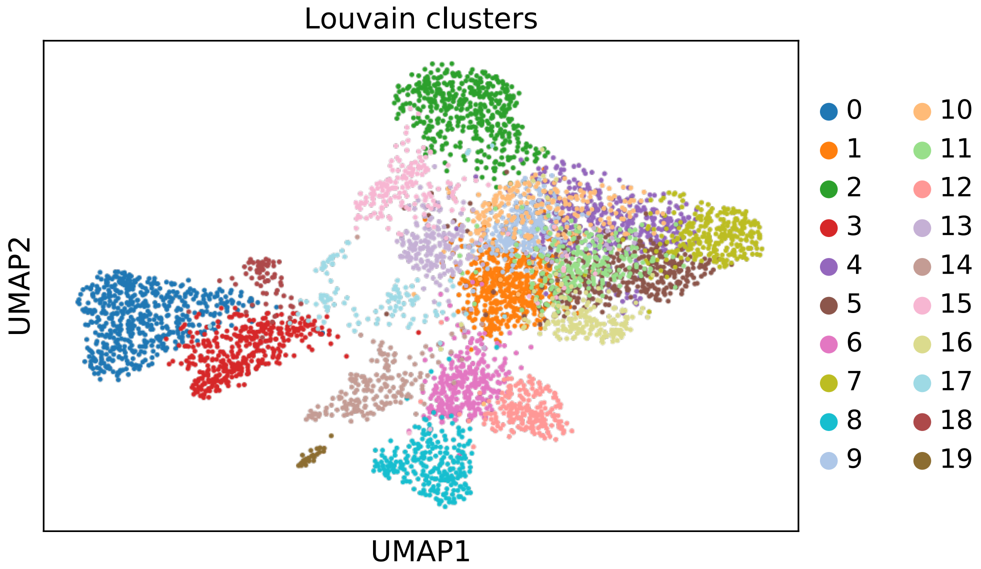
(a) *UMAP of HRO-2 with Louvain clusters*



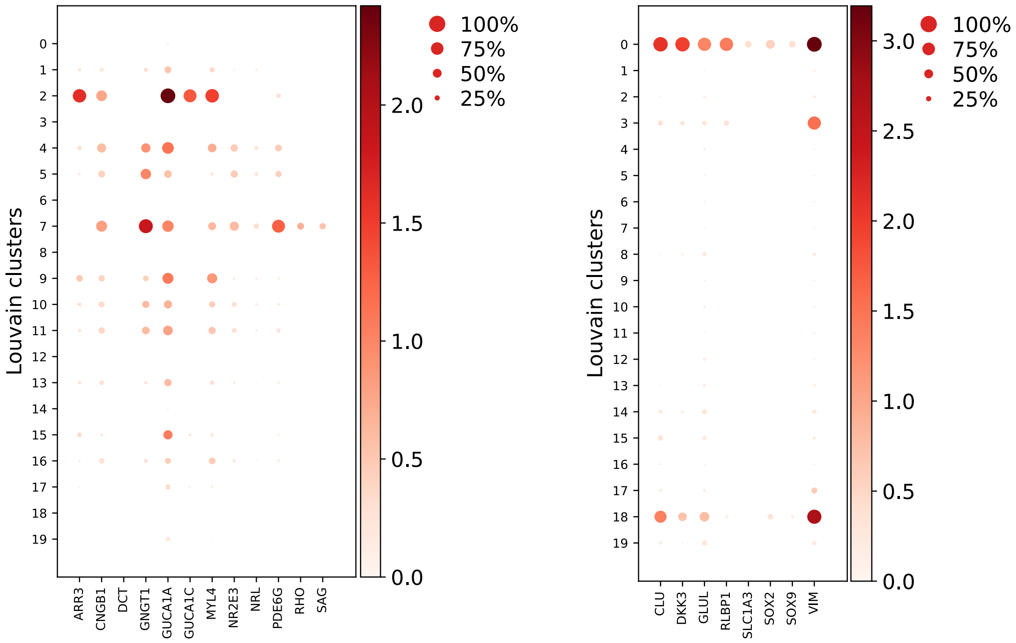
(b) *Dot plot of photoreceptor-associated genes* **(c)** *Dot plot of MG-associated genes*

Figure 5.2. *UMAP embedding of HRO sample 2 and expression pattern of selected marker genes.*

(a) UMAP of preprocessed HRO sample 2. In total, 21 different cell clusters were detected via the Louvain community detection algorithm. Expression of selected photoreceptor (b) and MG cell-associated genes(c) across 21 different Louvain clusters. The mean expression is indicated in the heatmap legend. The dot size corresponds to the percentage of cells expressing the gene inside the Louvain cluster.



(a) *UMAP of HRO-3 with Louvain clusters*



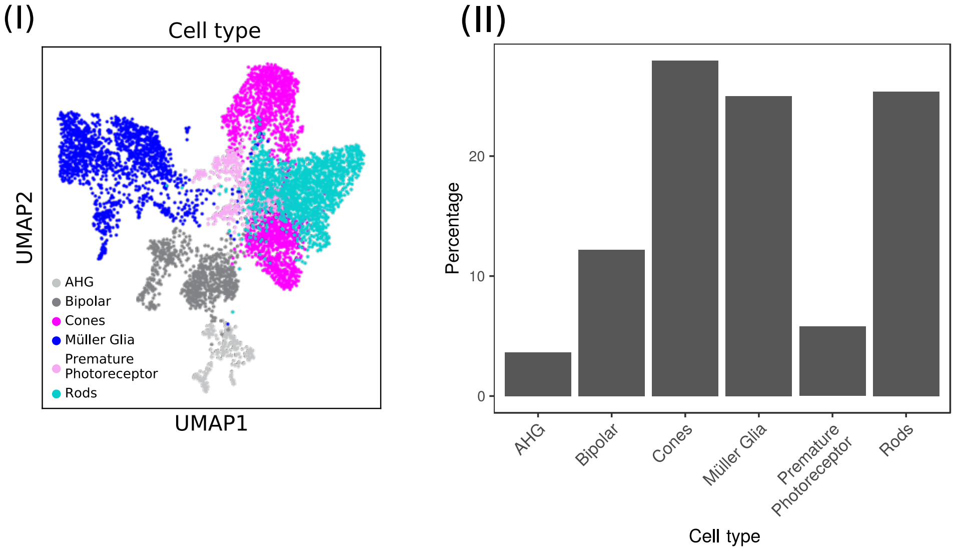
(b) *Dot plot of photoreceptor-associated genes* **(c)** *Dot plot of MG-associated genes*

Figure 5.3. *UMAP embedding of HRO sample 3 and expression pattern of selected marker genes.*
(a) UMAP of preprocessed HRO sample 3. In total, 20 different cell clusters were detected via the Louvain community detection algorithm. Expression of selected photoreceptor (b) and MG cell-associated genes (c) across 20 different Louvain clusters. The mean expression is indicated in the heatmap legend. The dot size corresponds to the percentage of cells expressing the gene inside the Louvain cluster.

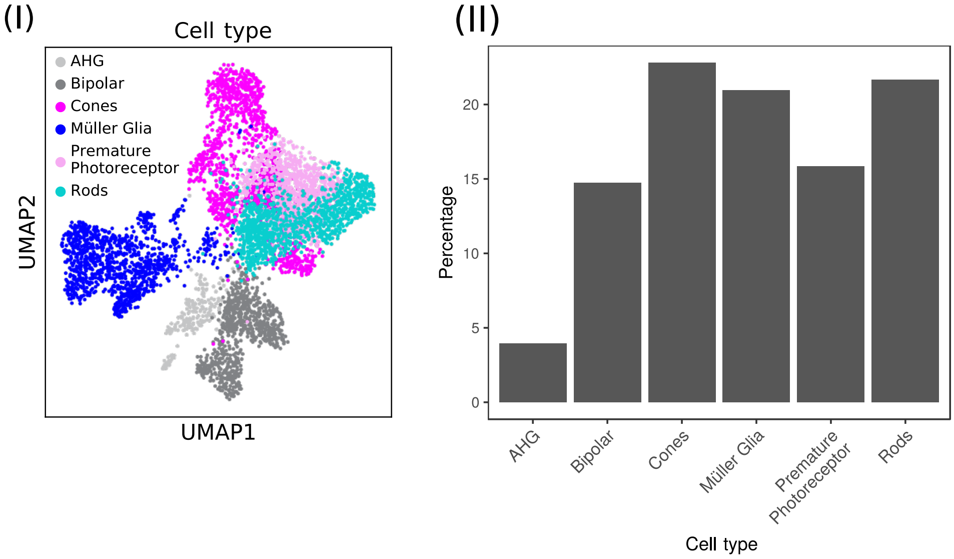
Apart from photoreceptors and MG cells, dot plot expression profiles were also calculated for bipolar cells, amacrine cells, horizontal cells, and photoreceptor progenitor cells. The latter cell type summarized a group of not yet fully developed photoreceptor cells. The corresponding illustrations can be found in the supplement for HRO-2 (Figure S-7) and HRO-3 (Figure S-8). For a better visibility, only a selected gene set per cell type was used, since the GOI lists contained more than 100 entries for certain cell types, which cannot be visualised. Also, cones and rods were compressed to the group of photoreceptors in the dot plots. As Amacrines, Horizontal and Ganglion cells were compressed into one annotation group, they were shown in one dot plot. So finally five different dot plots were included in this dissertation giving rise to six cell type annotations that were considered for annotation.

Using the information of the GOI-dot plot and expert knowledge, the Louvain clusters of both HRO samples were assigned to six different retinal cell types as depicted in Figure 5.4a(I) and 5.4b(I). Whereby photoreceptors (rods and cones) were located on the right-hand side in the Uniform Manifold Approximation and Projection (UMAP) embeddings, MG cells were found on the opposite. Amacrines, horizontal, and ganglion cells were assigned to the AHG group and were found on the bottom of the UMAP. The group of photoreceptors progenitor cells were detected between both photoreceptor subtypes of rods and cones. Those finding were comparable across both HRO samples.

Apart from the general arrangement inside the low dimensional embedding, also the relative cell type abundances were comparable across HRO-2 and HRO-3. An overview of these results is depicted in Figure 5.4a(II) and 5.4b(II). While cones represented the largest subgroup within the organoid system with 28% and 23%, respectively for HRO-2 and HRO-3, AHG were the smallest with around 3% in both data sets. Other larger cell groups were rods (25%, 22%) and MG cells (25%, 21%). The relative amount of premature photoreceptors largely varied between both organoid samples with 6% in HRO-2 and 16% in HRO-3. It should be noted that 8% more mature photoreceptors were annotated in HRO-2 than HRO-3, which might account to some extent for the nearly 10% fewer premature photoreceptors. Bipolar cells accounted for 12% (HRO-2) and 15% (HRO-3).



(a) *HRO-2 cell-type composition*



(b) *HRO-3 cell-type composition*

Figure 5.4. *HRO cell-type compositions.*

UMAP embeddings(I) and distribution of relative cell type quantities(II) in HRO-2 (a) and HRO-3 (b). (a) Using sets of GOI lists and comparing their expression pattern across 22 Louvain clusters, all 6665 cells were annotated to six retinal cell. Whereby cones represented the biggest subgroup in HRO-2, similar quantities were detected for rods and MG cells. AHG were the smallest organoid subgroup. (b) For HRO-3, six retinal cell types across 20 Louvain clusters (n=5730 cells) were annotated. Similar to HRO-2, comparable quantities of cones, rods, and MG cells were detected. In HRO-3, more premature photoreceptors than bipolar cells were found. Again the smallest cluster was represented by the AHG group.

Transfer learning annotation. While the previous cell annotation relied on expert knowledge as well as subjective decisions on the interpretation of the GOI expression, other cell annotation tools try to learn annotation patterns from alternative reference data sets. Classification of single cells by transfer learning (`CaSTLe`) is such a tool [Lieberman *et al.* 2018]. Briefly, it defines a set of informative genes in a reference data set which allows for the classification of a new data set using a reference data set. The whole approach is based on a random forest architecture.

As annotation reference, three different retinal data sets by Cowan *et al.* [2020] were used: one fully developed organoid and two adult, human tissue samples of the fovea and periphery. An overview of the data sets is provided in the workflow description in section 5.1.1 and the Supplementary Tables S-2 and S-3.

In order to assess how well the model can learn the features and predict the cell type, each reference data set was randomly split into a training and test set. While the classifier was trained on 80% of the data, the model was evaluated on the remaining 20%. In total, 3790, 3822, and 8733 cells were utilised in the foveal, peripheral, and organoid test sets. An overview of the sensitivity and specificity is given in Figure 5.5. Using the sensitivity measure, the number of false negatives can be quantified, while the specificity aims for the false positives. Across all reference data sets, the specificity remained close to one, corresponding to a low number of false positive classifications. The sensitivity values did vary strongly across the reference data sets and cell types. Generally, cell types with larger quantities, reached higher sensitivity values, such as rods, cones, and MG cells for the developed organoid data (Figure 5.5(A)). The overall lowest sensitivity values were reached for glycinergic amacrine cells (AC-Y) in the organoid data. Both adult Cowan data sets reached on average higher sensitivity values compared to the organoid data. Summing up these results, the classifier was able to reach overall high specificity values, and, except for two cell types, sensitivity values above 0.5.

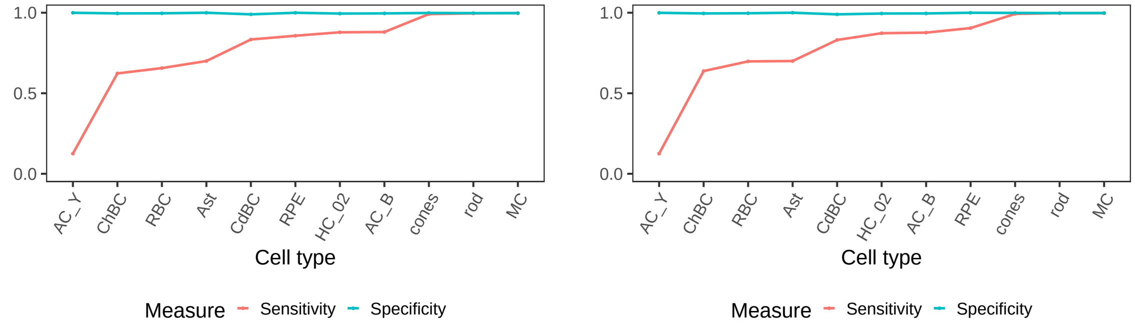
Using these pre-trained classifiers, both HRO samples were classified using the `CaSTLe` framework. The `CaSTLe`-annotation was then projected onto the low-dimensional UMAP representation derived by the `scanpy` pipeline described earlier. An overview of these results is shown in Figure 5.6. Using the classifier trained with the organoid data, depicted in Figures 5.5a and 5.5b, regionally defined cell clusters were found. Similar to the manual annotation results (Figure 5.4a and Figure 5.4b), photoreceptor cells were detected on the right-hand side of the UMAP plot, while MG cells were found on the opposite. Based on the fact, that more distinct cell types were present in the reference annotation, a higher resolution was gained for the cells located on the lower part of the UMAP compared to the manual annotation (HC, AC, CdBC and ChBC). However, since no photoreceptor progenitor cells were annotated in the Cowan reference, this cell type could not be detected in the `CaSTLe`-derived annotation.

A different picture was observable for both classifiers trained on the adult data sets. Figures 5.5(c-f) reveal less distinct cluster annotation. Generally, the coarse-grained arrangement of photoreceptors and MG cells was found for the foveal data (see Figure 5.6c and 5.6d). While cones and rods were located on the right-hand side, MG cells were located on the opposite site. However, far more rods were detected in originally bipolar (BC) and RPE cell locations. A completely novel cell class of micro-glia (uG) were found close to the MG cells.

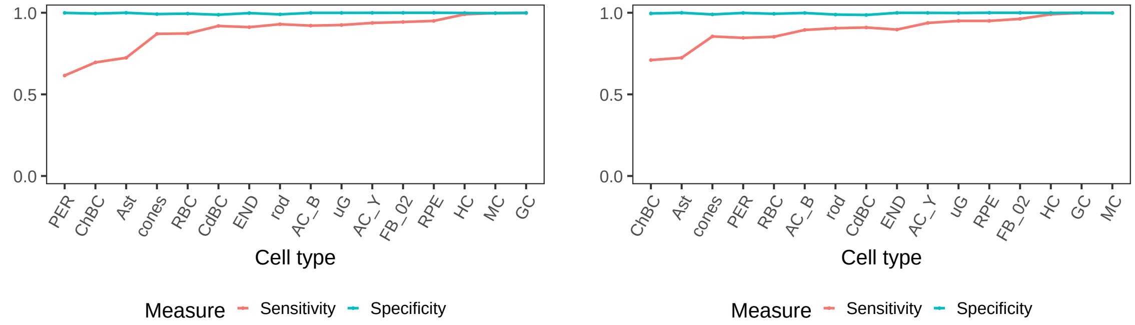
Using the adult periphery data classifier, a different annotation scheme was observed ((see Figure 5.6e and 5.6f)). Generally, the cell types were more scattered across the whole UMAP, and many RPEs were detected. In original MG cell locations, many horizontal cells (HC) were annotated. Overall, nearly no cones were detected.

In summary, using a transfer learning tool called `CaSTLe` and three different retina reference data sets, both HRO samples were annotated. While the annotation via the Cowan organoid data largely corresponds to the manual annotation, both adult data classifiers revealed scattered celltype locations. These results highlight the need of a suitable reference data set to generate meaningful annotations via machine learning tools. Due to the high correspondence between the annotation results and the developmental status of the data, the organoid data trained annotation was used for further downstream analysis and will be referred to as the `CaSTLe`-annotation.

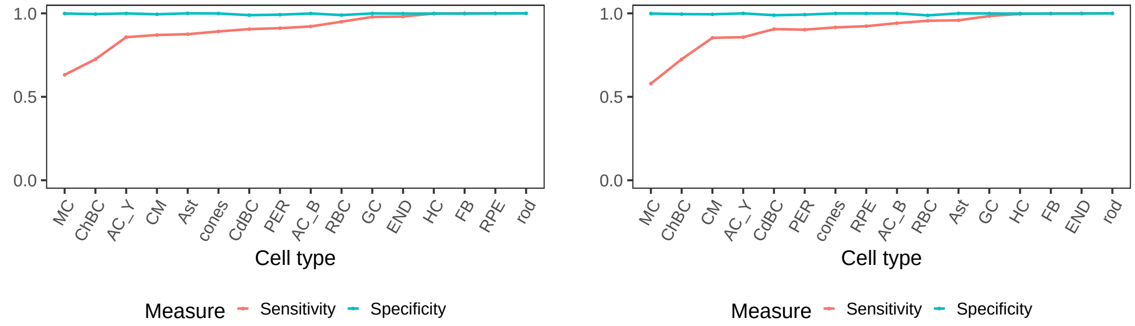
After describing the general spatial arrangement of both `CaSTLe`-annotated HRO samples, Figure 5.7 highlights the relative cell type abundances. Across both HRO samples, cones represented the bigger cell class with more than 30%, followed by MG cells (MC). Rods accounted for around 10% of the cells in both HRO samples. While slightly more CdBC and CdBC were annotated in HRO-3, the remaining cell types (AC-B, AC-Y, HC-02, RBC, and RPE) owned similar percentages across both HRO samples. Solely astrocytes (Ast) were only detected in small quantities in the HRO-2 sample.



(a) *Organoid data in HRO-2* **(b)** *Organoid data in HRO-3*



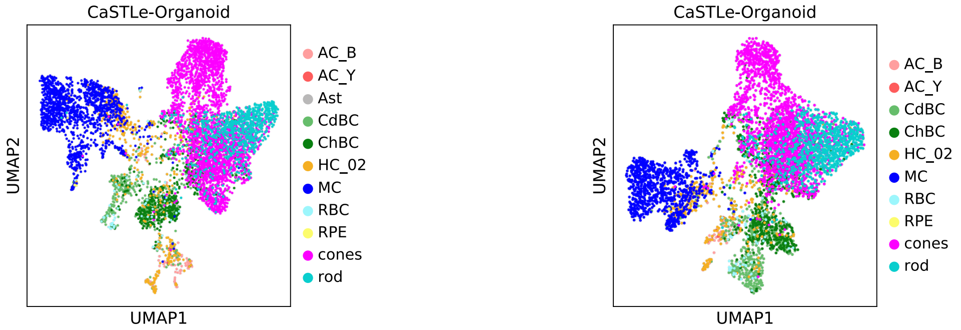
(c) *Foveal data in HRO-2* **(d)** *Foveal data in HRO-3*



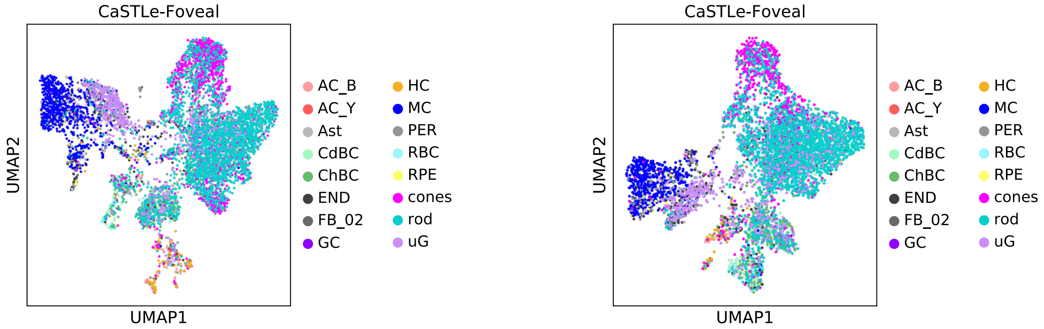
(e) *Peripheral data in HRO-2* **(f)** *Peripheral data in HRO-3*

Figure 5.5. *Evaluation of* `CaSTLe`*-classification model for both HRO control samples across all reference data sets.*

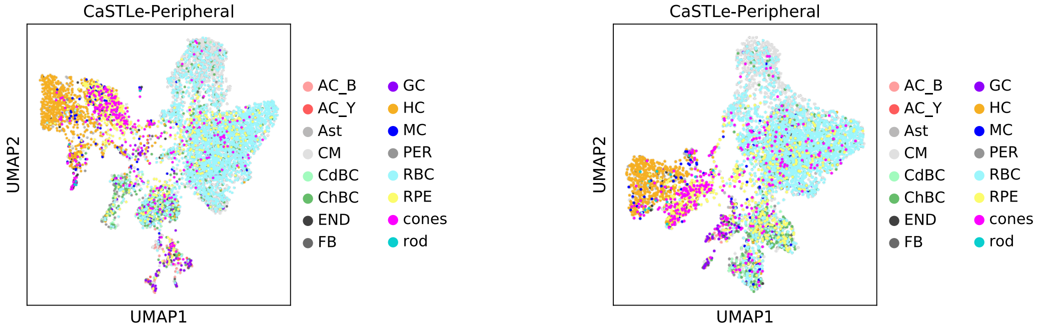
Sensitivity and specificity values using 20% of the subsetted Cowan-reference data sets during the classification of the HRO-2 (I) and HRO-3 (II). The result of the developed human retina organoid is shown in (a+b), the adult foveal data in (c+d) and the adult peripheral data in (e+f). While the specificity was constantly close to one, the sensitivity varied over the different data sets and HRO samples. However, the majority of cells reached a sensitivity value larger than 0.5. The lowest sensitivity was detected for AC-Y in both `CaSTLe`*-classifiers trained with the organoid data.*



(a) *Organoid data in HRO-2* **(b)** *Organoid data in HRO-3*



(c) *Foveal data in HRO-2* **(d)** *Foveal data in HRO-3*



(e) *Peripheral data in HRO-2* **(f)** *Peripheral data in HRO-3*

Figure 5.6. *Evaluation of `CaSTLe`-annotation for both HRO control samples across all reference data sets.*

(a+b) Annotation using the developed organoid data for training resulted in quite defined cell type locations. While photoreceptors were located on the right-hand side, MCs were located on the opposite side. (c+d) Less defined cell clusters were detected using the adult, foveal data for training `CaSTLe`. Still, the coarse-grained distribution of photoreceptors, MCs and remaining cells was similar to (a+b). Here, more rods were detected. (e+f) A scattered annotation was found for the adult periphery training data. In this case, a completely different spatial location was encountered. While mainly RBCs were detected on the right-hand side, HCs were found on the left-hand side.

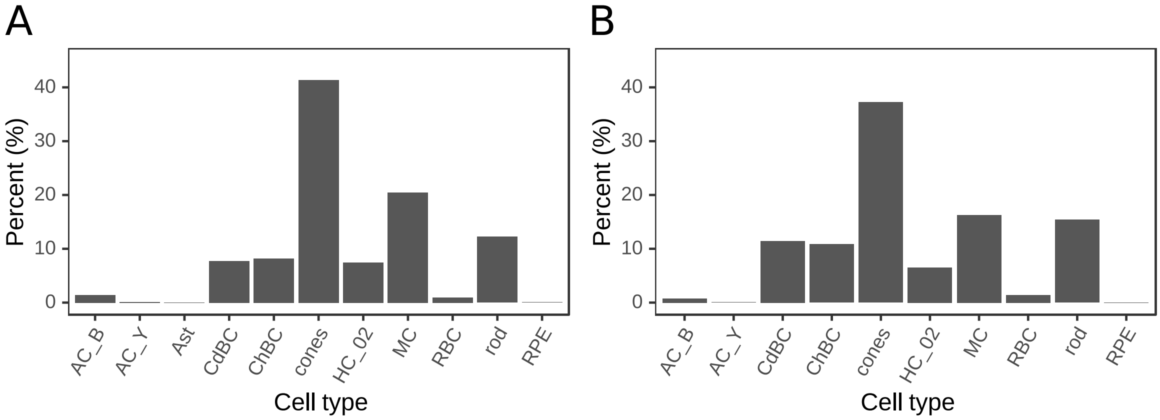


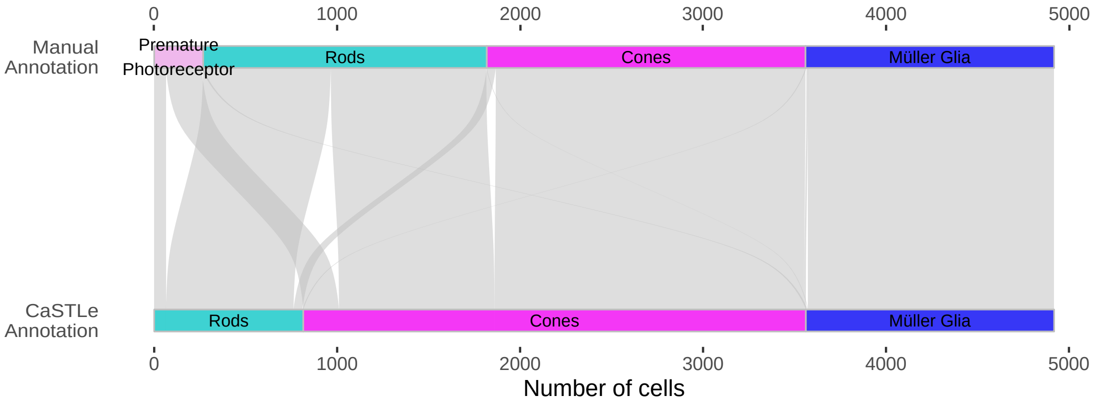
Figure 5.7. *Relative composition of both HRO samples using the Cowan organoid trained* `CaSTLe` *classifier.*

The percentage of `CaSTLe`*-annotated cell types in HRO-2(A) and HRO-3(B) is shown. While cones were the biggest class of cell types across both samples, MC and rods were the second and third largest groups, respectively. Slightly more CdBCs than ChBCs were annotated in HRO-3. AC-B,AC-Y, HC-02, RBC, and RPE showed equal quantities across both samples. Astrocytes were unique for HRO-2, though low in abundance.*

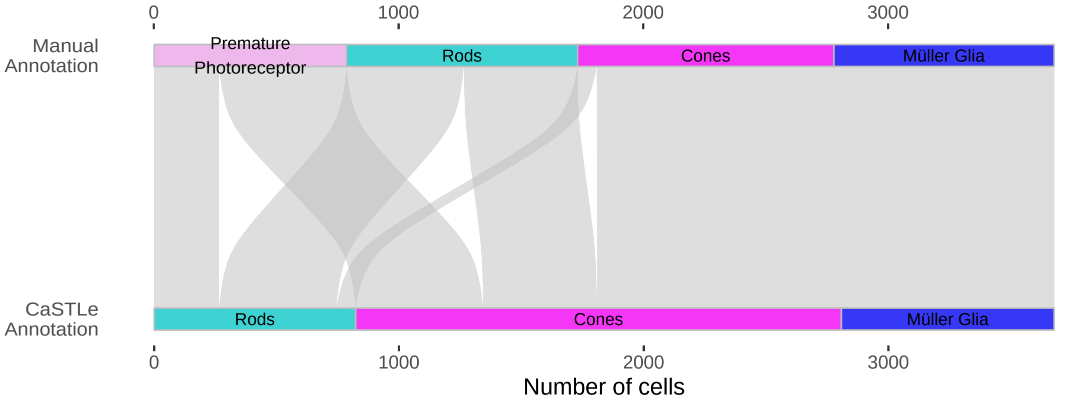
5.2.1 Comparison of manual and machine learning annotations

In the previous parts, two untreated, single-cell HRO systems were annotated using two different approaches. Using a manual annotation and transfer learning approach, various retinal cell types were identified. Here, both annotation results are compared to each other focussing on three major cell types: cones, rods and MG cells, since they represent the largest cell population within the HRO system. Figure 5.8 summarizes the consistency of these retinal cell types. In general, nearly all MG cells were similarly annotated via the `CaSTLe` pipeline and manual approach for both HRO-2(A) and HRO-3(B). Major differences of both annotation techniques can be seen for photoreceptor subtypes of rods and cones. When looking at the cell counts, there is already a rather large discrepancy between both annotation approaches: whereby in the manual annotation comparable quantities of cells were annotated as cones and rods, respectively, the `CaSTLe`-annotation yielded nearly more than three times more cones than rods. Cells being classified as cones in the manual approach were also mainly annotated as cones with `CaSTLe`. Only a small percentage was annotated as rods in the `CaSTLe`-annotation. However, nearly one-third of cells being classified as rods in the manual annotation were annotated as cones with `CaSTLe`, which largely explains the shift in the cone-ratio from 30% to 40%. From the premature photoreceptor population, also more cells were annotated as cones using `CaSTLe`.

An overall similar trend can be seen for HRO-3. Again, all manually annotated MG cells were also annotated as MG cells in the `CaSTLe`-workflow. The same was found for manual cones, which generally remained cones using the `CaSTLe`-annotation. Solely a small population was identified as rods in `CaSTLe`. As already stated for HRO-2, the manual-rods part nearly equally divide into `CaSTLe`-rods and cones. In HRO-3, the premature photoreceptor cluster contained more cells than in HRO-2 but the ratio migrating into `CaSTLe`-rods and cones remained comparable.



(a) *HRO-2*



(b) *HRO-3*

Figure 5.8. *Consistency of three major retina cell types across both annotation approaches. Per* `CaSTLe`*-annotated cone, rod, and MG cell the corresponding manual annotation result was extracted in HRO-2(a) and HRO-3(b). While* `CaSTLe`*-MG cells were also annotated as MG cells using the manual annotation, the* `CaSTLe`*-annotations for both, cones and rods were less consistent. Especially concerning the annotation results from HRO-3, many* `CaSTLe`*-cones were detected as premature photoreceptors and rods in the GOI-based annotation.*

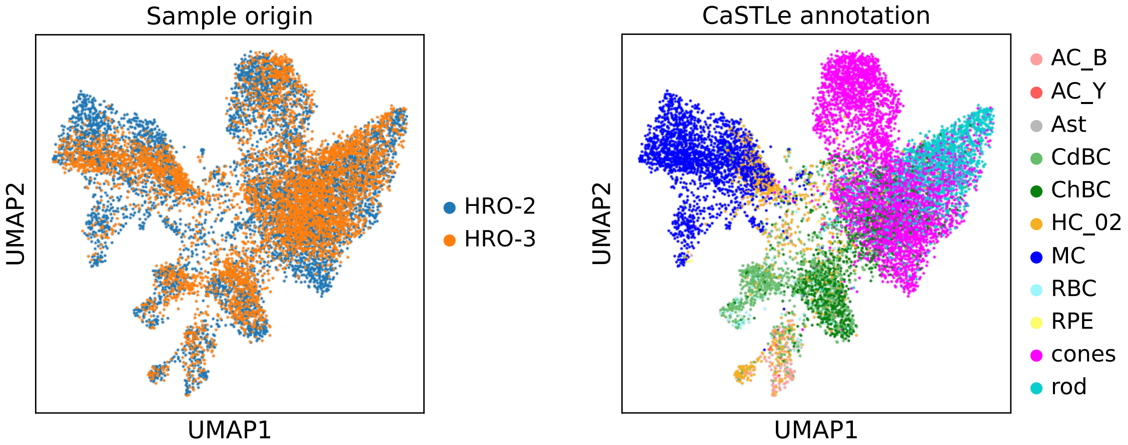
5.2.2 Comparison of HRO sample variance

Apart from the consistency of retinal cell types across annotation workflows, the cell cluster embedding of the two different HRO samples can be compared to each other. Therefore, both HRO samples were merged on top of each other, and the variance of cones, rods and MG cells was calculated. For this comparison, the `CaSTLe`-annotation was used. The results were summarized in Figure 5.9. Merging the HRO-3 sample on top of the HRO-2 UMAP embedding (Figure 5.9a and Figure 5.9b) revealed a general correspondence on sample and cell type level. Furthermore, a PCA of cones, rods, and MG cells for both HRO samples was calculated (Figure 5.9c-e).

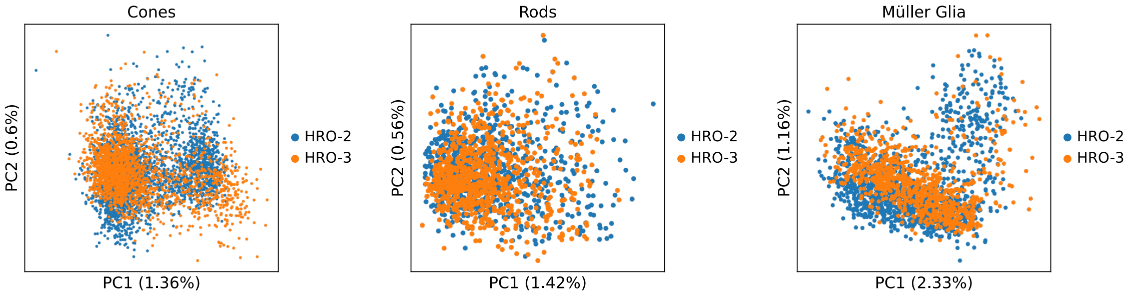
As shown in Figure 5.9c, both HRO samples revealed two slightly separated cone clusters. In HRO-2, however, both clusters seemingly contained a comparable amount of cells, which probably resembled the two cone clusters in the original UMAP embedding, cf. Figure 5.9b. In the control sample HRO-3, a larger discrepancy between both cone distributions was found, since most cones were located on the left part of the PCA and formed a rather sharply separated cluster.

Regarding rods, both HRO showed a very homogenous low dimensional embedding. This finding could also be supported by the embedding by sample origin (Figure 5.9a) and `CaSTLe`-annotation (Figure 5.9b).

Again a larger discrepancy between both samples was detected for MG cells (see Figure 5.9e). While HRO-2 MG cells were slightly more scattered, the HRO-3 appeared more compact in the centre. Comparing this result with the UMAP plots of the sample origin and `CaSTLe`-annotation, indeed HRO-3 MG-cells were more *compact*.



(a) *UMAP of merged HRO samples* **(b)** *UMAP of merged `CaSTLe`-annotations*



(c) *PCA of cones* **(d)** *PCA of rods* **(e)** *PCA of Müller Glia*

Figure 5.9. *Merged HRO controls data sets.*

UMAP embedding of merged HRO controls, coloured according to the sample origin (a) and the `CaSTLe` cell types (b). (c-e) PCA of three major retinal cell types coloured according to the batch origin. HRO-3 was merged onto the low dimensional embedding of UMAP-2. The first and second principal components indicate the percentage of variance explained. After merging both data sets, the results indicate a large correspondence on the levels of `CaSTLe`-annotations and HRO sample. While a larger discrepancy across both HRO samples was detected for cones (c), rods(d) and MG cells(e) show a similar distribution on the PCA plot.

5.3 Ambiguous correlation results of HRO towards fovea and periphery

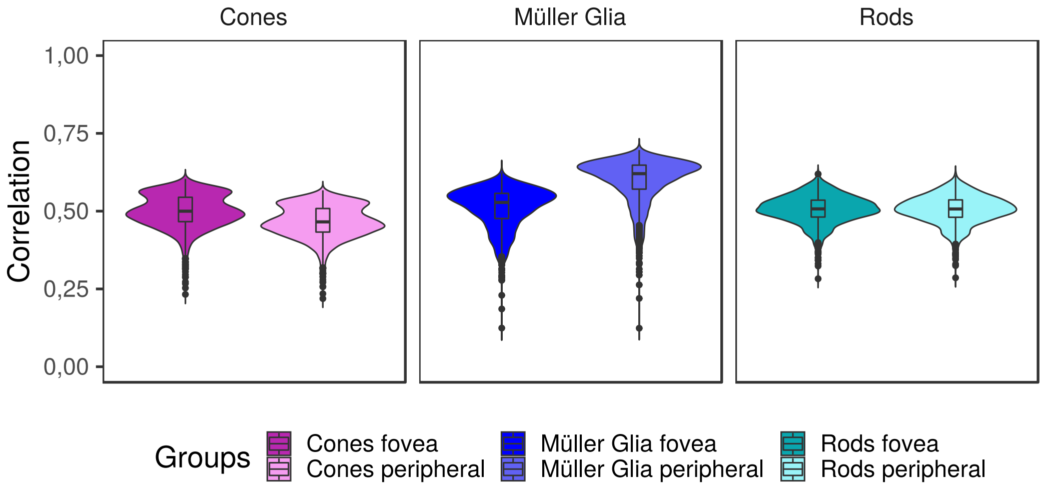
Using two different cell annotation approaches, two untreated HRO samples were quantitatively analysed. Though all major cell types were present in both HRO samples, it remained unclear if the organoid system is closer to the fovea or periphery of the retina. As a proxy, the number of rods and cones can be used, since the fovea has been described to be cone-rich and rod-sparse [Sakurai 2015]. Tackling this question, three different data sets were used. Again, the adult Cowan *et al.* [2020] single-cell data sets were included. Another human reference was used taken from Peng *et al.* [2019a], whereas a macaque data set was taken from Voigt *et al.* [2019]. All reference samples were collected from actual adult organs of the fovea and periphery. Focussing on photoreceptors and MG cells, the reference mean expression vectors from all three reference data sets were calculated for each cell type and both retinal regions. Therefore, the cell annotations provided in the respective Cowan *et al.*, Voigt *et al.*, and Peng *et al.* publications were used. In a further step, the `CaSTLe`-identified HRO photoreceptors and MG cells were correlated to these mean expression vectors. The reference data sets were not combined to infer the data set specific influence on the result.

Calculating the Pearson correlation of HRO photoreceptors and MG cells towards their peripheral and foveal counterparts of the Peng *et al.* [2019a] macaque reference data set, Figure 5.10 depicts the result of this analysis. Overall, HRO-2 (Figure 5.10a) and HRO-3 (Figure 5.10b) revealed comparable distributions of correlation values between foveal and peripheral comparisons. HRO-rods showed nearly identical median correlation values with 0.507 vs. 0.507 (HRO-2) and 0.528 vs. 0.529 (HRO3) to the foveal and peripheral reference samples, respectively. A minimal increase in correlation values towards fovea was detectable for cones in both HRO samples (HRO2: 0.500 vs. 0.466, HRO3: 0.487 vs. 0.453). The biggest difference in the mean correlation was observed for HRO-MG cells. Here, a larger trend towards the periphery was found with 0.528 versus 0.621 in HRO2 and 0.537 versus 0.635 in HRO3.

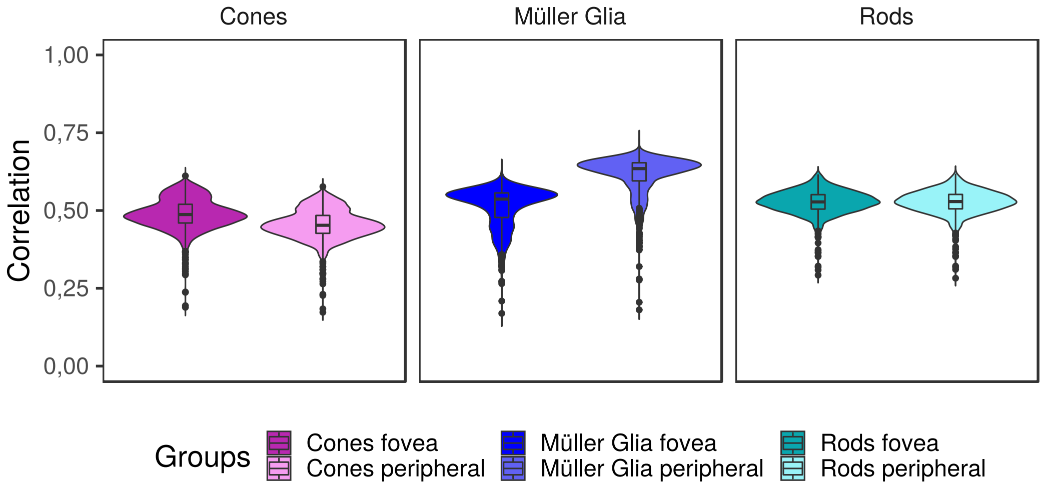
Switching from the macaque data to human reference samples, the results using the Voigt *et al.* [2019] data is depicted in Figure 5.11. This data set contained three individual human tissue samples. To derive better visibility, the results of the three human donors were combined. The expanded results are provided in Supplementary Figure S-9 and Figure S-10. Generally, the distribution of correlation values between foveal and peripheral rods as well as foveal and peripheral MG cells appeared more similar, while a larger difference was detected between both cone distributions. Again, HRO-cones revealed a higher median correlation of 0.597 in HRO-2 (Figure 5.11a) and 0.563 in HRO-3 (Figure 5.11b) towards fovea, while peripheral cones were found to have a lower median correlation of 0.283 and 0.267 in HRO2 and HRO3, respectively. A small difference in the distribution of correlation values can be found for MG cells. Reference foveal MG cells were found to have a minimally higher median correlation towards HRO-MG cells (HRO2: 0.617, HRO3: 0.514) than their peripheral counterparts (HRO2:0.583, HRO3: 0.473). The HRO-rods, however, have a comparable median correlation towards both foveal and peripheral rods with 0.505 and 0.488 in HRO2, and 0.476 and 0.458 in HRO3.

Regarding the individual donor results, generally similar correlation distributions were detected. Only slight differences were observable for donor-1, especially in the photoreceptors. Finally, Figure 5.12 summarizes the result of the human retina tissue reference data set by Cowan *et al.* [2020]. Similar to the Voigt data, the result of all three donors are included in Supplementary Figures S-11 and S-12. In contrast to both other data sets, neither of the three cell types revealed a trend towards fovea or periphery. The highest overall mean correlation values were detected for HRO-cones in HRO2 (Figure 5.12a) with 0.561 towards fovea and 0.550 towards the periphery. The lowest values were revealed by HRO-MG cells in HR03 (Figure 5.12b) with 0.545 towards fovea and 0.539 towards the periphery. Both HRO-rods owned median correlations of 0.506 to 0.507 for fovea or periphery in HRO-2, and 0.508 and 0.509 in HRO-3. Here, no difference across all individual donors was detected.

Summing up, the previously described results indicate that trends towards foveal or peripheral expression patterns largely depend on the reference data set and cell types. While nearly no difference in mean correlation values was observable for the Cowan *et al.* [2020] reference data, a strong trend of HRO-cones towards fovea was detected for the Voigt *et al.* [2019] data. The Peng *et al.* [2019a] data showed also a trend towards fovea, although far less pronounced. Regarding HRO-MG cells, a trend towards peripheral localisation was noticeable when using the macaque reference data set. This trend was, however, not observable in both human data sets. No difference in mean correlation values was detected for HRO-rods across all three reference data sets.

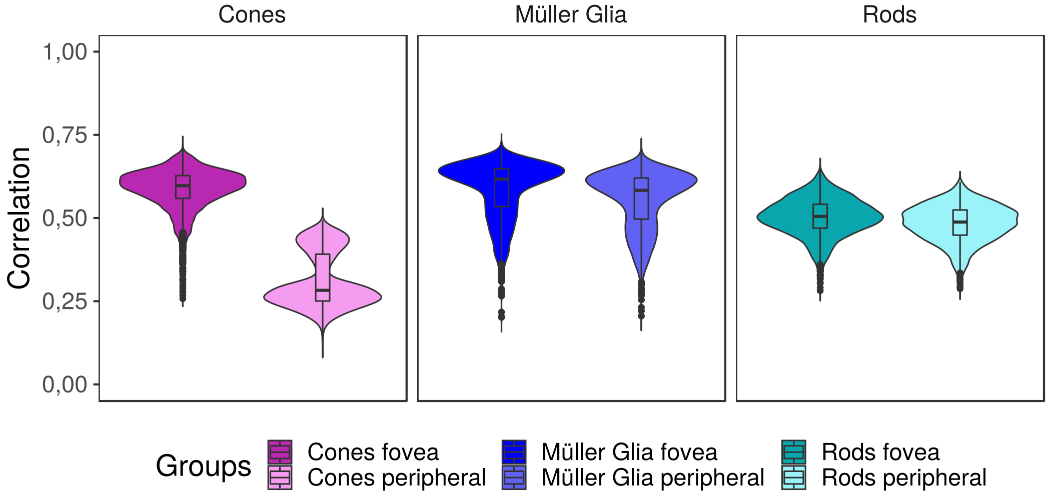


(a) *HRO-2 using `CaSTLe`-annotation*

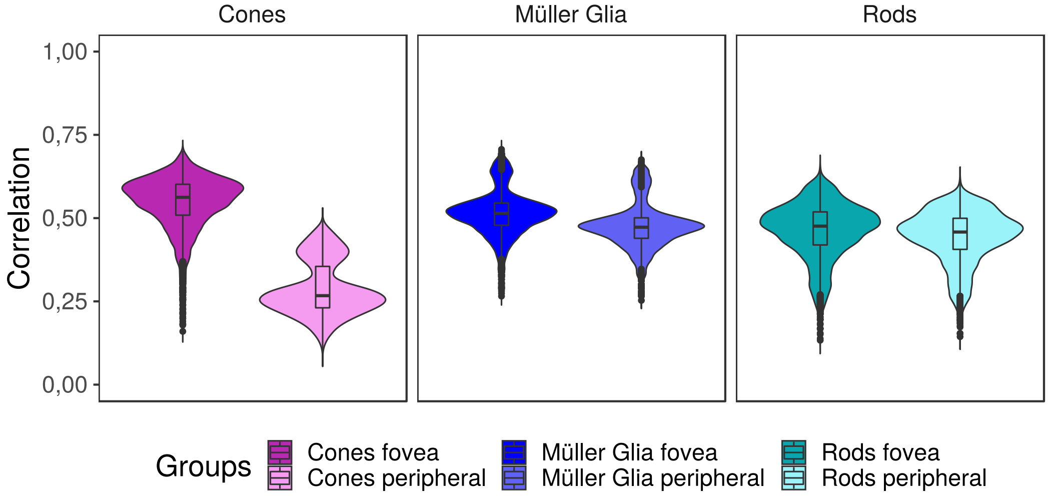


(b) *HRO-3 using `CaSTLe`-annotation*

Figure 5.10. *Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Peng* et al. *[2019a].*
Results of HRO-2 are shown in (a), whereby sample HRO-3 are shown in (b). Across samples, cells were annotated via the `CaSTLe`-annotation pipeline. The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells. Correlation distributions were comparable between foveal and peripheral expressions for both HRO-photoreceptors, though cones revealed a higher mean value towards fovea. A higher median correlation towards the periphery was detected for HRO-MG cells. These results were detected across both HRO samples.



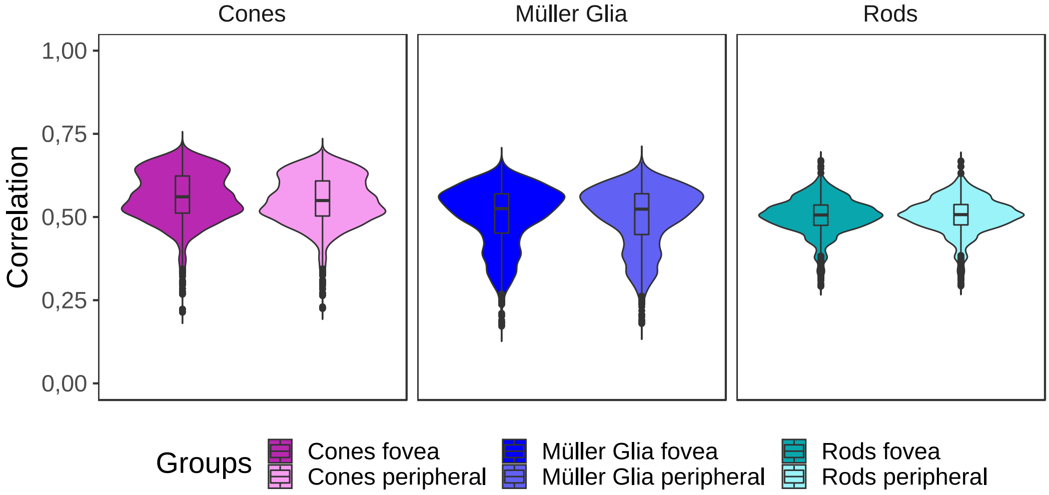
(a) *HRO-2 using* `CaSTLe`*-annotation*



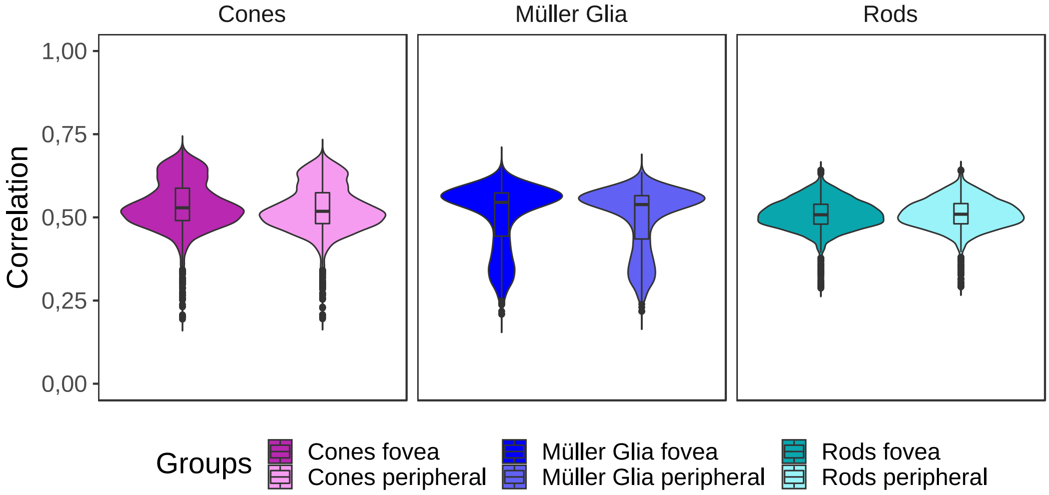
(b) *HRO-3 using* `CaSTLe`*-annotation*

Figure 5.11. *Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Voigt* et al. *[2019].*

Results of HRO-2 are shown in A, whereby sample HRO-3 are shown in B. Across samples, cells were annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells. Whereby HRO-cones owned higher mean correlation values towards fovea, HRO-rods again showed comparable values. Müller Glia cells revealed similar mean correlation values, with a slight trend towards the fovea. These results were detected across both HRO samples.*



(a) *HRO-2 using* `CaSTLe`*-annotation*



(b) *HRO-3 using* `CaSTLe`*-annotation*

Figure 5.12. *Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Cowan* et al. *[2020].*
Results of HRO-2 are shown in A, whereby sample HRO-3 are shown in B. Across samples, cells were annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of HRO cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells. Correlation distributions were highly similar between foveal and peripheral expressions for all three analysed HRO cell types. These results were detected across both HRO samples.*

5.4 RNA velocity analysis revealed dynamics within retina organoid systems

After characterising the cellular composition of the organoids, it should be furthermore investigated if developmental maturity of the organoids can be inferred. This maturity could also help to assess how well known developmental processes can be reproduced by the organoid systems. Using RNA velocity as a proxy for pseudotemporal ordering, the unsynchronized scRNA-seq data can be ordered by *maturity*. Together with a graph abstraction tool called `PAGA`, dynamics between cell types were investigated and visualised.

Briefly, the `CaSTLe`-annotated HRO samples were used as input for the `PAGA` algorithm, which detected cell clusters and calculated the connectivity between them. Based on this, a *new* UMAP embedding was calculated. Using the information of the RNA velocity analysis via `scVelo`, dynamics within and across the cell clusters were visualised. Additionally, the pseudotime, as well as the root and endpoint were calculated and indicated. The bolder the calculated edge, the stronger the connectivity between the two cell types.

The result of HRO-2 can be seen in Figure 5.13. For this analysis, the low dimensional UMAP embedding was recalculated based on the `PAGA` graph shown in Figure 5.13a. The original UMAP embedding results using the manual annotation are depicted in Supplementary Figure S-13. Using the RNA velocity information, and the `CaSTLe` cell annotation, the `PAGA` graph with directions was calculated. Generally, most trajectories ended in cones and rods. No connection ended in retinal pigment epithelium (RPE) and GABAergic amacine cells (AC-B). The reordered UMAP embedding resembled in general the *original* UMAP plot 5.13b. The inner retinal neurons (AC, BC and HC), photoreceptors (rods and cones), and MG cells were separated from each other. Revealing the dynamics within the HRO-2 samples, Figure 5.13c illustrates the regional development directions. Figure 5.13d shows the pseudotime* result of HRO2. Inferring the RNA velocities allowed to order the HRO cells based on their relative development. While the inner retinal neurons were overall *earlier*, rods appeared to be the most recent cell type. This finding was supported by the results from 5.13e and 5.13f, where the root and end point† of the data was identified, respectively. With AC, HC and some parts of the MC cluster were detected as the root cells, mature cones and rods indicated

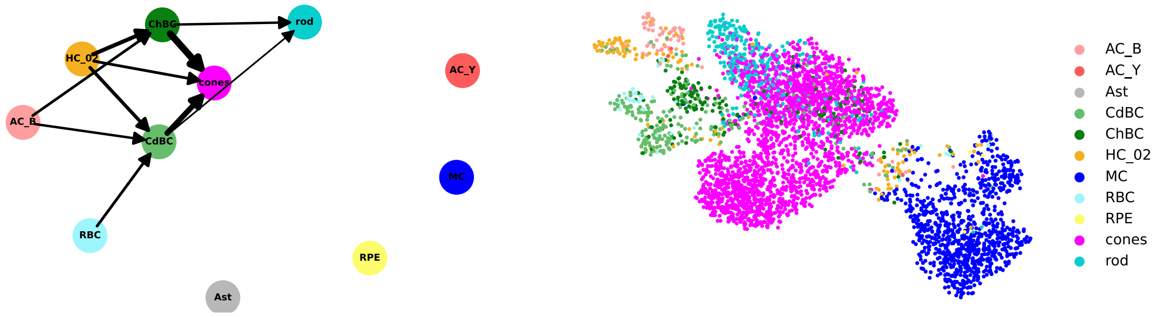
---

[*] Velocity pseudotime is a random-walk based distance measure on the velocity graph. After computing a distribution over root cells obtained from the velocity-inferred transition matrix, it measures the average number of steps it takes to reach a cell after start walking from one of the root cells. citation APIs

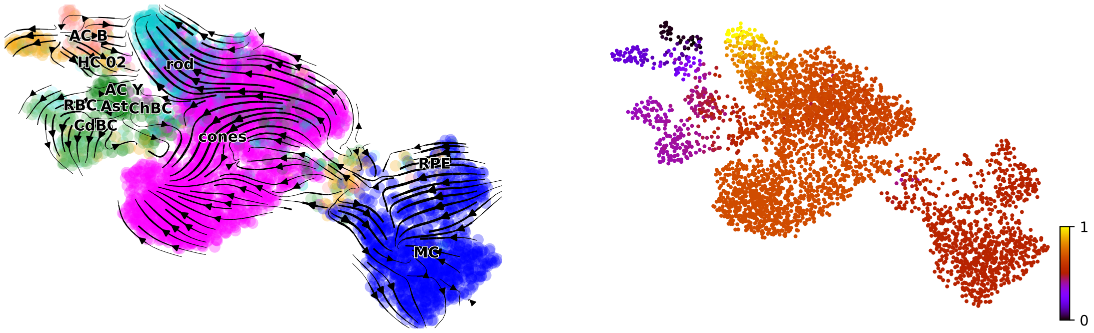
[†] The end points and root cells are obtained as stationary states of the velocity-inferred transition matrix and its transposed, respectively, which is given by left eigenvectors corresponding to an eigenvalue of 1, citation of tutorial

the end points. No endpoint was detected for the MC cluster.

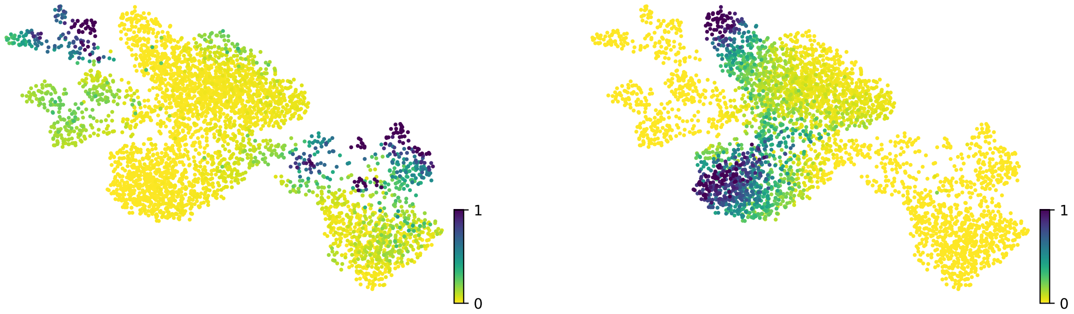
Similar results were detected for HRO-3 as depicted in Figure 5.14. Focussing first on the `PAGA`-graph shown in Figure 5.14a, again the boldest and most numerous connections ended in cones and bipolar cells (BC). No connection ended in RPE and MC cells. The recalculated UMAP embedding (Figure 5.14b), as well as the regional dynamics (Figure 5.14c), revealed one obvious difference to HRO-2. In HRO-3, both the `PAGA`-graph and the RNA-velocity streams indicated that the population of rods own a trajectory into the cone cluster. Here, the pseudotime analysis highlighted MG cells as the earliest and most cones as the latest cell type. Generally, there was a sharper distribution between *early* and *late* time points with only a few cells being somewhere in transition than in HRO-2. The root cell analysis however showed that the inner retinal neurons were *earlier*, similar to HRO2. Still, the most recent cells were within the photoreceptor cluster.



(a) *PAGA coarse-grained trajectory* **(b)** *UMAP using PAGA embedding*



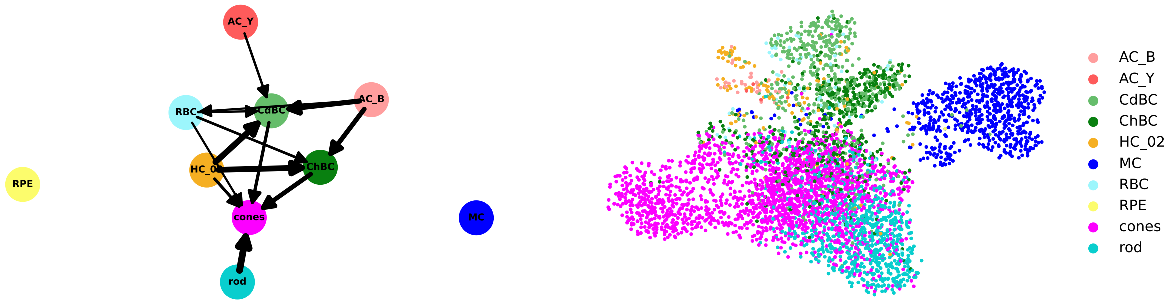
(c) *RNA velocity streams* **(d)** *Pseudotime*



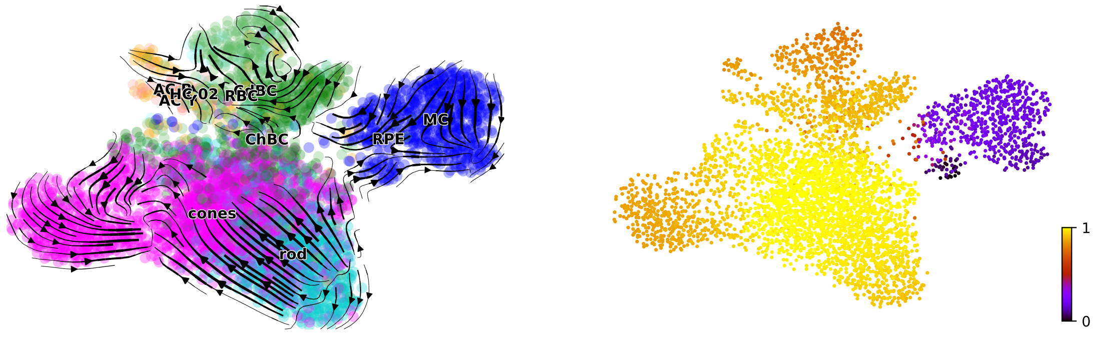
(e) *Root cell visualization* **(f)** *Endpoint visualization*

Figure 5.13. *Trajectory inference of HRO-2 using RNA velocity and PAGA.*

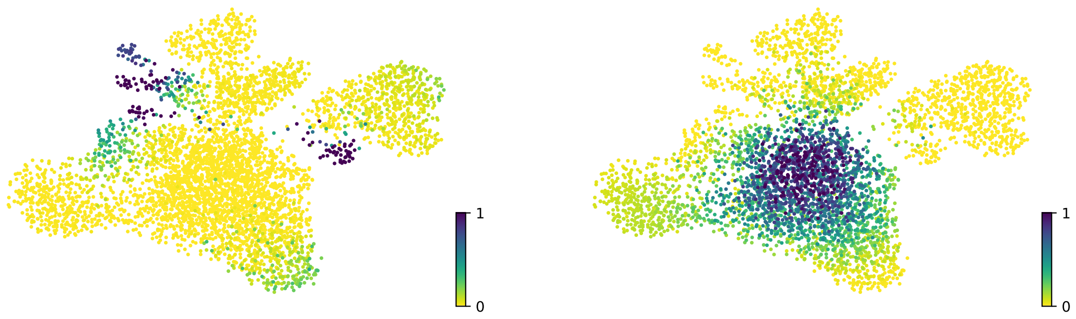
(a) Using the RNA velocity dynamics, PAGA calculated a coarse-grained developmental trajectory using the CaSTLe-annotation. (b) Based on this embedding, the UMAP was recalculated. (c) Regional developmental dynamics were visualised onto the UMAP embedding. (d) Ordering the HRO-2 cells via the RNA velocity calculations allowed to infer a pseudotime-variable. Dark violet indicated early whereby yellow represented late cells. (e+f) Using the pseudotime, root (e), as well as endpoints (f) in the data, were identified in HRO-2. Dark blue here indicated the root or endpoints, respectively.



(a) *PAGA coarse-grained trajectory* **(b)** *UMAP using PAGA embedding*



(c) *RNA velocity streams* **(d)** *Pseudotime*



(e) *Root visualization* **(f)** *Endpoint visualization*

Figure 5.14. *Trajectory inference of HRO-3 using RNA velocity and* `PAGA`*.*
(a) Using the RNA velocity dynamics, `PAGA` *calculated a coarse-grained developmental trajectory using the* `CaSTLe`*-annotation. (b) Based on this embedding, the UMAP was recalculated. (c) Regional developmental dynamics were visualised onto the UMAP embedding. (d) Ordering the HRO-3 cells via the RNA velocity calculations allowed to infer a pseudotime-variable. Dark violet indicated early whereby yellow represented late cells. (e+f) Using the pseudotime, root (e), as well as endpoints (f) in the data, were identified in HRO-3. Dark blue here indicated the root or endpoints, respectively.*

5.5 Discussion

To decipher the molecular pathways involved in AMD, a novel human retina organoid system called HRO was established. Here, several analysis steps have been applied on the single-cell level to initially prove the stability and validity of the organoid system to be adequate and effective for this task. Using two different cell cluster annotation approaches, and cell trajectory inference, the HRO composition and developmental status was highlighted.

Reliable cell type annotation requires meaningful external knowledge. In this dissertation, two annotation approaches were applied comparatively. In a first attempt, cell clusters in the HRO single-cell data were detected. Employing a collection of known marker genes, their expression pattern was compared across these cell clusters and expert knowledge was used to annotate these clusters based on their expression. Secondly, a transfer learning tool, called `CaSTLe` was employed which extracts useful features from a reference data set and uses them to train a random forest model. Finally, this classifier was used to annotate the HRO cells.

One advantage of the manual annotation workflow is its adaptability and flexibility. While machine learning tools heavily rely on suitable reference data sets, the analysis of marker genes can integrate the knowledge from many studies and previous experiments. As of for the HRO system, some cell clusters did reveal both, cone and rod specific expression patterns. Instead of forcing the assignment into fixed groups, an additional class of premature photoreceptors was created to suit the problem. Since the HRO system is still not fully mature, it appears more reasonable to not finally decide on the final entity of these *premature* cells.

Though the manual annotation approach appears to be uncomplicated, it harbours some limitations. Most prominently the manual annotation depends on the current state of knowledge and often (external) expert knowledge. Another problem arises from some discrepancies between protein level detection and gene expression [Liu *et al.* 2016]. While many marker genes were often defined on protein level, for example using antibody staining, the transcriptional signal must not be equally prominent. With more datasets being produced and analysed, more knowledge is gathered, which allows extracting marker genes for specific cells more rapidly and widely. However, the interpretation of the resulting expression patterns across cell clusters is not only time-consuming but also often not reproducible. As also visible in this dissertation, some *marker genes* for example show a very uniform expression across all clusters, which suggest that those may not be suitable for annotation on the RNA level (see Supplemental Figure S-7c). Moreover, the decision on whether a gene is considered to be *expressed high enough*, often underlies subjective, arbitrary thresholds. A marker gene-based annotation can also become very challenging when using very similar (sub)cell types. Like for the human retina organoid, rods and cones for example belong both to the class of photoreceptors. Therefore, marker genes for these cell types often contain re-occurring genes. Both, the decision to exclude all non-unique genes or leave them in, influences the final cell annotation.

Though it is more robust to annotate whole clusters instead of individual cells, the annotation result is highly dependent on the clustering itself. This problem becomes already prominent when switching either the clustering algorithm, the version of the analysis pipeline, or even switching the whole analysis tool. Due to this reason, the older version of `scanpy` was used for cell type annotation to ensure reproducibility of the cell clusters. Therefore, the transfer learning approach is less susceptible to these changes in cell clustering.

Being more robust in terms of low dimensional embedding, the performance of `CaSTLe` was highly dependent on the used reference dataset. In this dissertation, three different reference datasets by [Cowan *et al.* 2020] were used which mainly differed in the tissue maturity, origin, and sample site. While the developed organoid data revealed very similar annotation results as the manual workflow, no premature photoreceptor cluster was encountered since it was not included in the reference data set. Referring to the classification performance during training, very high specificity values were obtained, especially regarding rods, cones, and MCs. However, comparing the developmental status of the reference data sets to the HRO system, the Cowan *et al.* organoids were older with 30 and 38 weeks versus 28.5 weeks (200 days). Using two other reference data sets, more diverging annotation results were obtained. These data sets were taken from the fovea and periphery of an adult tissue sample, respectively. While the very general cell cluster layout detected in the manual approach was roughly maintained using the foveal data set, nearly no photoreceptor cells were annotated employing the peripheral reference data set. These results indicate that the choice of a suitable reference data set is the most critical step for this machine-learning based annotations. While the developed organoid data revealed very similar annotation results as the manual workflow, the results from both adult reference datasets were very different. To promote the availability of high-quality reference data sets, not exclusively limited to single-cell annotation, many efforts were undertaken in generating so-called *atlases* [Papatheodorou *et al.* 2020, Travaglini *et al.* 2020]. These databases contain next to expression data also high-quality annotation results, for example from Fluorescence Activated Cell Sorting (FACS). Yet these atlases must be filtered for certain variables such as developmental time points or target organisms. Depending on the size of these atlases, the resulting dataset might be very small, deducing the classification performance of `CaSTLe`, or other machine learning annotation tools.

Regarding this dissertation, no suitable atlas-data set was available by that time.

These reference data-sensitive results of the `CaSTLe` workflow do furthermore emphasize the usefulness of using two *independent* annotation approaches. Though the restriction to current knowledge was listed as a disadvantage of the manual annotation workflow, in-depth research on consensus and well-known marker genes could help to assess the quality of the annotation results derived from machine learning-based approaches. Altogether, it can be summarized that both, the manual annotation and the machine learning-based `CaSTLe`-workflow revealed overlapping results. Solely the ratio of rods and cones was changed, mainly due to the missing premature photoreceptor cluster. Due to its advantages towards reproducibility and the higher resolution in cell types, the `CaSTLe`annotation was used for all downstream tasks.

Premature photoreceptors appear to be mostly cones resulting in a cone-rich organoid system. As already stated in the previous part, different quantities of photoreceptors were detected after the manual annotation and `CaSTLe`. Whereby very similar quantities (HRO2: 28% cones, 25% rods, HRO-3: 23% cones, 22% rods) were detected for each photoreceptor cell-type using the manual annotation, 41% (HRO-2) and 37% (HRO-3) of HRO cells were labelled as cones via `CaSTLe`. Thereby, the amount of rods was reduced to 12% (HRO-2) and 15% (HRO-3). The main reason for this shift is the missing class of premature photoreceptors in the `CaSTLe`-reference data. Previously described results indicate that most cells from this manually assigned premature cluster shifted towards cones and only a smaller part towards rods. The largest discrepancies, however, could be observed for the manually assigned rod cells, where equal parts were assigned to rods and cones with `CaSTLe`. Although these differences might sound irritating at first, their low dimensional embedding revealed a convincing result: both previous cone sub-clusters were combined into a larger cone cluster that left a smaller rod population in proximity but still separated. Taking into consideration the RNA velocity streams of the original UMAP embedding and the manual annotation results, the `CaSTLe`-annotation can be supported. Especially for HRO-2, the velocity streams from the lower manual-cone cluster move upwards, over the premature cluster towards the upper cone location.

Regarding HRO-3, the interpretation of the velocity streams on the `CaSTLe` as well as manual annotation was less obvious. As indicated by the `PAGA`-graph, the rod population *transitions* into the cone cluster. Regarding the manual annotation embedding, the RNA velocity also highlighted a trajectory from rods to cones, spanning over the premature photoreceptor cluster. These results might indicate that perhaps more manually annotated rods belong to the premature photoreceptor cluster. In the case of the `CaSTLe`-annotation, perhaps only the outermost tip of the UMAP embedding (see Figure 5.14c) represents fully differentiated rods.

Together with the different amounts of premature photoreceptors from the manual annotation, one might suggest that the *maturity* between HRO-2 and HRO-3 was different. Whereby more premature photoreceptors were initially annotated in HRO-3, also developmental differences were highlighted by the RNA velocity analysis. As HRO-2 owned two end-points in cones and rods each, HRO-3 only revealed one which largely mapped to the `CaSTLe`-cone population.

HRO system may represent parafovea. Since AMD is understood to start in the parafoveal region of the retina, it was investigated if these organoids sufficiently resemble this foveal region. After annotating the HRO systems via the `CaSTLe` workflow, the expression of rods, cones, and MG cells was correlated to three adult human and macaque reference data sets. There, tissue samples from the foveal and peripheral regions of the retina were extracted and analysed via a single-cell transcriptomics platform. Using the provided annotation, mean expression vectors of cones, rods, and MG cells were calculated and correlated to `CaSTLe`-annotated photoreceptors and MG cells.

Generally, no clear and coherent trend across all reference datasets was observable. Whereas cones owned higher mean correlation values towards the foveal region in the human Voigt *et al.* data set [Voigt *et al.* 2021], no clear difference was observable for the other two data sets. One possible explanation is that the HRO systems exhibit so-called *parafoveal* characteristics (see Figure 5.15). This could partly be explained by the used sample size of the reference data sets. While Voigt *et al.* sampled a 2 mm area of fovea and periphery, both other datasets used only a 1.5 mm tissue sample. Due to the larger sample size of the Voigt data, parts of the parafoveal region might also have been extracted and analysed. The presence of this parafoveal area could explain the large differences in correlation distributions for cones in the Voigt data, which was not observable for both other datasets.

Opposing trends however were detected for MG cells. Whereas a higher median correlation towards the periphery was found in the macaque data, the Voigt *et al.* [2019] data suggests a foveal localisation. For the Cowan *et al.* [2020] reference data set, no difference could be found. This finding could be explained by the primate origin of the retina tissue in the Peng data, which ultimately leads to the question of how well findings between primates and humans can be extrapolated, even more so in MG cells. In the publication by Syrbe *et al.* [2017], a unique MG cell type of the primate inner fovea was described which is called the *Müller cell cone*. The dataset by Peng *et al.* did not explicitly discuss or annotate this special MG sub-cell type. Due to this primate-specific cell type located in the fovea, the Pearson correlation values might have revealed a periphery-like trend in MG cells.

While some trends were stated for cones and MG cells, the `CaSTLe`-annotated rod cells did not show any trend across all included reference expression vectors. A possible explanation may be given by the cell type compositions of the retinal sample sites. Generally, the fovea is a cone-rich, rod-sparse (nearly absent) region, located four degrees from the central fixation point [Sakurai 2015] (see Figure 5.15). With increasing distance from the fovea, rod cell counts become more prominent, whereas the cone density declines. Therefore, just by cell type distributions of the HRO-system after the `CaSTLe`-annotation, the data suggest a more parafoveal characteristic.

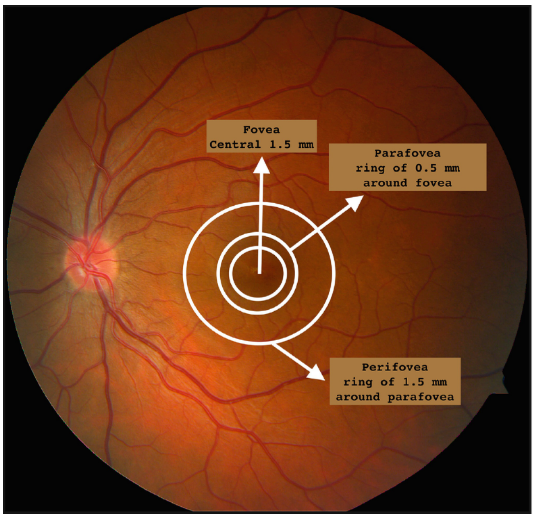


Figure 5.15. *The spatial location of the parafovea in the eye taken from Tsang & Sharma [2018].*
The peripheral region lies approximately 9mm away from the foveal centre.

Other problems of course may arise from the initial annotation of the reference datasets. As already described for the Cowan data, the authors also used a manual annotation based on marker gene expression. A similar workflow was performed for the primate data by Peng *et al.*. Voigt *et al.* used a combination of previous knowledge and the relation between cell clusters via a dendrogram*. The decisions on meaningful marker genes, thresholds of expression values and cluster algorithms heavily affect the final annotation result as it was already discussed in detail in the previous section. Therefore, the presented correlation analysis is biased from both sides: the HRO cell type annotation that was done with `CaSTLe`, and the predominantly expert knowledge-based cell type annotation that was done in the reference data.

Trajectory inference revealed overlapping results to literature. Diving deeper into developmental processes inside the retinal organoids, RNA velocity delivered interesting insights into major cell trajectories and temporal aspects. Using the ratio between unspliced and spliced transcripts, RNA velocity aids in ordering single cells along an artificial pseudotime axis, and hence helps to infer knowledge of developmental trajectories and timing within the organoid. Via this approach, information about cell developmental trajectories and timing could be ex-

---

*Based on transcriptional similarity

tracted. Due to the age of the HRO-system, mainly the concept of developmental trajectories was used to infer *maturity*, not cell transitions.

Within the HRO system, the graph abstraction algorithm `PAGA` and the information from RNA velocity highlighted a pseudotemporal ordering of the cells. Across both HRO samples, horizontal (HC) and amacrine cells (AC) were detected as *early* cells. While no incoming edges were detected for those two cell types, also the root cell visualization supported those findings. These findings correspond to findings by Sridhar *et al.* [2020] and Quinn & Wijnholds [2019] about the general developmental projection of human fetal retinal cells, see Figure 5.16. Starting from an initial progenitor population of *T1* cells, two major transitions were identified. While ganglion cells originate from the T1 cell group, amacrine and horizontal cells derived from the *T2* progenitor. Lastly, bipolar cells and photoreceptors derive from the *T3* cluster. Considering the temporal aspect, Quinn & Wijnholds [2019] reported that amacrine and horizontal cells are both amongst the *earliest* cell types within the retina cell genesis(see Figure 5.17). Similar results were stated by Cowan *et al.* [2020], who sequenced retinal organoids across seven time points. Also, horizontal and amacrine cells were present in the youngest organoids, followed by bipolar cells and cones.

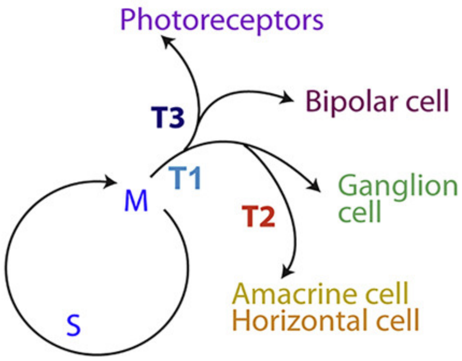


Figure 5.16. *Development of retinal cells taken from Sridhar* et al. *[2020]. T1, T2, and T3 describe three different transitional states of retinal cell populations. The circle represents the cell cycle stages of DNA synthesis (S) and Mitosis (M).*

As highlighted by the RNA velocity results of HRO-2, both photoreceptor cells represent the most recent cell types. However, in HRO-3 solely endpoints were calculated for the cone cell cluster, and a transition from the rod towards the cone population was identified via `PAGA`. As already mentioned in this discussion, differences in the velocity streams and quantities of premature photoreceptors, which were annotated in the manual approach, indicated different degrees of maturity in organoid development. Figure 5.17 also indicates that cones develop prior to rods, furthermore strengthening the indication that HRO-3 is less developed than HRO-2.

Generally, diverging results were obtained regarding the origin and timing of bipolar cells. While Quinn & Wijnholds [2019] indicated a rather late development, Cowan *et al.* [2020]

showed that bipolars emerged in the organoids at the same time point as cones. Moreover, the trajectory from which it originated was different across data sets. As indicated in Figure 5.16, bipolar cells were found close to the photoreceptor transition, whereas Cowan *et al.* [2020] revealed that they develop from the amacrine-horizontal-ganglion-axis.

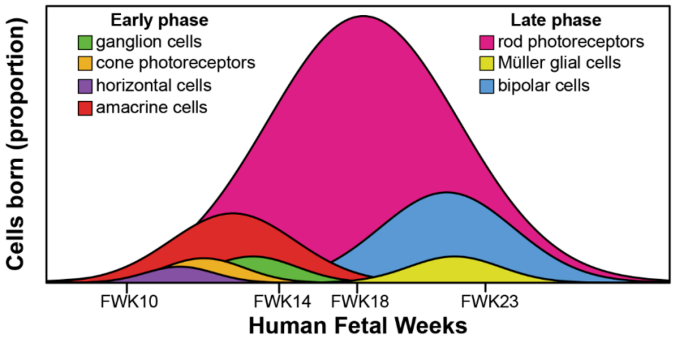


Figure 5.17. *The genesis of retinal cell types taken from Quinn & Wijnholds [2019]. Generally, two phases during cell genesis can be identified: Early phase (ganglion cells, cone photoreceptors, horizontal cells, and amacrine cells) and an overlapping late phase (rod photoreceptors, Müller glia cells, and bipolar cells. FWK stads for fetal week.*

Adding to the previous point, though all cell types were annotated in the HRO-system, no MG cells were identified as *most recent*, since they evolve slightly later than cones, according to Quinn & Wijnholds [2019]. Indeed, a few MG cells in HRO-2 revealed a rather early pseudotime point as calculated by RNA velocity. Though not included in Figure 5.16, other low dimensional embeddings of later fetal time points in the Sridhar *et al.* [2020], indicate a proximity to the progenitor cell cluster upstream of T1. This may indicate a rather early development of MG cells, with is contradicting to the results shown by Quinn & Wijnholds [2019]. These findings already point out the differences in organ development between the *in vivo* and *in vitro* tissues, as it was also indicated by Cowan *et al.* [2020]. Despite some discrepancies, the comparable results that have been described in the literature promote the usefulness of this data-driven analysis.

In general, it was demonstrated that using different annotation workflows, very comparable results can be achieved to characterize a neo-natal human retina organoid. While increasing the reproducibility of the annotation results via the machine learning tool `CaSTLe`, the manual cell assignments already provided some valuable information about the quantity of not yet fully developed photoreceptors. Solely relying on `CaSTLe` would not deliver this information. On cell cluster level, both HRO samples were very similar, however, differences on the developmental level between both samples were detected. Still, it remains unclear if the HRO systems resemble the para-foveal region of the retina, or not. Together with the RNA velocity analysis, general developmental trajectories of the retina were verified, which proposes the suitability of the HRO system to study AMD. To furthermore confirm its suitability, the differences between the organoid and adult retina should be overcome by for example improving the culture conditions.

Chapter 6

Conclusion

After the identification of the DNA structure in the '50s by Watson and Crick, it was not until the establishment of sequencing workflows by Frederick Sanger in the late '70s that the very basic structure of life was deciphered. While huge costs were faced on the computational as well as experimental side when sequencing the first human genome, published in 2001 [Lander *et al.* 2001], the emergence of so-called *next-generation sequencing* methods paved the way for excessive, cost-efficient, and quick sequencing. This novel, global DNA and RNA analysis allowed gaining major insights, revolutionising not only the fields of biology and medicine. A new era emerged with the establishment of single-cell-based assays. The combination of both the single-cell resolution and the general sequence analysis opened up another era in systems biology. In this final chapter, the previously demonstrated results will be briefly summarised and their novelty discussed. As already indicated throughout the whole dissertation, current challenges and limitations with respect to reproducibility will be pointed out, while proposing possible solutions. In the end, an outlook will be provided, suggesting potential starting points for follow-up projects.

Recapitulation of the results. To infer the influence of various levels of dropout and subsequent data imputation with respect to network inference on single-cell transcriptomics data, a synthetic data set generated and used. Starting from a downsampled bulk RNA-seq data set, the true correlation signals were identified via `WGCNA`, and subsequently masked by six artificial levels of dropout. The increase in dropout was associated with a stepwise eradication of the correlation structure until no preservation could be identified in the highest dropout level. By applying six different data imputation tools to the artificial dropout data sets, three major insights could be distilled: (1) Within low dropout scenarios, major correlation structures were still preserved, and hence allowing for direct network inference. (2) On the opposite side, all imputation tools failed to recover the original correlation structures but introduced large amounts of false correlation signals in the high dropout data sets and (3) in moderately sparse data sets, data imputation techniques proved beneficial to recover the correlation structure. In those moderately sparse data sets, the tool called `DCA`, a deep count autoencoder, revealed a peak of performance. These results highlighted a small *window of opportunity* for network inference after data imputation. Moreover, it was investigated if the set of imputation tools influenced cell clusters' annotation via marker genes. Using a human retina organoid data set by Kim *et al.* [2019], none of the included imputation tools interfered with cell cluster annotation.

`DCA` was applied to a human retina organoid based on these encouraging results to analyse the gene correlation networks in a biological context. Focussing on the complete as well as cell type-specific networks, `DCA` allowed fulfilling the scale-free topology criterion used by `WGCNA` which is a hallmark of biological networks. The derived cone- and rod-specific correlation networks were furthermore characterised and compared. Using the respective marker genes for these two cell types, unique cell type-specific gene modules could be identified for both subnetworks, indicating that `DCA` indeed enhanced the true cell type-specific expression signals buried within the dropped-out single-cell data. Additionally, hub genes were identified across modules and networks, preserving their hub gene status and regulating different correlation networks. One of them was EIF4EBP1, a eukaryotic translation initiation factor binding protein associated with the ErbB-signaling pathway. Upon the onset of age-related macular degeneration, a major cause of blindness in developed countries, ErbB was associated with causing cell death in retinal cells [Sheu *et al.* 2019].

Alongside a technical evaluation of data imputation and their benefits for deriving deeper resolution in network inference, the toolbox of the single-cell realm was also used to characterize a novel human retina organoid system called HRO. Two organoids were characterized, by employing two different cell annotation procedures based on a manual marker gene expression analysis or a transfer learning pipeline. Overall, the results were comparable, though the transfer learning tool `CaSTLe` revealed a cone-rich cell composition, while the manual annotation returned equal amounts of rods and cones. Both HRO samples indicated a good correspondence after merging both data sets, though slight shifts in cell embeddings were observable for Müller glia cells and cones.

Since many macular diseases are known to set off in the foveal-parafoveal region of the retina, it was investigated if the HRO-system exhibited a foveal or peripheral expression pattern within photoreceptor and Müller glia cells. Using reference samples of primate and human origin did not indicate a consistent trend in correlation towards either a foveal or a peripheral region.

Finally, the calculation of developmental trajectories within the HRO samples and the compression of the cell clusters into abstracted graphs helped to compare the HRO samples with respect to their developmental status. This analysis indeed highlighted differences between both control samples. While already more *premature photoreceptors* were annotated in HRO-3 using the manual cell annotation, the developmental transitions between retinal cell types identified different pseudotemporal characteristics. As rods and cones were detected as being the *most mature cell type* in HRO-2, solely cones were pinpointed in HRO-3, alongside an unusual trajectory from the rod into the cone populations. These results display differences within the developmental *status-quo*, which were not visible initially.

Novelty of this work. In this dissertation, a human retina organoid was analysed, providing a basis to study the disease mechanisms associated to age-related macular degeneration (AMD). The identified cell type quantities in the untreated organoids can also enlighten more subtle changes after the initialisation of the disease. Moreover, using RNA velocity, *baseline* developmental trajectories were inferred, which could serve as a reference.

Due to the many promises and possible applications in single-cell transcriptomics, many efforts were undertaken concerning method development, particularly to adjust and adapt to the specific sparse data characteristics. With more knowledge of the data characteristics, more efforts were undertaken for the identification of *best-practice* workflows.

Aside from plentiful prospects of single-cell transcriptomics, also many challenges and limitations arose. As gene correlation networks proved useful for conventional bulk RNA-seq data analysis, it was initially unfeasible on the single-cell data based on a lack of performance of the inference tools [Chen & Mar 2018]. Due to the huge amount of sparseness alongside the large data sets, well-established tools were not applicable anymore, while new algorithms lacked reproducibility and robustness. Based on these prerequisites, this dissertation sought to answer how different degrees of sparseness affect network inference if data imputation could facilitate network inference, and what useful insights can be gained from cell type-specific gene networks. Providing a robust benchmarking framework, various sparsity levels and imputation tools were applied on a reference data set, identifying a *window of opportunity* for low and moderate levels of dropout. Transferring these insights to a biological context, `DCA`-imputation of a retina organoid allowed to infer biologically meaningful cell type-specific gene correlation networks. However, also data imputation did not allow for correlation network inference from high dropout data sets, still retaining the potential of the data. These restrictions may be alleviated with technical improvements such as capturing the mRNA more efficiently and advanced method development in both, direct network inference and data imputation. With growing knowledge about the data structure, more suitable assumptions and models can be used. More recent advancements, for example, propose to step away from the dropout model and move towards a probabilistic model in which all transcripts within a cell are equally likely to be captured and sequenced [Breda *et al.* 2021].

Reproducibility of single-cell results. Though the many possibilities of single-cell transcriptomics and continuous tool development, one major constraint concerns the reproducibility of the results. As a disclaimer, this topic is not exclusive to the single-cell omics field; however, due to its recency, it becomes very eminent. In this section, general limitations will be discussed, finishing with a few suggestions on ensuring reproducibility in the context of single-cell transcriptomics.

With respect to this dissertation, concerns were encountered in the general reproducibility of results across pipelines and versions. For example, in the HRO analysis, the manual annotation pipeline was run using an older version of `scanpy`. The reason for this was a different number of detected clusters when rerunning the analysis in a higher version pipeline, which would have resulted in a complete revision of the annotation results used for downstream analysis. As already mentioned in the respective discussion, an annotation procedure not relying on a previous cell clustering would be more reproducible and robust. The obvious reasons lie in the recency of this field and the rapidly growing insights, which cause a frequent version update of single-cell workflows. This recency was moreover observable since, generally different pipelines and coding environments were used. While python and `R` are predominantly used in the single-cell transcriptomic universe, also two different preprocessing workflows were established, namely `scanpy` and `Seurat`. Though technically no restriction in the interoperability between these pipelines can be stated, it is more time-consuming to ensure proper data ex- and import. To ensure reproducibility of the results, it is more consistent to stay within the same analysis pipeline. However, exceeding the steps of basic data preprocessing and clustering to employ more downstream analysis, two different scenarios were equally likely: (1) Sticking to the same analysis pipeline for a tool with weaker performance reducing the expressiveness of the inferred result, or (2) switching the pipeline for a tool with the best performance but simultaneously sacrificing reproducibility. In the context of this dissertation, a trajectory inference method, `PAGA` was used, which was embedded in the `scanpy`-environment. While this specific tool revealed good performance in a benchmarking publication [Saelens *et al.* 2019], no generalisation can be implicated regarding other downstream tasks or tools.

Due to an initial absence of tools that could adapt for the specific single-cell characteristics, such as size and sparsity, a huge increase was met in method development. Now, systematic studies aim at comparing a subset of these tools as well as the parameter choices, helping to establish recommendations depending on the data structure and supplementary information [Wang *et al.* 2020, Soneson *et al.* 2021].

Adherence to FAIR principles. After stating concerns regarding the reproducibility of analysis workflows in the field of single-cell transcriptomics, data and script provision could alleviate these limitations. Generally, transcriptomics data aims to provide an untargeted and comprehensive picture of the gene expression pattern of a certain organism at a certain time point or treatment. No measurement is truly unbiased. Since these studies are both cost- and work-intensive, researchers tried to establish public databases, where other researchers could access and analyse the data without the need to re-generate it for themselves. Likewise, public code and workflow repositories such as Github and Gitlab were founded to allow easy code and script sharing, since bioinformatic analysis grew over the last decades. These databases should adhere to the FAIR principles, which state that data should be stored findable, accessible, interoperable, and reusable [Wilkinson *et al.* 2016]. However, during the course of this dissertation, major difficulties with at least one of these principles were encountered, or (sub-)results were not provided at all.

Regarding the latter point, often, the cell annotation results derived within the publications themselves were not published. That became most prominent when researching single-cell transcriptomics data sets for the annotation of the HRO system. In the Kim *et al.* [2019] publication moreover, solely a rough annotation workflow was provided instead of the actual script. The combination of incomplete data provision and script allocation severely reduces the reproducibility of the results. Moreover, a significant amount of time must be spent on re-generating the described annotation results. Oppositely, an optimal data supply was provided by Cowan *et al.* [2020]. Next to the expression data, the annotation and a visualisation platform were also published.

However, some of these limitations could be diminished by establishing *gold standard* data and code sharing workflows. Similar efforts were already undertaken regarding public databases by defining *best-practice* workflows or checklists to increase reproducibility and transparency of the results [Kolker *et al.* 2014, Kenneth M. Merz *et al.* 2020, Chervitz *et al.* 2011]. Apart from the (un-)processed expression data itself, which is already frequently hosted on Gene Expression Omnibus (GEO), also the corresponding `Seurat`, or `scanpy`-objects should be supplied, as all required information will be included at their correct position. To further increase the interoperability, the cell annotation should either be published alongside the expression data as a simple csv file or should directly be included within the `scanpy`- or `Seurat`-objects. The provision of the annotation data serves two important issues: (1) spare a lot of time when the original researcher or any other has to re-use the annotation and (2) ensures the same starting point for the follow-up analysis. Similar efforts were undertaken in the generation of so-called *cell type atlases* such as the single-cell Expression Atlas [Papatheodorou *et al.* 2020]. These databases seek to provide a collection of high-quality data sets that can be used to learn cell type annotation patterns. Alongside, it should become conventional to provide the bioinformatic analysis scripts on a public repository. While many publications already provide all scripts underlying the generated results in Github or Gitlab, some others do not provide them at all or only upon request. More common rules and requirements would drastically increase the reproducibility of the results.

In summary, due to the recency of single-cell transcriptomics, missing standard workflows raise major concerns in reproducibility. Though non-standard workflows could theoretically be backed up with script provision and building a respective container, it is rarely encountered. Comprehensive benchmarking of the implemented tools will reduce these concerns to strengthen the inferred results. Furthermore, expanding Minimum Information About Sequencing or Microarray Experiments (MINSEQE/ MIAME)* for data upload and code sharing will increase the studies' transparency and allow for effortless follow-up analysis by the single-cell community.

Outlook and perspectives. In this dissertation, the defined influence of data sparsity and imputation with respect to the inference of cell type-specific networks from single-cell transcriptomics data was stated, as well as providing an in-depth characterisation of a human retinal organoid model system.

Apart from the technical aspects, like reproducibility and data provision, which were already discussed before, the mindful biological interpretation represents another central challenge in single-cell data analysis. As for the `DCA`-imputed networks, a more in-depth analysis of the gene correlation networks is required. Besides the very variable cell type-specific networks derived from EIF4EBP1, also other more consensus networks should be characterised, for example, based on common photoreceptor regulators or housekeeping genes. Following this careful evaluation, other gene networks of interest should be investigated to decipher more subtle differences between rods and cones.

The same careful evaluation was applied to the HRO system. An initial analysis, given the

---

[*] `https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html`

possible analysis tools already provided high-resolution insights into the cell type quantities and developmental trajectories of the organoid system. Based on these trajectories, it was possible to approach developmental processes within the organoid model, highlighting overlaps to *in vivo* organ development. Likewise, differences were also identified, which may aid in improving the experimental conditions and differentiation protocols when culturing the organoids to resemble the human eye fully. Then, these organoids can be used to derive useful insights into the progression of AMD or other similar macular diseases.

Chapter 7

Supporting material

7.1 Data and Code availability

All R-markdown files, jupyter notebooks, Helper-functions and conda-environments used in this dissertation are deposited on Github (`https://github.com/lisbeth-dot-95/Dissertation`). In this repository, the scripts were ordered according to the structure presented in this dissertation. The gold data, as well as corresponding dropout data sets, are included in the R-package `scorrgoldnet` on Github (`https://github.com/lisbeth-dot-95/scorrgoldnet`)

All single-cell-objects and annotation results were uploaded to zenodo*. Both `Seurat`-objects used in Chapter 4, can be found under this doi:`https://doi.org/10.5281/zenodo.5519574`. The use of publicly available data sets was indicated in the respective workflow section of the chapters. Since both HRO data sets have not been published, their access is embargoed but will be released upon publishing. The preprocessed and annotated results can be found here:`https://doi.org/10.5281/zenodo.5519551`. The RNA velocity analysis results were deposited under the following doi:`https://doi.org/10.5281/zenodo.5525816`.

7.2 Chapter 3

To evaluate the effect of different levels of dropout on gene network inference, a benchmarking framework was established spanning six levels of dropout and seven imputation methods. In this section, supplementary information corresponding to Chapter three was included.

While analysing the reference data set, comprising the gold and six dropped-out data sets, the distribution of expression values was investigated. Figure S-1 depicts the full distribution pattern of logged expression values.

After proofing the suitability of the synthetic reference data set, seven imputation tools were applied to each dropout level and the effect on the correlation network preservation was analysed. Using the module preservation statistics, established by [Langfelder *et al.* 2011], two different thresholds of preservation were used. Values exceeding the threshold of ten were considered to be strongly preserved, whereby values below two represent non-preserved correlated gene groups. In between both thresholds, modules were considered moderately preserved. Here, Figure S-2 highlights the distribution of module preservation scores over different dropout levels and imputation tools. A compressed version using a log2-FC was used in the main part (see Figure 3.6).

Apart from the composite measure of the module preservation, also the distributions of the

---

* `https://zenodo.org/`

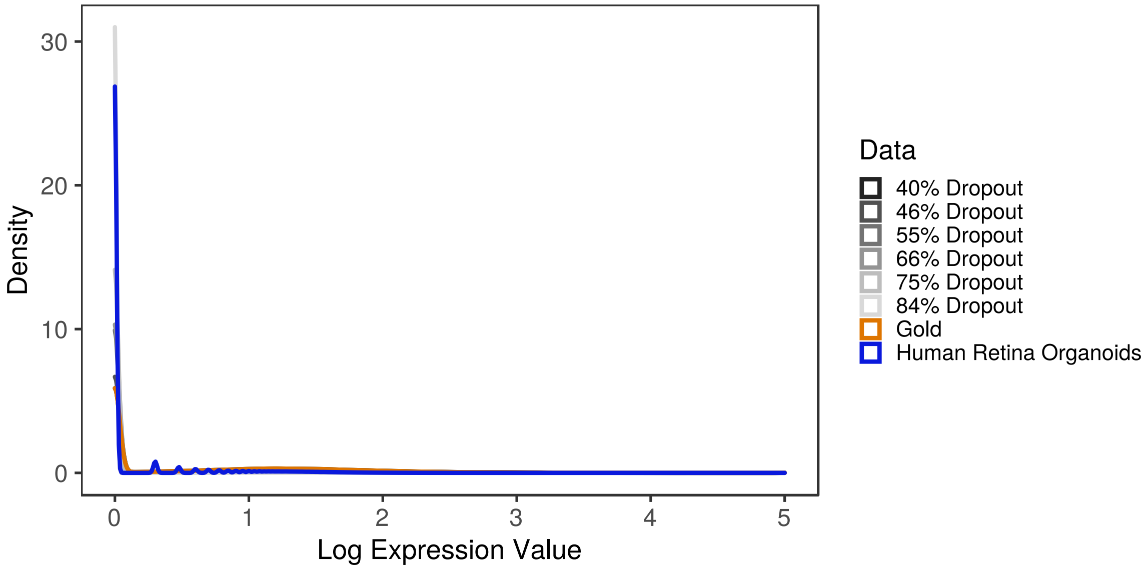


Figure S-1. *Distribution of logged expression values of all reference data sets and a human retina data set.*

The density of expression values across eight data sets is shown to contrast the impact of different dropout rates. The gold data is plotted in orange, all six dropout reference data sets are shown with a grey gradient and a biological data set is plotted in blue.

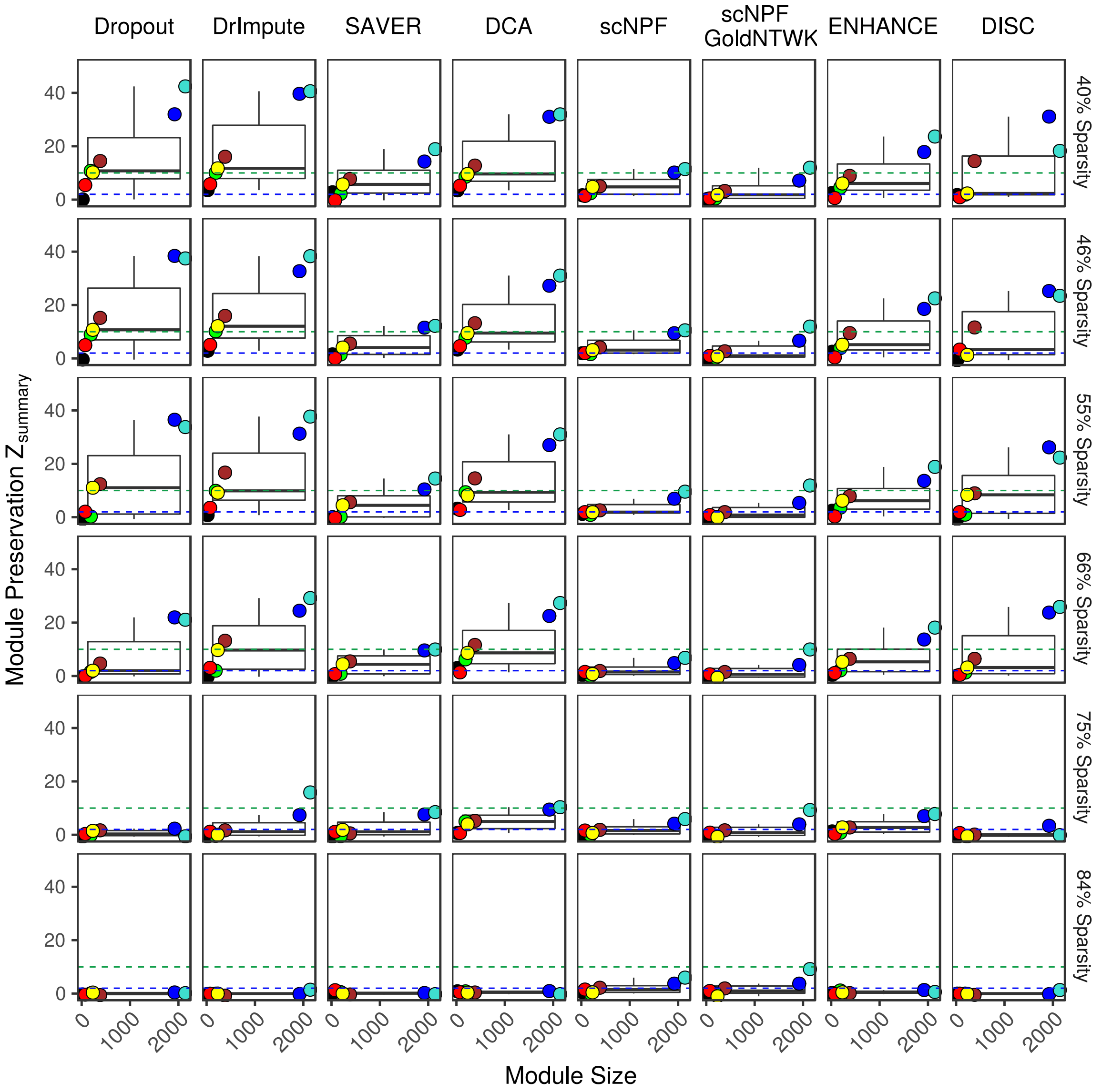


Figure S-2. *Box plots showing the behavior of module preservation across different dropout levels.*

The $Z_{summary}$ measure implemented in `WGCNA` *is a composite, permutation-based metric of various network density and connectivity measures. The blue and the green line indicate the threshold towards moderate and strong module preservation, respectively. Colouration of the dots corresponds to the individual modules. Dropout refers to the amount of artificially introduced non-true zeros in each of the reference data sets.*

topological overlap measured (TOM) values before and after imputation were analysed since other publications indicated a systematic increase in correlation values Andrews & Hemberg

[2019]. A summary is provided in Figure S-3.

While initially focussing on the effect of imputations on the gene-gene correlation, a human retina organoid data set by [Kim *et al.* 2019] was applied to investigate the effect on cell-cell correlations. Here, more specifically the *annotatability* using known marker genes before and after imputation was highlighted. Therefore, an automatic annotation pipeline was established which was described in section 3.1. The individual dot plots used for cell cluster annotations are depicted in Figure S-4.

Next to the percentages of annotated retinal cell types, here rods, cones, and Müller glia cells, also unambiguous clusters (mixed) and un-annotatable clusters were quantified. Previous work mentioned that imputations helped to reobtain original cell clusters [Eraslan *et al.* 2019]. Therefore, also the number of cell clusters after the Louvain community detection and the mean silhouette coefficient was included. An overview of the results is indicated in Table S-1.

Table S-1. *Results of retina annotation pipeline before and after imputation. The table shows the percentages of retina-specific cell types as well as not assignable and mixed clusters.*

	MSC	Clusters	Cones	Rods	MG cells	Not assignable	Mixed annotation
sparse	-0,01	8	26,03	21,60	24,34	28,04	0,00
`DrImpute`	-0,18	11	24,29	21,10	18,05	6,55	0,00
`SAVER`	-0,03	8	26,08	18,05	19,76	36,11	0,00
`DCA`	0,14	8	28,75	20,28	29,12	21,84	0,00
`scNPF`	-0,03	9	20,43	21,77	23,25	34,55	0,00
`scNPF`String	-0,02	8	20,65	21,03	18,50	39,82	0,00
`ENHANCE`	0,02	21	44,28	17,09	15,82	19,32	3,49

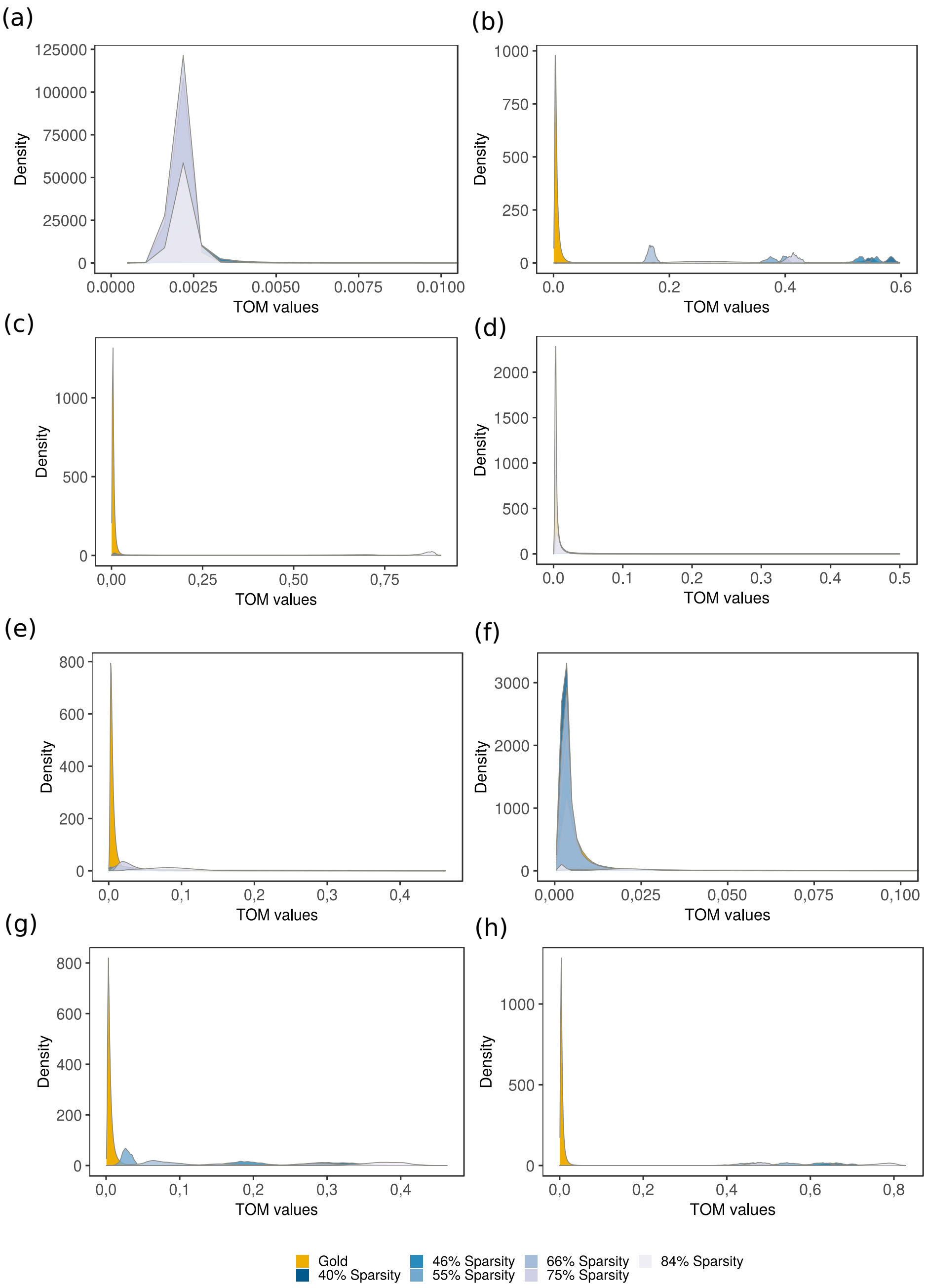
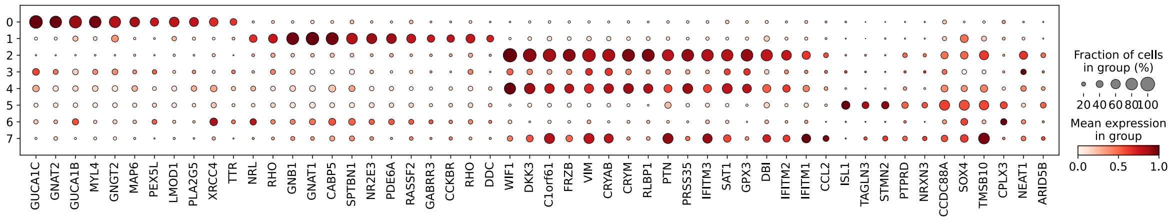
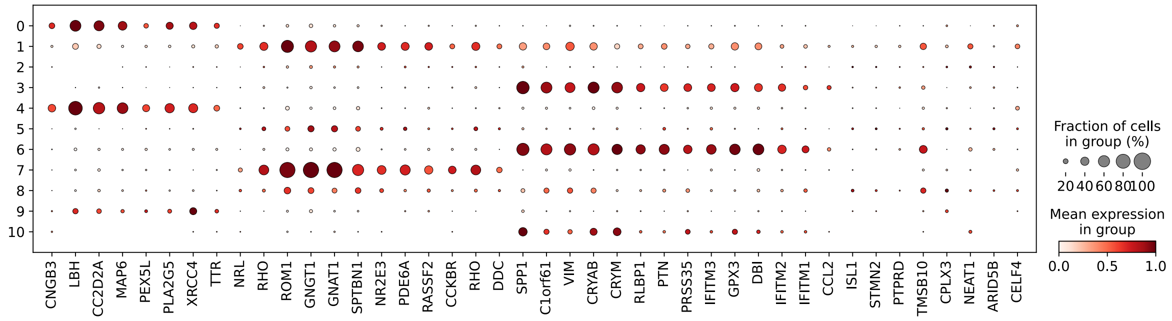


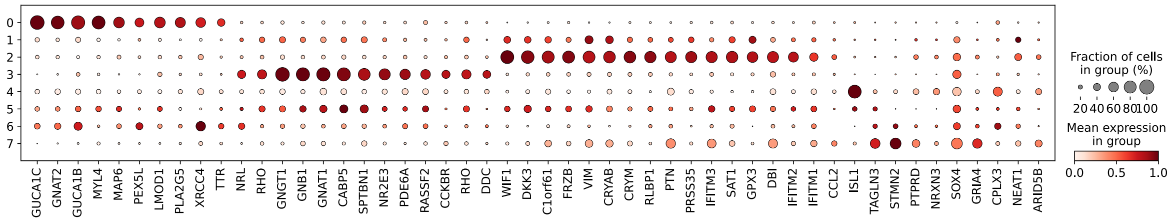
Figure S-3. *Density of TOM values before and after imputation of the synthetic data set.* After transforming expression data before and after imputation to TOM values via the gold data β value, positive edges were detected and compared during the edge recovery analysis. Colours correspond to the original information content of data (lightest blue - less information, highest dropout level). (a) Dropout data, (b) `DrImpute`, (c) `SAVER`, (d) `ENHANCE`, (e) `DCA`, (f) `DISC`, (g) `scNPF`, (h) `scNPF Gold`. Whereby some tools, such as `DrImpute` (b) and both `scNPF` approaches (g+h) produce high TOM values compared to the gold data, `DCA` (e) and `ENHANCE` (d) TOM densities were close to gold.



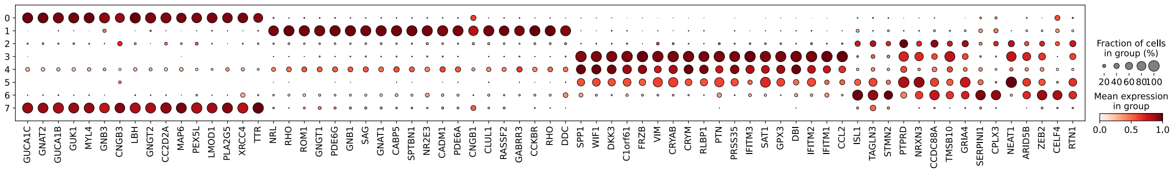
(a) *Sparse human retina organoids*



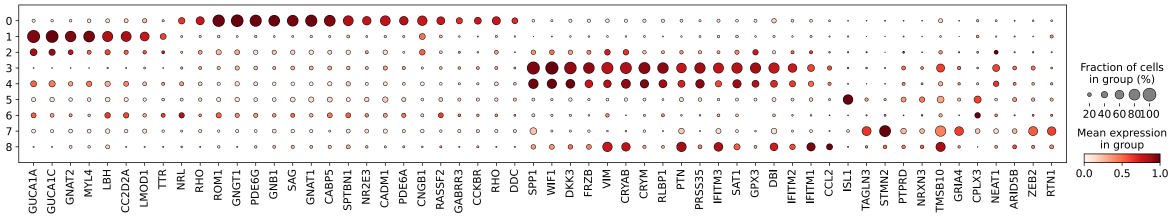
(b) `DrImputed` *imputed human retina organoids*



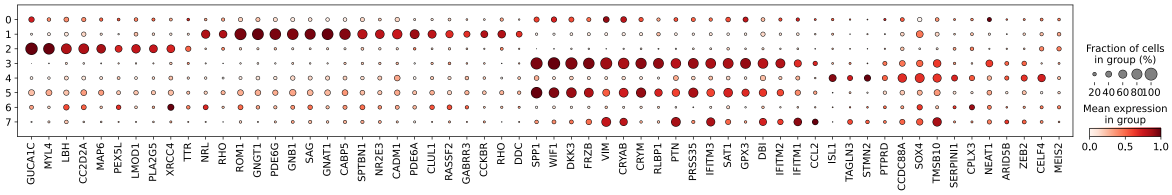
(c) `SAVER` *imputed human retina organoids*



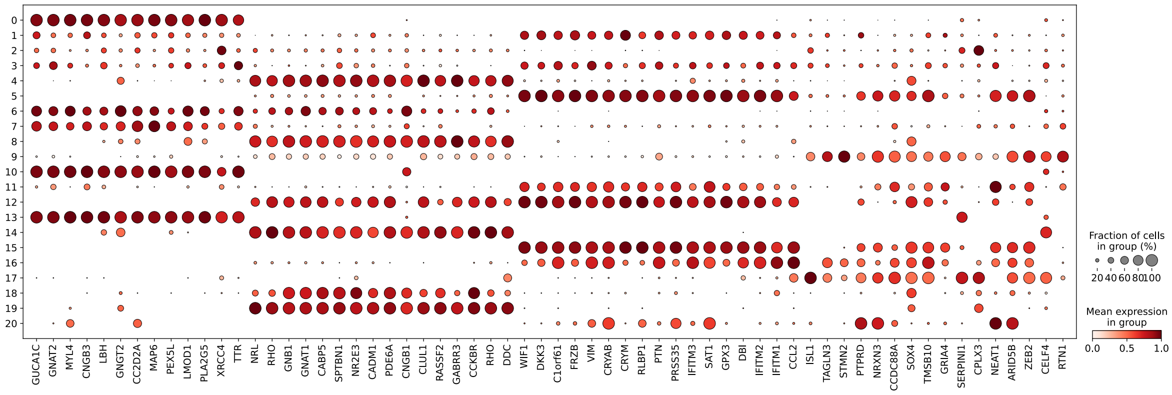
(d) *DCA imputed human retina organoids*



(e) *scNPF imputed human retina organoids*



(f) *scNPF String imputed human retina organoids*

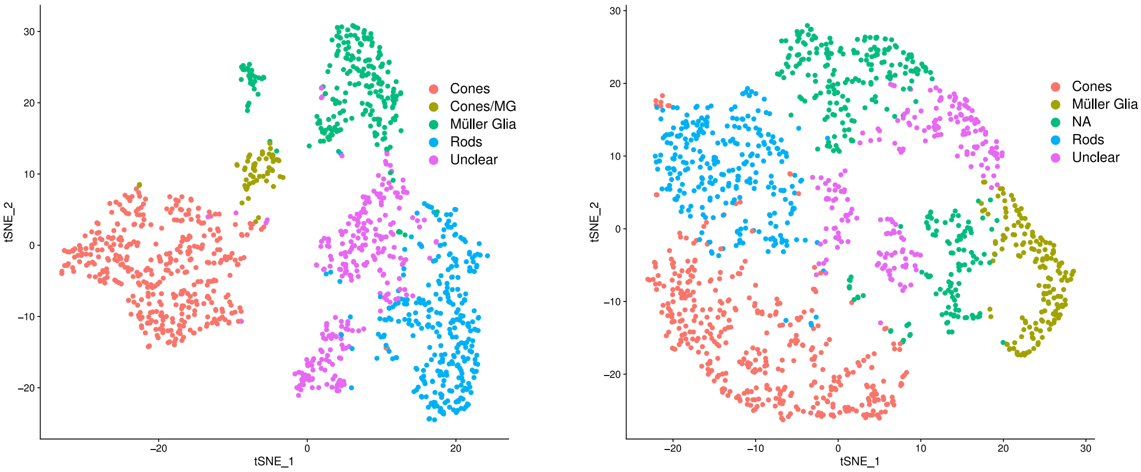


(g) *ENHANCE imputed human retina organoids*

Figure S-4. *Dot plot of marker gene expression in human retina organoids.*
All dot plots show the expression values of marker genes across Louvain clusters using the `scanpy` workflow before and after imputation. Colouring corresponds to the mean gene expression and the dot size to the percentage of cells per cluster expressing the respective gene. Expression values were scaled per Louvain cluster.

7.3 Chapter 4

Based on the results of Chapter three, it was evaluated if indeed cell type-specific gene correlation networks were inferable after `DCA`-imputation of an experimentally derived data set. Therefore, the human retina organoid data set by Kim *et al.* [2019] was used. Via the provided set of marker genes, cell clusters were annotated to either rod, cone, or Müller glia cell. An overview of the annotation results in the sparse and imputed data is provided in Figure S-5. Additionally, both `Seurat`-objects were made publicly available (see Section 7.1).

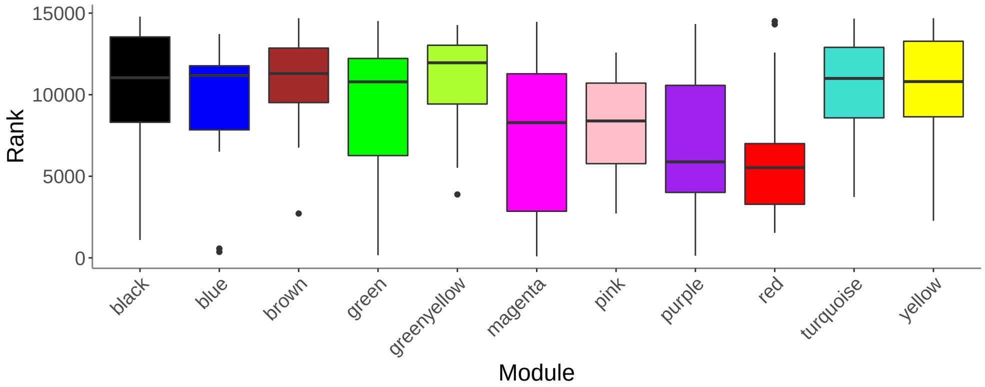


(a) *Unimputed data.* **(b)** *`DCA`-imputed data.*

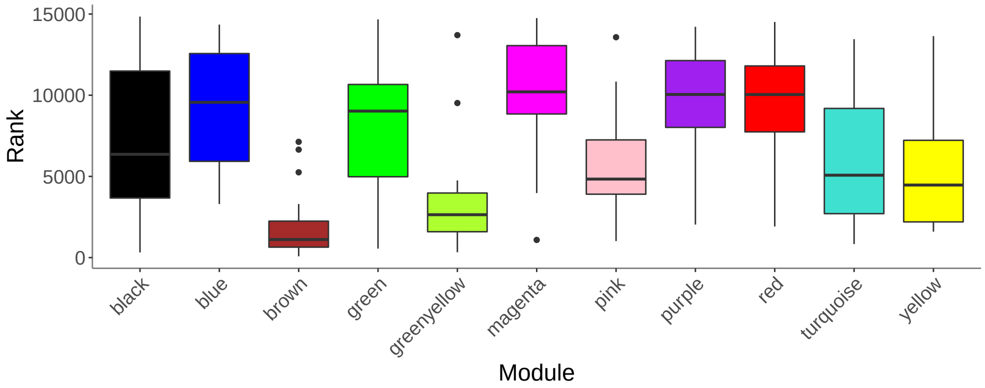
Figure S-5. *Annotation result of human retina data set.*

The raw and `DCA`-imputed data was preprocessed and visualized via the `Seurat`-pipeline. Using the set of provided marker genes, the expression pattern across Louvain clusters was analysed. Unclear clustering results correspond to the set of marker genes that were not specific for cones, rods and Müller Glia cells. If no expression signal was detected, cells were assigned to the NA cluster.

After stating that indeed scale-free gene correlation networks were inferred after `DCA`-imputation, the biological information of the modules was analysed. Hypothesising that the GOIs fulfil a central role within the respective cell type-specific networks, they should own a hub gene-like status. Clearly, one module in the rod- and cone-specific network revealed a low rank, corresponding to a high absolute MM-value. Assuming that these network configurations are cell type-specific, they should not be detected in the respective other network. In the main part of the thesis, the distribution of MM-values of the top20 hub genes from these modules was investigated. Figure S-6 shows an alternative representation, using the same ranking procedure used for Figure 4.4 These trends were in concordance with the insights gained in the main part.



(a) *Rod-module hub genes in cone network*



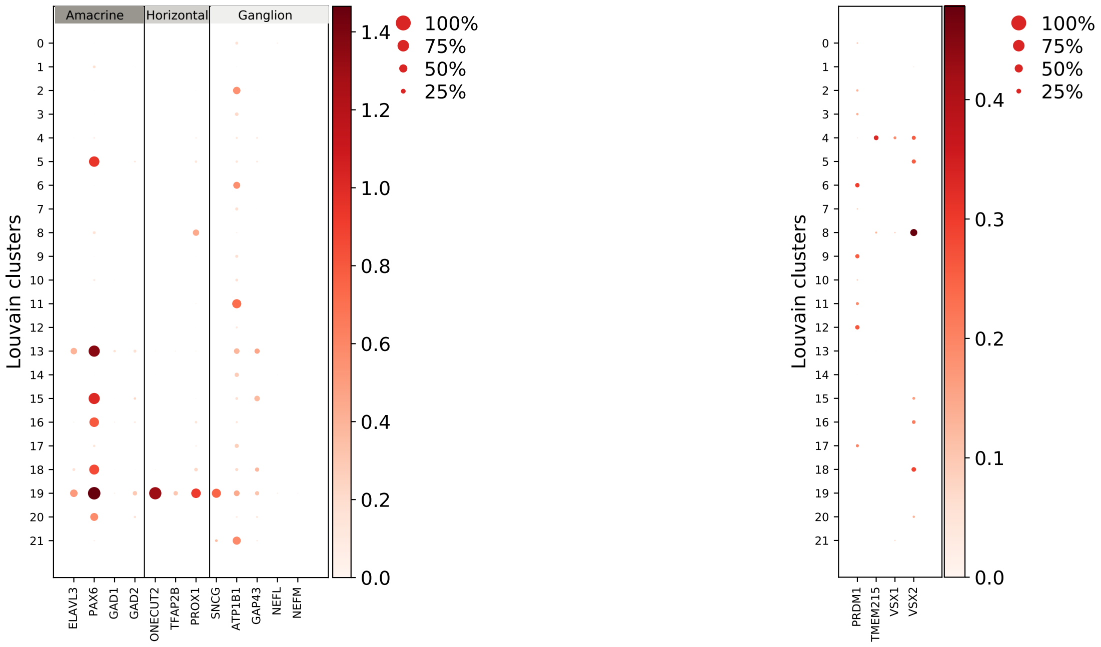
(b) *Cone-module hub genes in rod network*

Figure S-6. *Distribution of cone- and rod-specific module membership ranks in complementary network.*
After identifying the cone- and rod-specific modules in the respective network, the uniqueness of the hub genes in these modules is analysed. Therefore, the top 20 MM-genes per module are extracted and the distribution of their MM-ranks is compared across all other modules in the opposite network. Both, the MM values of the rod-greenyellow hub genes in the cone network (a), as well as the cone-purple hub genes in the rod network (b) show no module with high MM-values but low variance.

7.4 Chapter 5

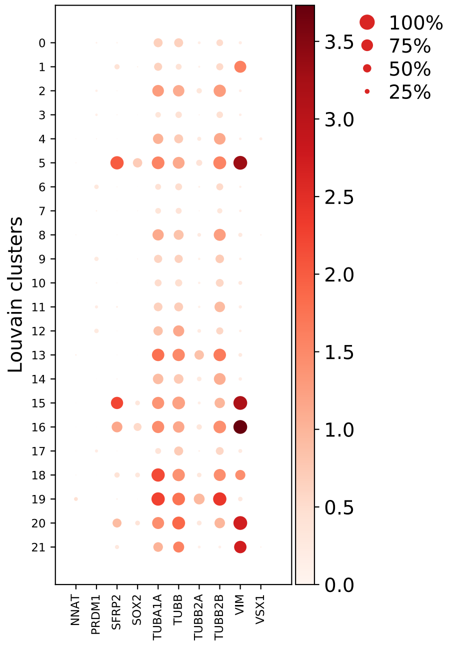
Age-related macular degeneration represents a major cause of blindness in developed countries and to date, neither the defined molecular pathway of the disease is known nor an efficient treatment option at hand. Using reprogrammed stem cells, organ-like structures, called organoids can be generated which own a sufficient cellular complexity and avoid animal testing. Here, a novel human retinal organoid system, named HRO was developed, which should ultimately allow to illuminate the defined disease progressions of AMD. Using the toolbox of single-cell transcriptomics, two untreated HRO controls were characterized. In a first step, the individual cells were annotated to retinal cell types. Here, a manual, as well as a machine learning pipeline, was applied. Figure S-7 and S-8 summarize the additional dot plots which were used in the manual approach to annotate Louvain clusters, detected in the low dimensional data.

Additionally, a transfer-learning tool called `CaSTLe` was used, which learns cell type characteristics from a reference data set to annotate unseen data. Three different reference data sets by [Cowan *et al.* 2020] were used spanning a developed organoid and two adult tissue samples. Regarding both adult tissue samples, regions of the fovea and periphery were extracted. A defined overview of the full Cowan-organoid data set is provided in Supplementary Table S-2. Likewise, the cell type counts regarding both adult samples is depicted in Supplementary Table S-3. For the main part, these high-resolution data sets were compressed to increase the classification performance. Therefore, the cell subgroup resolution was sacrificed, while larger superior groups were gathered. As an example, all cells belonging to the subgroup AC_B_0X were compressed to AC_B. Moreover, were all annotation artefacts (5-, 37-, and 38-) removed before training `CaSTLe`. Regarding the adult data sets shown in Supplementary Table S-3, all immune cells (MAST, MO, and NK) were also withdrawn from the training data. The final compressed overview is depicted in Table 5.1.



(a) *AHG cells HRO3*

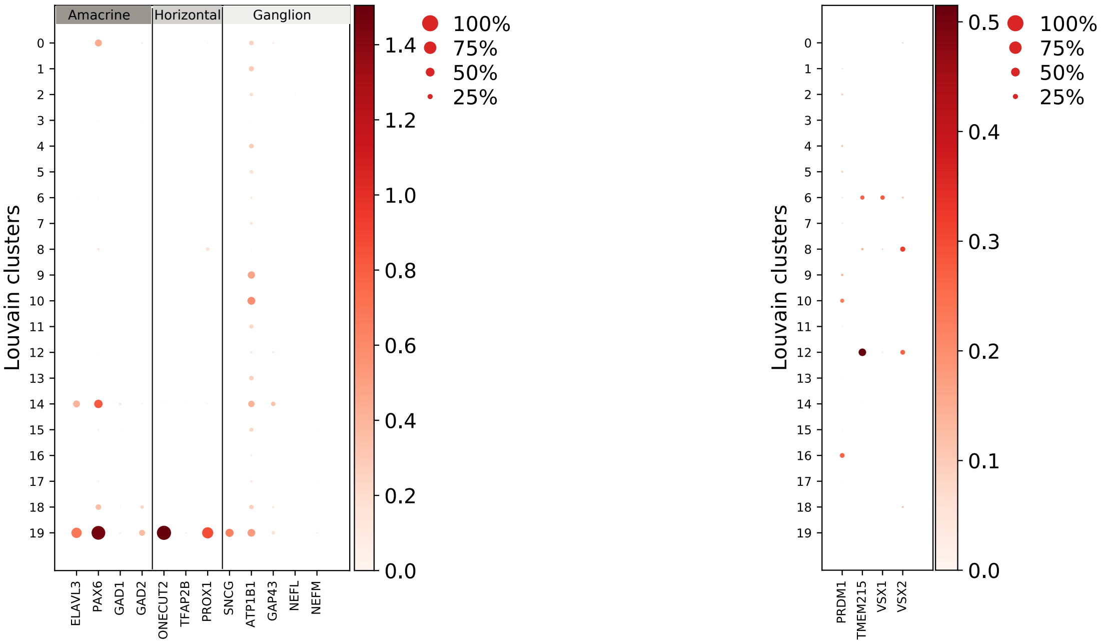
(b) *Bipolar cells HRO3*



(c) *Photoreceptor progenitor cells HRO3*

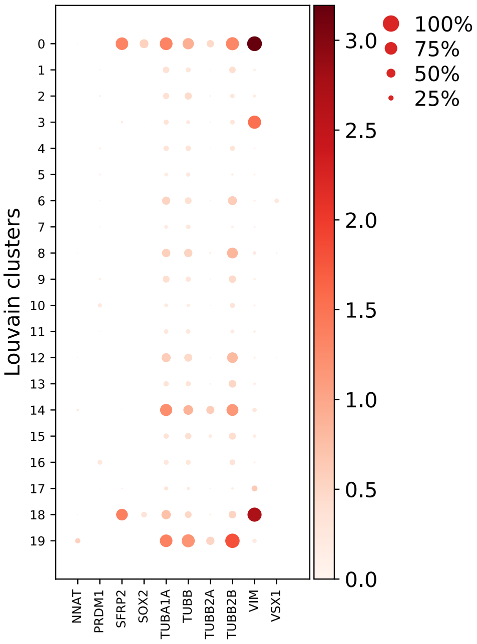
Figure S-7. *Dot plot of marker gene expression in HRO2.*

Expression of selected Amacrine-Horizontal-Ganglion (AHG) (a), Bipolar (b), and photoreceptor progenitor cell-associated genes (c) across 21 different Louvain clusters. The mean expression is indicated in the heatmap legend. The dot size corresponds to the percentage of cells expressing the gene inside the Louvain cluster.



(a) *AHG cells HRO3*

(b) *Bipolar cells HRO3*



(c) *Photoreceptor progenitor cells HRO3*

Figure S-8. *Dot plot of marker gene expression in HRO3.*

Expression of selected Amacrine-Horizontal-Ganglion (AHG) (a), Bipolar (b), and photoreceptor progenitor cell-associated genes (c) across 20 different Louvain clusters. The mean expression is indicated in the heatmap legend. The dot size corresponds to the percentage of cells expressing the gene inside the Louvain cluster.

Table S-2. *Overview of cell types and amounts of the Cowan* et al. *[2020] organoid data set. Legend: AC - Amacrines, Ast - Astrocytes, CdBC/ChBC - Bipolars, CM - Choroidal melanocyte, END - Endothelial cells, FB - Fibroblasts, GC - Ganglions, HC - Horizontals, MC - Müller Glia, PER - Pericytes, RBC - Rod bipolar cell, RPE - Retinal pigment epithelium, uG - Mircoglia*

Compressed		Original	
Cell type	Count	Cell type	Count
37-	78	37-	78
38-	58	38-	58
5-	216	5-	216
AC_B	1230	AC_B_01	141
		AC_B_02	29
		AC_B_04	88
		AC_B_05	59
		AC_B_06	167
		AC_B_07	17
		AC_B_08	32
		AC_B_09	13
		AC_B_10	142
		AC_B_11	11
		AC_B_12	253
		AC_B_13	33
		AC_B_15	51
		AC_B_16	20
		AC_B_17	107
		AC_B_18	67
AC_Y	81	AC_Y_01	58
		AC_Y_03	23
Ast	55	Ast	55
CdBC	1700	CdBC_01	283
		CdBC_02	1235
		CdBC_03	123
		CdBC_04	28
		CdBC_05	31
ChBC	658	ChBC_01	235
		ChBC_02	66
		ChBC_03	204
		ChBC_04	153
HC_02	1762	HC_02	1762
cones	12973	L/M cone	12892
		S cone	81
MC	10542	MC_01	10341
		MC_02	112
		MC_03	89
RBC	461	RBC	461
rod	13913	rod	13913
RPE	130	RPE	130

Table S-3. *Overview of cell types and amounts of the Cowan* et al. *[2020] adult data sets. Legend: AC - Amacrines, Ast - Astrocytes, CdBC/ChBC - Bipolars, CM - Choroidal melanocyte, END - Endothelial cells, FB - Fibroblasts, GC - Ganglions, HC - Horizontals, MC - Müller Glia, PER - Pericytes, RBC - Rod bipolar cell, RPE - Retinal pigment epithelium, uG - Mircoglia. Continued on next page.*

Foveal				Peripheral			
Compressed		Original		Compressed		Original	
Cell type	Count	Cell type	Count	Cell type	Count	Cell type	Count
AC_B	192	AC_B_01	29	AC_B	417	AC_B_01	30
		AC_B_08	39			AC_B_08	40
		AC_B_10	16			AC_B_10	17
		AC_B_11	43			AC_B_11	12
		AC_B_15	16			AC_B_15	32
		AC_B_16	17			AC_B_16	33
		AC_B_18	32			AC_B_18	99
AC_Y	123			AC_Y	330	AC_Y_01	130
						AC_Y_02	44
		AC_Y_03	123			AC_Y_03	156
Ast	149	Ast	149	Ast	172	Ast	172
CdBC	2058	CdBC_01	93	CdBC	2418	CdBC_01	416
		CdBC_02	364			CdBC_02	524
		CdBC_03	1247			CdBC_03	1070
		CdBC_04	354			CdBC_04	288
						CdBC_05	120
ChBC	398	ChBC_02	50	ChBC	1182	ChBC_02	440
		ChBC_03	14			ChBC_03	75
		ChBC_04	334			ChBC_04	564
						ChBC_01	103
END	368	END_01	204	END	208	END_01	102
		END_02	82			END_02	52
		END_03	82			END_03	54
FB_02	252	FB_02	252	FB_02	579	FB_02	579
GC	6086	GC_01	4338	GC	35	GC_01	23
		GC_04	232			GC_04	12
HC	1037	HC_01	144	HC	844	HC_01	261
		HC_02	893			HC_02	583
cone	1375	L/M cone	1339	cone	1202	L/M cone	1149
		S cone	36			S cone	53
MC	3886	MC_01	3886	MC	8207	MO_01	6536
						MO_02	1491
						MO_03	180
MO	137	MO_01	32	MO	300	MO_01	220
		MO_02	40			MO_02	57
		MO_03	65			MO_03	23
NK	130	NK	130	NK	44	NK	44

Foveal				Peripheral			
Compressed		Original		Compressed		Original	
PER	75	PER	75	PER	85	PER	85
RBC	1016	RBC	1016	RBC	4191	RBC	4191
rod	1894	rod	1894	rod	13029	rod	13029
RPE	84	RPE	84	RPE	186	RPE	186
TCell	299	TCell	299	TCell	160	TCell	160
				CM	157	CM	157
uG	172	uG	172	uG	171	uG	171

After identifying retinal cell types using these two pipelines, the `CaSTLe`-annotation was used to correlate HRO-photoreceptors and Müller glia cells to adult foveal and peripheral reference samples. In the course of this dissertation, three reference data sets were used including primate and human data. As both human reference data sets by [Cowan *et al.* 2020] and [Voigt *et al.* 2019] contained the information of three different donors, the distributions of the correlation values were illustrated individually. The results of the Voigt *et al.* data are depicted in Figure S-9 and Figure S-10. Similarly, all Cowan *et al.* related data results are shown in Figure S-11 and Figure S-12.

Though a temporal aspect of the data is lost while tissue preparation, RNA velocity allows approaching a so-called pseudotime. This information may help to order cells relative to this measure. Here, RNA velocity was used to compare the developmental status of both untreated HRO systems. In the main part of this dissertation, the `CaSTLe`-annotation was used to analyse the developmental trajectory across cell types. Additionally, Figure S-13 summarizes the RNA velocity streams using the manual annotation result.

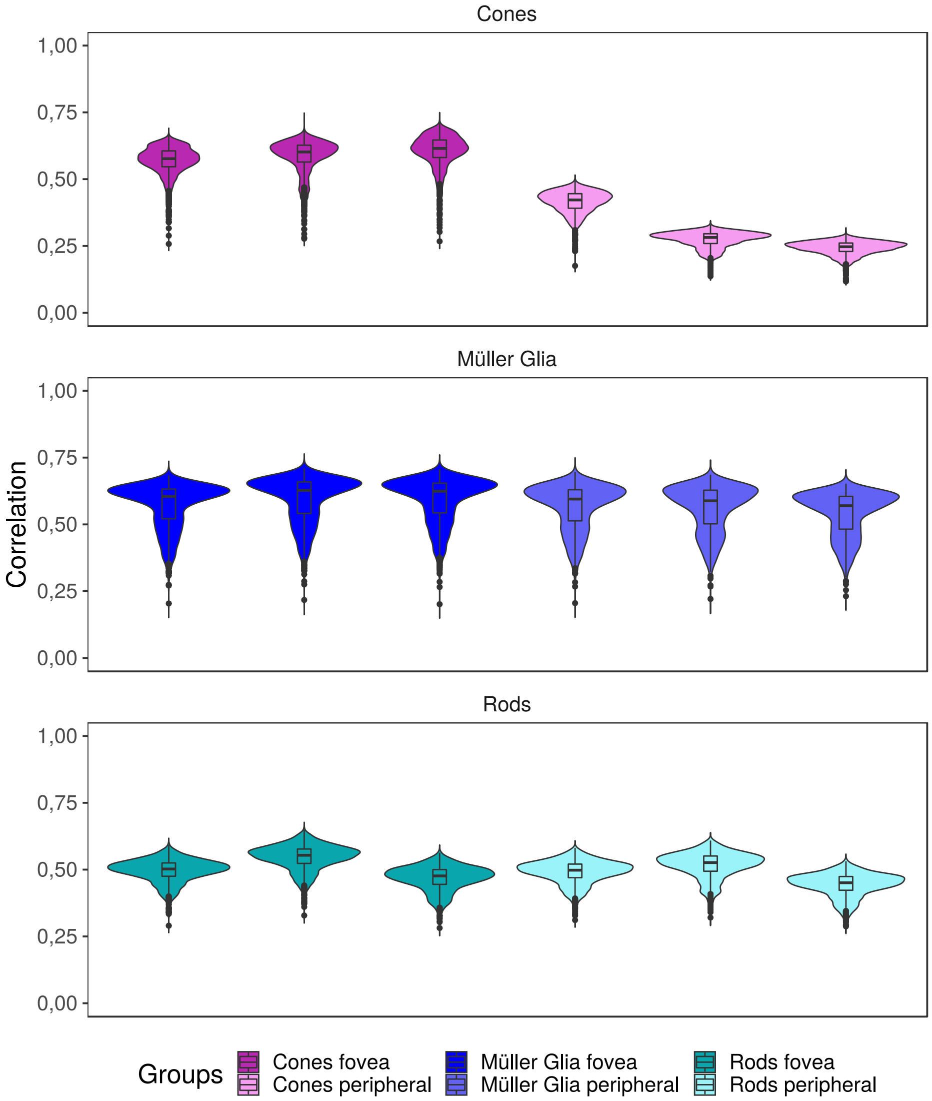


Figure S-9. *Distribution of Pearson correlation of HRO-2 cells and human tissue reference data taken from Voigt* et al. *[2019] per donor.*

Cells are annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells.*

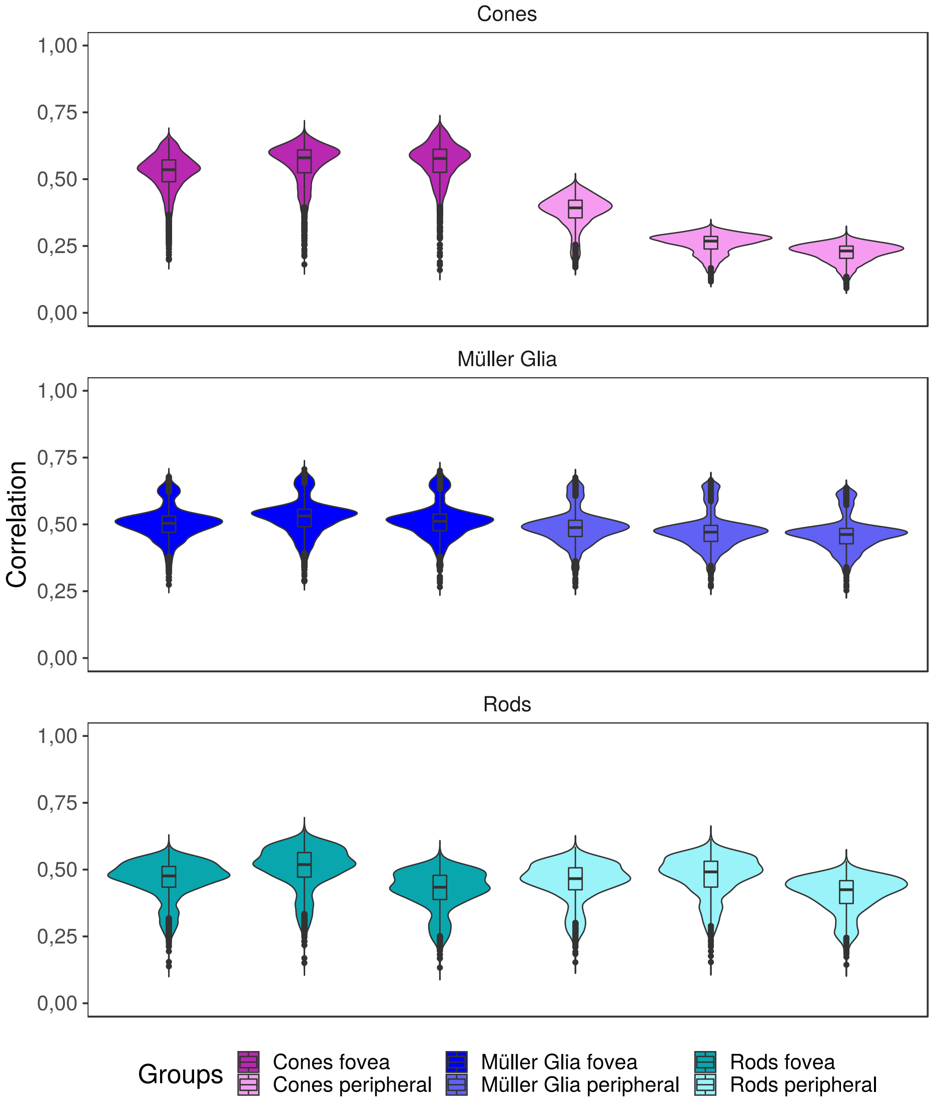


Figure S-10. *Distribution of Pearson correlation of HRO-3 cells and human tissue reference data taken from Voigt* et al. *[2019] per donor.*

Cells are annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells.*

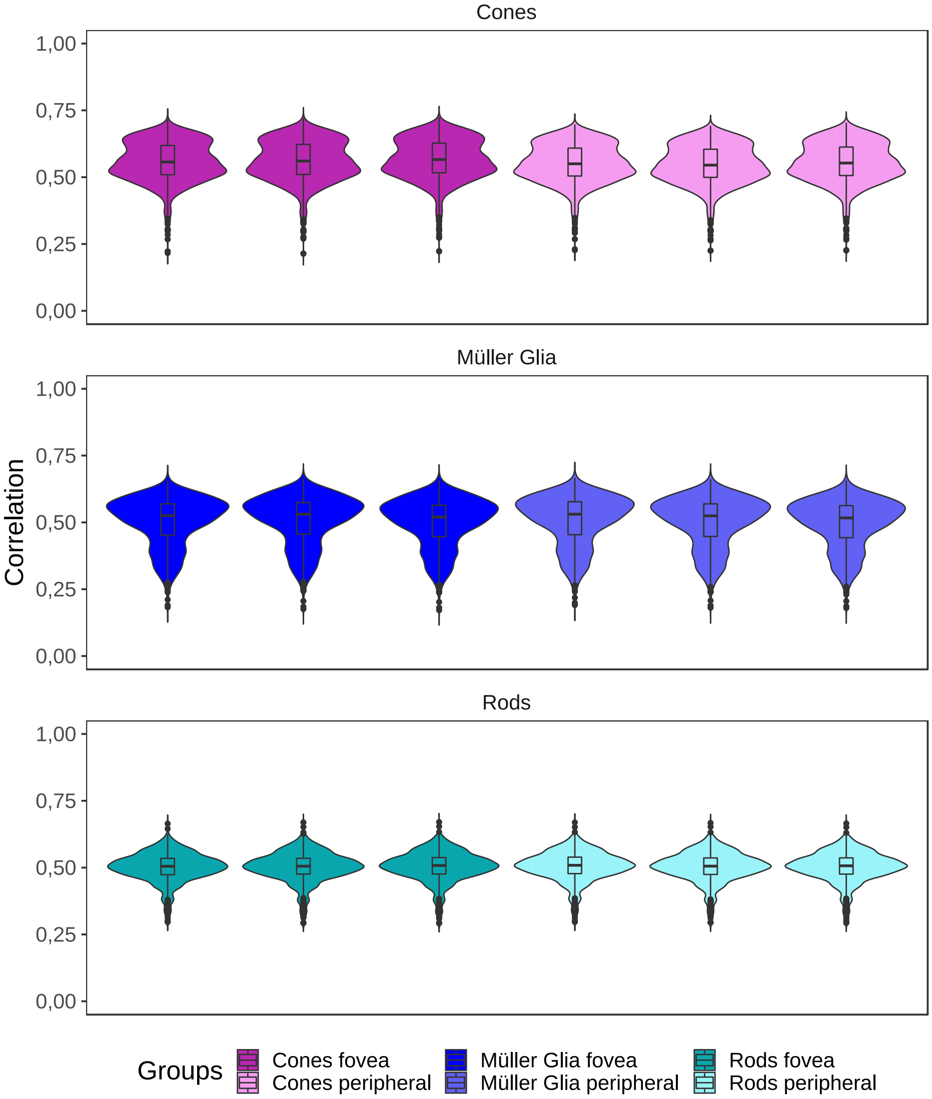


Figure S-11. *Distribution of Pearson correlation of HRO-2 cells and human tissue reference data taken from Cowan* et al. *[2020] per donor.*

Cells are annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells.*

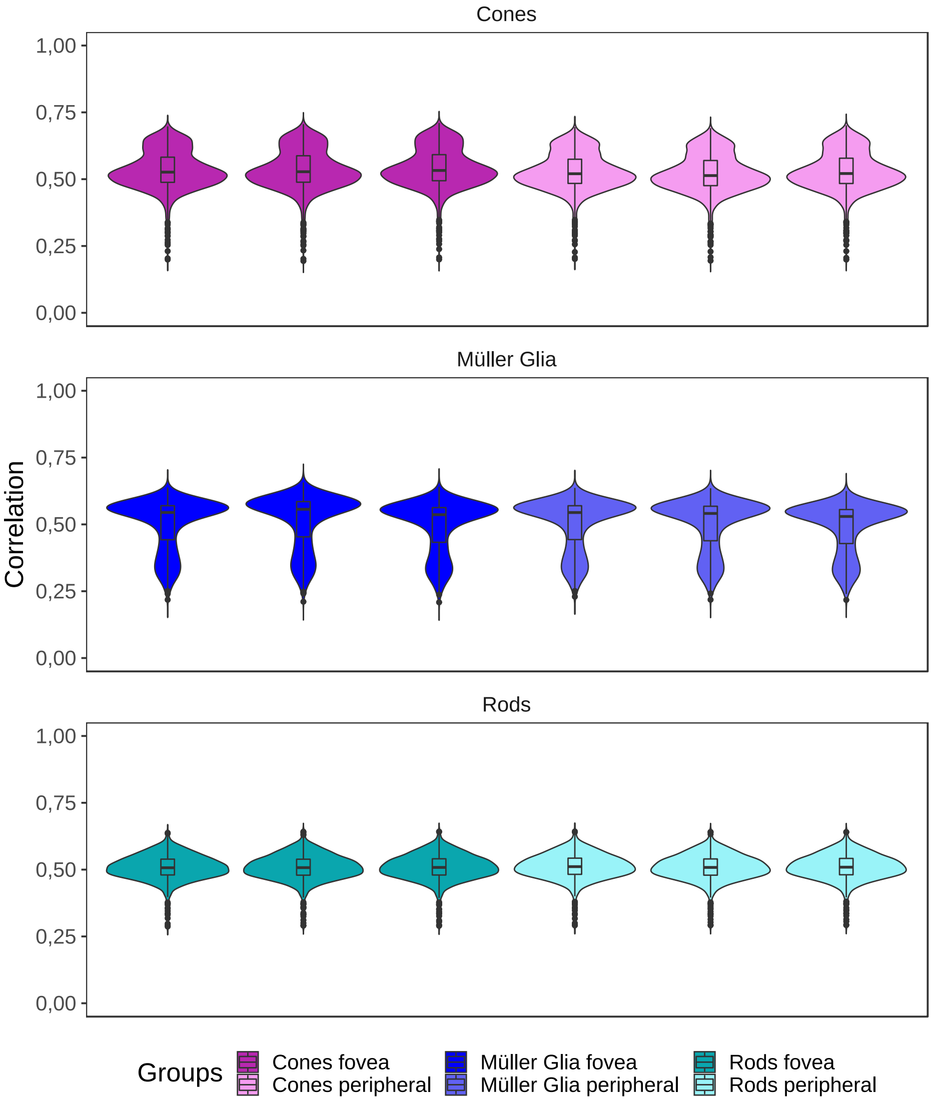
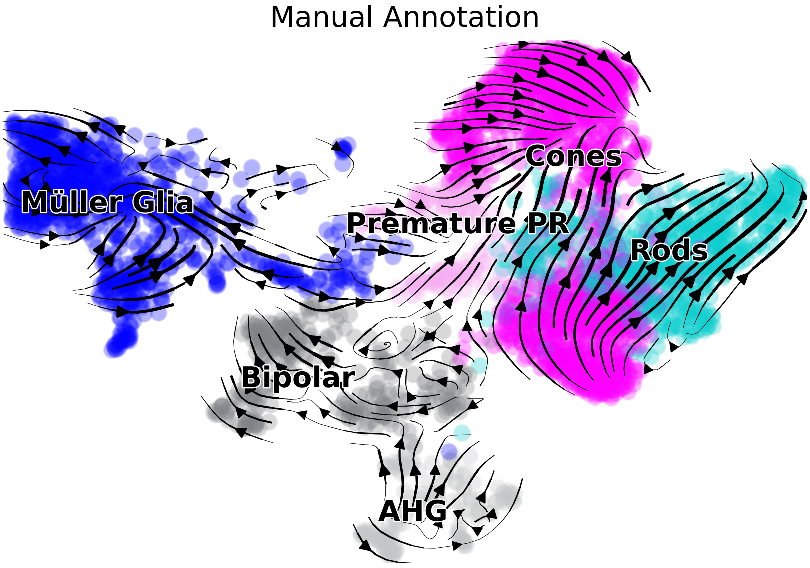
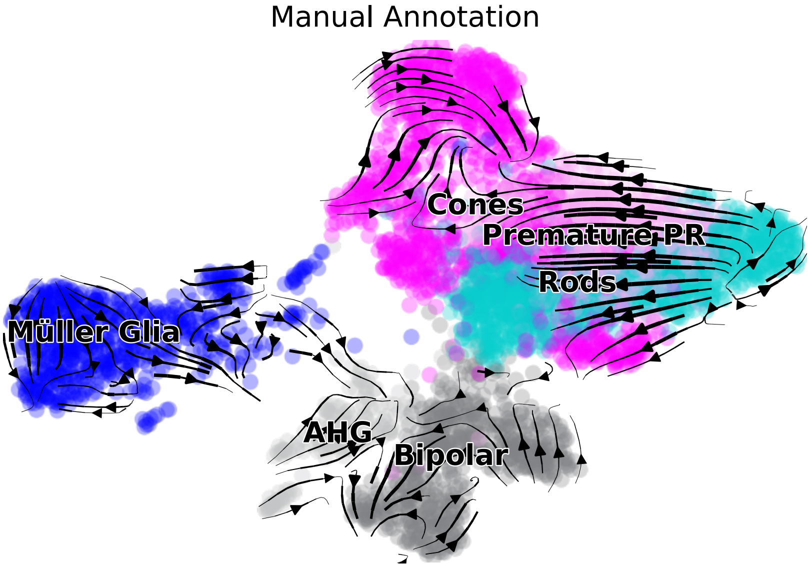


Figure S-12. *Distribution of Pearson correlation of HRO-3 cells and human tissue reference data taken from Cowan* et al. *[2020] per donor.*

Cells are annotated via the `CaSTLe`*-annotation pipeline The violin plot shows the Pearson correlation of organoid cones, rods and Müller Glia cells against the human reference vectors of foveal and peripheral cells.*



(a) *RNA velocity streams manual annotation of HRO-2*



(b) *RNA velocity streams manual annotation of HRO-3*

Figure S-13. *RNA velocity streams of manual annotation results.*

Using the calculated RNA velocity streams, the development of the premature photoreceptor cell type can be traced. Similar to the results of the `CaSTLe`*-annotation, differences across both HRO-samples were detected.*

List of Figures

1.1 Central Dogma of Molecular Biology, taken from Fu *et al.* [2014]. 3

1.2 Overview of single-cell preparation methods 7

1.3 Overview of individual tasks during gene network inference 12

1.4 Differences in resolution from bulk and single-cell RNA-seq data on the level of gene networks. 13

1.5 Composition of the human retina 14

1.6 Differences of a healthy eye and different AMD subtypes 16

3.1 Overcoming sparsity via imputation approaches to allow for network inference in single-cell transcriptomics data. 30

3.2 Workflow to investigate the impact of imputation of gene regulatory networks (GRN) in single-cell transcriptomic data. 31

3.3 Characteristics of the synthetic gold data. 39

3.4 Heatmaps of selected genes in the gold and three sparse data sets. 40

3.5 Reference data characteristics. 41

3.6 Z_{summary} before and after imputation of reference data sets. 44

3.7 Edge recovery trend before and after imputation. 45

3.8 Comparison of `Seurat` and `scanpy` preprocessing results. 48

3.9 Louvain clusters detected in human retina organoid data before and after imputation. 49

3.10 Effect of imputation on cell cluster annotation in human retina organoid data. 50

4.1 Evaluating the influence of `DCA` on inferring gene correlation networks from a human retina data set. 56
4.2 Scale-free topology estimation and gene cluster dendrograms for weighted gene correlation network analysis in the sparse organoid retina data. 61
4.3 Scale-free topology estimation and gene cluster dendrograms for weighted gene correlation network analysis in `DCA`-imputed organoid retina data. 62
4.4 Distribution of GOI ranks in the `DCA`-imputed retina sub-cell type networks. . 65
4.5 Distribution of cone- and rod-specific module membership values in complementary network. 66
4.6 Maximal module membership values of hub genes in the complementary networks. 67
4.7 Rod-specific subnetworks of EIF4EBP1. 70
4.8 Cone-specific subnetworks of EIF4EBP1. 71

5.1 Analysing and characterising the HRO systems. 81
5.2 UMAP embedding of HRO sample 2 and expression pattern of selected marker genes. 88
5.3 UMAP embedding of HRO sample 3 and expression pattern of selected marker genes. 89
5.4 HRO cell-type compositions. 91
5.5 Evaluation of `CaSTLe`-classification model for both HRO control samples across all reference data set. 94
5.6 Evaluation of `CaSTLe`-annotation for both HRO control samples across all reference data set. 95
5.7 Relative composition of both HRO samples using the Cowan organoid trained `CaSTLe` classifier. 96
5.8 Consitency of three major retinal cell types across both annotation approaches. 98
5.9 Merged HRO controls data sets. 100
5.10 Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Peng *et al.* [2019a]. 103
5.11 Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Voigt *et al.* [2019]. 104

5.12 Distribution of Pearson correlation of HRO-cells and human tissue reference data taken from Cowan *et al.* [2020]. 105

5.13 Trajectory inference of HRO-2 using RNA velocity and `PAGA`. 108

5.14 Trajectory inference of HRO-3 using RNA velocity and `PAGA`. 109

5.15 The spatial location of the parafovea in the eye taken from Tsang & Sharma [2018]. 114

5.16 Development of retinal cells taken from Sridhar *et al.* [2020]. 115

5.17 The genesis of retinal cell types taken from Quinn & Wijnholds [2019]. 116

S-1 Distribution of logged expression values of all reference data sets and a human retina data set. 129

S-2 Box plots showing the behaviour of module preservation across different dropout levels. 130

S-3 Density of TOM values before and after imputation of the synthetic data set. 132

S-4 Dot plot of marker gene expression in human retina organoids. 134

S-5 Annotation result of human retina data set. 135

S-6 Distribution of cone- and rod-specific module membership ranks in complementary network. 136

S-7 Dot plot of marker gene expression in HRO2. 138

S-8 Dot plot of marker gene expression in HRO3. 139

S-9 Distribution of Pearson correlation of HRO-2 cells and human tissue reference data taken from Voigt *et al.* [2019] per donor. 143

S-10 Distribution of Pearson correlation of HRO-3 cells and human tissue reference data taken from Voigt *et al.* [2019] per donor. 144

S-11 Distribution of Pearson correlation of HRO-2 cells and human tissue reference data taken from Cowan *et al.* [2020] per donor. 145

S-12 Distribution of Pearson correlation of HRO-3 cells and human tissue reference data taken from Cowan *et al.* [2020] per donor. 146

S-13 RNA velocity streams of manual annotation results. 147

List of Tables

3.1 Overview of published imputation approaches used in this thesis. 33
3.2 Overview of filtering thresholds used for data preprocessing. 36

4.1 Overview of preserved hub genes. 69
4.2 Overview of gene associations detected in the EIF4EBP1 subnetworks. 73

5.1 Overview of cell types and amounts of the Cowan *et al.* [2020] data sets. . . . 83

S-1 Results of retina annotation pipeline before and after imputation. 131
S-2 Overview of cell types and amounts of the Cowan *et al.* [2020] organoid data set. 140
S-3 Overview of cell types and amounts of the Cowan *et al.* [2020] adult data sets. 141

Bibliography

Acharya, U. R., Mookiah, M. R. K., Koh, J. E., Tan, J. H., Noronha, K., Bhandary, S. V., Rao, A. K., Hagiwara, Y., Chua, C. K. & Laude, A. (2016). Novel risk index for the identification of age-related macular degeneration using radon transform and DWT features. *Computers in Biology and Medicine*, **73**, 131–140.

Acquaah-Mensah, G. K., Agu, N., Khan, T. & Gardner, A. (2015). A regulatory role for the insulin- and BDNF-Linked RORA in the hippocampus: Implications for Alzheimer's disease. *Journal of Alzheimer's Disease*, **44**, 827–838.

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., Van Den Oord, J., Atak, Z. K., Wouters, J. & Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, **14**, 1083–1086.

Al-Zamil, M. & Yassin, S. A. (2017). Clinical Interventions in Aging Dovepress Recent developments in age-related macular degeneration: a review. *Clinical Interventions in Aging*, 12–1313.

Aldridge, S. & Teichmann, S. A. (2020). Single cell transcriptomics comes of age.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 2020 21:1*, **21**, 1–16.

Andrews, T. S. & Hemberg, M. (2019). False signals induced by single-cell imputation [version 2; peer review: 3 approved, 1 approved with reservations]. *F1000Research*, **7**, 1–35.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J. & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, **20**, 163–172.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, **3**, 78.

Barabási, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*.

Besharse, J. C. & Pfenninger, K. H. (1980). Membrane Assembly in Retinal Photoreceptors I. Freeze-fracture Analysis of Cytoplasmic Vesicles in Relationship to Disc Assembly. *THE JOURNAL Of CELL BIOLOGY*, **87**, 451–463.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**, P10008.

Booij, J. C., Baas, D. C., Beisekeeva, J., Gorgels, T. G. & Bergen, A. A. (2010). The dynamic nature of Bruch's membrane. *Progress in Retinal and Eye Research*, **29**, 1–18.

Breda, J., Zavolan, M. & van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, **39**, 1008–1016.

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

Butte, A. J. & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418–429.

Chen, G., Ning, B. & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis.

Chen, M., Muckersie, E., Robertson, M., Forrester, J. V. & Xu, H. (2008). Up-regulation of complement factor B in retinal pigment epithelial cells is accompanied by complement activation in the aged retina. *Experimental Eye Research*, **87**, 543–550.

Chen, S. & Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, **19**, 1–21.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S.-A., Stoeckert, C. J., Jr., Taylor, C. F., Taylor, R. & Ball, C. A. (2011). Data Standards for Omics Data: The Basis of Data Sharing and Reuse. *Methods in molecular biology (Clifton, N.J.)*, **719**, 31.

Cobb, M. (2015). Who discovered messenger RNA? *Current Biology*, **25**, R526–R532.

Cooper, G. M. (2000). Translation of mRNA.

Cowan, C. S., Renner, M., De Gennaro, M., Gross-Scherf, B., Goldblum, D., Hou, Y., Munz, M., Rodrigues, T. M., Krol, J., Szikra, T., Cuttat, R., Waldt, A., Papasaikas, P., Diggelmann, R., Patino-Alvarez, C. P., Galliker, P., Spirig, S. E., Pavlinic, D., Gerber-Hollbach, N., Schuierer, S., Srdanovic, A., Balogh, M., Panero, R., Kusnyerik, A., Szabo, A., Stadler, M. B., Orgül, S., Picelli, S., Hasler, P. W., Hierlemann, A., Scholl, H. P., Roma, G., Nigsch, F. & Roska, B. (2020). Cell Types of the Human Retina and Its Organoids at Single-Cell Resolution. *Cell*, **182**, 1623–1640.e34.

Crick, F. (1970). Central Dogma of Molecular Biology. Technical report.

Cruz, N. M., Song, X., Czerniecki, S. M., Gulieva, R. E., Churchill, A., Kim, Y. K., Winston, K., Tran, L. M., Diaz, M., Fu, H., Finn, L. S., Pei, Y., Himmelfarb, J. & Freedman, B. S. (2017). Organoid cystogenesis reveals a critical role of microenvironment in human polycystic kidneydisease. *Nature Materials 2017 16:11*, **16**, 1112–1119.

Curcio, C. A. (2001). Photoreceptor topography in ageing and age-related maculopathy. *Eye*, **15**, 376–383.

Dairawan, M. & Shetty, P. J. (2020). The Evolution of DNA Extraction Methods, 2020–2028.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, **10**, 1–14.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, **5**, 0054–0066.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. & Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature*, **260**, 500–507.

Franklin, R. & Gosling, R. (1953). Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.

Fu, Y., Dominissini, D., Rechavi, G. & He, C. (2014). Gene expression regulation mediated through reversible m 6 A RNA methylation.

Ghanbari, M., Erkeland, S., Xu, L., Colijn, J., Franco, O., Dehghan, A., Klaver, C. & Meester-Smoor, M. (2017). Genetic variants in microRNAs and their binding sites within gene 3'UTRs associate with susceptibility to age-related macular degeneration. *Human mutation*, **38**, 827–838.

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Christopher Love, J. & Shalek, A. K. (2017). Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, **14**, 395–398.

Gong, W., Kwak, I.-y., Pota, P., Koyano-nakagawa, N. & Garry, D. J. (2018). DrImpute Imputing dropout events in single cell RNA sequencing data — RNA-Seq Blog, 1–10.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, **13**, 845–848.

He, Y., Yuan, H., Wu, C. & Xie, Z. (2020). DISC: a highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. *Genome Biology 2020 21:1*, **21**, 1–28.

Herring, C. A., Banerjee, A., McKinley, E. T., Simmons, A. J., Ping, J., Roland, J. T., Franklin, J. L., Liu, Q., Gerdes, M. J., Coffey, R. J. & Lau, K. S. (2018). Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Systems*, **6**, 37–51.e9.

Hou, W., Ji, Z., Ji, H. & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, **21**, 1–30.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M. & Zhang, N. R. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nature Methods*, **15**, 539–542.

JM, C. & H, F. (2016). The Discovery of Reverse Transcriptase. *Annual review of virology*, **3**, 29–51.

Kamimoto, K., Hoffmann, C. M. & Morris, S. A. (2020). CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*.

Kenneth M. Merz, J., Amaro, R., Cournia, Z., Rarey, M., Soares, T., Tropsha, A., Wahab, H. A. & Wang, R. (2020). Editorial: Method and Data Sharing and Reproducibility of Scientific Results. *Journal of Chemical Information and Modeling*, **60**, 5868–5869.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., Itzkovitz, S., Colonna, M., Schwartz, M. & Amit, I. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell*, **169**, 1276–1290.e17.

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

Kim, J., Koo, B. K. & Knoblich, J. A. (2020). Human organoids: model systems for human biology and medicine.

Kim, S., Lowe, A., Dharmat, R., Lee, S., Owen, L. A., Wang, J., Shakoor, A., Li, Y., Morgan, D. J., Hejazi, A. A., Cvekl, A., DeAngelis, M. M., Jimmy Zhou, Z., Chen, R. & Liu, W. (2019). Generation, transcriptome profiling, and functional validation of cone-rich human retinal organoids. *Proceedings of the National Academy of Sciences of the United States of America*, **166**, 10824–10833.

Klein, R., Peto, T., Bird, A. & Vannewkirk, M. R. (2004). The epidemiology of age-related macular degeneration. *American Journal of Ophthalmology*, **137**, 486–495.

Kogelman, L. J., Cirera, S., Zhernakova, D. V., Fredholm, M., Franke, L. & Kadarmideen, H. N. (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Medical Genomics*, **7**.

Kolker, E., Özdemir, V., Martens, L., Hancock, W., Anderson, G., Anderson, N., Aynacioglu, S., Baranova, A., Campagna, S. R., Chen, R., Choiniere, J., Dearth, S. P., Feng, W.-C., Ferguson, L., Fox, G., Frishman, D., Grossman, R., Heath, A., Higdon, R., Hutz, M. H., Janko, I., Jiang, L., Joshi, S., Kel, A., Kemnitz, J. W., Kohane, I. S., Kolker, N., Lancet, D., Lee, E., Li, W., Lisitsa, A., Llerena, A., MacNealy-Koch, C., Marshall, J.-C., Masuzzo, P., May, A., Mias, G., Monroe, M., Montague, E., Mooney, S., Nesvizhskii, A., Noronha, S., Omenn, G., Rajasimha, H., Ramamoorthy, P., Sheehan, J., Smarr, L., Smith, C. V., Smith, T., Snyder, M., Rapole, S., Srivastava, S., Stanberry, L., Stewart, E., Toppo, S., Uetz, P., Verheggen, K., Voy, B. H., Warnich, L., Wilhelm, S. W. & Yandl, G. (2014). Toward More Transparent and Reproducible Omics Studies Through a Common Metadata Checklist and Data Publications. *OMICS : a Journal of Integrative Biology*, **18**, 10.

Kwon, W. & Freeman, S. A. (2020). Phagocytosis by the Retinal Pigment Epithelium: Recognition, Resolution, Recycling.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S. & Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, **560**, 494–498.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., de Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T. H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., de Ridder, J., Saliba, A. E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P. & Schönhuth, A. (2020). Eleven grand challenges in single-cell data science, volume 21. Genome Biology.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A. & Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature 2001 409:6822*, **409**, 860–921.

Langfelder, P. & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**.

Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Computational Biology*, **7**.

Lee, J. H., Gao, C., Peng, G., Greer, C., Ren, S., Wang, Y. & Xiao, X. (2011). Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circulation Research*, **109**, 1332–1341.

Lieberman, Y., Rokach, L. & Shay, T. (2018). CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE*, **13**, e0205499.

Liu, Y., Beyer, A. & Aebersold, R. (2016). Leading Edge Review On the Dependency of Cellular Protein Levels on mRNA Abundance.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, **13**, e1005457.

Luecken, M. D. & Theis, F. J. (2019). Current best practices in singlecell RNAseq analysis: a tutorial. *Molecular Systems Biology*, **15**.

Luo, Y., Coskun, V., Liang, A., Yu, J., Cheng, L., Ge, W., Shi, Z., Zhang, K., Li, C., Cui, Y., Lin, H., Luo, D., Wang, J., Lin, C., Dai, Z., Zhu, H., Zhang, J., Liu, J., Liu, H., Devellis, J., Horvath, S., Sun, Y. E. & Li, S. (2015). Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. *Cell*, **161**, 1175–1186.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**.

Mattick, J. S. & Makunin, I. V. (2006). Non-coding RNA.

McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.

Mcinnes, L., Healy, J. & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Technical report.

Miko, I. (2008). Gregor Mendel and the Principles of Inheritance — Learn Science at Scitable.

Molday, R. S. & Moritz, O. L. (2015). Photoreceptors at a glance. *Journal of Cell Science*, **128**, 4039.

Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. (2020). A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Briefings in Bioinformatics*, **00**, 1–15.

Ozsolak, F. & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M.-P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., Petryszak, R., PradaMedina, C. A., Talavera-López, C., Teichmann, S., Vizcaino, J. A. & Brazma, A. (2020). Expression Atlas update: from tissues to single cells. *Nucleic Acids Research*, **48**, D77–D83.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, **6**.

Patruno, L., Maspero, D., Craighero, F., Angaroni, F., Antoniotti, M. & Graudenzi, A. (2020). A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V., and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P., , and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. & Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, **12**, 2825—-2830.

Peng, T., Zhu, Q., Yin, P. & Tan, K. (2019a). SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biology*, **20**, 1–12.

Peng, Y. R., Shekhar, K., Yan, W., Herrmann, D., Sappington, A., Bryman, G. S., van Zyl, T., Do, M. T. H., Regev, A. & Sanes, J. R. (2019b). Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell*, **176**, 1222–1237.e22.

Perevozchikov, A. P., Kuznetsov, O. K. & Zerov, Y. P. (1973). RNA dependent RNA polymerase in virions of Rous sarcoma virus. *Doklady Biological Sciences*, **209**, 121–124.

Pevsner, J. (2015). Bioinformatics and Functional Genomics. Wiley Blackwell.

Plassschaert, L. W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A. M. & Jaffe, A. B. (2018). A single cell atlas of the tracheal epithelium reveals the CFTRrich pulmonary ionocyte. *Nature*, **560**, 377–381.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, **17**, 147–154.

Quinn, P. M. & Wijnholds, J. (2019). Retinogenesis of the Human Fetal Retina: An Apical Polarity Perspective. *Genes*, **10**.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

S, K., MJ, B., Z, H., M, S., A, W., F, S.-C., P, G., L, S., JS, F., D, H., Z, Q., M, H., WB, H., P, K., S, P., B, T. & JG, C. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, **574**, 418–422.

Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology 2019 37:5*, **37**, 547–554.

Sakurai, M. (2015). Parafovea.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463.

Schaffter, T., Marbach, D. & Floreano, D. (2011). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Shendure, J. (2008). The beginning of the end for microarrays? *Nature Methods*, **5**, 585–587.

Sheu, S.-J., Chen, J.-L., Bee, Y.-S., Lin, S.-H. & Shu, C.-W. (2019). ERBB2-modulated ATG4B and autophagic cell death in human ARPE19 during oxidative stress. *PLoS ONE*, **14**.

Soneson, C., Srivastava, A., Patro, R. & Stadler, M. B. (2021). Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology*, **17**, e1008585.

Sridhar, A., Hoshino, A., Finkbeiner, C. R., Chitsazan, A., Dai, L., Haugan, A. K., Eschenbacher, K. M., Jackson, D. L., Trapnell, C., Bermingham-McDonogh, O., Glass, I. & Reh, T. A. (2020). Single-Cell Transcriptomic Comparison of Human Fetal Retina, hPSC-Derived Retinal Organoids, and Long-Term Retinal Cultures. *Cell Reports*, **30**, 1644–1659.e4.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. In *Bioinformatics*, volume 18. Oxford University Press.

Syrbe, S., Kuhrt, H., Gärtner, U., Habermann, G., Wiedemann, P., Bringmann, A. & Reichenbach, A. (2017). Müller glial cells of the primate foveola: An electron microscopical study. *Experimental Eye Research*, **167**, 110–117.

Tan, Y. & Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, **9**, 207–213.e2.

Terakita, A. (2005). The opsins. *Genome Biology*, **6**, 213.

Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., Berry, G. J., Shrager, J. B., Metzger, R. J., Kuo, C. S., Neff, N., Weissman, I. L., Quake, S. R. & Krasnow, M. A. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature 2020 587:7835*, **587**, 619–625.

Tsang, S. H. & Sharma, T. (2018). Retinal Histology and Anatomical Landmarks. *Advances in Experimental Medicine and Biology*, **1085**, 3–5.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*, **19**, 575–592.

Van Der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Vaugon, M. (2006). Inequalities in Sobolev Spaces. *Encyclopedia of Mathematical Physics: Five-Volume Set*, 32–36.

Voigt, A. P., Mullin, N. K., Whitmore, S. S., DeLuca, A. P., Burnight, E. R., Liu, X., Tucker, B. A., Scheetz, T. E., Stone, E. M. & Mullins, R. F. (2021). Human photoreceptor cells from different macular subregions have distinct transcriptional profiles. *Human Molecular Genetics*, **30**, 1543.

Voigt, A. P., Whitmore, S. S., Flamme-Wiese, M. J., Riker, M. J., Wiley, L. A., Tucker, B. A., Stone, E. M., Mullins, R. F. & Scheetz, T. E. (2019). Molecular characterization of foveal versus peripheral human retina by single-cell RNA sequencing. *Experimental Eye Research*, **184**, 234–242.

Völkner, M., Kurth, T., Schor, J., Ebner, L. J., Bardtke, L., Kavak, C., Hackermüller, J. & Karl, M. O. (2021). Mouse Retinal Organoid Growth and Maintenance in Longer-Term Culture. *Frontiers in Cell and Developmental Biology*, **9**, 1–25.

Wagner, F., Barkley, D. & Yanai, I. (2019). Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis, 1–9.

Wang, C., Gao, X. & Liu, J. (2020). Impact of data preprocessing on cell-type clustering based on single-cell RNA-seq data. *BMC Bioinformatics 2020 21:1*, **21**, 1–13.

Wang, K., Li, H., Sun, R., Liu, C., Luo, Y., Fu, S. & Ying, Y. (2019). Emerging roles of transforming growth factor β signaling in wet age-related macular degeneration. *Acta biochimica et biophysica Sinica*, **51**, 1–8.

Wang, Q., Oh, J. W., Lee, H. L., Dhar, A., Peng, T., Ramos, R., Guerrero-Juarez, C. F., Wang, X., Zhao, R., Cao, X., Le, J., Fuentes, M. A., Jocoy, S. C., Rossi, A. R., Vu, B., Pham, K., Wang, X., Mali, N. M., Park, J. M., Choi, J. H., Lee, H., Legrand, J. M., Kandyba, E., Kim, J. C., Kim, M., Foley, J., Yu, Z., Kobielak, K., Andersen, B., Khosrotehrani, K., Nie, Q. & Plikus, M. V. (2017). A multi-scale model for hair follicles reveals heterogeneous domains driving rapid spatiotemporal hair growth patterning. *eLife*, **6**.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics.

Watson, J. D. & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356*, **171**, 737–738.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3:1*, **3**, 1–9.

Wolf, F. A., Angerer, P. & Theis, F. J. (2018). Open Access S CANPY : large-scale single-cell gene expression data analysis, 1–5.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L. & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, **20**, 1–9.

Wong, W. L., Su, X., Li, X., Cheung, C. M. G., Klein, R., Cheng, C. Y. & Wong, T. Y. (2014). Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, **2**, e106–e116.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J. Y., Horvath, S. & Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, **500**, 593–597.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N. & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, **5**.

Ye, L., Swingen, C. & Zhang, J. (2013). Induced Pluripotent Stem Cells and Their Potential for Basic and Clinical Sciences. *Current Cardiology Reviews*, **9**, 63.

Ye, W., Ji, G., Ye, P., Long, Y., Xiao, X., Li, S., Su, Y. & Wu, X. (2019). ScNPF: An integrative framework assisted by network propagation and network fusion for preprocessing of single-cell RNA-seq data. *BMC Genomics*, **20**, 1–16.

Yip, A. M. & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, **8**, 1–14.

Zappia, L., Phipson, B. & Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, **18**, 1–15.

Zhai, X., Xue, Q., Liu, Q., Guo, Y. & Chen, Z. (2017). Colon cancer recurrence-associated genes revealed by WGCNA co-expression network analysis. *Molecular Medicine Reports*, **16**, 6499–6505.

Zhang, L. & Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–14.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**, 1–12.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. & Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, **65**, 631–643.e4.

Curriculum Scientiae

Education:

since 10/2018	PhD student at the University of Leipzig and Helmholtz Centre for Environmental Research UFZ, Germany • Group of Dr. Jörg Hackermüller • Supervised by Prof. Peter F. Stadler, Chair of Bioinformatics • Thesis: *Network inference from sparse single-cell transcriptomics data*
10/2016 – 09/2018	Master student at the University of Cologne, Germany • M.Sc. Biological Sciences • Thesis: *Establishment of high throughput clone screening using CRISPR- Cas9.*
09/2013 – 09/2016	Bachelor student at the University of Applied Sciences Bonn-Rhine-Sieg, Germany • B.Sc. Applied Biology • Thesis: *Effect of LSD1 Silencing on miRNA Expression Pattern in Hepatic Stellate Cells.*

Practical Training:

11/2017 – 12/2017	Project course during master studies, Cologne Center for Genomics, Germany • Statistical Genetics and Bioinformatics group of Prof. Dr. Michael Nothnagel • Project: *Application of 1000 Genome Data onto three different polygenic risk score sets for Breast and Ovarian Cancer.*
09/2017 – 10/2017	Project course during master studies, Cologne Center for Genomics, Germany

- Functional Epigenomics group of Prof. Dr. Dr. Michal R. Schweiger
- Project: *Analysis of the distinct role of SF3B1 mutations during chronic lyphocytic leukemia*

IT-Knowledge:

OPERATING SYSTEMS:	UNIX, Mac, Linux, Windows
PROGRAMMING:	Python, R
MARKUP LANGUAGES:	Latex, HTML

Language Skills:

GERMAN:	native speaker
ENGLISH:	fluent
FRENCH:	basic knowledge

Publications

Journals:

Steinheuer L. M. & Canzler S.& and Hackermüller J. (2021).
Benchmarking scRNA-seq imputation tools with respect to network preservation highlights lack in performance. In preparation. doi: https://doi.org/10.1101/2021.04.02.438193 .

Völkner M. & Wagner F. & Steinheuer L. M.& Carido M. & Kurth T. & Yazbeck A. & Schor J. & Wieneke S. & Ebner L. & Del Toro Runzer C. & Taborsky D. & Zoschke K. & Vogt M. & Canzler S. & Hermann A.& Khattak S. & Hackermüller J. & Karl M. O. (2021).
HBEGF-TNF induces a complex outer retinal atrophy with macular degeneration hallmarks in human organoids. In revision.

Schubert, Kristin & Karkossa, Isabel & Schor, Jana & Engelmann, Beatrice & Steinheuer, Lisa Maria & Bruns, Tony & Rolle-Kampczyk & Ulrike & Hackermüller, Jörg & von Bergen, Martin. (2021).
A Multi-Omics Analysis of Mucosal-Associated-Invariant T Cells Reveals Key Drivers of Distinct Modes of Activation newblock *Frontiers in Immunology*, 2021, 12, 616967.

Yu, X & Steinheuer L. M. & Schmiel, M & Ulmer, B & Eischeid, H & Büttner, R & Odenthal, Margarete & Wang, Lingyu. (2018).
The epigenetic writer LSD1 controls hepatic cell cycle progression by alteration of global expression profiles and protein signatures. *Zeitschrift für Gastroenterologie*, 56. E2-E89. 10.1055/s-0037-1612769.

Yu, X & Steinheuer L. M. & Schmiel, M & Ulmer, B & Eischeid, H & Buettner, R & Odenthal, Margarete & Wang, Lingyu. (2018). **The lysine-specific histone demethylase LSD1 controls cell cycle progression in liver fibrosis and cancer by alteration of protein expression profiles.** *Zeitschrift für Gastroenterologie*, 56. E2- E89. 10.1055/s-0037-1612806.

Wang, Lingyu & Steinheuer L. M. & Ulmer, B & Schmiel, M & Yu, X & Eischeid, H & Büttner, R & Odenthal, Margarete. (2018).
The lysine specific histone demethylase LSD1 in activated Hepatic Stellate Cells contributes to liver fibrosis. *Zeitschrift für Gastroenterologie*, 56. E2-E89. 10.1055/s-0037-1612682.

Wang, Lingyu & Yu, X. & Schmiel, M. & Steinheuer L. M. & Eischeid, H. & Ulmer, B. & Büttner, R. & Odenthal, Margarete. (2018).
The impact of the epigenetic writer LSD1 in the cell cycle control in liver fibrosis and hepatocellular carcinoma. . *Journal of Hepatology*, 68. S690. 10.1016/S0168-8278(18)31640-4.

Wang, Lingyu & Steinheuer L. M. & Ulmer, B & Yu, X & Eischeid, H & Buettner, R & Odenthal, Margarete. (2016).
The Epigenetic Modifications by the Histone Demethylase LSD1 in Hepatic Stellate Cells Contribute to Liver Fibrosis. *Zeitschrift für Gastroenterologie*, 54. 1343-1404. 10.1055/s-0036-1597373.

Conferences / Seminars:

sparse2Big Final Meeting (Presenter)
Steinheuer L. M.: *Data imputation in moderately sparse data can facilitate network inference*
09/2021; Schliersee, Germany

Symposium on Interdisciplinary Research in Mathematics and Life Sciences (Poster - Poster price)
Steinheuer L. M., Canzler S., and Hackermüller J.: *Data imputation in moderately sparse data can facilitate network inference*
08/2021; Bonn, Germany

Invited speaker AG Thurley Retreat (Presenter)
Steinheuer L. M.: *Data imputation in moderately sparse data can facilitate network inference*
08/2021; Bonn, Germany

Symposium - Single Cell Biology (Poster)
Steinheuer L. M., Canzler S., and Hackermüller J.: *Benchmarking scRNA-seq imputation tools with respect to network inference highlights deficits in performance*
03/2021; online

Department Seminar Molecular Systems Biology (Presenter)
Steinheuer L. M.: *Looking for networks - An update on network inference from single-cell transcriptomics data*
02/2020; Leipzig, Germany

IP Exposome Plenary Meeting (Presenter)
Steinheuer L. M.: *Diving into the individual cell - exploiting the potential of single-cell RNA transcriptomics*
10/2019; Leipzig, Germany

Conference - The Identy and Evolution of Cell Types (Poster)
Steinheuer L. M., Canzler S., and Hackermüller J.: *Alleviating sparsity on scRNA-seq data using imputation approaches to unmask gene correlation structures*
05/2019; Heidelberg, Germany

Workshop Computational Single Cell Genomics (Poster)
Steinheuer L. M., Canzler S., and Hackermüller J.: *Network inference from sparse scRNA-seq data - Evaluation of imputations -*
03/2019; Munich, Germany

sparse2Big Meeting (Presenter)
Steinheuer L. M.: *Network inference from sparse scRNA-seq data - Evaluation of imputations -*
03/2021; Munich, Germany

Department Seminar Molecular Systems Biology (Presenter)
Steinheuer L. M.: *Sparse2Big- From sparse to big data*
03/2019; Leipzig, Germany

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Lisa Maria Steinheuer

Leipzig, 30.09.2021