ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master's Thesis Econometrics and Management Science

Uncovering the genetics of asthma: A comparison of gene regulatory networks

Author:
Clemens P.D. Hendrickx (451192)

External Supervisor:
MSc. M. van Breugel

Academic Supervisor:
prof. dr. R. Dekker

Second assessor:
dr. P. Wan

Abstract

Asthma is a long-term inflammatory disease of the airways that is caused in large part by genetic factors. This study investigates this genetic component that is known to involve a large number of interacting genes. To analyse changes in these interactions between asthmatic and healthy individuals, two large networks of genetic interactions are constructed and compared. We propose a new method for comparing networks that aims to uncover biologically meaningful differences. This method uses gene affinity profiles to embed information on the genes and their position in the networks, which are subsequently compared between the networks. The comparison and subsequent analyses yield a set of 40 genes whose internal interaction shows interesting differences between the networks and some of which have been linked to asthma in previous research. Future research can build on these results by using the identified set of genes as a starting point for more targeted studies on genetic interaction. Additionally, this study adds to the currently limited body of research on network comparison by proposing a new method for embedding gene information and comparing genetic interaction networks.

Wednesday 2nd March, 2022

Preface

This thesis describes the research conducted by the author, Clemens Hendrickx, as part of the Master's degree in Econometrics and Management Science at the Erasmus University Rotterdam. The research described in this document was conducted as part of the research collaboration between MIcompany, an artificial intelligence and data company headquartered in Amsterdam, and the University Medical Center Groningen (UMCG)[1].

The role of the author in this research collaboration was to conduct a master's thesis study that approaches uncovering the genetic component of asthma from a new unexplored direction and that could possibly serve as the basis for a publishable article. To this end the greatest care has been taken in every step of the research to ensure the validity of the results. This is also why the data preparation phase of the study was done very thoroughly and is described in great detail both in the main text and in the appendices. The choices made throughout the study were based on extensive reviews of the relevant literature and intermediate results were extensively validated. At certain points medical experts from the UMCG (see Appendix A) were consulted to inspect intermediate results or when a choice needed to be made that was outside the expertise of the author, often concerning topics in cellular biology and genomics. We would like to note that the steps taken and choices made in this study are entirely the author's own unless it is explicitly stated that a choice was made in consultation with these experts.

I would like to thank MIcompany for offering me the opportunity to conduct this research in which I was able to combine my passion for econometrics and machine learning with my deep interests in medicine and biology. In particular, I would like to thank M. van Breugel for his guidance throughout this project. Furthermore, this study also would not have been possible without the continuous support and involvement of the team of the UMCG, consisting of prof. dr. G. Koppelman, dr. ir. M.C. Nawijn and M. Berg, to whom I am extremely grateful. Finally, my gratitude goes out to prof. dr. R. Dekker, my supervisor from the Erasmus University Rotterdam, for his guidance and feedback throughout the whole process of writing this thesis.

[1] More specifically: the Groningen Research Institute for Asthma and COPD (GRIAC), which operates as part of the UMCG.

Table of contents

1 Introduction

Asthma is a chronic inflammatory disease of the airways that affects people from all over the world. It is estimated that around 262 million people worldwide suffer from the disease and that around 460 thousand people die because of it every year (Global Burden of Disease Study, 2020). The prevalence of the disease has seen a very strong increase in recent decades, a trend that is expected to continue in the foreseeable future. Even though the disease is very prevalent and has been known to humanity for a long time, the diagnosis of asthma still leaves much to be desired, with recent years seeing numerous publications on the over-, under- and mis-diagnosis of the disease (Heffler et al., 2015, 2018; Kavanagh et al., 2019). Current treatments for asthma are only able suppress the symptoms rather than curing this chronic disease (El-Husseini et al., 2020). For around 5 to 10 percent of patients these treatments are not enough to control the symptoms (Busse et al., 2000). The reason for this is that asthma is a complex disease that results from the interplay of many genetic and environmental factors (Ober & Hoffjan, 2006; Martinez, 2007; Halapi & Bjornsdottir, 2009; Thomsen, 2015). The exact causes of asthma are still not well understood and in particular the large genetic component[2] of this disease has proven to be difficult to explain. What makes the genetic component so difficult to understand is that the genetic mechanisms behind the disease are known to be polygenic, meaning that they involve a large number of interacting genes rather than a few isolated genes (Thomsen, 2015; Ober & Hoffjan, 2006; Martinez, 2007).

The ability to research these genetic mechanisms greatly increased with the introduction of single-cell RNA[3] sequencing (scRNA-seq) data in 2009 (Tang et al., 2009). This new type of genetic data differs from the classical bulk sequencing data in that it reports the different levels of genetic expression in individual cells, rather than the average expression per gene over all cells in a batch. The more granular nature of this data allows for much more detailed analyses of genetic mechanisms, which has contributed greatly to the popularity of this type of sequencing technology in the recent decade. With single-cell sequencing technology becoming widely available only recently, it has only been in the last few years that we have seen publications in the field of asthma research that utilise this technology. Most of these early papers focus on specific types of airway cells (L. Wang et al., 2021; H. Li et al., 2021; Royer & Cook, 2021). Braga et al. (2019) showed that the cells of healthy and asthmatic individuals show substantial

[2]Estimates from twin pair studies suggest that around 50%-60% of asthma is heritable (Skadhauge et al., 1999).

[3]Ribonucleic acid (RNA) is one of the main two types of molecules that carry genetic information, the other one being DNA.

differences in genetic expression.

The current literature on the genetics behind asthma tells us that genes do not operate in isolation (Thomsen, 2015) and calls for analysing their expression in relation to each other (El-Husseini et al., 2020). As a result of this, the field is moving towards a more interaction focused view where entire networks of genes are analysed. While it is known that genes operate in networks, there is much still unknown about what these networks look like exactly and how they work biologically. For this reason, researchers have to rely on computational reconstructions of these gene regulatory networks (GRNs) that try to capture the underlying mechanics that govern the genetic expression in cells (Heijink et al., 2020). The power of these reconstructions lies not in the accuracy with which each of the individual links between genes in these networks directly corresponds to an underlying biological reality, which can still have some degree of uncertainty to it. Rather, the power of these reconstructions lies in their ability to capture the larger interaction structures between different genes. It is creating these reconstruction that we refer to in this thesis when we talk about network construction or network inference. These networks are mathematically represented by weighted graphs. The nodes represent the expression levels of specific genes. The edges and their weights can be seen as the associations between genes, where the precise nature of this association depends on the mathematical model used for the construction of the graph. A more detailed description of these networks can be found in Section 4.3. The comparison of these networks between asthmatic and healthy individuals can provide valuable insights into how genetic interaction is different for asthmatic individuals. This could both help us to understand the genetic mechanisms behind this chronic disease better and could yield target genes for further research on drug development El-Husseini et al. (2020).

Even though computationally reconstructing these networks has been an active and promising field of research in the past decade, only very little research on the genetics behind asthma is done using a network based approach. We were able to find a few papers in which a basic correlation network is constructed (Jackson et al., 2020a; Seumois et al., 2020) and only one publication in which a case-control comparison between networks is performed (Banerjee et al., 2021). Yet it is precisely this type of network based studies that Heijink et al. (2020) identify in their paper on the mechanisms driving asthma development as the promising next step for the field that may ultimately lead to a deeper understanding of these drivers and that may help to identify possible treatments. El-Husseini et al. (2020) too calls for research on identifying the genetic factors of asthma and the biological networks within which they function.

This is what motivates us to conduct the large-scale network comparison study of this paper. Given the great promise and novelty of this type of research in the context of asthma genomics, we conduct a study in which construct these networks that consist of thousands of genes using state of the art techniques and compare them between the asthmatic and control groups. The goal of this study is to identify groups of genes whose interaction differs the most between asthmatic and healthy individuals. The interaction in these groups of genes could then be investigated more deeply and could be used as the starting point for future more targeted studies. In this way the identification of these groups could aid in the discovery of genetic interaction structures that play a role in the development of asthma. To this end, we formulate the following research question:

Which genes show the largest difference in their local interaction between asthmatic and healthy individuals?

To answer this question, we perform a comparative analysis of two computationally reconstructed gene regulatory networks, one belonging to an asthmatic subject group and the other to a control group, using a self-developed method for large-scale gene network comparison. The medical relevance for the field of asthma research of conducting such a network comparison study lies in its potential for identifying sets of genes for future more targeted research on asthma and the development of treatments. In addition to this medical relevance, this study also bears methodological relevance for the broader field of gene network analysis. This methodological relevance lies in the fact that for as far as we know no comprehensive approaches to large-scale gene network comparison exist. While there do exist many general methods for comparing networks (see Tantardini et al. (2019) for an overview), we were not able to find any that are suitable for reaching our goal of identifying nodes that differ in their local interaction structures between two networks. Our method is specifically designed for this and could potentially be used in similar contexts for research on diseases other than asthma.

The different steps that make up this study are divided into three parts: data preparation, network inference and network comparison. Figure 1 shows an overview of these steps and we will cover each part briefly in to following. The first part is that of data preparation. Here we start with raw scRNA-seq data containing the expression levels of genes in cells sampled from human subjects in the airway. As stated before, the highly granular nature of single-cell data enables us to perform analyses on a very detailed level. However, this also poses some issues

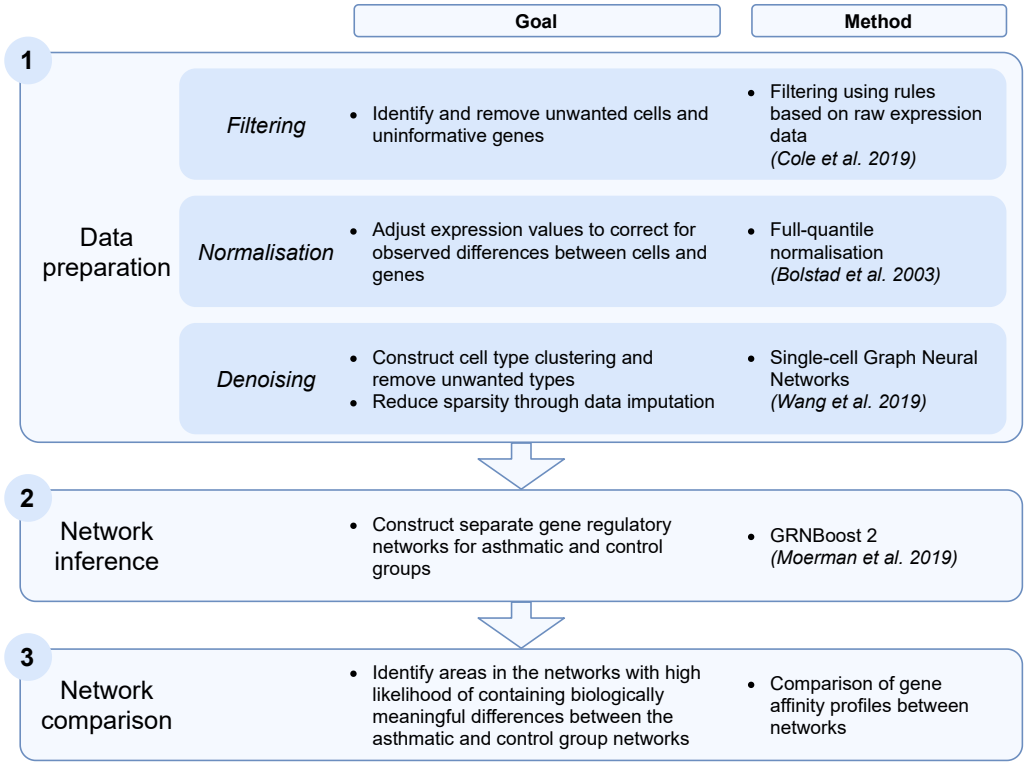


Figure 1. A schematic flow of the different parts that make up this study.

that can, when not properly dealt with, prevent us from constructing networks that accurately represent the underlying biological mechanisms and thus prevent us from performing a good comparison. These issues fall into three categories: Firstly, there are the issues relating to the quality of parts of the data, for example measurements on broken or dead cells, that need to be filtered out. Secondly, issues regarding the comparability of observations, where differences between some sets of observations caused by the sequencing technology call for adequate normalisation of the data. Lastly, there is a large amount of noise in the data caused by a variety of factors, one of which is the presence of measurement errors. Before we can construct the networks, our raw data has to go through an extensive data preparation process in which each of these issues is dealt with.

The second part of the study concerns constructing the networks for each subject group from the prepared data. Different algorithms for the inference of networks from genomic sequencing data have been proposed over the years. In this paper, we use the regression based GRNBoost2 (Moerman et al., 2019) algorithm that builds a network of genes based on the predictive relations it finds between the expression levels of genes in the data. It gives high weight to edges that correspond to a strong predictive relation between two genes' expression levels by

statistically comparing these gene relations within one type of cells with each other. The constructed networks are very large, consisting of thousands of nodes and tens of thousands of edges.

In the third and last part, the constructed networks are compared between the subject groups. Comparing networks of this size and complexity is no trivial task and different types of comparisons are possible depending on the goal of the comparison. The goal in this paper is to find how the interaction of genes differs between the asthmatic and control groups and how this could be linked to the pathology of asthma. In the context of the constructed networks, this means that we aim to find groups of genes that are connected differently to the other genes between the networks. We opted for an approach focused on the local connections of individual genes in our networks. Local connections in this case being defined as the paths from the gene of interest to easily accessible other genes. These are paths that are either short (consisting of few edges), strong (consisting of high weight edges) or a combination of the two. A localised approach was chosen because we can most confidently attribute changes in the local connections of genes in our networks to actual biological differences between the subject groups[4]. To perform such a comparison we developed our own method for large-scale gene network comparison. This method embeds nodes (genes) into gene affinity profiles to allow for the direct comparison of the connectivity structure in these local neighbourhoods of the genes between the two networks. Such a comparison is able to identify genes that show the largest difference in local connectivity between the two networks. These genes then serve as candidate genes for further analysis and inspection of their neighbourhood in the networks to look for biologically meaningful insights.

In this type of research where large networks are compared, taking steps to ensure the validity of the results is of the utmost importance. When comparing networks of this size, everything has to be done to minimise the chance of obtaining spurious results. Numerous steps have been taken to this end: All choices, when possible, are based on insights from existing literature and our approach was shown to experts in the field. Intermediate results were checked to see if the steps that were taken had the desired effect and were also shown to experts when needed. Additionally, only the strongest results from the network comparison that pass stringent robustness requirements were reported. One such example is that the selected results are consistently found over a range of different values for the network comparison hyperparameters. Finally, we also perform external validation by linking our results to existing results in the asthma genomics research literature.

[4]This choice was made in consultation with G.H. Koppelman, M.C. Nawijn and M. Berg, see Appendix A.

The rest of this paper is organised as follows: Section 2 gives a description of the problems encountered in each of the three parts of the study and provides an overview of the relevant literature relating to these problems. The data used in this researched as well as the first filtering step in the preparation process is described in Section 3. The methods used in each subsequent step of the study are described in Section 4. The results of the study are presented in Section 5. In Section 6, we conclude and formulate an answer to our research question and explore the implications and limitations of this study.

2 Problem description

This section describes the different challenges encountered on the path to answering our research question, gives an overview of the relevant literature and briefly describes the approach taken to overcome these challenges. As quite a few hurdles had to be overcome in the process of answering the research question, we chose to address these in this chapter separately from the explanation of the methods used to overcome them, which are given later in Section 4. This section is divided into three subsections: the data preparation, the inference of the networks and the comparison of the networks. These sections correspond to the different parts of this study as illustrated in Figure 1.

2.1 Data preparation

The first phase of the research is that of data preparation. For the analyses in this paper, we make use of single-cell RNA-sequencing data. This type of data captures the levels of expression for each different gene in a single cell. The data takes the form a read count matrix in which each row represents a different individual cell and each column represents a gene. The entries in the matrix are the counts of how many times the genetic material associated to that gene is found in that specific cell. Higher values hence correspond to higher levels of gene expression. This type of data is known to be challenging to work with and requires extensive pre-processing before it can be used. We can not overstate the importance of this pre-processing phase, with Vieth et al. (2019) going as far as to claim that the implementation of the right pre-processing pipeline can have the same effect on the research outcome as quadrupling the sample size would have. Based on literature, we identify three categories of problems that need to be dealt during data preparation: data quality, data comparability and data noisiness. Each of these problems is covered separately in the sections below.

2.1.1 Data quality

The first problems that we encounter when working with single-cell data have to do with the quality of some parts of the data. Due to the highly detailed nature of single-cell data, our data set may contain observations on low-quality cells or uninformative genes. The presence of such undesirable observations belonging to either low-quality cells or uninformative genes is a source of noise that can distort the signal that we aim to measure. Therefore, the identification and removal of these observations (which we will call filtering from this point onwards) is a crucial step in preparing the data (G. Chen et al., 2019). Low-quality observations are those coming from broken, dead or mixed cells and the presence of such observations is a known feature of

raw single-cell data (Ilicic et al., 2016). The data from these observations cannot be said to be representative for that of regular living cells, which we aim to investigate. Such cells can be identified by exceptionally high or low levels of genetic expression and by the proportion of mitochondrial genetic material found (Bacher & Kendziorski, 2016). Uninformative genes are those that show very low levels of expression and thus provide us with little information. We would expect that quite a number of genes in our data to show very little to no expression in our data. This is because we have measurements on all the genes in the human genome (which consists of around 32 thousand genes), many of which likely play no role in the workings of the lung cells that we analyse in this study. Through filtering out these genes we can both increase the quality of our data and combat zero inflation (Cole et al., 2019). The implementation of this filtering is described in more detail in the chapter describing the data used (Section 3).

2.1.2 Data comparability

Another data issue we run into after having removed low quality observations has to do with the comparability of different observations. Our data set contains the raw counts of the number of times the genetic material of specific genes is measured in a single cell. Aside from the level of expression of the gene in question, some of the differences in the raw counts can be due to other factors. It is for example said that levels of the read counts are related to gene length and that they can vary between batches due to differences in the sequencing depth[5] (Risso et al., 2011). Because of these factors the raw counts are not directly comparable over all observations (Risso et al., 2011) and some transformation of the data is needed to allow for an honest comparison between observations.

This transformation, which is called normalisation, is identified in the literature as an essential step to be able to properly analyse the differences in transcript levels (Robinson & Oshlack, 2010; Risso et al., 2011; Anders & Huber, 2012; Bacher et al., 2017; Cole et al., 2019; Vieth et al., 2019). The goal of normalisation is to account for observed differences in measurements between samples and genes that are due to either technical reasons or unwanted biological effects (e.g. batch effects), but at the same time to preserve the biological effects of interest (Cole et al., 2019). We aim to adjust these systematic variations to make the expression counts comparable across samples (Bacher et al., 2017). Rather than have the absolute raw measurements, we analyse the normalised counts that reflect more of a relative expression level within cells. One of the problems that we run into when trying to do this is that we do not know exactly

[5]The sequencing depth is the total number of reads that are produced for a given batch.

what all the factors are that cause these (slight) differences. The process of accounting for differences while at the same time preserving the biological variation of interest thus requires a careful consideration of methods and implementation when normalising scRNA-seq data, since one risks transforming the data in a way that distorts the signal of interest. Finding the right balance between these two objectives is no easy task and different techniques have been shown to have different impacts on downstream analyses (Cole et al., 2019; Vieth et al., 2019). How and which method we chose for this normalisation step is described in Section 4.1.

2.1.3 Data noisiness

The last data issue that we have to take care of before we can construct the networks is that of the high levels of noise in the data. The presence of this noise makes it more difficult to measure the gene relationships in the data accurately, which makes identifying the biological signals needed for constructing networks in the next step more difficult. The two major sources of noise that we identify are that of high data sparsity and cell type heterogeneity.

Data sparsity (dropouts)

The raw expression data contained a high proportion of zero entries, namely around 88 percent. This is in line with sparsity levels typical for this kind of data, which are generally between 85 and 95 percent (Van Dijk et al., 2018). Even though the filtering step does lead to some reduction in the sparsity, the filtered data set still contains around 75 percent zero entries (see Table 1). In this data set, an entry can be zero either because that particular gene is not expressed in that cell at the moment of measurement (a true zero) or because of a lack of detection by the sequencing method (a false or dropout zero). This high number of zeros introduces great imbalance in the data, with large parts of the data having one and the same value. This can lead to complications in downstream analyses. The probability of dropouts occurring is difficult to pinpoint as these differ across sequencing techniques. However, it has been shown that the Smart-seq2 protocol used to create our data set suffers less than more recent methods like 10X genomics, with a rough estimate of the dropout rate being somewhere from zero to two percent (X. Wang et al., 2021).

We tackle the issue of data sparsity in two ways: Firstly, by focusing the analysis on a subset of highly variable genes which on average have a lower proportion of zero entries. Secondly, by performing an imputation step on the data in which we transform the data to a more workable format and in the process correct for some of the errors introduced into the data by dropout

zero's. See Section 4.2 for a detailed description.

Cellular heterogeneity

The data that we are using comes with an initial rough labelling of cells by type, indicating that the data set consists of different cell types. The identified types are: basal, secretory, ciliated, immune and ionocyte cells. Because different cell types have different functions, they show different patterns of genetic expression. These differences in expression are an additional source of noise in the data that can potentially obstruct uncovering clear signals in downstream analyses. The issue of noise introduced by cellular heterogeneity is tackled by clustering the cells on cell type and focusing our analysis on a selection of cells, see Section 4.2 for more details.

2.2 Network inference

The second phase of the study concerns the construction of the two networks from our prepared data set, one for our asthmatic and one for our control group. The goal of our networks is to capture the relations between these genes. To this end, the nodes in these networks represent (expression levels of) genes, the edges represent the relations between these genes and the weights on the edges represent the strengths of these relations. The nature of the relationship captured by the edges of the network depends on the method used for its construction. A fast and inexpensive option that has seen some use (Jackson et al., 2020a; Seumois et al., 2020) is to simply set the weights on the edges equal to the correlation between the expression of the genes. Because this is often not sufficient to capture the complex relations between genes, more advanced methods have been proposed. These make use of probabilistic models, information theory or another algorithmic approach to capture more complex relations between genes. These more advanced methods can quickly become too computationally expensive as one tries to construct larger networks using more data. The challenge in the network inference phase lies in selecting the right method and implementing this in a scalable and computationally feasible manner. In Section 4.3 we motivate our choice of method and describe how it works.

2.3 Network comparison

The last phase of this study concerns the comparison of the constructed networks in order to identify regions with the highest likelihood of containing biologically meaningful differences in the genetic interaction structures for asthmatic individuals. This is very complicated given the size and complexity of the networks in question, and we identify two important aspects that

need to be considered Firstly, we need to choose on what scale (global, local or something in-between) that we want to compare the networks. Secondly, we need to define the type of quantifiable difference between the networks that is biologically meaningful for a comparison on that scale. These two choices are treated separately in the sections below.

2.3.1 Scale of the comparison

A great many types of comparisons are possible when comparing two large networks with thousands of nodes and tens of thousands of edges. It is easy to lose oneself in all the possible options that exist for performing such a comparison and it is key to not lose sight of what information you are looking for and to pick the approach that is best suited for this (Wills & Meyer, 2020). A global approach could be preferred when you are interested in comparing the networks as a whole. On the other end of the spectrum are the comparisons of structures on the smallest levels such as the degrees of individual nodes. Something in-between would involve identifying clusters or modules in the networks and then comparing those. The only application of any of these types of network comparisons in the field of asthma that we could find was in Banerjee et al. (2021). In that paper, they identified hubs as the genes with high degrees of connectivity and their comparison consisted of comparing the sets of identified hub genes between the networks.

In the context of this study, our ultimate goal is to identify the regions in the network with the largest differences in interaction whose genes could serve as target genes in future research. The networks in this study are constructed such that nodes represent genes, edges between nodes represent a predictive relation between their expression levels and the edge weights represent the strength of this relation. With the goal of identifying regions of interest, we opt for an approach in which we compare each genes' connectivity with its neighbourhood between the asthmatic and the control network. In terms of scale, such an approach can be considered a local approach because we compare each gene with its counterpart in the other network. However, each of these comparisons compares information on the connectivity of the neighbourhood around that gene and is thus not entirely local. Alternatives to focusing on each gene individually are either a global comparison of the networks, which would not yield the type of insights that we are looking for, or a module based approach in which clusters of genes in the network are identified and compared. This last option is also not well suited for our purpose since it is heavily dependent on the way the modules are constructed and when comparing different modules you have essentially multiplied the number of network comparisons that you have to do.

2.3.2 Defining a meaningful comparison

Given that we want to focus on comparing each gene individually, it then becomes a question of how to compare the genes in the different networks in a meaningful way. Our approach is based on the idea that a type of difference between our asthmatic and control networks which is biologically meaningful is one where the same genes in different networks are connected differently to the genes in their local vicinity[6]. This is how we define the biologically meaningful difference between the networks that we look for in this study. We would like to note however that by choosing this definition we do not deny the existence of other types of biologically meaningful differences between these types of networks. This definition was chosen because it is one that is relatively straightforward, measurable and confirmed by experts. In the following we describe the intuition behind our approach to measuring this type of difference, see Section 4.4 for a detailed description of the methodology behind this approach.

When you look at the network from the perspective of a single node (gene), one way to summarise its role in the network is by looking at how well it is connected to all the other nodes. The ease with which one is able to reach other nodes from this node and over which paths are what defines that node within the network. It is this way of summarising nodes within a network that we use to compare individual nodes between networks. For each gene, we embed its relation to all the other genes within the network by computing for each gene pair how strongly they are connected to each other. We call such embeddings gene affinity profiles, where affinity is a metric that is higher if two genes are connected by more paths, stronger paths (consisting of higher weight edges), shorter paths or a combination of the three. Since edges and their weights in our networks represent a found predictive relation between genes' expression levels in the data, paths over these edges can be interpreted as potential regulatory relations between genes (Huynh-Thu et al., 2010). Embedding the relevant information of these nodes into these profiles provides us with a straightforward path to quantifying our comparison, since many methods for quantifying the difference between one-dimensional vectors exist.

Before we can start comparing these profiles between the two networks, there is one important step that still has to be taken before this comparison can be considered biologically meaningful. That is putting an emphasis on local connections in the gene affinity profiles. Earlier we stated that the paths over the network that are summarised by these affinity profiles can

[6]This view is based on our own understanding of genetic regulation and was confirmed by experts from the UMCG. These experts are G.H. Koppelman, M.C. Nawijn and M. Berg, see Appendix A.

be seen to represent regulatory relations. However, it is important to know that the likelihood that a path actually captures a true biological regulatory signal decreases sharply as its length increases. This is because it is unlikely that a gene has a regulatory link with another gene that has to pass through many other genes first[7]. Even in cases in which such long indirect links would exist, the way the networks are constructed in this study would not allow for such links to be reliably distinguished from false ones that are the result of noise. For these reasons, we choose to weigh the gene affinity profiles in such a way that comparing them emphasises differences in local connections.

The final step in this phase of the research is then to compare these now re-weighted affinity profiles for each gene between the network of the asthmatic group and the network of the control group. If the profiles of one gene differ much more than the others, this is an indication that the gene in question has disproportionately large differences in the way it is connected between the two networks to its close neighbours compared the other genes. In this study, these genes are referred to as candidate genes. From the way these candidate genes are identified, we can assume that the candidate gene and the genes in its neighbourhood interact differently in the case of asthma and that further inspection of these regions in the network might be promising for understanding the genetic drivers of asthma. See Section 4.4 for a detailed description of the network comparison methodology.

[7] See note 6

3 Data

The goal of this paper is to answer our research question and to gain insights into the genetic drivers behind asthma. To reach this goal, we analyse data on the genetic expression of asthmatic and non-asthmatic individuals provided to us by the University Medical Centre Groningen (UMCG). Over the span of multiple years, the UMCG has built up an extensive single-cell sequencing database containing data gathered from different subject groups and body tissues using multiple sequencing methods. Single-cell sequencing is relatively new way of collecting genomic data that allows for getting a much more detailed picture of the state of genetic expression in cells. In this study, we use a data set on the genetic expression of cells sampled in the airways of asthmatic and healthy human subjects. More specifically, bronchial brush biopsies where performed to "brush" cells from the surface of the airways (called the airway epithelium). The levels of genetic expression in these sampled cells were measured using the Smart-seq2 protocol (Picelli et al., 2014) which resulted in a data set containing the expression levels of genes in each individual cell. Our reasoning for choosing this data set is two-fold: Firstly, the tissue from which the cells are sampled, the epithelium, is identified as critical for the process through which asthma develops. Secondly, Smart-seq2 is known for its high robustness and reliability (X. Wang et al., 2021). It is shown to suffer less from the dropout problem and is generally less noisy than our other option: the more recent method called 10X genomics (Qiu, 2020; X. Wang et al., 2021). An important assumption underpinning this case-control type study is that the only structural difference in the data between the two groups is the presence or absence of an asthma diagnosis. We feel confident in making this assumption because the control group was carefully selected to allow for such a study and this data set has been used in other studies at the UMCG.

Working with single-cell data in this study came with a lot of challenges. This resulted in a lot of time going into getting the data in the right format and implementing our methodology. Please see Appendix B for details on the implementation, the challenges encountered and a link to the code repository.

After compiling the right data set for this research, we end up with the raw read count matrix and a table with metadata. The read count matrix consists of 3094 rows, each representing a different donor cell, and 32,012 columns, representing all of the genes in the human genome (*Human Genome Project FAQ*, 2020).

The entries of this matrix represent the genetic expression of a gene in one donor cells, with each value being the number of times the genetic material associated with that gene (RNA) is

detected in that cell. These counts range from 0 to 1,643,083 and a very large part of them (88.28%) are zeros. We will go into more detail on this zero-inflation later in the paper. The meta data that we have on cells consists of: a donor identifier, the health status of the donor (asthmatic or non-asthmatic), the sex of the donor, a batch identifier[8] and cell type label. We must note that we are warned by the creator of the cell type labels[9] to not rely completely on this labelling when discriminating between cell types known to be similar.

Before we can start analysing the data, we have deal with the fact that some parts of the data are of less interest to us or are of lower quality, as was described in section 2.1.1. As was done in Cole et al. (2019), we decide to filter out these genes and observations to reduce the noise in our data. An added bonus of performing a filtering step is that we also reduce the size of our data set, which reduces the computational burden of subsequent analyses.

We perform filtering of our data in two steps: cell filtering and gene filtering. These steps were performed in this order for because we find that it makes the most intuitive sense to remove low-quality cells before identifying and removing uninformative genes rather than the other way around and because this order was also the order chosen in Cole et al. (2019).

The filtering is done based on the read counts in the raw data. Each of the rules used was decided upon in consultation with experts from the UMCG. We ended up choosing to first filter out the cells in which less than 1000 or more than 8000 genes were detected and the cells that contained more than 20 percent mitochondrial RNA. After the cell filtering, we filtered out all genes that are not detected in at least ten cells and that are not highly expressed[10] in at least five cells. The gene filtering rules were inspired by the rules used in Cole et al. (2019).

As a result of the filtering, the size of our data set was reduced quite substantially. Table 1 shows some descriptive statistics of data set after filtering. We first see that the donors, the different sexes and the number of cells sampled are evenly distributed over both subject groups. We can also see that cells are split evenly over subject groups for the individual batches. An important effect of the filtering step is that we see that the number of genes per cell is now reduced from 32,012 to 14,548. The removal of these uninformative genes not only improves the average quality of our observations, it also greatly reduces the size of our data set. The percentage of zero entries in the data before filtering was 88.28 percent, which falls well in the 85 to 95 percentage range identified as typical for this type of data in Van Dijk et al. (2018). Table 1 shows that after filtering, this percentage is reduced to 75 percent with no large dif-

[8] This is an identifier of the sequencing run in which the measurement took place.

[9] This person is M. Berg, see Appendix A.

[10] A gene is considered highly expressed in a cell if the number of reads is higher than the upper-quantile of the non-zero elements of the count matrix (Cole et al., 2019).

ference between subject groups. Lastly, we see a difference between the subject groups in the average amount of genetic material detected in each cell (the average read count per cell). This difference in average read count and the likely presence of other observable unwanted differences leads us to choose to normalise the data to correct for these differences. This is discussed in detail in Section 4.1.

Table 1

Descriptive statistics of the filtered data set

	Asthmatic	Control	Full data
Number of donors	9	9	18
Males (Females)	7 (2)	7 (2)	14 (4)
Total number of cells	1666	1428	3094
Cells in 1^{st} batch	649	479	1128
Cells in 2^{nd} batch	467	256	723
Cells in 3^{rd} batch	468	619	1087
Cells in 4^{th} batch	82	74	156
Percentage zero entries	75.11	74.96	75.04
Average read count per cell	38.86	30.71	35.09
Genes per cell	14548	14548	14548

4 Methodology

This section gives an overview of the methods and techniques that we use to answer our research question. It consists of four subsections, each of which describes a different part of the study as shown in Figure 1. The first two subsections describe the data preparation steps of normalisation and data denoising. In the third subsection we explain how the networks are constructed. The final subsection describes how the networks are compared and how the results needed for answering our research question are obtained.

4.1 Normalisation

The comparability of our observations is one of the issues outlined in Section 2 that has to be dealt with before our data is ready for analysis. As a recap, the levels of the raw counts in our data might vary between observations not only because of differences in the level of genetic expression, but these differences might also be partially driven by other factors. We choose to normalise our data in order to mitigate these effects. See Section 2.1.2 for a more detailed motivation of why we normalise our data. In this section we describe how we choose the normalisation method and give a description of how this method works.

4.1.1 Choosing the normalisation procedure

The choice of which method to use for the normalisation of our data is not trivial task since no single best normalisation procedure for single-cell data exists and procedures have been shown to perform differently depending on the data at hand (Cole et al., 2019; Vieth et al., 2019). For that reason, we choose to use the approach outlined in Cole et al. (2019) to systematically compare different normalisation procedures and select the best performing configuration based on objective metrics.

We compare a total of sixteen different configurations and score them using five performance metrics. In this section, we only report the main results of this comparison. A detailed description of the complete comparison and the results can be found in Appendix C.1. The results of this comparison lead us to conclude that that applying full-quantile (FQ) normalisation (Bolstad et al., 2003) without batch correction is the best choice for our data set, proving to be a top performer for four out of the five metrics. The fact that FQ normalisation scored the best on the metric for preserving biological variation is of particular importance because our ultimate

goal is to compare the genetic expression of the two subject groups. We find it encouraging that method that proved best in our comparison is the same as the one found for the Smart-seq data set in Cole et al. (2019).

Having identified FQ normalisation without batch correction as the best method for our data set, we proceed with the data normalised using this method for the remainder of this paper. This data set is from here on referred to as the normalised data set. By comparing the descriptive statistics before and after normalisation, we conclude that the normalisation had the desired effect and that it was implemented correctly. See Appendix C.2 for a description of this comparison. A brief description of how this method works is given in the section below.

4.1.2 Full-quantile normalisation

In this section we will provide a brief explanation of the idea behind FQ normalisation and how it works. This description is based on the one by Bolstad et al. (2003), please see the original paper for a more detailed description.

What full-quantile normalisation strives to make the distribution of the observed gene counts the same for all the cells. The idea behind this method comes from the quantile-quantile plot, which shows a straight diagonal if the distribution of the values in the two compared vectors is the same. FQ normalisation generalises this concept to n dimensions, with n being the number of cells in our data set. The n dimensional analogue for this straight diagonal in two dimensions is the line along the line given by unit vector $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$. All of the n gene count vectors of our sampled cells can be said to have the same distribution if plotting their quantiles in n dimensional space gives this straight line (Bolstad et al., 2003). This brings us to the core idea of this method, which is to make all observations have the same distribution by projecting their quantiles onto this diagonal (see Bolstad et al. (2003) for a formal description of this idea of projecting the quantiles).
The following algorithm achieves this projection and thus normalises the gene count vectors by giving them the same distribution (Bolstad et al., 2003):

1. let X be the $p \times n$ read count matrix where each gene is a row and each cell is a column;
2. sort each column of X to get X_{sort};
3. calculate the means of all the rows in X_{sort} and assign this mean to all elements of that row to get X'_{sort};
4. get $X_{normalised}$ by rearranging each column of X'_{sort} to have the same ordering as the

original X.

A simple way of understanding the above algorithm, and thus FQ normalisation, is that it replaces the maximum value of each cell by the average of the maximum values of all cells, and the same for the second highest value, third highest value and so on. An important note on our implementation of FQ normalisation is that after having obtained the normalised data matrix, we set all entries that had been zero in the raw data to be zero too in the normalised matrix. This is done to reduce the risk of introducing false signals in the data.

4.2 Denoising

The final step of the data preparation process outlined in Section 2 consists of dealing with issues relating to data noisiness. Put briefly, our data is highly sparse and contains observations on different types of cells with expected heterogeneous expression patterns. Both of these aspects of the data are sources of noise that can distort the signals in the data. In order to reduce the level of noise in our data and combat these two sources of noise, we take three specific steps to denoise our data: Firstly, we focus our analysis on the most highly varying genes. Secondly, we impute the data to reduce sparsity. Thirdly, we cluster observations based on cell type and remove undesirable observations. The first two of these aim to combat data sparsity and the third one tackles the issue of cellular heterogeneity. In the following subsections, we first explain separately what each of the three denoising steps entails. After this, we motivate our choice for the chosen method to implement these steps and a description of how this technique works. Lastly, we explain how the output of this step is used to filter out undesired cell types.

4.2.1 Focusing on most highly variable genes

The first way we tackle data sparsity is by removing all but the 2,000 most highly varying genes. The assumption here is that highly variable genes are more informative and carry more information. By focusing on this subset we improve the signal-to-noise ratio of the data. Focusing on the most informative genes in this way is common practice when working with this kind of data and the number of genes included is often chosen to strike a balance between including as much information as possible and computational feasibility. Pratapa et al. (2020) showed that using more genes generally results in better networks but also warned that methods often become too computationally expensive, especially beyond 1,000 genes. A best-practice guide for working with scRNA-seq data recommended using between one and five thousand highly variable genes (Luecken & Theis, 2019). We opted for selecting the 2,000 most highly varying genes as this falls nicely in this range. See Appendix D.1 for a description of the effects of this

step on our data.

4.2.2 Data imputation

The second way we reduce data sparsity is by imputing the data, that is: we try to correct the data to mitigate the effect of wrong entries. In the case of our data set, these wrong entries are most often the dropout zeros. Imputation of scRNA-seq data to reduce sparsity has been common practice for many years in this type of research and many different methods have been proposed, each with its own advantages and disadvantages. An important aspect of imputation that we want to note is that of the slight difference in the meaning of the word imputation in a general statistical context and in the context of scRNA-seq data. In general statistical context, imputation often refers to the process of replacing missing data values with substituted values. In that case, the missing values are easily identifiable in the data. In the context of scRNA-seq data, the goal is still to replace missing values. However, the difference here is that the missing values (or dropouts) are represented in the data by a zero and are thus not easily distinguishable from true zero values.
This problem of identifying the missing values adds an extra layer of difficulty when trying to "correct" the data. As a result of this, imputation methods specifically developed for scRNA-seq data are unable to just substitute the values of the missing data points. Many of these methods resort to approaches where they transform the entire data set and thus also modify other entries. When done right, such transformations improve the signal-to-noise ratio, do not introduce false signals and remove as little signals of interest as possible in the data.

For the imputation in this paper we use the method called single-cell Graph Neural Network (scGNN; J. Wang et al., 2021). This method is also for the cell-type clustering described in the following section. It is because of this combined purpose of this method that in this and the next section we first describe how its output is used for our research purposes. See Section 4.2.4 for our motivation for choosing this method and Section 4.2.5 for a description of the method.

4.2.3 Cell-type clustering

The last way we reduce the noise in our data is clustering the observations on cell type and subsequently focusing our analyses on cells of specific types. We chose in this paper to focus on basal and secretory cells, and thus to remove the other cell types from our data set after having identified them through clustering. These two cell types are chosen for a number of

reasons: Firstly, these two cell types are chosen because are expected to show very similar genetic expression profiles. The reason for this is biological. We are told by an expert from UMCG[11] that cells of these types exist on a spectrum where one type can transition into the other and no clear line of separation exists. This makes distinguishing between these two types based on their genetic expression a task that is not only difficult but also not necessary for the purposes of this research, since both cell types are expected to have similar expression profiles. Thus having both types in the data is not expected to increase the level of noise by much. Secondly, these two cell types are expected to represent largest fraction of the data[12]. Thus focusing on these cell types means ignoring the least amount data. Thirdly, secretory cells in particular are said to be relevant cells to analyse when researching asthma (Erle & Sheppard, 2014). Lastly, although they are not easily distinguishable from each other, they are expected to be distinguishable from other cell types based the expression data.

Our data set already contains an initial labelling of the cells, which was created based on the expression levels of certain marker genes. We use this labelling for identifying and filtering out the more clearly distinguishable cell types, namely: immune cells and ionocytes. For filtering out the remaining undesirable cell type, the ciliated cells, we choose not to rely solely on this initial labelling because of its questionable accuracy for some of the border cases between ciliated and secretory or basal cells. For this reason we choose to use the clustering constructed and outputted by our imputation method (scGNN), see This is why we choose to construct our own cell-type clustering. This clustering is constructed using the same method that we use for the imputation (scGNN, see Sections 4.2.4 and 4.2.5).

4.2.4 Choosing the clustering and imputation method

We employ scGNN as was described in J. Wang et al. (2021) for both clustering the cells and imputing the data. Using this approach, we simultaneously deal with the increased sparsity caused by dropout zeros and are able to construct a better cell type clustering for isolating the secretory and basal cells. A useful feature of scGNN is that it is designed in such a way that the found cell clustering can enhance the imputation performance and vice versa. See Section 4.2.5 for a description of how scGNN works. The existence of efficient implementations of this method for usage on large datasets combined with its ability to have imputation and clustering benefit from each other provide two initial strong arguments for choosing this method. Because employing scGNN serves two different purposes in this paper, our motivation for choosing this

[11] This person is M.C. Nawijn, see Appendix A.
[12] Approximately 62 percent based on the initial labelling.

method can be also split into imputation and clustering specific arguments. These arguments are described in the remainder of this subsection.

Our reasons for choosing scGNN for imputing our data are the following: The first reason is that the method does not assume any statistical distribution or relationships for the gene expression data or the dropout events. This makes it an approach to both imputation and clustering that is hypothesis free, which is a desirable property when doing this type of research. This distribution free property is not shared by many of the other methods, with for example scImpute (W. V. Li & Li, 2018), SAVER (Huang et al., 2018) and DCA (Eraslan et al., 2019) all making one or more distributional assumption. The second reason is because we believe it to be the most promising of the deep-learning based methods. This class of imputation methods has seen many publications in recent years and has become one of the most prominent in the domain of single-cell data imputation (Arisdakessian et al., 2019; Lopez et al., 2018; Eraslan et al., 2019; Talwar et al., 2018). These methods all make use of either (deep) autoencoders or some other type of neural network and often require very few distributional assumptions. What distinguishes scGNN from other methods in this class is that it makes use of additional information from the data to regularise the autoencoder, this can increase its ability to improve the signal-to-noise ratio in terms of embedding biologically meaningful information. This additional information is extracted from data and consists of the output of a Left Truncated Mixture Gaussian (LTMG; Wan et al., 2019) model to incorporate information on gene regulatory signals, inferred cell-cell relationships from the clustering procedure and inferred cell types from the clustering procedure. Lastly, in J. Wang et al. (2021) it is shown to outperform other popular imputation methods, among others: MAGIC (Van Dijk et al., 2018), SAVER (Huang et al., 2018), scImpute (Huang et al., 2018), scVI (Lopez et al., 2018), DCA (Eraslan et al., 2019) and DeepImpute (Arisdakessian et al., 2019).

Our choice for using scGNN to cluster our observations is because of the following three reasons. Firstly, this clustering technique is specifically designed for clustering scRNA-seq data by cell type, which is not the case for many of the alternatives. These alternative methods for clustering consist of applying common clustering methods on the expression data directly (e.g. k-means or the density based DBSCAN). Secondly, rather than clustering on the (normalised) expression counts directly, the method uses inferred cell-cell relationships from a cell graph to obtain clusters. We expect this biologically motivated approach to outperform traditional clustering methods in our context as the method clusters on embeddings of biologically meaningful

information relevant for inferring cell-cell relationships. Thirdly, experiments show significant improvement in cell clustering performance compared to existing scRNA-seq analytical frameworks (for example Seurat with the Louvain algorithm) (J. Wang et al., 2021).

4.2.5 Single Cell Graph Neural Networks (scGNN)

In this section, we describe how scGNN works. We intentionally do not go into too much detail as this method is not the main focus of this thesis. See J. Wang et al. (2021) for a complete description of the inner working of this method.

As input, it takes the filtered and normalised expression data from the previous step. This data is then ln(1+x) transformed and the 2,000 most highly varying genes are selected. The resulting pre-processed input matrix serves as input for three separate parts of the scGNN architecture: the Left-truncated mixed Gaussian (LTMG) model, the scGNN iteration cycle and the imputation autoencoder. This architecture is represented schematically in figure 2a and we will briefly explain each of these parts.

At the start of the scGNN method, an LTMG model (Wan et al., 2019) is fitted on the pre-processed data to quantify gene regulatory signals over different transcriptional regulatory states. The expression level of each gene in every single cell is modeled as a mixture of several left-truncated Gaussian distributions. The left truncation of the distributions allows the model to better handle zero entries (both true zeros and dropouts) and low expression values. The different distributions correspond to the heterogenous gene expression states among cells that serve as an approximation of the gene's varied transcriptional regulatory states (Wan et al., 2019). After the model is fitted, each expression value in the data is assigned to the distribution with the highest probability. It is this assignment that serves as the main output of the LTMG model for scGNN and is used as regularisation in both the feature and imputation autoencoders through penalising the loss function, which causes it to treat each gene differently.

The data is then sent through an iteration cycle consisting of three auto-encoder based steps: In the first step, the input expression matrix is encoded to a lower feature space and subsequently reconstructed using the feature autoencoder. As can be seen in figure 2b, the output dimensions of the first and second layer of the encoder are 512 and 128, and 128 and 512 for the decoder. In the encoder, the ReLU activation function is used, whereas the decoder employs a sigmoid activation function. The loss function of the feature autoencoder contains a

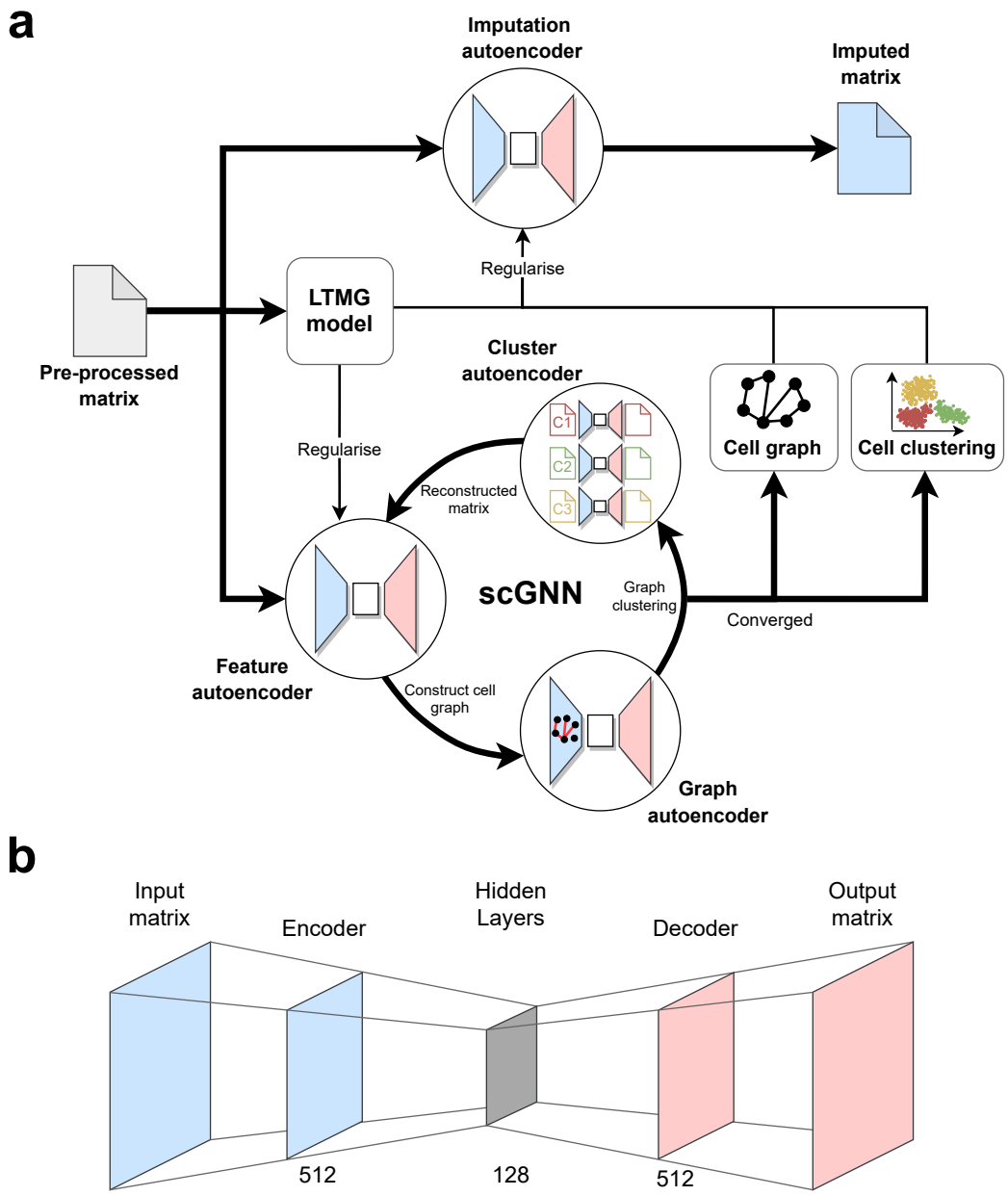


Figure 2. Subfigure **a** shows the architecture of scGNN as a whole. Subfigure **b** shows and the architecture of the feature, imputation and cluster autoencoders. The output dimensions of the encoder and decoder are 512 × 128 and 128 × 512, respectively. This figure was based on figures 1 and 2b in J. Wang et al. (2021).

regularisation term with the output from the LTMG model. The embedding found by the encoder is used to construct a k-nearest neighbor graph based on the Euclidean distances between cells. The graph is pruned using the Isolation Forest algorithm (Liu et al., 2008) to obtain a more biologically meaningful graph.

The adjacency matrix from the pruned graph and the learned embedding from the feature en-

coder are used as inputs for the graph autoencoder. This autoencoder finds a low-dimensional representation of the topological information of the pruned graph, which is then used to cluster the cells using k-means.
In the next step, the reconstructed expression matrix from the feature autoencoder is split according to the clusters found using the graph autoencoder. Each of these cell-clusters is then encoded and decoded using separate cluster autoencoders. These autoencoders have the same architecture as the feature autoencoder, only without the regularisation term from the LTMG model. The other difference with the feature encoder step in the iterative process is in this cluster based step the reconstructed matrix out of the cluster autoencoders serves as the output rather than the low-dimensional feature embedding. The full reconstructed matrix is the concatenation of the results from all autoencoders belonging to different clusters and this matrix serves as the new input for the feature autoencoder in the next cycle.
This cycle is continued until convergence is observed in the constructed cell graph and the cell clustering. When this point is reached, the resulting cell graph and cell clustering from the last iteration serve as output of the iteration cycle.

Once the iteration cycle has converged, the original pre-processed matrix is run through the imputation autoencoder which both imputes dropout zeros and denoises the data using the found relationships between the cells. This imputation autoencoder has the same architecture as the feature autoencoder (Figure 2b). In addition to the gene regulation regularisation term from the LTMG model, the loss function of the imputation autoencoder also contains three additional regularisation terms. The first two of these terms come from the output of the iteration cycle. One is based on the constructed pruned cell graph. For every cell, this regularisation term makes cells that are within one edge of that cell in the graph carry extra weight in the imputation process. The other one is based on the cell clustering. In a similar to way the previous regularisation term, it makes cells within the same cluster carry extra weight in the imputation process. The third additional regularisation term is an L1 regularisation term which introduces more sparsity in the model by reducing the number of non-zero weights and hereby increases the generalisation performance of the autoencoder.
The final output of the entire procedure that is used for subsequent steps in this research consists of both the imputed expression matrix and the clustering of the cells. In Appendix D.2, we describe the effects of this imputation step on our data as well as the results of the validation tests that we performed to check if this step did not distort the data.

4.2.6 Cell-type filtering

We use both the output of scGNN and the initial cell-type labelling to remove cell types that are not the focus of this study. Through comparing the output clustering with the initial cell-type labelling we also validate the performance of our clustering method. We do this by looking at the distribution of these labels over the new found clusters. If our clustering method performed well, we would see the clusters' ability to discriminate between ciliated and secretory and basal cells to highly correspond to that of the initial labelling. Or put more concretely: when our new found clusters are dominated either by cells initially labelled as ciliated or of cells initially labelled as secretory or basal. We perform the cell-type filtering by keeping only clusters that are dominated by cells initially labelled as secretory and/or basal. From that point on we consider all cells in the remaining data set to be either secretory or basal, which are the focus types of this research. As stated before, a further distinction between these types based on genetic expression is neither necessary nor biologically feasible. Since the initial labelling was said to be less trustworthy in some edge cases, we would expect a small proportion of cells to be labelled differently by this procedure than they were labelled initially. Later inspection of the output confirmed this expectation and we refer to Appendix D.3 for a detailed description of the results of the cell-type filtering step.

4.3 Network inference

After our data is fully prepared for analysis, we continue with inferring two gene regulatory networks from this data, one for the asthmatic and one for the control group. As outlined in Section 2.2, the method used for the construction of these networks has a big influence on the type of relationships and interactions captured in the network. In the following, we will first provide an overview of which different approaches to network inference exist, then describe why we chose the method used for this study and lastly describe how this method works.

4.3.1 Overview of the different methods for network inference

The specific type of relationship captured in network depends on the method used for its construction. The inference of these networks from expression data has been an active area of research for over 20 years (Pratapa et al., 2020). The introduction of single-cell data saw a big increase in researchers' capabilities to reconstruct GRNs. In their 2019 survey on GRN inference, Huynh-Thu & Sanguinetti classify these methods into two general classes: probabilistic models and data driven methods.

Probabilistic network models

Methods for network inference in this class start from an explicit probabilistic model of the data and construct the networks based on this. The two most common network types in the probabilistic class are Gaussian graphical models and Bayesian networks. The edges in Gaussian graphical models represent partial correlations between the level of expression of genes, but this type of network has only seen limited usage in biological research. Bayesian networks, in which the edges represent conditional dependence, have been much more popular for constructing biological networks (e.g. GRNs and protein interaction networks) and have even seen usage years before the availability of single-cell data (N. Friedman et al., 2000). This is in part because of the ease with which prior information can be incorporated. The downside of these probabilistic models is that they rely on formulating an explicit model of the data in terms of probabilities, which requires making numerous assumptions. A downside specific to Bayesian networks is that they do not allow for cycles in the constructed network, which is unfortunate given the high prevalence of cyclic processes in biological system. Lastly, constructing a Bayesian network is notoriously expensive and proven to be an NP-hard problem in Cooper (1990). Although recent advances in efficient implementations have tried to combat this problem, these networks are still less practical to use on the larger data sets that are common in single-cell research like ours.

Data driven network models

Unlike the probabilistic network models, models in this class start with assuming a fully connected network and estimate the weight on each edge directly from the data. The result from this approach is a weighted network from which the final network topology can be obtained through careful thresholding (Huynh-Thu & Sanguinetti, 2019). Networks constructed using data driven methods can be subdivided into three types: correlation networks, relevance networks and regression networks (Huynh-Thu & Sanguinetti, 2019).

Correlation networks use standard correlation measures to estimate the associations between genes. The often linear nature of statistical correlation means that it is unlikely to capture the more complex relations between genes and these networks often have little predictive power due their inability to distinguish between direct and indirect interactions. One notable exception to this linear constraint is presented in Guo et al. (2014), where they infer networks using non-linear distance correlation. Due to their low computational cost and easy of construction, these correlation networks are still quite common in research.

Relevance networks are able to capture more subtle dependencies than correlation networks by

calculating the degree of dependence between two genes based on mutual information. Methods in this category are among the most widely used in literature, with a recent publication in field of asthma by Banerjee et al. (2021) making use of a partial correlation and information theory (PCIT) algorithm (Reverter & Chan, 2008) and the PIDC (partial information decomposition and context) algorithm (Chan et al., 2017) being among the top performers in study comparing network inference methods (Pratapa et al., 2020). They are only slightly more computationally intensive than correlation networks. However, these mutual information based techniques fall short of capturing predictive relations of genetic expression (Huynh-Thu & Sanguinetti, 2019). Also, when the sample size is medium to small, the estimation of the joint probabilities needed for the construction of the networks might be highly sensitive to noise.

The last type, the regression network, has seen a huge increase in popularity in recent years as the development of highly scalable implementations allowed them to be used effectively on the huge datasets that are common in scRNA-seq research. These networks try explain the expression of one gene by regressing it on all the others. This approach allows for the construction of weighted directed graphs that even have some predictive capability. Although regression networks can be more computationally expensive to construct than correlation and relevance networks, it makes up for that in its ability to capture more complex relations, whereas correlation and relevance networks only focus on pairwise dependencies. The fact that they require much less assumptions than probabilistic network models also makes regression based methods more attractive. For these reasons, it is not surprising that regression based network inference methods are among the most popular and scalable approaches for reconstructing GRNs (Huynh-Thu & Sanguinetti, 2019).

4.3.2 Choosing the network inference method

In this study, we choose to use GRNBoost2 Moerman et al. (2019) for the inferring the gene regulatory networks from our data. Our reasons for choosing this method are given below.

We choose GRNBoost2 because the literature on network inference indicates that is not just one of the best regression based methods for network inference, but also one of the best performing methods overall. The fact that it is a regression based network inference method is good because these methods require much less assumptions than probabilistic network models (see previous section). Regression based methods also stand out from the other methods that require few assumptions because they are able to capture high-order conditional dependencies

between gene expression patterns (Huynh-Thu & Sanguinetti, 2019). Due to the way that they are constructed, these networks have some predictive capability, meaning that given a subset of genes, one can predict the expression levels of the remaining genes (Huynh-Thu & Sanguinetti, 2019). These networks are also able to capture directional relationships and can handle cycles (Huynh-Thu et al., 2010). Given the common occurrence of feedback loops in biology, having a network that can handle cycles is a desirable property.

GRNBoost2 is the latest development of a line of inference methods in the regression network class that inherits much of its high performance from its predecessors GENIE3 (Huynh-Thu et al., 2010) and GRNBoost (Aibar et al., 2017). Like its nine years older predecessor GENIE3, GRNBoost2 improves on the idea of the regression network by replacing the linear regression model with a more flexible set of regression trees. This enables the method to also capture non-linear relations between the expression levels of different genes, which allows the resulting networks to paint a richer picture of the interactions between genes. What makes GRNBoost2 an improvement upon its predecessors is that it is a much faster and more scalable implementation of the same idea that achieves the same results. Through the use of gradient boosting and other computational improvements (see Section 4.3.3), this method is much better able to deal with data from high-throughput gene profiling technologies, like single-cell RNA-seq, for which it was specifically developed.

A common way to get a sense of how a network inference method performs is by running it on the data from the DREAM5 competition (a competition in which biological network inference algorithms competed against each-other) and comparing its performance to that of the other contestants. In such a comparison, GRNBoost2 showed excellent performance equal to its predecessor GENIE3 (Moerman et al., 2019), who not only was the top performer in the DREAM5 competition (Marbach et al., 2012) but also in the earlier DREAM4 competition (Huynh-Thu et al., 2010). An extensive comparison of network inference algorithm by Pratapa et al. (2020) identified GRNBoost2, GENIE3 and PIDC as the gene regulatory network inference methods of choice. Their tests showed these three consistently being the best performers in terms of accuracy. These test also showed that GRNBoost2 was much more computationally efficient than PIDC and GENIE3 and also to be less sensitive to dropouts than PIDC and GENIE3 (Pratapa et al., 2020). It was also shown that GRNBoost2 consistently produced nearly identical outputs when ran on the same data set multiple times (Pratapa et al., 2020), which is a desirable property for methods with a stochastic element like GRNBoost2 to have.

Based on these arguments, we conclude that GRNBoost2 is the best option to use as network inference method in this research.

4.3.3 GRNBoost2

In this section, we provide a description of how GRNBoost2 works. This description is divided into three parts: First, we describe the main structure of the GRNBoost2 approach to network inference. Secondly, we provide a slightly more detailed description of how the edge weights are estimated in GRNBoost2. Lastly, we describe the networks outputted by the method and the thresholding step needed to obtain the final networks such that it can be used in downstream analyses. Like in Section 4.2.5, we do not go into full detail of the exact inner workings the method. For a detailed description we refer to the original papers introducing GRNBoost2 (Moerman et al., 2019), and its predecessors GENIE3 (Huynh-Thu et al., 2010) and GRNBoost (Aibar et al., 2017).

Main structure

The goal of GRNBoost2 is to infer gene regulatory networks based on gene expression data. The this is done is by using the data to assign weights to assumed regulatory links from a gene to another gene, with the goal that high value weights correspond to actual regulatory interactions (Huynh-Thu et al., 2010).

The idea of GRNBoost2 is to determine separately for each gene which of the other genes serve as its regulators. It defines the identification of such regulatory relations as the problem of determining the subset of genes whose expression levels directly influence or can be used to predict the expression of the target gene (Huynh-Thu et al., 2010). In other words, it approaches it as a problem of determining feature importance.

Being a regression based GRN inference method, GRNBoost2 achieves the construction of a network of the regulatory relations between p genes by dividing it into p different regression problems. In each of these sub-problems, the expression level of one gene is predicted from the expression levels of all the other genes in the data set. Because this approach involves the training of p prediction models, the Huynh-Thu et al. stressed the importance of choosing an algorithm that is fast and requires little manual tuning. This method uses the Gradient Boosting Machines (GBM; J. H. Friedman, 2001) implementation from the XGBoost library (T. Chen & Guestrin, 2016) for each of these regression problems (Aibar et al., 2017). The GBM model trained for each sub-problem uses boosting (Freund et al., 1999) to combine multiple shallow

decision trees to form one strong prediction model. These GBM regression models are a great choice for this purpose because in addition to being fast and not requiring any hyperparameters to be set manually, they also make no assumptions about the target function, are able to deal with interacting features and non-linearity and have straightforward ways of computing feature importance. The type of GBM used in GRNBoost2 employs an early-stopping regularisation strategy to ensure that each regression model contains only just enough trees. Trees that do not seem to show an improvement of the model are aborted early to prevent unnecessary computations. Estimates from the original paper state that around 80% fewer trees are constructed using this approach than when using its predecessor GENIE3, which employs Random Forests without this regularisation (Moerman et al., 2019). This increase in efficiency that using GBMs with early-stopping regularisation provides over Random Forests is a very useful, if not necessary, improvement given that to the large scale adoption of single-cell sequencing data has come with much higher volumes of data. In addition to the improvements described above, GRNBoost2 comes with an efficient ready-to-use Python implementation of the method. Because each prediction model is built for each of the genes seperately, the method is highly parallelisable by nature. The GRNBoost2 algorithm greatly exploits this parallelisability to achieve faster performance along with some other computational tricks implemented using the Dask (Rocklin, 2015) parallel computing library for Python. As we described in Section 4.3.2, the increased efficiency of GRNBoost2 comes at almost no cost in terms of performance when comparing it to GENIE3, with both methods performing equally well in tests.

Having trained a prediction model for the expression of one of the genes, it then computes the importance of the features from the prediction model to serve as the weights in the network. We explain how this feature importance is computed in the next section. The final network is realised when the p sets of importance scores are combined to form one network. This procedure is summarised by the following algorithm:

1. for each of the p genes in the expression data set:
   - train a GBM regression model for predicting that genes' expression level using the expression levels of the other $p - 1$ genes as predictors;
   - compute the feature importance of all the predictors in the model;
2. combine the feature importance scores of the p different models into one network.

Estimating edge weights through feature importance

As can be seen in the algorithm given above, the feature importance of the predictors is calculated for each of the p GBM regression models. Sławek & Arodź (2013) showed that, like the

Random Forests used in GENIE3, GBM regression models can also be used for estimating feature importance in a GRN inference context (Moerman et al., 2019). To do this, the following approach is used:
First, they calculate $I(\mathcal{N})$ using (1), which gives the total reduction of the variance of the output variable at each node in a tree caused by the split:

$$I(\mathcal{N}) = |S|Var(S) - |S_t|Var(S_t) - |S_f|Var(S_f) \tag{1}$$

Here, S denotes the set of the samples that reach node $\mathcal{N}$, S_t (resp. S_f) denotes the subset of S for which the test is true (resp. false). $Var(.)$ denotes the variance of the output variable in a subset and $|.|$ denotes the cardinality of a set of samples (Huynh-Thu et al., 2010).
Then, the importance of a particular feature in a single tree is computed by summing up the I values of all the nodes in that tree in which that feature is used to split. A feature is given an importance value of zero if it is not used in any of the nodes.
Finally, the feature importance of a predictor over all trees in one of the GBM regression models is calculated by taking the average of the importance scores of that predictor over all trees in that model.

An important consequence of calculating feature importance in this way is that the scale of the importance scores found in one of the prediction models is linked to the variance of gene predicted by that model. If not properly addressed, this will introduce a bias in the found regulatory links in favour of highly variable genes. To avoid this bias, the gene expressions are normalised to all have unit variance before training the prediction models. This normalisation ensures that the weights found in the p different regression problems are comparable and can be safely combined into one network (Huynh-Thu et al., 2010).

The final step in the construction of the network is to combine all feature importance scores for all the genes into one network. In this process, a directed edge from a source gene to a target gene is constructed if the feature importance of the source gene in predicting the expression of the target gene was found to be non-zero. The weight given to this edge is the feature importance score that was found. The resulting network is thus a weighted directed network.

Obtaining the final networks through setting an edge weight threshold
After running GRNBoost2 on our data, we end up with two networks, one for each subject

group. These networks consist of a set of nodes and a set of directed weighted edges. The nodes correspond to the set of genes in our post-imputation data set. The edges correspond to the regulatory links found by GRNBoost2 between the genes using the feature importance approach described earlier.

As was noted in both Huynh-Thu et al. (2010) and Huynh-Thu & Sanguinetti (2019), an important final step when construction GRNs using data based methods like GRNBoost2 is thresholding the networks outputted by the method. As any link with non-zero weight found by the method is included in these networks, the resulting networks will contain a very high number of edges when no cutoff value for minimum edge weight is chosen. Huynh-Thu et al. (2010) stress the importance of implementing a cutoff before using the networks for any practical purposes as it will likely be filled with noise without it. The question of what this threshold should be is no trivial one and should be treated with great care, as the inclusion or exclusion of edges can potentially have great effects on downstream analyses. As we will explain in Section 4.4, our approach to comparing the networks after setting this threshold is designed to be as robust as possible to this choice. In addition to designing our analyses as robustly as possible, we also check if this is really the case by running our analyses for a large range of possible cutoff values and only report results that we reliably find over this range. The cutoff used in this study is presented in Section 5.1 as well as the motivation as to why this value was chosen.

4.4 Network comparison

The final step in our analysis is the comparison of the constructed networks. The goal is to perform this comparison in such a way that it is most likely to identify biologically meaningful differences in the genetic interactions between the asthmatic and control groups. As stated in Section 2.3, it is the links that genes in the networks have with genes in their close vicinity in the network that have the highest likelihood of capturing the underlying regulatory signals. Differences in these local connections therefore have the highest likelihood of capturing biologically meaningful differences[13]. This is why our approach to comparing the networks is designed to identify so-called candidate genes whose neighbourhoods in both networks display the greatest difference in their internal connectivity. This is done through embedding each of the nodes into its own gene affinity profile over the network. These profiles are a self-developed type of embedding that represent the nodes in a GRN as one-dimensional vectors. Comparing each gene's profile in the asthmatic network with its counterpart in the control network allows us to

[13]This view is based on our current understanding of genetic regulation and was confirmed by experts from the UMCG. These experts are G.H. Koppelman, M.C. Nawijn and M. Berg, see Appendix A.

estimate how large the difference in the internal connectivity of the neighbourhoods is between the networks. The goal is to subsequently compare these estimates between the different genes to gauge which genes' neighbourhoods show the largest difference and can thus be identified as candidate genes. The candidate genes and the genes in their neighbourhoods can be used as the starting point for more targeted and detailed analysis of the interaction structures.

The remainder of this section is built up as follows: First, we describe in detail what these affinity profiles are and how they are constructed. We then explain how these profiles are compared between the networks and how the outcomes are compared with each other. Finally, we describe how these results will be used to answer our research question and how they can be used to obtain biological insights.

4.4.1 Constructing gene affinity profiles

In this section we describe how we embed the relevant information of genes in our networks into one-dimensional vectors called gene affinity profiles. As is often the case when embedding information, this embedding does not achieve perfect retention of all of the information in the original format, nor is that the goal of the embedding. Rather, the goal is to capture the relevant information for each of the genes in the network in a separate data structure that allows for easy and meaningful comparison. Achieving this goal can be seen as to consist of three parts, which make up the three steps in the construction of the gene affinity profiles: The first of these is mapping the relevant node information to a one-dimensional vector. This is based on the idea that the relevant information of the genes in the networks can be captured by how well they are connected to the other genes in the network. This is achieved by computing its distance to all other nodes in the network. The second is transforming these vectors in such a way that they are comparable between different genes. This is done by converting the distance vector of each gene into a relative distance vector. The third and final step is a transformation that ensures that comparing the profiles emphasizes differences in local connections. This is done by transforming the relative distances into similarities (or affinities) using the non-linear Gaussian radial basis function (RBF; Vert et al., 2004) kernel. The following two subsections describe each of these steps in detail.

Distances over the network

The first step in constructing the gene affinity profiles is computing the distances between all genes over the networks. The resulting distance vector for each gene indicates the ease with

which one can reach each of the other genes in the network from that gene. As stated before, the assumption that we make here is that embedding nodes based on their distance profile in the network captures the relevant information for detecting biologically meaningful differences. In Section 2.3, we defined a biologically meaningful difference as when we the same genes in different networks are connected differently in terms of the genes in their local vicinity. We feel confident in making this assumption for two reasons: First, because these distance profiles can be used to assess whether there is a difference in how well a gene is reachable from another gene, which we can use as a proxy for measuring whether or not genes are connected differently. Secondly, because these distance profiles include the distances from one gene to all other genes. This means that, by definition, also this information for the genes in close vicinity is included.

For computing these distances, we use the effective resistance (ER) distance between nodes as was described in Ellens et al. (2011). The intuition behind this metric is that it treats the graph as an electrical circuit with resistors between the nodes. The electrical resistance of these resistors (i.e. the edges) is based on the original edge weights. In the case of our networks, these resistances are defined as the reciprocal of the edge weights (Ellens et al., 2011) since we want high weight (high predictive power) to correspond to low resistance. The distance between two nodes is then computed as the total electrical resistance over this circuit if a current was flowing between these nodes. This distance metric thus considers all possible paths between two points. In that way, it can be seen as a way to measure how well a pair of nodes are connected in the network, this is what we refer to when we talk about connectivity in this paper. The effective resistance was already proven to be a proper distance function in Klein & Randić (1993), where the authors showed that it satisfies the non-negativity, symmetry and triangle inequality conditions. Our reason for choosing to use effective resistance instead of the much better known shortest path distance is because it is a better metric for connectivity over a network and because of its robustness to adding or removing edges. Rather than giving the length of the single shortest path, the resistance distance considers all possible paths between two nodes and in doing so also measures the robustness of the reachability of one node from another (Wills & Meyer, 2020). This robustness is an important property for this study because, as noted in Section 4.3, we have to choose a threshold below which we remove all edges from the network. Even minor deviations in the choice of this threshold can mean that a great many edges are included or excluded from the network. We want the impact of small changes in the choice of this threshold to also be small. Using a brief example, we will intuitively show that this is the case for the resistance distance, but not for the shortest path distance. Imagine a

connected graph that does not contain a direct edge between nodes a and b. Adding an edge connecting a and b could have a large impact on the shortest path distance, since this edge could replace the current shortest path if it is a faster alternative. The effect on the effective resistance of adding such an edge is much less. This follows from Lemma 4.7 in Ellens et al. (2011), where we see that if the initial ER distance between the nodes is R_{ab} and the newly added edge has resistance r_{ab} then the new ER distance between the nodes becomes:

$$R'_{ab} = \frac{1}{\frac{1}{r_{ab}} + \frac{1}{R_{ab}}} \tag{2}$$

We see that adding a new edge adds a new possible path from a to b, which is now included in the new ER distance. But unlike for the shortest path distance, all other possible paths are also still part of the computation.

Before we can compute the distances between nodes over the networks, we have to make sure our networks satisfy two conditions: Each of the networks needs to be undirected and connected.

The networks outputted by GRNBoost2, our network inference method, are directed (see Section 4.3.3). In our network comparison step however, we convert these networks to undirected networks. The reason that we chose to work with undirected networks has to do with the implementation of the effective resistance metric. This metric was developed for usage on undirected graphs, which also best fits the electrical circuit analogy on which the metric is based. Efforts have been made to extend the concept to directed networks, most notably in Young et al. (2015a,b). Because our initial networks are directed we have spend considerable time finding ways to reliably implement a measure of effective resistance on directed graphs. We found however that extending the concept to a directed graph represents quite a big leap from the original idea and that we could find no examples where this is applied to weighted directed graphs. Thus choosing to maintain directionality in our networks would mean that we have to let go of the weights. In the context of our networks, we felt that the computed edge weights were a more valuable and reliable piece of information than the directionality of the edges. This is supported by the fact that when inspecting the networks we noticed that in the vast majority of cases for every edge between nodes in one direction there was an edge in the opposite direction of similar magnitude. We thus convert our directed networks into undirected networks. This is achieved substituting the directed edges between two nodes with one undirected edge whose weight is the average of the weights of the directed edges. Cases where the edge in one of the directions does not exist are treated as though there do exist two edges where one has zero weight.

The second condition states that each network needs to be a connected graph. A connected graph is one where there exists at least one path from every node to every other node over the edges of the graph. Without this connectivity property, some nodes would have no possible paths to reach each other and thus we would not be able to compute a distance between them[14], nor would assigning any distance to such non-existing paths make sense. This connectivity requirement is an important determinant in selecting the threshold below which we remove edges whose weights aren't high enough. As you increase the threshold, more edges are removed and it becomes more likely that the network becomes disconnected. The selection of this threshold is an important choice as we explained at the end of Section 4.3.3 to which we designed our network comparison to be as robust as possible. We use the connectivity requirement to set an upper-bound on the range of cutoff values that are feasible for our data set. See Section 5.1 for more details on the choice of this threshold. A straightforward way of testing whether a graph is connected is by looking at the *algebraic connectivity* of a graph, which is defined as the second smallest eigenvalue of the (weighted) Laplacian matrix of a graph (Ellens et al., 2011). We use this way of checking the connectivity over using for example an algorithm for this because the Laplacian matrix of the graph and its eigenvalues are also central to the computation of the resistance distances as we will explain below. Using algebraic connectivity is thus more efficient. We choose a cutoff threshold such that the graph is connected, this is the case when the algebraic connectivity is greater than zero. One note on this procedure is that we automatically filter out individual genes that might be individually separated from the network for a certain threshold. Thus only when the graph is split into two sub-graphs each consisting of at least two genes and one edge do we consider the graph disconnected.

Having ensured that our networks are both undirected and connected we can compute the effective resistances over the network. Various methods for calculating this distance metric exists, but in this paper we use the approach as was described in Ellens et al. (2011). In the following we describe the computations necessary for obtaining the distance matrix, for a detailed description and explanation on how to compute this metric, please see Vos (2016).
The Laplacian matrix is central to calculating the effective resistances between all the nodes in a graph. For a weighted graph $G = (V, E)$, with edge weights $w_{ij} \in \mathbb{R}_{>0}$ for edge $(i, j) \in E$ and

[14]This is the case for all distance functions over graphs. In some applications, this problem is solved by setting the distance between such nodes to be infinite. This does not work for our purposes since we intend to perform calculations with these values later.

$w_{ij} = 0$ otherwise, the weighted Laplacian matrix is given by (Vos, 2016):

$$L_{ij} = \begin{cases} \sum_k w_{ik} & \text{if } i = j, \\ -w_{ij} & \text{otherwise.} \end{cases} \tag{3}$$

The weights used here are defined as the reciprocal of the resistances over individual edges (Vos, 2016). This means in our case that we can construct the Laplacian directly from our weighted networks. In similar way to the adjacency matrix, this Laplacian matrix fully characterises a graph such that the full graph can be reconstructed given either of these matrices (Ellens et al., 2011) and thus that the Laplacian contains the same amount of information as the original graph. Let $\lambda_1, \lambda_2, ..., \lambda_n$ be the n eigenvalues of weighted Laplacian matrix L ordered from smallest to largest and let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ be the corresponding eigenvectors. We know that these eigenvalues are real, non-negative and that the smallest one is zero because the Laplacian matrix is symmetric, positive semi-definite and its rows sum up to zero (Ellens et al., 2011). As noted earlier, the second smallest eigenvalue (λ_2) is also known as the algebraic connectivity and should be larger than zero for this study.

The effective resistance between nodes a and b is given by (Ellens et al., 2011):

$$R_{ab} = (\mathbf{e}_a - \mathbf{e}_b)^\top L^+ (\mathbf{e}_a - \mathbf{e}_b) = L^+_{aa} - 2L^+_{ab} + L^+_{bb} \tag{4}$$

In this, L^+ is the Laplacian pseudo-inverse, defined in Vos (2016) as the unique matrix satisfying:

$$L^+ \mathbf{v}_1 = 0 \quad \text{and} \quad L^+ \mathbf{v}_i = \frac{1}{\lambda_i} \mathbf{v}_i \text{ for } i \neq 1. \tag{5}$$

As was shown Section 2.2 in Vos (2016), this pseudo-inverse can be computed using

$$L^+ = U D^+ U^\top \tag{6}$$

Here, U has $\mathbf{v}_i$ as its i^{th} column and is D^+ a diagonal matrix with $D^+_{11} = 0$ and $D^+_{ii} = \frac{1}{\lambda_i}$ for $i \neq 1$.

Using equations 4, 5 and 6, we calculate the resistances distance between all pairs of nodes in both the asthmatic and control networks. The result is two distance matrices, one for each network, with the distances on all off-diagonal elements and with the diagonal equal to zero. Each row in these matrices contains the resistance distances from a gene to all other genes (including itself). These rows are what we referred to earlier when talking about the distance vectors that embed the relevant information of that gene in the network.

We validated our computed distances by comparing it with the results of an alternative way of computing them. This method, which was proposed in Young et al. (2015a), does not involve the computation of eigenvalues but rather involves solving the Lyapunov equation (Dullerud & Paganini, 2013). The fact that this alternative yielded us the exact same distances, but at much higher computational costs, let us to conclude that we had correctly calculated the resistance distances over the networks.

Converting to relative distances

The second step in the construction of the gene affinity profiles is converting the distance vectors computed in the previous step to relative distance vectors. This is done to improve the comparability of the gene embeddings by ensuring that in later steps we can safely compare the outcomes between different genes. The distances that make up the original resistance distance vectors might differ in terms of scale over the different genes in the networks. If not corrected for, this could lead to some of the differences found in later steps being due to such a scale difference between genes rather than measuring which genes show the largest change in local connectivity when compared to the control network. Working with relative distances per gene means that each value now represents the distance to another gene in the network in relation to its distances to all other genes in the network. This allows us to compare the outcomes of our gene affinity profile comparison later on a more equal footing and apply the same RBF kernel transformation to all values in the data set. Converting the absolute effective resistances into relative effective resistances is done by dividing each element in the distance vector of a gene by the sum of that vector. If V is the set of all genes (nodes) our network and R_{ab} is the resistance distance from gene a to gene b, the relative resistance distance is given by:

$$\hat{R}_{ab} = \frac{R_{ab}}{\sum_{k \in V} R_{ak}} \tag{7}$$

This transformation is applied to all distance vectors for each of the genes in both networks. The resulting relative distance vectors have the same dimensionality as the absolute distance vectors and are stored in the similar distance matrices.

Affinities with local emphasis

The goal of the third and final step in constructing the gene affinity profiles is to make sure that comparing them between networks emphasizes differences in connectivity within their neighbourhood. This is achieved by transforming the relative resistance distances into affinities

(similarities) using the Gaussian RBF kernel. We chose to use this function for the following reasons: Firstly, the kernel is a decreasing function in terms of its input, as can be seen in figure 3. This means that its output can be interpreted as a measure of similarity when given the distance between two points as input (Vert et al., 2004). This is important for our purposes because it means that the property central to our use of distance vectors, which was that they indicate which genes are closer than others to that gene, is preserved in the transformation. The second reason is that its derivative (slope) is decreasing in absolute terms over the range of our data and for our choice of parameters. This is illustrated in Figure 3, where we can see that the minimum value of the input lies beyond the inflection point of the curve. This is a very important feature for the purposes of this research, because it means that through this transformation we can put more emphasis on changes in the distances to genes that are in close vicinity of the gene in question. By letting all relative distance values in both of the networks undergo the exact same RBF kernel transformation, we map the values belonging to shorter distances to a wider range of numbers than those belonging to long distances. Through this feature we achieve our goal of emphasising the local structures when comparing between the networks.

The RBF kernel used for this transformation is defined as (Vert et al., 2004):

$$k_{RBF}(\hat{R}_{ab};\sigma) = \exp\left(-\frac{\hat{R}_{ab}^2}{2\sigma^2}\right). \tag{8}$$

Here, σ is a scaling parameter that determines the shape of the kernel and is thus critical in setting the degree by which local connections are emphasized. Choosing the right value for this parameter is no trivial task as it requires a way to quantify the degree of local emphasis that is desired for our research and to link this to the value of the parameter.

We ended up choosing the value of this parameter using a heuristic that we developed to link the value of the scale parameter to the size of the local neighbourhood. In developing this heuristic, we started with defining a way to quantify what a desirable level of local emphasis would be using the size of the local neighbourhood of a gene. Neighbourhood size in this context is defined as the number of genes whose relation (relative distance) to the gene in question should be given extra weight by our method in the comparison. In consultation with experts from the UMCG[15], a neighbourhood size of 50 was chosen. This choice is motivated by the following arguments: Firstly, it ensures that the majority of the differences that we will measure

[15]These experts are G.H. Koppelman, M.C. Nawijn and M. Berg, see Appendix A.

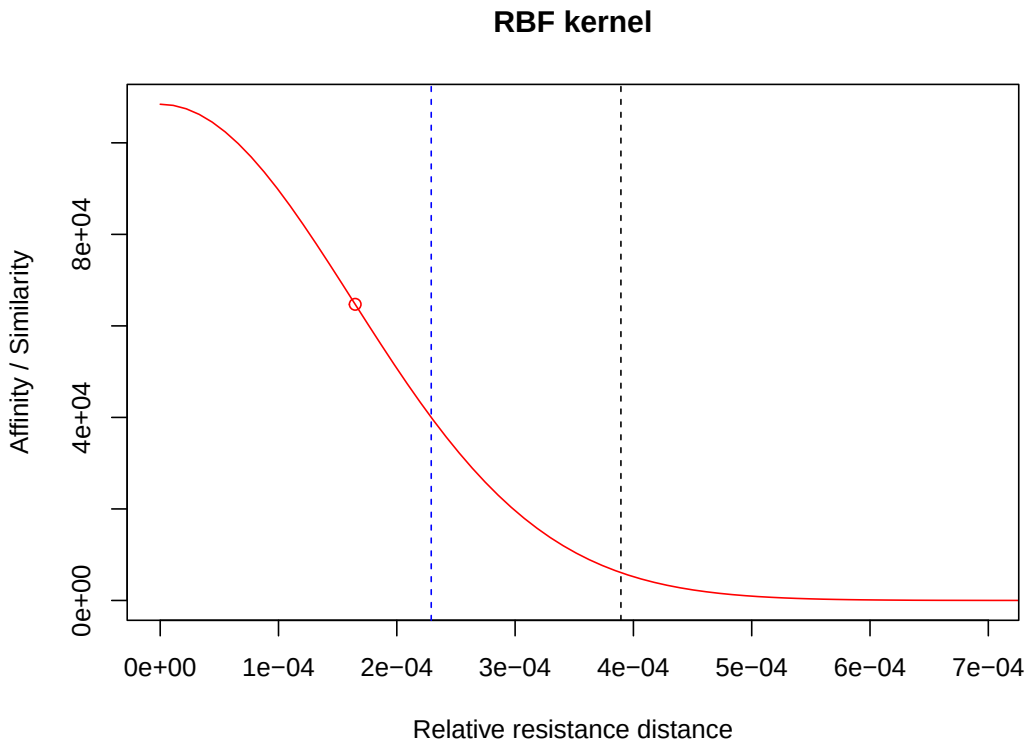


Figure 3. A plot of the shape of the Gaussian radial basis function kernel in terms of its input, where the scaling parameter was set using the procedure described in Section 4.4.1. The red dot signals the inflection point on the curve beyond which the curve is concave, which in this specific case means the derivative is strictly decreasing in absolute terms. The dotted blue line shows the minimum relative resistance distance in our data set. The dotted black line is the estimated neighbourhood border, which calculated as the average minimum distance from a gene in the network within which 50 other genes lie.

in our comparison lie in the interactions between genes in this neighbourhood. This relatively modest neighbourhood size of 50 makes it possible to manually analyse these neighbourhoods later and thus to be able to interpret our results. Secondly, because the experts confirmed that this size is a reasonable estimate of the number of genes that a gene is expected to have regulatory interaction with. Our results proved robust to variations in this parameters, with values from 25 to 75 yielding roughly the same results.

The next step is to link this neighbourhood size to the value of the scale parameter σ. The goal of the RBF transformation is to make our comparisons more sensitive to changes in the relative distance within the local neighbourhood. We achieve this by choosing σ such that the

distances to genes who fall within the neighbourhood get mapped to points with the steepest slope by the kernel and that other distances get mapped to points with (close to) zero slope. By doing this we ensure that when we later compare the gene affinity profile of a gene with its counterpart in the other network, a small difference in relative distance to a gene within the neighbourhood will result in a larger difference in affinity than the same small difference in relative distance to a gene that lies far outside the neighbourhood. As we can see in Figure 3, both the kernel output and its slope are monotonically decreasing (in absolute terms) over the range of our data. This indicates two important things: It means that the RBF kernel maps lower inputs to higher values and it means that the slope of the curve is steeper at the higher values. It is this link between higher output value and steeper slope that we use to set the scaling parameter. This is done in the following way:

First, we calculate the average relative distance to the border of our neighbourhood of predetermined size. This means that for every gene in both networks we calculate the minimum distance within which 50 other genes lie and compute the average over all these values. We call this average the estimated neighbourhood border.

The second step is finding a σ for which distances under the estimated neighbourhood border get mapped by the kernel to much higher values than those above it. This is achieved by setting σ equal to the value for which cumulative density of the distances that fall within the neighbourhood border represent a five times larger fraction of the total density than the fraction that these distances represent in the total number of distances in our data. If we define $\hat{R}_{total}$ as the set of all the computed relative resistance distances in both the networks and $\hat{R}_{neighbourhood} \subseteq \hat{R}_{total}$ the subset of only the distances within the estimated neighbourhood border, this is equal to finding σ that solves the following equation for magnification factor $M = 5$:

$$M = \frac{k_{RBF}(\hat{R}_{neighbourhood}; \sigma)/k_{RBF}(\hat{R}_{total}; \sigma)}{\#\hat{R}_{neighbourhood}/\#\hat{R}_{total}}. \tag{9}$$

The reason that a magnification factor of five was chosen is because experimentation showed it to be a reasonable value in the range of possible values. The magnification factor has the same, but opposite, effect on the scaling parameter as the choice of neighbourhood size has: A higher factor results in a smaller σ with more weight on the closest genes to the gene in question. The largest possible value for this factor given our data set was around nine, above which equation 9 had no solution. Inspection of the behaviour of the above described procedure for different choices of magnification factor between the values of one and nine showed that it is robust to changes in this factor, with values between three and eight producing roughly the same results.

At the end of this last step in constructing the gene affinity profiles in which we transformed the relative resistance distances to gene affinities, we end up with two gene affinity profiles for each gene, one belonging to the asthmatic network and to the control network. Each profile thus takes the form of a vector containing its affinities for all other genes in its network.

4.4.2 Pairwise comparison of genes between the networks

Having constructed the gene affinity profiles for all genes in both the asthmatic and the control group networks, the next step is to compare these profiles between the networks to assess which genes' neighbourhoods in the two networks show the greatest difference in internal connectivity and can be labelled as candidate genes.

Each of the gene affinity profile takes the form of a numeric vector with its length equal to the number of genes in the network. The profiles are constructed in such a way that computing the difference between each of the entries highlights differences in local connectivity. We summarise this into one score for each gene through computing the L2 (euclidean) distance[16] between the two gene affinity profiles of the different networks.

We end up with a score for each gene that measures the difference between the gene affinity profiles of the different networks. Genes that score high on this metric are more likely to have large differences in the way that they are connected to other genes in their neighbourhood.

4.4.3 Validation and interpretation of results

The goal of our network comparison was identify regions in the networks that showed the greatest differences in the connectivity of their neighbourhood. The resulting L2 scores that we have obtained from comparing the gene affinity profiles of each of the genes in the networks give us an indication of the size of this difference. The main results in this thesis are obtained by selecting the genes with the largest L2 scores and inspecting them and their neighbourhoods in both networks manually.

The genes with large L2 scores that we end identifying as candidate genes are those that are consistently identified as outliers over a range of hyperparameter choices (network weight cutoff, neighbourhood size and magnification factor). What constitutes an outlier is determined by the interquartile range rule (Tukey et al., 1977). This rule states that, if we let

[16] We also checked using the L1 distance. This did not change the outcomes of the comparison.

Q_1 and Q_3 denote the lower and upper quartiles of the L2 scores for all genes with that choice of hyperparameters, the range outside which a value is considered an outlier is given by $[Q_1 - 1.5 * (Q_3 - Q_1); Q_3 + 1.5 * (Q_3 - Q_1)]$.

Rather than basing our selection of candidate genes on which genes are outliers with large scores, we would prefer to be able to assess their statistical significance and select our results based on that. We spend a lot of time trying to in some way assess the statistical significance of these scores and also experimented with other metrics and test statistics to measure this difference. The problem that we kept running into was the fact that the values in the gene affinity profiles are not independent because they are all based on the resistance distance over the network. This meant that the independence assumption that is central to many statistical tests is violated. We tried a great many different types of non-parametric testing (bootstrapping, permutation testing, etc.), but ended up having to conclude that we could not find a proper way to assess statistical significance within the time-frame of this thesis.
We do check if our results are robust to changes in the chosen hyperparameters as a form of internal validation and to give us a sense as to whether the results can be attributed to random chance.

The output of this comparison are the names of genes whose neighbourhoods in the asthmatic and control network have the largest difference in their internal interaction structure. This means that can now define regions of interest in the networks whose change in interaction structure between asthma and control can be explored to extract biological insights. Each of the candidate genes identifies such a region of interest, which we define as the union of all the genes that are part of that candidate gene's neighbourhood in either the asthmatic or the control network. Having identified these regions of interest greatly reduces the size of your problem when you are comparing large networks with thousands of nodes. However, the interpretation of the differences in connectivity within these regions between the subject groups (i.e. between the networks) remains a highly complex task. This is largely because comparing changes in connectivity between networks of 50 to 100 nodes can still prove very challenging. It is still an open question of how to best go about biologically interpreting the changes in the interaction pattern of these regions. This is a question that lies on the cutting edge of the field of genomics and one that has not seen many publications. We will attempt to formulate such an in depth biological interpretation with the team at UMCG in the coming months, but this lies outside the scope of this thesis.

However, we will include a limited biological interpretation of the changes in these regions. The purpose of this interpretation is to give the reader some sense of what the road ahead in this research looks like and to serve as a kind of external validation on the results found in this study. In this step, we manually go over the identified regions to assess whether they seem plausible, whether aspects of these regions can be linked to previous literature on the genetics of asthma and how they could help formulate new hypotheses for future research.

5 Results

In this section, we provide the main results of this study, that is, we describe the networks that we constructed and the outcomes of the comparison of these networks. We do not include a description of the results from the data preparation steps in this section, but the interested reader can find the intermediate results and validation of the normalisation and denoising steps in Appendices C and D respectively.

5.1 Network construction

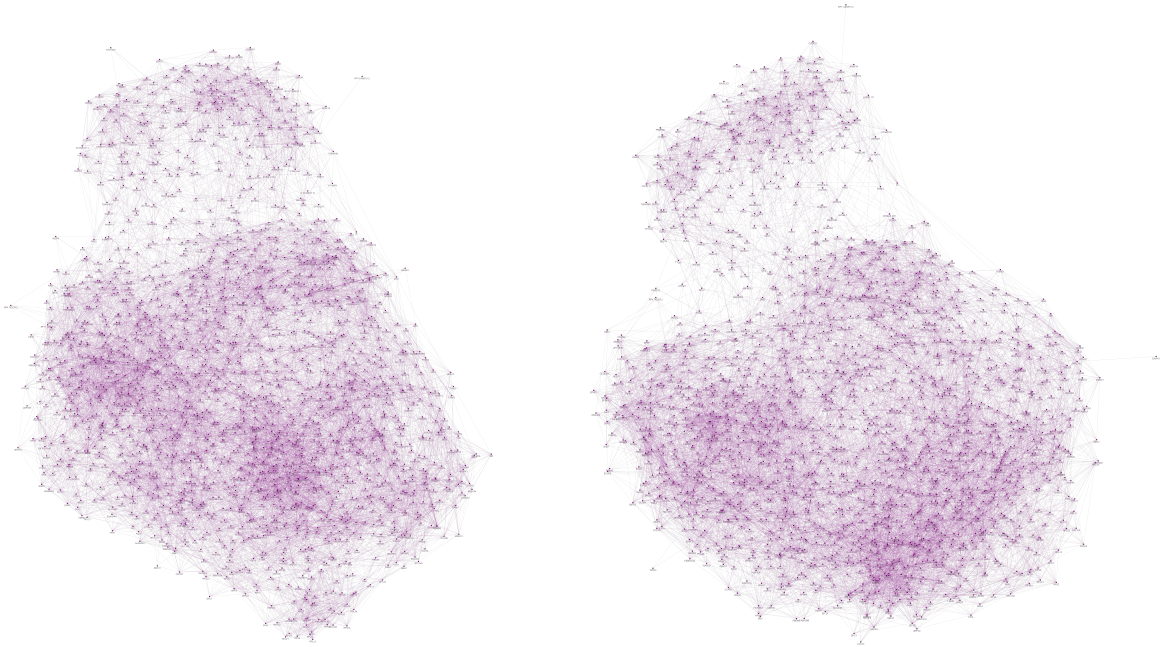
In this section we provide a description of the networks constructed for the comparison in this study. It covers the results from the network inference step and the subsequent steps of removing directionality and thresholding.

After the clustering and imputation phase, the data is split up based on the health status of the donor. A seperate GRN is estimated using GRNBoost2 for both the asthmatic subject group (1032 cells) and the control group (886 cells). Each network consists of nodes, directed edges and weights on the edges. Before we are ready to compare these directed networks outputted by GRNBoost2, two things need to be done. The first is converting them into undirected networks. This is achieved by averaging edges that go in opposite directions and this is a necessary for our network comparison (see Section 4.4).
The second step is removing edges below a certain threshold. As was described in Section 4.3, the networks outputted by GRNBoost2 report all edges between genes for which any non-zero relation could be found. This leads to a very high number of edges in the network having a (very) low weight. Without thresholding, the directed networks each contain around 1.45 million edges representing around 80 percent of the maximum number of possible edges in an undirected network of that size. The weights range from just above zero to 142, with 95.7 percent having a weight lower than one. These values have no direct interpretation but they signify

the relative strengths of the relations found between the genes. In order to lower the chances of our comparison being dominated by these low weight noisy edges, we focus only on the strong edges found by removing all edges below a certain threshold. In this study, we choose to set this threshold at 10. We chose the number 10 because experimentation showed that it strikes a good balance between removing a lot of weak edges and not removing too many medium strength edges that could contain meaningful information. To protect us from reporting any spurious results we test the robustness of the results from the comparison to this choice by running our comparison for cutoff values ranging from 5 to 22 (sensitivity analysis). Here, the upper-bound (22) is set by the point above which the networks cease to be connected which is determined by computing the algebraic connectivity (see Section 4.4). The lower-bound is defined as the lowest cutoff that still removes a large enough portion of the low weight edges. For our chosen value five, this corresponds to removing 98 percent of edges. The results of the network comparison presented in the next section prove to be very robust to this choice. Our comparison method consistently identifies the same regions as having the highest difference in local connectivity for cutoffs between 6 and 16[17].

We end up with our final two networks to be compared after having converted the networks into undirected networks and having removed all edges with weights below the threshold. These networks each contain 1897 nodes that are connected by 19227 and 18511 edges for asthma and control respectively. The reason that the number of nodes in the final networks is down to 1897 from the 2000 selected in Section 4.2.1 is for two reasons. Firstly, the imputation reduced nearly a hundred genes' expression to zero, which meant that they carried no information anymore for the network inference method and were automatically ignored. See Appendix D.2 for a more detailed exploration of this result. The second reason is that setting the threshold at 10 caused around three genes to lose all their edges and were therefore no longer considered to be part of the network. Even though implementing the threshold of 10 thus resulted in an almost 99 percent reduction in the number of edges, these networks are still much too large and complex to be interpreted by hand. Figure 4 shows a side-by-side graphical representation of the final two networks to be compared. This figure gives us a sense of the size and complexity of the networks and shows why a scalable approach to network comparison like the one chosen for this study is necessary.

[17]Sometimes the candidate genes found differ slightly in their relative ranking for different cutoff values, but they consistently stand out at the top.



(a) Asthma *(b)* Control

Figure 4. A graphical representation of the networks used in the network comparison. The images were made using the Cytoscape software environment (Shannon et al., 2003).

5.2 Network comparison

In this section we present the results from the network comparison that we described in Section 4.4. In the first subsection, we describe the main results of this study that are used to answer our research question. From the networks described in the previous section, we constructed gene affinity profiles for each of the genes. The main results of this study are the candidate genes that result from our comparison of these gene affinity profiles and the subsequent identification of a region of interest in the network based on the neighbourhoods of these candidate genes. The second subsection provides a brief exploration of the genes in this region of interest and gives an example of what analysing the differences in interaction in this region between the asthmatic and control group could look like.

5.2.1 Identification of regions of interest

The main results of this study are obtained from comparing the gene affinity profiles of the genes in the asthmatic network with their counterparts in the control network. To find the regions with the largest change in their internal interaction structure, we calculated the euclidean (L2) distances between these vectors. Figure 5 shows a box plot of the distribution of these distances in which outliers are plotted as dots.

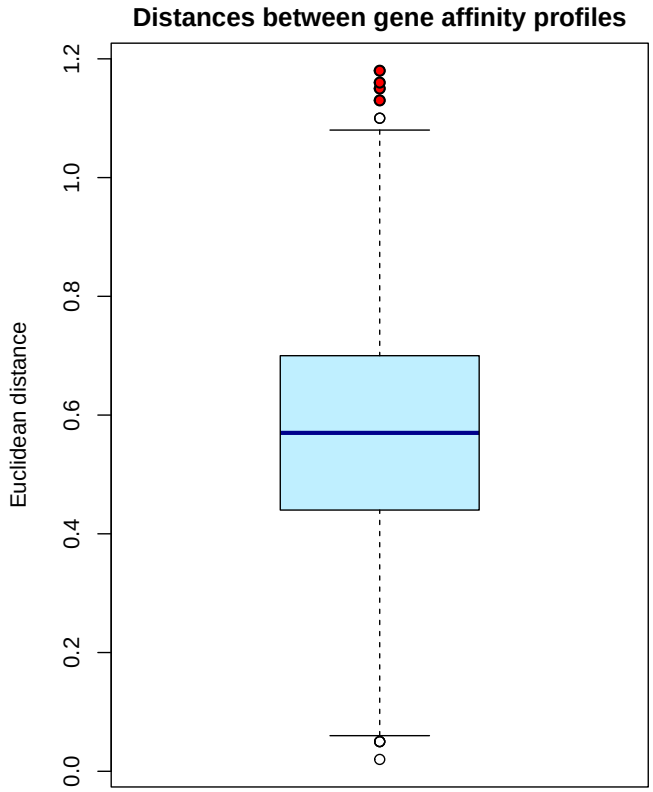


Figure 5. A boxplot depicting the distribution of the L2 distances of all the genes in the networks. Marked in red are the four largest scores that belong to the genes: NDUFA5, ALDOA, SNX3 and FAM3D.

Identification of candidate genes

Four candidate genes are marked red in Figure 5: NDUFA5 (1.18), ALDOA (1.16), SNX3 (1.15) and FAM3D (1.13). These are the genes that showed the largest difference in their gene affinity profiles between the networks and whose neighbourhoods are thus assumed to have the largest and most interesting changes in connectivity between the asthmatic and control groups. These candidate genes each identify a region of interest in the networks that is the union of its neighbourhood in the asthmatic network and the control network. In other words to avoid confusion, we use the term "neighbourhood" to refer to the set of 50 closest genes to a gene in either the asthmatic network or the control network and we use the term "region (of interest)" to refer to the union of these neighbourhoods.

We selected these four genes as candidates because they were identified as outlier genes with the largest difference in their gene affinity profiles over a wide range of the hyperparameter

choices. They consistently showed up as outliers with the four largest distances for different choices in the cutoff threshold (from 6 to 16), neighbourhood size (25 ranging to 75) and magnification factor (from 3 to 8). This robustness to the choices of hyperparameters serves as a form of internal validation of the method and makes us confident that results found are not merely due to us choosing one particular set of hyperparameters. We do not include the other two large outliers visible in the top of Figure 5 because they do not show up as consistently as the other four over the different choices of hyperparameters. The outliers at the bottom of the figure are also not included in the results for two reasons: Firstly, because the goal of this study was to identify specifically regions in the networks that show large differences in local connectivity. It is this goal to which our network comparison method is tailored. Secondly, because they do not show up consistently as outliers at the bottom for other choices of hyperparameters and thus them being outliers for this configuration might be due to chance.

Links with existing literature

Before we start exploring the regions in the networks around these candidate genes, we briefly look into these genes themselves. The aim here is to see if they have been linked to asthma in previous research, which could serve as external validation of our results. We were aided in the process of finding these links by experts from the UMCG[18] because of their knowledge on asthma genomics and familiarity with the relevant literature. We found that in Jevnikar et al. (2019) NDUFA5 was identified as part of the IL6-IL6R trans-signalling signature, a genetic pathway that has been linked to asthma (Jevnikar et al., 2019; El-Husseini et al., 2020). O'Neil et al. (2011) identified the protein ALDOA (which is encoded by the gene ALDOA) as one of five proteins that are involved in inflammatory and immunological disorders in asthmatic individuals. Furthermore, Fang et al. (2012) found that ALDOA showed strong differential genetic expression in asthma. We were not able to find studies directly linking SNX3 to asthma. Bosco et al. (2010) observed that FAM3D is associated with asthmatic airway obstruction and identified as being part of the epithelial differentiation pathway involved with asthma. Kaneko et al. (2013) also describes FAM3D as being part of a network of genes associated with asthma whose top functions involve: cell-mediated immune response, cellular development, cellular function and maintenance. The identification of these links between asthma and the candidate genes is already a good sign that our search has yielded relevant results.

[18]These experts are G.H. Koppelman, M.C. Nawijn and M. Berg, see Appendix A.

The neighbourhoods of the candidate genes

Our method identified these four candidate genes because their regions in the networks are likely to contain interesting differences in terms of their internal interactions. Earlier, we defined these regions of interest as the union between their neighbourhoods of 50 closest genes in the networks. This means that the genes in each region can be subdivided into three groups: genes that are only in the neighbourhood of a candidate gene in the asthmatic network, genes that are only in the neighbourhood of a candidate gene in the control network, and genes that are part of the neighbourhood in both networks. For all candidate genes, these groups contain on average 30, 30 and 20 genes respectively, with average total size of the region of interest being around 80 genes.

When comparing the sets of genes that make up these four regions we notice a high degree of overlap. Computing the percentages of common genes between the regions, we find that the regions of NDUFA5 and ALDOA have slightly under 60 percent of genes in common. Their commonality is higher with the region of SNX3, which has around 70 percent of genes in common with both NDUFA5 and ALDOA. The region of FAM3D seems to be the most disjoint from the rest, sharing around 34 percent of genes with each of the other regions. This high overlap in genes might point to the fact that the four regions identified by our method are all picking up on the same general region in the network rather than four disjoint ones. When we further explore the different neighbourhoods in both networks that make up these regions, we see that 3 genes are present in all neighbourhoods. Furthermore, we find that 21 genes are present in all the neighbourhoods in the asthmatic network and 16 genes in all the neighbourhoods in the control network. We conclude that these genes, which are listed in Table 2, are the genes that show the largest difference in their interaction between asthmatic and healthy individuals because they consistently show up in the neighbourhoods of the candidate genes. It is this group of genes that constitute the intersections of the neighbourhoods in either the asthma or control network that we report as the main results of this study and that could serve as a starting point of future research.

The composition of this reported group is of course dependent on us choosing to intersect the neighbourhoods of all four candidate genes. If we chose to intersect only the neighbourhoods of three of the four candidate genes then more genes would be included in this group. The opposite also holds, where the group would shrink in size if we intersected it with the neighbourhoods of another gene. For the precise composition of the neighbourhoods in both networks of each of the four candidate genes we refer to Appendix E. We chose to report the intersection in the

Table 2

Genes consistently found in the neighbourhoods of NDUFA5, ALDOA, SNX3 and FAM3D

Network	Gene names
Asthma only	ALPL, ATP5F1, CANX, CNDP2, COPS6, EI24, FKBP1A, GOLM1, MFSD1, MTCH1, PDLIM5, PPP2CA, RAB1A, RAB2A, RHOA, SAP18, TM9SF2, TPI1, UBE2D3, XRCC6, YTHDF2
Control only	ALDOA, ATP6AP2, DDOST, FAM3D, N4BP2L2, NDUFA5, PAFAH1B2, PSMD1, RASSF6, SELT, SEPT7, SF3B1, SLC31A1, SNX3, XRCC5, YIPF6
Both	C1orf63, TMED2, TMEM230

neighbourhoods of all four candidate genes because the genes in this intersection show up so consistently gives us confidence in that the changes in interaction between these genes capture a biologically meaningful difference. Thus by focusing on this overlap we report the results that we are most confident about. An added benefit of focusing on this intersection is that it reduces our final results to a group of 40 genes, which makes the exploration done in the next section easier. We do not reduce the size of this group even further by intersecting it with the neighbourhoods of another gene because we only robustly identified four candidate genes.

5.2.2 Exploration of the region of interest

In this section we explore the region in the networks identified in the previous section that consists of the genes in Table 2, divided into two parts. In the first part, we report the links that we were able to find between these genes and existing literature. In the second part, we present an example of what analysing the difference in interaction between between asthmatic and control in this region could look like.

5.2.3 Links with existing literature

In order to check if the genes of interest identified in this study (see Table 2) have been linked to asthma in previous studies, we conducted a brief but broad search of the literature. The goal of this search is to provide some external validation of our results by showing that the genes identified in this study have been linked to asthma before. It is for this reason that, when going through the literature, we looked for instances in which the authors describe a clear link between (one of) the genes and asthma in the main body of the paper. This last part is important because it is not uncommon for the appendices of papers in this field to include vast lists on the outcomes of tests conducted on thousands of genes. These lists often contain hundreds

of statistically significant results[19]. We therefore chose to only report the links between genes and asthma that were explicitly mentioned in the main body of the papers because we felt that appearing among hundreds of significant results in an appendix did not constitute strong enough evidence of a link.

Table 3

Links found in literature between reported genes and asthma

Gene	Link with asthma	Source
ALPL	Differentially expressed for asthma and identified as a biomarker for a particular asthma inflammatory phenotype;	Baines et al. (2011, 2014); Reddy & Covar (2016); Berthon et al. (2017)
CANX	Dysregulated in asthmatic subjects and associated with impaired lung function among other symptoms;	Pathinayake et al. (2021)
FKBP1A	Differentially expressed for asthma; Part of a genetic pathway related to asthma;	Persson et al. (2015) Osei-Kumah et al. (2011)
MFSD1	Identified as a potential target gene for treatment of asthma due to its role in airway contractility;	McGraw et al. (2006)
PPP2CA	Differentially expressed for asthma;	Shu et al. (2021)
RAB1A	Linked to aspirin intolence in asthmatic patients;	Park et al. (2014)
RHOA	Identified as being a promising new target for asthma treatment;	Chiba et al. (2009); Zhang et al. (2020)
SLC31A1	Differentially regulated for asthma;	Jackson et al. (2020b)
TM9SF2	Higher methylation level in this gene is associated with increased odds of asthma;	Breton et al. (2016)
XRCC6	Differentially expressed for asthma;	Shu et al. (2021)
YTHDF2	Differentially regulated for asthma;	Sun et al. (2021)

Note. This list does not include the links that we reported earlier for the candidate genes.

Table 3 shows the links that we found in this search. In order to avoid repetition, we do not include the links between candidate genes NFUDA5, ALDOA and FAM3D that were reported in the previous section. One might notice that there is little overlap in the papers in which these links are found. We think that this is a result of the vastly different scope and approach taken in our study when compared to these earlier studies. As is common for most asthma research, these studies all investigate a specific type, aspect or symptom of asthma and use methods specifically designed for that goal. Such research requires extensive knowledge on the underlying biology. This is different from the study conducted in this thesis, where we aimed to

[19]Significance often being determined using the Benjamini-Hochberg adjusted p-values to decrease the false discovery rate.

discover a group of genes that showed interesting differences in their interaction using a data-driven approach requiring only minimal biological background knowledge. As we mentioned in the introduction, such an approach focusing on changes in interaction constitutes a completely new way of analysing genetic data that picks up on different signals in the data than earlier approaches. This also makes it not unreasonable for the genes identified in our study to be linked to asthma in a range of different types of studies before.
It would have been very interesting if the genes that we found were all identified in the same (type of) study, which could have meant that we had stumbled directly upon an earlier discovered genetic mechanism relating to asthma. The fact that they are not points to fact that our approach is analysing aspects of the data that were out of reach for the methods in other studies.

Our final note on the links reported in Table 3 is that they all (except for SLC31A1) belong to genes that were found in the neighbourhoods of the candidate genes in the asthmatic network and not the control network (see Table 2). That they belong to this subgroup of our results suggests that they are much closer to each other in the asthmatic network and thus are likely to be much more connected in asthmatic patients. We will see further evidence of this increased connectivity in the next section. The fact that we find these links with literature for this group of genes, the ones found in the neighbourhoods in the asthmatic networks, shows that it is worth investigating this group further. A deeper investigation of the results could focus on either interpreting the biological functions of these genes or exploring the interaction changes between them. Unfortunately, both our understanding of the biology underpinning all these links and the time available to us are too limited to be able to do a biological interpretation and for example identify a common denominator between the links reported in this section. However, we will provide a brief exploration of the changes in connectivity in the next section.

5.2.4 Connectivity difference between the networks

In this section, we explore the difference in connectivity between the two networks for the identified region that consists of the genes in Table 2. We must note that exploring the changes in internal connectivity for groups of this size and using this to formulate a biological interpretation gets complex very quickly. For the scope of this thesis, we limit ourselves to a basic exploration using a visualisation that aims to serve as an example of where one could start when formulating such an interpretation.

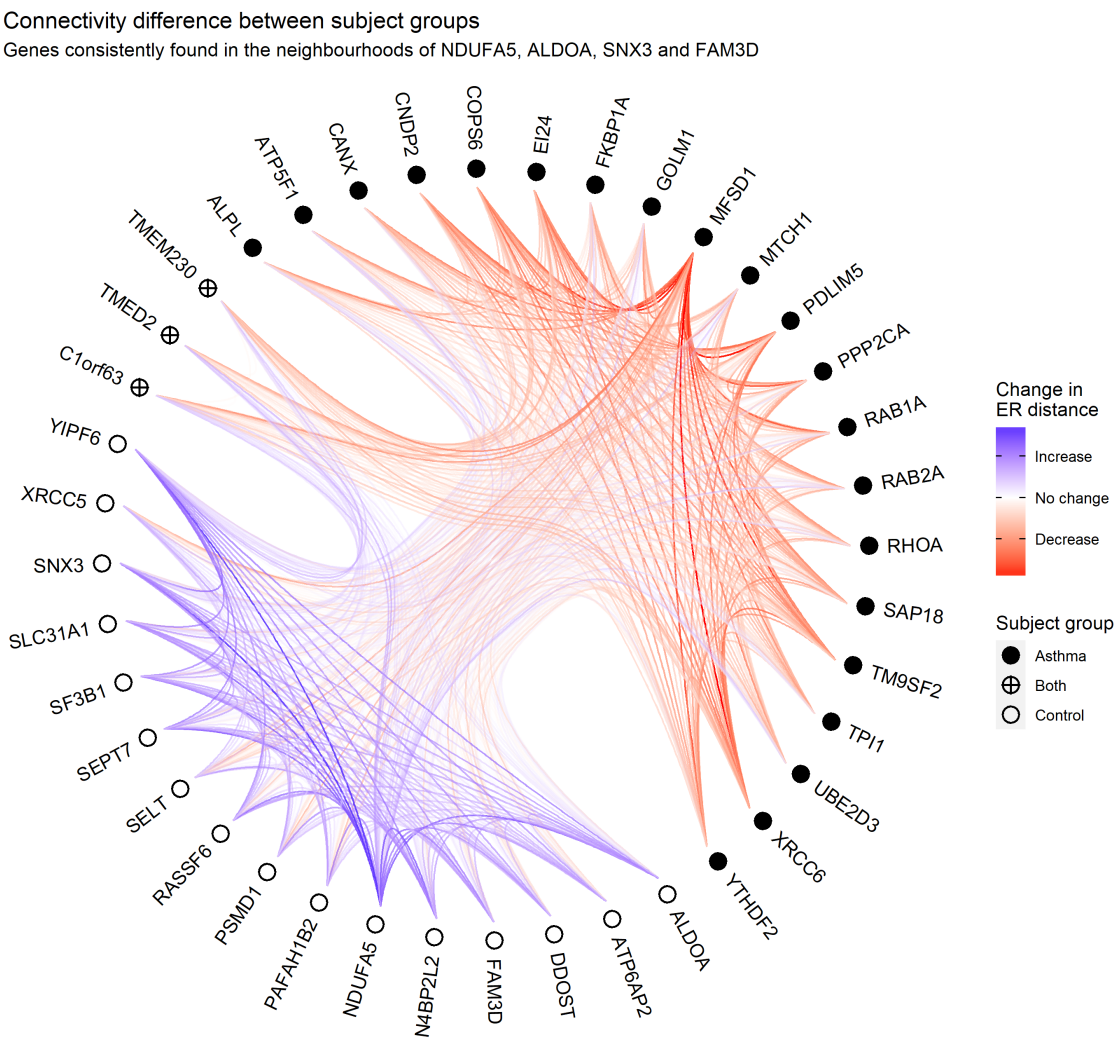


Figure 6. A visualisation of the difference in gene connectivity between the asthmatic and the control networks for the genes consistently found in the neighbourhoods of NDUFA5, ALDOA, SNX3 and FAM3D. Node filling shows if a gene was consistently found to be part of the neighbourhoods in either the asthmatic network, the control network or both. A red (blue) colored edge shows that the distance between genes is smaller (larger) in the asthmatic network.

Figure 6 shows this visualisation of the difference in the connectivity between the identified genes. Each node corresponds to one of the genes from Table 2 and are distributed in groups over the circular graph based on whether the gene was found in the neighbourhoods in the asthmatic network (black), control network (white) or both (crossed). The edges in this graph visualise the changes in connectivity between pairs of genes when comparing the asthmatic network to the control network. Here, we used the ER distance between two genes as a measure of how well they are connected just as we did in other parts of this study. In order to better show the differences between the three identified subgroups of genes in our results (denoted by

the different fillings of the nodes), the edges between genes are bundled according to group. This is also why there are three white areas visible in the graph. The colour of the edges drawn between the nodes shows the difference in this connectivity when comparing the asthmatic network to the control network. A pair of genes whose ER distance in the asthmatic network is smaller than the control network are connected by a red edge. This decrease in distance thus means increased connectivity in the asthmatic network. Conversely, genes whose connectivity decreases are connected by a blue edge. The size of this change in connectivity is denoted by the intensity of the edge. The less the connectivity of a pair of genes differs between the asthmatic and the control network, the more transparent and white the edge becomes (thus becoming less visible). For pairs of genes whose connectivity differs a lot between asthmatic and control, the opposite occurs and their edges become less transparent and more red or blue.

Looking at the edges in the figure, we notice that the strongest increases in pairwise connectivity (the darkest red edges) all seem to be between the genes that are in the neighbourhoods of the candidate genes in the asthmatic network. This is not surprising given the fact that these genes were close enough to the candidate genes in the asthmatic network to be in the neighbourhood but not close enough in the control network to be in the neighbourhood. We also notice the opposite result of the strongest decreases in pairwise connectivity (the darkest blue edges) all seem to be between genes from the neighbourhoods of the candidate genes in the control network. This is again not surprising for the same but opposite reason that was given for the red edges.

An interesting conclusion that we draw based on this figure is that it seems that the genes denoted by the black nodes all seem to be much more strongly connected in the asthmatic network. Given the links with asthma that we found for some of these genes in the previous section, this increased connectivity points to increased genetic interaction between these genes in asthmatic individuals, which in turn could play a role in the development of asthma. Further research is needed to explore what exactly this role might be. A possible explanation could be that these genes that show increased interaction are members of one or multiple genetic pathways that see increased activity in asthmatic cells. Both the identification of such pathways and linking these to biological functions in cells can help in formulating a more structured explanation as to how these genes might be "cooperating" in allowing asthma to develop.

A concrete example of a study that could follow this thesis is one where networks are analysed consisting only of the genes with black nodes in Figure 6 and related genes. Which genes are considered related genes can be determined either based on the networks constructed in this

study or on associations with other genes found in previous studies. If one is able to identify (parts of) known pathways in these networks, one could use the method proposed in Grimes et al. (2019) to quantify the differences of these known gene regulatory pathways between the asthmatic and control groups.

Lastly, we notice a very high number of (dark) red edges connecting MFSD1 to other genes. This points to the fact that MFSD1 has a much more central position in this group of genes for asthmatic individuals, suggesting that it might have a central role in the mechanisms behind asthma. It might therefore be particularly interesting to investigate MFSD1 and genes related to it in future research.

6 Conclusion

In this thesis, we conducted a study using an entirely network based approach to compare the genetic interaction in the airway cells of asthmatic individuals with that of healthy individuals. The field of asthma genomics is quickly moving towards analysing genes and their inferred interactions in a network context (El-Husseini et al., 2020). With the field of gene regulatory network construction seeing many publications in recent years, this study goes one step further by not just creating but also comparing these networks between an asthmatic and a control group. It is with such a comparison that we aimed to answer the following research question:

Which genes show the largest difference in their local interaction between asthmatic and healthy individuals?

The study conducted to answer this question consisted of many different steps that lie at the crossroads of different fields, combining topics in econometrics, statistics, machine learning and biology. We divided these steps of the research up into three parts.
In the first two parts, we formulated and implemented an approach for the preparation of the single-cell sequencing data and the construction of gene regulatory networks based on the existing knowledge from the literature. The main challenge here was getting familiar with all the different methods, determining which ones best suited our purposes and implementing them correctly. In the process of choosing the methods in these parts of the study, we worked with topics from a wide range of fields. Examples are: the normalisation of genetic data, different clustering algorithms, feature embeddings, imputation using (deep) autoencoders, the inference of networks from expression data, working on high-performance computing clusters, gradient

boosting machines, variable importance in tree models and more. After these first two parts, we had constructed two large genetic interaction networks from our prepared data set.
In the third part, we compared the constructed networks with the goal of identifying regions of genes in the networks that have a high likelihood of displaying interesting differences between the subject groups. Having noticed the lack of existing methods for performing a large-scale network comparison with such a goal, we proposed and implemented a novel self-developed approach to network comparison that relies on embedding genes into gene affinity profiles. Our challenge here was identifying the type of comparison needed for our study, developing an effective methodology for this and actually implementing it in a scalable way. To reach this goal, we combined knowledge from topics like: graph embeddings, dimensionality reduction, spectral graph theory, different similarity measures, graph distances and different approaches to non-parametric testing.

Our network comparison yielded four candidate genes (NDUFA5, ALDOA, SNX3 and FAM3D) that each identify a region of interest based on their neighbourhoods in the networks. The subsequent inspection of these neighbourhoods resulted in the identification of a group of 40 genes (see table 2) that were consistently found in the neighbourhood of the candidate genes in either the asthmatic network, the control network or both. This group of 40 genes serves as the main result of this study and as the answer to our research question. The genes in this group are the ones identified in this study as having the largest difference in their interaction between the two subject groups. The fact that they stand out in this way combined with the fact that some of them were linked to asthma in previous literature gives us confidence that there are actual differences in genetic regulatory interaction for these genes in asthmatic cells. Our brief exploration of this group of genes already led to the potentially interesting insight that some of these genes seem to be much closer connected in the asthmatic network and that it might be increased cooperation between these genes that plays a role in the development of asthma.

The implications of the study conducted in this thesis are two-fold. On the one hand, the results of this study can help narrow down the scope of future studies and can even serve as starting point for a more targeted investigation of their genetic interaction. On the other hand, the new method proposed in this study can serve as an example of a fully data-driven approach to compare large networks with the goal of identifying regions showing large differences in local connectivity, both in the field of asthma or more broadly in the field of genomics. In addition to

building on our methodology for this specific goal in network analysis, we would also like to recommend researchers to explore the possibilities of large-scale network comparison in a broader sense. As this study shows, such broad comparisons of large networks can help researchers discover new interesting differences between the interaction of genes for different subject groups. The main benefit of this approach over the more common targeted approaches is that it only requires a minimal amount of biological assumptions. This allows the data to speak for itself and opens the door to potentially surprising new insights.

Finally, we would like to mention three limitations of this study. The first is the need for a more thorough validation of the results. In order to turn the results from this thesis into actual biological insights, we recommend first subjecting the results to a more thorough validation and subsequently perform a deeper analysis of the results in terms of its biological implications. An idea for further validation is repeating the entire process on random subsets of our data to see if the results agree or, if possible, on a different data set.
The second is that further analysis of the results is still needed to be able to draw actual biological conclusions. We already provided some examples of what the start of further analysis of the results with the goal of formulating a biological interpretation could look like in Section 5.2.2. There, we recommended diving deeper into the observed increased connectivity between some of the genes identified. This increased connectivity could be due to increased cooperation of these genes, potentially in known genetic interaction structures like genetic pathways. Other ideas could be to infer new networks from the data using only the identified genes or to subject the expression data of these genes to statistical testing directly. In particular, we would recommend investigating the genes found in the neighbourhoods for the asthmatic subject group (see table 2) given that a quick review of the literature already yielded many links between those genes and asthma. The last limitation has to do with the fact that the comparison which yielded the results of this study focused on one specific type of difference between networks. Limiting this study to only this type was necessary for defining a reasonable scope of the thesis, but it should not be seen as a denial of the existence of other types of biologically meaningful differences between networks like these. Other types of comparisons could complement the results found in this study and help the field get closer to understanding the genetic mechanisms behind asthma.

In conclusion, this study yielded results relevant for further understanding the genetic mechanism behind asthma and obtained these results using a new proposed method for comparing large genetic interaction networks.

7 References

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., ... others (2017). Scenic: single-cell regulatory network inference and clustering. *Nature methods*, *14*(11), 1083–1086.

Anders, S., & Huber, W. (2012). Differential expression of rna-seq data at the gene level–the deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, *10*, f1000research.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., & Garmire, L. X. (2019). Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome biology*, *20*(1), 1–14.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., ... Kendziorski, C. (2017). Scnorm: robust normalization of single-cell rna-seq data. *Nature methods*, *14*(6), 584.

Bacher, R., & Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, *17*(1), 1–14.

Baines, K. J., Simpson, J. L., Wood, L. G., Scott, R. J., Fibbens, N. L., Powell, H., ... Gibson, P. G. (2014). Sputum gene expression signature of 6 biomarkers discriminates asthma inflammatory phenotypes. *Journal of allergy and clinical immunology*, *133*(4), 997–1007.

Baines, K. J., Simpson, J. L., Wood, L. G., Scott, R. J., & Gibson, P. G. (2011). Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. *Journal of Allergy and Clinical Immunology*, *127*(1), 153–160.

Banerjee, P., Balraj, P., Ambhore, N. S., Wicher, S. A., Britt, R. D., Pabelick, C. M., ... Sathish, V. (2021). Network and co-expression analysis of airway smooth muscle cell transcriptome delineates potential gene signatures in asthma. *Scientific Reports*, *11*(1), 1–16.

Berthon, B. S., Gibson, P. G., Wood, L. G., MacDonald-Wicks, L. K., & Baines, K. J. (2017). A sputum gene expression signature predicts oral corticosteroid response in asthma. *European Respiratory Journal*, *49*(6).

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193.

Bosco, A., Ehteshami, S., Stern, D. A., & Martinez, F. D. (2010). Decreased activation of inflammatory networks during acute asthma exacerbations is associated with chronic airflow obstruction. *Mucosal immunology*, *3*(4), 399–409.

Braga, F. A. V., Kar, G., Berg, M., Carpaij, O. A., Polanski, K., Simon, L. M., ... others (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature medicine*, *25*(7), 1153–1163.

Breton, C. V., Gao, L., Yao, J., Siegmund, K. D., Lurmann, F., & Gilliland, F. (2016). Particulate matter, the newborn methylome, and cardio-respiratory health outcomes in childhood. *Environmental epigenetics*, *2*(2), dvw005.

Busse, W. W., Banks-Schlegel, S., & Wenzel, S. E. (2000). Pathophysiology of severe asthma. *Journal of Allergy and Clinical Immunology*, *106*(6), 1033–1042.

Chan, T. E., Stumpf, M. P., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, *5*(3), 251–267.

Chen, G., Ning, B., & Shi, T. (2019). Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, *10*, 317.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chiba, Y., Tanabe, M., Goto, K., Sakai, H., & Misawa, M. (2009). Down-regulation of mir-133a contributes to up-regulation of rhoa in bronchial smooth muscle cells. *American journal of respiratory and critical care medicine*, *180*(8), 713–719.

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., ... Yosef, N. (2019). Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell systems*, *8*(4), 315–328.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, *42*(2-3), 393–405.

Dullerud, G. E., & Paganini, F. (2013). *A course in robust control theory: a convex approach* (Vol. 36). Springer Science & Business Media.

El-Husseini, Z. W., Gosens, R., Dekker, F., & Koppelman, G. H. (2020). The genetics of asthma and the promise of genomics-guided drug target discovery. *The Lancet Respiratory Medicine*.

Ellens, W., et al. (2011). Effective resistance and other graph measures for network robustness. *MS thesis*.

Ellens, W., Spieksma, F. M., Van Mieghem, P., Jamakovic, A., & Kooij, R. E. (2011). Effective graph resistance. *Linear algebra and its applications*, *435*(10), 2491–2506.

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), 1–14.

Erle, D. J., & Sheppard, D. (2014). The cell biology of asthma. *Journal of Cell Biology*, *205*(5), 621–631.

Fang, Z., Tian, W., & Ji, H. (2012). A network-based gene-weighting approach for pathway analysis. *Cell research*, *22*(3), 565–580.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601–620.

Gandolfo, L. C., & Speed, T. P. (2018). Rle plots: Visualizing unwanted variation in high dimensional data. *PloS one*, *13*(2), e0191629.

Global Burden of Disease Study. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *the Lancet*, *396*(10258), 1204 – 1222.

Grimes, T., Potter, S. S., & Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. *Scientific reports*, *9*(1), 1–12.

Guo, X., Zhang, Y., Hu, W., Tan, H., & Wang, X. (2014). Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PloS one*, *9*(2), e87446.

Halapi, E., & Bjornsdottir, U. S. (2009). Overview on the current status of asthma genetics. *The clinical respiratory journal*, *3*(1), 2–7.

Heffler, E., Crimi, C., Mancuso, S., Campisi, R., Puggioni, F., Brussino, L., & Crimi, N. (2018). Misdiagnosis of asthma and copd and underuse of spirometry in primary care unselected patients. *Respiratory medicine*, *142*, 48–52.

Heffler, E., Pizzimenti, S., Guida, G., Bucca, C., & Rolla, G. (2015). Prevalence of over-/misdiagnosis of asthma in patients referred to an allergy clinic. *Journal of Asthma*, *52*(9), 931–934.

Heijink, I. H., Kuchibhotla, V. N., Roffel, M. P., Maes, T., Knight, D. A., Sayers, I., & Nawijn, M. C. (2020). Epithelial cell dysfunction, a major driver of asthma development. *Allergy*, *75*(8), 1902–1917.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., ... Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, *15*(7), 539–542.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. Retrieved from http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html

Human genome project faq. (2020, Feb 24). National Human Genome Research Institute. Retrieved from https://www.genome.gov/human-genome-project/Completion-FAQ (Accessed: 04-05-2021)

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, *5*(9), e12776.

Huynh-Thu, V. A., & Sanguinetti, G. (2019). Gene regulatory network inference: an introductory survey. In *Gene regulatory networks* (pp. 1–23). Springer.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell rna-seq data. *Genome biology*, *17*(1), 1–15.

Jackson, N. D., Everman, J. L., Chioccioli, M., Feriani, L., Goldfarbmuren, K. C., Sajuthi, S. P., ... others (2020a). Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. *Cell Reports*, *32*(1), 107872.

Jackson, N. D., Everman, J. L., Chioccioli, M., Feriani, L., Goldfarbmuren, K. C., Sajuthi, S. P., ... others (2020b). Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. *Cell reports*, *32*(1), 107872.

Jevnikar, Z., Östling, J., Ax, E., Calvén, J., Thörn, K., Israelsson, E., ... others (2019). Epithelial il-6 trans-signaling defines a new asthma phenotype with increased airway inflammation. *Journal of allergy and clinical immunology*, *143*(2), 577–590. Retrieved from https://eprints.soton.ac.uk/421475/1/IL6_TS_UBIOPRED.pdf

Kaneko, Y., Yatagai, Y., Yamada, H., Iijima, H., Masuko, H., Sakamoto, T., & Hizawa, N. (2013). The search for common pathways underlying asthma and copd. *International journal of chronic obstructive pulmonary disease*, *8*, 65.

Kavanagh, J., Jackson, D. J., & Kent, B. D. (2019). Over-and under-diagnosis in asthma. *Breathe*, *15*(1), e20–e27.

Klein, D. J., & Randić, M. (1993). Resistance distance. *Journal of mathematical chemistry*, *12*(1), 81–95.

Li, H., Wang, H., Sokulsky, L., Liu, S., Yang, R., Liu, X., ... Zhang, G. (2021). Single-cell transcriptomic analysis reveals key immune cell phenotypes in the lungs of patients with asthma exacerbation. *Journal of Allergy and Clinical Immunology*, *147*(3), 941-954.

Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, *9*(1), 1–9.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413–422).

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, *15*(12), 1053–1058.

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

Lun, A. T., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, *17*(1), 1–14.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, *9*(8), 796–804.

Martinez, F. D. (2007). Genes, environments, development and asthma: a reappraisal. *European Respiratory Journal*, *29*(1), 179–184.

McGraw, D. W., Fogel, K. M., Kong, S., Litonjua, A. A., Kranias, E. G., Aronow, B. J., & Liggett, S. B. (2006). Transcriptional response to persistent β2-adrenergic receptor signaling reveals regulation of phospholamban, which alters airway contractility. *Physiological genomics*, *27*(2), 171–177.

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., & Aerts, S. (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, *35*(12), 2159–2161.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, *5*(7), 621–628.

Ober, C., & Hoffjan, S. (2006). Asthma genetics 2006: the long and winding road to gene discovery. *Genes & Immunity*, *7*(2), 95–100.

O'Neil, S. E., Sitkauskiene, B., Babusyte, A., Krisiukeniene, A., Stravinskaite-Bieksiene, K., Sakalauskas, R., ... Lötvall, J. (2011). Network analysis of quantitative proteomics on asthmatic bronchi: effects of inhaled glucocorticoid treatment. *Respiratory research*, *12*(1), 1–15.

Osei-Kumah, A., Smith, R., Jurisica, I., Caniggia, I., & Clifton, V. (2011). Sex-specific differences in placental global gene expression in pregnancies complicated by asthma. *Placenta*, *32*(8), 570–578.

Park, J.-S., Heo, J.-S., Chang, H. S., Choi, I. S., Kim, M.-K., Lee, J.-U., ... Park, C.-S. (2014). Association analysis of member ras oncogene family gene polymorphisms with aspirin intolerance in asthmatic patients. *DNA and cell biology*, *33*(3), 155–161.

Pathinayake, P. S., Waters, D. W., Nichol, K. S., Brown, A. C., Reid, A. T., Hsu, A. C.-Y., ... others (2021). Endoplasmic reticulum-unfolded protein response signalling is altered in severe eosinophilic and neutrophilic asthma. *Thorax*.

Persson, H., Kwon, A. T., Ramilowski, J. A., Silberberg, G., Söderhäll, C., Orsmark-Pietras, C., ... others (2015). Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles. *Journal of Allergy and Clinical Immunology*, *136*(3), 638–648.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, *9*(1), 171–181.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, *17*(2), 147–154.

Qiu, P. (2020). Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, *11*(1), 1–9.

Reddy, M. B., & Covar, R. A. (2016). Asthma phenotypes in childhood. *Current opinion in allergy and clinical immunology*, *16*(2), 127–134.

Reverter, A., & Chan, E. K. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, *24*(21), 2491–2497.

Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC bioinformatics*, *12*(1), 1–17.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, *11*(3), 1–9.

Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference* (Vol. 130, p. 136).

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.

Royer, D. J., & Cook, D. N. (2021). Regulation of immune responses by nonhematopoietic cells in asthma. *The Journal of Immunology*, *206*(2), 292–301.

Seumois, G., Ramírez-Suástegui, C., Schmiedel, B. J., Liang, S., Peters, B., Sette, A., & Vijayanand, P. (2020). Single-cell transcriptomic analysis of allergen-specific t cells in allergy and asthma. *Science immunology*, *5*(48).

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–2504.

Shu, H., Li, Y., Xu, H., Yin, Q., Song, J., Zheng, M., & Li, H. (2021). Interferon characterization associates with asthma and is a potential biomarker of predictive diagnosis. *Bioscience reports*, *41*(3).

Skadhauge, L., Christensen, K., Kyvik, K., & Sigsgaard, T. (1999). Genetic and environmental influence on asthma: a population-based study of 11,688 danish twin pairs. *European Respiratory Journal*, *13*(1), 8–14.

Sławek, J., & Arodź, T. (2013). Ennet: inferring large gene regulatory networks from expression data using gradient boosting. *BMC systems biology*, *7*(1), 1–13.

Staff webpage of the University of Groningen. (2021, Nov 29). University of Groningen. Retrieved from https://www.rug.nl/staff/ (Accessed: 02-012-2021)

Sun, D., Yang, H., Fan, L., Shen, F., & Wang, Z. (2021). m6a regulator-mediated rna methylation modification patterns and immune microenvironment infiltration characterization in severe asthma. *Journal of Cellular and Molecular Medicine*, *25*(21), 10236–10247.

Talwar, D., Mongia, A., Sengupta, D., & Majumdar, A. (2018). Autoimpute: Autoencoder based imputation of single-cell rna-seq data. *Scientific reports*, *8*(1), 1–11.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... others (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, *6*(5), 377–382.

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific reports*, *9*(1), 1–19.

Thomsen, S. F. (2015). Genetics of asthma: an introduction for the clinician. *European clinical respiratory journal*, *2*(1), 24643.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., ... others (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, *174*(3), 716–729.

Vert, J.-P., Tsuda, K., & Schölkopf, B. (2004). A primer on kernel methods. *Kernel methods in computational biology*, *47*, 35–70.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., & Hellmann, I. (2019). A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, *10*(1), 1–11.

Vos, V. S. S. (2016). *Methods for determining the effective resistance* (Unpublished doctoral dissertation). Masters thesis, 20 December.

Wan, C., Chang, W., Zhang, Y., Shah, F., Lu, X., Zang, Y., ... others (2019). Ltmg: a novel statistical modeling of transcriptional expression states in single-cell rna-seq data. *Nucleic acids research*, *47*(18), e111–e111.

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., ... Xu, D. (2021). scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature Communications*, *12*(1), 1–11.

Wang, L., Netto, K. G., Zhou, L., Liu, X., Wang, M., Zhang, G., ... Yang, M. (2021). Single-cell transcriptomic analysis reveals the immune landscape of lung in steroid-resistant asthma exacerbation. *Proceedings of the National Academy of Sciences*, *118*(2).

Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021). Direct comparative analyses of 10x genomics chromium and smart-seq2. *Genomics, Proteomics & Bioinformatics*.

Wills, P., & Meyer, F. G. (2020). Metrics for graph comparison: a practitioner's guide. *PloS one*, *15*(2), e0228728.

Young, G. F., Scardovi, L., & Leonard, N. E. (2015a). A new notion of effective resistance for directed graphs—part i: Definition and properties. *IEEE Transactions on Automatic Control*, *61*(7), 1727–1736.

Young, G. F., Scardovi, L., & Leonard, N. E. (2015b). A new notion of effective resistance for directed graphs—part ii: Computing resistances. *IEEE Transactions on Automatic Control*, *61*(7), 1737–1752.

Zhang, Y., Saradna, A., Ratan, R., Ke, X., Tu, W., Do, D. C., ... Gao, P. (2020). Rhoa/rho-kinases in asthma: from pathogenesis to therapeutic targets. *Clinical & Translational Immunology*, *9*(5), e1134.

Appendices

A Experts consulted

In this section, we provide a brief description of the experts that where consulted in the thesis process for biological and medical questions outside of the authors expertise. The information in this section was obtained from the *Staff Webpage of the University of Groningen* (2021).

G.H. Koppelman

Gerard Koppelman is a professor, pediatric pulmonologist and the head of the section of Pediatric Pulmonology and Pediatric Allergology at the UMCG and Faculty of Medical Sciences at the University of Groningen. His expertse is in pediatric pulmonology, genetics and environmental factors in asthma and atopy, and functional genetics. He is also the program director of GRIAC, the Groningen Research Institute of Asthma and chronic obstructive pulmonary disease (COPD).

M.C. Nawijn

Martijn Nawijn is an associate professor at the UMCG and Faculty of Medical Sciences at the University of Groningen. His expertise is in molecular genetics, cellular immunology, transgenesis and functional genomics. He leads an independent research group that researches the genetics drivers of the inception and remission of asthma and COPD using genome sequencing data.

M. Berg

Marijn Berg is a PhD-student in the field of bioinformatics at the UMCG and Faculty of Medical Sciences at the University of Groningen. His research and publications focus on asthma and COPD and make use of single-cell sequencing genomics data.

B Notes on implementation

A large proportion of the time spent on this thesis research went into implementing the methodology correctly in code. Each part of the research came with its own challenges and every implementation step was meticulously tested for correctness before we started on the next one. A conservative estimate of the total hours spent programming all the code needed to conduct this study is around 200 hours. Because of this large investment of time and effort, we wrote this section to highlight some of the challenges encountered in the implementation and to give a description of the different scripts used.

B.1 Implementation challenges encountered

Numerous challenges arose whilst working with single-cell sequencing data and trying to implement the numerous methods used in this study. This of demanded very large time investments to get things working and required us to learn many new programming skills. Examples of this are learning to work with the Bioconductor software packages (Huber et al., 2015) that are central to bioinformatics research or using the biological network visualisation tool Cytoscape (Shannon et al., 2003). Because the total list of challenges encountered is too long include in this document, we limit ourselves to describing just the following two.

B.1.1 Getting the data in the right format

Getting the data in the right format and matching the metadata to the expression data ended up being quite a lengthy process. At first we tried working with a collection of raw data files from the database. Multiple issues arose, however, when working with these raw data files. Firstly, the metadata was spread out over many different files that did not always have uniquely identifying keys linking cells in these files to each other to the cells in the count matrix. Secondly, we noticed that the metadata files were sometimes a concatenation of files in which the columns were not matched properly, which resulted in some of the columns not containing one single type of variable. Matching the count data with the metadata is necessary to be able to extract the correct observations from the data set and to not include observations from for example other types tissue or sequenced using other sequencing methods. After having made a serious effort to construct our data set from the raw data, using among other methods regular expressions and pattern matching to try to construct working keys, we had to conclude that this was not feasible with the current data set in the time frame of this thesis. We then opted to use switch from using these raw data files provided to me to using a different raw data set

that had been compiled by one of the researchers[20] at the UMCG from the same database and that had been used in earlier asthma research. From this data set we were able compile the observations relevant for our research and with pattern matching we were able to obtain batch information on the observations.

B.1.2 Computational power

Due to the large size of our data set and high computational costs of methods used in this paper, it was clear from the start that we would not be able to run all computations locally. For this reason, we got access to the high-performance computing (HPC) cluster of the UMCG[21]. This greatly expanded the range of methods that were computationally feasible for this research and significantly cut down on computation time. However, this did also mean having to learn how to work on a HPC cluster: learning about distributed computing, working in a Linux Shell and both writing and executing jobs using shell scripts on a distributed system. Getting familiar with these technologies took some time and to speed things up we followed a course on high-performance computing at the University of Groningen. The time spent becoming proficient in working on a HPC cluster was not wasted, since we were able to run almost all of the heavy computations on the cluster and we can confidently say that this study would not have been possible without it.

B.2 Description of scripts

This subsection gives a brief description of each of the scripts used for the research described in this thesis. These code scripts can be found in the following GitHub repository: Thesis_GRN_Comparison_CH.git. The repository follows the same structure as this thesis, with folders for the different parts and steps of the research. Scripts whose name ends in `_hpc` are written to be run on the HPC cluster.

part_1_data_preparation

filtering

`script_1.1.1_raw_data_processing.R`

R script that imports the raw data sources and merges them into the right format. The output

[20] This person is M. Berg, see Appendix A.

[21] More specifically: we used the Gearshift HPC cluster of the UMCG. For more information visit `http://docs.gcc.rug.nl/gearshift/`.

is one table for the gene expression data and one for the metadata.

`script_1.1.2_data_filtering.R`

R script that implements the data filtering steps described in Section 3. The output is the filtered expression data set.

normalisation

`script_1.2.1_normalisation_hpc.R`

R script that implements the scone workflow from Cole et al. (2019) on the HPC cluster. This workflow implements a variety of normalisation configurations and scores these configurations using different metrics. This comparison is described in detail in Appendix C. It outputs a data object containing both the scores from the comparison as the different normalised data sets.

`script_1.2.2_run_normalisation_comparison_hpc.sh`

Shell script used for running script 1.2.1 on the HPC cluster. It does not have any output itself.

`script_1.2.3_normalisation_comparison_output_analysis.R`

R script in which the output of the normalisation comparison is analysed and the best performer is chosen. It outputs the data set normalised using FQ-normalisation.

denoising

`script_1.3.1_run_scGNN_prework_hpc.sh`

Shell script that runs the prework routine of scGNN on the HPC cluster. It prepares the data for being used in scGNN by for example training the LTGM model. It outputs the data in the format necessary for running scGNN in script 1.3.2.

`script_1.3.2_run_scGNN_clustering_and_imputation_hpc.sh`

Shell script that runs scGNN on the normalised (and prepared) data on the HPC cluster. Its relevant outputs are the cell-type clustering and the imputed expression data set.

`script_1.3.3_scGNN_output_selection.R`

R script that creates the relevant subset of the imputed data set by filtering out cell types based on the cell-type clustering of scGNN. The output is the imputed expression data set.

`script_1.3.3_scGNN_output_analysis.R`

R script that analyses the effects of the imputation step on the data. It outputs the results of several tests and statistics which are described in Appendix D.

part_2_network_inference

`script_2.1_network_inference_hpc.py`

Python script that builds the networks from the imputed data set by running GRNBoost2 on the HPC cluster.

`script_2.2_run_network_inference_hpc.sh`

Shell script used for running script 2.1 on the HPC cluster. It does not have any output itself.

part_2_network_comparison

`script_3.1_prep_network_comparison.R`

R script that prepares the networks outputted by script 2.1 for the network comparison step. It converts the edge lists that GRNBoost2 outputted to weight matrices and creates `.csv` files for visualising the networks in Cytoscape. The outputs are these weight matrices and the `.csv` files.

`script_3.2_cutoff_and_effective_resistance.R`

R script that implements part of the network comparison. It implements the edge weight cutoff threshold and computes the effective resistance distance between all genes in the networks. It outputs the effective resistance distance matrices of both networks.

`script_3.3_create_and_compare_affinity_profiles.R`

R script that finishes the creation of the gene affinity profiles, implements the comparison of the profiles and reports the results. It outputs the results from the comparison and the boxplot presented in Section 5.2.

`script_3.4_visualise_connectivity_differences.R`

R script that creates the visualisations of the connectivity differences for the neighbourhoods of the different candidate genes (see Appendix E) and of the overlapping genes reported in Section 5.2.2.

C Intermediate results: Normalisation

In this section, we describe the data-driven process used for choosing the best normalisation procedure and provide a brief comparison of the normalised data set with the raw data set.

C.1 Choosing the normalisation procedure

This section describes the process of selecting the normalisation procedure. We based this selection on the approach described in Cole et al. (2019).

In this study, we compare a total of sixteen different configurations for normalising our data. An overview of the different normalisation methods used can be found in Table 4. The choice for including these normalisation methods in the comparison was made based on expected performance, technical diversity, computational feasibility and ease of implementation[22]. In addition to the normalisation of the data, we also analyse the effect of actively correcting for batch-effects, which means that we apply each normalisation method both with and without active batch-effect correction. Such batch effects may arise due to the fact that different sets of cells are processed on different lanes in the sequencing process, which are known to be able to cause some differences in count levels between lanes (Risso et al., 2011). However, it must be noted that many of the methods already implicitly correct a portion of these batch effects and it might thus be the case that an additional active corrective step might not be necessary in all cases.

Table 4

Normalisation techniques included in the comparison

Name	Description	Details
CLR	Centered log-ratio normalisation	$CLR(x) = [ln\frac{x_1}{g(x)}; ...; ln\frac{x_n}{g(x)}]$ where $x = [x_1; ...; x_n]$ and $g(x) = \sqrt[n]{x_1 x_2 ... x_n}$ (geometric mean)
DESeq	Relative log-expression scaling normalisation	Anders & Huber (2012)
EFF	Scaling by the number of detected genes	Divide each cell count by the number of detected genes.
FQ	Full-quantile normalisation	Bolstad et al. (2003)
scran	Simple deconvolution normalisation	Lun et al. (2016)
SUM	Sum scaling normalisation	Mortazavi et al. (2008)
TMM	Trimmed mean of M values	Robinson & Oshlack (2010)
None	No normalisation	-

[22]For example: SCnorm (Bacher et al., 2017) was excluded because of the large computational costs (an issue that was also raised in Cole et al. (2019) for this method) and implementing it in a computationally feasible way was not possible within the timeline of this thesis.

The resulting normalised data sets are scored using five of the performance metrics developed in Cole et al. (2019) that relate to different aspects of the distribution of gene expression. In order to get the best picture of the performance of each method, we used as many of the performance metrics from Cole et al. (2019) as possible. However, we did exclude the three metrics that concerned the association with controlgenes and quality control metrics, because we were advised not to rely on controlgenes in this research by an expert from UMCG[23] and because our data set did not contain quality control metrics. We do not deem the exclusion of these metrics as problematic because we find the other metrics more important for the purposes of this research and because Cole et al. (2019) found that preexisting batch classifications are better proxies for inner-batch effects than factors computed from quality control metrics or negative control genes for their Smart-seq data set. Table 5 provides a brief overview of the metrics used in this paper. These metrics are also described below, please see Cole et al. (2019) for a more detailed description.

The first three metrics asses each configuration by how well the cells can be grouped according to factors of wanted and unwanted variation. By assessing the quality of clusters we get a sense of how well certain types of variation are preserved in the data. Cluster quality is measured using silhouette widths Rousseeuw (1987), defined with respect to the Euclidean distance over the first three expression principal components. The other two metrics compare the global distributional properties between the samples. For these metrics, low scores that indicate similar global expression distributions are desirable. These metrics use the properties of the gene-level *relative log-expression* (RLE) rather than the log-counts directly as these are found to be more informative in Gandolfo & Speed (2018). Table 5 shows an overview of the metrics used for the normalisation method performance comparison. Each of these metrics yields a score for each normalisation procedure. Because of differences between the metrics in what constitutes a good score, we convert the scores to rank scores and compare these across procedures to select the best performing option.

Table 6 shows the results of the comparison of normalisation methods, where the methods are presented in order of average rank. As expected, we can see that applying batch-correction improves the score of the BATCH_SIL metric, which measures the removal of batch structure. Based the results in Table 6, we conclude that full-quantile (FQ) normalisation (Bolstad et al., 2003) without batch correction is the best performing normalisation method for our data set. We draw this conclusion based on the following:

[23]This person is M. Berg, see Appendix A.

Table 5

Performance metrics used for scoring normalisation procedures

Name	Property measured	Details
BIO_SIL	The preservation of biological difference	Clusters defined by subject group
BATCH_SIL	The removal of batch structure	Clusters defined by batch
PAM_SIL	The preservation of single-cell heterogeneity	Clusters defined by PAM clustering
RLE_MED	Global distribution between samples	Mean squared median RLE
RLE_IQR	Global distribution between samples	Variance of inter-quartile range of RLE

In the table, FQ normalisation clearly is a top performer for four of the five total performance metrics. Here, we would like to draw additional attention to the fact that it is the best method for preserving biological variation, indicated by its score on the BIO_SIL metric. The high score on this metric is particularly important for this research because our ultimate goal is to compare the genetic expression of the two subject groups. Its high performance in preserving biological variation and below average performance in removing batch structure are indicative of the trade-off between these two objectives that was mentioned earlier (see Section 2.1.2). We find the below average performance on removing batch structure acceptable for our research because of the following two reasons:
Firstly, because we favour preservation of biological variation in the trade-off mentioned earlier. Secondly, because we were assured by the researcher from the UMCG who compiled this data set[24] that the batch effects in this data, if present at all, were likely very small because special steps were taken in the sequencing procedure to minimise these effects. This claim is supported by the fact that the unranked scores of BATCH_SIL are all very close.

C.2 Descriptive statistics of the normalised data set

In the following, we will take a brief look at the results of the FQ normalisation (see Section 4.1) by comparing the descriptive statistics of the normalised data set (Table 7) with those from before normalisation (Table 1).

Here we see, as we would expect, no changes in the number of donors, the number of genes per cell and the cell distributions over the subject groups and batches. The goal of normalisation was to correct for unwanted observable differences in the data. One of these differences was the average read count per cell. In Table 1, we found that these were equal to 38.86, 30.71 and 35.09 for the asthmatic group, control group and full data set respectively. After normalisation, the

[24] This person is M. Berg, see Appendix A.

Table 6

Results of the comparison of normalisation procedures

Normalisation Method	Score rank					Average Rank
	BIO_SIL	BATCH_SIL	PAM_SIL	RLE_MED	RLE_IQR	
FQ	1	10	2	1.5	2	3.3
FQ *(batch)*	5	8	4	1.5	1	3.9
SUM	3	9	11	4	4	6.2
SUM *(batch)*	9	5	13	3	3	6.6
DESeq *(batch)*	12	2	7	5	8	6.8
DESeq	4	7	9	7	10	7.4
TMM	2	14	5	9	9	7.8
CLR	6	12	8	10	6	8.4
TMM *(batch)*	7	13	10	6	7	8.6
CLR *(batch)*	11	11	12	8	5	9.4
EFF *(batch)*	14	3	3	15	15	10
None	15	4	6	14	12	10.2
EFF	13	6	1	16	16	10.4
None *(batch)*	16	1	14	11	11	10.6
SCRAN *(batch)*	8	16	15	12	14	13
SCRAN	10	15	16	13	13	13.4

Note. For each performance metric, the five highest ranked scores are displayed in green.

Table 7

Descriptive statistics of the normalised data set

	Asthmatic	Control	Full data
Number of donors	9	9	18
Males (Females)	7 (2)	7 (2)	14 (4)
Total number of cells	1666	1428	3094
Cells in 1st batch	649	479	1128
Cells in 2nd batch	467	256	723
Cells in 3rd batch	468	619	1087
Cells in 4th batch	82	74	156
Percentage zero entries	75.11	74.96	75.04
Average read count	33.79	33.78	33.78
Genes per cell		14548	

average read counts per cell are now 33.79, 33.78 and 33.78 for those same groups. We thus see that the normalisation eliminated this and potentially other observable differences in the data. The fact that our normalisation method scored best on the BIO_SIL metric gives us confidence that the amount of biological variation removed is as small as possible.

D Intermediate results: Data denoising

This section presents an overview of the outcome of the denoising steps taken in the data preparation phase of the research.

D.1 Removal of immune cells and ionocytes, and gene filtering

First, the immune and ionocyte cells are removed from the data. In the original labelling, only 44 cells (1.4 percent) were identified as such. After this, the 2000 most highly variable genes were selected. The resulting data set consists of data on 3050 cells with each 2000 genes. By focusing on only the highest varying genes, the sparsity in the data is reduced from 75.04 percent to 42.30 percent (see Table 8).

D.2 Imputation

The goal of imputing the data was to reduce the sparsity in the data even further by transforming the data to correct for inaccuracies. As can be seen in Table 8, the percentage of zero entries in the data greatly decreased after the imputation step. The reduction of zero entries from 42.30 percent down to 7.25 percent causes our data to be more balanced and no longer be dominated by zeros. This is beneficial as it is expected to aid our network construction method later in inferring gene-gene relationships. Later in this section we perform several checks to asses whether the imputation had any adverse effects on the data. An important note here is that these percentages are relative to the filtered data set of 3050 cells and 2000 genes, meaning that approximately 2.15 million zero values were replaced with non-zero values. If we put this number in the context of the size of the original data set before filtering, we see that this represents around 2 percent of the data, which is in line with the expected rate of dropouts X. Wang et al. (2021), and that we decreased the overall sparsity in terms of the original data from 88 percent to 86 percent.

As is common to most popular single-cell imputation methods, in the process of imputing our data also non-zero values were changed. Such a transformation can improve the quality of the data when it reduces the noise in the data, the biological signal of interest is maintained and no false signals are introduced. In the following we perform some comparative checks on the data to get a sense in what ways the imputation changed the data. Although we must note that it is extremely difficult to understand the precise nature of the transformation of the data when using deep learning based methods like autoencoders. Because the goal of this research is ultimately to compare the genetic interactions in the asthmatic and control groups, our checks

Table 8

Descriptive statistics of the pre- and post-imputation data set

	Asthmatic	Control	Full data
Percentage zero entries (Pre)	41.97	42.68	42.30
Percentage zero entries (Post)	7.22	7.30	7.25
Average log read count (Pre)	2.582	2.558	2.571
Average log read count (Post)	2.584	2.559	2.572
Average %Δ per cell	0.189	0.143	0.168
Total number of cells	1635	1415	3050
Genes per cell	2,000 *(Highest variance)*		

mainly consist of comparing aspects of the data before and after imputation and whether or not we observe differences in these changes for our two subject groups. Observing similar transformations during imputation in both subject groups would indicate the method not treating the groups differently, which would give us confidence in the imputation process not distorting the signals in the data and not introducing false signals.

In Table 8, we see that the average log read count increases only very slightly during the imputation step and that this result does not differ for our two subject groups. We observe a similar result when looking at the average percentage change of the log read count per cell. When looking at the data in terms of changes per gene, we see that a subset of 96 out of the 2000 input genes had their expression completely reduced to zero. After splitting on subject group and the removal of ciliated cells (see Section D.3), this was 98 and 100 genes for asthmatic and control respectively. All 98 genes from the asthmatic group were also present in the 100 from the control group. This subset was the same for both our case and control group. A further inspection of the expression data of these genes before imputation showed that these were generally the genes with the highest percentage of zero's (on average 86.22 percent, more than twice the average of the total set of genes). The low pre-imputation levels of expression of these genes can also be seen in the average log read count, which was 0.437 (0.449 and 0.441 for asthma and control respectively), far below the total average of 2.571. Of the remaining 1904 genes, the average change in log read count per gene was 1.102 percent (1.102 and 1.102 for asthma and control respectively). There were 111 genes (the same for asthma and control) whose log read count changed more than 10 percent in absolute terms, though most of these changes were only slightly larger than 10 percent.

Because we ultimately want to analyse the relations between genes, the last check we do is to see how the expressions of genes changed in relation to each other. To do this, we calculated the average log read count of each gene both before and after imputation. We then ranked these averages from lowest to highest for pre- and post-imputation separately. By calculating the correlation between these ranks, also known as Spearman's rank correlation coefficient, we can asses how well the relationship between the average read counts can be described using a monotonic function, or put differently: how much the ranking of the average expression of the genes remained the same. The resulting correlation is 0.995 (0.994 and 0.993 for asthma and control respectively). This result shows that relative levels of expression where not heavily affected by imputation and gives confidence in believing that gene-gene relationships remained stable.

D.3 Clustering

The goal of the clustering was to split the data into smaller subsets from which we can select the ones containing basal and secretory cells. We used scGNN to cluster the cells by type based on their genetic expression (see Section 4.2.3). The default settings of scGNN have the number of clusters be determined by the Louvain algorithm. Under these settings the method tries to find a balance between finding good fitting clusters and minimising the number of clusters, which resulted in around five clusters being identified. In consultation with a researcher at UMCG[25], we decided to fix the number of clusters at ten. We did this because for our purposes finding more better fitting clusters is actually preferred as it allows us to more precisely select groups of cells belonging to the desired types.

Table 9 shows the clusters identified by scGNN compared to the the initial labelling already present in the data. We see that each of clusters found by scGNN is clearly dominated by cells labelled initially as either ciliated or basal and secretory. As we discussed in Section 4.2.6, this serves as validation of the good performance of our clustering method because it agrees in a majority of cases with the initial labelling. Because we are warned not to rely entirely on the initial labelling for separating ciliated from basal and secretory cells, we use this comparison to determine which cell type is most likely to be the one identified by each of the clusters found by scGNN.

[25]This person is M. Berg, see Appendix A.

Table 9

scGNN clusters compared to rough initial clustering and subject group

scGNN clustering	Initial labelling		Subject group		Total
	Ciliated	Basal & Secretory	Asthmatic	Control	
I	11	225	123	113	236
II	130	6	60	76	136
III	5	374	212	167	379
IV	1	96	51	46	97
V	692	13	394	311	705
VI	47	42	36	53	89
VII	31	51	46	36	82
VIII	8	179	112	75	187
IX	198	4	112	90	202
X	0	937	489	448	937
Total	1123	1927	1635	1415	3050

Note. The most prevalent cell type(s) according to the initial clustering in each cluster is (are) marked green.

Based on which cell type of the initial labelling dominates in the scGNN clusters, we conclude that the clusters I, III, IV, VII, VIII and X are the clusters with basal and secretory cells. The other scGNN clusters are considered to consist only of ciliated cells and are removed from the data set. This means that compared to the initial labelling, we exclude 65 cells that this labelling would mark as basal or secretory and include 56 cells that would have been marked as ciliated cells. This means that around four percent of cells are assigned a different type than that by initial labelling. Due to the unreliability of the initial labelling for edge cases, we continue under the assumption that the split we found using scGNN more accurately represents the underlying biological reality. After having these results checked by a researcher at UMCG with the relevant biological expertise[26], we removed the the ciliated cells and are left with 1918 cells that are concluded to be only basal and secretory cells (1032 asthmatic and 886 control). Having identified the cells that are of the types that are the focus of this research, there is still a small chance that a few cells are misclassified and are thus wrongly included or removed. Given that our data set consists of thousands of cells, we expect the effect of a few misclassified cells to be minimal and to cause no significant issues for downstream analyses.

Table 9 also shows that the subject groups are quite evenly distributed over the scGNN clusters. We see that the proportion of asthmatic cells in the selected clusters (containing the 1918 cells) is 53.86 percent, which is very close to that in the complete data (53.61 percent).

[26]This person is M. Berg, see Appendix A.

E Additional results: Regions of candidate genes

This appendix gives an overview of the regions of interest mentioned in Section 5.2 that belonged to the four genes found in our network comparison: NDUFA5, ALDOA, SNX3 and FAM3D. Details about the composition of these regions and the changes in internal connectivity can be seen in Figures 7, 8, 9 and 10. These figures are constructed in the same way as Figure 6.

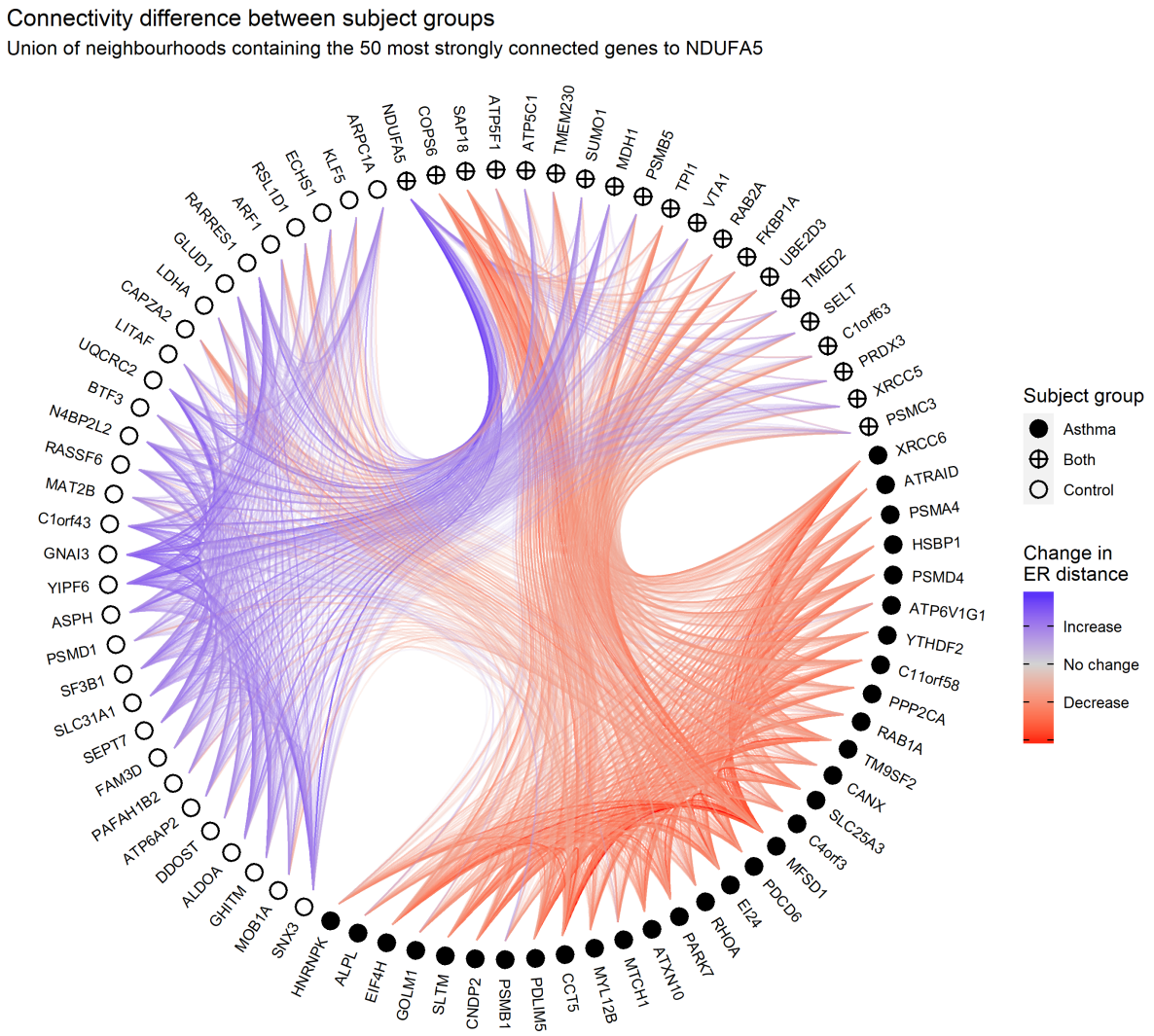


Figure 7. A visualisation of the difference in gene connectivity between the asthmatic and the control networks in the region of NDUFA5. Node filling shows if a gene belongs to the neighbourhood of NUDFA5 in either the asthmatic network, the control network or both. A red (blue) colored edge shows that the distance between genes is smaller (larger) in the asthmatic network. Only edges belonging to the top 50% of absolute changes in ER distance are drawn to reduce clutter.

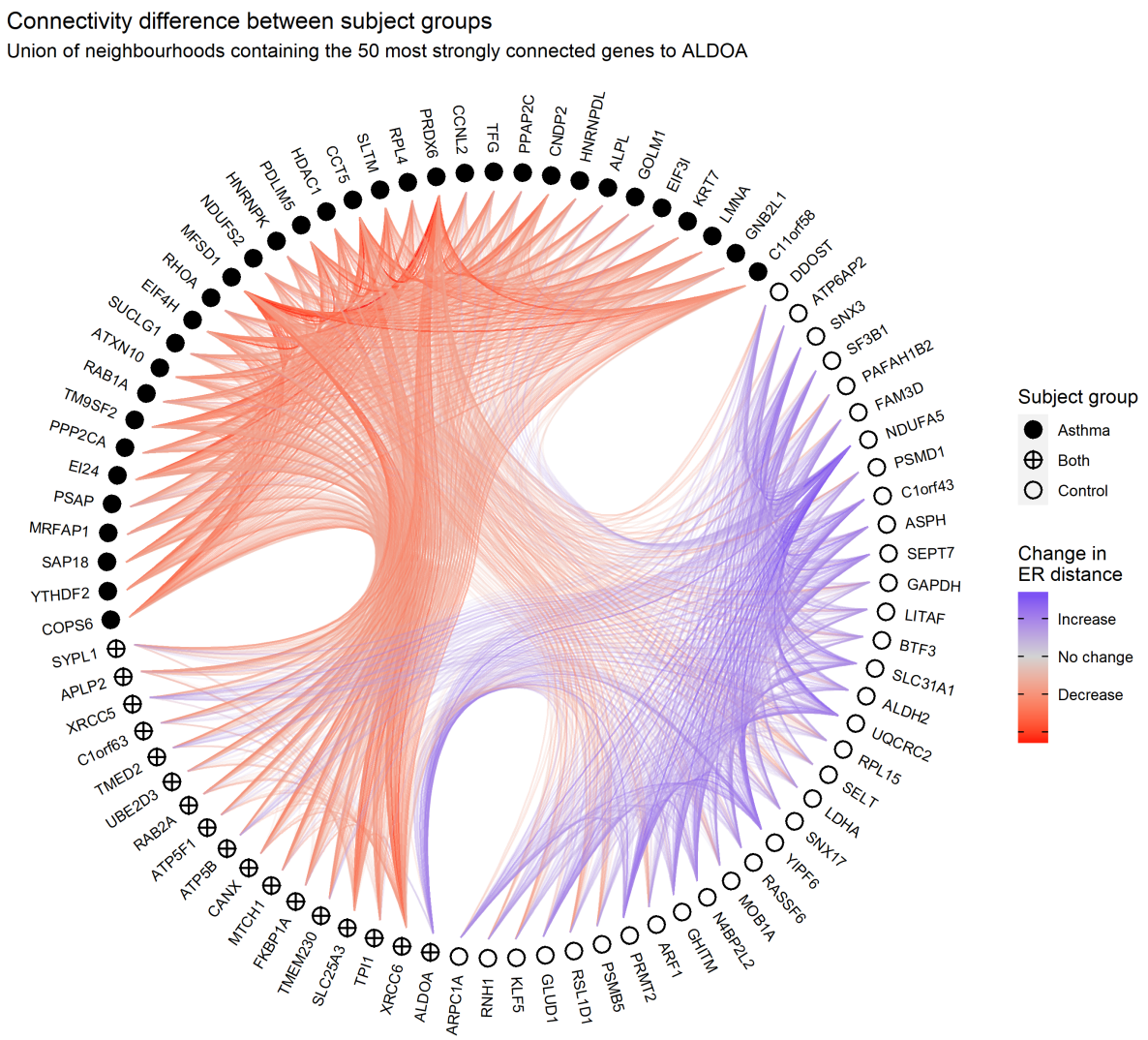


Figure 8. A visualisation of the difference in gene connectivity between the asthmatic and the control networks in the region of ALDOA. Node filling shows if a gene belongs to the neighbourhood of ALDOA in either the asthmatic network, the control network or both. A red (blue) colored edge shows that the distance between genes is smaller (larger) in the asthmatic network. Only edges belonging to the top 50% of absolute changes in ER distance are drawn to reduce clutter.

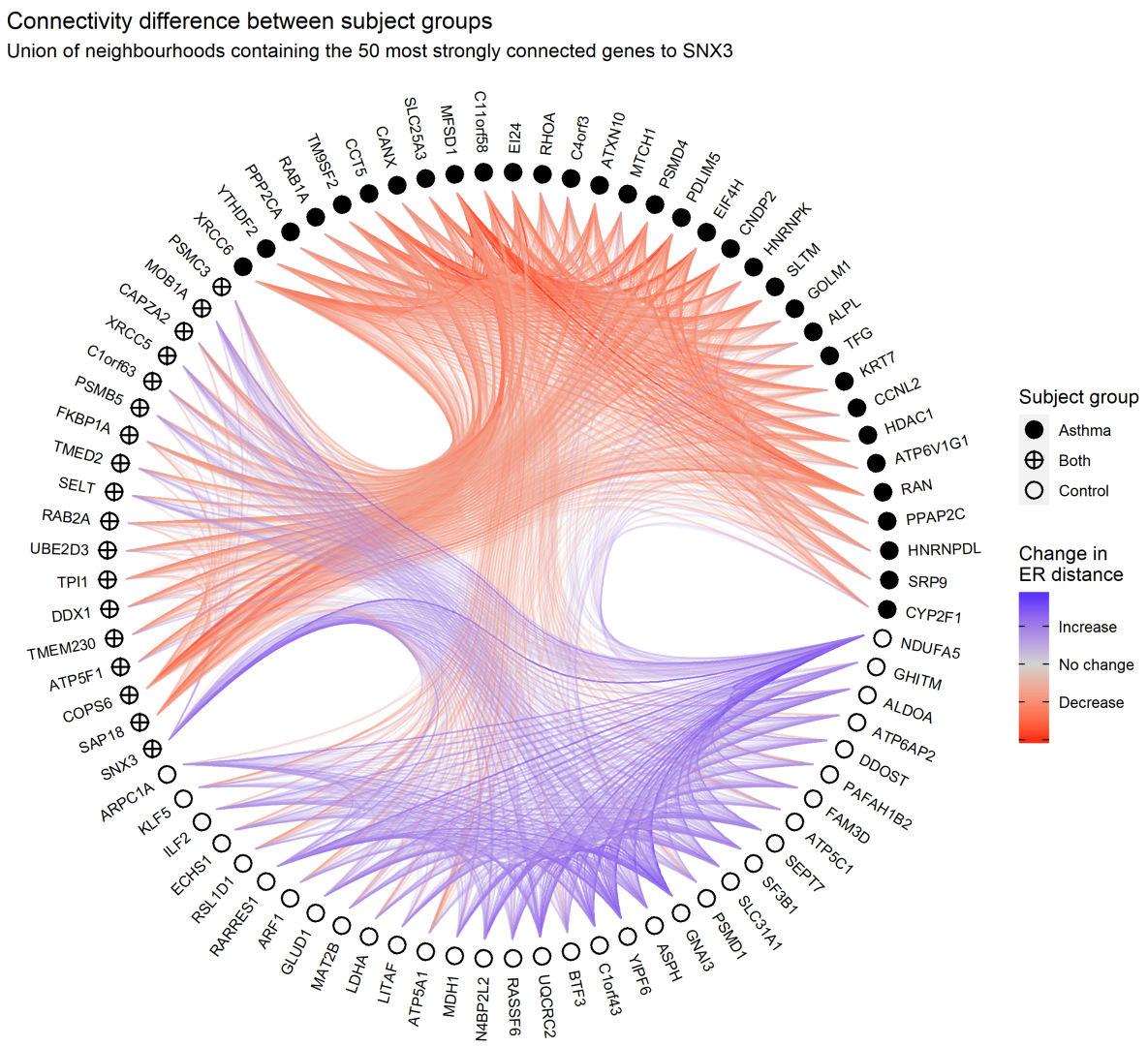


Figure 9. A visualisation of the difference in gene connectivity between the asthmatic and the control networks in the region of SNX3. Node filling shows if a gene belongs to the neighbourhood of SNX3 in either the asthmatic network, the control network or both. A red (blue) colored edge shows that the distance between genes is smaller (larger) in the asthmatic network. Only edges belonging to the top 50% of absolute changes in ER distance are drawn to reduce clutter.

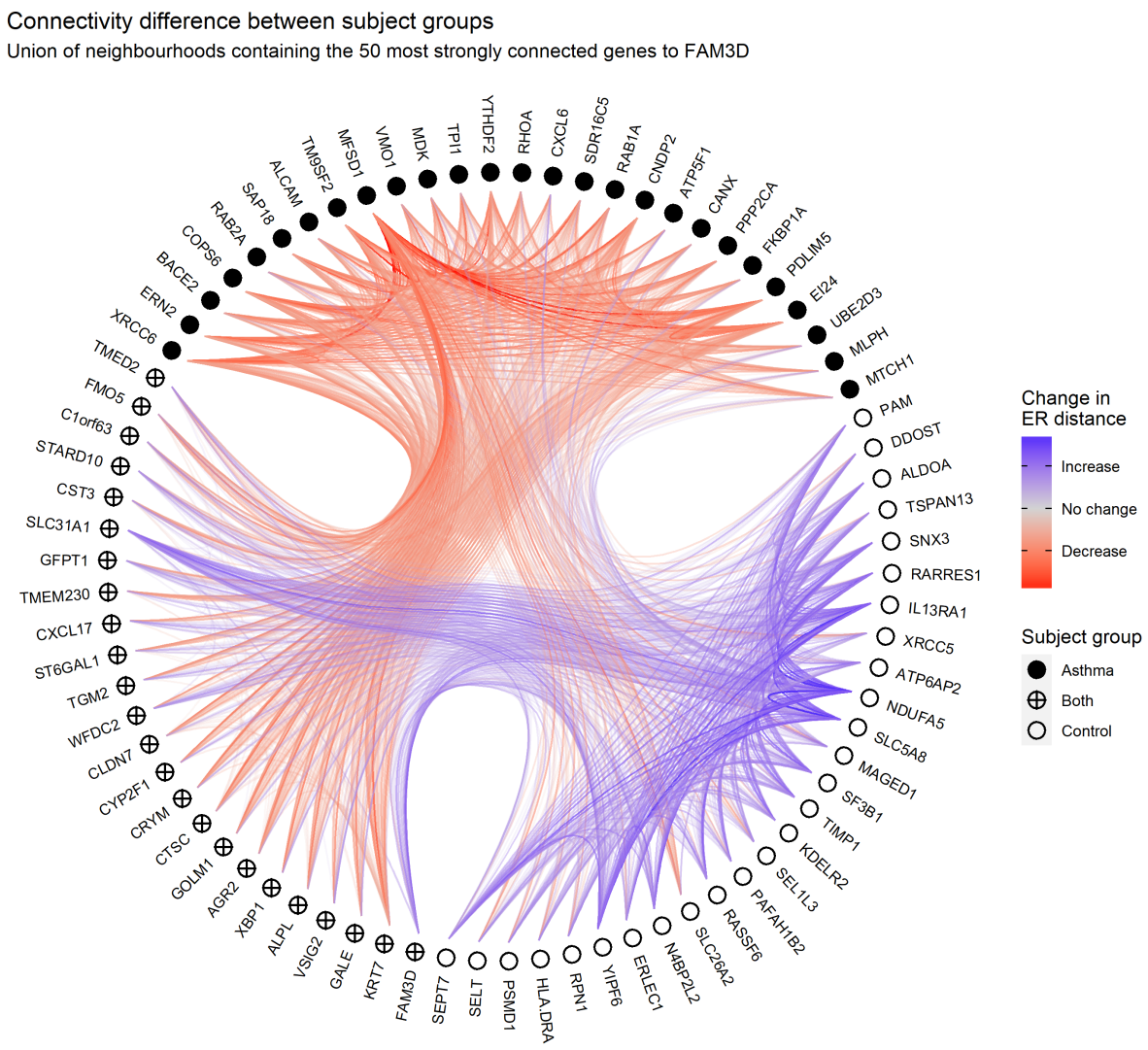


Figure 10. A visualisation of the difference in gene connectivity between the asthmatic and the control networks in the region of FAM3D. Node filling shows if a gene belongs to the neighbourhood of FAM3D in either the asthmatic network, the control network or both. A red (blue) colored edge shows that the distance between genes is smaller (larger) in the asthmatic network. Only edges belonging to the top 50% of absolute changes in ER distance are drawn to reduce clutter.