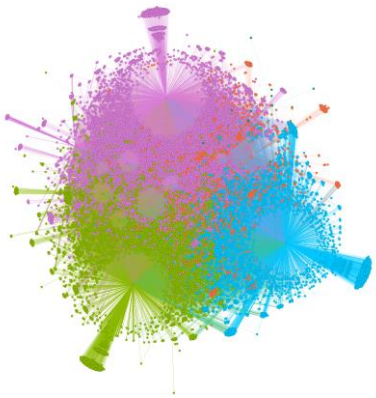


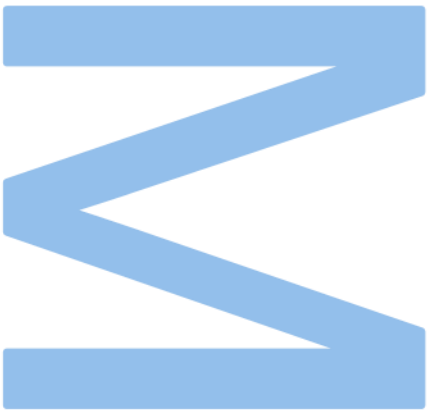
Unravelling the Complexity of Human Disease: Transcriptomic Networks of Phenotype – Gene Expression Data

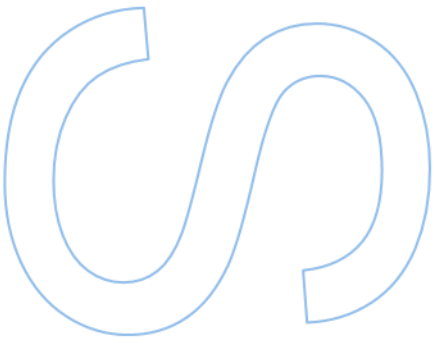
Darmit Kumar
Bioinformatics and Computational Biology
Department of Computer Sciences
2022/2023

Supervisors
Pedro Gabriel Dias Ferreira, FCUP & INESC TEC

Co-supervisors
Luísa Maria de Sousa Mesquita Pereira, IPATIMUP & i3S
Pedro Manuel Pinto Ribeiro, FCUP & INESC TEC


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





“Everything is interaction and reciprocity” -

Alexander Von Humboldt

Sworn Statement

I, Darmit Manish Kumar, born in 25/04/2000, resident in Porto, Portugal, phone number 916747565, of Portuguese nationality, bearer of Identification Card No.14961743, enrolled in the Master Degree Bioinformatics and Computational Biology at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Darmit Manish Kumar

30/06/2023

Acknowledgements

I would like to extend a thank you to Pedro G. Ferreira for believing in me and suggesting the topic of this dissertation, guiding me through the process and helping me when I was feeling lost. Your orientation and companionship were invaluable. Also, to Pedro Ribeiro for introducing me to the world of network science (”Networks everywhere”) during the network science course and for accepting to tutor me. To Luísa Pereira, I am very grateful for inviting me to work with her group at i3S, tireless support, friendship, and many shared laughs. A very fun and charismatic friend who always made time to help me. I learned a lot from the three of you, and sincerely hope we can keep in touch and work together in the future.

I am thankful to the group of Genetic Diversity at i3S, for taking me in and for all the direct and indirect contributions to this work. Your friendship and help were invaluable. Namely to: Rita Oliveira, Susana Seixas, Ana Magalhães, Joana Pereira, Verónica Fernandes, Carla Santos, Ricardo Pinto, Nicole Mineiro, Fernanda Rosário and Emanuel Matias.

Of course, I am thankful to all the professors who I contacted with during my master’s and bachelor’s degree for the dedication and knowledge you have provided me with. I would also like to thank Isabel Oliveira, my high school teacher, as in your classes I developed a wonder for the field of biology.

To all my friends that made this period of my life that much more special and fulfilling: no words can express the support you have given me through both the challenging and joyful moments, amidst remarkable adventures and occasional misadventures.

To my family I owe the biggest thank you and so much more. For being there for me since day one - 23 years of unconditional love, encouraging me to follow all my hopes and dreams, and making them possible.

Resumo

A regulação da expressão genética é determinada pela informação genética codificada no genoma, e é responsável pela determinação das caraterísticas visíveis de cada tecido. As doenças complexas resultam, normalmente, de alterações na expressão e regulação genética que podem resultar de fatores de origem genética ou de fatores ambientais e de estilo de vida que aumentam a predisposição a risco de doença. Características ou fenótipos demográficos como a idade, sexo, índice de massa corporal e ancestralidade estão comumente associadas a níveis de incidência e quadros clínicos diferenciados. Desta forma, o estudo da desregulação do transcriptoma em diferentes tecidos é um fator chave para a compreensão dos mecanismos moleculares associados à expressão do fenótipo.

A partir da análise de dados do consórcio *Genotype-Tissue Expression*, Garcia-Pérez e colegas (2022) procuraram compreender como é que a variação na expressão individual de cada gene se relaciona com alterações fenotípicas, disponibilizando a lista de genes com expressão diferencial em vários fenótipos e para cada um dos tecidos disponíveis. Neste trabalho, estendemos esta análise com uma metodologia de ciência de redes para caraterizar os sistemas e mecanismos de expressão individual, o comportamento pleiotrópico dos genes e a relação entre fenótipos demográficos e clínicos. Para o efeito, construímos um multi-grafo bipartido, que reflete a interligação de 18 649 genes, 17 doenças e 4 fenótipos demográficos em 46 tecidos, a que chamámos **PheGEx: rede humana de relação Fenótipo – Expressão Génica**. Analisámos a rede como um todo e comparámos os subgrafos específicos de cada tecido. Por último, interrogámos, de forma sistemática, pares de fenótipos através da aplicação do enriquecimento funcional de grupos de genes.

A nossa análise permite obter uma perspetiva integrada do sistema de expressão genética em amostras de humano. Para cada tecido e para os dados no seu todo, identificámos genes centrais e periféricos, e detetámos comunidades de genes e fenótipos. Encontrámos ainda genes, processos e vias distintivos ou subjacentes a grupos de risco, e que são comuns ou específicos de determinadas doenças.

Palavras-chave: doenças complexas, ciência de redes, expressão diferencial de genes, consórcio genotype tissue expression, epidemiologia, interações genéticas

Abstract

The regulation of gene expression is determined by the genetic information encoded in the genome and is responsible for determining the observable characteristics of each tissue. Complex diseases typically result from changes in gene expression and regulation, which can stem from genetic factors or environmental and lifestyle factors that increase the risk of disease. Demographic characteristics or phenotypes such as age, gender, body mass index, and ancestry are commonly associated with varying levels of incidence and distinct clinical presentations. Therefore, the study of transcriptome dysregulation in different tissues is crucial for understanding the molecular mechanisms associated with phenotype expression.

Based on data analysis from the *Genotype-Tissue Expression* consortium, Garcia-Pérez and colleagues (2022) aimed to comprehend how variations in individual gene expression relate to phenotypic changes. They provided a list of genes exhibiting differential expression across various phenotypes and for each available tissue. In this study, we expand upon their analysis using a network science methodology to characterize individual expression systems and mechanisms, the pleiotropic behaviour of genes, and the relationship between demographic and clinical phenotypes. To achieve this, we constructed a bipartite multi-graph that represents the interconnections among 18,649 genes, 17 diseases, and 4 demographic phenotypes across 46 tissues. This network, named **PheGEx: Human Phenotype-Gene Expression Relationship Network**, was analysed as a whole, and tissue-specific subgraphs were compared. Lastly, we systematically examined pairs of phenotypes by applying functional enrichment analysis to gene clusters.

Our analysis offers an integrated perspective on the gene expression system in human samples. For each tissue, as well as the dataset as a whole, we identified central and peripheral genes and we discovered gene and phenotype communities. We also identified genes, processes, and pathways that are distinctive or underlying to risk groups, and that are common or specific to certain diseases.

Keywords: complex diseases, network science, differential gene expression, genotype tissue expression consortium, epidemiology, genetic interactions

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALA	ACUTE LOBULAR ATELECTASIS
avPathLength	AVERAGE PATH LENGTH
BMI	BODY MASS INDEX
BP	BIOLOGICAL PROCESSES
CD	COMPLEX DISEASES
CSV	COMMA SEPARATED VALUE
C1	FIRST CERVICAL VERTEBRA
CPFE	COMBINED PULMONARY FIBROSIS AND EMPHYSEMA
dbGAP	DATABASE OF GENOTYPES AND PHENOTYPES
DEG	DIFFERENTIALLY EXPRESSED GENE
D1	DATASET 1
D2	DATASET 9
T1D	TYPE 1 DIABETES MELLITUS
T2D	TYPE 2 DIABETES MELLITUS
FDR	FALSE DISCOVERY RATE
FGEA	FUNCTIONAL GENE ENRICHMENT ANALYSIS
GML	GRAPH MODELLING LANGUAGE
GO	GENE ONTOLOGY
GO:BP	GENE ONTOLOGY BIOLOGICAL PROCESSES
GTEx	GENOTYPE - TISSUE EXPRESSION
HLA	HUMAN LEUKOCYTE ANTIGEN
HT	HASHIMOTO THYROIDITIS
KEGG	KYOTO ENCYCLOPEDIA OF GENES AND GENOMES
LCC	LARGE CONNECTED COMPONENT
lncRNA	LONG NON-CODING RNA
logFC	THE LOGARITHM BASE TWO OF THE FOLD CHANGE
MHC	MAJOR HISTOCOMPATIBILITY COMPLEX
miRNA	MICRO RNA
mRNA	MESSENGER RNA
nComponents	NUMBER OF COMPONENTS
nEdges	NUMBER OF EDGES
nNodes	NUMBER OF NODES
PCR	POLYMERASE CHAIN REACTION
PheGEx	PHENOTYPE GENE EXPRESSION NETWORK

RNA-seq	RNA SEQUENCING
SSNES	SUPRAPUBIC SKIN THAT IS NOT EXPOSED TO SUN
SNP	SINGLE NUCLEOTIDE POLYMORPHISM
TPM	TRANSCRIPTS PER MILLION
wavPathLength	WEIGHTED AVERAGE PATH LENGTH
wavDegree	WEIGHTED AVERAGE DEGREE
wDiameter	WEIGHTED DIAMETER

Chapter 1

Introduction

Human health and development are complex processes that result from interactions within and among multiple levels of biological organisation[1]. These interactions involve molecular mechanisms that require continuous and precise regulation. The dysfunction in the regulation of one or multiple genes can disrupt the genotype-phenotype relationship, which may give rise to disease [2,3]. Therefore, the study of these processes is essential for a better understanding of disease aetiology, identification of therapeutic targets, and discovery of development and disease biomarkers[2,4–6]. It is also observed that whilst Mendelian disorders result from variations in coding regions, common diseases usually result from changes in expression and regulation [7–9]. Therefore, the study of *gene expression regulation* is essential.

Human traits have frequently been associated with altered gene expression in the respective tissue. Thus, having a holistic view of the interplay between genes, tissues and traits is an important goal. *Network science* provides an integrative methodology to study real-world systems using graphs. It can help to incorporate multiple dimensions and types of data in a single model. It provides insight on the behaviour and structure, and on the role of each component in the context of the larger system. Therefore, it is an essential tool to study complex large biological systems.

The recent release of whole phenotype associated gene expression datasets by Garcia-Pérez, R. et al. 2022[10] shed new light on the individual phenotype – regulation interconnection. The description of gene expression for 17 clinical and 4 demographic phenotypes on 46 tissues from the analysis of 781 donors provides a unique opportunity to study the human transcriptome in detail. However, a system-wide understanding requires the incorporation of the multiple dimensions in a single model.

In this work, we apply *bipartite multigraph* methods to develop a human phenotype associated expression model of the organization and structure of these relationships, generating a new resource for the study of the transcriptome and its role in disease. With this analysis, we characterize the phenotype-regulation system across tissues, compare tissue-specific sub-systems, describe the processes involved in Hashimoto thyroiditis (HT), multiple lower tract lung injuries and diabetes, and shed light on the processes that underlie demographic clinical profile variations.

1.1. Gene Expression and Regulation

The DNA, through the process of gene expression, is responsible for the visible characteristics of organisms – the phenotype[11]. However, complex characteristics and behaviour are only possible due to the dynamic and powerful regulation of the expression of various genes[6]. In response to some perturbation of the system, genes will be differentially regulated, hence functionally modifying the molecular networks and/or interactions among them[1]. In this way, differentially expressed genes (DEGs) are an intrinsic occurrence that drive a myriad of biological processes (i.e., response to immune challenges, cell differentiation and morphogenesis), phenotypical differences (for example, age and sex-related traits) and pathological conditions (i.e., diabetes and steatosis)[6,12,13]. Hence, multiple phenotypes can result from the same set of genetic information[1,6].

Another crucial aspect is that the specialization of various types of cells, which form the diverse organs in an organism, also relies on the complex control of gene expression. Housekeeping genes play a vital role in regulating essential functions in all cells and are expressed in every cell. On the other hand, genes encoding specialized functions are selectively active in specific cells and inactive in others. Currently, the estimated number of protein-coding genes in the human genome is approximately 20,000[14], with recognized housekeeping genes ranging from 1,878 to 3,804[15,16]. Additionally, other types of genes, including long non-coding RNAs, have been identified as potentially significant. Although their annotation is still limited, they are now acknowledged for their biological relevance[17].

As the process of gene expression (copying a segment of DNA into RNA) is called transcription, the assembly and quantity of transcripts of a cell is called transcriptome. The control of gene expression can be done by different processes and at diverse points: epigenetic mechanisms, transcription initiation, transcript processing, transportation of messenger RNA (mRNA), regulation by noncoding regions of the RNA, translation, and post-translation.[6,13] The goal of this field is the complete understanding and exhaustive record of transcript molecules: mRNA, micro RNA (miRNA), long non-coding RNA (lncRNA), small RNA, splicing patterns, the structure of genes, post-transcriptional modifications, and changes in transcript levels[6,18]. The understanding of these aspects is of paramount importance to comprehending allelic heterogeneity, pleiotropy of complex traits and interpreting the functional components of the genome.

1.1.1. RNA Sequencing

The continuous advancement of sequencing technologies and computational power has resulted in the rise of research projects that systematically publish large biological datasets, allowing for a better understanding of expression, regulation, and the discovery of disease mechanisms[19]. For the study of the transcriptome, the most widely employed and vital techniques are Microarrays (and even these are becoming currently discontinued) and RNA sequencing (RNA-seq), which allow for a genome-wide analysis[6,19,20]. These increasingly cheaper high-throughput characterisations are displacing the traditional analysis of a small number of gene transcripts by quantitative real-time polymerase chain reaction (PCR) technology[6].

1.1.1.1. Genotype - Tissue Expression Project

Despite the colossal advances in sequencing and knowledge of the human genome, the majority of genetic changes are found outside protein-coding regions and often outside of genes as well[21]. This poses a great challenge to the discernment of mechanisms responsible for the emergence of the phenotype. Furthermore, the characterization of regulatory mechanisms is further hampered by the common inaccessibility of implicated tissues[22]. And most large-scale transcriptomic datasets lack high-resolution phenotyping[11].

With these issues in mind, the Genotype-Tissue Expression (GTEx, https://gtexportal.org/) consortium was designed to study the relation between genetic variation and gene expression[23]. This consortium genotyped, sequenced the transcriptome and profiled gene expression on multiple tissues from hundreds of donors[24]. The project relies on post-mortem samples from deceased donors to maximize the number of available tissues per individual[22].

The free availability of the data published by the consortium facilitates the discovery of regions of the genome that impact gene expression and the extent to which it occurs[24]. Additionally, it enables further investigation into the complex relationship between genetics and disease[23]. This, in turn, helps comprehend heritable susceptibility[24] to disease and make noteworthy advancements in the diagnosis, treatment, and prevention of a wide range of medical conditions.

While the published data by the consortium provides valuable insights, it is important to consider some associated limitations. The logistics of the project entails the incorporation of a bias in the data produced: (1) as the majority of the sampled population is of

advanced age; (2) between the time of death and tissue sample collection, there are changes in the expression of genes that must be accounted for[25]; (3) the biospecimen database mostly possesses diseased individuals[22]; and (4) relationships obtained from gene expression profiling analysis are often disease-induced rather than disease-causing[8,26,27].

The latest GTEx open-source RNA-seq data is included in the GTEx Analysis Version 8 (GTEx V8)[24]. The data (Figure 1) includes 17,382 RNA-seq samples from 948 post-mortem donors spanning 54 tissues. Additionally, it contains the genotype obtained from whole-genome sequencing of 838 donors. Donors' age ranges from 20 to 70, with most being elderly, with 557 being from the biological sex male and 281 the biological sex female. For each of the sample donors in the project the presence of several traits was recorded. Some clinical traits included are HT, pneumonia, emphysema, atelectasis, fibrosis, diabetes, pancreatic saponification, atrophy, hyperplasia, gynecomastia, spermatogenesis, steatosis, atherosclerosis, gastritis, and testicular sclerosis. The recorded demographic phenotypes are sex, age, BMI, and ancestry.

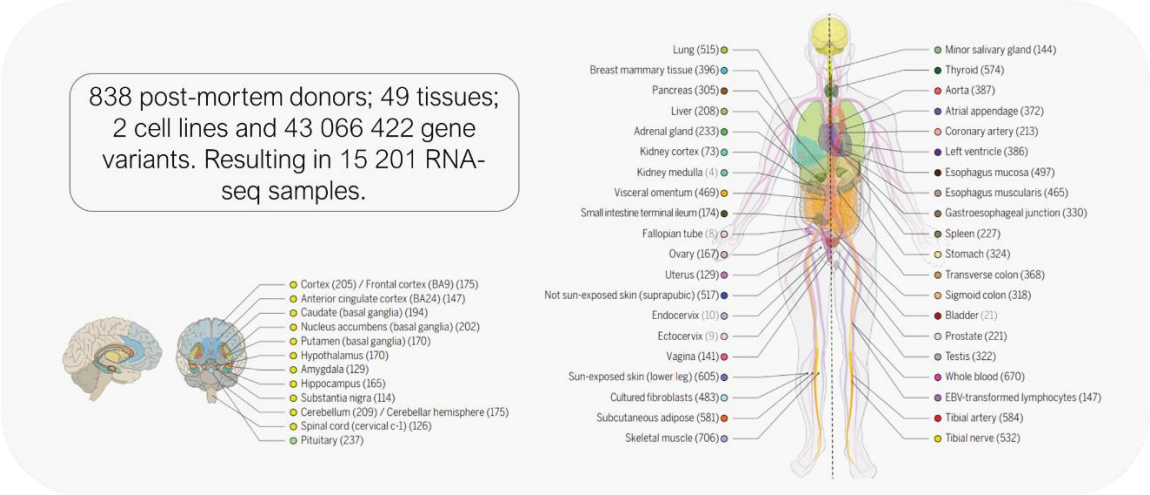


Figure 1 – GTEx atlas of genetic regulatory effects. Final dataset statistics are displayed in the text box on the top left corner. In the two illustrations, the 54 tissue types analysed by the consortium are visible with colour coding. Adapted from The GTEx Consortium [24].

1.1.1.2. Landscape of Expression Across Human Traits

While most previous expression studies relied on a restricted number of tissues for the analysis, Garcia-Pérez and colleagues (2022)[10], by analysing GTEx data, thoroughly report the effects of demographic factors on expression, the impact of clinical traits on their tissue of origin and evaluate whether the impact of demographic factors is system

wide or tissue specific. They found that interacting contributions from BMI, age, ancestry, and sex are rare. On the other hand, additive influences are pervasive. It was shown that expression variation is mostly genetically driven. The authors also conducted an extensive study of diabetic neuropathy and, with the aid of computer vision annotation of diabetic nerve slide images, retrieved multiple genes correlating tissue pathology with the condition.

1.2. Complex Diseases

Mendel's principles of segregation and independent assortment, the main assumptions of inheritance, are the foundation of our modern understanding of heredity. However, multiple biological processes lead to traits that violate these rules, namely: reduced penetrance, variable expressivity, phenotype definition, gene-environment, gene-gene, and polygenic interactions[28,29]. Due to the ubiquity of these phenomena, the patterns discovered by Mendel tend to be rare in disease contexts [7–9,30]. Instead, most common pathologies are categorized as complex diseases (CD) or multifactorial diseases, many of which do not follow clear inheritance models and can develop over time[1,8,9,31]. CDs result from the combination of multiple genetic, environmental and lifestyle factors[7–9,31] (Figure 2). Therefore, the intricacy of CDs is higher since the combined action of multiple factors may result in phenotypically similar states[8]. Hence, genetics is only part of the risk of developing a disease[31].

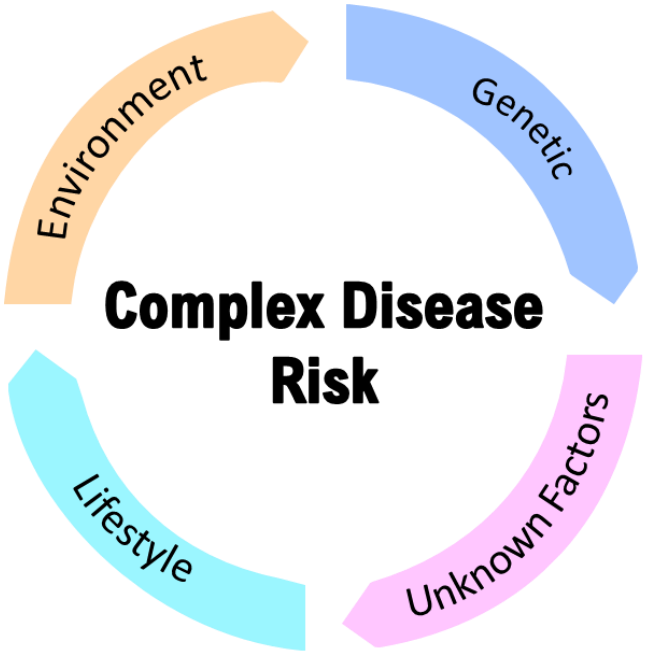


Figure 2 – Complex disease common risk factor decomposition.

Since individual factors and aetiology of these conditions are hard to derive, a common strategy is the study of molecular interactions that are related to the phenotype, such as genetics, epigenetics, differential expression, proteomics, and the effect of environmental factors, among others[6,8,9,32–35]. The study of related molecular mechanisms often relies on the comparison of "healthy" individuals with diseased ones, named as control and case cohorts[8,36,37]. Common techniques are the use of genome-wide association analysis and genome-wide expression profiling, to interrogate the genome and the transcriptome, respectively[3,6,19,20,24,33–35]. Genome-wide RNA expression analysis plays a fundamental role in biomedical and genomic research as it enables the identification of DEGs that may be implicated in disease processes and altered disease susceptibility.

Distinguishing causal contributions among characteristics that are disease-induced is an obstacle inherent in this class of traits[8,26,27]. Since phenotypes in CDs are heterogeneous, the analysis is hindered by confounding factors, such as population structure, cell type composition bias, effects and other unknown health factors[1,8,26,27]. To overcome these obstacles one strategy is to have a large and varied enough sample to include all possible sources of variation and identify the genes or interactions that most correlate to the disease[8]. Some common CDs are asthma, osteoporosis, and multiple sclerosis[31]. Other examples include Alzheimer's disease, scleroderma, Parkinson's disease, connective tissue diseases, kidney diseases and autoimmune diseases[31]. To better understand CDs, novel strategies to study related molecular mechanisms and distinguishing causal factors are needed.

In the following subsections we provide a short description of some of the complex traits in the datasets used in this work.

1.2.1. Hashimoto Disease

HT is an autoimmune thyroid disorder that causes chronic inflammation of the thyroid tissue. It is a complex disease that is an archetype for auto immune diseases[38,39]. Epidemiologically, it occurs with a much higher frequency in women than in men (4-10 times), it is most common among older demographics and displays geographic heterogeneity[40]. Furthermore, an increased incidence of auto-immune thyroiditis has been observed worldwide[40]. There is evidence for genetic susceptibility due to an increased number of cases in siblings of affected patients and there is a significantly higher concordance rate for auto-immune thyroid disorders among monozygotic twins than in dizygotic ones (29-55% and 0-7%, respectively)[39,40]. Even though strong epidemiological correlations with demographic factors such as sex, ancestry and age

have been well established, their influence and interaction with the development of the condition are not understood yet. Some studies hypothesize that the higher incidence among women is due to the condition with X-linked genes and skewed X-inactivation[41]. It is postulated that genetic alterations in conjunction with epigenetic and environmental factors might be triggering the autoimmune disease[38,39]. Radiation, iodine intake, drugs, infections, pollutants, pregnancy, and different chemicals are some environmental factors associated with the disease[38,40,42]. Several causal genes have been identified[39,42] but with a small individual impact on the development of the pathology[39]. On the other hand, combined action is responsible for the development of the disorder[39].

1.2.2. Lower Tract Lung Infections

Diseases of the respiratory system are highly prevalent and are associated with high morbidity and mortality rate across both industrialized and developing countries[43–46]. These are highly dependent on environmental and lifestyle factors and multiple epidemiological profiles can converge on similar phenotypes[44,47]. Some of the most preponderant susceptibility and severity markers are smoking and air pollution [44,46–48].

Pneumonia is an inflammatory response of the lung tissue due to the infection of the lower respiratory tract and typically leads to the accumulation of fluids and impairment of respiratory function[46,49]. Causes are generally of microbial or viral nature[46,50]. However, this distinction is hard to make in a clinical setting[46,50]. Research and diagnostics of the aetiology and progress of the condition are restrained as specimens are generally not from the affected tissue, hence hindering the identification of the causal pathogen[46,50]. Individuals below 5 and above 65 years old are notably vulnerable to this disease[49].

Emphysema is a chronic obstructive pulmonary disease where the progressive destruction of the alveoli parenchyma leads to an enlargement of the airspace[44,48]. It can sometimes coexist with fibrosis of the lower lung in the condition termed combined pulmonary fibrosis and emphysema (CPFE)[48].

Atelectasis is a common condition affecting about 90% of patients subjected to general anaesthesia[51,52]. Affected patients often have a collapse of around 5% of the lung area, however, this number can exceed 15-20%[51]. Therefore, the development of acute lobar atelectasis (ALA) is somewhat predictable in patients who might be predisposed to segmental lung collapse due to factors, such as a higher BMI or age[53]. Nonetheless, the pathophysiology of ALA is still not completely comprehended[53]. Proposed mechanisms are complete absorption of the airspace when the patient is exposed to pure oxygen;

compression due to the decreased intrapulmonary pressure; and due to decreased surfactant production in the alveoli[51,52].

1.2.3. Diabetes and Saponification

Diabetes mellitus is a heterogeneous metabolic pathology, characterized by a complex aetiology[54]. There are four categories of the disease. However, the two main classes are type 1 and type 2. In type 1 diabetes mellitus (T1D) the pancreatic beta-cells are destroyed due to the dysfunction of the immune system[54]. There is an increased risk for other autoimmune diseases in patients with type 1A diabetes. This association might be due to the strong relationship of the disorder with the human leukocyte antigen (HLA)[54]. However, idiopathic diabetes, or type 1B, are characterized by not having HLA associations[54]. It is uncommon for patients with this type of diabetes to be obese when diagnosed[54]. Type 2 diabetes mellitus (T2D) is the most common. Patients display insulin resistance and beta-cell dysfunction[54]. The pathogenesis of T2D is marked by genetic predisposition, a strong association for the higher incidence among close relatives and among some ethnicities, and environmental factors. Here obesity is a very high-risk factor due to its relevance in homeostatic regulation[54].

Saponification is a disease that also affects the pancreatic tissue[55]. When acute pancreatitis occurs, lipolytic fat necrosis with spots of saponification can ocurr[55].

1.2.4. Fibrosis and Atrophy

Fibrosis results from a rheumatic autoimmune condition and can affect multiple organs[56]. The condition is caused by the overproduction, accumulation, and remodelling of extracellular matrix proteins[56]. Fibrosis can be triggered by multiple disorders, being a common outcome to many of those diseases[48,56].

Atrophy is a condition where there might be a reduction in the size and function of a given tissue, such as skeletal muscle, testicle, and thyroid tissues[57–59]. It may arise from exposure to chemicals or drugs, lesion, inflammation and generally reduced biogenesis and destruction ratio[58,59].

1.2.5. Hyperplasia and Gynecomastia

Hyperplasia refers to the growth of a given tissue via the increase in the number of cells[60]. It can occur as a natural biological response, or it can correspond to a pathological

development, creating lesions. These can be mild and reversible or, in some cases, become more serious and lead to cancer[60,61]. In some cases, the synergy of genetic predisposition and diet can enhance this effect[60–62]. The pathology can impact different organs, for example, thyroid, endometrium, and adipose tissue[60–62]. In the thyroid, it is a consequence of the disruption of the feedback mechanism of thyroid-stimulating hormone and thyrotropin-releasing hormone[62]. The enlargement of breasts in men, gynecomastia, is generally a benign development of the mammary glands distinguished by hyperplasia of epithelium ducts and the growth of the connective tissues surrounding the duct[63].

1.2.6. Spermatogenesis

The process of sperm production in the seminiferous tubules is known as spermatogenesis[64]. The defective production or impairment of this process can cause male infertility. The reduced sperm count is a pathology termed oligospermia[65]. This condition might arise from the synergetic action of genetic, environmental and lifestyle factors, as strong associations with stress, exposure to heat, pesticides, air pollution, and Y chromosome microdeletions have been established[65].

1.2.7. Steatosis

Steatosis is a non-alcoholic fatty liver CD[66]. It is highly associated with T2D and obesity, and there are several proposed genetic risk factors. More severe conditions can lead to the simultaneous development of liver fibrosis[66]. The combination of the two has a poorer prognosis with a higher likelihood of cirrhosis and hepatocellular carcinoma. Epidemiologically it displays a higher incidence in men, which may be due to the greater distribution of fat around the abdomen and higher hepatic lipase activity in men[66]. It seems that it is more common among some ethnicities, such as Korean, and it is widespread in industrialized and developed countries[66].

1.2.8. Atherosclerosis

Chronic plaque accumulation, atherosclerosis, is the most frequent cause of pathologies in blood vessels[67]. The stenosis restricts oxygen delivery to the rest of the body resulting in myocardial infarction and stroke[67]. The build-up plaque generates a chronic immune response, hence creating a strong autoimmune component to the pathology profile of this disease[67]. Its' occurrence is strongly correlated with plasma and low-density

lipoprotein cholesterol and apolipoprotein levels. Genetic studies have identified multiple single nucleotide polymorphisms (SNPs) as risk factors. Atherosclerosis also displays an increased incidence among groups of smokers, hypertense or obese, or diabetes mellitus patients[67].

1.2.9. Gastritis

The inflammation of the gastric tissue, gastritis[68,69], commonly results from *Helicobacter pylori* infection, stress, drug use or from autoimmune damage to parietal cells[68]. Nonetheless, other bacterial agents, viruses, fungi, parasites, dietary factors, alcohol, bile reflux, radiation, food sensitivity and other idiopathic factors might also converge into a similar phenotype, even though these are rare in severe cases[69,70]. There are two main types of the disease, atrophic and non-atrophic[68,69]. The presence of the atrophy of the mucosa, through the dysfunction of gastric glands, constitutes a worse clinical profile due to the higher risk of progression to gastric cancer[69]. Gastritis is also reported to have higher incidence rates with the increase in age[71].

1.2.10. Testicular Sclerosis

Sclerosis is the stiffening of the tissue[72]. When the affected tissue is the testis, it commonly refers to the damage of the seminiferous tubules through thickening of the lamina propria[72,73], which may result in organ dysfunction[73]. Sclerosis of the Sertoli cell, even though very rare, is also possible in cases of testicular cancer[74].

1.3. Network Science

Systems are ubiquitous in the natural and built world, encompassing a wide range of phenomena, structures, interconnections, and ecosystems (*sensu latu*)[75]. Real-world systems, however, are intricate, vast, sparse, and difficult to comprehend and visualize[75,76]. They encompass electrical grids, communication infrastructures, street roads, social networks, scientific article co-authorship, interactions between genes, proteins, metabolites, neurons, and countless others.

In the field of systems biology, the ever-growing size and number of biological datasets created the need for more efficient processing, mining, and unravelling of complex multidimensional organic systems. For example, the study of disease aetiology needs to embrace the pleiotropic and heterogeneous nature of complex diseases where

phenotypes are likely to reflect syndromic states composed of multiple disease subtypes[9]. It has been verified that even in “simple” mendelian diseases, non-integrative strategies may lead to shortcomings since patients with the same SNP can present distinct phenotypes[9]. Such is the example of sickle-cell anaemia, where the same mutation can lead to different clinical profiles[9]. Further comprehension of disease under the context of genetic, molecular, and environmental context is essential. Therefore, in contrast with traditional reductionist approaches, systems biology problems need to be tackled with a holistic and quantitative method.

Network Science is a field of research, dedicated to the modelling and study of complex systems, which are represented as graphs. This system representation strategy enables the condensation of large amounts of data and the production of intuitive visualizations of complex sparse data. Looking at the system as a whole can provide more insights than simple statistical metrics such as mean and variance[33,77]. This approach offers an additional advantage over traditional machine learning algorithms by leveraging relationship information and intensity, thereby preserving the internal structure of the dataset[77]. Graphs exhibit a similar structure and topology across various domains. This similarity allows for the development of common analysis strategies and offers a methodology and space for different knowledge areas to converge using a common language on the exploration of their respective datasets[9,75,77]. Such is the case of the production of differential and disease-associated networks for a given patient which have been important for the development of precision treatments[1,32,33,78,79]. Nonetheless, the field should be aware that employing network science has the drawback of requiring some degree of simplification of biological systems[1]. The dynamic representation of natural systems involves the study and record of multiple different conditions to produce several static graphs[1]. Despite these limitations, the gains clearly justify the pursuit of this approach.

Network components or states are called *nodes* (Figure 3, panel A)[9,75]. Interactions among nodes are represented as *edges* (Figure 3, panel A). Edges can represent probability, connection, influence, and correlation, among others. Nodes and edges can have one or multiple categorical attributes that further describe the system (i.e., node/edge type or ranking; Figure 3, panel B). There can be a weight associated to each edge that denotes the intensity of the interactions [9,75]. Often represented with the thickness of the edge (Figure 3, panel C). Continuous variables that describe the entity (i.e., centrality measures) can be mapped to the size of the node (Figure 3, panel C).

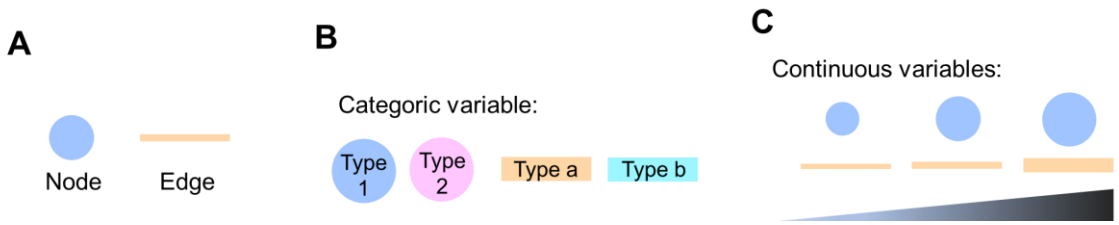


Figure 3 – Graph elements. Panel A: illustration of a node and an edge. Panel B: visual mapping of categorical variables by assigning a different colour to each category. This strategy is used in both nodes and edges. Panel C: Continuous variables are normally mapped to the size of the node or thickness of the edge. The thickness usually corresponds to the weight.

A graph G is denoted as G = (V, E), where V is the set of nodes or vertexes, and E is the set of edges. When relationship direction is not applicable, the network is termed undirected [9,75]. A directed graph (or *digraph*) implies that interactions have a specific origin and end point, so edges are represented as arrows instead of lines (Figure 4)[9,75]. Graphs that allow for multiple edges to be established among nodes (multi-edges) or edges that connect nodes to themselves (self-loops) are called *multigraphs* (Figure 4)[80]. They are, otherwise, known as *simple graphs* (Figure 4).

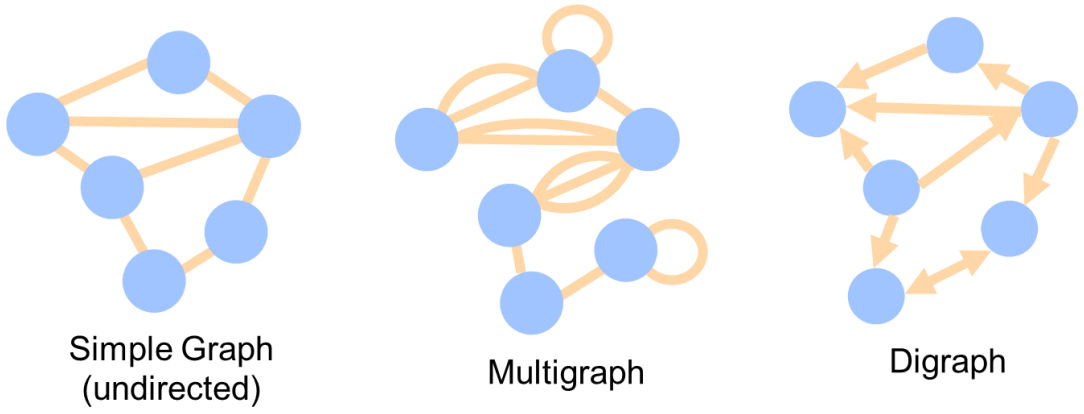


Figure 4 – Common network types: simple graphs, multigraphs, and digraphs. In the multigraph several edges can be observed between a single pair of nodes and self-loops are present. In the digraph the orientation of the edges represents the direction of the relationship between the pair of nodes.

In a graph not all nodes are accessible to all others. Each group of nodes where all nodes are reachable to one another is called a *component*. If the largest component contains a large portion of the nodes in the network, it is called a *giant component* or a *large connected component* (LCC).

Bipartite graphs are composed of two disjoint sets of nodes (Figure 5)[9,75,81]. Here, edges can only be established between a node from one set and a node from another set. Therefore, each set of nodes can be interpreted as a dimension[9,81]. The bipartite graph

can then be projected onto one of the dimensions, often by connecting nodes from the chosen dimension that are connected to the same node in the opposite set[9,81]. This definition can be extended to a multipartite graph by increasing the number of disjoint sets that cannot be connected within themselves[75].

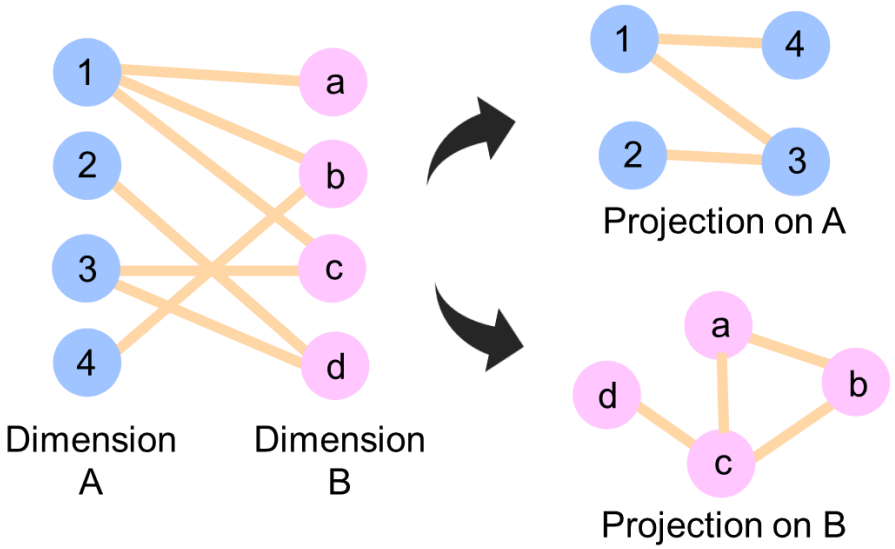


Figure 5 – Bipartite graphs and projections. In a bipartite network there are two sets of nodes or dimensions. Connections can only be established from one set to another. The standard process for the transformation of a bipartite network into its projections is also showcased. Here nodes from the same dimension are connected to one another if they have a common neighbour in the bipartite network.

From a graph, we can obtain an *ego network* with depth N for any given node. Given a graph G, a node *A* and depth N, the ego network of A is constituted by all nodes reachable from A in N steps (Figure 6).

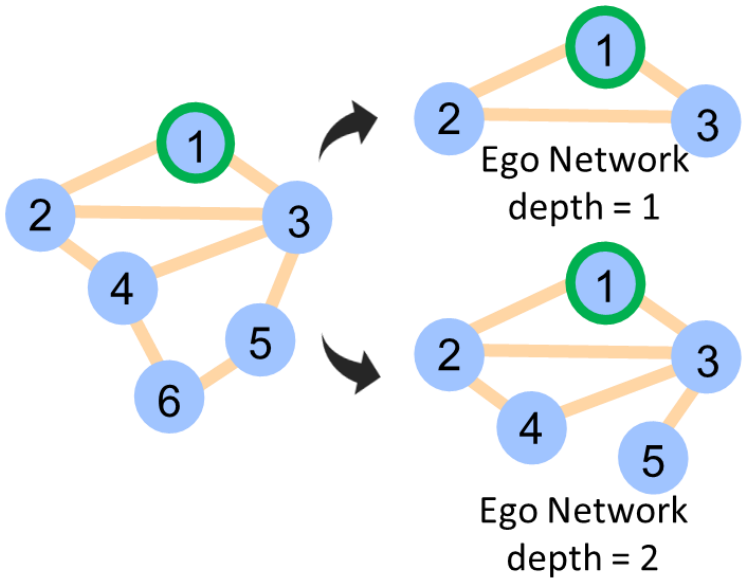


Figure 6 – Ego-networks. An ego-network operator is applied to node 1 of the network on the left. On the right, we exemplify the result with two ego-networks. The first with depth 1 and the second with depth 2.

Another operator commonly employed in the field of network science is the simplification of multigraphs. Where multiple edges between two nodes are compiled into one, and self-loops are removed. The number of edges (nEdges) between two nodes can be kept as the weight of the replacement connection (Figure 7).

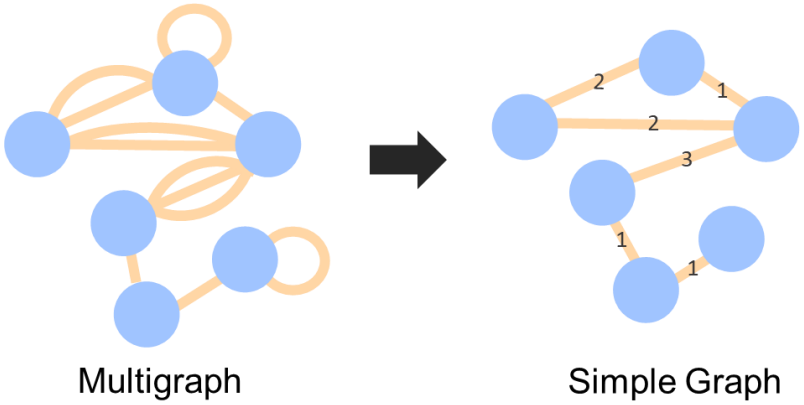


Figure 7 – Simplification of multigraphs. The process of simplification of a multigraph is exemplified. Multi-edges are converted into a single edge and self-loops are removed. The weights of the final weighted simple graph are displayed on top of each edge for exemplification purposes.

Network analysis includes the identification of communities through the process of unsupervised clustering of nodes[75,82]. *Communities* are groups of data objects connected strongly among themselves but weakly with the rest of the network[9,75,82] (Table 1).

Table 1 - Popular community finding algorithms[82].

Algorithm	Description
Girvan-Newman	Hierarchical divisive algorithm that progressively removes edges with the highest betweenness value.
Louvain	(1) Modularity is maximized by joining nodes with one of their neighbours. (2) Communities are collapsed into single nodes with self-loops. Both steps are repeated until no modularity increases are possible.
CFinder	Finds overlapping communities through the method of clique percolation.
Random Walks	Similarity between nodes is estimated, often, using a random walk with a restart strategy.

Identifying the nodes which are the most relevant to the network according to a pre-established definition of importance is accomplished with *centrality* measures[9,75] (Table 2).

Table 2 - Different centrality measures and respective definitions.

Metric	Description
Degree	Number of connections of each node.
Betweenness	Number of times a node is involved in a shortest path between two nodes.
Closeness	Average distance between a node and all others.
Eccentricity	Maximum distance between a node and any other node.
Harmonic	Sum of the inverse distance between a node and all other nodes.
Eigen Vector	Reflects the relative importance of the node by taking into account the centrality of the node and its neighbours.
PageRank	Recursive measure based on the number of incoming links and on the importance of its neighbours until convergence is achieved.
HITS	Measures the authority and hub scores of nodes based on the structure of incoming and outgoing links. Nodes pointed to by many authoritative nodes are more authoritative, while nodes with many outgoing links to other authoritative nodes are good hubs.

1.4. Functional Gene Enrichment Analysis

The coordinated action of functionally related genes is often necessary and employed to accomplish their respective roles in the cell[83]. This may be important in CDs, where it is postulated that pathologies arise from the interaction of multiple genes that have individual weak impact[83,84]. Genome-wide studies frequently generate large lists of genes, rendering it difficult to reach biological interpretations. The technique Functional

Gene Enrichment Analysis (FGEA), by considering patterns in a large set instead of individual genes, targets the detection of coordinated expression changes[83,85].

FGEA is a computational method that determines if a set of genes or genetic variants is significantly associated with specific biological functions or pathways[86]. It helps identify the functional implications and potential roles of genes or variants by assessing their enrichment in known functional categories[86], or Gene Ontology (GO)[85,87] or Kyoto Encyclopedia of Genes and Genomes (KEGG)[88] terms.

The enrichment is calculated by comparing a set of genes or genetic variants of interest to a reference set, such as the entire genome, using statistical methods like the hypergeometric test or Fisher's exact test[86]. The analysis determines the statistical significance of the overlap between the genes of interest and functional categories or pathways, providing insights into the potential functional implications of the genes or variants[86].

Chapter 2

Material and Methods

2.1. Dataset

Garcia-Pérez, R. et al. 2022[10] explored the landscape of expression of demographic and clinical phenotypes of the GTEx database. The multiple differential expression metrics and results were published online with open access (10.5281/zenodo.6797627) in the form of 10 datasets.

In this paper, a restricted dataset was built from the public release GTEx V8[24], employing the methods outlined in this subsection. The study authors downloaded transcripts per million (TPM) quantifications for all protein-coding and lncRNA biotypes, excluding pseudoautosomal regions. Genes with a TPM larger than 1 and that have more than 10 reads in at least 20% of sampled tissues were considered. Expression data was then restricted to tissue sources with at least 100 RNA-seq samples; that include demographic information (sex, age, BMI and ancestry) and that possess metadata for covariates of the expression and splicing analysis (parameters of donor’s death, ischemic time, RNA integrity number and sequencing quality control metrics). The above-mentioned demographic trait, ancestry, is the genetically inferred ancestry. However, only European American, and African American samples were included, offering a reduced ability to

explore expression variation across ancestries. As a result, the authors quantified gene expression variation with demographic factors for 22,967 genes (18,185 protein-coding and 4,782 lncRNA), on 781 donors, spanning 46 different tissues.

The authors annotated clinical traits of each donor from medical records available in the database of Genotypes and Phenotypes (dbGap) which were then made to correspond to all tissues from that donor. Histopathological annotations were manually created and are tissue-specific, therefore the identified condition corresponds only to the tissue sample. Phenotype terms with related or similar pathologies were sometimes combined into a single term. Such was the case of the Atherosclerosis phenotype that was made to include artery samples annotated as: atherosis, atherosclerosis, calcification, Mönke-bergs arteriosclerosis and/or sclerosis. Spermatogenesis related pathologist comments were analysed and modified to distinguish samples with normal and low levels of spermatogenesis. Some clinical annotations are only available for a subset of donors and samples, therefore only traits with more than 5 affected samples per tissue were considered. Traits were also removed when directly related to the donor's cause of death, as these are strongly correlated with the hardy scale index. Patients diagnosed with T1D expressing insulin were also deleted. So, from the initial 48 different clinical traits, authors ended up with only 17 in the final dataset.

Garcia-Pérez, R. et al. detected DEGs in the derived dataset, through the limma-voom framework. A linear-regression model was implemented for each of the 46 tissue sources and adjusted for the technical covariates above-mentioned. The regression model also included cell type heterogeneity and sequencing batch PEER factors that were inferred as hidden influences through the exploration of unknown sources of variation. The log(cpm) expression values in each tissue were assessed for their statistical significance for every regressor. Thus, a first regression model was created for the demographic factors, including the PEER factors, age, ancestry, BMI and sex for every gene in each tissue. When a tissue was not present in both biological sexes, the regressor sex variable was omitted.

$$expression(log2cpm) \sim technical\ covariates + Age + BMI + Sex + Ancestry$$

For the second model, Garcia-Pérez et al. matched the clinical traits with an origin tissue, based on the known aetiology of the disease. Histopathological derived traits were matched to the tissue that was used for the diagnosis and in cases where there is more than one origin tissue the trait was matched to the tissue with the highest sample size. Furthermore, only tissue samples with available annotations for all clinical traits were

included. Finally, the model was applied to examine DEGs related to the clinical traits in their tissue of origin.

$$expression(log2cpm) \sim technical\ covariates + Age + BMI + Sex + Ancestry + clinical\ trait(s)$$

In this work, we used datasets 1 (D1) and 9 (D2) published by Garcia-Pérez, R. et al. 2022 [10].

- **D1: characterizes DEGs across tissues with the demographic phenotypes age, sex, ancestry, and BMI.**
- **D2: characterizes DEGs across tissues with the clinical traits from GTEx V8 with sufficient sample size in the tissues of origin.**

Two crucial metrics reported in the data are the logarithm base two of the fold change (logFC) and the corrected multiple testing p value using the false discovery rate obtained through Benjamin-Hochberg procedure (FDR).

As expected, the combined datasets (Figure 8) possess expression data for all tissues for all the demographic traits, age, ancestry, and BMI except for sex on tissues that are exclusive to one of the sexes. T2D is the only clinical trait represented in all tissues and T1D is also mostly well represented. Other than diabetes, only the clinical phenotypes fibrosis and atrophy have data for more than one tissue.

In summary, the data to be analysed in this work comprises:

- Two datasets, D1 and D2, of genes differentially associated with 4 demographic factors and 17 clinical phenotypes;
- Clinical phenotypes are HT, pneumonia, emphysema, atelectasis, fibrosis, T1D, T2D, pancreatic saponification, atrophy, hyperplasia, gynecomastia, spermatogenesis, steatosis, atherosclerosis, gastritis, testicular sclerosis, and congestion;
- Demographical phenotypes are sex, age, BMI, and ancestry;
- Derived from the analysis of about 23,000 genes on 46 tissues from 768 donors;
- A total of 4,233,722 analysed samples for tissue-trait pairs.

Figure 8 shows the representation of the number of DEGs per phenotype and tissue pair.

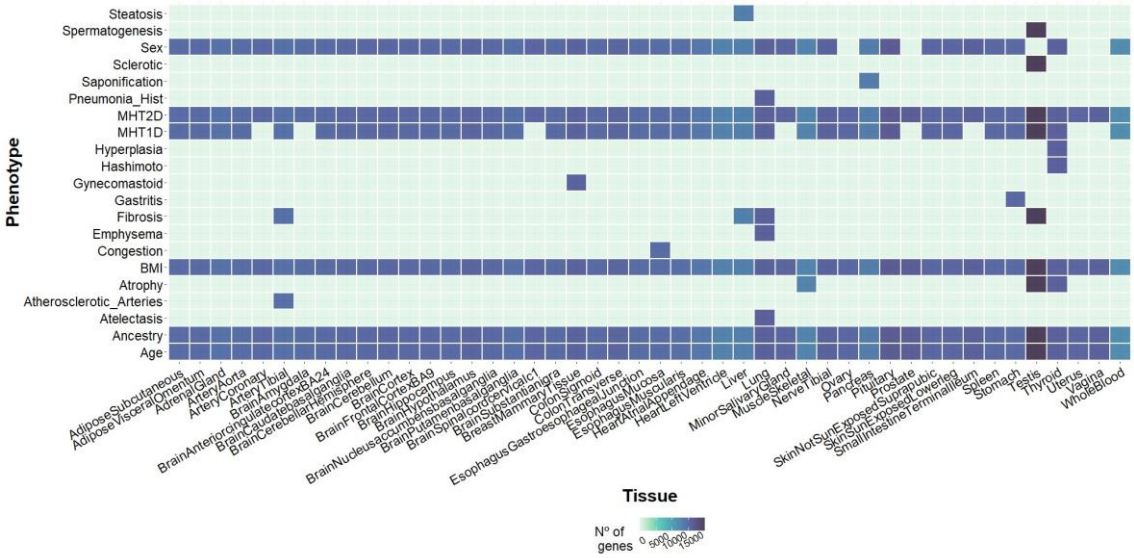


Figure 8 – Number of genes analysed per tissue – phenotype pair. Adapted from D1 and D2 published by Garcia-Pérez, R. et al. 2022.

2.2. Phenotype – Gene Expression Network

We obtained the Phenotype-Gene Expression network (**PheGEx**) list of DEGs by iteratively retrieving gene–tissue–trait triplets with an FDR below 0.05 (Figure 9). It should be noted that D2 only analysed the tissues of origin of the considered pathologies and, due to the authors' effort to reduce demographic bias, both demographic factors and pathologies were included in the regression analysis of clinical traits. However, D2 aims to describe the differential expression associated with diseases. Thus, we only used it to retrieve DEGs related to clinical traits. For each triplet we also obtained the associated logFC and the tissue where the DEG was detected. We used the logFC to derive a logical variable "overexpressed" that is true when logFC > 0 (over expressed genes) or false when logFC < 0 (under expressed genes). A second variable was produced with the absolute logFC. The PheGEx structure was produced using the function "graph_from_data_frame" from the R package igraph. The tissue and "overexpressed" variables were incorporated as edge type attributes and the absolute logFC was used to encode the edge weight. PheGEx was then saved in a comma separated value (CSV) and a graph modelling language (GML) file. The network was analysed and visualized using igraph and the software Gephi[89], using the CSV and GML files, respectively.

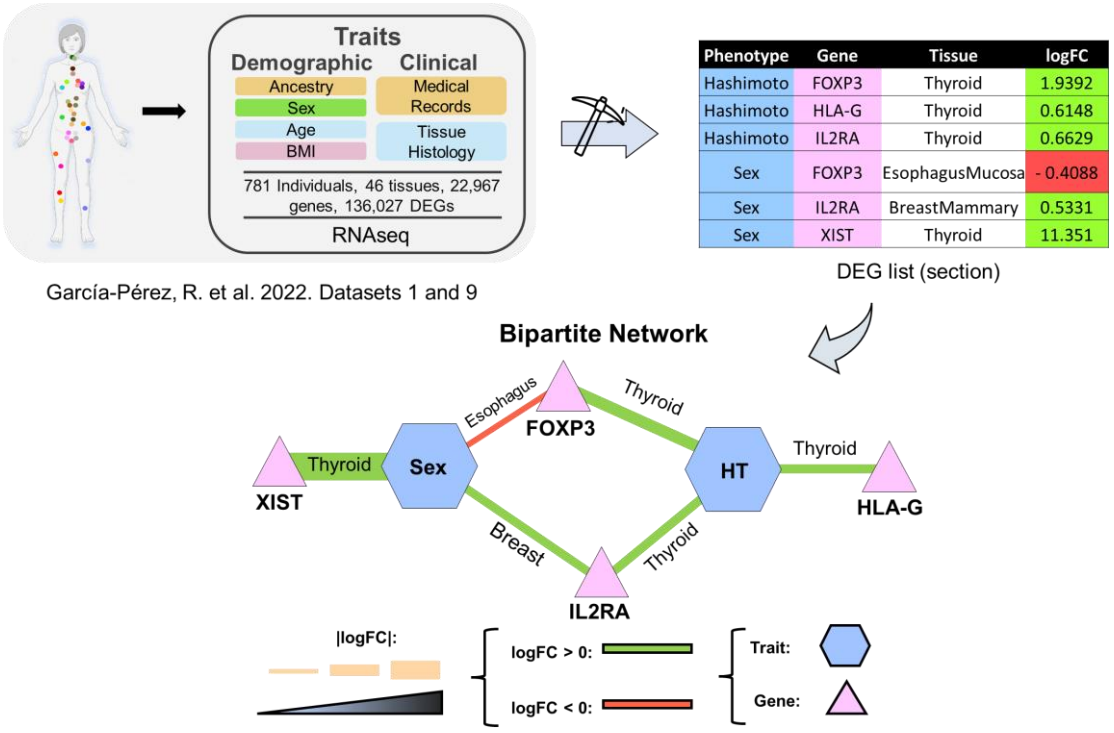


Figure 9 – Workflow for obtaining PheGEx. A DEG list with trait-gene-tissue triplets and logFC are mined from D1 and D2 (adapted from García-Pérez, R. et al. 2022[10]). The list is then used to build the network (An example short list is used for the purposed of this diagram)

Following the construction of the network, we programmatically calculated tissue-specific subgraphs for all 46 tissues in the data. In this analysis, we aimed to characterize each tissue-specific network topology, behaviour and how they compare to one another (Figure 10). This was achieved with the use of the edge attribute tissue. We applied the igraph function “subgraph.edges” (with “delete.vertices” set to “TRUE”) to restrict edge types and remove unconnected nodes. We then iteratively determined the number of nodes (nNodes) and nEdges, the diameter, centralization, density, average path length (avPathLength), weighted average degree (wavDegree) and the number of components (nComponents) of each tissue. All metrics were obtained using the corresponding igraph functions, except for the average weighted degree. This metric was calculated by dividing the sum of the weighted degree of all nodes by the nNodes – using the functions “strength”, “sum” and “length”. The diameter and the avPathLength were calculated twice, one obtaining the standard function implementation and the second also considering edge weights. For this analysis, we considered the weight as the inverse of the logFC, so that stronger connections translate into closer nodes with smaller distances in the weighted metrics.

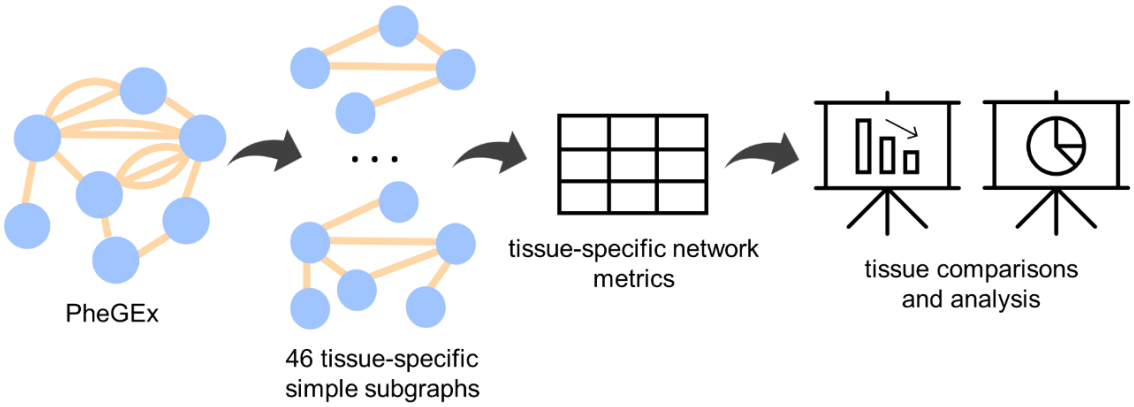


Figure 10 – Workflow for obtaining tissue-specific subgraphs from PheGEx. We then analyse tissue-specific expression and compare them.

2.3. Gene Set Enrichment Analysis

The network representation of the expression-disease system exposes the behaviour of diseases and showcases the interplay and commonalities between disease-demographic and disease-disease trait pairs. With this analysis, we can explore the pathological profiles, whilst controlling for the factors age, BMI, ancestry, and sex (Figure 11). To obtain a biological interpretation of these relationships we applied an over representation analysis to the set of shared DEGs among each pair of phenotypes within the tissue of origin and across tissues, to grasp what are the processes occurring at both levels. We strive to unravel the processes responsible for the demographic and genetic risk factors associated with CDs, whilst better understanding the traits profile and aetiology. We used the R package clusterProfiler[90] with the GO biological processes database (GO:BP) and the KEGG pathway repository.

We focused our functional study on 3 clinical phenotypes: the HT disease, lung injuries and diabetes. Hence, we used the thyroid, lung, and pancreas graphs (derived in the tissue-metric analysis), respectively, to investigate relationships on tissue of origin and the PheGEx network to interpret biological function across tissues. The tissue-specific networks were also exported and further analysed and compared using Gephi.

This strategy allowed us to investigate biological processes (BP) and pathways on the source of the biological dysfunction, as well as the effect of differential expression across tissues in the profile of the disease.

The clusters of DEGs enriched in this analysis were obtained by calculating the list of genes shared exclusively by the two considered traits (degree=2 in a simplified version

of the network). This was done to exclude genes that are involved in a very high number of demographic and pathologies, which tend to reflect mostly cell division, metabolism, and other house-keeping tasks. This phenomenon might be explained due to their relevance in post injury settings and general relevance in biological processes.

To isolate the set of genes shared by two phenotypes of interest, we first filtered the network by keeping only genes with a degree equal to two, using the “subgraph” function. Using this subgraph, we obtained the ego networks for both phenotypes and then calculated the intersection of the two networks with the base R functions “Reduce” and “intersect”.

Using this strategy, we programmatically calculated and applied an over-representation analysis to the set of shared DEGs, in each of the four networks on all pairs of phenotypes.

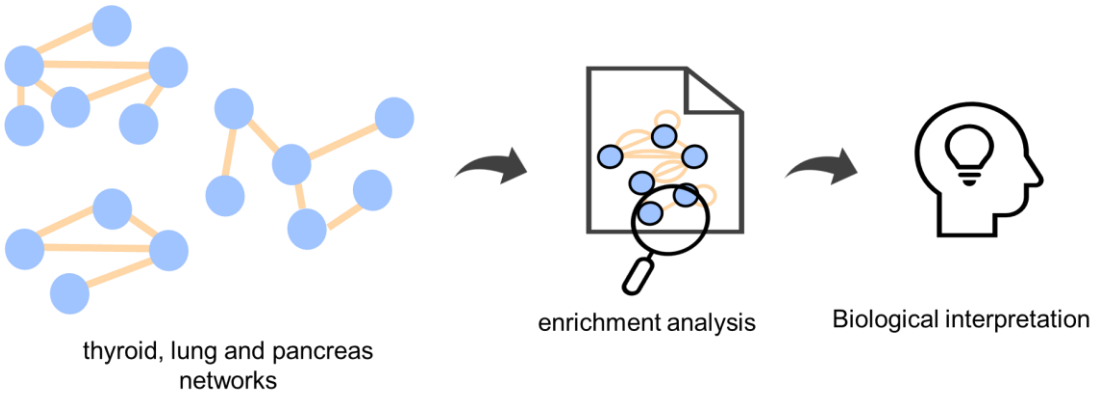


Figure 11 – Enrichment analysis strategy. The thyroid, lung and pancreas with HT, lung injuries and diabetes, respectively, are used as study cases for the biological interpretation of trait relationships employing an enrichment analysis.

Chapter 3

Results

3.1. Phenotype – Gene Expression Network Analysis

The development of the bipartite model of the gene – disease differential system resulted in the production of the PheGEx (Figure 12), which is formed by a single LCC. The network has 18,670 nodes (18,649 genes + 21 phenotypes) and 136,027 edges (differential expression instances). There is a total of 46 unique tissues in the network encoded as edge attributes. The most represented tissue is the thyroid on about 9% of

the network edges, followed by suprapubic not sun exposed skin (SSNES). The least represented is the first cervical vertebra (C1) on just 0.04%. 50% of the edges encode gene over expression and 50% under expression. The density is a metric that translates the fraction of connections between genes and diseases that exist out of all possible ones. Hence, a density of 1 would correspond to a network where nodes connect to one another, and a density of 0 to a graph where no genes are significantly differentially expressed in any disease. The PheGEx graph is sparse with a density of 0.001. However, this is still relatively high given the size of the network. The maximum shortest path between two nodes is of 6 steps (diameter) and the avPathLength between vertexes is of 2.18. Thus, all nodes are easily accessible to one another. The minimum longest shortest path (radius) between two nodes is of 3 steps, so there is at least one node that can access all others at the utmost of 3 steps. Each node is connected, in average, to 14.57 other nodes (average degree). Furthermore, each node has, in average, a 3.52 logFC strength distributed through the relationships with other nodes (wavDegree). The network has a low modularity of 0.23, describing 5 communities. The first includes the phenotype BMI; the second contains sex, spermatogenesis, sclerotic, gastritis, saponification, pneumonia, emphysema, atelectasis, steatosis, and gynecomastia; the third only ancestry; the fourth age, T1D, T2D, HT, atrophy, hyperplasia, fibrosis, and atherosclerotic arteries; and the fifth just congestion.

Figure 12 – PheGEx visual representation. The colour of the nodes reflects the clusters found using the Louvain algorithm. The edge thickness is proportional to the edge weight (logFC).

The edges with the highest weight are dominated by connections involving the phenotype sex across various genes, often including the same genes in multiple tissues (*XIST*, *KDM5D*, *EIF1AY*, *USP9Y* and *DDX3Y*). The genes with the largest degree, weighted degree and eigenvector are *RPS4X*, *RP13-216E22.4*, *EIF1AX* and *SMC1A*, all of which are strongly associated to the phenotype sex and often with ancestry as well. These genes are all located in the sexual chromosomes, X and Y, thus having high differential expression when comparing the two biological sexes. The betweenness centrality highlights the gene *SPATA18*, along with *RP13-216E22.4*, *ANKRD65*, *MIR4458HG*, *RP11-706O15.1*, *ZMAT3* and *LINC01006*. The closeness metric values mostly favoured genes, such as *PLAG2G16*, *SPTBN1*, *IGF1*, *ANKS1A* and *PPP1R26*. A total of 95 genes are highlighted as peripheric by the eccentricity statistic with an

assigned value of 6. In contrast with an average eccentricity value of 4 on the genes with the highest closeness centrality and an eccentricity of 3 on most highly connected phenotypes (i.e., HT, sex, BMI, age, and pneumonia).

In summary, the data to be analysed in this work comprises:

- A single LCC, 18,649 genes, 21 traits (4 demographic + 17 clinical) and 136,027 phenotype associated differential expression connections across 46 tissues;
- Genes are equally over and downregulated in the network;
- Thyroid and SSNES are the most represented tissues;
- High interconnectivity with low avPathLength and diameter;
- 5 communities were described with a modularity of 0.23;
- Central genes are dominantly associated with sex.

3.2. Tissue-Specific Network Characterization

From our analysis of each tissue included in the dataset by Garcia-Pérez, R. et al. 2022 (Figure 13, Supplementary Table 1, Supplementary Table 2), we found that the tissues: thyroid, SSNES, breast mammary, adipose subcutaneous tissues and the tibial nerve involve both the largest nNodes and DEGs.

The brain tissues, C1, amygdala, hippocampus, substantia nigra, frontal cortex BA9 and anterior cingulate cortex BA24, implicate the least number of both nodes and DEGs. These tissue networks are the densest, having the largest ratio of differential expression instances among the involved nodes. This behaviour is matched by other brain related tissues, with 9 out of the 10 densest networks corresponding to brain related tissues.

The networks that are the least dense are the breast mammary tissue, the thyroid and the SSNES, with less than 0.03% of the possible differential expression instances among the involved nodes. The brain stubstantia nigra, amygdala, anterior cingulate cortex BA 24 and the frontal cortex BA9, and the left heart ventricle are the tissues with the most disconnected networks, having 3 components each. On the other hand, having a single LCC is much more common, with 22 networks displaying this topology.

Uterus, brain hypothalamus, ovary, and minor salivary gland showcase the highest degree centralization values, returning values close to one. Whilst stomach, liver, brain hippocampus, and oesophagus muscularis have values around 0.3, indicating low dependency of these networks on a central node.

The networks with highest wavDegree are the tibial nerve and tibial artery. The tissue with the lowest wavDegree is the C1 vertebra, followed by the coronary artery. This suggests that, in the tibial nerve and tibial artery, each connection between a gene and a phenotype are intermediated by differential expression processes with a lower fold change, whilst in the C1 and coronary artery have lower average fold changes.

The weighted diameter (wDiameter) of C1 is of just 11.8. The terminal ileum on the small intestine, has the second lowest wDiameter of about 160. This metric is a measure of the longest shortest path considering the sum of the weights as the distance. These two tissues have particularly low values, even when compared to other brain tissues and networks. The networks with the highest weighted average path length (wavPathLength) are the aorta artery and the sigmoid colon. C1, brain frontal cortex BA9, coronary artery and pituitary have the lowest wavPathLength, that are at least, one order of magnitude below the other networks. Since we are using the inverse of the weight to calculate these metrics, we interpret lower values as being closer due to the higher strength in connection (strongly associated or with higher associated fold change in the expression of the gene).

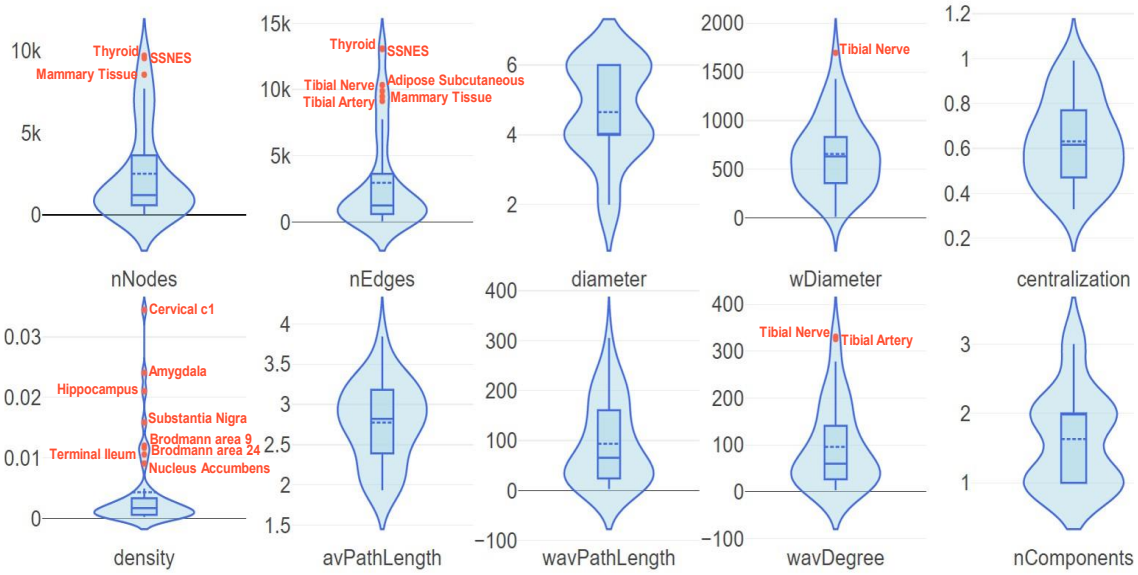


Figure 13 – Tissue-specific Network Metrics and Distribution. Each violin plot corresponds to a network metric. The outliers are discriminated using red markers and labels. In the boxplot we include the upper and lower fences, the mean (dashed line), the median (solid line), and the first and the third quartiles are annotated. An idea of the distribution is given with the surrounding trace.

In summary, the comparison of tissues revealed:

- Tissues are heterogeneous in their composition and structure;
- Most tissues form a single LCC, being highly interconnected;
- The distribution of brain tissues across metrics stands out, as most implicate a low number of DEGs and genes and tend to have disconnected components.

3.3. Biological Interpretation of Network Relationships

Using the enrichment objects created for the PheGEx and the three tissue-specific networks, we were able to get interesting insights into the encoded biological interactions. In this work, we focused our efforts on the auto immune disease HT, lung injuries and diabetes. We investigated the factors and nature of the relationships within the origin tissues and across tissues, potentiating detection and understanding of clinical profiles, and assessing cross-trait impact and interaction.

3.3.1. Hashimoto thyroiditis study case

The thyroid network (Figure 14) has 2 components: a main LCC and a very restricted second one formed by T2D and two more genes. The network has 9,677 nodes and 13,091 edges, an average degree of 2.7, an average weighted degree of about 1, a diameter of 6, and a density of approximately 0. The edges with the largest weight continue to be between sex and the genes *XIST*, *USP9Y*, *KDM5D*. The edges with the largest weight, including HT, connect to *CD79A*, *CXCL13* and *JCHAIN*, all related with the immune system. A total of 20 genes have a degree of 4, the largest for a gene in the network. The genes *TMEM56*, *PCDHB14*, *MYOC*, *RAB11FIP4*, *EOGT*, *LIMD1*, *NEGR1* and *RCAN3*, with diverse functions, have simultaneously the largest betweenness, eigenvector and closeness centrality in the LCC. The eccentricity however is dominated by the nodes connected exclusively to the BMI, including *TSHR*, *LIAS* and *SP1*.

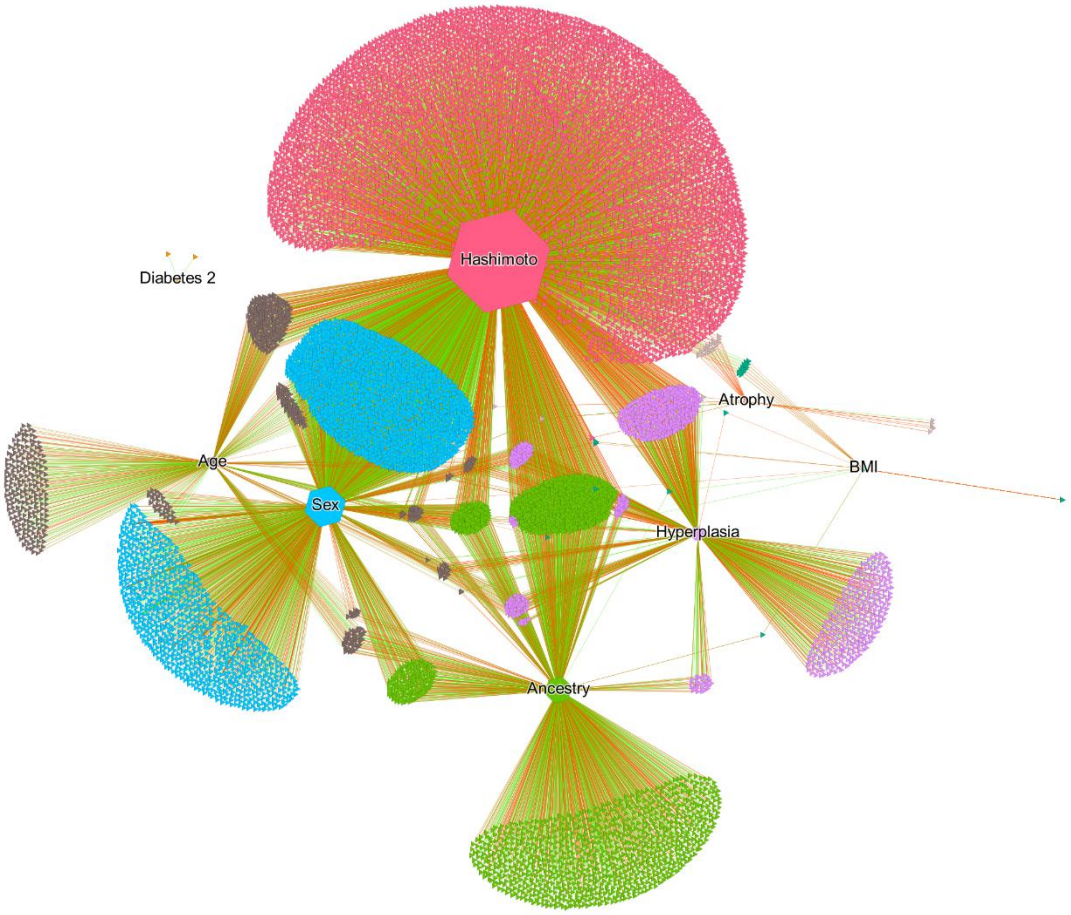


Figure 14 – Thyroid network. Phenotype nodes are shaped as hexagons and gene nodes as triangles. To enhance clarity, only phenotype nodes are labelled. The size of the node corresponds to the respective eigenvector centrality. The colour reflects the clusters found using the Louvain algorithm. The edge thickness is proportional to the edge weight (absolute logFC) and the edge colour is red when it reflects under expression and green for over expression.

The phenotype pair HT-sex (Supplementary Figure 1), in the thyroid tissue, returned mostly immune response related biological processes, when using the GO database. For KEGG the results also include immunity terms (i.e., Natural killer cell mediated cytotoxicity, T cell receptor signalling pathway primary immunodeficiency), among several viral and bacterial infection related instances. The enrichment of HT-sex in the global PheGEx, however, did not return any significant terms.

HT-age (Supplementary Figure 2) also only showed significant enrichment for the thyroid specific shared DEGs. GO returned four terms related to vision. It also highlighted the pathways Cushing syndrome, colorectal cancer, apoptosis, platinum drug resistance and *p53* signalling pathway.

HT-ancestry (Supplementary Figure 3) only retrieved KEGG pathways. On the thyroid an infection related term was retrieved, and PheGEx revealed the “Amyotrophic lateral sclerosis” pathway across tissues.

HT-BMI (Supplementary Figure 4) only returned results across tissues mostly relating to cAMP and cGMP signalling, processing and presenting of antigens, major histocompatibility complex (MHC) class 1 and the endoplasmic reticulum.

HT-hyperplasia (Supplementary Figure 5) emphasized extracellular organization, proliferation, growth, morphogenesis, and development. It also included an eye development term. It did not return any KEGG significant pathways. It did not retrieve any significant results across tissues.

The enrichment of the genes exclusively differentially expressed with HT (Supplementary Figure 6) revealed within the thyroid 473 processes and 25 pathways. It also returned 8 KEGG terms on the human disease expression network. Within the origin tissue: cellular size, adhesion, cell component disassembly, several signalling pathways, cell cycle, infection, morphogenesis, and a few leukocyte regulation terms were identified. Across tissues several compound metabolism, biosynthesis of steroid hormone, chemical carcinogenesis and two cytochrome P450 KEGG pathways were obtained.

In summary, the HT functional assessment revealed:

- HT-sex and just HT showed mostly immune system related terms;
- HT-age emphasized age related health issues, such as cancer-related pathways and eyesight processes;
- The analysis suggested metabolic variation in the disease between European Americans and Black Americans;
- Signalling and MHC are indicated as differentiated with BMI;
- HT-Hyperplasia reveals common cell organizations and structure events to both diseases.

3.3.2. Lung Injuries

The network (Figure 15) has 6,343 nodes, 7,739 edges and a single giant component. In lung, the main clinical nodes are pneumonia and atelectasis, followed by emphysema and fibrosis. For demographics, ancestry, sex, BMI, and age were observed as nodes. It has a diameter of 6, the density is ~0 and the avPathLength is 3.21. Each node is connected, in average, to 2.44 other nodes, with an average weight of 0.73.

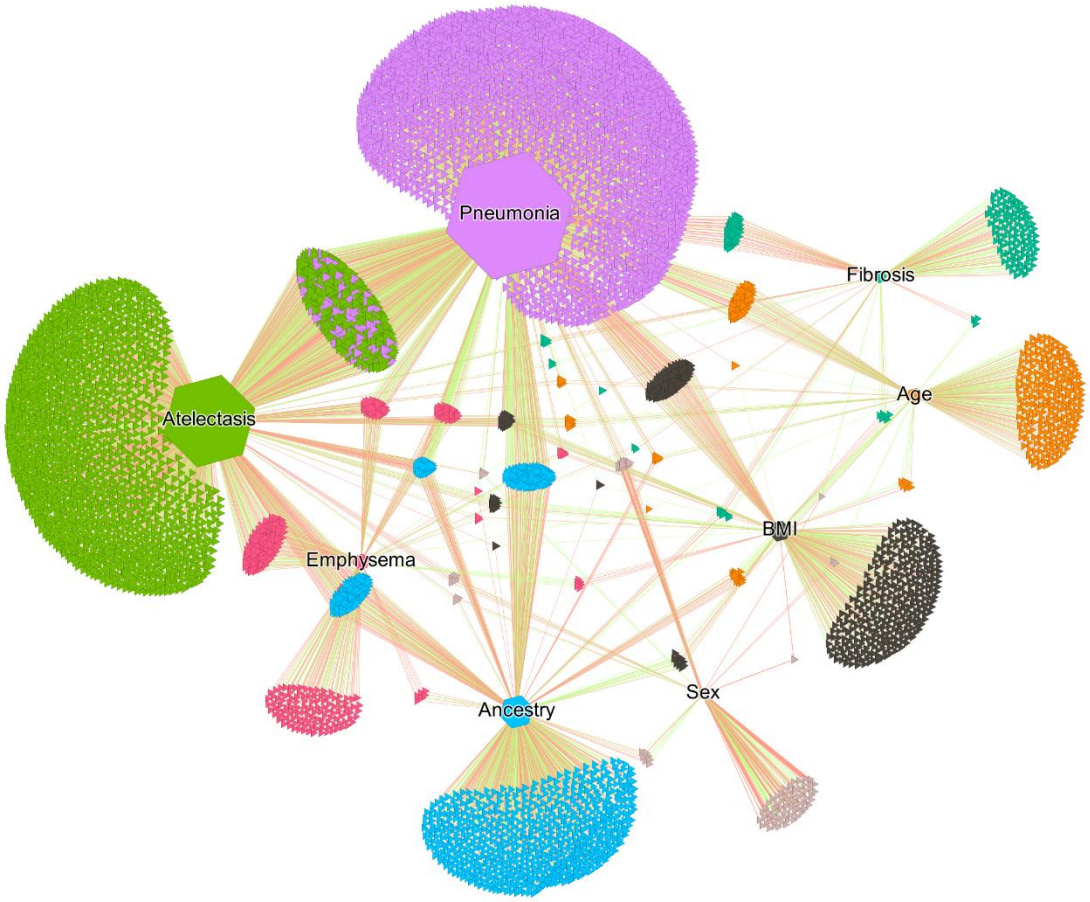


Figure 15 – Lung Network. Phenotype nodes are shaped as hexagons and gene nodes as triangles. To enhance clarity, only phenotype nodes are labelled. The size of the node corresponds to the respective eigenvector centrality. The colour reflects the clusters found using the Louvain algorithm. The edge thickness is proportional to the edge weight (logFC) and the edge colour is red when it reflects under expression and green for over expression.

The edges with the highest weight related to pneumonia are with the genes *PI3*, *DEFB4A* and *SAA2* – that participate in the response against microbes and inflammation – with weights between 1.5 and 2. For atelectasis the genes *MS4A8*, *ERICH3* and *CBLN4* stand out with weights of about 1.3. For Emphysema the genes *JCHAIN*, *RP1-79C4*.4 and *LINC01506* are the strongest with values of about 0.6. With fibrosis, the genes *CPXM1*, *SFRP2* and *ADGRF1* display the highest absolute log FC, with values between 0.78 and 1.04, and are related with cell interaction and signalling.

120 genes have degree 3, the largest in the network for genes. *CPXM1*, *PLA2G1B* and *PTX3* stand out for having the highest weighted degree among them. Several genes performed better than most phenotypes for the closeness and eigenvector centralities. *PYGL*, *SELENBP1*, and *PLA2G1B* are the top performing genes for these metrics. The genes with the largest betweenness are *AURKA*, *KLHL29* and *CDKN3*.

We first explored the commonalities of biological terms between the different types of injuries. Emphysema-pneumonia (Supplementary Figure 7) were characterized exclusively by the BP Golgi lumen acidification and by the oxidative phosphorylation KEGG pathway in the lung tissue. Across tissues no significant biological terms were found. Atelectasis-pneumonia (Supplementary Figure 8) obtained 6 BPs and pathways on the lung tissue. The GO terms were related to epithelium, and tube development and formation; 6 KEGG terms were also obtained, including “tight junction” and “DNA replication”. Across tissues, 3 pathways were found related to diabetes, *IL-17* signalling and viral protein interaction with cytokine. No biological significance was found for the set of shared DEGs established between atelectasis-emphysema.

No significant biological terms were found between pneumonia-sex. For pneumonia-age (Supplementary Figure 9), in the lung tissue, we found 56 BPs and 1 KEGG pathway, mostly related to cell cycle. Whilst, across tissues, pneumonia-age returned terms related with DNA and immune response. Neither ancestry nor BMI had enrichment results with pneumonia in the tissue-specific analysis. Across tissues, pneumonia with BMI (Supplementary Figure 10) returned several significant terms of diverse cell functions. The analysis of pneumonia-ancestry (Supplementary Figure 11) returned 4 diverse pathways across tissues.

Emphysema had no enriched sets with the demographic trait sex within the lung tissue, however across tissues (Supplementary Figure 12) it returned 24 BPs and 3 KEGG pathways, all related to the immune system. Emphysema-age lung-specific (Supplementary Figure 13) returned mainly DNA replication biological processes and includes response to molecules of bacterial origin and lipopolysaccharides but did not produce enrichment results in the PheGEx network. Emphysema also did not enrich with ancestry in the PheGEx network, while in the lung (Supplementary Figure 14) it returned pathways related with cell division and lipid metabolism. Emphysema-BMI (Supplementary Figure 15) revealed cell cycle and cell morphology in the lung.

The analysis of atelectasis did not provide results with any demographic factor except for sex (Supplementary Figure 16) on the tissue-specific lung network – reporting the TGF-beta signalling pathway.

We then analysed the impact of fibrosis along with the previous 3 injuries. Fibrosis-emphysema (Supplementary Figure 17) resulted in regulation of axonogenesis and chemotaxis, cell migration and protein localization to vacuole. Atelectasis-fibrosis (Supplementary Figure 18) provided several microtubule and microvillus processes, cell-cell adhesion, transmembrane transport, the remaining relate to the sensory perception

of sound and light. Fibrosis-pneumonia (Supplementary Figure 19) retrieved the processes “respiratory gaseous exchange by respiratory system”, “sphingolipid metabolic process” and “membrane lipid metabolic process”.

The enrichment of Emphysema alone (Supplementary Figure 20) in the lung did not provide any information, however across tissues we identified AMP processes, nucleoside biosynthetic processes, DNA replication initiation and pathways related to DNA replication, repair, recombination, and nucleobase metabolism. Another pathway identified was the “Fanconi anaemia pathway”.

Atelectasis exclusive (Supplementary Figure 21) enrichment did not retrieve any results in the larger context of all human tissues. On the lung, a protein processing in the endoplasmic reticulum pathway was obtained, along with many cilia related processes.

Pneumonia (Supplementary Figure 22) provided 10 processes across tissues and 389 processes and 87 pathways within lung specific shared DEGs. We found humoral immune response, positive regulation of response to external stimulus, defence to Gram-positive bacterium, fibrinolysis, coagulation, and haemostasis, across tissues. In the lung KEGG pathway returned multiple infection, cancers, and other diseases, along with some immune related terms. GO BPs displayed the term “wound healing” along with immune, signalling, adhesion and replication pathways and processes.

In summary, lung injuries functional assessment revealed:

- Lung injuries – demographic factors relationship mostly displayed DNA, cell cycle and signalling pathways;
- Emphysema returned DNA replication and repair along with biosynthetic processes;
- Atelectasis DEGs underlined events related to cilium structures;
- Exclusive pneumonia DEGs enrichment showed cell keeping terms along with infection, immune terms and some blood related terms.

3.3.3. Diabetes mellitus

The pancreas network (Figure 16) was simpler, with 1,571 nodes and 1,655 edges, and involving T1D and T2D, ancestry, sex, and age. Saponification was not linked to the LCC. The average degree was 2.1, the diameter 4, the avPathLength 3.32 and the density 0.001. The most peripheral nodes are the ones shared by sex and ancestry, with an eccentricity of 4. The most central node, according to eigenvector, closeness and betweenness, in the *LCC* is *SPATA18*, a key regulator for mitochondrial quality.

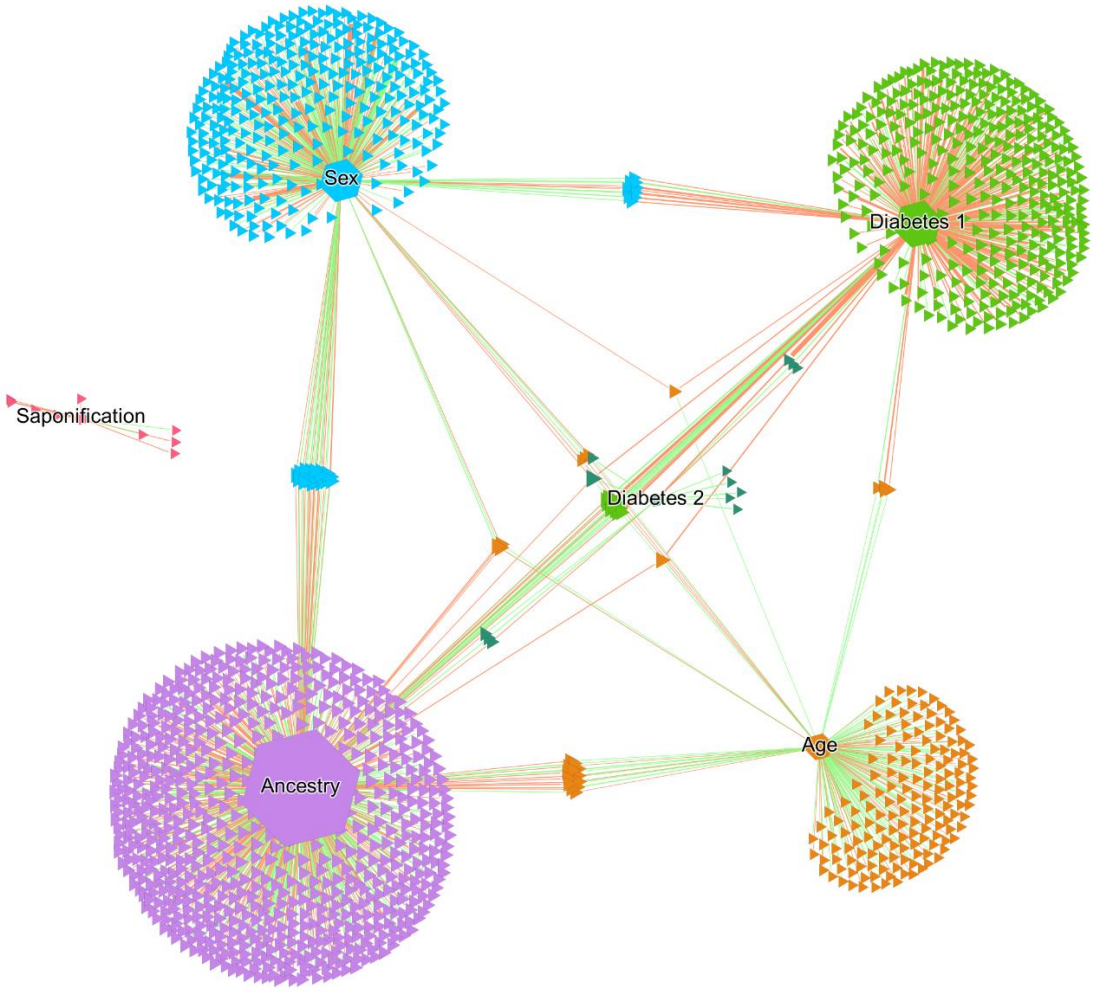


Figure 16 – Pancreas network. Phenotype nodes are shaped as hexagons and gene nodes as triangles. To enhance clarity, only phenotype nodes are labelled. The size of the node corresponds to the respective eigenvector centrality. The colour reflects the clusters found using the Louvain algorithm. The edge thickness is proportional to the edge weight (logFC) and the edge colour is red when it reflects under expression and green for over expression.

We first looked into the set of shared DEGs between both types of diabetes (Supplementary Figure 23) and several significant terms were detected. Across tissues, signalling, receptor regulation, and bone processes were highlighted. In the pancreas, metabolism related with energy was detected.

Curiously, BMI did not have any DEGs specific to pancreatic tissue. Across tissues, T1D-BMI (Supplementary Figure 24) showed “cellular response to testosterone stimulus” and “ER overload response”. Other terms relate to response to ionizing radiation, protein lipidation, lipoprotein biosynthetic processes, response to ketone, fibroblast growth, liver and lung development, response to biotic stimulus and macroautophagy. T2D-BMI (Supplementary Figure 25) only showed 2 protein folding related processes.

Neither type of diabetes enriched with age across tissues, within the pancreas only T1D returned results (Supplementary Figure 26) mostly related to cell cycle phase transitions, the remaining referring to mitochondrial genome maintenance, dendrite extension and signalling by p53.

T2D did not enrich with sex for pancreatic and global networks. T1D-sex (Supplementary Figure 27) returned one KEGG term for each network, “Butanoate metabolism” across tissues and “Nitrogen metabolism” in pancreas.

Ancestry-T1D (Supplementary Figure 28) in pancreas returned hormone response, polysaccharide metabolism and regulation, and nervous system related terms. Across tissues, visual perception, embryonic development, nervous system, differentiation, and some mitosis and DNA related BPs were found. KEGG highlighted diabetes, pluripotency, and gap junction. No terms were found for ancestry-T2D on the PheGEx. On pancreas (Supplementary Figure 29) only GO BP terms were found. These returned mostly immune and pseudopodium terms and other diverse processes.

The enrichment of the exclusive T1D DEGs (Supplementary Figure 30) resulted, generally, in the processes: hormone and peptide transport and secretion, secretion regulation, nervous system signalling, carbohydrate homeostasis gluconeogenesis and pancreatic cell differentiation. Various KEGG pathways related to diabetes, insulin and signalling were retrieved.

The analysis of the DEGs exclusive to T2D (Supplementary Figure 31) resulted, in pancreas, in the outputs: stimulus, immune cell chemotaxis, regulation of signal transduction, negative regulation of immune system processes, growth factor stimulus and multiple bone morphogenesis proteins terms. Across tissues, the KEGG pathways “Olfactory transduction” and “Homologous recombination”.

In summary, diabetes functional assessment revealed:

- BMI does not have any DEGs in the pancreatic tissue;
- The enrichment of T1D DEGs returned multiple terms related to diabetes, insulin, the pancreas and carbohydrates;
- T2D exclusive DEGs enrichment showed immune, growth factor and bone associated processes;
- Diabetes – demographic factors returned few and varied terms.

3.3.4. Further restricting some gene sets

Some connections in the full bipartite network (across tissues), did not provide insightful enrichment results. Such was the case of the relationship between diabetes and demographic factors, and HT-sex. Since their understanding is paramount to the profile of the diseases, we further analysed these relationships by obtaining new, additionally restricted, gene sets. We applied a degree 2 restriction directly to the multigraph. This restriction leads to an increased specificity of the sets. However, the reduced set size has the downside that the robustness of the analysis is decreased and that some relevant genes might also be removed. Even so, new relevant terms were found for several of the reanalysed relationships.

Previously we had not found any enrichment results for type T1D-age. Using this methodology (Supplementary Figure 32), we retrieved 8 KEGG pathways, including “Type 1 diabetes mellitus”, “Endocrine resistance”, “Estrogen signalling pathway” and “Parathyroid hormone synthesis, secretion and action” with the genes *HBEGF* and *PTPRN*.

T1D-sex (Supplementary Figure 33) had only revealed the metabolism of butanoate and nitrogen. The *de novo* enrichment expanded on these results with 12 KEGG terms. The novel pathways included the metabolism of several chemicals (i.e., taurine, fructose mannose, alanine, aspartate, and glutamine), the term “Type I diabetes mellitus” and “Glycolysis / Gluconeogenesis”.

For T2D-sex (Supplementary Figure 34), two pathways related with the nervous system were identified along with 16 BPs related to vitamin metabolism, glutamate, blood brain barrier, placenta development and cell development.

We found BPs for T2D-BMI (Supplementary Figure 35) related to lipoproteins, endoplasmic reticulum, and autophagy.

T2D-ancestry (Supplementary Figure 36) the pathways “Insulin secretion” and “Complement and coagulation cascades”. We also found 22 GO terms, including glucagon and endoplasmic reticulum, complement activation, cyclase and lyase activity, and negative regulation of blood pressure.

With this method, sex-HT (Supplementary Figure 37), retrieved 5 processes. From these, 3 pertain to T cell tolerance induction and 2 to “*CD4*-positive, alpha-beta T cell proliferation”. These terms were traced to the genes *FOXP3* and *IL2RA*.

In summary, the complementary functional analysis revealed:

- T1D with sex, age and ancestry now reveals multiple diabetes and hormone-related pathways;
- T2D retrieved new terms with ancestry, sex and BMI, mostly related with metabolic processes;
- HT-sex showed important immune processes that can be traced to the genes *FOXP3* and *IL2RA*.

Chapter 4

Discussion

4.1. The PheGEx Network

The analysis of the PheGEx network showed that differentially expressed genes in the human disease and demography associated expression system are equally over and under expressed. The formation of a single giant component, the diameter of 6 and the avPathLength of 2.16 indicates a high degree of interconnection between phenotypes through differential expression, emphasizing the complexity of human health and disease. Even though, the density is of just 0.001, but given its dimension, it is somewhat dense, which constrains network processing and visualization.

Thyroid is the tissue responsible for the most connections in the network. Due to the nature of the network, tissues with clinical phenotypes might be expected to have a much larger proportion of differentially expressed genes. However, the pancreatic tissue, that also has hormonal function and clinical trait nodes, is associated with only 1.22% of the connections in the network, compared to 9.62% in the thyroid. These results suggest that expression changes with demographic and clinical trait variation are particularly high in the thyroid. The tissue responsible for the second most connections is the SSNES with 9.52% of the edges. In this tissue, sex and ancestry are responsible for most connections and have a higher associated weight. This indicates that specifically sex and ancestry, in this tissue, are responsible for a considerable portion of expression changes in the network.

The analysis of gene centrality in the network reveals that genes with a high centrality, across measures, are mainly related to sex, followed by ancestry. These results suggest a high importance for the role of differential expression in the dimorphisms between

sexes and ancestry-driven variation in the human species. Several entries in the literature report on the importance of regulation of expression for the development of sexual variation despite the high genomic similarity between sexes[91–93].

The modularity value of 0.23 is not very high, therefore the found communities must be approached carefully. However, the aggregation of both types of diabetes, HT, atrophy, hyperplasia, fibrosis, and atherosclerotic arteries is a positive indicator. As both types of diabetes remained together, and conditions often associated with the thyroid also appear in the same aggregate. The appearance of HT and diabetes together might be explained by the autoimmune nature of the thyroid disease and diabetes, and by the hormonal implications in both instances.

Sex, spermatogenesis gynecomastia, sclerotic, gastritis, saponification, pneumonia, emphysema, atelectasis, and steatosis clustered in the same group. The combination of the sex related traits: spermatogenesis, gynecomastia, sclerotic testicles, and sex in the same group corresponds to the expected results. The same is true for the inclusion of all lung injuries within the same cluster. Gastritis being characterized by inflammation, and being often caused by an infection, might explain the aggregation with lung injuries. Saponification, in this dataset, is only found in the pancreas with just 11 DEGs, where none are shared with another phenotype in that tissue. The analysis of its depth 2 ego network across tissue reveals 10 out of the 11 saponification DEGs are shared with sex. This, most likely, explains the association of the disease with this group. Nonetheless, the aggregation of sex related pathologies, steatosis and lung injuries remains unclear.

The isolated traits BMI, ancestry and congestion within their own clusters is unexpected, and the clustering algorithm should be revised to include further conditions that might improve resolution.

4.2. Tissue-Metrics Dataset

In this analysis, nervous system tissues tended to have high density, high average weighted degree, low number of DEGs and nodes, and often have 3 components with the only 3 network traits being age, sex, and ancestry separated between components. Hence, this class of tissues tends to have little or no differentially expressed genes in common between demographic factors, justifying the high density values in many of them. Brain tissues with a larger nEdges and nodes display a more typical behaviour. Consequently, most brain tissues stand out for the lack of interconnection. It should also

be noted that there are no genes significantly differentially expressed with BMI in the brain.

The tibial nerve and artery, on the other hand, have a large number of DEGs, but a lower wDiameter indicating complex behaviour with many interactions, but a lower associated logFC. This could reflect a smaller individual contribution from each gene.

The presence of a different number of clinical phenotypes across tissues and the different nComponents makes the comparison among them, with the current methodology, in some metrics, more complex.

4.3. Enrichment results

The complementary strategy of analysing relationships within and among tissues simultaneously proved to be very thorough, extensive, and effective. The roles of genes and phenotypes in the expression system became more interpretable, rendering large amounts of data more approachable. Thus, several processes and pathways with the associated genes were retrieved. The derived data provides new insights into the epidemiology and pathogenesis of diseases, and clues that can result in candidate causal genes and processes. Overall, we found that we obtained multiple terms and genes that have an obvious association with the disease and some genes and pathways that already have been well described in the literature. The use of a second enrichment methodology, when appropriate, proved effective at retrieving better results on the PheGEx network, even if with less support.

The analysis of HT returned several immunological processes for genes that are unaffected by other phenotypes. Our results suggests that the increased incidence on women might be related to immune and infection processes and pathways, due to the enrichment results of DEGs associated to both HT and sex. The analysis of HT-sex proved the capability of our methodology to retrieving crucial genes and terms from the clusters. Namely, the restricted FGEA of HT-sex retrieved the genes *FOXP3* and *IL2RA* and included the “T cell tolerance induction” and “CD4-positive, alpha-beta T cell proliferation” pathways. *FOXP3* is an X-linked gene that is an important transcription factor for T regulatory cells[94] and is strongly described in the literature as a key for auto-immune diseases of the thyroid[39,42,94]. Furthermore, the over activity of CD-positive T regulatory cells has been described as a main factor in the pathology of HT[94]. T-regulatory cells are responsible for self-tolerance[94].

Viral infection terms were also present in HT-age. The presence of infection terms might support the literature, where some instances of HT profiles are reported to start after a first infection clinical case.

On the other hand, HT-age analysis returned several pathways commonly associated with ageing and cancer. The eyesight collected terms might constitute a novel clue into the processes and risk factors that lead to the development of thyroid-associates eye diseases[95]. HT-hyperplasia disclosed cellular organization and division changes shared by both conditions. Since HT can lead to the damage of the thyroid, it can also result on thyroid hyperplasia (or goitre). The analysis of this relationship might be particularly relevant for an improved understanding of this clinical progression.

Across lung injuries, the enrichment study retrieved mostly cell cycle and repair processes, along with some epithelium and cilium terms. In emphysema, lipid related terms with ancestry and age were present. Since, there are several studies suggesting that the pathology is associated with the derangement of lipids, the results seem to be somewhat supported and might underlie risk groups for emphysema[96]. Atelectasis, being most often caused in the context of surgery due to the application of anaesthesia might explain the reduced enrichment results with demographic factors. The analysis of DEGs exclusive to Atelectasis, retrieved multiple cilium related terms. These structures are relevant to the movement of mucus in the lung, and it is known that the disease can come about from the blockage of the airways with mucus or a foreign object[53,97].

The absence of BMI, an established key factor for T2D[98], associated DEGs on the pancreatic tissue seems to indicate that this relationship occurs specifically across tissues. The restricted FGEA results can be traced to the genes *MYH1* and *NTSR2*, both of which have already been well described in the literature in the context of T2D[99,100]. The presence of several nervous system related terms during the enrichment of diabetes may be due to neuropathic diabetes. The relationship in question has already been extensively examined by the authors of this dataset, García-Pérez, R. and their collaborators, within the given data.

With the restricted FGEA of the T1D exclusive DEGs, we obtained glycolysis, endocrine resistance, gluconeogenesis, and metabolism of sugars processes. Since these are highly characteristic of the disease, and seem to be unaffected by demographic factors, the corresponding genes can be further investigated to better understand their role in the profile of the disease.

Chapter 5

Conclusion and Future Work

As future work, we would like to test improvements on the network projections. The current network structure and standard projection methodology leads to the production of poor network projections due to the high dimension and interconnectivity of the PheGEx. Since the phenotype projection would have both demographic and clinical factors, the interpretation capability would also be diminished. To solve these challenges, in the future, we would like to develop separate projections for demographic and clinical traits and derive a sparse gene projection, by implementing a more restrictive conversion function (i.e., only considering edges with a weight above a certain threshold). To obtain a sparse phenotype projection, one could, for example, only consider connections between traits if they share genes that can be significantly enriched.

The community finding algorithm used could be reapplied with a lower resolution to find more clusters that are, consequently, more specific. An overlapping community finding algorithm with an implementation for bipartite networks could be used instead, to increase the biological interpretation and quality of the results, as this would allow for a better exploration of the network, highlighting the genes and traits that belong to multiple classes.

Even though we were able to characterize the system with the functional enrichment of genes, the methodology could be further improved to simplify the process, avoiding the need for the reanalysis of some relationships. The fact that several genes in our network are lncRNAs with poor annotation and scarce presence in the literature implies that their impact was often overlooked by the enrichment algorithm. The enhanced ability of our methodology to investigate diseases will be improved over time. As more data will be made available, the network can be expanded and the ability to decompose regulation of gene expression will increase.

With this work, we were able to model and characterize the human disease-gene expression system overall and in each tissue. The PheGEx comprises a novel tool for researchers to conduct studies on the relationship of regulation of gene expression with disease and influence from demographic factors. A successful methodology for the interpretation of disease graphs was developed and implemented, leading to an extensive elucidation of shared and exclusive pathways, processes, and genes from the conditions HT, T1D, T2D, pneumonia, atelectasis, and emphysema. Our results can, in

the future, be used to investigate the occurrence of risk groups and risk factors (i.e., the increased incidence of HT among female and older populations, and the relationship of BMI and T2D). Furthermore, the same methodology could be applied to other clinical traits present in the PheGEx and on other disease networks.

References

1. Singh, A. J., Ramsey, S. A., Filtz, T. M. & Kioussi, C. Differential gene regulatory networks in development and disease. *Cellular and Molecular Life Sciences* **75**, 1013–1025 (2018).

2. Zhang, Y. *et al.* Identifying network biomarkers of cancer by sample-specific differential network. *BMC Bioinformatics* **23**, 1–19 (2022).

3. Vidal, M., Cusick, M. E. & Barabási, A. L. Interactome Networks and Human Disease. *Cell* **144**, 986–998 (2011).

4. Fan, Z., Zhou, Y. & Ressom, H. W. MOTA: Network-based multi-omic data integration for biomarker discovery. *Metabolites* **10**, (2020).

5. Zuo, Y. *et al.* INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. *Methods* **111**, 12–20 (2016).

6. Raghavachari, N. & Garcia-Reyero, N. Overview of gene expression analysis: Transcriptomics. in *Methods in Molecular Biology* vol. 1783 1–6 (Humana Press Inc., 2018).

7. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics 2002 3:10* **3**, 779–789 (2002).

8. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology 2017 18:1* **18**, 1–15 (2017).

9. Loscalzo, J., Barabási, A.-L. & Silverman, E. K. *Network Medicine*. (Harvard University Press, 2017).

10. García-Pérez, R. *et al.* The landscape of expression and alternative splicing variation across human traits. *Cell Genomics* **3**, 100244 (2022).

11. Alasoo, K. Clinical traits impacting human tissue transcriptomes. *Cell Genomics* **3**, (2023).

12. Schmitz, S. U., Grote, P. & Herrmann, B. G. Mechanisms of long noncoding RNA function in development and disease. *Cellular and Molecular Life Sciences 2016 73:13* **73**, 2491–2509 (2016).

13. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

14. Nurk, S. *et al.* The complete sequence of a human genome. *Science (1979)* **376**, 44–53 (2022).

15. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science (1979)* **350**, 1096–1101 (2015).

16. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569–574 (2013).

17. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–1789 (2012).

18. Zhicheng, D. & Yan, C. Transcriptomics: Advances and approaches. *REVIEW • Transcriptomics: Advances and approaches. Sci China Life Sci* **56**, 960–967 (2013).

19. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11* **20**, 631–656 (2019).

20. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology 2016 17:1* **17**, 1–19 (2016).

21. Hrdlickova, B., de Almeida, R. C., Borek, Z. & Withoff, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1910–1922 (2014).

22. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* **13**, 311–317 (2015).

23. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics 2013 45:6* **45**, 580–585 (2013).

24. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. **369**, 1318–1330 (2020).

25. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications 2018 9:1* **9**, 1–15 (2018).

26. Porcu, E. *et al.* Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications 2021 12:1* **12**, 1–9 (2021).

27. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun* **10**, (2019).

28. Hern, L. M. & Bidichandani, S. I. What Mendel did not discover: exceptions in Mendelian genetics and their role in inherited human disease. *J Okla State Med Assoc* **97**, 12–17 (2004).

29. Dean, M. Approaches to Identify Genes for Complex Human Diseases: Lessons From Mendelian Disorders. doi:10.1002/humu.10259.

30. Hidalgo, C. A., Blumm, N., Barabási, A. L. & Christakis, N. A. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Comput Biol* **5**, e1000353 (2009).

31. Hunter, D. J. Gene–environment interactions in human diseases. *Nature Reviews Genetics 2005 6:4* **6**, 287–298 (2005).

32. Ni, Y., He, J. & Chalise, P. Integration of differential expression and network structure for 'omics data analysis. *Comput Biol Med* **150**, 106133 (2022).

33. Yi, H. & Ma, S. Assisted differential network analysis for gene expression data. *Genet Epidemiol* **45**, 604–620 (2021).

34. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).

35. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics 2008 9:5* **9**, 356–369 (2008).

36. Bell, R. W. Continuously distributed random variables in factorial designs. *Exp Aging Res* **18**, 47–50 (1992).

37. Song, J. W. & Chung, K. C. Observational Studies: Cohort and Case-Control Studies. *Plast Reconstr Surg* **126**, 2234 (2010).

38. Eschler, D. C., Hasham, A. & Tomer, Y. Cutting edge: The etiology of autoimmune thyroid diseases. *Clin Rev Allergy Immunol* **41**, 190–197 (2011).

39. Hasham, A. & Tomer, Y. Genetic and epigenetic mechanisms in thyroid autoimmunity. *Immunol Res* **54**, 204–213 (2012).

40. Ragusa, F. *et al.* Hashimotos' thyroiditis: Epidemiology, pathogenesis, clinic and therapy. *Best Pract Res Clin Endocrinol Metab* **33**, 101367 (2019).

41. Kronzer, V. L., Bridges, S. L. & Davis, J. M. Why women have more autoimmune diseases than men: An evolutionary perspective. *Evol Appl* **14**, 629–633 (2021).

42. Zaletel, K. & Gaberscek, S. Hashimotos Thyroiditis: From Genes to the Disease. *Curr Genomics* **12**, 576–588 (2011).

43. Jin, X. *et al.* Global burden of upper respiratory infections in 204 countries and territories, from 1990 to 2019. *EClinicalMedicine* **37**, (2021).

44. Viegi, G., Maio, S., Pistelli, F., Baldacci, S. & Carrozzi, L. Epidemiology of chronic obstructive pulmonary disease: Health effects of air pollution. *Respirology* **11**, 523–532 (2006).

45. Troeger, C. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis* **17**, 1133–1161 (2017).

46. Cilloniz, C., Martin-Loeches, I., Garcia-Vidal, C., Jose, A. S. & Torres, A. Microbial Etiology of Pneumonia: Epidemiology, Diagnosis and Resistance Patterns. *International Journal of Molecular Sciences 2016, Vol. 17, Page 2120* **17**, 2120 (2016).

47. Wang, M. *et al.* Association Between Long-term Exposure to Ambient Air Pollution and Change in Quantitatively Assessed Emphysema and Lung Function. *JAMA* **322**, 546–556 (2019).

48. Hage, R., Gautschi, F., Steinack, C. & Schuurmans, M. M. Combined Pulmonary Fibrosis and Emphysema (CPFE) Clinical Features and Management. *Int J Chron Obstruct Pulmon Dis* **16**, 167 (2021).

49. Torres, A. *et al.* Pneumonia. *Nature Reviews Disease Primers 2021 7:1* **7**, 1–28 (2021).

50. Jain, S. Epidemiology of Viral Pneumonia. *Clin Chest Med* **38**, 1 (2017).

51. Hedenstierna, G. & Edmark, L. Mechanisms of atelectasis in the perioperative period. *Best Pract Res Clin Anaesthesiol* **24**, 157–169 (2010).

52. Randtke, M. A., Andrews, B. P. & Mach, W. J. Pathophysiology and Prevention of Intraoperative Atelectasis: A Review of the Literature. *Journal of PeriAnesthesia Nursing* **30**, 516–527 (2015).

53. Marini, J. J. Acute Lobar Atelectasis. *Chest* **155**, 1049–1058 (2019).

54. Banday, M. Z., Sameer, A. S. & Nissar, S. Pathophysiology of diabetes: An overview. *Avicenna J Med* **10**, 174–188 (2020).

55. Smith, J. P., Arnoletti, J. P., Varadarajulu, S. & Morgan, D. E. Post-pancreatitis Fat Necrosis Mimicking Carcinomatosis. *Radiol Case Rep* **3**, 192 (2008).

56. Wynn, T. A. & Ramalingam, T. R. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat Med* **18**, 1028 (2012).

57. Kostoglou-Athanassiou, I. & Ntalles, K. Hypothyroidism - new aspects of an old disease. *Hippokratia* **14**, 82 (2010).

58. Pesce, C., Reale, A. & Sanguineti, G. The Pathology of Testicular Atrophy. *Surv Synth Pathol Res* **5**, 500–511 (1986).

59. Sartori, R., Romanello, V. & Sandri, M. Mechanisms of muscle atrophy and hypertrophy: implications in health and disease. *Nature Communications 2021 12:1* **12**, 1–12 (2021).

60. Jo, J. *et al.* Hypertrophy and/or Hyperplasia: Dynamics of Adipose Tissue Growth. *PLoS Comput Biol* **5**, e1000324 (2009).

61. Mills, A. M. & Longacre, T. A. Endometrial hyperplasia. *Semin Diagn Pathol* **27**, 199–214 (2010).

62. Perez-Montiel, M. D. & Suster, S. The spectrum of histologic changes in thyroid hyperplasia: a clinicopathologic study of 300 cases. *Hum Pathol* **39**, 1080–1087 (2008).

63. Sansone, A., Romanelli, F., Sansone, M., Lenzi, A. & Di Luigi, L. Gynecomastia and hormones. *Endocrine* **55**, 37–44 (2017).

64. Gilbert, S. F. Spermatogenesis. in *Developmental Biology* (Sinauer Associates, 2000).

65. Gumus, E., Kati, B., Pelit, E. S., Ordek, E. & Ciftci, H. A different look at genetic factors in individuals with non-obstructive azoospermia or oligospermia in our

research study: To whom, which threshold, when, in what way? *Rev Int Androl* **19**, 41–48 (2021).

66. Bettermann, K., Hohensee, T. & Haybaeck, J. Steatosis and Steatohepatitis: Complex Disorders. *Int. J. Mol. Sci* **15**, 9924–9944 (2014).

67. Wolf, D. & Ley, K. Immunity and Inflammation in Atherosclerosis. *Circ Res* **124**, 315–327 (2019).

68. Kulnigg-Dabsch, S. Autoimmungastritis. *Wiener Medizinische Wochenschrift* **166**, 424–430 (2016).

69. Rugge, M. *et al.* Gastritis: The histology report. *Digestive and Liver Disease* **43**, S373–S384 (2011).

70. El-Zimaity, H. M. T. *et al.* Enterococcus gastritis. *Hum Pathol* **34**, 944–945 (2003).

71. Sipponen, P. & Maaroos, H. I. Chronic gastritis. *https://doi.org/10.3109/00365521.2015.1019918* **50**, 657–667 (2015).

72. Harris, R. D., Chouteau, C., Partrick, M. & Schned, A. Prevalence and Significance of Heterogeneous Testes Revealed on Sonography. *http://dx.doi.org/10.2214/ajr.175.2.1750347* **175**, 347–352 (2012).

73. Apa, D. D., Çayan, S., Polat, A. & Akbay, E. Mast Cells and Fibrosis on Testicular Biopsies in Male Infertility. *http://dx.doi.org/10.1080/01485010290099183* **48**, 337–344 (2009).

74. Gomha, F. S. *et al.* The sclerosing Sertoli cell tumor of the testis is an extremely rare entity. *Ther Adv Urol* **12**, (2020).

75. Barabási, A. L. & Pósfai, M. *Network Science*. (Cambridge University Press, 2016).

76. Menczer, F., Fortunato, S. & Davis, C. A. A First Course in Network Science. *A First Course in Network Science* (2020) doi:10.1017/9781108653947.

77. Ochodkova, E., Zehnalova, S. & Kudelka, M. Graph construction based on local representativeness. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10392 LNCS**, 654–665 (2017).

78. Wei, C., Xu, A., Yu, H., Chen, G. & Chen, Y. Gene Analysis of Renal Tumor Based on Differential Network. *Proceedings - 9th International Conference on*

Information Technology in Medicine and Education, ITME 2018 182–186 (2018) doi:10.1109/ITME.2018.00049.

79. Jacob, L., Neuvial, P. & Dudoit, S. More power via graph-structured tests for differential expression of gene networks. *https://doi.org/10.1214/11-AOAS528* **6**, 561–600 (2012).

80. Li, P., Yu, J., Liu, J., Zhou, D. & Cao, B. Generating weighted social networks using multigraph. *Physica A: Statistical Mechanics and its Applications* **539**, 122894 (2020).

81. Goh, K. Il *et al.* The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685–8690 (2007).

82. Fortunato, S. Community detection in graphs. *Phys Rep* **486**, 75–174 (2010).

83. Wang, L., Jia, P., Wolfinger, R. D., Chen, X. & Zhao, Z. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics* **98**, 1–8 (2011).

84. Ghaffar, A. & Nyholt, D. R. Genome-wide imputed differential expression enrichment analysis identifies trait-relevant tissues. *Front Genet* **13**, 3715 (2023).

85. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25 (2000).

86. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol* **8**, e1002375 (2012).

87. The Gene Ontology resource: enriching a GOld mine. doi:10.1093/nar/gkaa1113.

88. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462 (2016).

89. Gephi: An Open Source Software for Exploring and Manipulating Networks | Proceedings of the International AAAI Conference on Web and Social Media. https://ojs.aaai.org/index.php/ICWSM/article/view/13937.

90. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

91. Gershoni, M. & Pietrokovski, S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol* **15**, 1–15 (2017).

92. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science (1979)* **348**, 660–665 (2015).

93. Aguet, F. *et al.* The impact of sex on gene expression across human tissues. *Science (1979)* **369**, (2020).

94. Kalantar, K. *et al.* Association of Foxp3 gene variants with risk of Hashimoto's thyroiditis and correlation with anti-tpo antibody levels. *Acta Endocrinol (Copenh)* **15**, 423–429 (2019).

95. Neag, E. J. & Smith, T. J. 2021 update on thyroid-associated ophthalmopathy. *Journal of Endocrinological Investigation 2021 45:2* **45**, 235–259 (2021).

96. Chen, H. *et al.* Lipid metabolism in chronic obstructive pulmonary disease. *International Journal of COPD* **14**, 1009–1018 (2019).

97. Zerella, J. T., Dimler, M., McGill, L. C. & Pippus, K. J. Foreign body aspiration in children: value of radiography and complications of bronchoscopy. *J Pediatr Surg* **33**, 1651–1654 (1998).

98. Khan, M. A. B. *et al.* Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *J Epidemiol Glob Health* **10**, 107–111 (2019).

99. Li, J. *et al.* Neurotensin is an anti-thermogenic peptide produced by lymphatic endothelial cells. *Cell Metab* **33**, 1449-1465.e6 (2021).

100. Dreher, S. I. *et al.* Tgf-β induction of mir-143/145 is associated to exercise response by influencing differentiation and insulin signaling molecules in human skeletal muscle. *Cells* **10**, 3443 (2021).

Annex

Supplementary Table 1 – Tissue-specific network metrics. Weighted diameter, centralization, average path length, weighted average path length and average weighted degree.

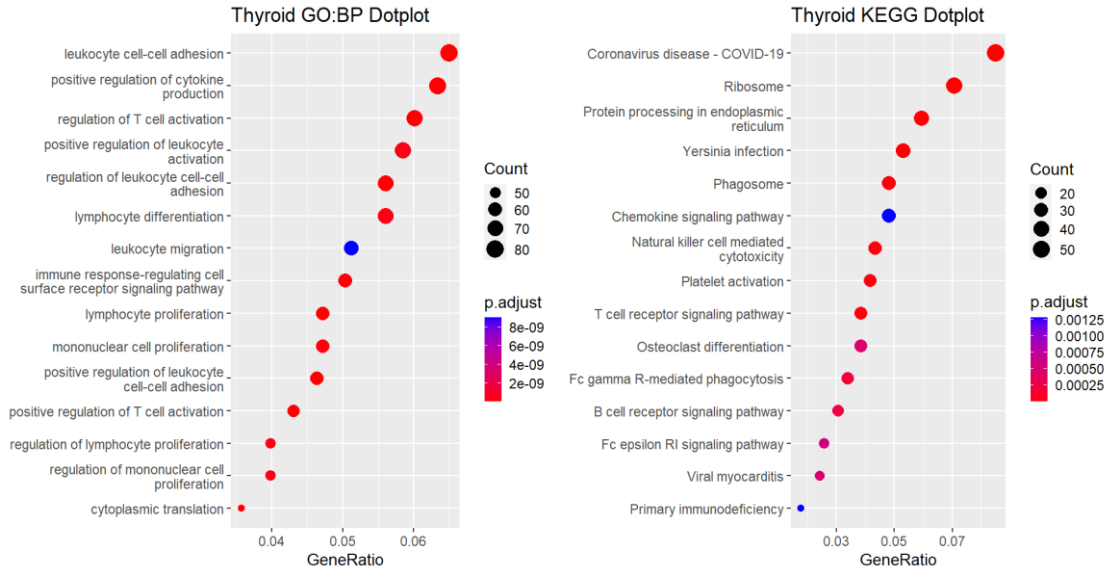
Tissue	Weighted Diameter	Centralization	Average Path Length	Weighted Average Path Length	Average Weighted Degree
BrainSpinalcordcervicalc1	11,82	0,63	1,94	3,39	3,29
BrainFrontalCortexBA9	301,31	0,64	1,98	7,01	17,63
ArteryCoronary	222,59	0,89	2,39	8,95	8,74
Pituitary	241,43	0,91	2,32	10,63	10,37
BrainAmygdala	206,02	0,44	1,93	15,05	25,98
SmallIntestineTerminalIleum	160,29	0,61	2,96	15,39	10,9
HeartLeftVentricle	389,92	0,66	2,87	16	15,47
Spleen	377,69	0,84	2,52	17,52	14,35
BreastMammaryTissue	747,51	0,91	2,33	19,1	26,58
Thyroid	991,21	0,75	2,72	19,14	35,22
Liver	218,09	0,36	3,71	20,81	16,89
BrainNucleusaccumbensbasalganglia	390,13	0,54	2,70	24,29	51,93
Lung	691,3	0,49	3,21	26,3	30,29
BrainCaudatebasalganglia	314,28	0,77	2,78	27,16	20,49
SkinNotSunExposedSuprapubic	465,33	0,58	2,80	28,05	55,4
EsophagusGastroesophagealJunction	537,61	0,64	2,99	30,39	24,05
HeartAtrialAppendage	691,39	0,41	3,27	33,11	32,33
Vagina	347,9	0,80	2,61	37,19	26,86
SkinSunExposedLowerleg	1084,43	0,59	2,89	40,48	54,25
Pancreas	670,8	0,46	3,32	42,31	38,58
BrainHippocampus	227,04	0,37	2,77	43,41	30,43
EsophagusMucosa	1061,05	0,70	2,91	61,18	63,25
Testis	1209,65	0,49	3,26	62,99	68,28
BrainSubstantianigra	262,44	0,52	1,96	67,76	50,78
BrainCerebellarHemisphere	526,53	0,40	3,28	71,46	56,64
BrainPutamenbasalganglia	414,28	0,47	3,18	87,14	69,79
MuscleSkeletal	1016,76	0,41	3,31	95,33	135,85
BrainCerebellum	625,11	0,45	3,25	97,43	80,53
AdiposeVisceralOmentum	792,44	0,58	3,17	97,69	106,46
EsophagusMuscularis	1102,99	0,39	3,29	118,39	114,41
Prostate	642,11	0,69	2,84	119,32	107,05
WholeBlood	825,87	0,41	3,34	121,09	120,8
BrainAnteriorcingulatecortexBA24	357,75	0,56	1,97	136,86	100,67
AdiposeSubcutaneous	1267,4	0,56	3,05	151,25	211,66
ColonTransverse	773,18	0,51	3,10	160,77	140,48
Stomach	832,05	0,33	3,84	168,66	114,21
MinorSalivaryGland	604,36	0,94	2,22	177,25	172,2
AdrenalGland	812,51	0,75	2,76	183,82	176,76

BrainHypothalamus	519,9	0,98	2,09	200,03	197,46
Ovary	764,01	0,94	2,21	200,67	199,82
BrainCortex	603,71	0,78	2,57	205,41	187,75
Uterus	669,53	0,99	2,03	210,53	210,25
ArteryTibial	1427,93	0,67	2,86	222,99	277,54
NerveTibial	1695,06	0,69	2,77	239,51	330,89
ColonSigmoid	838,31	0,66	2,93	256,95	234,73
ArteryAorta	1302,56	0,87	2,45	305,22	325,8

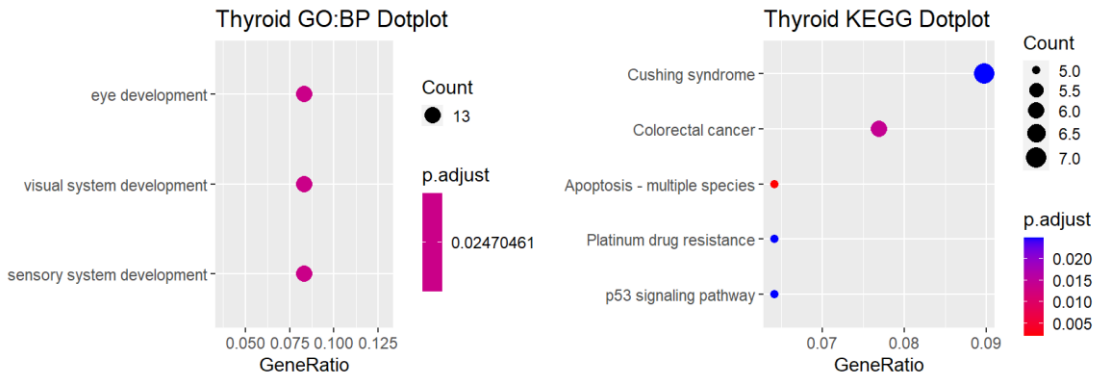
Supplementary Table 2 – Tissue-specific network metrics. Number of nodes, edges and components, the diameter, and the density

Tissue	Nº of Nodes	Nº of Edges	Diameter	Density	Nº of Components
BrainSpinalcordcervicalc1	57	55	2	0,034	2
BrainFrontalCortexBA9	164	161	2	0,012	3
ArteryCoronary	999	1003	6	0,002	2
Pituitary	2228	2259	6	0,001	1
BrainAmygdala	81	78	2	0,024	3
SmallIntestineTerminalIleum	191	192	4	0,011	1
HeartLeftVentricle	1144	1193	6	0,002	3
Spleen	722	727	4	0,003	2
BreastMammaryTissue	8520	9471	6	0,000	1
Thyroid	9677	13091	6	0,000	2
Liver	697	724	6	0,003	1
BrainNucleusaccumbensbasalganglia	220	219	4	0,009	2
Lung	6343	7739	6	0,000	1
BrainCaudatebasalganglia	453	457	6	0,004	1
SkinNotSunExposedSuprapubic	9525	13048	6	0,000	1
EsophagusGastroesophagealJunction	686	697	6	0,003	2
HeartAtrialAppendage	1697	1790	4	0,001	2
Vagina	599	600	4	0,003	1
SkinSunExposedLowerleg	5797	6806	4	0,000	2
Pancreas	1571	1655	4	0,001	2
BrainHippocampus	94	92	4	0,021	2
EsophagusMucosa	2039	2129	6	0,001	2
Testis	3318	3596	6	0,001	2
BrainSubstantianigra	124	121	2	0,016	3
BrainCerebellarHemisphere	840	858	4	0,002	1
BrainPutamenbasalganglia	407	409	4	0,005	2
MuscleSkeletal	4741	5972	4	0,001	1

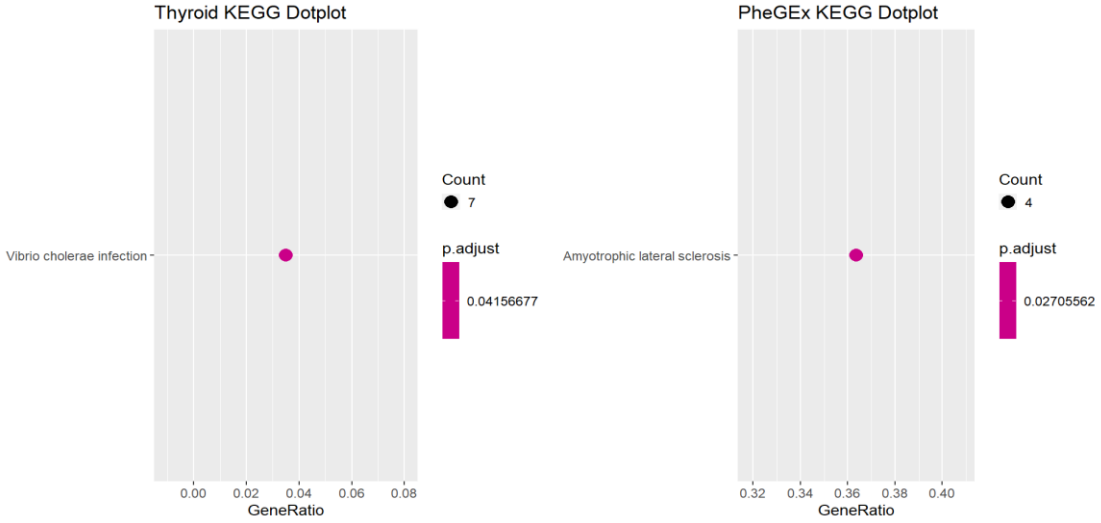
BrainCerebellum	689	694	4	0,003	2
AdiposeVisceralOmentum	3807	4271	4	0,001	2
EsophagusMuscularis	1694	1757	6	0,001	1
Prostate	466	466	4	0,004	1
WholeBlood	1974	2130	4	0,001	1
BrainAnteriorcingulatecortexBA24	169	166	2	0,012	3
AdiposeSubcutaneous	7669	10337	6	0,000	1
ColonTransverse	870	881	4	0,002	1
Stomach	911	927	6	0,002	2
MinorSalivaryGland	1692	1699	6	0,001	1
AdrenalGland	1287	1308	6	0,002	2
BrainHypothalamus	2156	2158	6	0,001	1
Ovary	2615	2651	4	0,001	1
BrainCortex	791	792	4	0,003	2
Uterus	3630	3633	4	0,001	1
ArteryTibial	7310	9130	6	0,000	2
NerveTibial	7346	9884	4	0,000	1
ColonSigmoid	1004	1015	4	0,002	1
ArteryAorta	6360	6986	6	0,000	1



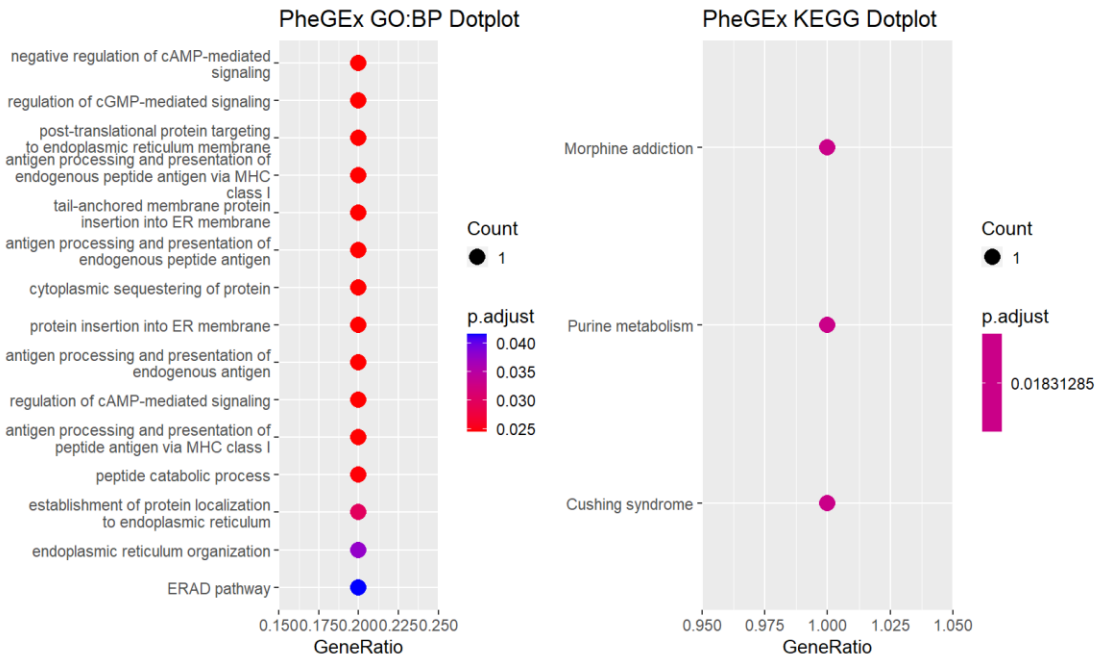
Supplementary Figure 1 – HT-sex enriched only for KEGG and GO:BP in the thyroid. Up to 15 terms are shown. The GO:BP plot shows simplified results (the clusterProfiler function “simplify” was used).



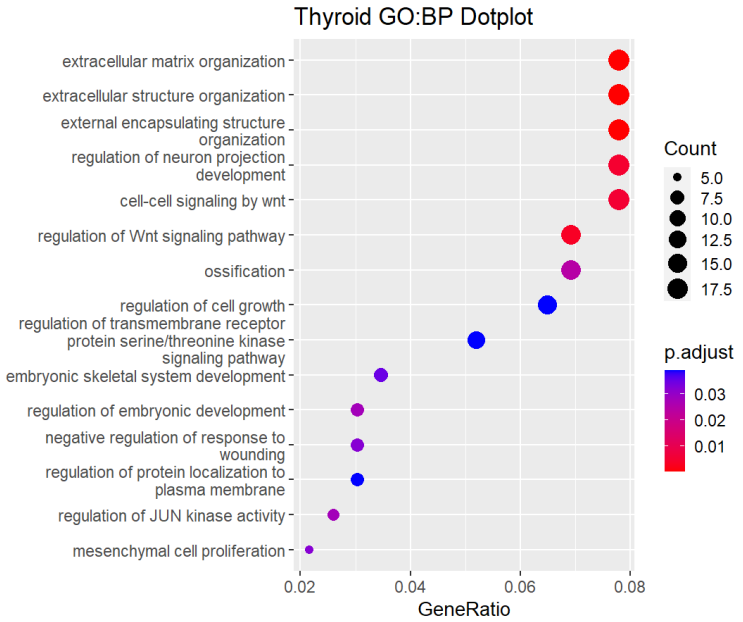
Supplementary Figure 2 – HT-age enriched only for KEGG and GO:BP in the thyroid. Up to 15 terms are shown. The GO:BP plot shows simplified results (the clusterProfiler function “simplify” was used).



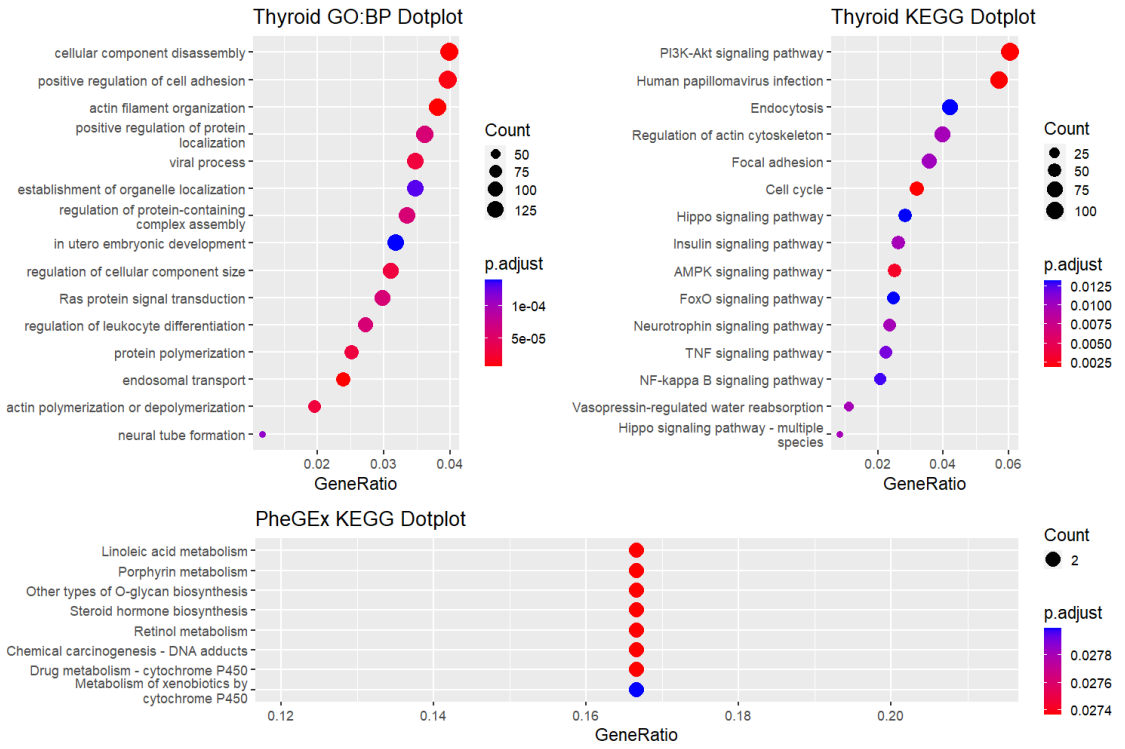
Supplementary Figure 3 – HT-ancestry enriched only for KEGG and GO:BP in the thyroid. Up to 15 terms were included. The GO:BP plot shows simplified results (the clusterProfiler function “simplify” was used).



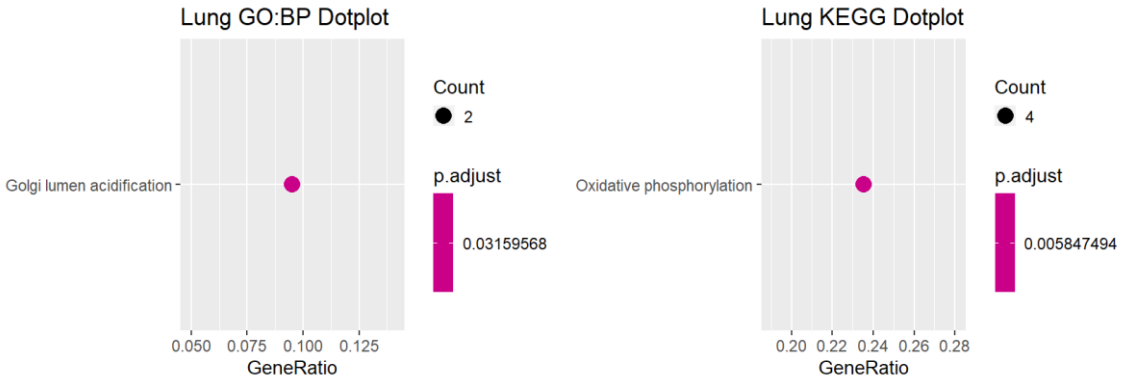
Supplementary Figure 4 – HT-BMI enriched only for KEGG and GO:BP across tissues. Up to 15 terms were included. The GO:BP plot shows simplified results (the clusterProfiler function "simplify" was used).



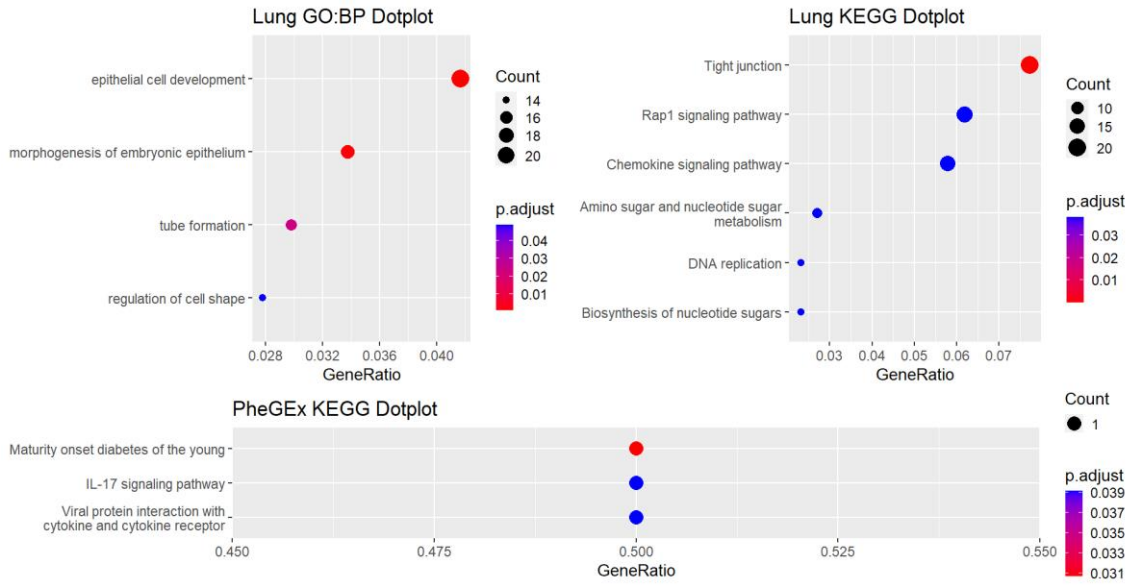
Supplementary Figure 5 – HT-hyperplasia enriched just for GO:BP in the thyroid. Up to 15 terms are shown. The GO:BP plot shows simplified results (the clusterProfiler function "simplify" was used).



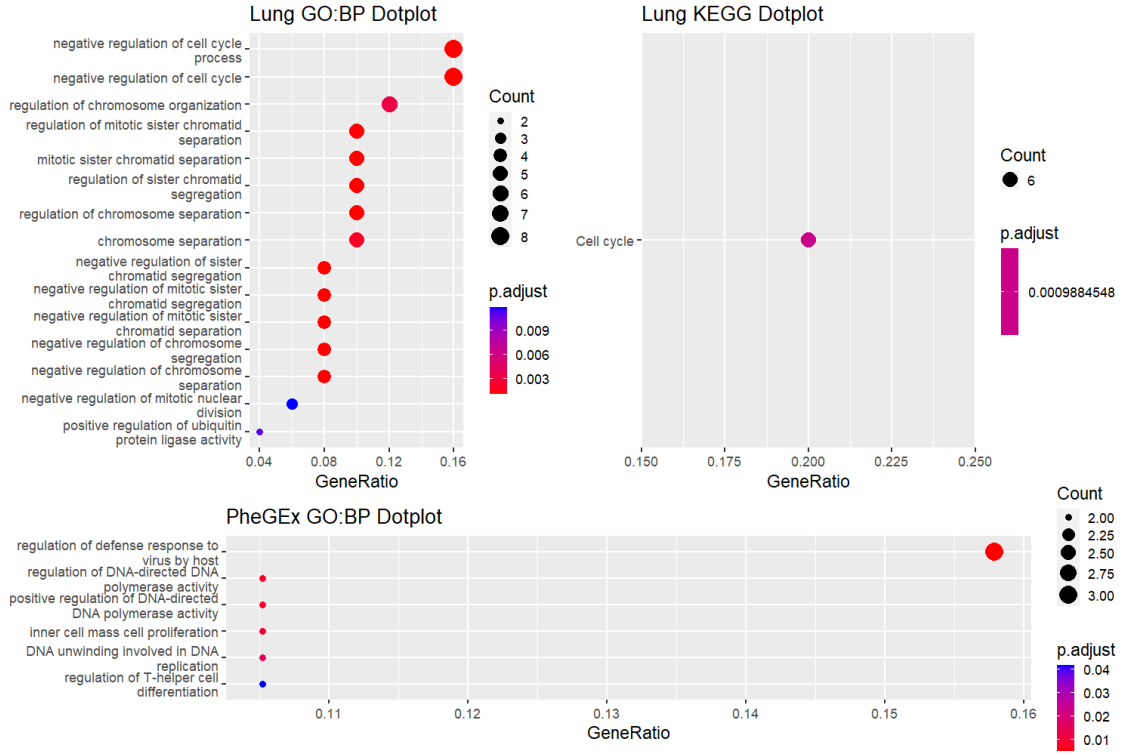
Supplementary Figure 6 – HT enrichment results. The thyroid enrichment results are on top, whilst on the bottom we have the PheGEx results. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



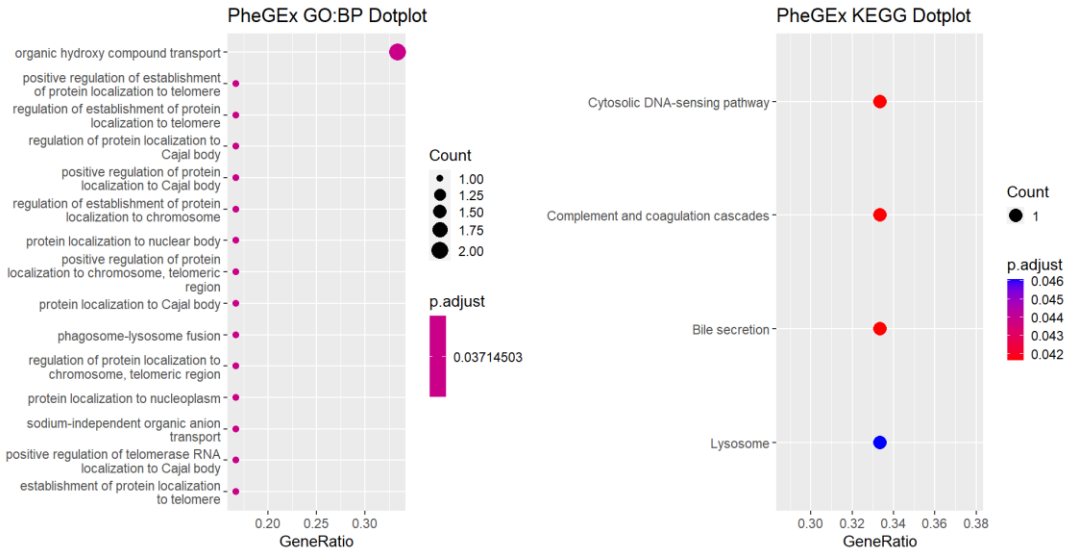
Supplementary Figure 7 – pneumonia-emphysema only enriched for KEGG and GO:BP on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



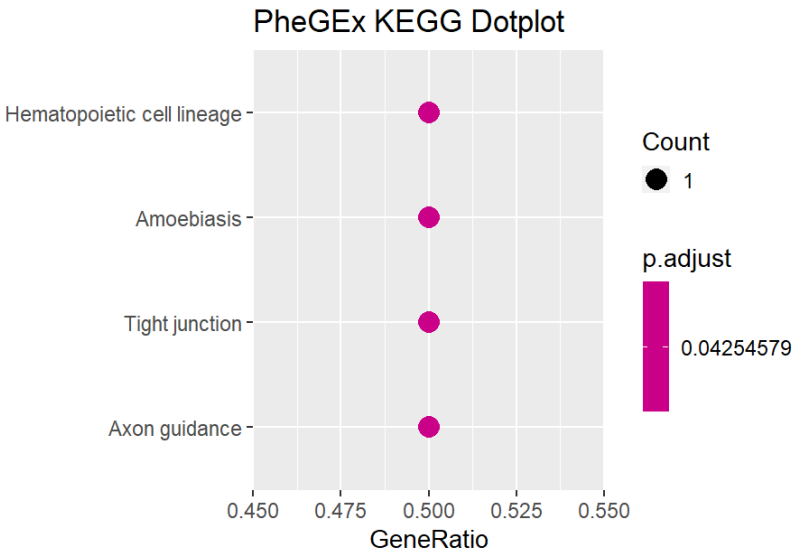
Supplementary Figure 8 – pneumonia-atelectasis enrichment results. The lung enrichment results are on top, whilst the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



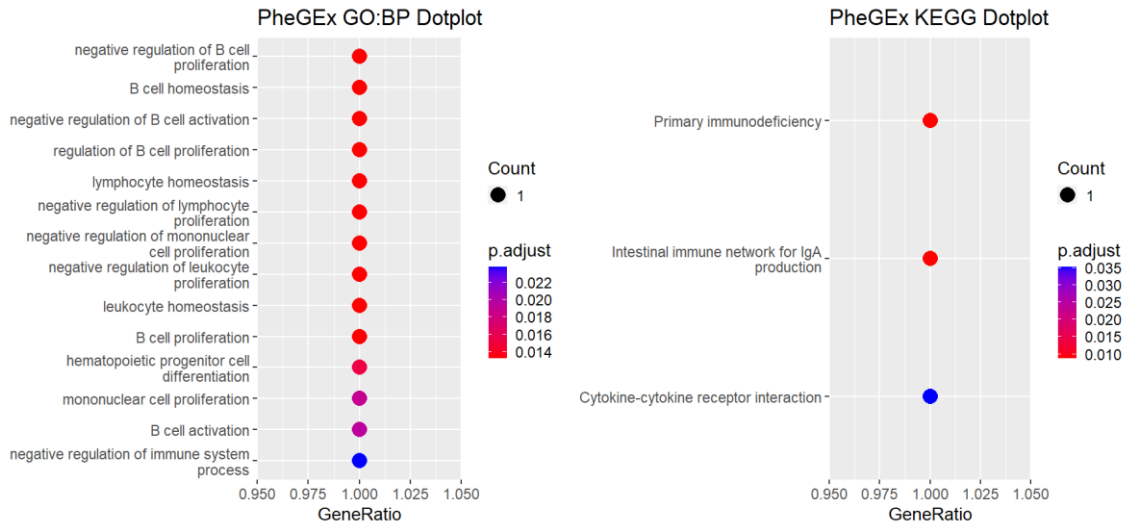
Supplementary Figure 9 – pneumonia-age enrichment results. The lung enrichment results are on top, whilst on the bottom we have the PheGEx results. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



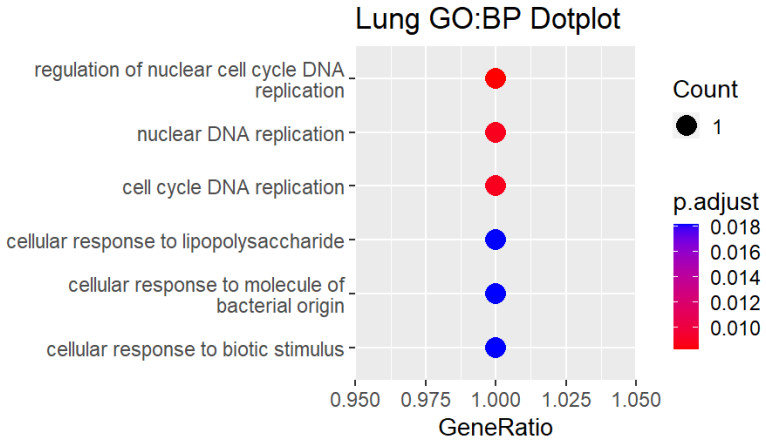
Supplementary Figure 10 – pneumonia-BMI only enriched for KEGG and GO:BP across tissues. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



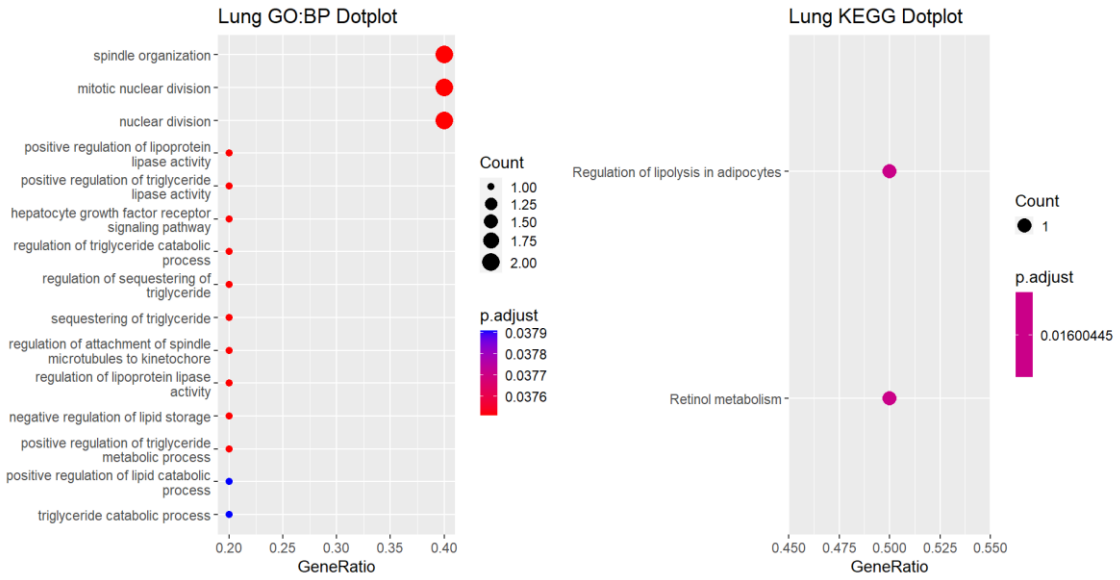
Supplementary Figure 11 – pneumonia-ancestry only enriched for KEGG across tissues. Up to 15 terms were included.



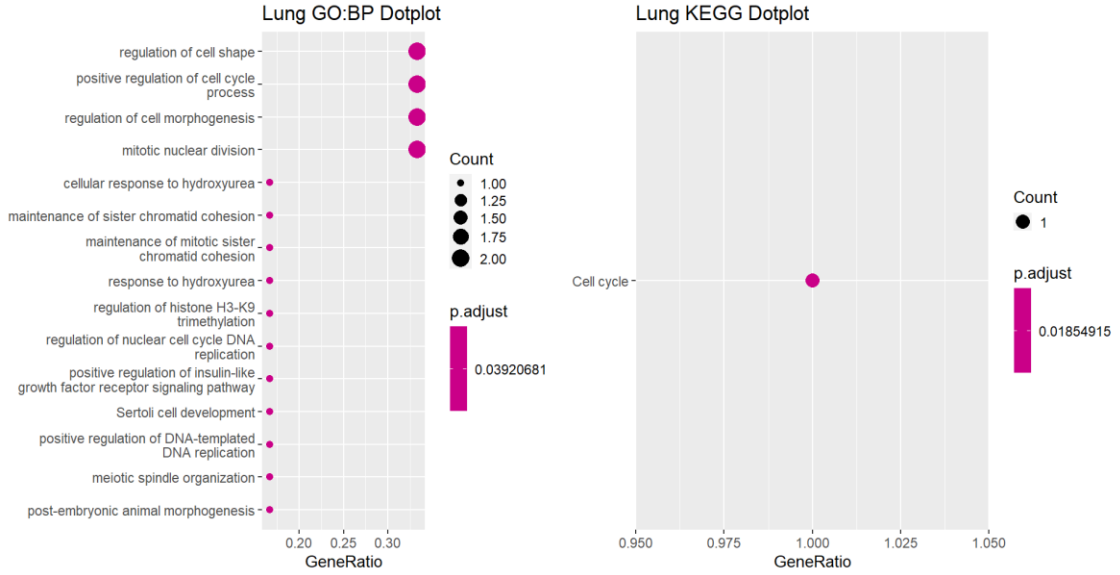
Supplementary Figure 12 – emphysema-sex only enriched for KEGG and GO:BP across tissues. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



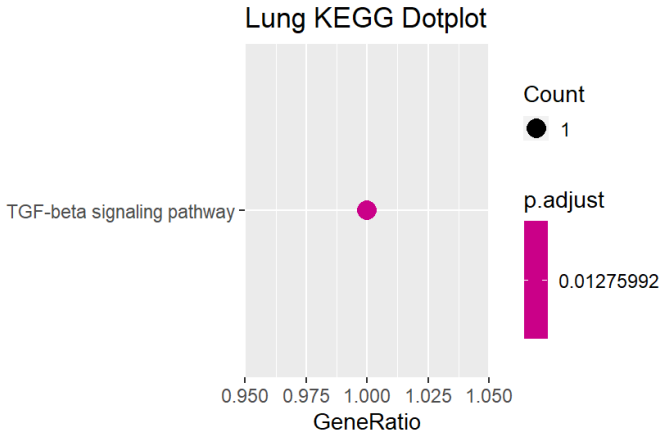
Supplementary Figure 13 – emphysema-age only enriched for GO:BP on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



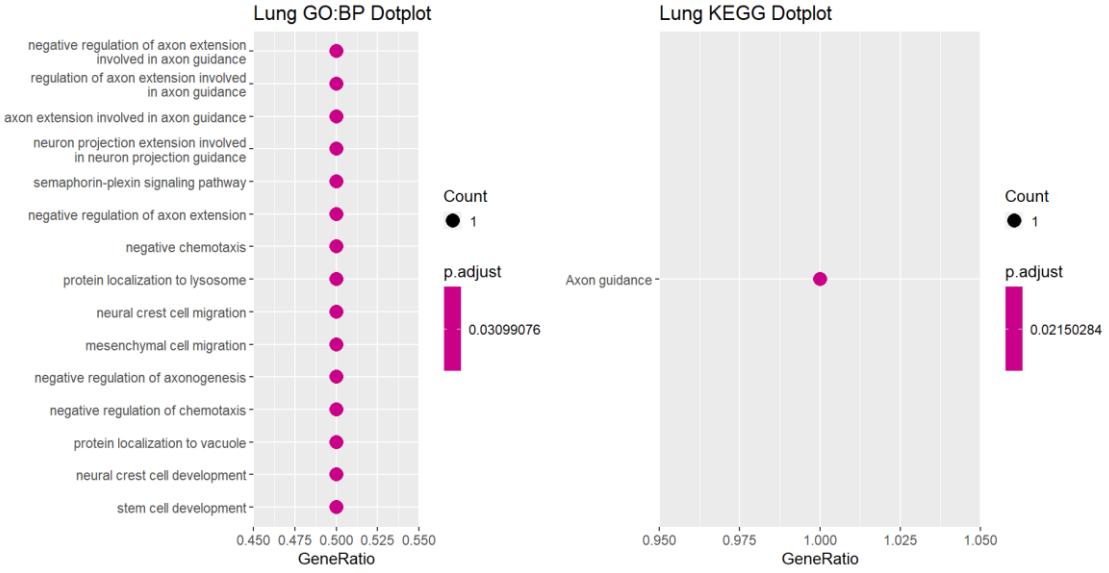
Supplementary Figure 14 – emphysema-ancestry only enriched for KEGG and GO:BP on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



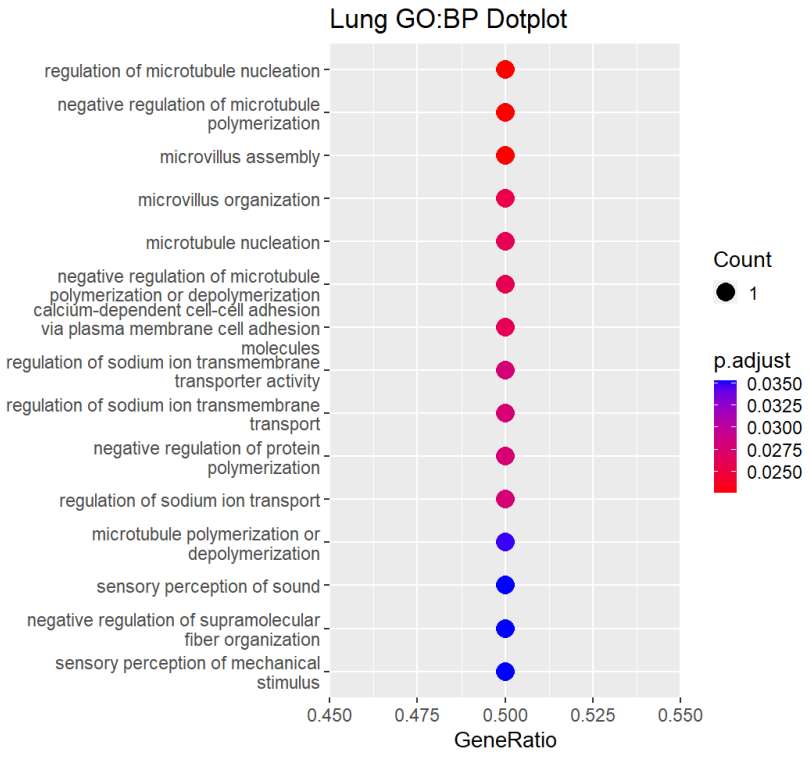
Supplementary Figure 15 – emphysema-BMI only enriched for KEGG and GO:BP on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



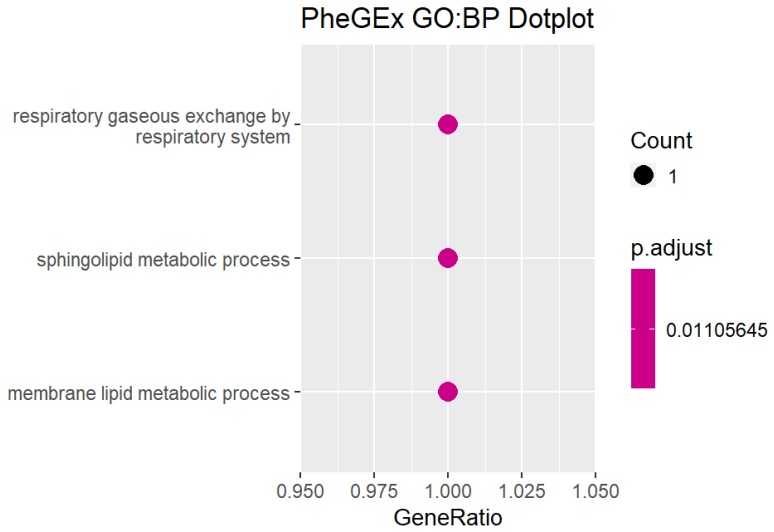
Supplementary Figure 16 – atelectasis-sex only enriched for KEGG on the lung. Up to 15 terms were included.



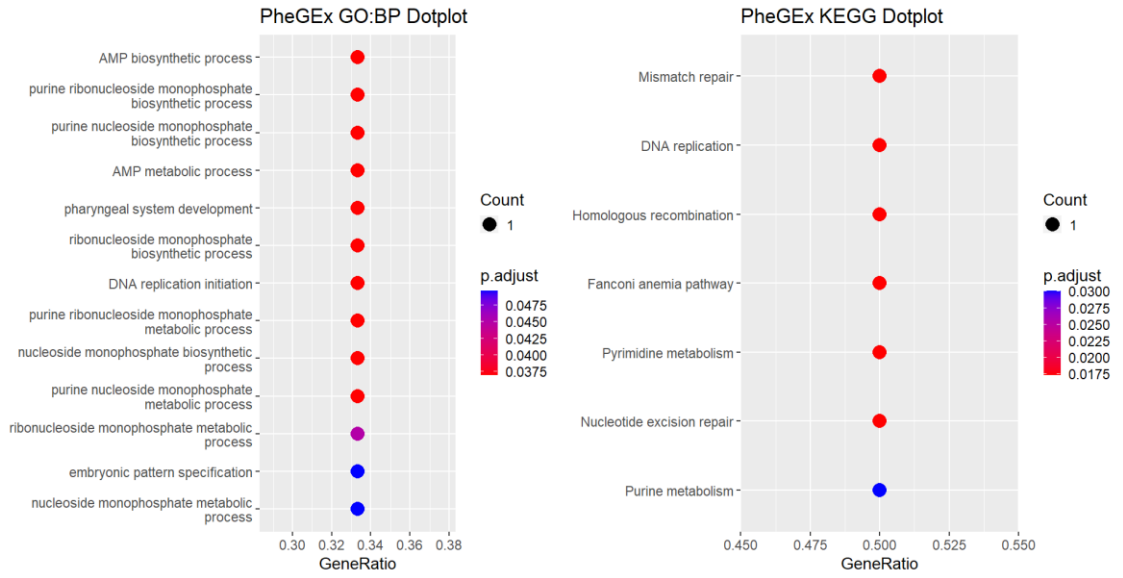
Supplementary Figure 17 – fibrosis-emphysema only enriched for GO:BP and KEGG on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



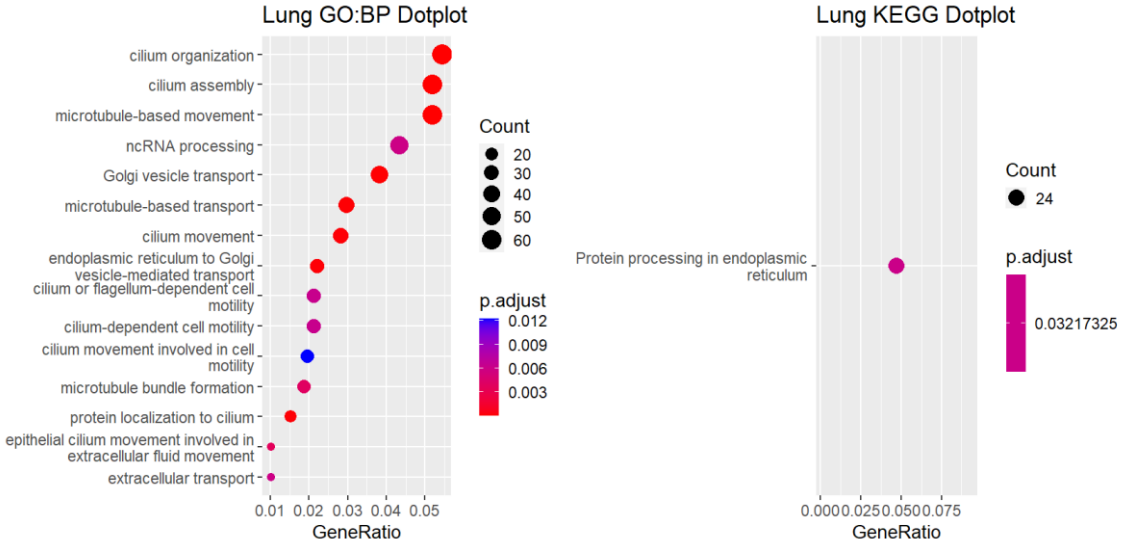
Supplementary Figure 18 – fibrosis-atelectasis only enriched for GO:BP on the lung. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function "simplify" was used).



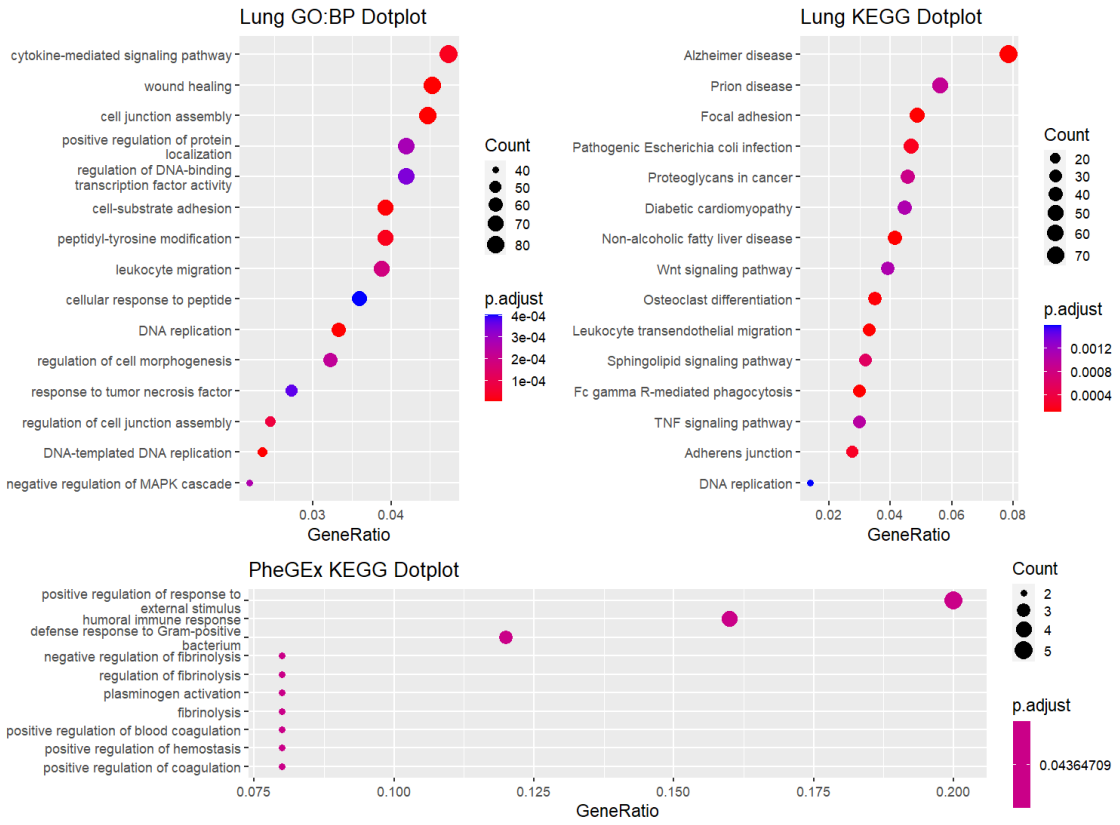
Supplementary Figure 19 – fibrosis-pneumonia only enriched for GO:BP across tissues. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function "simplify" was used).



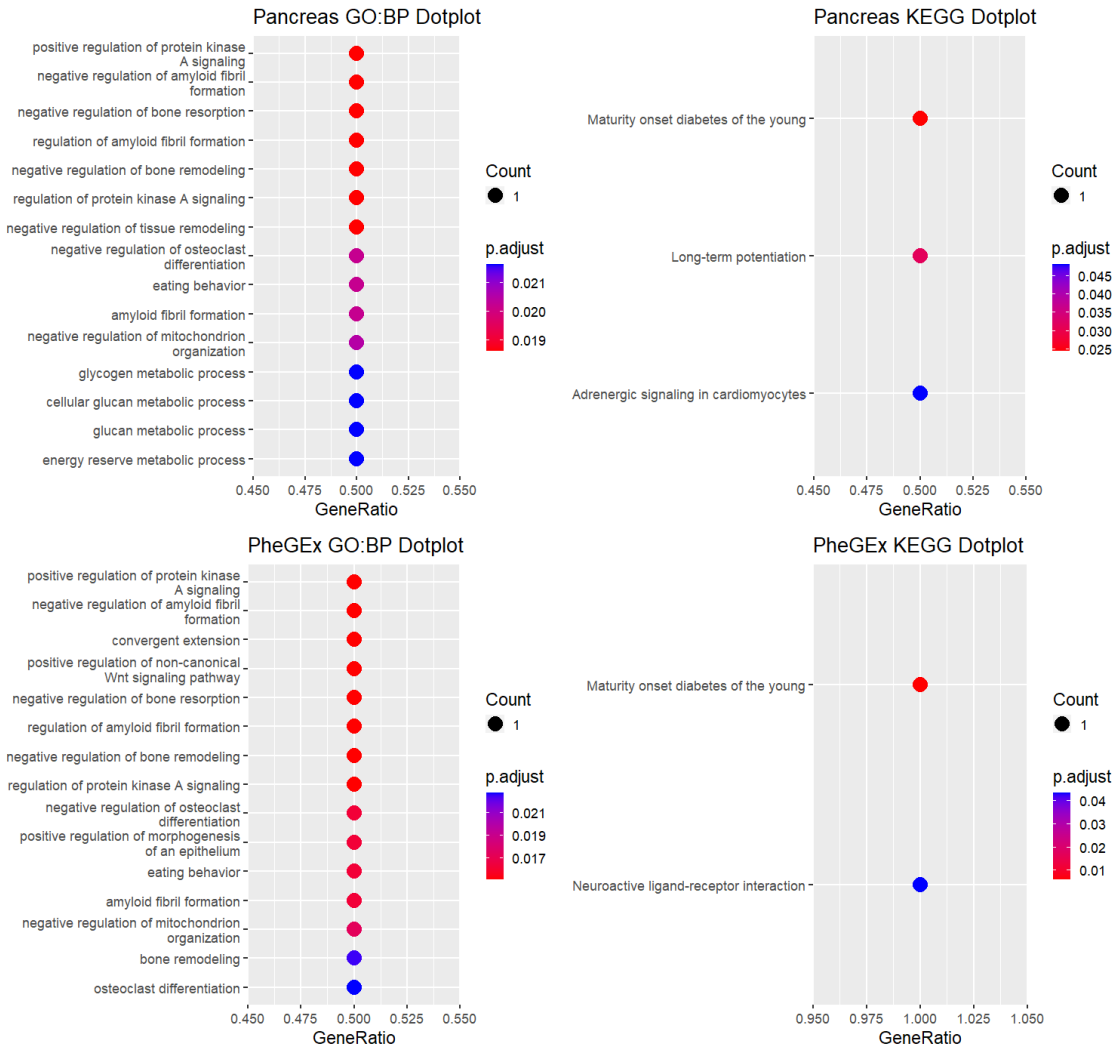
Supplementary Figure 20 – emphysema only enriched for KEGG and GO:BP across tissues. Terms were only obtained for the PheGEx network. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



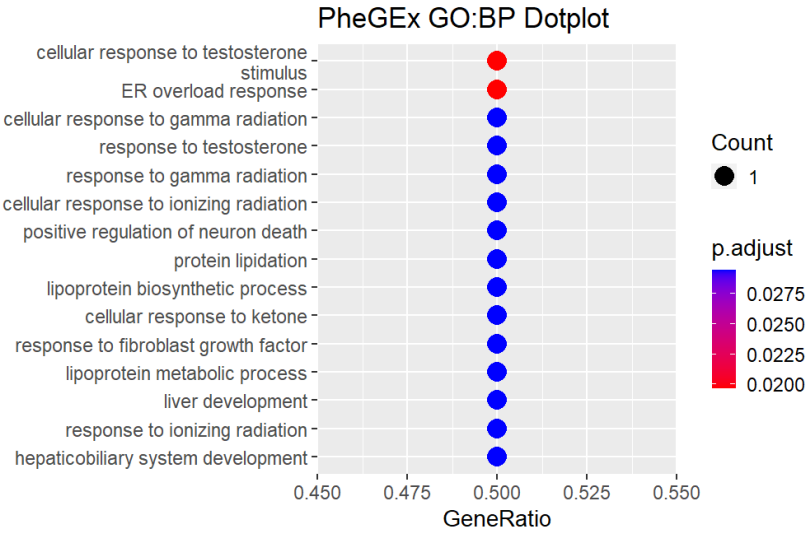
Supplementary Figure 21 – atelectasis only enriched for KEGG and GO:BP on the lung. Terms were only obtained in the lung tissue. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



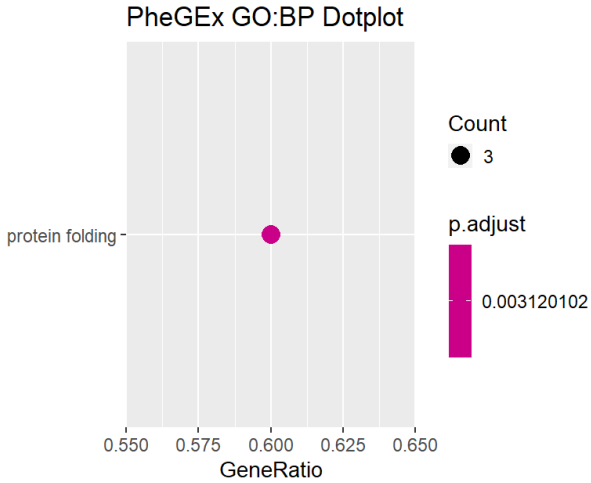
Supplementary Figure 22 – pneumonia enrichment results. The enrichment results on the Lung are on top and the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



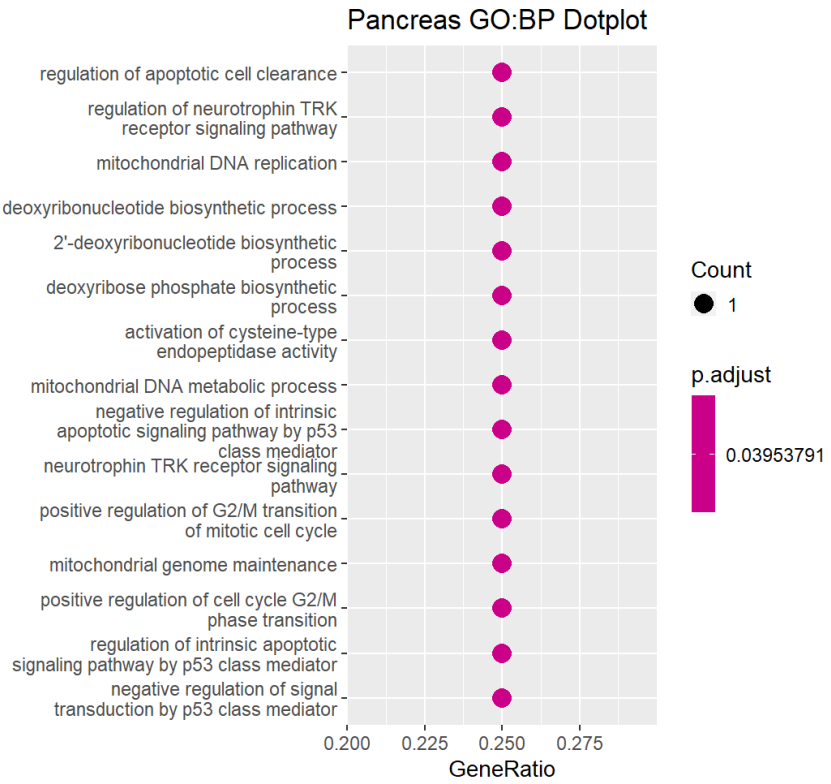
Supplementary Figure 23 – T1D-T2D enrichment results. The enrichment results on the Pancreas are on top and the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



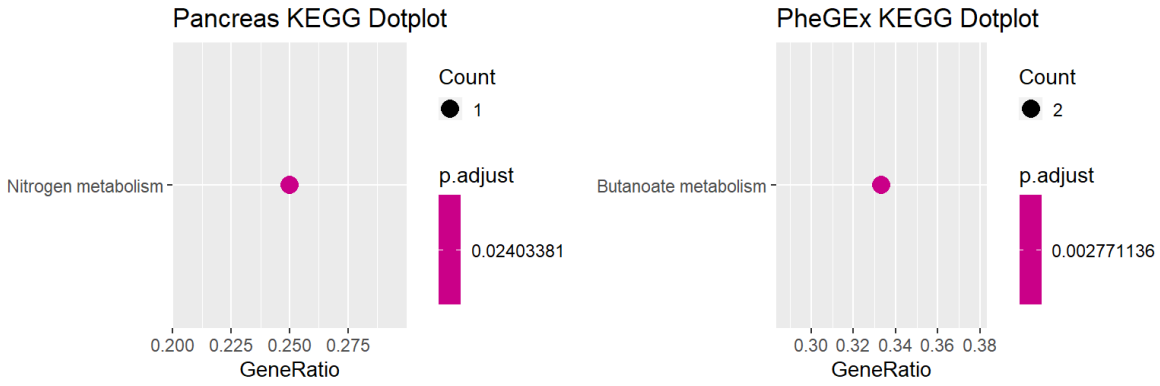
Supplementary Figure 24 – T1D-BMI only enriched with GO:BP across tissues. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



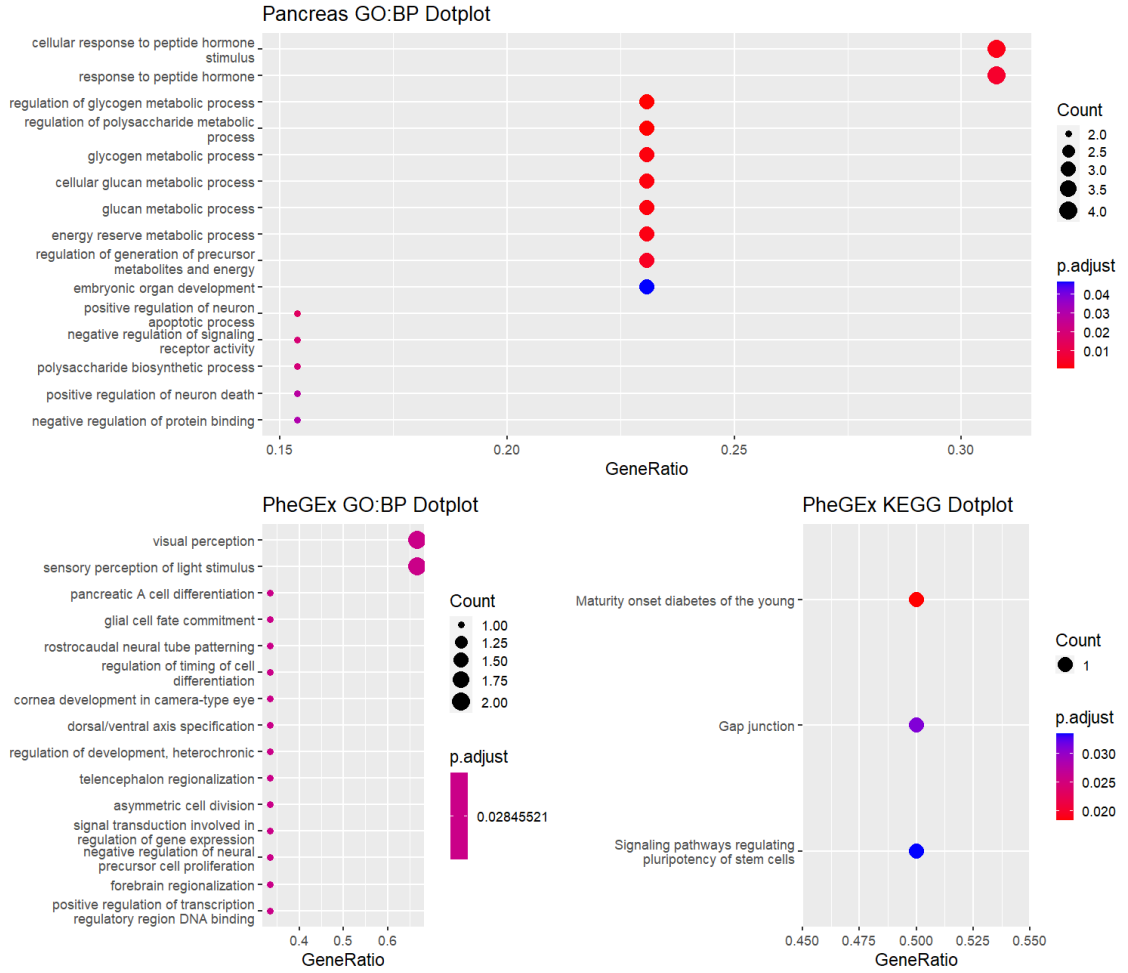
Supplementary Figure 25 – T2D-BMI only enriched with GO:BP on the PheGEx. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



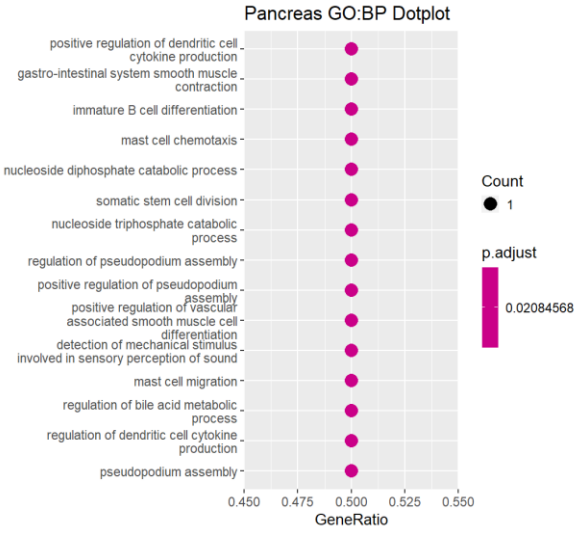
Supplementary Figure 26 – T1D-age only enriched with GO:BP in the pancreatic tissue. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function "simplify" was used).



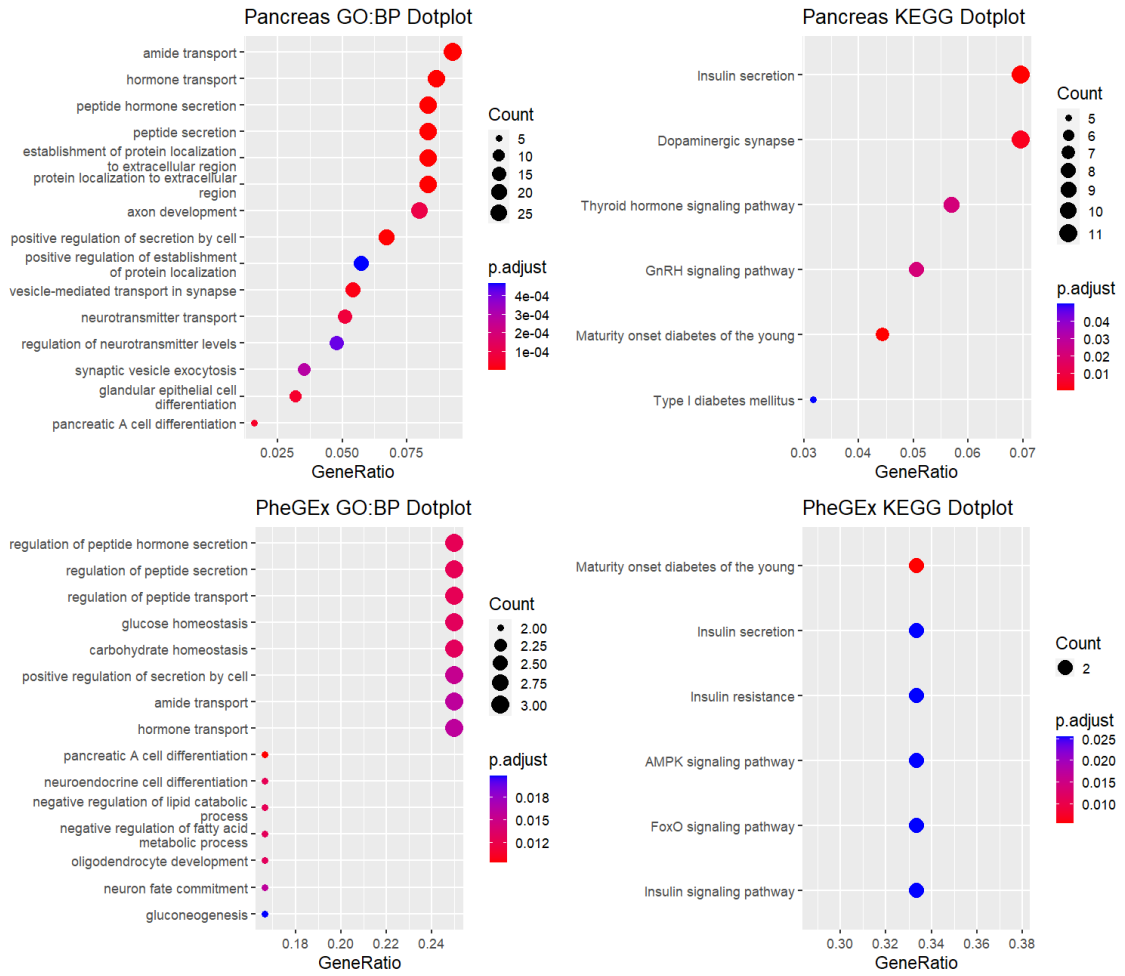
Supplementary Figure 27 – T1D-sex only enriched with KEGG in both the pancreatic tissue and on the PheGEx. Up to 15 terms were included.



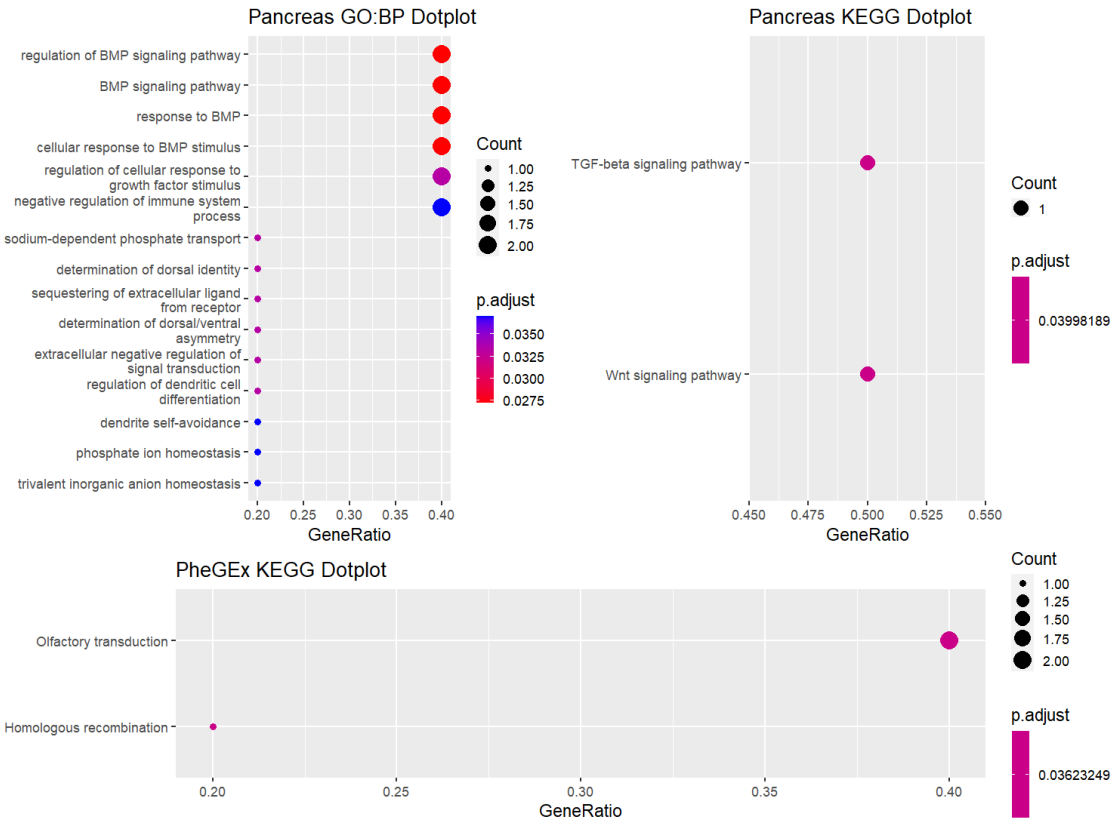
Supplementary Figure 28 – T1D-ancestry enrichment results. The enrichment results on the pancreas are on top and the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



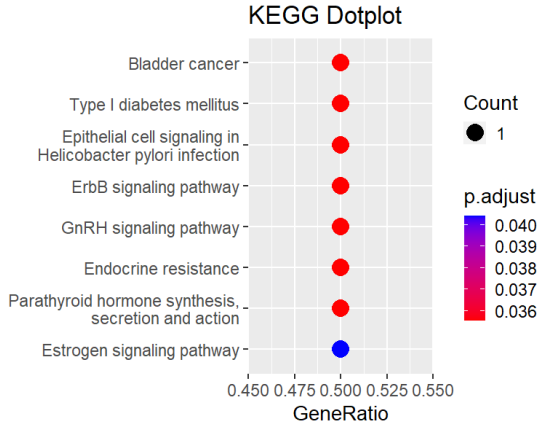
Supplementary Figure 29 – T2D-ancestry only enriched with GO:BP in the pancreatic tissue. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



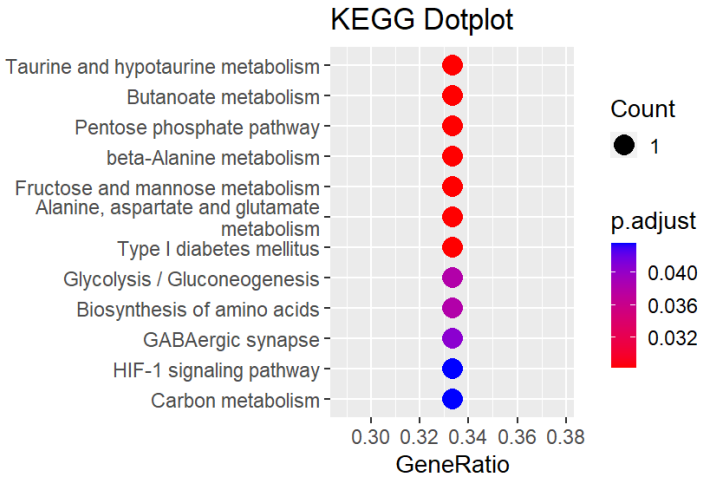
Supplementary Figure 30 – T1D enrichment results. The enrichment results on the Pancreas are on top and the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



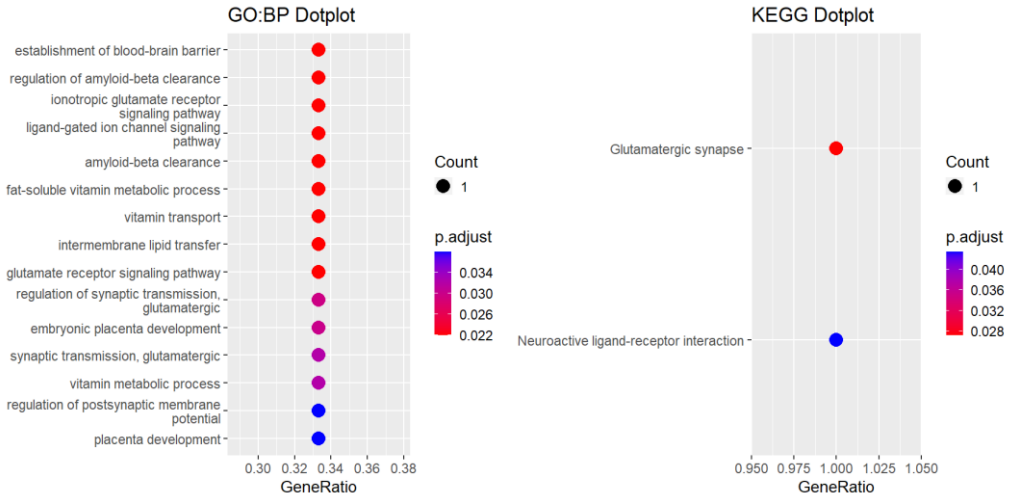
Supplementary Figure 31 – T2D enrichment results. The enrichment results on the Pancreas are on top and the PheGEx results are on the bottom. Up to 15 terms were included. The GO:BP plot shows the simplified results (clusterProfiler function “simplify” was used).



Supplementary Figure 32 – T1D-age restricted enrichment only returned GO:BP terms. Up to 15 terms were included.

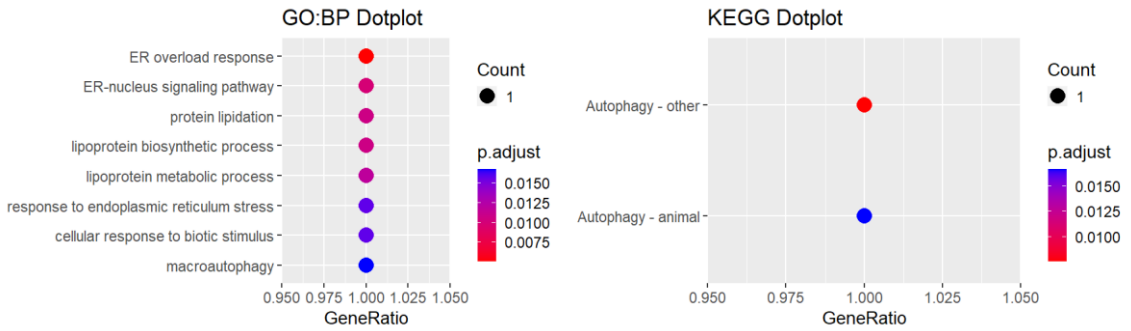


Supplementary Figure 33 – T1D-sex restricted enrichment only returned KEGG terms. Up to

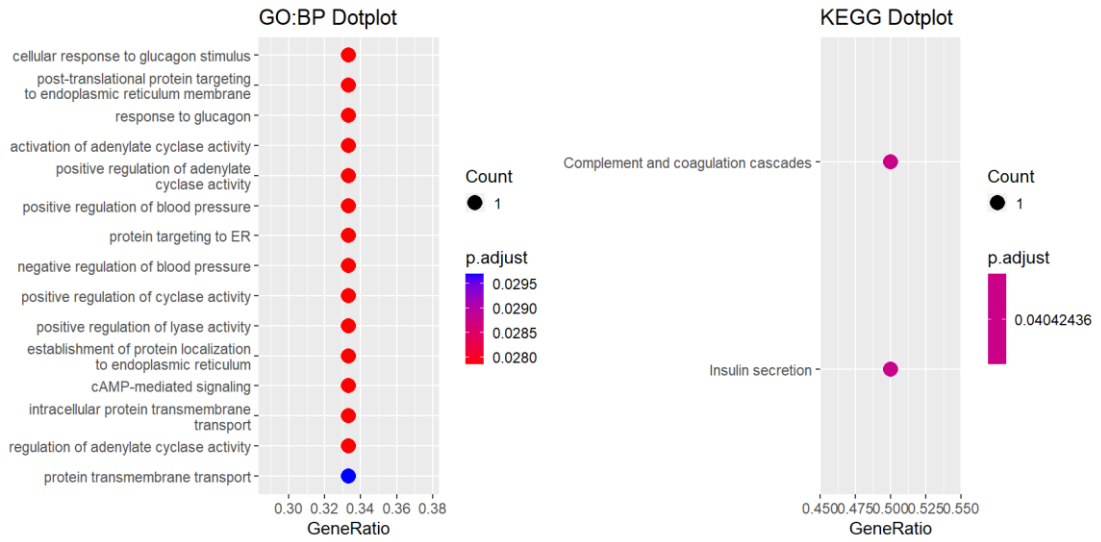


Supplementary Figure 34 – T2D-sex restricted enrichment GO:BP and KEGG terms. Up to 15 terms were included.

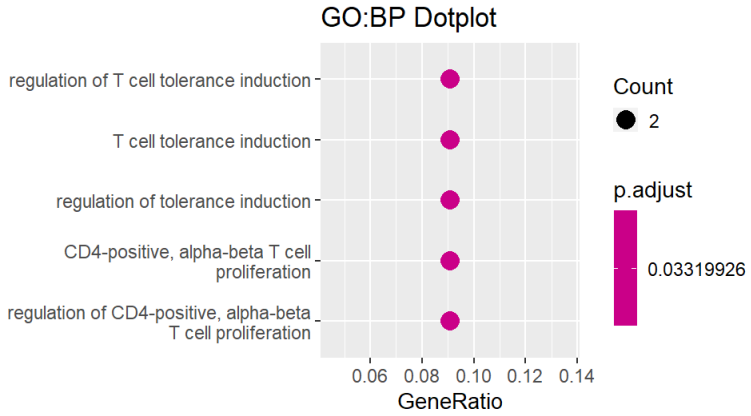
15 terms were included.



Supplementary Figure 35 – T2D-BMI restricted enrichment GO:BP and KEGG terms. Up to 15 terms were included.



Supplementary Figure 36 – T2D-ancestry restricted enrichment GO:BP and KEGG terms. Up to 15 terms were included.



Supplementary Figure 37 – HT-sex restricted enrichment only returned GO:BP terms. Up to 15 terms were included.