Universitat Oberta de Catalunya

uoc.edu

Pathway and network analysis of Alzheimer's disease

Ximena Sofia Lemus Maulen
Master's Degree in Bioinformatics and Biostatistics
Area 2

Jaime Sastre Tomas
Marc Maceira Duch

June 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Pathway and network analysis of Alzheimer's disease*
Nombre del autor:	*Ximena Sofia Lemus Maulen*
Nombre del consultor/a:	*Jaime Sastre Tomas*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Master's Degree in Bioinformatics and Biostatistics*
Área del Trabajo Final:	*Area 2*
Idioma del trabajo:	English
Número de créditos:	15 (TFM)
Palabras clave	*Complex diseases, pathway and network analysis, comparative study of software tools*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

Actualmente existen un tipo de estudios llamados GWAS que han permitido identificar varias variantes genéticas asociadas a enfermedades complejas. Sin embargo, en la mayoría de casos, se desconoce qué mecanismos biológicos participan entre la ocurrencia de las variantes genéticas y el desarrollo de la enfermedad.

Nuestra propuesta es la utilización de análisis de rutas biológicas y de redes génicas junto con la integración de resultados de diferentes datos ómicos. Para ello, seleccionamos listas de genes representativas de datos transcriptómicos y genómicos de previos estudios de Alzheimer, y comparamos los resultados obtenidos con los programas: ConsensusPathDB, DAVID EnrichmentMap, g:Profiler, IPA, MAGMA, Reactome y WebGestalt.

Algunos de los procesos biológicos sobrerrepresentados en la enfermedad de Alzheimer estaban relacionados con respuestas inmunológicas, procesos proteico-lipídicos, desarrollo neuronal, ciclo del ácido cítrico, péptido beta-amiloide, y metabolismo de ácidos nucleicos. La comparación de métodos de redes génicas no fue concluyente, pero los resultados que se obtuvieron coinciden con los resultados de los estudios de los datos en los basamos nuestro proyecto.

Una conclusión fue que el análisis de diferentes datos biológicos junto con la aplicación de análisis de rutas biológicas y génicas aporta una perspectiva

sistémica y funcional de la enfermedad de Alzheimer. Otra conclusión fue que el número de resultados significativos difiere entre programas de rutas biológicas a pesar de analizar los mismos datos y aplicar el mismo tipo de método. Sin embargo, generalmente los resultados más relevantes coinciden entre programas si se selecciona la misma base de datos, pero el nivel de significancia varia.

Abstract (in English, 250 words or less):

Nowadays, a type of studies called GWAS have identified several risk alleles associated with different complex diseases. However, in most of the cases, the identity of the biological mechanisms involved in the development of complex diseases since the occurrence of genetic variants is yet unknown.

Our proposal is to conduct pathway and network analysis together with the integration of results from different omics data. We selected to use lists of genes representative of transcriptomic and genomic data from previous Alzheimer's disease studies. Moreover, we decided to compare the software tools: ConsensusPathDB, DAVID, EnrichmentMap, g:Profiler, IPA, MAGMA, Reactome, and WebGestalt.

Some of the pathways enriched in Alzheimer's disease were related to immune responses, protein-lipid processes, citric acid cycle, amyloid-beta peptide, neuronal development, and nucleic acid metabolism. The comparison of network analyses was inconclusive. However, our results coincide with the results seen in the studies of the data we base our project on.

One of our conclusions was that pathway and network analysis together with the analysis of different biological data gave us a systemic and functional insight into Alzheimer's disease. Another conclusion was that different number of significant results are obtained when using different tools, despite analysing the same data and applying the same pathway analysis method. Nevertheless, in general, if the same database is selected, the most relevant results coincide between tools, but the level of significance differs.

Table of Contents

List of Figures

List of Tables

1 Abstract

The effect of multiple genetic variants contributes to the development of complex diseases. Nevertheless, in most of the cases their functional roles are yet to be discovered. The integration of different types of biological data together with the usage of pathway and network analysis has been proposed to unravel the biological mechanisms behind the development of complex diseases from a systemic and functional point of view.

For our project, we decided to compare: (i) the results from different tools that offer pathway and/or network analysis, and (ii) the results obtained when different types of biological data are used for pathway and network analysis. We selected Alzheimer's disease as an example of complex disease, gene lists representative of transcriptomic and genomic data from previous Alzheimer's disease studies, and eight software tools to compare. The novelty of our project is the double comparison we carried out and one of questions we want to answer: whether there is any loss in biological information when pathway and network analysis are carried out using only one tool and only one type of biological data.

Some of the pathways found to be enriched in Alzheimer's disease were processes related to: synaptic signalling, neuronal development, transmembrane transport, citric acid cycle, cellular respiration, intestinal lipid absorption, protein-lipid processes, amyloid-beta peptide, and immunological responses. The comparison of network analysis methods was not conclusive but, the results obtained coincided with the results of the studies of the data we rely on to carry out our project.

In conclusion, pathway and network analysis together with the analysis of different biological data gave us a systemic and functional insight into Alzheimer's disease. Different number of significant results between tools, despite analysing the same data and applying the same pathway analysis method. Nevertheless, in general, the most relevant results coincide when using different software tools if the same database is selected for the analysis. But the level of significance differs.

2 Introduction

2.1 Background and motivation

Complex human diseases are influenced by many genetic and environmental factors [1, 2, 3, 4]. Alzheimer's disease (AD) is a neurodegenerative disease, the most common cause of dementia, and is an example of a complex disorder [5, 6, 7]. Several studies have been carried out to discover what genetic, environmental and lifestyle factors are involved in AD aetiology and

pathogenesis [1, 8]. For example, genome-wide association studies (GWAS) have had an important role in identifying single nucleotide polymorphisms (SNPs) associated with AD, other complex diseases, and complex traits [9, 10].

Most of the genetic variants associated with a particular disease identified by GWAS are common variants of small effect sizes, but some of moderate effect sizes can also be detected [1, 8]. However, part of the genetic architecture of complex diseases is not explained by GWAS because GWAS results are: associations of risk alleles with a phenotype. Those genetic variants, individually or together, explain only a small proportion of the phenotypic variance due to genetic factors [1, 8, 11]. A systems approach to complement GWAS findings in unraveling the genetic and functional basis of complex disorders has previously been proposed [8, 12, 13]. Together with the integration of multiple-omics and interaction data, pathway and network analysis can give a valuable functional insight into how such SNP markers relate to each other and what collectively effects have in complex diseases development [8, 12, 14].

There exist several software tools to carry out pathway and network analysis [12]. However, a potential problem of having a wide range of tools and no standard procedure to perform pathway and network analysis is the heterogeneity of results. Each program uses different types of input data, gene identifiers, gene set definitions, pathway annotation databases, and statistical methods [8, 12]. Therefore, in this project we would like to select a subset of the available software, to perform pathway and network analysis using different types of biological data from previous studies of AD, and to compare the results obtained.

2.2 Objectives

- To provide an in-depth description of the software characteristics and requirements for pathway and network analysis
- To compare the output of different software tools using microarray data, RNA-seq data, and GWAS summary statistics
- To demonstrate the role of network and pathway analysis in our understanding of complex diseases such as Alzheimer's disease
- To discover any loss in biological information when pathway and network analyses are carried out using only one software tool and only one type of biological data in studies of complex diseases

2.3 Research approach

Pathway and network analysis give a global and functional insight into genes' roles, the relationships between genes and the trait or disorder of study in an

increasingly automated manner thanks to the usage of several databases with numerous gene annotations and other relevant biological information [12, 15]. Previously, the only way to get a mechanistic understanding of a gene —in other words, to know in which biological pathways it has a role and with which genes it interacted–, was by carrying out exhaustive literature reviews and experiments [15]. Therefore, we will take advantage of databases that have integrated different biological information and conduct pathway and network analysis to have a functional and global comprehension of genes and molecular mechanisms involved in AD development.

An ideal systems perspective of AD would be achieved by carrying out pathway and network analysis using multiple levels of biological data; for example: genomics, transcriptomics, epigenomics, proteomics, and metabolomics data [16]. Nevertheless, in order to conduct such an extensive analysis it is essential to have tools available that can handle each of the data types, at least one dataset of each biological level, time, and funding. Moreover, by increasing the number of levels of biological data, we increase the number of covariates and potential confounders that must be taken into account and thus we increase the complexity of the analysis [16].

In this project we aim to explain the missing genetic basis of AD from a systems point of view, and due to lack of funding and short completion period we will only analyse genomics and transcriptomic data. By doing so, we will have a global-integrative perspective of the genetic basis of AD. At the same time, we will:

- Avoid deviation from the genomic level.
- Exclude the interaction between environmental and genetic factors, the effect of post-translational modifications, and several other factors that would increase the study´s complexity and affect the results due to their highly dynamic behaviour [16].
- Avoid results that are strongly dependent on sample preparation [17] and the type of tissue extracted [16].

The datasets selected to be used as references for the current project and as representatives of transcriptomic and genomic data are from Blalock et al. [18], Nativio et al. [19], and Kunkle et al. [20]. These datasets were chosen because they are freely available and because in [18, 19, 20], the authors worked with samples of controls and of patients with AD patients, and performed pathway enrichment analyses in their studies. Furthermore, [20] and [19] are recent studies, and in [19], a multi-omics strategy was conducted. Having those three studies as direct references helped us to have a better understanding of AD when using different types of biological data.

We selected the software tools mentioned in Project management section because each of the tools has unique characteristics and different approaches to perform pathway and network analysis, provides extensive tutorials or

documentation, and is freely available or offers a trial period. In addition, some of these tools are more flexible because they can work with several data types. Others are connected through Cytoscape's apps, which facilitates data integration of different platforms.

2.4 Project management

The three different types of data used in our study are:

- Microarray data [18]
- RNA-seq data [19]
- GWAS summary statistics [20]

The software tools chosen to analyse or visualise the data are:

- ConsensusPathDB [21, 22]
- DAVID [23, 24]
- EnrichmentMap [25] (Cytoscape [26] Plugin)
- g:Profiler [27]
- IPA [QIAGEN Inc.] [28]
- MAGMA [29]
- Reactome [30, 31, 32, 33]
- WebGestalt [34, 35, 36, 37]

In addition, in some cases R was also used to preprocess the datasets and create figures [38]. The Methodology section contains more information about the software tools and the datasets used.

Some of the key factors to carry out the project successfully were:

- Short learning periods to understand how each of the software works in general and which assumptions they rely on
- No technical problems or short waiting periods due to unavailability or updates of web-servers
- Proper time management to have enough time to preprocess each of the datasets, carry out each of the proposed analysis using the selected software, prepare visual reports of the results and interpret them.

Figure 1 shows a Gantt chart with two main sections. The name of the tasks are shown on the right of each task bar. The tasks related to the analysis of the project can be seen on the top part in dark blue, and the tasks related to the writing of the project are marked in light blue at the bottom part. Moreover, there are two yellow rhombi included in the Gantt chart which represent important days. One of them was the deadline to activate the evaluation license given by QIAGEN to use IPA, and the other rhombus represents the deadline to hand in the dissertation.

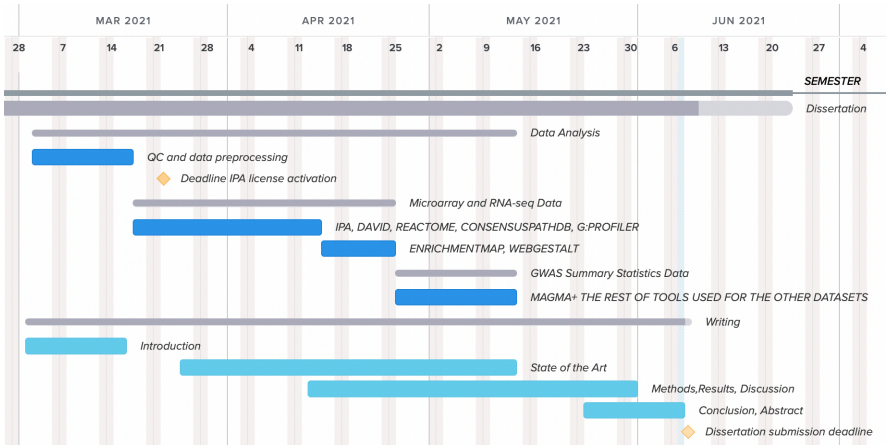


Figure 1: Gantt chart. Tasks related to the analysis of the datasets of the project are shown in dark blue. Tasks related to the writing of the project are shown in light blue, and the yellow rhombi represent important dates.

2.5 Contribution of the project to the field of study

This work illustrates the importance of applying a systems approach to have a better understanding of the biological factors involved and the effect of their interrelationships in AD aetiology and pathogenesis. Moreover, this project gives an overview of some of the available software tools and their respective approaches to perform pathway and network analysis. Lastly, the comparison of results when using the same program and different omics data of AD as input, or vice-versa, gives the opportunity to decipher whether the biological information obtained is the same overall independently of the data type or software tool used.

2.6 Brief description of the content of the next sections

In the State of the Art section, a literature review of pathway and network analyses and previous comparative studies will be shown. In the Methodology section, the general approach used to carry out the project, key information of the studies the datasets derive from, and a detailed explanation of the software tools and data used for each analysis can be found. In the Results section tables and figures together with other outcomes obtained after carrying out the analyses will be shown. The interpretation of the results will be done in the

Discussion section and a synthesis of the research topic and the significance of our findings can be found in the Conclusions section. In addition, in the Conclusions section deviations from the research project proposal, limitations of our work, questions unanswered and suggestions for future projects will be also covered. Finally, in the Glossary section the abbreviations used along the dissertation and their meaning can be found.

3 State of the Art

3.1 Introduction to some key words

The words ”pathway”, ”network”, and ”gene set”, are often used as synonyms in the literature. The fact that they are used interchangeably —when they are not synonyms despite being close in meaning — leads to create confusion among researchers when trying to integrate different biological information, create new methods or tools, interpret data, and communicate their findings. Therefore, the definition of each of those terms is not trivial and should always be given until a consensus on their definition is reached.

We will start by defining in a general way what is a biological pathway. Afterwards, as an example, we will put in context gene sets, pathways, and networks when working with genes. Despite using as an example a specific context, the definitions for each of these words will still be broad. So that the sense of these words can be later applied to other scenarios and specific definitions can be added to them. Lastly, while putting in context all of these concepts, we will also give some examples of factors that have an effect on their definitions and as a result give them more specific meanings.

A biological pathway could be broadly defined as a sequence of biological events, performed by a group of biological entities, to create a new product or induce a cellular change [11, 39, 40]. So, a sequence of biological events needs to have a specific start and endpoint to be considered a biological pathway [12].

A gene set could be defined as a pathway if they are involved in the same biological pathway [11, 13, 39]. Whereas a group of gene sets could be defined and visualised as a network if the genes are biologically related, for example, through the interaction of some of them, through the interaction of their products, or through the biological pathways they form part of [11, 12, 15, 41]. Gene sets also have to fulfil other requirements to be considered pathways or networks but they vary from case to case. However, even if the definitions of pathways and networks still seem very similar, usually pathways have stricter definitions and the extent of pathways is much smaller than the extent that networks can cover. In addition, networks require interaction between their components whereas pathways not necessarily.

The definition of gene set is the most difficult and ambiguous — but its definition

is key— because gene sets are the units of pathway and network analyses [42]. A group of genes is considered a gene set if they are somehow related. However, the genes will be considered to be related depending on the definition given by the tool selected to conduct the analysis, the gene set databases used, the individual performing the analyses, or the hypothesis being tested [12].

After defining and putting in context gene sets, pathways, and networks, we will explain what are pathway and network analyses. In addition, we will explain other biological contexts where these type of analyses can be applied to.

3.2 Pathway analysis

The aim of pathway analysis, if we apply it to the example given at the beginning of this section, would be to identify statistical significant gene sets (that can be defined as pathways) related to a phenotype, trait or disease of interest, from lists of ”unrelated” genes —but of interest based on previous experiments or studies—. At first, the genes are unrelated for us but after conducting a pathway analysis, we might discover that there are gene sets or pathways significantly related to our phenotype than other pathways or solitary genes.

Pathway analysis are also known as functional enrichment analysis because this type of analysis not only can be applied to gene sets–biological pathways relationships, but also to other relationships between gene sets and other gene–product properties of interest such as gene ontology (GO) terms [43]. If we take again as an example genes, but instead we are interested in identifying statistical significant gene sets that have common molecular functions, we could still apply pathway analysis. Nevertheless, in this case, instead of considering gene sets as pathways because they form part of the same biological pathway, we would consider gene sets as pathways because the genes of the gene sets share the same molecular functions.

Pathway analysis methods can be divided into two main groups. However, depending on the perspective from which the methods can be described, the name of the categories and their characteristics are different. Below I will enumerate three different ways for describing and classifying pathway analyses.

1. Based on the usage or non-usage of existing biological pathway knowledge, we can divide pathway analysis methods in topology-based tests and non-topology-based tests. The latter group of tests can also be can also be called gene set analysis methods [15, 43, 44, 45]. Topology-based methods take into account in the analysis prior information available of biological pathways such as positions and roles of genes, interactions of genes or gene products, direction of gene signaling, and other biological information known [15, 44, 45].

2. Based on the null hypothesis being tested, we can divide pathway analysis methods in competitive or enrichment tests, and self-contained or

association tests [11, 12, 13, 42]. In the competitive method, the null hypothesis states that the gene set considered pathway shows the same degree of association with the phenotype of study as other genes that do not form part of that pathway [11, 12, 13, 15]. Whereas the null hypothesis for the self-contained type states that there is no association between the gene set considered pathway and the phenotype of study [11, 12, 13, 15].

3. Based on the type of gene list to be analysed, pathway analysis methods can be divided into rank-based and non-rank-based tests [15, 43]. Non-rank-based tests take as input only a gene list of interest to be analysed. Whereas ranked tests take as input a list of genes and some quantitative data associated with each of the genes on the list. The quantitative variable permits to rank genes in terms of importance following certain chosen criteria [15].

There exist several statistical methods to conduct pathway analyses. Nevertheless, they usually fall in one of the two categories from the three points of view mentioned above. In the Methodology and Results sections we will explain the pathway analyses we have used with each software tool.

3.3 Network analysis

At the beginning of the section, we gave as an example that gene sets could be defined as networks. Not only groups of genes can be represented as networks but any biological system and other complex systems in which there are many entities interacting with each other can be represented as networks [41]. The aim of network analysis is to help to integrate great amounts of different information, facilitate interpretation of complex systems, and sometimes even predict how the different entities that form part of the network interact with each other [41, 46, 47].

All networks have nodes and edges but, depending on the field and what is being analysed, their visualisation and topology are different [41]. In other words, the representation and the properties of the network, substructures within a network or clusters of nodes highly inter-connected called modules, nodes, and edges vary from case to case [41]. However, in general, the main idea behind nodes is that they represent the entities that act or receive an action and edges represent interactions, overlaps, or relationships between entities [41, 47].

For example, in biology networks can represent several relationships between: genes, proteins, metabolites, biological pathways, and drugs [41, 47]. Nevertheless, networks can also represent interactions, signals or any other combination of relationships between the entities mentioned above [41, 47].

In our case, we will mainly work with interactions between genes, gene–product properties, and biological pathways. Moreover, the type of network visualisations we will use will be: undirected ball-and-stick diagrams and directed acyclic

graphs (DAG) [41]. Lastly, we will explain in more detail the network analyses we have used with each software tool in the Methodology and Results sections. There exist several network visualisations and methods to conduct network analyses [41] but with the software tools chosen and the time available for this project, we have only been able to explore two different graphs and three different network analyses. Therefore, we will only cover the methods and network visualisations used because the rest of existing approaches go beyond the scope of this project.

3.4 Previous studies

To the best of our knowledge, there are few comprehensive comparative studies of pathway and network analyses [13, 42, 44, 48, 49]. All of them agree that comparison of methods is complicated because there exist several different tools and methods to conduct pathway or network analyses. At least, more than 70 different methods might exist to date [44]. To give an idea of how dynamic the bioinformatics field is, out of the eight software tools we have chosen to explore in this project, only half of them have been mentioned in at least one of the comparative studies carried out until now.

Of all comparative studies found, some of them have used real data and others simulated data but, it seems that competitive methods of gene-set analysis [12, 42] and topology-based methods [48, 49] have a better performance than their opposites. Between rank-based [13, 48, 49] and non-rank-based methods [42] it is not clear which type of analysis performs better. However, all comparative studies are aware of the great amount of different factors that could have influenced their results and the limitations of their studies.

Future comparative studies will benefit from accessing new methods and tools with bigger databases. Nevertheless, it is quite likely that data integration and curation will still be an issue. Therefore, it is key that more comparative studies are carried out to confirm the results seen until now, to keep an updated picture of the state of the field, and to decipher how different and reliable are the results obtained when using different methods and programs.

4 Methodology

4.1 General approach

The main steps to conduct this project were quality control and data preprocessing, pathway analysis, and, in some cases, network analysis.

After the quality control and data preprocessing step, the workflow was to use as input either gene lists alone or gene lists together with some quantitative data

from three different datasets, analyse the data using different parameters or statistical analyses, and compare the results from a qualitatively point of view. The results could be information about potential biological pathways represented by genes present in our gene lists (pathway analysis), networks of biological pathways that overlap between them (network analysis), or both. These results were compared across different software tools and datasets. Table 1 shows information about the programs used, the datasets analysed with each of them, and the different methods used to analyse the data.

Depending on the information available on each dataset and on the program used to analyse the data, different variables were selected to conduct the pathway analyses. However, the main variable for this type of analysis was a gene identifier. For those cases where quantitative data were required, expression values, p-values, or adjusted p-values for multiple testing were used.

4.2 Data

4.2.1 Microarray data

Some of the data obtained in [18] was used in this project as a representative of microarray and transcriptomics data. They extracted RNA from the hippocampal CA1 gray matter of 30 postmortem human brain samples, used Affymetrix Human Genome U133 plus 2.0 arrays and HGU133 annotation data (October 2003) [18]. Out of the 30 samples, 11 were from males and 19 were from females. The average age of the individuals was 86.3 years and they were divided into four categories of different AD severity based on two AD marker scales: MiniMental Status Exam (MMSE) and neurofibrillary tangle (NFT) density. AD severity is negatively correlated with MMSE results and positively correlated with NFT counts. Seven individuals were classified as having severe AD, eight moderate AD, seven incipient AD, and eight as controls [18].

The data that we used for our analyses can be found in the Supplemental Table 1 of [18]. For all 3,465 genes whose expression profiles were significantly correlated with MMSE scores, NFT counts, or both across all 30 individuals, their respective gene symbol, gene description, probeset identifier, Pearson's correlation coefficients with MMSE scores and with NFT counts, and Pearson's tests p-values are provided in the dataset. Genes with low expression in AD were positively correlated with MMSE results and negatively correlated with NFT scores; whereas genes with high expression in AD were negatively correlated with MMSE results and positively correlated with NFT scores.

To conduct our analyses, we decided to use gene symbols as our main gene identifiers and analysed up-regulated and down-regulated genes separately. In addition, for those software tools that offered rank-based methods, we selected the p-values obtained in the Pearson's tests for MMSE results as our variable to rank the genes. Gene symbols were selected as our main gene identifiers

because all software tools chosen for this project gave the option to work with this type of identifier and to facilitate the comparison of results between software tools and datasets analysed. P-values obtained for the MMSE correlation coefficients were selected as our variable to rank the genes because the classification of patients into the four categories was mostly based on MMSE scores, and also because in [18] was found that MMSE correlation coefficients in comparison with NFT correlation coefficients explained a greater percentage of the variability of gene expression.

Nevertheless, when we used DAVID we worked with probeset identifiers instead because less than 80% of our gene symbols coincided with the gene symbols of their database. When we used IPA, we analysed all genes together, and we worked with probeset identifiers and gene symbols to increase the number of annotated genes to carry out our analysis.

4.2.2 RNA-seq data

In this project, some of the data in [19] were used as representatives of RNA-seq and transcriptomics data. In part of the study, transcriptomics data from the lateral temporal lobe of 30 postmortem human brain samples were analysed. 12 of the samples were from patients with AD, 10 from healthy older subjects, and eight were also from healthy individuals but younger. Most of the samples were from males and the mean age for each of the three groups was 68, 68 and 52 years, respectively.

To perform our pathway and network analysis, we used the results of the differential gene expression between the samples of patients with AD and old controls (q-value<0.05) of [19], data publicly available in GEO repository under accession identifier GSE153873. Nativio et al. [19] aligned the RNA-seq reads to the Genome Reference Consortium Human build 37.75 (GRCh 37.75), used RefSeq gene annotation, normalised the data using Evaluation of External RNA Controls Consortium (ERCC) spike-in control transcripts to account for global transcriptional changes happening between AD and old healthy individuals samples, and found a total of 855 differentially expressed genes. In the dataset, information about 421 up-regulated genes and 434 down-regulated genes in AD can be found separately. For each gene, their respective gene symbol, log_2 read count normalised per gene length and ERCC spike-in in AD samples and old controls, the log_2 fold change of AD samples with respect to old controls, p-value and false discovery rate (FDR) adjusted p-value obtained when identifying differentially expressed genes is provided.

4.2.3 Genome-Wide association study summary statistics data

The summary statistics of a genome-wide association meta-analysis of late-onset Alzheimer's disease (LOAD) conducted by The International Genomics of Alzheimer's Project (IGAP) group [20] was chosen as a representative of genomics data in this study. A sample of 94,437 subjects, among which 59,163 were controls and 35,274 were patients clinically diagnosed of LOAD, was used for the whole meta-analysis [20]. In [20], 20 loci already known to be associated with LOAD were confirmed and five new risk loci were identified. The chromosome and SNP positions were based on GRCh37, assembly hg19, and the variants annotation were based on RefSeq.

The summary statistics of [20] can be found in the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) under accession identifier NG00075. Their study was divided into different stages, but for our work only the results of the discovery stage were used. The results of stage 1 include imputed and genotyped data of 41,944 controls and 21,982 patients with LOAD. In the dataset, for each SNP, the chromosome and position where they are found, their reference SNP number (rsID), the risk allele, the non-effect allele, the effect size for the risk allele, the standard error and p-value for the effect size are provided.

We only analysed 10,534,426 SNPs out of 11,480,632 genetic variants because 946,206 SNPs did not have an rsID. Moreover, we analysed common and rare variants together because allele frequencies are not provided in the GWAS summary statistics due to data privacy. However, this should not be an inconvenience because Kunkle et al. [20] found that there was a positive and significant correlation between the gene association results from the pathway analysis using common and rare variants separately.

Table 1: Summary table with information about pathway or network methods used and datasets analysed with each software tool. All software tools are freely available, except IPA. The names of the pathway or network methods shown in the table are the same names the creators of the programs use to call or describe their respective analyses. The latest year of release makes reference to the program not to the latest year when their databases where updated.

Software tools (current version and/or latest year of release)	**Datasets analysed**	**Pathway (P) or network (N) analyses used**
ConsensusPathDB (v34, 2019) [21, 22]	Microarray, RNA-seq and Genome-wide association study summary statistics	Over-representation analysis[‡¶](P) and Wilcoxon enrichment analysis[†‡¶](P)
DAVID (v6.8, 2016) [23, 24]	Microarray, RNA-seq and Genome-wide association study summary statistics	Functional annotation analysis[‡¶](P)
EnrichmentMap (v3.3.2, 2021) [25]	Microarray, RNA-seq and Genome-wide association study summary statistics	Network-based visualisation of gene-set enrichment results (N)
g:Profiler (v.e103_eg50_p15_68c0e33, 2019) [27]	Microarray, RNA-seq and Genome-wide association study summary statistics	Over-representation analysis[‡¶](P) and Incremental enrichment analysis[†‡¶](P)
IPA[*] (2021) [28]	Microarray and RNA-seq	Core analysis[‡¶](P)(N)
MAGMA (v1.09a, 2021) [29]	Genome-wide association study summary statistics	Competitive gene-set analysis[‡¶](P)
Reactome (v76, 2021) [30, 31, 32, 33]	Microarray, RNA-seq and Genome-wide association study summary statistics	Over-representation analysis[‡¶](P)
WebGestalt (2019) [34, 35, 36, 37]	Microarray, RNA-seq and Genome-wide association study summary statistics	Over-representation analysis[‡¶](P), Gene-set enrichment analysis[†‡¶](P) and Network topology-based analysis (N)

[*]Proprietary software

[†]Ranked pathway analysis method

[‡]Competitive pathway analysis method

[¶]Non-topology-based pathway analysis method

4.3 Software tools

In this section information about each software tool, methods, and parameters used will be explained. In addition, specific steps done with each dataset to work with each software tool will be explained. The order in which programs are presented was selected to facilitate the understanding of part of our workflow. The chronological order in which the analyses were conducted can be seen in Figure 1.

Some of the parameters selected to analyse the data are selected by default by each software tool but if the users want to select a different value or option it is possible for them to change them. Those options or parameters chosen or modified by us —not kept as default—, will have an $^{(nd)}$ next to each of them. For the rest of options or parameters it can be assumed that were left as default.

4.3.1 IPA

Only the microarray and RNA-seq dataset could be analysed using IPA [28] with an evaluation license. During the trial period, we could only explore some of the settings and results due to the lack of funding and the restrictions related to the usage of an evaluation license, which we accepted in order to include IPA in our list of software tools for this project.

For the microarray dataset-related analyses, we submitted to IPA probeset identifiers, official gene symbols, genes' correlation with MMSE and their respective p-value, and genes' correlation with NFT and their respective p-value. For the RNA-seq dataset analysis, we submitted to IPA the official gene symbols, the differential gene expression between AD and old control samples, their respective p-value and adjusted p-value.

Due to the restrictions related to the usage of an evaluation license, incorrect selection of reference set or different selection of species for the analyses of the microarray and RNA-seq dataset, the results from IPA will not be included in the tables shown but, they will still be discussed. Not only the results obtained in pathway or network analysis are important but also the data and settings chosen to obtain results. The results obtained with IPA will help to emphasise the importance of selecting the appropriate settings in pathway and network analysis to avoid obtaining misleading results.

For the core analyses conducted using the RNA-seq data, the settings selected were:

- Reference set: Ingenuity knowledge base
- Relationships to consider: direct and indirect relationships.

- Interaction networks: 35 molecules per network and 25 networks per analysis.
- Species: Humans, mice, rat, and others.

Whereas the settings selected for the core analyses carried out using the microarray data were:

- Reference set: User dataset(nd).
- Relationships to consider: direct and indirect relationships.
- Interaction networks: 35 molecules per network and 25 networks per analysis.
- Species: Humans.
- Any cutoff to apply to the submitted dataset(nd): genes' correlation with MMSE p-value < 0.05.

4.3.2 MAGMA

Only the GWAS summary statistics dataset was analysed using MAGMA [29]. The first step consisted in mapping SNPs to genes (annotation step). Followed by an association analysis between genes and the phenotype of interest (gene analysis), and ended with a mapping from genes to gene-sets and an association analysis between gene-sets and the phenotype of interest (competitive gene-set analysis).

For the annotation step, a file with SNPs identifiers, their gene location (base pair position and chromosome), and the p-values for their effect sizes from the GWAS summary statistics dataset was submitted to MAGMA. In addition, gene locations file (human genome build 37) provided on CTGLab's (creators of MAGMA) website was also submitted and used for this step. Lastly, an annotation window(nd) of 35 kilobases (kb) upstream and 10 kb downstream around genes was used in the annotation step. The window setting was chosen based on one of the analysis settings selected and applied in [20].

For the gene analysis step, the file obtained as output from the annotation step and a file with SNP identifiers and their respective p-values for their effect sizes from the GWAS summary statistics dataset was submitted to MAGMA. In addition, a reference genome data file was also required for this step. MAGMA provides 1000 Genome data files from different populations; for our case, the European file was used as our reference genome data. The sample size used to obtain the statistics for the SNPs was also submitted to MAGMA because it was needed for the gene analysis (n=63,925). Lastly, the model selected for the gene analysis was the SNP-wise mean model.

A total of four gene-set analyses were carried out. Each of them required the output from the gene analysis step and a GMT file with information about genes and gene-sets from a database of interest. The GMT files used were obtained from the Molecular Signatures Database v7.4 [50, 51]. GMT files with gene-set information from KEGG [52, 53, 54], Reactome [30, 31, 32, 33], BioCarta [55], and GO [56, 57] databases were used for our gene-set analyses.

The gene list from the GWAS summary statistics dataset used for the analyses conducted with the rest of software tools is the one obtained as an output from the MAGMA's gene analysis. Nevertheless, before using it as input for the other analyses, the gene list was filtered. Only the genes with a p-value lower than 0.05 were kept and used for the rest of analyses.

4.3.3 g:Profiler

A total of five gene lists from the microarray, RNA-seq, and GWAS summary statistics datasets were analysed using g:Profiler [27]. Up-regulated and down-regulated genes from the microarray and RNA-seq datasets were analysed separately, and only one gene list from the GWAS summary statistics dataset was analysed. In all cases, the official gene symbol was used as gene identifier. g:Profiler offers two pathway analyses: Over-representation analysis (ORA) and Incremental enrichment analysis.

For the ORA the settings used were:

- Options: Homo sapiens in organism option
- Advanced options: All results$^{(nd)}$, no evidence codes$^{(nd)}$, only annotated genes in statistical domain scope, SCS threshold in significance threshold, user threshold 0.05, and numeric identifiers treated as Entrez gene identifiers.
- Data sources: GO molecular function terms, GO biological processes terms, GO cellular component terms, no electronic GO annotations$^{(nd)}$ [56, 57], KEGG [52, 53, 54], Reactome [30, 31, 32, 33], WikiPathways [58], CORUM [59], Human Protein Atlas (HPA) [60, 61, 62], TRANSFAC [63], miRTarBase [64], Human phenotype ontology (HP) [65].

For the Incremental enrichment analysis genes were ranked in descending order of significance. No quantitative data were uploaded to the program, only ranked gene lists were submitted to g:Profiler. To rank the genes from the microarray dataset, the quantitative variable used was the p-value for the genes' correlations with MMSE. For the genes from the RNA-seq dataset, the adjusted p-value for the differential expression of genes between AD and old control samples was used; and for the genes from the GWAS summary statistics dataset, the p-value from MAGMA's gene analysis. The settings used for this type of analysis were the same as for the ORA but including the ordered query in the options section$^{(nd)}$.

4.3.4 EnrichmentMap (Cytoscape Plugin)

Data from all three datasets were analysed using EnrichmentMap [25, 26]. The data submitted to EnricmentMap were results from g:Profiler analyses and data obtained directly from the datasets.

Over-representation analyses and Incremental enrichment analyses were rerun in g:Profiler but with the following settings:

- Options: Homo sapiens in organism option. Ordered query option(nd) was also selected but only for the Incremental enrichment analyses.
- Advanced options: All results(nd), no evidence codes(nd), only annotated genes in statistical domain scope, SCS threshold in significance threshold, user threshold 0.05, and numeric identifiers treated as Entrez gene identifiers.
- Data sources(nd): GO biological processes terms, no electronic GO annotations [56, 57], and Reactome [30, 31, 32, 33].

The results were downloaded using the Generic Enrichment Map format directly from g:Profiler. In addition, we downloaded from g:Profiler GMT files which had information about the gene sets used in GO and Reactome databases, and merged them in one file.

For all analyses, the GMT file with information about the gene sets used in GO and Reactome databases was submitted to EnrichmentMap. ORA and Incremental enrichment analysis results from g:Profiler were analysed separately in EnrichmentMap. However, for each type of analysis, up-regulated and down-regulated genes results were submitted to EnrichmentMap jointly. Lastly, for each dataset and type of analysis additional gene data were also submitted to EnrichmentMap. In all cases, the gene identifiers used were the official gene symbols.

For the microarray dataset-related analyses, the up-regulated and down-regulated gene lists ranked in descending order of significance together with their respective correlations with MMSE p-value were also submitted to EnrichmentMap. For the GWAS summary statistics dataset-related analyses, a gene list also ranked in descending order of significance together with their respective p-value from MAGMA's gene analysis was submitted to EnrichmentMap. Lastly, for the RNA-seq dataset-related analyses, the up-regulated and down-regulated gene lists ranked in descending order of significance together with their respective adjusted p-value for the differential expression of genes between AD and old control samples was used when the Incremental enrichment analysis results from g:Profiler were submitted to EnrichmentMap. Whereas when the results from the ORA results from g:Profiler were submitted to EnrichmentMap, the gene identifiers together with their

respective gene expression from up-regulated and down-regulated genes in AD and old control samples were also submitted to EnrichmentMap.

In EnrichmentMap the settings to create all networks were:

- Number of nodes[(nd)]: FDR q-value cutoff and p-value cutoff equal to one.
- Number of edges[(nd)]: Dataset edges automatic, overlap chosen as metric, overlap cutoff 0.5.

Once the networks were created, the FDR q-value cutoff was decreased and/or the overlap cutoff was increased until a few nodes and the edges that connected them could be visualised properly.

4.3.5 ConsensusPathDB

All three datasets were analysed using ConsensusPathDB [21, 22]. Among the methods that ConsensusPathDB offers, we used two out of the three gene set analyses they offer: ORA and Wilcoxon enrichment analysis. The third gene set analysis, called induced network modules, could not be used due to technical problems with their server.

For the ORA, two gene lists (up-regulated and down-regulated genes were analysed separately) from the microarray and RNA-seq datasets and one gene list from the GWAS summary statistics data set were analysed. Official gene symbols were used as gene identifier. The settings that could be chosen were related to: Network neighborhood-based entity sets, Pathway-based sets and GO categories, and Protein complex-based gene sets. For our analyses, we selected:

- Network neighborhood-based entity sets: 1-next neighbors set radius[(nd)], two minimum set size, zero minimum connectivity index, two minimum overlap with input list, and 0.05 p-value cutoff[(nd)].
- Pathway-based sets: pathways as defined by all pathway databases offered in ConsensusPathDB[*], two minimum overlap with input list, and 0.05 p-value cutoff[(nd)].

  [*]Name of the pathway databases offered in ConsensusPathDB: PharmGKB [66], HumanCyc [67], INOH [68], Reactome [30, 31, 32, 33], KEGG [52, 53, 54], Small Molecule Pathway Database (SMPDB) [69, 70], Edinburgh Human Metabolic Network (EHMN) [71], WikiPathways [58], NetPath [72], SignaLink [73], BioCarta [55], and Pathway Interaction Database (PID) [74]. However, only results from Reactome, KEGG and WikiPathways databases were selected for comparison with the results from the other programs.

- GO categories: GO level 2–5 for all GO categories (biological process, molecular function and cellular component)(nd) [56, 57], and 0.05 p-value cutoff(nd) for all GO levels.
- Protein complex-based gene sets: No settings were selected for this option(nd).

For the Wilcoxon enrichment analysis, only the RNA-seq dataset was analysed. Two gene lists (up-regulated and down-regulated genes were analysed separately and the official gene symbol was used as gene identifier) together with the gene expression levels for each gene in old control and AD samples were submitted. The other two datasets were not analysed using this method because gene expression levels for two different phenotypes were required and we did not have that information neither for the GWAS summary statistics dataset nor for the microarray dataset.

The settings chosen for the analysis of the RNA-seq dataset using the Wilcoxon enrichment analysis were:

- Network neighborhood-based entity sets: 1-next neighbors set radius(nd), four minimum set size, zero minimum connectivity index, four minimum overlap with input list, and 0.05 p-value cutoff(nd).
- Pathway-based sets: pathways as defined by all pathway databases offered in ConsensusPathDB*, four minimum overlap with input list, and 0.05 p-value cutoff(nd). *Name of the pathway databases offered in ConsensusPathDB: PharmGKB [66], HumanCyc [67], INOH [68], Reactome [30, 31, 32, 33], KEGG [52, 53, 54], SMPDB [69, 70], EHMN [71], WikiPathways [58], NetPath [72], SignaLink [73], BioCarta [55], and PID [74]. However, only results from Reactome, KEGG and WikiPathways databases were selected for comparison with the results from the other programs.
- GO categories: GO level 2–5 for all GO categories (biological process, molecular function and cellular component)(nd) [56, 57], and 0.05 p-value cutoff(nd) for all GO levels.
- Protein complex-based gene sets: No settings were selected for this option(nd).

4.3.6 DAVID

We used the Functional annotation analysis to analyse our three datasets in DAVID [23, 24]. We submitted two gene lists from the microarray and RNA-seq dataset and one gene list from the GWAS summary statistics dataset. For the RNA-seq and GWAS summary statistics dataset we used the official gene symbols as gene identifier; for the microarray dataset we used the probeset

identifier as gene identifier. For all cases we selected to submit our gene list as gene list (not as background or reference set) and selected *Homo sapiens*(nd) as species.

In the Functional annotation analysis, annotation categories and databases, from which to base your analysis on, can be chosen. The available categories in DAVID are: disease, functional categories, GO, general annotations, literature, main accessions, pathways, protein domains, protein interactions and tissue expression. For all our analyses we only selected the following categories, databases or options:

- Disease: OMIM [75].
- Functional categories(nd): up keywords.
- GO: direct GO terms for biological processes, direct GO terms for molecular functions, and direct GO terms for cellular components [56, 57].
- Pathways: BBID [76], BioCarta [55], KEGG(nd) [52, 53, 54], and Reactome [30, 31, 32, 33].
- Protein domains(nd): No options were selected for this category.

Among the tools available to use in the Functional annotation analysis, the Functional annotation clustering tool and the Functional annotation chart tool —which is an ORA method— were used. When the Functional annotation chart tool was selected, the thresholds chosen to obtain results were: EASE score equal to 0.1 and minimum gene number for each term (count) equal to two. When the Functional annotation clustering tool was selected, the settings chosen to obtain results were classification stringency medium and high(nd). The thresholds selected for the medium classification stringency were:

- Kappa similarity: similarity term overlap equal to three and similarity threshold equal to 0.50.
- Classification: initial and final group membership equal to three and multiple linkage threshold 0.50.
- Enrichment thresholds: EASE score 0.05(nd) and one, respectively.

The thresholds selected for the high classification stringency(nd) were the same as the ones selected for the medium classification stringency except:

- Kappa similarity: similarity threshold equal to 0.85.

4.3.7 Reactome

We used the Gene list analysis, which is an ORA, and project to human to analyse our three datasets in Reactome [30, 31, 32, 33]. We analysed two gene lists (up-regulated and down-regulated genes were analysed separately) from the microarray and RNA-seq datasets and one gene list from the GWAS summary statistics dataset. In all cases we used the official gene symbols as gene identifier.

For the microarray and RNA-seq dataset, we also submitted quantitative data. The quantitative variables uploaded were only used for visualisation purposes, they were not used in Reactome's statistical analyses. For the microarray dataset the quantitative data we uploaded were: the genes' correlation with MMSE and p-value, and genes' correlation with NFT scores and p-value. For the RNA-seq dataset we uploaded: the differential gene expression between AD and old control samples, the p-value, and adjusted p-value. The options selected were the genes' correlation with MMSE for the microarray dataset (Figures **??** and **??**) and the differential expression between AD and old control samples for the RNA-seq dataset (Figures **??** and **??**) to colour the Voronoi diagrams showed in the Results section.

4.3.8 WebGestalt

Gene lists from all three datasets were analysed using WebGestalt [34, 35, 36, 37]. Up-regulated and down-regulated genes from the microarray and RNA-seq datasets were analysed separately; and only one gene list from the GWAS summary statistics dataset was analysed. In all cases, the official gene symbol was used as gene identifier. The three type of analyses that WebGestalt offers are: ORA, Gene-set enrichment analysis (GSEA) and Network topology-based analysis (NTA).

For the ORA the settings used were:

- Reference gene list: genome-protein coding$^{(nd)}$.
- Functional databases$^{(nd)}$: GO cellular components no redundant terms, GO molecular functions no redundant terms, GO biological processes no redundant terms [56, 57], KEGG [52, 53, 54], and Reactome [30, 31, 32, 33], WikiPathways [58], and OMIM [75].
- Advanced parameters: minimum of genes per category equal to five, maximum of genes equal to 2000, Bonferroni-Hochberg multiple testing correction, significance level top 15 most significant results$^{(nd)}$(and to download all significant results we selected 0.05 FDR option), number of non-redundant sets expected from the weighted set cover algorithm equal to 10, number of categories visualised in the report 40$^{(nd)}$, and continuous colour for the DAGs.

For the NTA, the settings used were:

- Functional database(nd): Protein-protein interaction (PPI) BioGRID [77, 78].
- Advanced parameters: Network expansion used as network construction method(nd), number of highlighted seed genes equal to ten, significance level top ten most significant results, and seeds to be highlighted.

For the GSEA, apart from uploading gene lists from all three datasets, quantitative data were also uploaded from all datasets to rank genes in descending order of significance. For the microarray dataset, the quantitative variable used was the p-value for the genes' correlations with MMSE. For the RNA-seq dataset, the adjusted p.value for the differential expression of genes between AD and old control samples was used. Lastly, for the GWAS summary statistics dataset, the p-value from MAGMA's gene analysis.

The settings used for this type of analysis were:

- Functional databases(nd): GO cellular components no redundant terms, GO molecular functions no redundant terms, GO biological processes no redundant terms [56, 57], KEGG [52, 53, 54], and Reactome [30, 31, 32, 33]. However, for the GSEA, all databases were selected one by one. In other words, one analysis for each functional database selected.
- Advanced parameters: minimum of genes per category equal to five, maximum of genes equal to 2000, significance level top ten most significant results, number of permutations equal to 1000, exponential scaling factor in enrichment score equal to one, the mean will be used as collapsing method for dealing with duplicated identifiers, number of non-redundant sets expected from the weighted set cover algorithm equal to ten, number of categories visualised in the report 40(nd), and continuous colour for the DAGs.

5 Results

For the pathway and network analyses, we mainly worked with gene lists from the microarray, RNA-seq, and GWAS summary statistics datasets. Based on the official gene symbols provided in each dataset, 98 genes from the RNA-seq dataset (total number of genes in the dataset 855) coincide with the genes found in the microarray dataset (total number of genes in the dataset 3,645). However, when using IPA, we found that a total of 120 genes match between the genes from the microarray and RNA-seq datasets (Figure 2). Among them, several genes that were found to be positively correlated with MMSE in [18] (shown in green colour in Figure 2) are located in the nucleus, and several genes that were found to be negatively correlated with MMSE (shown in red colour in Figure 2)

are located in the cytoplasm. In addition, a great proportion of genes positively and negatively correlated with MMSE seem to be located somewhere else other than in the nucleus, cytoplasm, plasma membrane or extracellular space.

After mapping the SNPs found in the GWAS summary statistics dataset to genes using MAGMA (total number of genes annotated 1,763), a total of 60 genes between the RNA-seq dataset and the GWAS summary statistics dataset overlap. Whereas a total of 214 genes between the microarray dataset and GWAS summary statistics match.

The longest gene list is the one from the microarray dataset, followed by the one obtained from the GWAS summary statistics and finally by the one from the RNA-seq dataset. The greatest number of gene matches between datasets' gene lists coincides to be between the two longest gene lists analysed, the gene lists from the microarray and the GWAS summary statistics dataset. The difference in length of the gene lists submitted to conduct our pathway analyses will be important to take into account for the discussion of the results.

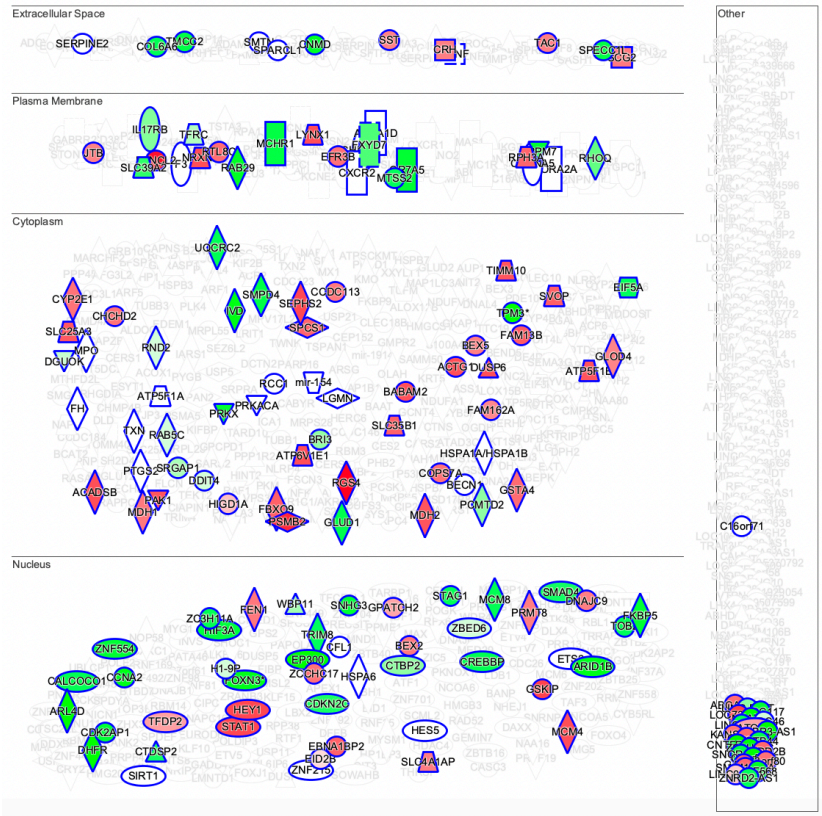


Figure 2: Overlap and cell location of gene–product molecules found between genes from the microarray and RNA-seq datasets when using IPA. The colours represent the correlation of genes with MiniMental Status Exam; where red indicates a negative correlation and green a positive correlation. Each type of molecules is represented by different shapes. Rhombi represent enzymes and peptidases; squares, cytokines and growth factors; ovals, transcription regulators and transmembrane receptors; rectangles, ion channels, ligand-dependent nuclear receptors, and G-protein coupled receptors; the rest of shapes represent transporters or other type of molecules. In the background, the rest of gene–product molecules found in the RNA-seq dataset are also shown in light grey colour.

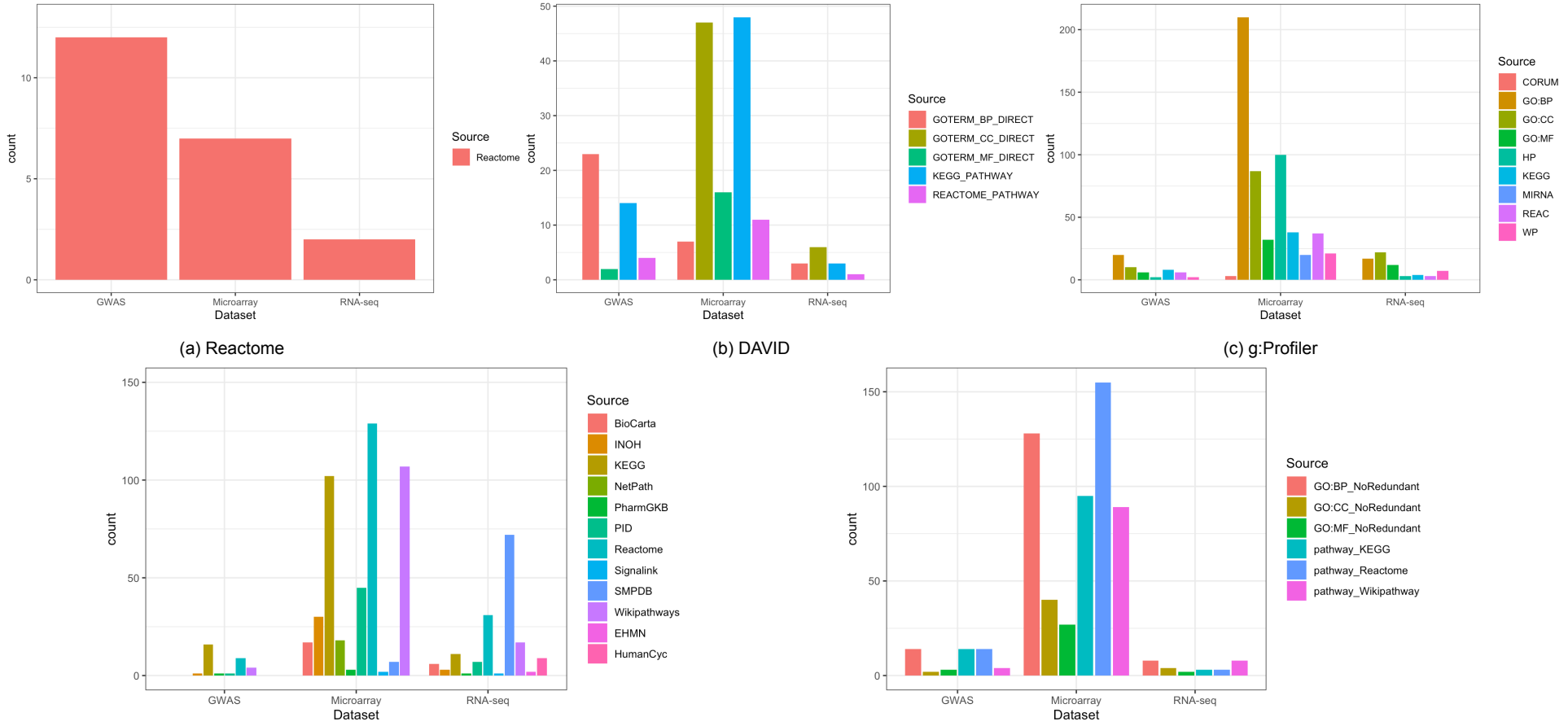
5.1 Pathway analysis

5.1.1 Overview

All pathway analysis methods used for this project are competitive (or enrichment) and non-topology-based tests. In addition, the majority of them are non-ranked methods. Only three of the pathway analysis methods used are ranked methods. In Figure 3 can be found some of the results obtained when using ORA methods —which are competitive, non-topology-based, and non-ranked pathway analysis methods. Each of the subfigures from Figure 3 show the number of statistically significant results obtained (after correcting for multiple testing) for each of the datasets analysed when using different software tools. Moreover, each subfigure has its own legend and scale. The different colours of the subfigures represent the databases from which each of the respective tools relies on to obtain biological information for the ORA.

In the Methodology section we mentioned all the databases selected from each software tool to base our analyses on. Figure 3d does not show the ConsensusPathDB's results based on GO terms and Figure 3c does not show g:Profiler's results based on TF and HPA sources because the number of some of the results obtained from those sources were too large in comparison with the rest to be shown in the same figure. The number of statistically significant results for the GO terms obtained when using ConsensusPathDB were: GO for biological processes 1495 (microarray dataset), 98 (RNA-seq dataset), and 107 (GWAS dataset); GO for molecular functions 239 (microarray dataset), 44 (RNA-seq dataset), and 24 (GWAS dataset); GO for cellular components 361 (microarray dataset), 60 (RNA-seq dataset), and 39 (GWAS dataset). The number of statistically significant results obtained when using g:Profiler based on the TF source were: 576 (microarray dataset), nine (RNA-seq dataset) and one (GWAS dataset). The results obtained based on HPA: 279 (microarray dataset), and six (RNA-seq dataset). Figure 3b does not show the results obtained from up keywords source because that option was only of interest for the Functional annotation clustering tool of DAVID, but not for the ORA.

In general, from Figure 3 it can be observed that more statistically significant results were obtained from the microarray dataset, independently from the software tool used, except when using Reactome tool. Moreover, the number of statistically significant results obtained from the GWAS summary statistics dataset and RNA-seq dataset are more similar than when comparing them with the results from the microarray dataset. However, the number of statistically significant results either between datasets when using the same tool or between tools when analysing the same dataset based on the same source, notably differ.



(a) Reactome

(b) DAVID

(c) g:Profiler

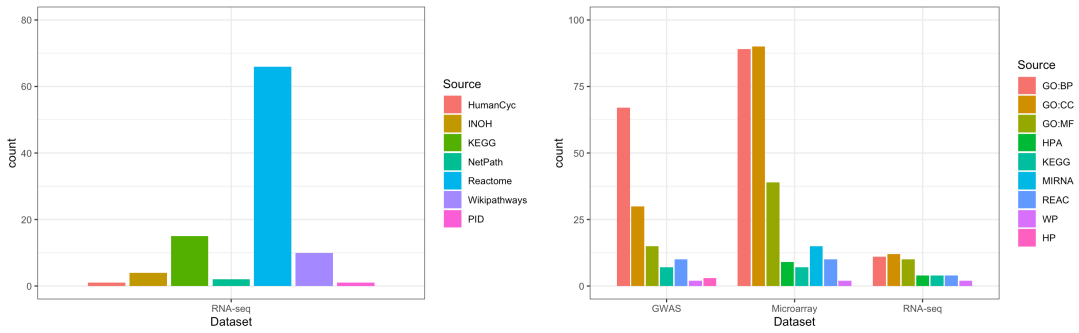
(d) ConsensusPathDB

(e) WebGestalt

Figure 3: Bar plots showing the number of statistically significant results obtained (results corrected for multiple testing) after analysing all three datasets using over-representation analyses in Reactome, DAVID, g:Profiler, ConsensusPathDB, and WebGestalt. Each subfigure has its own legend and scale. In all software tools, except when using g:Profiler, the method of Benjamini-Hochberg was used to control for multiple testing and a false discovery rate 5% threshold was selected. When using g:Profiler, their SCS method was chosen to control for multiple testing and a 0.05 threshold was selected. REAC stands for Reactome, WP for WikiPathways, GO:BP for GO terms for biological processes, GO:MF for GO terms for molecular functions, and GO:CC for GO terms for cellular components.

The number of statistically significant (p-value < 0.05) results obtained when analysing the GWAS dataset with MAGMA were: 13 (source BioCarta), 656 (source GO terms), seven (source KEGG), and 74 (source Reactome). The results obtained with MAGMA are not shown in Figure 3 because MAGMA does not use an ORA and test statistics such as: the hypergeometric test (ConsensusPathDB, g:Profiler, Reactome, WebGestalt), and Fisher's Exact test or adaptations of it (DAVID, IPA, WebGestalt). MAGMA also uses a competitive analysis but instead utilises a linear regression approach on genes taking into account for example gene density, gene size, and gene-gene correlations. MAGMA also offers a self-contained analysis. Nevertheless, the latter option was not used for this project.

The number of statistically significant results (after correcting for multiple testing) when using rank-based methods are shown in Figure 4. ConsensusPathDB, g:Profiler and WebGestalt are the only tools, among the programs chosen for this project, that offered rank-based pathway analysis methods. The results obtained when using WebGestalt are not shown in Figure 4 because only three results passed the significance level chosen (FDR 5%): one GO for cellular components (microarray dataset), one for GO molecular functions (RNA-seq dataset) and one KEGG pathway (RNA-seq dataset). In the case of WebGestalt, the number of statistically significant results obtained was much lower when we used the GSEA method than when we used the ORA.



(a) ConsensusPathDB (b) g:Profiler

Figure 4: Bar plots showing the number of statistically significant results obtained (results corrected for multiple testing) after analysing the respective datasets using rank-based methods in ConsensusPathDB and g:Profiler. Each subfigure has its own legend and scale. When using ConsensusPathDB, the method of Benjamini-Hochberg was used to control for multiple testing and a false discovery rate 5% threshold was selected. When using g:Profiler, their SCS algorithm was chosen to control for multiple testing and a 0.05 threshold was selected. REAC stands for Reactome, WP for WikiPathways, GO:BP for GO terms for biological processes, GO:MF for GO terms for molecular functions, and GO:CC for GO terms for cellular components.

If we compare the number of statistically significant results shown in Figure 3d and in Figure 4a when analysing the RNA-seq dataset, we can observe that the number of results substantially vary. The results obtained for the GO terms are not shown in Figure 4a because in comparison to the rest of results obtained, they were too large to be shown in the same figure. The results obtained for the GO terms were: GO for biological processes 1335, GO for molecular functions 215, and GO for cellular components 269. The results obtained for the GO terms also noticeably differ in number when using the ORA or the Wilcoxon enrichment

analysis.

In the case of g:Profiler, the number of statistically significant results obtained also differs when using ORA or the Incremental enrichment analysis, specially for GO terms. In Figure 4b, the results for the Incremental enrichment analysis from g:Profiler are shown. The results based on TF source are not shown in Figure 4b because of their difference in magnitude in comparison to the rest of results shown in the figure. The number of statistically significant results obtained based on TF source were: 531 (microarray dataset), three (RNA-seq dataset), and five (GWAS summary statistics).

5.1.2 A closer look at the most relevant results

In Tables 3–7, we show the most relevant results obtained (after correcting the results for multiple testing) with all software tools, except with IPA and MAGMA, when using ORA. The results are shown sorted by tool, database and in descending order of level of significance. The ratios shown in the column of Entities found refers to the number of entities matched between the gene list submitted and the database knowledge with respect to the total number of entities known to be part of the respective biological pathway (database knowledge). Tables 2 and 3 show the results obtained when analysing the microarray dataset. Whereas, Tables 4 and 5 show the results obtained when analysing the RNA-seq dataset, and Tables 6 and 7 the results obtained when analysing the GWAS summary statistics dataset.

In Tables 2 and 3, there are several results that coincide between software tools. However, the results for the up-regulated genes vary more from one tool to the other. Whereas, the results for the down-regulated genes are more consistent between programs, specially when basing the results on KEGG's database (even the ranking based on FDR or SCS coincides). In general, the results are more similar between tools when the same database is selected but, some results based on KEGG's and Reactome's or on KEGG's and WikiPathways' databases also coincide despite the number of entities found and ratio vary.

Some interesting results are the results obtained when the up-regulated genes from the microarray dataset were analysed using Reactome's tool (Table 2), and the rest of tools based on Reactome's database (Tables 2 and 3). Reactome's tool results do not pass the threshold for significance; whereas the same results obtained when using a different software tool based on Reactome's database or not pass the significance threshold.

Table 2: Most relevant pathways found in up-regulated (above double line) and down-regulated (below double line) genes from the microarray dataset using over-representation analysis from Reactome and g:Profiler. The thresholds chosen for multiple testing methods were FDR 5% and SCS experiment-wide α=0.05.

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
Cohesin Loading onto Chromatin	7 / 10	FDR: 0.852	R-R
FOXO-mediated transcription	30 / 110	FDR: 0.852	R-R
Molecules associated with elastic fibres	14 / 38	FDR: 0.852	R-R
FOXO-mediated transcription	19 / 66	SCS: 1.61×10^{-4}	g:P-R
Gene expression (Transcription)	160 / 1435	SCS: 1.91×10^{-4}	g:P-R
RNA Polymerase II Transcription	145 / 1301	SCS: 7.93×10^{-4}	g:P-R
Molecules associated with elastic fibres	13 / 37	SCS: 9.90×10^{-4}	g:P-R
FoxO signaling pathway	27 / 131	SCS: 1.53×10^{-4}	g:P-K
Pathways in cancer	68 / 529	SCS: 9.24×10^{-4}	g:P-K
Cellular senescence	27 / 156	SCS: 4.54×10^{-3}	g:P-K
Focal adhesion	31 / 200	SCS: 1.12×10^{-2}	g:P-K
VEGFA-VEGFR2 Signaling Pathway	61 / 437	SCS: 7.84×10^{-3}	g:P-W
Sarcoma	32 / 165	SCS: 9.38×10^{-3}	g:P-H
Neurotransmitter receptors and postsynaptic signal transmission	53 / 231	FDR: 5.00×10^{-3}	R-R
Unblocking of NMDA receptors, glutamate binding and activation	14 / 27	FDR: 5.00×10^{-3}	R-R
Protein-protein interactions at synapses	28 / 93	FDR: 6.00×10^{-3}	R-R
Neuronal System	91 / 487	FDR: 6.00×10^{-3}	R-R
Activation of NMDA receptors and postsynaptic events	31 / 113	FDR: 7.00×10^{-3}	R-R
Transmission across Chemical Synapses	68 / 341	FDR: 8.00×10^{-3}	R-R
Neuronal System	71 / 400	SCS: 3.51×10^{-13}	g:P-R
Transmission across Chemical Synapses	52 / 259	SCS: 2.14×10^{-11}	g:P-R
Neurotransmitter receptors and postsynaptic signal transmission	39 / 196	SCS: 4.74×10^{-8}	g:P-R
Protein-protein interactions at synapses	24 / 86	SCS: 1.60×10^{-7}	g:P-R
Glutamatergic synapse	28 / 114	SCS: 1.24×10^{-8}	g:P-K
Dopaminergic synapse	29 / 131	SCS: 8.26×10^{-8}	g:P-K
Retrograde endocannabinoid signaling	30 / 148	SCS: 3.93×10^{-7}	g:P-K
Long-term potentiation	19 / 67	SCS: 1.01×10^{-6}	g:P-K
Disruption of postsynaptic signalling by CNV	13 / 34	SCS: 1.09×10^{-5}	g:P-W
Brain-Derived Neurotrophic Factor signaling pathway	28 / 144	SCS: 1.13×10^{-5}	g:P-W
Epileptic encephalopathy	31 / 100	SCS: 2.54×10^{-10}	g:P-H

adj., adjusted; DB, database; FDR, false discovery rate; g:P, g:Profiler; K, KEGG; R, Reactome; W, WikiPathways; H, Human Phenotype ontology; SCS, g:Profiler algorithm to correct for multiple testing.

Table 3: Most relevant pathways found in up-regulated (above double line) and down-regulated (below double line) genes from the microarray dataset using over-representation analysis from ConsensusPathDB, WebGestalt, and DAVID. The threshold chosen for multiple testing correction was FDR 5%.

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
Molecules associated with elastic fibres	12 / 31	FDR: $5.53x10^{-4}$	CPDB-R
Elastic fibre formation	12 / 36	FDR: $1.75x10^{-3}$	CPDB-R
Gene expression (Transcription)	13 / 38	FDR: $5.03x10^{-3}$	CPDB-R
FoxO signaling pathway	28 / 132	FDR: $1.68x10^{-4}$	CPDB-K
Pathways in cancer	65 / 526	FDR: $3.23x10^{-3}$	CPDB-K
Cellular senescence	28 / 132	FDR: $4.46x10^{-3}$	CPDB-K
Type 2 papillary renal cell carcinoma	11 / 34	FDR: $3.66x10^{-3}$	CPDB-W
TGF-beta Signaling Pathway	24 / 132	FDR: $4.46x10^{-3}$	CPDB-W
Gene expression (Transcription)	151 / 1429	FDR: $8.12x10^{-5}$	WG-R
RNA Polymerase II Transcription	135 / 1292	FDR: $3.31x10^{-4}$	WG-R
Molecules associated with elastic fibres	13 / 38	FDR: $3.57x10^{-4}$	WG-R
FoxO signaling pathway	28 / 132	FDR: $8.12x10^{-5}$	WG-K
VEGFA-VEGFR2 Signaling Pathway	60 / 431	FDR: $1.10x10^{-4}$	WG-W
PPARA activates gene expression	25 / 113	FDR: 0.028	D-R
Cohesin Loading onto Chromatin	7 / 10	FDR: 0.028	D-R
TGF-beta signaling pathway	21 / 84	FDR: $2.87x10^{-3}$	D-K
FoxO signaling pathway	28 / 134	FDR: $2.87x10^{-3}$	D-K
Neuronal System	68 / 368	FDR: $5.07x10^{-6}$	CPDB-R
Transmission across Chemical Synapses	49 / 224	FDR: $8.67x10^{-6}$	CPDB-R
Protein-protein interactions at synapses	32 / 88	FDR: $1.27x10^{-5}$	CPDB-R
FoxO signaling pathway	42 / 132	FDR: $8.67x10^{-6}$	CPDB-K
Dopaminergic synapse	29 / 131	FDR: $1.20x10^{-4}$	CPDB-K
Amyotrophic lateral sclerosis (ALS)	21 / 51	FDR: $1.54x10^{-4}$	CPDB-K
Insulin Signaling	48 / 160	FDR: $8.67x10^{-6}$	CPDB-W
Neuronal System	70 / 368	FDR: 0	WG-R
Transmission across Chemical Synapses	52 / 227	FDR: 0	WG-R
Neurotransmitter receptors and postsynaptic signal transmission	36 / 156	FDR: $3.65x10^{-10}$	WG-R
Glutamergic synapse	27 / 114	FDR: $4.54x10^{-8}$	WG-K
Dopamine Neurotransmitter Release Cycle	10 / 23	FDR: 0.023	D-R
Glutamatergic synapse	29 / 114	FDR: $2.37x10^{-6}$	D-K
Retrograde endocannabinoid signaling	27 /101	FDR: $2.37x10^{-6}$	D-K
Long-term potentiation	21 / 66	FDR: $3.57x10^{-6}$	D-K

adj., adjusted; DB, database; FDR, false discovery rate; CPDB, ConsensusPathDB; D, DAVID; K, KEGG; R, Reactome; W, WikiPathways; H, Human Phenotype ontology; WG, WebGestalt.

In Tables 4 and 5, also several results coincide between software tools either when the same database was selected or not. However, those results that match, either when comparing tools or databases, only in some cases pass the significance threshold.

Gene expression (Transcription), RNA Polymerase II Transcription, and FOXO-related pathways appear among the most relevant pathways found when analysing up-regulated genes from the microarray and RNA-seq datasets (Tables 2–5). Of those, the results that do not reach statistical significance are the ones obtained when using Reactome's tool (Tables 2 and 4).

In Tables 6 and 7, also several results for the GWAS summary statistics dataset coincide between programs either when the same database was selected or not. All results shown in Tables 6 and 7 pass the significance threshold. Among the statistically significant results found for the GWAS dataset, none of them coincide with the results found for the microarray or the RNAdataset.

Among the statistically significant results obtained when we used rank-based methods, some results coincide with those we obtained when we analysed the microarray, RNA-seq and GWAS summary statistics dataset using ORA. In the case of the microarray dataset, the results that match between the results shown in Tables 2 and 3, and the results obtained when using the Incremental enrichment analysis method from g:Profiler are: FOXO-mediated transcription, Glutamergic synapse, Retrograde endocannabinoid signaling, Transmission across Chemical Synapses, Neuronal System, Protein-protein interactions at synapses, Disruption of postynaptic signalling by CNV, and Unblocking of NMDA receptors, glutamate binding and activation. When we used GSEA method from WebGestalt, none of the results obtained reached statistical significance.

When we analysed the RNA-seq dataset using the Incremental enrichment analysis method from g:Profiler, the results that reached statistical significance and match with those shown in Tables 4 and 5 are: Cell cycle, Regulation of FOXO transcriptional activity by acetilation, RNA Polymerase II Transcription, TCA cycle, and Carbon metabolism. When we used the Wilcoxon enrichment analysis from ConsensusPathDB, the results that reached statistical significance and that coincide with the results shown in Tables 4 and 5 are: TCA cycle and respiratory electron transport, Metabolism, Amino Acid metabolism, Cell cycle, Mitotic G1-G1/S phases, HSF1-dependent transactivation, Gene expression (Transcription), RNA Polymerase II Transcription, Attenuation phase, S Phase, Selenium Micronutrient Network, and RUNX2 regulates chondrocyte maturation. Lastly, when using GSEA method from WebGestalt, only one result reached statistical significance (without counting results obtained based on GO terms). The result, Transcriptional misregulation in cancer, was obtained when selecting KEGG as database. However, it does not coincide with any of the results shown in Tables 4 and 5.

Table 4: Most relevant pathways found in up-regulated (above double line) and down-regulated (below double line) genes from the RNA-seq dataset using over-representation analysis from Reactome, g:Profiler, and DAVID. The thresholds chosen for multiple testing methods were FDR 5% and SCS experiment-wide α=0.05.

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
G0 and Early G1	6 / 38	FDR: 0.483	R-R
RUNX2 regulates chondrocyte maturation	3 / 7	FDR: 0.483	R-R
Regulation of FOXO transcriptional activity by acetilation	4 / 16	FDR: 0.483	R-R
Regulation of FOXO transcriptional activity by acetylation	4 / 10	SCS: 4.44×10^{-3}	g:P-R
RNA Polymerase II Transcription	38 / 1301	SCS: 0.031	g:P-R
Gene expression (Transcription)	40 / 1435	SCS: 0.053	g:P-R
Cell cycle	9 / 124	SCS: 9.86×10^{-3}	g:P-K
Cell cycle	9 /124	FDR: 0.075	D-K
HTLV-I infection	11 / 254	FDR: 0.333	D-K
MECP2 regulates transcription of neuronal ligands	6 / 13	FDR: 0.009	R-R
Attenuation phase	9 / 47	FDR: 0.030	R-R
HSF1-dependent transactivation	9 / 47	FDR: 0.106	R-R
Mitochondrial protein import	9 / 69	FDR: 0.175	R-R
HSF1 activation	7 / 43	FDR: 0.175	R-R
Citric acid cycle (TCA cycle)	7 / 50	FDR: 0.339	R-R
Citric acid cycle (TCA cycle)	6 / 22	SCS: 2.41×10^{-3}	g:P-R
MECP2 regulates transcription of neuronal ligands	3 / 7	SCS: 0.148	g:P-R
Carbon metabolism	14 / 116	SCS: 2.39×10^{-5}	g:P-K
Citrate cycle (TCA cycle)	7 / 30	SCS: 3.09×10^{-4}	g:P-K
Metabolic pathways	56 / 1490	SCS: 2.13×10^{-3}	g:P-K
Amino Acid metabolism	12 / 91	SCS: 5.92×10^{-5}	g:P-W
Citric acid cycle (TCA cycle)	6 / 18	SCS: 1.80×10^{-4}	g:P-W
Selenium Micronutrient Network	9 / 91	SCS: 0.018	g:P-W
Abnormality of acid-base homeostasis	23 / 363	SCS: 0.012	g:P-H
Increased serum lactate	15 / 172	SCS: 0.014	g:P-H
Citric acid cycle (TCA cycle)	6 / 19	FDR: 0.017	D-R
Mitochondrial protein import	7 / 54	FDR: 0.221	D-R
Carbon metabolism	13 / 113	FDR: 3.00×10^{-3}	D-K
Citrate cycle (TCA cycle)	7 / 30	FDR: 5.73×10^{-3}	D-K
Metabolic pathways	50 / 1219	FDR: 5.73×10^{-3}	D-K

adj., adjusted; DB, database; FDR, false discovery rate; g:P, g:Profiler; K, KEGG; R, Reactome; W, WikiPathways; D, DAVID; H, Human Phenotype ontology; SCS, g:Profiler algorithm to correct for multiple testing.

Table 5: Most relevant pathways found in up-regulated (above double line) and down-regulated (below double line) genes from the RNA-seq dataset using over-representation analysis from ConsensusPathDB and WebGestalt. The threshold chosen for multiple testing correction was FDR 5%.

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
Mitotic G1-G1/S phases	8 / 104	FDR: 0.021	CPDB-R
HSF1-dependent transactivation	5 / 36	FDR: 0.021	CPDB-R
Gene expression (Transcription)	36 / 1373	FDR: 0.024	CPDB-R
RNA Polymerase II Transcription	33 / 1236	FDR: 0.027	CPDB-R
Cell cycle	9 / 124	FDR: 0.019	CPDB-K
Initiation of transcription and translation elongation at the HIV-1 LTR	4 / 32	FDR: 0.038	CPDB-W
Cell Cycle	7 / 120	FDR: 0.044	CPDB-W
Regulation of gene expression by Hypoxia-inducible factor	3 / 11	FDR: 0.430	WG-R
G0 and Early G1	4 / 27	FDR: 0.430	WG-R
S Phase	9 / 161	FDR: 0.490	WG-R
RUNX2 regulates bone development	4 / 32	FDR: 0.490	WG-R
Cell cycle	9 / 124	FDR: 0.242	WG-K
Integrated Breast Cancer Pathway	9 / 152	FDR: 0.430	WG-W
Initiation of transcription and translation elongation at the HIV-1LTR	4 / 32	FDR: 0.490	WG-W
Citric acid cycle (TCA cycle)	6 / 22	FDR: 3.00×10^{-4}	CPDB-R
The citric acid (TCA) cycle and respiratory electron transport	12 / 173	FDR: 5.52×10^{-3}	CPDB-R
Metabolism	60 / 1972	FDR: 6.26×10^{-3}	CPDB-R
Citrate cycle (TCA cycle)	7 / 30	FDR: 2.09×10^{-4}	CPDB-K
Pyruvate metabolism	5 / 39	FDR: 0.026	CPDB-K
Citric acid cycle (TCA cycle)	6 / 17	FDR: 2.09×10^{-4}	CPDB-W
Amino Acid metabolism	11 / 91	FDR: 2.09×10^{-4}	CPDB-W
Selenium Micronutrient Network	9 / 83	FDR: 1.76×10^{-3}	CPDB-W
Citric acid cycle (TCA cycle)	6 / 22	FDR: 1.29×10^{-3}	WG-R
Carbon metabolism	13 / 116	FDR: 3.17×10^{-4}	WG-K
Citrate cycle (TCA cycle)	7 / 30	FDR: 6.68×10^{-4}	WG-K
Metabolic pathways	48 / 1305	FDR: 4.32×10^{-3}	WG-K
Amino Acid metabolism	13 / 91	FDR: 6.89×10^{-5}	WG-W
Citric acid cycle (TCA cycle)	6 / 18	FDR: 4.41×10^{-4}	WG-W
Urea cycle and associated pathways	5 / 21	FDR: 0.011	WG-W

adj., adjusted; DB, database; FDR, false discovery rate; CPDB, ConsensusPathDB; K, KEGG; R, Reactome; W, WikiPathways; WG, WebGestalt.

Table 6: Most relevant pathways found in genes from the genome-wide association study summary statistics dataset using over-representation analysis from Reactome, g:Profiler and DAVID. The thresholds chosen for multiple testing methods were FDR 5% and SCS experiment-wide α=0.05

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
Interferon gamma signaling	85 / 250	FDR: 7.20×10^{-10}	R-R
Interferon Signaling	114 / 394	FDR: 1.26×10^{-9}	R-R
Interferon alpha/beta signaling	61 / 188	FDR: 2.65×10^{-6}	R-R
Endosomal/Vacuolar pathway	35 / 82	FDR: 7.60×10^{-6}	R-R
Antigen Presentation: Folding, assembly and peptide loading of class I MHC	38 / 102	FDR: 4.74×10^{-5}	R-R
TRAF6 mediated IRF7 activation	21 / 43	FDR: 4.51×10^{-4}	R-R
Interferon Signaling	39 / 191	SCS: 4.64×10^{-4}	g:P-R
Interferon alpha/beta signaling	17 / 66	SCS: 2.89×10^{-2}	g:P-R
Regulation of IFNA signaling	9 / 22	SCS: 3.46×10^{-2}	g:P-R
TRAF6 mediated IRF7 activation	10 / 27	SCS: 3.71×10^{-2}	g:P-R
Chylomicron remodeling	6 / 10	SCS: 4.38×10^{-2}	g:P-R
Plasma lipoprotein assembly	8 / 18	SCS: 4.41×10^{-2}	g:P-R
Epstein-Barr virus infection	37 / 198	SCS: 4.87×10^{-4}	g:P-K
Autoimmune thyroid disease	15 / 49	SCS: 9.93×10^{-4}	g:P-K
Tuberculosis	32 / 175	SCS: 3.41×10^{-3}	g:P-K
Influenza A	31 / 169	SCS: 4.25×10^{-3}	g:P-K
Hematopoietic cell lineage	21 / 95	SCS: 4.75×10^{-3}	g:P-K
Cholesterol metabolism	13 / 50	SCS: 2.54×10^{-2}	g:P-K
Statin Pathway	13 / 33	SCS: 3.01×10^{-4}	g:P-W
SARS coronavirus and innate immunity	10 / 31	SCS: 3.56×10^{-2}	g:P-W
Frontotemporal dementia	12 / 26	SCS: 1.25×10^{-3}	g:P-H
Xanthomatosis	11 / 27	SCS: 1.61×10^{-2}	g:P-H
Regulation of IFNA signaling	11 / 26	FDR: 0.029	D-R
Interferon alpha/beta signaling	18 / 67	FDR: 0.029	D-R
Autoimmune thyroid disease	18 / 52	FDR: 2.22×10^{-4}	D-K
Herpes simplex infection	35 / 183	FDR: 1.77×10^{-3}	D-K
Tuberculosis	33 / 177	FDR: 3.64×10^{-3}	D-K
Measles	27 / 133	FDR: 3.64×10^{-3}	D-K

adj., adjusted; DB, database; FDR, false discovery rate; g:P, g:Profiler; K, KEGG; R, Reactome; W, WikiPathways; D, DAVID; H, Human Phenotype ontology; SCS, g:Profiler algorithm to correct for multiple testing.

Table 7: Most relevant pathways found in genes from the genome-wide association study summary statistics dataset using over-representation analysis from ConsensusPathDB and WebGestalt. The threshold chosen for multiple testing correction was FDR 5%.

Pathway name	**Entities found**	**adj. p-value**	**Tool-DB**
Interferon Signaling	38 / 158	FDR: 5.40×10^{-6}	CPDB-R
Interferon alpha/beta signaling	19 / 70	FDR: 8.97×10^{-4}	CPDB-R
Regulation of IFNA signaling	11 / 26	FDR: 8.97×10^{-4}	CPDB-R
Chylomicron remodeling	6 / 10	FDR: 6.73×10^{-3}	CPDB-R
Plasma lipoprotein assembly	8 / 19	FDR: 9.40×10^{-3}	CPDB-R
Interferon gamma signaling	20 / 94	FDR: 1.02×10^{-2}	CPDB-R
Epstein-Barr virus infection	41 / 201	FDR: 5.82×10^{-5}	CPDB-K
Autoimmune thyroid disease	18 / 53	FDR: 6.51×10^{-5}	CPDB-K
Herpes simplex infection	36 / 185	FDR: 1.07×10^{-3}	CPDB-K
Tuberculosis	34 / 179	FDR: 1.07×10^{-3}	CPDB-K
Measles	27 / 132	FDR: 1.97×10^{-3}	CPDB-K
Toll-like receptor signaling pathway	23 / 104	FDR: 2.27×10^{-3}	CPDB-K
Factors involved in megakaryocyte development and platelet production	10 / 18	FDR: 1.59×10^{-4}	CPDB-W
Statin Pathway	13 / 31	FDR: 1.68×10^{-4}	CPDB-W
Toll-like Receptor Signaling Pathway	23 / 102	FDR: 1.97×10^{-3}	CPDB-W
Regulation of Toll-like receptor Signaling Pathway	28 / 143	FDR: 3.03×10^{-3}	CPDB-W
Interferon Signaling	44 / 197	FDR: 2.01×10^{-5}	WG-R
TRAF6 mediated IRF7 activation	12 / 29	FDR: 1.95×10^{-3}	WG-R
Interferon alpha/beta signaling	19 / 69	FDR: 2.21×10^{-3}	WG-R
Regulation of IFNA signaling	11 / 26	FDR: 2.21×10^{-3}	WG-R
Epstein-Barr virus infection	41 / 201	FDR: 2.99×10^{-4}	WG-K
Autoimmune thyroid disease	18 / 53	FDR: 2.99×10^{-4}	WG-K
Herpes simplex infection	35 / 185	FDR: 3.45×10^{-3}	WG-K
Tuberculosis	34 / 179	FDR: 3.84×10^{-3}	WG-K
Statin Pathway	41 / 201	FDR: 2.99×10^{-4}	WG-W

adj., adjusted; DB, database; FDR, false discovery rate; CPDB, ConsensusPathDB; K, KEGG; R, Reactome; W, WikiPathways; WG, WebGestalt.

Some of the results obtained when analysing the GWAS summary statistics dataset coincide when using the ORA and the Incremental enrichment analysis from g:Profiler. The results that match between those shown in Tables 6 and 7 and the statistically significant results obtained with g:Profiler are: Autoimmune thyroid disease, Cholesterol metabolism, Chylomicron remodeling, Hematopoietic cell lineage, Plasma lipoprotein assembly, and Statin Pathway. When using GSEA from WebGestalt, no results (without counting GO terms) reached statistical significance.

Lastly, the most statistically significant results (p-value<0.05) obtained when using the competitive gene-set analysis from MAGMA and Reactome as source are: Reelin Signaling Pathway, VLDL assemblance, VLDL clearance, RHOG GTPase cycle, and Diseases associated with O-glycosylation of proteins. The results when KEGG was selected as a source: Autoimmune thyroid disease, GnRH Signaling Pathway, Base excision repair, Regulation of autophagy, and Pyrimidine metabolism. If we compare the results shown in Tables 6 and 7 with the results obtained with MAGMA, only Autoimmune thyroid disease coincides. Among the results that reached statistical significance when we analysed the GWAS summary statistics dataset and used a rank-based test (Incremental enrichment analysis from g:Profiler), the results that coincide with those obtained with MAGMA are: Autoimmune thyroid disease, VLDL clearance, VLDL assembly, and NR1H2 and NR1H3-mediated signaling.

5.2 Network analysis

Network analyses were conducted using EnrichmentMap, IPA, and WebGestalt. IPA and WebGestalt, do not need the results from another software tool to perform their network analyses. Whereas EnrichmentMap requires data from other programs which can perform pathway analyses to conduct its network analysis.

In our case, we used the results from g:Profiler based on GO terms for biological processes and Reactome's database to create the networks with EnrichmentMap. Figures 5–10 show the networks created with EnrichmentMap using overlap as metric. Figures 11–13 show DAGs created with the NTA method from WebGestalt. The results obtained with IPA will be presented and discussed in the Discussion section.

5.2.1 EnrichmentMap

We tried to maintain the same cutoffs for the nodes and edges for all networks to facilitate comparison. Nevertheless, for visualisation and readability purposes, depending on the number of results obtained when creating the networks, we selected a different FDR (q-value) cutoff for the nodes and/or a different overlap cutoff for the edges.

One big network was created using the ORA results from g:Profiler, a $1x10^{-7}$ q-value threshold for the nodes, and a 0.8 overlap threshold for the edges (Figure 5). Metabolic-related processes or pathways predominate among the results that passed the q-value threshold in up-regulated genes from the microarray dataset. Whereas, pathways or processes related to neuronal development and synaptic signaling predominate among the results that reached the q-value threshold in down-regulated genes from the microarray dataset. From the results observed, Regulation of biological process and Regulation of cellular process seem to be the cores of the network that connect metabolic-related pathways or processes with neuronal development-related and synaptic signaling-related processes or pathways.

Two small networks and one big network were created using the Incremental enrichment results from g:Profiler, a $1x10^{-4}$ q-value threshold for the nodes, and a 0.6 overlap threshold for the edges (Figure 6). Neuronal development-related and synaptic signaling-related processes or pathways are also present in these networks. In comparison to Figure 5, results related to metabolism do not appear but other processes related to cations and transmembrane transport appear. In addition, all nodes and edges shown passed the q-value and edges cutoffs only in down-regulated genes from the microarray dataset.

Figures 7 and 8 share pathways or processes related to the citric acid cycle, cellular respiration, and metabolism among the results obtained from g:Profiler when down-regulated genes from the RNA-seq dataset were analysed using ORA and Incremental enrichment analysis, respectively. In addition, the same overlap edge cutoff (0.6) could be selected to connect the nodes of the networks. In Figure 8 other nodes appear among the results for the down-regulated genes from the RNA-seq dataset. Some are connected between them and others appear isolated from the rest. Nevertheless, Figures 7 and 8 mostly differ by the results obtained when up-regulated genes were analysed in g:Profiler; the only result that coincides is the regulation of FOXO transcriptional activity by acetylation.

Positive regulation of peptidyl-serine phosphorylation of STAT protein and protein-lipid-related pathways or processes are results that Figures 9 and 10 have in common. Moreover, the same overlap cutoff, with value equal to 0.6, could be selected to create the edges of the networks. In Figure 9 other pathways or processes related to intestinal lipid absorption and STAT protein appear among the results for genes from the GWAS summary statistics dataset. Whereas in Figure 10 a relatively big network formed with pathways or processes related to amyloid-beta peptide appears.

Figure 5: Network created with EnrichmentMap using overlap as metric after analysing the microarray data using the Over-representation analysis from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold $1x10^{-7}$ (q-value) for the nodes and 0.8 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the purple edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The blue arrows next to some nodes indicate that the results of the up-regulated genes for those biological pathways or biological processes passed the q-value cutoff selected. The colour of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (ochre) to 1 (white). The right half of the nodes represents the q-value obtained in the analysis of the up-regulated genes and the left half of the nodes represents the q-value obtained in the analysis of the down-regulated genes. If one of the halves of the node is painted with grey, it means that the respective biological pathway or biological process was not among the results obtained when either the up-regulated or down-regulated genes were analysed in g:Profiler.

Figure 6: Network created with EnrichmentMap using overlap as metric after analysing the microarray data using the Incremental enrichment method from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold $1x10^{-4}$ (q-value) for the nodes and 0.6 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the purple edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The colour of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (ochre) to 1 (white). The right half of the nodes represents the q-value obtained in the analysis of the up-regulated genes and the left half of the nodes represents the q-value obtained in the analysis of the down-regulated genes. If one of the halves of the node is painted with grey, it means that the respective biological pathway or biological process was not among the results obtained when either the up-regulated or down-regulated genes were analysed in g:Profiler.

Figure 7: Network created with EnrichmentMap using overlap as metric after analysing the RNA-seq data using the Over-representation analysis from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold 0.05 (q-value) for the nodes and 0.6 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the blue edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The blue arrows next to some nodes indicate that the results of the up-regulated genes for those biological pathways or biological processes passed the q-value cutoff selected. The colour of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (orange) to 1 (white). The right half of the nodes represents the q-value obtained in the analysis of the up-regulated genes and the left half of the nodes represents the q-value obtained in the analysis of the down-regulated genes. If one of the halves of the node is painted with grey, it means that the respective biological pathway or biological process was not among the results obtained when either the up-regulated or down-regulated genes were analysed in g:Profiler.

Figure 8: Network created with EnrichmentMap using overlap as metric after analysing the RNA-seq data using the Incremental enrichment method from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold 0.2 (q-value) for the nodes and 0.6 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the blue edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The blue arrows next to some nodes indicate that the results of the up-regulated genes for those biological pathways or biological processes passed the q-value cutoff selected. The colour of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (orange) to 1 (white). The right half of the nodes represents the q-value obtained in the analysis of the up-regulated genes and the left half of the nodes represents the q-value obtained in the analysis of the down-regulated genes. If one of the halves of the node is painted with grey, it means that the respective biological pathway or biological process was not among the results obtained when either the up-regulated or down-regulated genes were analysed in g:Profiler.

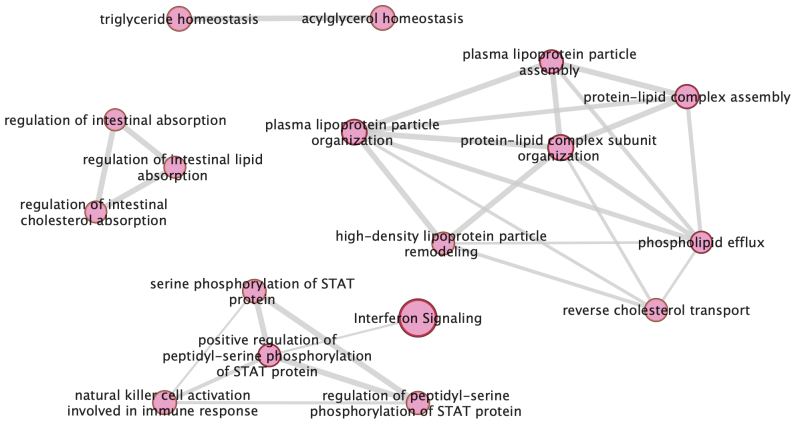


Figure 9: Network created with EnrichmentMap using overlap as metric after analysing the genome-wide association study summary statistics data using the Over-representation analysis method from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold 0.15 (q-value) for the nodes and 0.6 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the grey edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The colour of the outer part of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (red) passing through mid values (orange) to 1 (yellow).

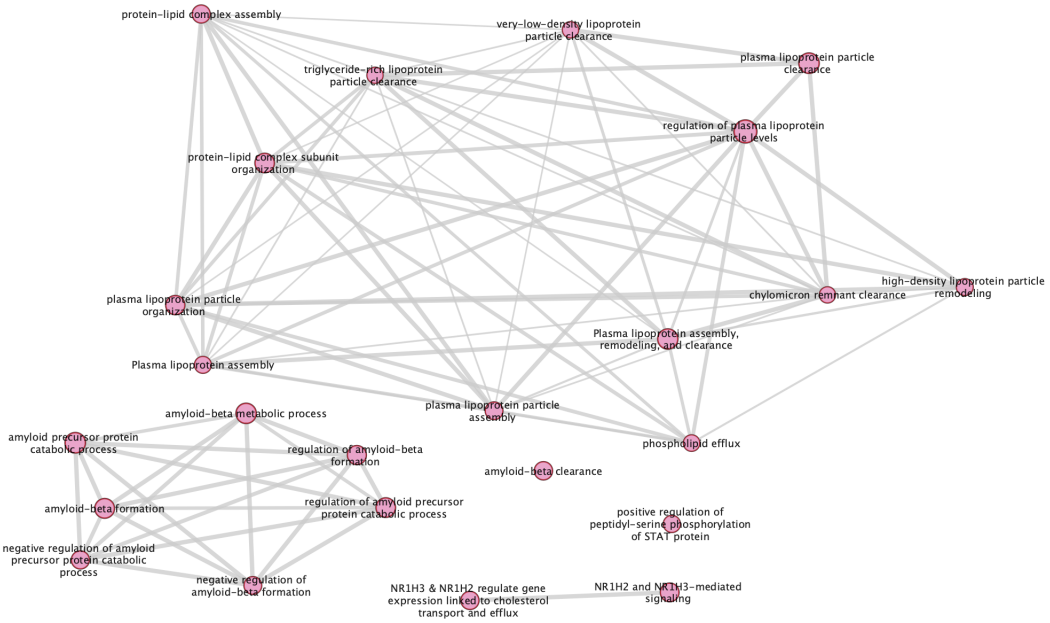


Figure 10: Network created with EnrichmentMap using overlap as metric after analysing the genome-wide association study summary statistics data using the Incremental enrichment method from g:Profiler. The cutoffs used to obtain the network shown were: false discovery rate threshold $2x10^{-3}$ (q-value) for the nodes and 0.6 overlap for the edges. The size of nodes represents the number of genes known to be part of that biological pathway or biological process; the bigger the node, the more genes are known to form part of that biological pathway or biological process. The width of the grey edges represents the number of genes that two nodes share; the thicker the line, the more genes the two nodes have in common. The colour of the outer part of the nodes represents the q-value of the biological pathway or biological process; the scale used goes from zero (red) passing through mid values (orange) to 1 (yellow).

5.2.2 WebGestalt

Figures 11–13 show DAGs with enriched GO terms for biological processes in the sub-network created after using the Network expansion method from WebGestalt NTA. Network expansion method and NTA from WebGestalt [34, 35, 36, 37] use random walk analysis to create a sub-network based on the gene–product knowledge of a selected source and the gene list submitted. In our case, we selected to create our sub-network based on PPI from BioGRID. The output of the algorithm of WebGestalt will give a sub-network created with some genes from our gene list and the most probable gene neighbours that are connected to those genes from our gene list.

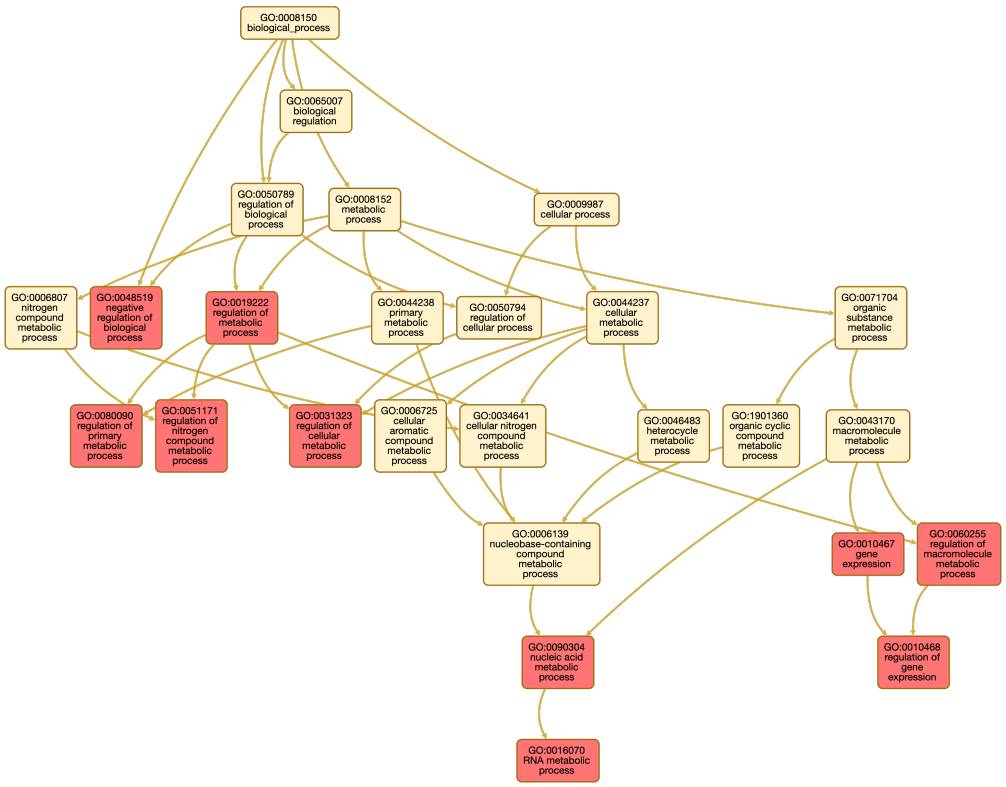
In the case of the GWAS summary statistics dataset, the results we obtained using the NTA from WebGestalt partially coincide with the results obtained using EnrichmentMap (Figures 9 and 10). Biological processes related to immunological responses predominate in the results shown in Figure 13. Whereas in Figures 9 and 10, apart from showing some pathways or processes related to immunological responses, also protein-lipid, amyloid-beta, and intestinal lipid absorption-related results appear.

For the microarray and RNA-seq datasets, in general, the results observed in Figures 11 and 12 (results obtained using the NTA from WebGestalt) coincide with those observed in Figures 5–8 (results observed in the networks created with EnrichmentMap when analysing the microarray or RNA-seq dataset).

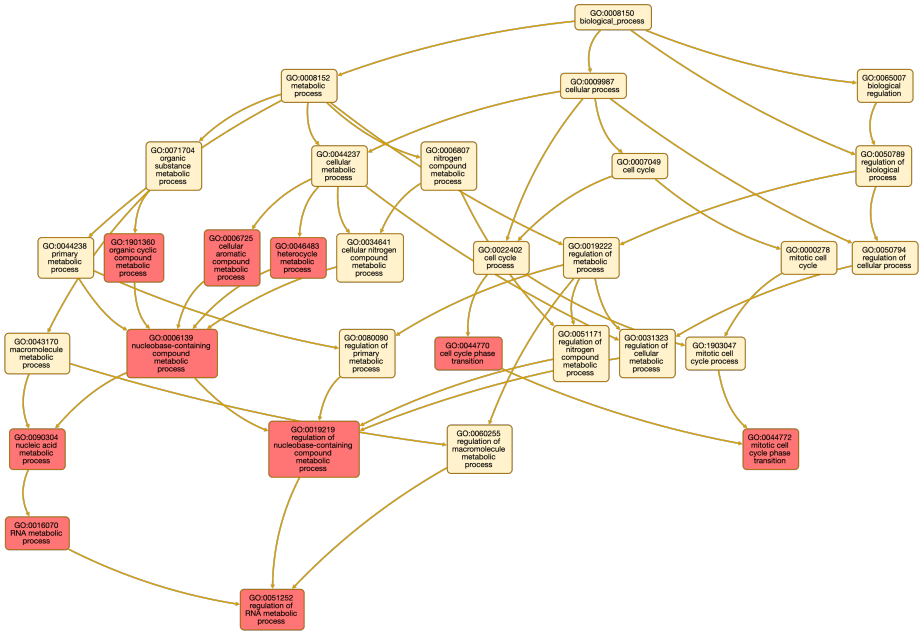
(a) Microarray dataset

(b) RNA-seq dataset

Figure 11: Directed acyclic graph showing the ten most relevant gene ontology biological processes in the sub-network created from the down-regulated genes from the microarray and RNA-seq datasets, respectively, using the Network topology analysis and Network expansion method from WebGestalt. The gene ontology biological processes (GO:BP) shown in yellow show other GO:BP related to any of the enriched GO:BP found in the sub-network.



(a) Microarray dataset



(b) RNA-seq dataset

Figure 12: Directed acyclic graph showing the ten most relevant gene ontology biological processes in the sub-network created from the up-regulated genes from the microarray and RNA-seq datasets, respectively, using the Network topology analysis and Network expansion method from WebGestalt. The gene ontology biological processes (GO:BP) shown in yellow show other GO:BP related to any of the enriched GO:BP found in the sub-network.

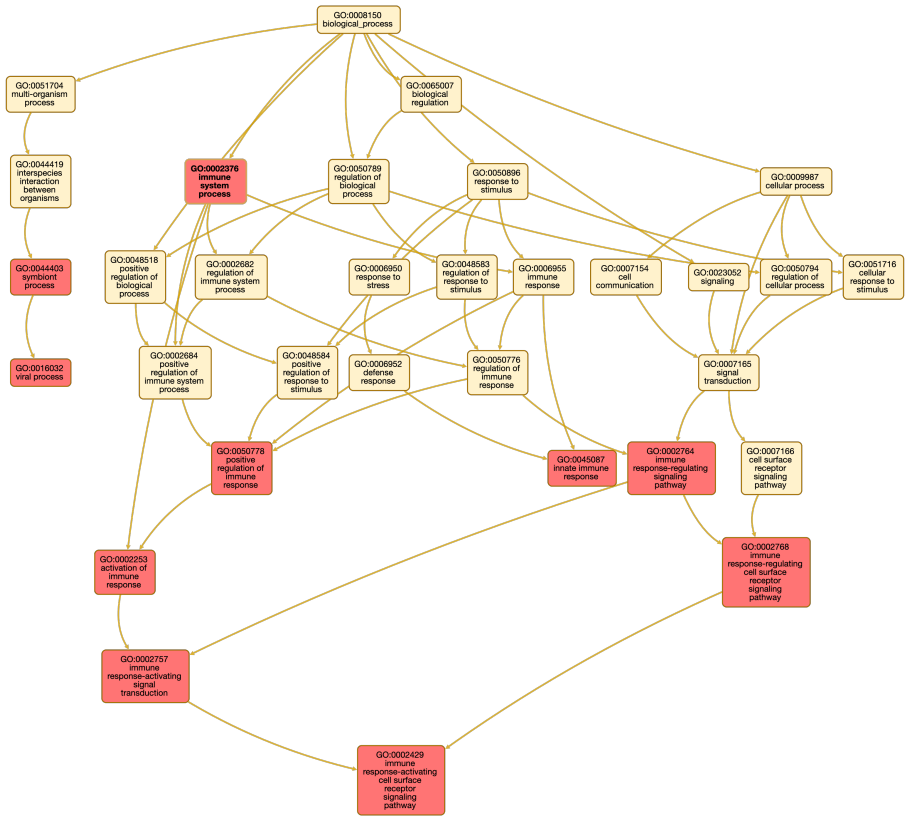


Figure 13: Directed acyclic graph showing the ten most relevant gene ontology biological processes in the sub-network created from the genes from the genome-wide association study summary statistics dataset using the Network topology analysis and Network expansion method from WebGestalt. The gene ontology biological processes (GO:BP) shown in yellow, show other GO:BP related to the enriched GO:BP found in the sub-network shown in red.

6 Discussion

We found that the analysis of different types of biological data helped us to have a better systemic understanding of AD because each type of biological data provides us with different knowledge related to AD. Pathway and network analysis together with the analysis of different types of biological data gave us a functional and systemic insight into AD pathogenesis. Our results from pathway analysis and network analysis are complimentary. Pathway analysis results gave us a list of gene sets, which can be considered pathways, and that are more related to AD than other genes that do not form part of those pathways; and network analysis results extended our pathway analysis results providing us with knowledge on how those pathways are connected between them and what enriched pathways are found when taking into account the neighbours from the genes that form part of those pathways. The results from the microarray dataset showed us that synaptic signalling, neuronal development, and transmembrane transport pathways or processes are enriched in down-regulated genes in AD, and that nucleic acid metabolic processes or pathways are enriched in up-regulated genes in AD. The results from the RNA-seq dataset showed us that pathways or processes related to the citric acid cycle, cellular respiration, and metabolism are enriched in down-regulated genes in AD, and that transcription,

cell cycle, and metabolism of nucleic acids are enriched in up-regulated genes in AD. Lastly, the results from the GWAS summary statistics dataset showed us that pathways related to intestinal lipid absorption, amyloid-beta peptide, protein-lipid processes, and immunological responses are enriched in AD.

We obtained a different number of statistically significant results when analysing the same data and using the same pathway analysis method but with different software tools. Moreover, we found that if we had selected the same database to base our analyses on while carrying out the same pathway analysis method but with different tools, in general, the most relevant results obtained coincided between tools but, with different levels of significance. In some cases, some of the most relevant results obtained with ORA were also present in rank-based pathway analysis methods. Nevertheless, we do not have enough evidence to confirm which type of pathway analysis performs better because we could conduct rank-based analysis with only three software tools and one of them could only be applied to the RNA-seq dataset. But the trend seems to indicate that fewer significant results are obtained when rank-based methods are used.

There are several factors that could have had an effect on the different number of significant results obtained with each software tool, apart from the usage of different pathway analysis methods and the type of data supported by each program. One factor is the usage of different versions of the databases the programs rely on for the definition of biological pathways, GO categories, diseases and other terms. A second factor, is the identifier match between programs, databases, and the data submitted. Related to the latter, despite submitting the same data to each program, the number of entities used for the analyses can differ not only because some entities could not be found but also because each program deals differently with entities that are duplicated or entities that have ambiguous annotations. A third factor is the number of parameters and what parameters users can change to perform their pathway analysis. Lastly, a fourth factor is the usage of different multiple testing methods, none, or the usage of algorithms to reduce biological redundancy within the results.

Our results shown in Figures 5, 6 and 11a confirm part of the GO terms for biological processes seen in [18] when down-regulated genes were analysed. However, our GO terms for biological processes do not coincide with the results observed in [18] when we analysed up-regulated genes. In addition, our results shown in Figure 7, 11b and 12b confirm the GO terms for biological processes obtained with DAVID in [19]. These findings also contribute to confirm that when comparing the most relevant results obtained with different tools but using the same type of pathway analysis method and datatabase, overall the results coincide. Lastly, most of our results shown in Figure 10 confirm the GO terms for biological processes obtained with MAGMA in [20]. Nevertheless, the results we obtained with MAGMA do not coincide with the results observed in [20]. This could be due to two reasons related to the methodology we applied. We did not restrict the sizes of the gene sets for our pathway analysis nor used adaptive permutation in the gene analysis step. In addition, with respect to the year when [20] was published, we believe that we have used a different version of MAGMA.

In the current version (v1.09a), an update of the SNP-wise mean model (used for the gene analysis step) has been made, which could have had an effect and made our results differ from the results observed in [20].

While carrying out our pathway analyses we discovered that some software tools give you the option to provide your own reference set or use the one provided by each of them. For pathway analysis, the selection of the proper reference set is essential to obtain non-misleading results. The first software tool that we used for our project was IPA and because of our lack of experience in conducting pathway analyses and misinterpretation of the documentation, we selected the incorrect reference set option for the analysis of the microarray dataset. The selection of the reference set depends on the experiment that it has been carried out. For example, if we have used a microarray chip with probesets only related to cancer and we have all the results of the microarray, then we can upload all the results to the program and select that we are also providing the reference set (with probesets only related to cancer) for our pathway analysis. In this example, if we select the knowledge of the database as the reference set, the results that we will obtain will be biased enriched in cancer pathways. At the beginning of the project, we did not understand in which pathway analysis cases we had to select the usage of the knowledge database or our own.

For the microarray dataset analysis, we selected that we were providing our own reference set for our pathway analysis, instead of selecting the IPA knowledge as reference set because we were providing a regular gene list of interest but not a whole reference set with some potential interesting genes in it. Despite of this, we obtained significant results without adjusting the p-values for multiple testing. For this reason, we decided to discuss the results obtained with IPA to show the importance of the correct selection of reference sets in pathway analyses. The three most significant pathway results (p-value<0.05) were: cAMP-mediated signaling, synaptic long term potentiation, and melatonin signaling. In the analysis of the RNA-seq data we selected the reference set of IPA but we based our results on different species to start our analysis as general as possible. The three most significant results (p-value<0.05) were: TCA cycle (eukaryotic), isoleucin degradation I, and valine degradation I.

Once submitted the data to IPA, IPA carries out automatically a pathway and network analysis. The biological functions of the two networks with the highest scores found for the microarray dataset were: (i) gene expression, RNA damage and repair, and RNA-post transcriptional modification, and (ii) infectious diseases, neurological disease, and organismal injury and abnormalities. The results for the RNA-seq dataset were: (i) amino acid metabolism, small molecule biochemistry and cellular assembly organisation, and (ii) protein synthesis, amino acid metabolism, and cell cycle. Some of the results have appeared among the results obtained with the rest of software tools. However, because of the restrictions imposed while using an evaluation license of IPA we could not repeat the analyses to compare results and their significance level when choosing proper settings or similar settings to the rest of analyses carried out. Therefore, the results obtained with IPA should be taken with caution and as an example of what can happen if incorrect settings are selected.

For the comparison of pathway analysis results between tools, we expected heterogeneity in the results. Nevertheless, the difficulty to compare results between programs from the syntactic, semantic, and biological point of view was higher than expected. Moreover, biological redundancy of GO terms and biological pathways also contributed to increase the difficulty in the comparison of results between tools. Each software tool provided the results with different data structures and different information. In addition, the description of the pathways differed between databases.

The methods of the network analyses we selected to carry out in this project are not comparable. In general, the approach and the results provided from WebGestalt seem to be mirroring the approach of EnrichmentMap. EnrichmentMap starts the analysis from pathway analysis results and finishes creating networks and sub-networks with those results. Whereas, WebGestalt starts the analysis creating a sub-network from a network and ends giving, for example, pathways that are enriched in the sub-network. In addition, WebGestalt uses protein-protein interaction information to conduct its NTA; EnrichmentMap does not use more biological information to conduct the network analysis. Lastly, the final number of network analyses performed is too small to conclude something else apart from the matching of some of our results with the results seen in [18, 19, 20].

7 Conclusions

Pathway and network analysis together with the analysis of different types of biological data gave us a global and functional understanding of AD pathogenesis from lists of genes which seemed to be unrelated. The results from our pathway and network analysis coincide with those obtained by Blalock et al. [18], Nativio et al. [19] and Kunkle et al. [20]. Pathways related to synaptic signalling, neuronal development, and transmembrane transport were found in the analysis of the microarray dataset. Pathways related to citric acid cycle, cellular respiration, nucleic acid metabolism, and cell cycle were found in the analysis of the RNA-seq dataset, and pathways related to amyloid-beta peptide, intestinal lipid absorption, immunological responses, and protein-lipid processes were found in the analysis of the GWAS summary statistics dataset. All of those processes are among the results that are enriched in AD.

Our results show that a different number of statistically significant results is obtained when using different software tools, despite analysing the same data and using the same pathway analysis method. Nevertheless, in general, the most relevant pathways coincide between tools when the same database and pathway analysis method are selected, but not with the same level of significance.

We need to conduct more network analyses to properly compare their methods and results. With the current results, we only have evidence to conclude that

most of the results obtained coincide with the results seen in [18, 19, 20].

7.1 Limitations

One of the limitations of this work is that we could not include in the analysis more software tools and other AD data from which patient information could be identified due to the short period of time available, data privacy, or lack of funding. Due to the lack of funding, we had strict usage permission of proprietary software to analyse the data selected. Therefore, we could not analyse the GWAS summary statistics dataset with IPA nor repeat the analyses for the microarray and RNA-seq datasets selecting the appropriate settings for our pathway and network analyses in the end.

One limitation of the analysis of the GWAS summary statistics dataset is that we had to rely on the mapping step and gene analysis of MAGMA. The usage of MAGMA was essential to obtain a list of genes (potentially related with AD) which we could use for the rest of pathway and network analyses we wanted to conduct with the other software tools selected. With the data available, MAGMA was one of the software tools available that fulfilled our needs because it could not only convert a list of SNPs into genes and perform a gene analysis but also it could perform a gene-set analysis. Nevertheless, the gene list used for our pathway and network analyses could have been different if we had chosen a different program.

Out of the five network analysis methods planned to be conducted, only two could be carried out successfully. The network analysis from IPA could be performed but we chose the incorrect settings for the analysis. The Induced network modules from ConsensusPathDB could not be conducted due to specific problems with this method and its server. WGCNA's workflow could only be applied to the microarray dataset but it required several modifications to analyse the data with the tool. Some important steps of the analysis with WGCNA had to be skipped due to the data available to conduct the analysis. As a result of all the modifications done to the general workflow and the outcomes obtained, we concluded that the results were not reliable. Therefore we decided to exclude WGCNA from the project.

Related to our pathway analyses, the data available, and the time available to carry out the project, several analyses could not be explored; many parameters were left as default. Consequently, the effect of some of the parameters on the results is unknown. In addition, for the pathway analyses, only two different multiple testing correction methods were used: FDR and SCS (g:Profiler's algorithm). More analyses using SCS algorithm are needed to confirm the level of significance of the results obtained.

A last constraint is that pathway annotation databases evolve rapidly and have different update frequencies. Consequently, the results that we showed correspond to the time frame when the project has taken place and they are

subject to change.

7.2 Future perspectives

Future comparative studies should explore the usage of text-clustering algorithms to reduce biological redundancy and group similar results together — from the biological, syntactic and semantic point of view. The usage of text-clustering algorithms might help to have a better control of the number of coincident results between databases and tools. The comparative study that finds a proper method to compare pathway results will have the opportunity to create a database and unify pathway information from different sources using key words and unique identifiers.

More software tools that provide topology-based and rank-based methods should be included in other comparative studies. Due to data availability and time restrictions, we could not explore and test more network analysis methods nor topology-based and rank-based pathway analysis methods. In addition, some of our analyses should be repeated using different parameters to verify their effect on pathway and network analysis results.

Different multiple testing correction methods in pathway and network analysis should be explored. Specially, methods such as the algorithm of g:Profiler, SCS, which take into account the structure behind gene set definitions because biological entities are not entirely independent from each other.

We would like to suggest to pathway analysis users that they use more than one gene identifier for their gene lists of interest; and we would like to recommend to current or future tool creators of pathway analysis methods that they consider allowing users to upload more than one gene identifier for the same gene list. Using more than one gene identifier for pathway analysis should help to increase the matching of gene annotations between the database and the data provided by the users.

7.3 Project deviations

As a result of prioritisation on the analyses, writing tasks were postponed. Nevertheless, giving priority to the analyses gave us the opportunity to amend some errors in time and rerun the analyses or look for alternatives.

As it has been explained in the Limitations section, one of deviations from the original plan was the number of network analyses to be carried out. The network analyses from ConsensusPathDB, WGCNA, and IPA could not be performed successfully due to problems with their server, the data available to perform the analyses, or restrictions in the number of analyses and time available to conduct the analyses, respectively.

At the beginning of the project, PINBPA [79] was chosen to be included in our study. However, after looking for more information to start running some of the analyses, we finally decided to exclude PINPBA from the project because it had little documentation available and some of it was redundant or confusing. Instead, we looked for another software tool to carry out enough number of analyses for our comparative study. We found WebGestalt, which was more known in the field, had more documentation, was recently updated, and could perform several analyses such as: an ORA, a rank-based pathway analysis, and a network analysis.

When we planned the software tools and analyses to be conducted, we understood from the documentation of Cytoscape that we could conduct network analyses with the pathway analysis results obtained from the software tools already chosen. Nevertheless, Cytoscape plug-ins were needed to perform network analyses and specific formats of data as well. After looking for more documentation, we found a workflow to conduct pathway analyses with g:Profiler and use the results obtained with specific formats (GMT files) to carry out network analyses in EnrichmentMap, a Cytoscape plug-in [15]. Therefore, g:Profiler and EnrichmentMap were included in the project.

Lastly, the creation of figures and tables summarising in a representative way the results from each of the software tools, methods, and databases selected required more time than expected. However, dealing with different data structures, tools, and information provided by each tool and database gave us the opportunity to be aware of issues related to data integration of different sources and comparison of results from the biological, syntactic, and semantic point of view.

In conclusion, several modifications and amendments had to be done to carry out the project. The enormous scope of the project was initially underestimated during planning, and therefore several comparisons between datasets, databases, software tools, methods, and settings, as mentioned in the Discussion and Limitations sections, had to be left out.

8 Glossary

- AD: Alzheimer's disease
- DAG: Directed acyclic graph
- FDR: False discovery rate
- GO: Gene ontology
- GSEA: Gene set enrichment analysis
- GWAS: Genome-wide association study
- IGAP: The International Genomics of Alzheimer's Project
- kb: kilobase
- LOAD: Late-onset Alzheimer's disease
- MMSE: MiniMental Status Exam
- NIAGADS: The National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site
- NFT: Neurofibrillary tangle
- NTA: Network topology-based analysis
- ORA: Over-representation analysis
- QC: Quality control
- SNP: Single nucleotide polymorphism
- rsID: Reference SNP number

9 References

[1] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, Oct 2009.

[2] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *New England Journal of Medicine*, vol. 360, no. 17, pp. 1759–1768, 2009.

[3] A. Dempfle, A. Scherag, R. Hein, L. Beckmann, J. Chang-Claude, and H. Schäfer, "Gene–environment interactions for complex traits: definitions, methodological requirements and challenges," *European Journal of Human Genetics*, vol. 16, pp. 1164–1172, Oct 2008.

[4] K. J. Mitchell, "What is complex about complex disorders?," *Genome Biology*, vol. 13, p. 237, Jan 2012.

[5] M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, and N. L. Pedersen, "Role of Genes and Environments for Explaining Alzheimer Disease," *Archives of General Psychiatry*, vol. 63, pp. 168–174, 02 2006.

[6] T. D. Bird, "Genetic aspects of alzheimer disease," *Genetics in Medicine*, vol. 10, pp. 231–239, Apr 2008.

[7] L. M. Bekris, C.-E. Yu, T. D. Bird, and D. W. Tsuang, "Genetics of alzheimer disease," *Journal of geriatric psychiatry and neurology*, vol. 23, pp. 213–227, Dec 2010. 21045163[pmid].

[8] J. N. Weiss, A. Karma, W. R. MacLellan, M. Deng, C. D. Rau, C. M. Rees, J. Wang, N. Wisniewski, E. Eskin, S. Horvath, Z. Qu, Y. Wang, and A. J. Lusis, "“good enough solutions” and the genetics of complex diseases," *Circulation Research*, vol. 111, no. 4, pp. 493–504, 2012.

[9] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham, B. Grenier-Boley, G. Russo, T. A. Thornton-Wells, N. Jones, A. V. Smith, V. Chouraki, C. Thomas, M. A. Ikram, D. Zelenika, B. N. Vardarajan, Y. Kamatani, C.-F. Lin, A. Gerrish, H. Schmidt, B. Kunkle, M. L. Dunstan, A. Ruiz, M.-T. Bihoreau, S.-H. Choi, C. Reitz, F. Pasquier, P. Hollingworth, A. Ramirez, O. Hanon, A. L. Fitzpatrick, J. D. Buxbaum, D. Campion, P. K. Crane, C. Baldwin, T. Becker, V. Gudnason, C. Cruchaga, D. Craig, N. Amin, C. Berr, O. L. Lopez, P. L. De Jager, V. Deramecourt, J. A. Johnston, D. Evans, S. Lovestone, L. Letenneur, F. J. Morón, D. C. Rubinsztein, G. Eiriksdottir, K. Sleegers, A. M. Goate, N. Fiévet, M. J. Huentelman, M. Gill, K. Brown, M. I. Kamboh, L. Keller, P. Barberger-Gateau, B. McGuinness, E. B. Larson, R. Green, A. J. Myers, C. Dufouil, S. Todd, D. Wallon, S. Love, E. Rogaeva, J. Gallacher, P. St George-Hyslop, J. Clarimon, A. Lleo, A. Bayer, D. W. Tsuang, L. Yu, M. Tsolaki, P. Bossù, G. Spalletta, P. Proitsi, J. Collinge, S. Sorbi, F. Sanchez-Garcia, N. C. Fox, J. Hardy, M. C. D. Naranjo, P. Bosco, R. Clarke, C. Brayne, D. Galimberti, M. Mancuso, F. Matthews, S. Moebus, P. Mecocci, M. Del Zompo, W. Maier, H. Hampel, A. Pilotto, M. Bullido, F. Panza, P. Caffarra, B. Nacmias, J. R. Gilbert, M. Mayhaus, L. Lannfelt, H. Hakonarson, S. Pichler, M. M. Carrasquillo, M. Ingelsson, D. Beekly, V. Alvarez, F. Zou, O. Valladares, S. G. Younkin, E. Coto, K. L. Hamilton-Nelson, W. Gu, C. Razquin, P. Pastor, I. Mateo, M. J. Owen, K. M. Faber, P. V. Jonsson, O. Combarros, M. C. O'Donovan, L. B. Cantwell, H. Soininen, D. Blacker, S. Mead, T. H. Mosley, D. A. Bennett, T. B. Harris, L. Fratiglioni, C. Holmes, R. F. A. G. de Bruijn, P. Passmore, T. J. Montine, K. Bettens, J. I. Rotter, A. Brice, K. Morgan, T. M. Foroud, W. A. Kukull, D. Hannequin, J. F. Powell, M. A. Nalls, K. Ritchie, K. L. Lunetta, J. S. K. Kauwe, E. Boerwinkle, M. Riemenschneider, M. Boada, M. Hiltunen, E. R. Martin, R. Schmidt, D. Rujescu, L.-S. Wang, J.-F. Dartigues, R. Mayeux, C. Tzourio, A. Hofman, M. M. Nöthen, C. Graff, B. M. Psaty, L. Jones, J. L. Haines, P. A. Holmans, M. Lathrop, M. A. Pericak-Vance, L. J. Launer, L. A. Farrer, C. M.

van Duijn, C. Van Broeckhoven, V. Moskvina, S. Seshadri, J. Williams, G. D. Schellenberg, P. Amouyel, E. A. D. I. (EADI), Genetic, E. R. in Alzheimer's Disease (GERAD), A. D. G. C. (ADGC), C. for Heart, and A. R. in Genomic Epidemiology (CHARGE), "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature Genetics*, vol. 45, pp. 1452–1458, Dec 2013.

[10] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu, N. Voyle, P. Proitsi, A. Witoelar, S. Stringer, D. Aarsland, I. S. Almdahl, F. Andersen, S. Bergh, F. Bettella, S. Bjornsson, A. Brækhus, G. Bråthen, C. de Leeuw, R. S. Desikan, S. Djurovic, L. Dumitrescu, T. Fladby, T. J. Hohman, P. V. Jonsson, S. J. Kiddle, A. Rongve, I. Saltvedt, S. B. Sando, G. Selbæk, M. Shoai, N. G. Skene, J. Snaedal, E. Stordal, I. D. Ulstein, Y. Wang, L. R. White, J. Hardy, J. Hjerling-Leffler, P. F. Sullivan, W. M. van der Flier, R. Dobson, L. K. Davis, H. Stefansson, K. Stefansson, N. L. Pedersen, S. Ripke, O. A. Andreassen, and D. Posthuma, "Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk," *Nature Genetics*, vol. 51, pp. 404–413, Mar 2019.

[11] P. Holmans, "7 - statistical methods for pathway analysis of genome-wide data for association with complex genetic traits," in *Computational Methods for Genetics of Complex Traits* (J. C. Dunlap and J. H. Moore, eds.), vol. 72 of *Advances in Genetics*, pp. 141–179, Academic Press, 2010.

[12] P. Y. Kao, K. H. Leung, L. W. Chan, S. P. Yip, and M. K. Yap, "Pathway analysis of complex diseases for gwas, extending to consider rare variants, multi-omics and interactions," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1861, no. 2, pp. 335–353, 2017.

[13] M. Evangelou, A. Rendon, W. H. Ouwehand, L. Wernisch, and F. Dudbridge, "Comparison of methods for competitive tests of pathway analysis," *PLOS ONE*, vol. 7, pp. 1–10, 07 2012.

[14] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, pp. 3587–3595, 06 2005.

[15] J. Reimand, R. Isserlin, V. Voisin, M. Kucera, C. Tannus-Lopes, A. Rostamianfar, L. Wadi, M. Meyer, J. Wong, C. Xu, D. Merico, and G. D. Bader, "Pathway enrichment analysis and visualization of omics data using g:profiler, gsea, cytoscape and enrichmentmap," *Nature Protocols*, vol. 14, pp. 482–517, Feb 2019.

[16] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari, "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences," *Briefings in Bioinformatics*, vol. 19, pp. 286–302, 11 2016.

[17] Q. Zhang, C. Ma, M. Gearing, P. G. Wang, L.-S. Chin, and L. Li, "Integrated proteomics and network analysis identifies protein hubs and network alterations in alzheimer's disease," *Acta Neuropathologica Communications*, vol. 6, p. 19, Mar 2018.

[18] E. M. Blalock, H. M. Buechel, J. Popovic, J. W. Geddes, and P. W. Landfield, "Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient alzheimer's disease," *Journal of Chemical Neuroanatomy*, vol. 42, no. 2, pp. 118–126, 2011. Gene Expression in Neurologic and Psychiatric Disorders.

[19] R. Nativio, Y. Lan, G. Donahue, S. Sidoli, A. Berson, A. R. Srinivasan, O. Shcherbakova, A. Amlie-Wolf, J. Nie, X. Cui, C. He, L.-S. Wang, B. A. Garcia, J. Q. Trojanowski, N. M. Bonini, and S. L. Berger, "An integrated multi-omics approach identifies epigenetic alterations associated with alzheimer's disease," *Nature Genetics*, vol. 52, pp. 1024–1035, Oct 2020.

[20] B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, A. Boland, M. Vronskaya, S. J. van der Lee, A. Amlie-Wolf, C. Bellenguez, A. Frizatti, V. Chouraki, E. R. Martin, K. Sleegers, N. Badarinarayan, J. Jakobsdottir, K. L. Hamilton-Nelson, S. Moreno-Grau, R. Olaso, R. Raybould, Y. Chen, A. B. Kuzma, M. Hiltunen, T. Morgan, S. Ahmad, B. N. Vardarajan, J. Epelbaum, P. Hoffmann, M. Boada, G. W. Beecham, J.-G. Garnier, D. Harold, A. L. Fitzpatrick, O. Valladares, M.-L. Moutet, A. Gerrish, A. V. Smith,

L. Qu, D. Bacq, N. Denning, X. Jian, Y. Zhao, M. Del Zompo, N. C. Fox, S.-H. Choi, I. Mateo, J. T. Hughes, H. H. Adams, J. Malamon, F. Sanchez-Garcia, Y. Patel, J. A. Brody, B. A. Dombroski, M. C. D. Naranjo, M. Daniilidou, G. Eiriksdottir, S. Mukherjee, D. Wallon, J. Uphill, T. Aspelund, L. B. Cantwell, F. Garzia, D. Galimberti, E. Hofer, M. Butkiewicz, B. Fin, E. Scarpini, C. Sarnowski, W. S. Bush, S. Meslage, J. Kornhuber, C. C. White, Y. Song, R. C. Barber, S. Engelborghs, S. Sordon, D. Voijnovic, P. M. Adams, R. Vandenberghe, M. Mayhaus, L. A. Cupples, M. S. Albert, P. P. De Deyn, W. Gu, J. J. Himali, D. Beekly, A. Squassina, A. M. Hartmann, A. Orellana, D. Blacker, E. Rodriguez-Rodriguez, S. Lovestone, M. E. Garcia, R. S. Doody, C. Munoz-Fernadez, R. Sussams, H. Lin, T. J. Fairchild, Y. A. Benito, C. Holmes, H. Karamujić-Čomić, M. P. Frosch, H. Thonberg, W. Maier, G. Roshchupkin, B. Ghetti, V. Giedraitis, A. Kawalia, S. Li, R. M. Huebinger, L. Kilander, S. Moebus, I. Hernández, M. I. Kamboh, R. Brundin, J. Turton, Q. Yang, M. J. Katz, L. Concari, J. Lord, A. S. Beiser, C. D. Keene, S. Helisalmi, I. Kloszewska, W. A. Kukull, A. M. Koivisto, A. Lynch, L. Tarraga, E. B. Larson, A. Haapasalo, B. Lawlor, T. H. Mosley, R. B. Lipton, V. Solfrizzi, M. Gill, W. T. Longstreth, T. J. Montine, V. Frisardi, M. Diez-Fairen, F. Rivadeneira, R. C. Petersen, V. Deramecourt, I. Alvarez, F. Salani, A. Ciaramella, E. Boerwinkle, E. M. Reiman, N. Fievet, J. I. Rotter, J. S. Reisch, O. Hanon, C. Cupidi, A. G. Andre Uitterlinden, D. R. Royall, C. Dufouil, R. G. Maletta, I. de Rojas, M. Sano, A. Brice, R. Cecchetti, P. S. George-Hyslop, K. Ritchie, M. Tsolaki, D. W. Tsuang, B. Dubois, D. Craig, C.-K. Wu, H. Soininen, D. Avramidou, R. L. Albin, L. Fratiglioni, A. Germanou, L. G. Apostolova, L. Keller, M. Koutroumani, S. E. Arnold, F. Panza, O. Gkatzima, S. Asthana, D. Hannequin, P. Whitehead, C. S. Atwood, P. Caffarra, H. Hampel, I. Quintela, Á. Carracedo, L. Lannfelt, D. C. Rubinsztein, L. L. Barnes, F. Pasquier, L. Frölich, S. Barral, B. McGuinness, T. G. Beach, J. A. Johnston, J. T. Becker, P. Passmore, E. H. Bigio, J. M. Schott, T. D. Bird, J. D. Warren, B. F. Boeve, M. K. Lupton, J. D. Bowen, P. Proitsi, A. Boxer, J. F. Powell, J. R. Burke, J. S. K. Kauwe, J. M. Burns, M. Mancuso, J. D. Buxbaum, U. Bonuccelli, N. J. Cairns, A. McQuillin, C. Cao, G. Livingston, C. S. Carlson, N. J. Bass, C. M. Carlsson, J. Hardy, R. M. Carney, J. Bras, M. M. Carrasquillo, R. Guerreiro, M. Allen, H. C. Chui, E. Fisher, C. Masullo, E. A. Crocco, C. DeCarli, G. Bisceglio, M. Dick, L. Ma, R. Duara, N. R. Graff-Radford, D. A. Evans, A. Hodges, K. M. Faber, M. Scherer, K. B. Fallon, M. Riemenschneider, D. W. Fardo, R. Heun, M. R. Farlow, H. Kölsch, S. Ferris, M. Leber, T. M. Foroud, I. Heuser, D. R. Galasko, I. Giegling, M. Gearing, M. Hüll, D. H. Geschwind, J. R. Gilbert, J. Morris, R. C. Green, K. Mayo, J. H. Growdon, T. Feulner, R. L. Hamilton, L. E. Harrell, D. Drichel, L. S. Honig, T. D. Cushion, M. J. Huentelman, P. Hollingworth, C. M. Hulette, B. T. Hyman, R. Marshall, G. P. Jarvik, A. Meggy, E. Abner, G. E. Menzies, L.-W. Jin, G. Leonenko, L. M. Real, G. R. Jun, C. T. Baldwin, D. Grozeva, A. Karydas, G. Russo, J. A. Kaye, R. Kim, F. Jessen, N. W. Kowall, B. Vellas, J. H. Kramer, E. Vardy, F. M. LaFerla, K.-H. Jöckel, J. J. Lah, M. Dichgans, J. B. Leverenz, D. Mann, A. I. Levey, S. Pickering-Brown, and A. P. Lieberman, "Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing," *Nature Genetics*, vol. 51, pp. 414–430, Mar 2019.

[21] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB: toward a more complete picture of cell biology," *Nucleic Acids Research*, vol. 39, pp. D712–D717, 11 2010.

[22] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig, "The ConsensusPathDB interaction database: 2013 update," *Nucleic Acids Research*, vol. 41, pp. D793–D800, 11 2012.

[23] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, pp. 1–13, 11 2008.

[24] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature Protocols*, vol. 4, pp. 44–57, Jan 2009.

[25] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader, "Enrichment map: A network-based method for gene-set enrichment visualization and interpretation," *PLOS ONE*, vol. 5, pp. 1–12, 11 2010.

[26] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: A software environment for integrated models of biomolecular interaction networks,” *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[27] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Research*, vol. 47, pp. W191–W198, 05 2019.

[28] A. Krämer, J. Green, J. Pollard, Jack, and S. Tugendreich, “Causal analysis approaches in Ingenuity Pathway Analysis,” *Bioinformatics*, vol. 30, pp. 523–530, 12 2013.

[29] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, “Magma: Generalized gene-set analysis of gwas data,” *PLOS Computational Biology*, vol. 11, pp. 1–19, 04 2015.

[30] A. Fabregat, K. Sidiropoulos, G. Viteri, O. Forner, P. Marin-Garcia, V. Arnau, P. D’Eustachio, L. Stein, and H. Hermjakob, “Reactome pathway analysis: a high-performance in-memory approach,” *BMC Bioinformatics*, vol. 18, p. 142, Mar 2017.

[31] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D’Eustachio, “The reactome pathway knowledgebase,” *Nucleic Acids Research*, vol. 48, pp. D498–D503, 11 2019.

[32] G. Wu, E. Dawson, A. Duong, R. Haw, and L. Stein, “ReactomeFIViz: a cytoscape app for pathway and network-based data analysis,” *F1000Research*, vol. 3, p. 146, Sept. 2014.

[33] G. Wu and R. Haw, “Functional interaction network construction and analysis for disease discovery.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 1558, pp. 235–253, 2017. Provided by Reactome. Citation Accessed on Mon Mar 01 2021.

[34] B. Zhang, S. Kirov, and J. Snoddy, “WebGestalt: an integrated system for exploring gene sets in various biological contexts,” *Nucleic Acids Research*, vol. 33, pp. W741–W748, 07 2005.

[35] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013,” *Nucleic Acids Research*, vol. 41, pp. W77–W83, 05 2013.

[36] J. Wang, S. Vasaikar, Z. Shi, M. Greer, and B. Zhang, “WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit,” *Nucleic Acids Research*, vol. 45, pp. W130–W137, 05 2017.

[37] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, “WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs,” *Nucleic Acids Research*, vol. 47, pp. W199–W205, 05 2019.

[38] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[39] N. G. G. R. Institute. *Biological Pathways Fact Sheet*, 2020. [Online].

[40] S. R. Chowbina, X. Wu, F. Zhang, P. M. Li, R. Pandey, H. N. Kasamsetty, and J. Y. Chen, “Hpd: an online integrated human pathway database enabling systems biology studies,” *BMC Bioinformatics*, vol. 10, p. S5, Oct 2009.

[41] A. Ma’ayan, “Introduction to network analysis in systems biology,” *Science Signaling*, vol. 4, no. 190, pp. tr5–tr5, 2011.

[42] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma, “The statistical properties of gene-set analysis,” *Nature Reviews Genetics*, vol. 17, pp. 353–364, Jun 2016.

[43] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, “Pathway analysis: State of the art,” *Frontiers in Physiology*, vol. 6, p. 383, 2015.

[44] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, “Identifying significantly impacted pathways: a comprehensive review and assessment,” *Genome Biology*, vol. 20, p. 203, Oct 2019.

[45] P. Khatri, M. Sirota, and A. J. Butte, “Ten years of pathway analysis: Current approaches and outstanding challenges,” *PLOS Computational Biology*, vol. 8, pp. 1–10, 02 2012.

[46] Z. Hu, E. S. Snitkin, and C. DeLisi, “VisANT: an integrative framework for networks in systems biology,” *Briefings in Bioinformatics*, vol. 9, pp. 317–325, 05 2008.

[47] S. I. Berger, J. M. Posner, and A. Ma’ayan, “Genes2networks: connecting lists of gene symbols using mammalian protein interactions databases,” *BMC Bioinformatics*, vol. 8, p. 372, Oct 2007.

[48] M. Bayerlová, K. Jung, F. Kramer, F. Klemm, A. Bleckmann, and T. Beißbarth, “Comparative study on gene set and pathway topology-based enrichment methods,” *BMC Bioinformatics*, vol. 16, p. 334, Oct 2015.

[49] J. Ma, A. Shojaie, and G. Michailidis, “A comparative study of topology-based pathway enrichment analysis methods,” *BMC Bioinformatics*, vol. 20, p. 546, Nov 2019.

[50] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, pp. 1739–1740, 05 2011.

[51] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. Mesirov, and P. Tamayo, “The molecular signatures database hallmark gene set collection,” *Cell Systems*, vol. 1, no. 6, pp. 417–425, 2015.

[52] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, pp. 27–30, 01 2000.

[53] M. Kanehisa, “Toward understanding the origin and evolution of cellular organisms,” *Protein Science*, vol. 28, no. 11, pp. 1947–1951, 2019.

[54] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, “KEGG: integrating viruses and cellular organisms,” *Nucleic Acids Research*, vol. 49, pp. D545–D551, 10 2020.

[55] D. Nishimura, “Biocarta,” *Biotech Software & Internet Report*, vol. 2, no. 3, pp. 117–120, 2001.

[56] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25–29, May 2000.

[57] T. G. O. Consortium, “The Gene Ontology resource: enriching a GOld mine,” *Nucleic Acids Research*, vol. 49, pp. D325–D334, 12 2020.

[58] M. Martens, A. Ammar, A. Riutta, A. Waagmeester, D. Slenter, K. Hanspers, R. A. Miller, D. Digles, E. Lopes, F. Ehrhart, L. J. Dupuis, L. A. Winckers, S. Coort, E. L. Willighagen, C. T. Evelo, A. R. Pico, and M. Kutmon, “WikiPathways: connecting communities,” *Nucleic Acids Research*, vol. 49, pp. D613–D621, 11 2020.

[59] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp, “CORUM: the comprehensive resource of mammalian protein complexes—2019,” *Nucleic Acids Research*, vol. 47, pp. D559–D563, 10 2018.

[60] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén, “Tissue-based map of the human proteome,” *Science*, vol. 347, no. 6220, 2015.

[61] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnström, B. Glimelius, T. Sjöblom, P.-H. Edqvist, D. Djureinovic, P. Micke, C. Lindskog, A. Mardinoglu, and F. Ponten, “A pathology atlas of the human cancer transcriptome,” *Science*, vol. 357, no. 6352, 2017.

[62] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, and E. Lundberg, "A subcellular map of the human proteome," *Science*, vol. 356, no. 6340, 2017.

[63] E. Wingender, "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation," *Briefings in Bioinformatics*, vol. 9, pp. 326–332, 04 2008.

[64] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang, "miRTarBase: a database curates experimentally validated microRNA–target interactions," *Nucleic Acids Research*, vol. 39, pp. D163–D169, 11 2010.

[65] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, and P. N. Robinson, "The Human Phenotype Ontology in 2021," *Nucleic Acids Research*, vol. 49, pp. D1207–D1217, 12 2020.

[66] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein, "Pharmacogenomics knowledge for personalized medicine," *Clinical Pharmacology & Therapeutics*, vol. 92, no. 4, pp. 414–417, 2012.

[67] M. Trupp, T. Altman, C. A. Fulcher, R. Caspi, M. Krummenacker, S. Paley, and P. D. Karp, "Beyond the genome (btg) is a (pgdb) pathway genome database: Humancyc," *Genome Biology*, vol. 11, p. O12, Oct 2010.

[68] S. Yamamoto, N. Sakai, H. Nakamura, H. Fukagawa, K. Fukuda, and T. Takagi, "INOH: ontology-based highly structured database of signal transduction pathways," *Database*, vol. 2011, 11 2011. bar052.

[69] A. Frolkis, C. Knox, E. Lim, T. Jewison, V. Law, D. D. Hau, P. Liu, B. Gautam, S. Ly, A. C. Guo, J. Xia, Y. Liang, S. Shrivastava, and D. S. Wishart, "SMPDB: The Small Molecule Pathway Database," *Nucleic Acids Research*, vol. 38, pp. D480–D487, 11 2009.

[70] T. Jewison, Y. Su, F. M. Disfany, Y. Liang, C. Knox, A. Maciejewski, J. Poelzer, J. Huynh, Y. Zhou, D. Arndt, Y. Djoumbou, Y. Liu, L. Deng, A. C. Guo, B. Han, A. Pon, M. Wilson, S. Rafatnia, P. Liu, and D. S. Wishart, "SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database," *Nucleic Acids Research*, vol. 42, pp. D478–D484, 11 2013.

[71] H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin, "The edinburgh human metabolic network reconstruction and its functional analysis," *Molecular Systems Biology*, vol. 3, no. 1, p. 135, 2007.

[72] K. Kandasamy, S. S. Mohan, R. Raju, S. Keerthikumar, G. S. S. Kumar, A. K. Venugopal, D. Telikicherla, J. D. Navarro, S. Mathivanan, C. Pecquet, S. K. Gollapudi, S. G. Tattikota, S. Mohan, H. Padhukasahasram, Y. Subbannayya, R. Goel, H. K. Jacob, J. Zhong, R. Sekhar, V. Nanjappa, L. Balakrishnan, R. Subbaiah, Y. L. Ramachandra, B. A. Rahiman, T. K. Prasad, J.-X. Lin, J. C. Houtman, S. Desiderio, J.-C. Renauld, S. N. Constantinescu, O. Ohara, T. Hirano, M. Kubo, S. Singh, P. Khatri, S. Draghici, G. D. Bader, C. Sander, W. J. Leonard, and A. Pandey, "Netpath: a public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, p. R3, Jan 2010.

[73] D. Fazekas, M. Koltai, D. Türei, D. Módos, M. Pálfy, Z. Dúl, L. Zsákai, M. Szalay-Bekő, K. Lenti, I. J. Farkas, T. Vellai, P. Csermely, and T. Korcsmáros, "Signalink 2 – a signaling pathway resource with multi-layered regulatory networks," *BMC Systems Biology*, vol. 7, p. 7, Jan 2013.

[74] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: the Pathway Interaction Database," *Nucleic Acids Research*, vol. 37, pp. D674–D679, 10 2008.

[75] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's Online Mendelian Inheritance in Man (OMIM®)," *Nucleic Acids Research*, vol. 37, pp. D793–D796, 10 2008.

[76] K. G. Becker, S. L. White, J. Muller, and J. Engel, "BBID: the biological biochemical image database ," *Bioinformatics*, vol. 16, pp. 745–746, 08 2000.

[77] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, pp. D535–D539, 01 2006.

[78] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2019 update," *Nucleic Acids Research*, vol. 47, pp. D529–D541, 11 2018.

[79] L. Wang, T. Matsushita, L. Madireddy, P. Mousavi, and S. E. Baranzini, "PINBPA: Cytoscape app for network analysis of GWAS data," *Bioinformatics*, vol. 31, pp. 262–264, 09 2014.

10 Appendices

The results from DAVID (pages 61-65) and Reactome (Page 66) will be shown below. The rest of results due to their extent can be found on the following repository: https://github.com/xim56/project2021

Category	Term	Count	PValue	Pop Hits	FDR	Dataset
GOTERM_BP_DIRECT	GO:0007268~chemical synaptic transmission	44	6.88E-08	240	0.000255	Down-Micro
GOTERM_BP_DIRECT	GO:0007269~neurotransmitter secretion	17	5.16E-07	51	0.000955	Down-Micro
GOTERM_BP_DIRECT	GO:0034220~ion transmembrane transport	38	8.51E-07	210	0.001050483372	Down-Micro
GOTERM_CC_DIRECT	GO:0005829~cytosol	372	4.95E-19	3315	3.35E-16	Down-Micro
GOTERM_CC_DIRECT	GO:0014069~postsynaptic density	51	2.08E-16	184	7.05E-14	Down-Micro
GOTERM_CC_DIRECT	GO:0043209~myelin sheath	42	1.67E-13	152	3.77E-11	Down-Micro
GOTERM_CC_DIRECT	GO:0030054~cell junction	78	8.12E-12	459	1.37E-09	Down-Micro
GOTERM_CC_DIRECT	GO:0045211~postsynaptic membrane	47	2.11E-11	211	2.85E-09	Down-Micro
GOTERM_CC_DIRECT	GO:0030425~dendrite	61	1.19E-10	335	1.34E-08	Down-Micro
GOTERM_CC_DIRECT	GO:0005737~cytoplasm	487	5.76E-10	5222	5.57E-08	Down-Micro
GOTERM_CC_DIRECT	GO:0030424~axon	45	1.44E-09	222	1.22E-07	Down-Micro
GOTERM_CC_DIRECT	GO:0005739~mitochondrion	158	1.78E-09	1331	1.34E-07	Down-Micro
GOTERM_CC_DIRECT	GO:0016020~membrane	223	4.55E-07	2200	3.08E-05	Down-Micro
GOTERM_CC_DIRECT	GO:0043005~neuron projection	40	1.93E-06	237	0.000119	Down-Micro
GOTERM_CC_DIRECT	GO:0005794~Golgi apparatus	102	2.41E-06	863	0.000136	Down-Micro
GOTERM_CC_DIRECT	GO:0008021~synaptic vesicle	21	1.01E-05	92	0.000528	Down-Micro
GOTERM_CC_DIRECT	GO:0043025~neuronal cell body	46	1.56E-05	315	0.000755	Down-Micro
GOTERM_CC_DIRECT	GO:0032281~AMPA glutamate receptor complex	11	1.81E-05	28	0.000817	Down-Micro
GOTERM_CC_DIRECT	GO:0048786~presynaptic active zone	11	2.58E-05	29	0.001091254851	Down-Micro
GOTERM_CC_DIRECT	GO:0070062~extracellular exosome	261	4.10E-05	2811	0.00163220031	Down-Micro
GOTERM_CC_DIRECT	GO:0045202~synapse	30	6.34E-05	181	0.002385173578	Down-Micro
GOTERM_CC_DIRECT	GO:0043195~terminal bouton	15	0.000137	62	0.004895096284	Down-Micro
GOTERM_CC_DIRECT	GO:0030672~synaptic vesicle membrane	14	0.000145	55	0.004895096284	Down-Micro
GOTERM_CC_DIRECT	GO:0005874~microtubule	41	0.000436	311	0.0140436366	Down-Micro
GOTERM_CC_DIRECT	GO:0005753~mitochondrial proton-transporting ATP synthase complex	8	0.00055	21	0.01693936691	Down-Micro
GOTERM_CC_DIRECT	GO:0098793~presynapse	14	0.000703	64	0.02019358175	Down-Micro
GOTERM_CC_DIRECT	GO:0005856~cytoskeleton	46	0.000716	371	0.02019358175	Down-Micro
GOTERM_CC_DIRECT	GO:0030426~growth cone	20	0.000837	116	0.02267471949	Down-Micro
GOTERM_CC_DIRECT	GO:0071782~endoplasmic reticulum tubular network	6	0.001120348856	12	0.02890589435	Down-Micro
GOTERM_CC_DIRECT	GO:0048471~perinuclear region of cytoplasm	68	0.001152820011	621	0.02890589435	Down-Micro
GOTERM_CC_DIRECT	GO:0032809~neuronal cell body membrane	7	0.001384519435	18	0.03347570205	Down-Micro
GOTERM_CC_DIRECT	GO:0005743~mitochondrial inner membrane	51	0.001703623568	441	0.03905785072	Down-Micro
GOTERM_CC_DIRECT	GO:0042734~presynaptic membrane	13	0.001730776251	62	0.03905785072	Down-Micro
GOTERM_CC_DIRECT	GO:0031594~neuromuscular junction	12	0.002020757661	55	0.04413073989	Down-Micro
GOTERM_CC_DIRECT	GO:0005622~intracellular	127	0.002198112435	1332	0.0465038162	Down-Micro
GOTERM_MF_DIRECT	GO:0005515~protein binding	814	4.42E-19	8785	5.86E-16	Down-Micro
GOTERM_MF_DIRECT	GO:0044325~ion channel binding	29	1.42E-08	113	9.40E-06	Down-Micro
GOTERM_MF_DIRECT	GO:0005516~calmodulin binding	35	1.76E-06	189	0.000778	Down-Micro
GOTERM_MF_DIRECT	GO:0044822~poly(A) RNA binding	127	4.36E-06	1129	0.00144694324	Down-Micro
GOTERM_MF_DIRECT	GO:0005524~ATP binding	152	0.000103	1495	0.02740263126	Down-Micro
GOTERM_MF_DIRECT	GO:0003924~GTPase activity	35	0.000173	234	0.03815365142	Down-Micro
GOTERM_MF_DIRECT	GO:0008565~protein transporter activity	16	0.000262	72	0.04883419915	Down-Micro
GOTERM_MF_DIRECT	GO:0004842~ubiquitin-protein transferase activity	44	0.000295	329	0.04883419915	Down-Micro
KEGG_PATHWAY	hsa04724:Glutamatergic synapse	29	1.97E-08	114	2.37E-06	Down-Micro
KEGG_PATHWAY	hsa04723:Retrograde endocannabinoid signaling	27	2.25E-08	101	2.37E-06	Down-Micro
KEGG_PATHWAY	hsa04720:Long-term potentiation	21	5.08E-08	66	3.57E-06	Down-Micro
KEGG_PATHWAY	hsa04728:Dopaminergic synapse	29	2.78E-07	128	1.47E-05	Down-Micro
KEGG_PATHWAY	hsa05014:Amyotrophic lateral sclerosis (ALS)	17	4.89E-07	50	2.07E-05	Down-Micro
KEGG_PATHWAY	hsa05033:Nicotine addiction	14	4.83E-06	40	0.00017	Down-Micro
KEGG_PATHWAY	hsa04727:GABAergic synapse	19	6.02E-05	85	0.001782402194	Down-Micro
KEGG_PATHWAY	hsa05010:Alzheimer's disease	29	6.76E-05	168	0.001782402194	Down-Micro
KEGG_PATHWAY	hsa04024:cAMP signaling pathway	32	9.47E-05	198	0.002221062349	Down-Micro
KEGG_PATHWAY	hsa04730:Long-term depression	15	0.000132	60	0.002795282395	Down-Micro
KEGG_PATHWAY	hsa05032:Morphine addiction	19	0.000153	91	0.002928079106	Down-Micro
KEGG_PATHWAY	hsa03060:Protein export	9	0.000199	23	0.003500253627	Down-Micro
KEGG_PATHWAY	hsa04721:Synaptic vesicle cycle	15	0.00023	63	0.003740386273	Down-Micro
KEGG_PATHWAY	hsa04713:Circadian entrainment	19	0.000269	95	0.004054673029	Down-Micro
KEGG_PATHWAY	hsa05031:Amphetamine addiction	15	0.000385	66	0.005418549284	Down-Micro
KEGG_PATHWAY	hsa05132:Salmonella infection	17	0.000476	83	0.00607826666	Down-Micro
KEGG_PATHWAY	hsa04020:Calcium signaling pathway	28	0.00049	179	0.00607826666	Down-Micro
KEGG_PATHWAY	hsa04071:Sphingolipid signaling pathway	21	0.000728	120	0.008531355184	Down-Micro
KEGG_PATHWAY	hsa05131:Shigellosis	14	0.000953	64	0.01058659588	Down-Micro
KEGG_PATHWAY	hsa04360:Axon guidance	21	0.001504350678	127	0.01587089965	Down-Micro

Category	Term	Count	PValue	Pop Hits	FDR	Dataset
KEGG_PATHWAY	hsa04010:MAPK signaling pathway	34	0.001690568774	253	0.01670647852	Down-Micro
KEGG_PATHWAY	hsa04261:Adrenergic signaling in cardiomyocytes	22	0.001801949993	138	0.01670647852	Down-Micro
KEGG_PATHWAY	hsa04114:Oocyte meiosis	19	0.001821085337	111	0.01670647852	Down-Micro
KEGG_PATHWAY	hsa04022:cGMP-PKG signaling pathway	24	0.002034635579	158	0.0178878378	Down-Micro
KEGG_PATHWAY	hsa04141:Protein processing in endoplasmic reticulum	25	0.002303447104	169	0.01944109356	Down-Micro
KEGG_PATHWAY	hsa04540:Gap junction	16	0.00259684665	88	0.02107440935	Down-Micro
KEGG_PATHWAY	hsa04922:Glucagon signaling pathway	17	0.003316274618	99	0.02591607202	Down-Micro
KEGG_PATHWAY	hsa04666:Fc gamma R-mediated phagocytosis	15	0.004397603865	84	0.03046179248	Down-Micro
KEGG_PATHWAY	hsa04725:Cholinergic synapse	18	0.00444837994	111	0.03046179248	Down-Micro
KEGG_PATHWAY	hsa04726:Serotonergic synapse	18	0.00444837994	111	0.03046179248	Down-Micro
KEGG_PATHWAY	hsa05120:Epithelial cell signaling in Helicobacter pylori infection	13	0.004475429227	67	0.03046179248	Down-Micro
KEGG_PATHWAY	hsa04921:Oxytocin signaling pathway	22	0.004998055413	150	0.03176179042	Down-Micro
KEGG_PATHWAY	hsa04664:Fc epsilon RI signaling pathway	13	0.005066933205	68	0.03176179042	Down-Micro
KEGG_PATHWAY	hsa04810:Regulation of actin cytoskeleton	28	0.005268543434	210	0.03176179042	Down-Micro
KEGG_PATHWAY	hsa04015:Rap1 signaling pathway	28	0.005268543434	210	0.03176179042	Down-Micro
KEGG_PATHWAY	hsa04370:VEGF signaling pathway	12	0.006051406259	61	0.03457561255	Down-Micro
KEGG_PATHWAY	hsa04914:Progesterone-mediated oocyte maturation	15	0.006063022106	87	0.03457561255	Down-Micro
KEGG_PATHWAY	hsa04070:Phosphatidylinositol signaling system	16	0.007354017597	98	0.04083415034	Down-Micro
KEGG_PATHWAY	hsa04931:Insulin resistance	17	0.007870059206	108	0.04257903827	Down-Micro
KEGG_PATHWAY	hsa04924:Renin secretion	12	0.008738991388	64	0.04609817957	Down-Micro
KEGG_PATHWAY	hsa04912:GnRH signaling pathway	15	0.009031414857	91	0.04647874475	Down-Micro
KEGG_PATHWAY	hsa04722:Neurotrophin signaling pathway	18	0.009758251831	120	0.04706595405	Down-Micro
KEGG_PATHWAY	hsa00620:Pyruvate metabolism	9	0.009805140289	40	0.04706595405	Down-Micro
KEGG_PATHWAY	hsa05212:Pancreatic cancer	12	0.009814701319	65	0.04706595405	Down-Micro
REACTOME_PATHWAY	R-HSA-212676:R-HSA-212676	10	3.10E-05	23	0.02283169138	Down-Micro
REACTOME_PATHWAY	R-HSA-1445148:R-HSA-1445148	19	5.62E-05	83	0.02283169138	Down-Micro
REACTOME_PATHWAY	R-HSA-438066:R-HSA-438066	8	0.000167	17	0.02966711836	Down-Micro
REACTOME_PATHWAY	R-HSA-399719:R-HSA-399719	8	0.000167	17	0.02966711836	Down-Micro
REACTOME_PATHWAY	R-HSA-977441:R-HSA-977441	7	0.000237	13	0.02966711836	Down-Micro
REACTOME_PATHWAY	R-HSA-5628897:R-HSA-5628897	18	0.000255	85	0.02966711836	Down-Micro
REACTOME_PATHWAY	R-HSA-181429:R-HSA-181429	8	0.000255	18	0.02966711836	Down-Micro
REACTOME_PATHWAY	R-HSA-112314:R-HSA-112314	7	0.000387	14	0.03937511543	Down-Micro
UP_KEYWORDS	Phosphoprotein	818	6.48E-40	8246	2.60E-37	Down-Micro
UP_KEYWORDS	Alternative splicing	926	2.54E-24	10587	5.10E-22	Down-Micro
UP_KEYWORDS	Acetylation	371	5.78E-20	3424	7.72E-18	Down-Micro
UP_KEYWORDS	Synapse	78	2.49E-19	357	2.50E-17	Down-Micro
UP_KEYWORDS	Cytoplasm	449	3.11E-12	4816	2.49E-10	Down-Micro
UP_KEYWORDS	Cell junction	96	4.73E-11	675	3.16E-09	Down-Micro
UP_KEYWORDS	Postsynaptic cell membrane	39	5.92E-10	179	3.39E-08	Down-Micro
UP_KEYWORDS	Transport	202	2.53E-08	1978	1.27E-06	Down-Micro
UP_KEYWORDS	Nucleotide-binding	185	4.80E-08	1788	2.14E-06	Down-Micro
UP_KEYWORDS	Protein transport	80	8.42E-08	610	3.38E-06	Down-Micro
UP_KEYWORDS	Epilepsy	28	1.71E-07	127	6.24E-06	Down-Micro
UP_KEYWORDS	Mental retardation	45	2.37E-06	299	7.91E-05	Down-Micro
UP_KEYWORDS	Lipoprotein	96	4.34E-06	852	0.000134	Down-Micro
UP_KEYWORDS	ATP-binding	142	4.68E-06	1391	0.000134	Down-Micro
UP_KEYWORDS	Coiled coil	273	6.12E-06	3036	0.000159	Down-Micro
UP_KEYWORDS	Ubl conjugation pathway	80	6.36E-06	680	0.000159	Down-Micro
UP_KEYWORDS	Calmodulin-binding	28	6.73E-06	152	0.000159	Down-Micro
UP_KEYWORDS	Methylation	107	1.20E-05	1001	0.000249	Down-Micro
UP_KEYWORDS	Mitochondrion	117	1.24E-05	1119	0.000249	Down-Micro
UP_KEYWORDS	Membrane	602	1.24E-05	7494	0.000249	Down-Micro
UP_KEYWORDS	Neurodegeneration	41	3.97E-05	293	0.000759	Down-Micro
UP_KEYWORDS	Transferase	160	0.000112	1708	0.002044824611	Down-Micro
UP_KEYWORDS	Kinase	79	0.00016	735	0.002790409099	Down-Micro
UP_KEYWORDS	Prenylation	26	0.000284	168	0.004739264675	Down-Micro
UP_KEYWORDS	Exocytosis	14	0.000674	67	0.01080512133	Down-Micro
UP_KEYWORDS	Cytoskeleton	108	0.001080080271	1138	0.01665816111	Down-Micro
UP_KEYWORDS	Cell projection	73	0.001502520959	721	0.02223987308	Down-Micro
UP_KEYWORDS	Stress response	17	0.001552908843	100	0.02223987308	Down-Micro
UP_KEYWORDS	Synaptosome	12	0.00178831374	57	0.0247280624	Down-Micro
UP_KEYWORDS	Magnesium	58	0.002164957145	552	0.0289382605	Down-Micro
UP_KEYWORDS	Microtubule	34	0.002388814974	280	0.03090047757	Down-Micro

Category	Term	Count	PValue	Pop Hits	FDR	Dataset
UP_KEYWORDS	GTP-binding	39	0.003504935191	343	0.04392121911	Down-Micro
UP_KEYWORDS	Golgi apparatus	78	0.004178299978	812	0.04904721287	Down-Micro
UP_KEYWORDS	Ligase	39	0.004264798533	347	0.04904721287	Down-Micro
UP_KEYWORDS	Serine/threonine-protein kinase	43	0.004280928804	393	0.04904721287	Down-Micro
GOTERM_BP_DIRECT	GO:0006351~transcription, DNA-templated	238	1.28E-07	1955	0.000561	Up-Micro
GOTERM_BP_DIRECT	GO:0006355~regulation of transcription, DNA-templated	182	7.67E-06	1504	0.01678586361	Up-Micro
GOTERM_BP_DIRECT	GO:0007179~transforming growth factor beta receptor signaling pathway	23	1.27E-05	92	0.01854335813	Up-Micro
GOTERM_BP_DIRECT	GO:0006397~mRNA processing	34	3.96E-05	179	0.04339125471	Up-Micro
GOTERM_CC_DIRECT	GO:0005654~nucleoplasm	349	7.61E-15	2784	5.58E-12	Up-Micro
GOTERM_CC_DIRECT	GO:0005737~cytoplasm	576	1.07E-13	5222	3.94E-11	Up-Micro
GOTERM_CC_DIRECT	GO:0005634~nucleus	584	5.21E-12	5415	1.28E-09	Up-Micro
GOTERM_CC_DIRECT	GO:0016020~membrane	258	4.57E-08	2200	8.39E-06	Up-Micro
GOTERM_CC_DIRECT	GO:0005925~focal adhesion	65	3.12E-07	391	4.58E-05	Up-Micro
GOTERM_CC_DIRECT	GO:0005829~cytosol	359	3.77E-07	3315	4.61E-05	Up-Micro
GOTERM_CC_DIRECT	GO:0016607~nuclear speck	36	3.94E-05	201	0.004128314219	Up-Micro
GOTERM_CC_DIRECT	GO:0005913~cell-cell adherens junction	50	5.73E-05	323	0.005256548451	Up-Micro
GOTERM_CC_DIRECT	GO:0071141~SMAD protein complex	6	0.000204	8	0.0166311449	Up-Micro
GOTERM_CC_DIRECT	GO:0005856~cytoskeleton	53	0.000272	371	0.019968261	Up-Micro
GOTERM_CC_DIRECT	GO:0043231~intracellular membrane-bounded organelle	72	0.000452	558	0.02780516355	Up-Micro
GOTERM_CC_DIRECT	GO:0030027~lamellipodium	28	0.000481	160	0.02780516355	Up-Micro
GOTERM_CC_DIRECT	GO:0001725~stress fiber	14	0.000492	54	0.02780516355	Up-Micro
GOTERM_CC_DIRECT	GO:0070062~extracellular exosome	287	0.000562	2811	0.02948330697	Up-Micro
GOTERM_CC_DIRECT	GO:0015629~actin cytoskeleton	34	0.000909	218	0.04450094234	Up-Micro
GOTERM_MF_DIRECT	GO:0005515~protein binding	936	1.17E-18	8785	1.54E-15	Up-Micro
GOTERM_MF_DIRECT	GO:0044822~poly(A) RNA binding	163	2.05E-10	1129	1.35E-07	Up-Micro
GOTERM_MF_DIRECT	GO:0003779~actin binding	49	5.00E-06	278	0.00218625267	Up-Micro
GOTERM_MF_DIRECT	GO:0003682~chromatin binding	61	1.69E-05	391	0.005550125839	Up-Micro
GOTERM_MF_DIRECT	GO:0003700~transcription factor activity, sequence-specific DNA binding	119	0.000133	961	0.03494416981	Up-Micro
GOTERM_MF_DIRECT	GO:0031994~insulin-like growth factor I binding	7	0.000273	12	0.04633792148	Up-Micro
GOTERM_MF_DIRECT	GO:0070410~co-SMAD binding	7	0.000273	12	0.04633792148	Up-Micro
GOTERM_MF_DIRECT	GO:0030618~transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity	5	0.000283	5	0.04633792148	Up-Micro
KEGG_PATHWAY	hsa04350:TGF-beta signaling pathway	21	2.07E-05	84	0.002868933681	Up-Micro
KEGG_PATHWAY	hsa04068:FoxO signaling pathway	28	2.17E-05	134	0.002868933681	Up-Micro
KEGG_PATHWAY	hsa04520:Adherens junction	18	7.94E-05	71	0.0069855959	Up-Micro
KEGG_PATHWAY	hsa05200:Pathways in cancer	56	0.000178	393	0.01177757207	Up-Micro
REACTOME_PATHWAY	R-HSA-1989781:R-HSA-1989781	25	7.09E-05	113	0.02765716326	Up-Micro
REACTOME_PATHWAY	R-HSA-2470946:R-HSA-2470946	7	8.87E-05	10	0.02765716326	Up-Micro
REACTOME_PATHWAY	R-HSA-2129379:R-HSA-2129379	13	9.33E-05	38	0.02765716326	Up-Micro
UP_KEYWORDS	Phosphoprotein	944	5.37E-45	8246	2.54E-42	Up-Micro
UP_KEYWORDS	Alternative splicing	1094	2.57E-33	10587	6.08E-31	Up-Micro
UP_KEYWORDS	Nucleus	576	7.69E-18	5244	1.21E-15	Up-Micro
UP_KEYWORDS	Acetylation	406	3.52E-17	3424	4.18E-15	Up-Micro
UP_KEYWORDS	Cytoplasm	528	1.75E-15	4816	1.66E-13	Up-Micro
UP_KEYWORDS	Ubl conjugation	231	2.76E-15	1705	2.18E-13	Up-Micro
UP_KEYWORDS	Coiled coil	352	5.63E-13	3036	3.81E-11	Up-Micro
UP_KEYWORDS	Transcription regulation	284	7.22E-13	2332	4.28E-11	Up-Micro
UP_KEYWORDS	Transcription	289	1.54E-12	2398	8.10E-11	Up-Micro
UP_KEYWORDS	Isopeptide bond	159	7.10E-12	1132	3.36E-10	Up-Micro
UP_KEYWORDS	Zinc	268	3.52E-09	2348	1.51E-07	Up-Micro
UP_KEYWORDS	Repressor	90	1.05E-08	592	4.15E-07	Up-Micro
UP_KEYWORDS	Zinc-finger	210	2.04E-08	1781	7.42E-07	Up-Micro
UP_KEYWORDS	Metal-binding	377	1.20E-07	3640	4.06E-06	Up-Micro
UP_KEYWORDS	RNA-binding	88	6.59E-06	665	0.000208	Up-Micro
UP_KEYWORDS	Cytoskeleton	135	7.14E-06	1138	0.000211	Up-Micro
UP_KEYWORDS	Polymorphism	1063	1.01E-05	12043	0.000281	Up-Micro
UP_KEYWORDS	Methylation	119	2.40E-05	1001	0.000632	Up-Micro
UP_KEYWORDS	Proto-oncogene	39	3.81E-05	236	0.000951	Up-Micro
UP_KEYWORDS	Host-virus interaction	55	5.77E-05	385	0.001367294883	Up-Micro
UP_KEYWORDS	mRNA splicing	41	6.91E-05	260	0.001558851164	Up-Micro
UP_KEYWORDS	Actin-binding	42	0.000106	274	0.002283775321	Up-Micro
UP_KEYWORDS	Nucleotide-binding	189	0.000124	1788	0.00255110745	Up-Micro
UP_KEYWORDS	Activator	82	0.000132	661	0.002573101311	Up-Micro
UP_KEYWORDS	mRNA processing	48	0.000136	332	0.002573101311	Up-Micro

Category	Term	Count	PValue	Pop Hits	FDR	Dataset
UP_KEYWORDS	Chromosomal rearrangement	48	0.000156	334	0.002852199151	Up-Micro
UP_KEYWORDS	Guanine-nucleotide releasing factor	27	0.000165	149	0.002889103729	Up-Micro
UP_KEYWORDS	Spliceosome	24	0.000218	127	0.003697683485	Up-Micro
UP_KEYWORDS	Apoptosis	68	0.000276	536	0.004508161889	Up-Micro
UP_KEYWORDS	Chromatin regulator	42	0.00029	287	0.00458757219	Up-Micro
UP_KEYWORDS	DNA-binding	209	0.000369	2050	0.005649690177	Up-Micro
UP_KEYWORDS	Transferase	177	0.000536	1708	0.007936026453	Up-Micro
UP_KEYWORDS	Endosome	61	0.000593	481	0.008524102161	Up-Micro
UP_KEYWORDS	Alternative promoter usage	18	0.000849	90	0.01182924589	Up-Micro
UP_KEYWORDS	Bromodomain	11	0.000877	39	0.01187947458	Up-Micro
UP_KEYWORDS	Alport syndrome	5	0.001249076398	7	0.01644617257	Up-Micro
UP_KEYWORDS	Proteoglycan	12	0.001615370302	49	0.02069420332	Up-Micro
UP_KEYWORDS	Cell projection	82	0.001854320439	721	0.02313020758	Up-Micro
UP_KEYWORDS	mRNA transport	19	0.00215814745	106	0.02622979208	Up-Micro
UP_KEYWORDS	ATP-binding	142	0.003304593268	1391	0.03915943023	Up-Micro
GOTERM_BP_DIRECT	GO:0006099~tricarboxylic acid cycle	8	1.34E-06	29	0.002217213925	Down-RNA
GOTERM_BP_DIRECT	GO:0006091~generation of precursor metabolites and energy	9	9.99E-06	53	0.008253151258	Down-RNA
GOTERM_BP_DIRECT	GO:0055114~oxidation-reduction process	28	6.92E-05	592	0.03813263474	Down-RNA
GOTERM_CC_DIRECT	GO:0005739~mitochondrion	68	6.77E-13	1331	2.22E-10	Down-RNA
GOTERM_CC_DIRECT	GO:0070062~extracellular exosome	94	1.38E-07	2811	2.17E-05	Down-RNA
GOTERM_CC_DIRECT	GO:0005743~mitochondrial inner membrane	28	1.99E-07	441	2.17E-05	Down-RNA
GOTERM_CC_DIRECT	GO:0043209~myelin sheath	14	9.80E-06	152	0.000803	Down-RNA
KEGG_PATHWAY	hsa01200:Carbon metabolism	13	1.39E-05	113	0.003005597319	Down-RNA
KEGG_PATHWAY	hsa00020:Citrate cycle (TCA cycle)	7	6.19E-05	30	0.005726992237	Down-RNA
KEGG_PATHWAY	hsa01100:Metabolic pathways	50	7.92E-05	1219	0.005726992237	Down-RNA
REACTOME_PATHWAY	R-HSA-71403:R-HSA-71403	6	4.58E-05	19	0.01735516729	Down-RNA
UP_KEYWORDS	Mitochondrion	54	5.48E-10	1119	1.92E-07	Down-RNA
UP_KEYWORDS	Transit peptide	31	1.20E-07	536	2.10E-05	Down-RNA
UP_KEYWORDS	Oxidoreductase	32	2.18E-07	582	2.55E-05	Down-RNA
UP_KEYWORDS	Acetylation	104	3.85E-07	3424	3.38E-05	Down-RNA
UP_KEYWORDS	Mitochondrion inner membrane	17	5.72E-05	270	0.004016016987	Down-RNA
UP_KEYWORDS	Tricarboxylic acid cycle	6	7.25E-05	24	0.004242088891	Down-RNA
UP_KEYWORDS	NADP	13	0.00022	186	0.01103486086	Down-RNA
UP_KEYWORDS	Flavoprotein	10	0.000504	122	0.02213298575	Down-RNA
GOTERM_CC_DIRECT	GO:0005654~nucleoplasm	70	8.44E-05	2784	0.01437238472	Up-RNA
GOTERM_CC_DIRECT	GO:0005634~nucleus	117	0.000105	5415	0.01437238472	Up-RNA
UP_KEYWORDS	Nucleus	124	3.39E-07	5244	9.30E-05	Up-RNA
UP_KEYWORDS	Transcription regulation	63	2.15E-05	2332	0.002947865844	Up-RNA
UP_KEYWORDS	Transcription	63	4.96E-05	2398	0.004526745462	Up-RNA
UP_KEYWORDS	DNA-binding	54	0.000193	2050	0.01318680792	Up-RNA
UP_KEYWORDS	Cell cycle	24	0.000263	650	0.01441811019	Up-RNA
UP_KEYWORDS	Repressor	22	0.00047	592	0.02145023704	Up-RNA
GOTERM_BP_DIRECT	GO:0002286~T cell activation involved in immune response	13	5.06E-08	22	0.000111	GWAS
GOTERM_BP_DIRECT	GO:0033141~positive regulation of peptidyl-serine phosphorylation of STAT protein	12	8.11E-08	19	0.000111	GWAS
GOTERM_BP_DIRECT	GO:0002323~natural killer cell activation involved in immune response	12	8.11E-08	19	0.000111	GWAS
GOTERM_BP_DIRECT	GO:0033344~cholesterol efflux	13	3.18E-07	25	0.000325	GWAS
GOTERM_BP_DIRECT	GO:0070328~triglyceride homeostasis	13	5.43E-07	26	0.000374	GWAS
GOTERM_BP_DIRECT	GO:0043691~reverse cholesterol transport	11	5.47E-07	18	0.000374	GWAS
GOTERM_BP_DIRECT	GO:0042100~B cell proliferation	14	1.11E-06	32	0.000648	GWAS
GOTERM_BP_DIRECT	GO:0033700~phospholipid efflux	9	5.97E-06	14	0.003058936901	GWAS
GOTERM_BP_DIRECT	GO:0030183~B cell differentiation	19	8.70E-06	66	0.003960350062	GWAS
GOTERM_BP_DIRECT	GO:0034380~high-density lipoprotein particle assembly	7	1.03E-05	8	0.004210241428	GWAS
GOTERM_BP_DIRECT	GO:0043330~response to exogenous dsRNA	13	1.60E-05	34	0.005951472193	GWAS
GOTERM_BP_DIRECT	GO:0006959~humoral immune response	17	1.83E-05	57	0.00625583716	GWAS
GOTERM_BP_DIRECT	GO:0060337~type I interferon signaling pathway	18	2.24E-05	64	0.006569725046	GWAS
GOTERM_BP_DIRECT	GO:0042632~cholesterol homeostasis	18	2.24E-05	64	0.006569725046	GWAS
GOTERM_BP_DIRECT	GO:0002250~adaptive immune response	30	2.57E-05	148	0.00703243608	GWAS
GOTERM_BP_DIRECT	GO:0060338~regulation of type I interferon-mediated signaling pathway	11	3.47E-05	26	0.00888618101	GWAS
GOTERM_BP_DIRECT	GO:0034375~high-density lipoprotein particle remodeling	8	0.000127	15	0.03072093765	GWAS
GOTERM_BP_DIRECT	GO:0045087~innate immune response	61	0.000172	430	0.03913021977	GWAS
GOTERM_BP_DIRECT	GO:0008203~cholesterol metabolic process	17	0.000186	68	0.0401647151	GWAS
GOTERM_BP_DIRECT	GO:0051607~defense response to virus	30	0.000196	165	0.0401647151	GWAS
GOTERM_BP_DIRECT	GO:0048261~negative regulation of receptor-mediated endocytosis	6	0.000221	8	0.04312454211	GWAS

Category	Term	Count	PValue	Pop Hits	FDR	Dataset
GOTERM_BP_DIRECT	GO:0033081~regulation of T cell differentiation in thymus	5	0.000265	5	0.04888240629	GWAS
GOTERM_BP_DIRECT	GO:0042157~lipoprotein metabolic process	12	0.000274	38	0.04888240629	GWAS
GOTERM_MF_DIRECT	GO:0005132~type I interferon receptor binding	12	1.60E-08	17	2.10E-05	GWAS
GOTERM_MF_DIRECT	GO:0004872~receptor activity	38	5.71E-05	217	0.03740726816	GWAS
KEGG_PATHWAY	hsa05320:Autoimmune thyroid disease	18	8.56E-07	52	0.000222	GWAS
KEGG_PATHWAY	hsa05168:Herpes simplex infection	35	1.36E-05	183	0.001767507616	GWAS
KEGG_PATHWAY	hsa05152:Tuberculosis	33	4.30E-05	177	0.003642935069	GWAS
KEGG_PATHWAY	hsa05162:Measles	27	5.79E-05	133	0.003642935069	GWAS
KEGG_PATHWAY	hsa05164:Influenza A	32	7.54E-05	174	0.003642935069	GWAS
KEGG_PATHWAY	hsa04620:Toll-like receptor signaling pathway	23	8.44E-05	106	0.003642935069	GWAS
KEGG_PATHWAY	hsa04623:Cytosolic DNA-sensing pathway	16	0.000282	64	0.01043084834	GWAS
KEGG_PATHWAY	hsa05160:Hepatitis C	25	0.000395	133	0.01278055514	GWAS
KEGG_PATHWAY	hsa04622:RIG-I-like receptor signaling pathway	16	0.000783	70	0.02252401686	GWAS
KEGG_PATHWAY	hsa04640:Hematopoietic cell lineage	18	0.001089870036	87	0.02822763393	GWAS
KEGG_PATHWAY	hsa04975:Fat digestion and absorption	11	0.001390965164	39	0.03275090705	GWAS
KEGG_PATHWAY	hsa04650:Natural killer cell mediated cytotoxicity	22	0.001669347106	122	0.03421659806	GWAS
KEGG_PATHWAY	hsa04380:Osteoclast differentiation	23	0.001816977557	131	0.03421659806	GWAS
KEGG_PATHWAY	hsa05150:Staphylococcus aureus infection	13	0.001849545841	54	0.03421659806	GWAS
REACTOME_PATHWAY	R-HSA-912694:R-HSA-912694	11	5.12E-05	26	0.02929734955	GWAS
REACTOME_PATHWAY	R-HSA-909733:R-HSA-909733	18	7.45E-05	67	0.02929734955	GWAS
REACTOME_PATHWAY	R-HSA-933541:R-HSA-933541	12	9.98E-05	33	0.02929734955	GWAS
REACTOME_PATHWAY	R-HSA-983231:R-HSA-983231	24	0.000156	112	0.03432145982	GWAS
UP_KEYWORDS	Phosphoprotein	793	2.55E-08	8246	1.27E-05	GWAS
UP_KEYWORDS	Alternative splicing	967	6.95E-06	10587	0.001629083992	GWAS
UP_KEYWORDS	Polymorphism	1085	9.85E-06	12043	0.001629083992	GWAS
UP_KEYWORDS	Immunity	68	6.66E-05	500	0.008255331787	GWAS
UP_KEYWORDS	Lipid transport	21	0.000102	96	0.01015793538	GWAS
UP_KEYWORDS	Sushi	15	0.000147	56	0.0121409963	GWAS
UP_KEYWORDS	Antiviral defense	23	0.000202	116	0.01431515102	GWAS
UP_KEYWORDS	Sugar transport	12	0.000385	41	0.02203312701	GWAS
UP_KEYWORDS	Amyotrophic lateral sclerosis	10	0.0004	29	0.02203312701	GWAS
UP_KEYWORDS	Host-virus interaction	52	0.00062	385	0.03075224533	GWAS
UP_KEYWORDS	Complement pathway	10	0.000691	31	0.0311778076	GWAS

Pathway identifier	Pathway name	#Entities found	#Entities total	Entities pValue	Entities FDR	Dataset
R-HSA-112314	Neurotransmitter receptors and postsynaptic signal transmission	53	231	5.46E-06	0.005190778659	Down-Micro
R-HSA-438066	Unblocking of NMDA receptors, glutamate binding and activation	14	27	5.94E-06	0.005190778659	Down-Micro
R-HSA-6794362	Protein-protein interactions at synapses	28	93	9.74E-06	0.005668613727	Down-Micro
R-HSA-112316	Neuronal System	91	487	1.46E-05	0.006376016045	Down-Micro
R-HSA-442755	Activation of NMDA receptors and postsynaptic events	31	113	1.98E-05	0.006918253698	Down-Micro
R-HSA-112315	Transmission across Chemical Synapses	68	341	2.61E-05	0.007605384493	Down-Micro
R-HSA-438064	Post NMDA receptor activation events	27	96	4.37E-05	0.01089276926	Down-Micro
R-HSA-9022702	MECP2 regulates transcription of neuronal ligands	6	13	8.40E-06	0.009004104465	Down-RNA
R-HSA-3371568	Attenuation phase	9	47	5.56E-05	0.02979663296	Down-RNA
R-HSA-877300	Interferon gamma signaling	85	250	4.01E-13	7.20E-10	GWAS
R-HSA-913531	Interferon Signaling	114	394	1.40E-12	1.26E-09	GWAS
R-HSA-909733	Interferon alpha/beta signaling	61	188	4.43E-09	2.65E-06	GWAS
R-HSA-1236977	Endosomal/Vacuolar pathway	35	82	1.70E-08	7.60E-06	GWAS
R-HSA-983170	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	38	102	1.32E-07	4.74E-05	GWAS
R-HSA-933541	TRAF6 mediated IRF7 activation	21	43	1.51E-06	0.000451	GWAS
R-HSA-202427	Phosphorylation of CD3 and TCR zeta chains	20	45	1.03E-05	0.00263086831	GWAS
R-HSA-202430	Translocation of ZAP-70 to Immunological synapse	19	42	1.33E-05	0.002980270843	GWAS
R-HSA-389948	PD-1 signaling	19	45	3.34E-05	0.006650808541	GWAS
R-HSA-1236974	ER-Phagosome pathway	46	173	4.91E-05	0.008538288711	GWAS
R-HSA-202433	Generation of second messenger molecules	22	59	5.24E-05	0.008538288711	GWAS
R-HSA-9029569	NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux	22	66	0.000248	0.03693960813	GWAS