CO-EXPRESSION PAIRS AND MODULES (CoEX-PM): A SHINY APPLICATION AND AN EXAMPLE CASE STUDY ON CHROMOGRANINS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
NEUROSCIENCE

By
Tuğberk Kaya
September 2018

CO-EXPRESSION PAIRS AND MODULES (CoEX-PM): A SHINY APPLICATION AND AN EXAMPLE CASE STUDY ON CHROMOGRANINS

By Tuğberk Kaya
September 2018

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Özlen Konu Karakayalı (Advisor)

Michelle Marie Adams

Aybar Can Acar

Approved for the Graduate School of Engineering and Science:

Ezhan Karasan
Director of the Graduate School

ABSTRACT

CO-EXPRESSION PAIRS AND MODULES (CoEX-PM): A SHINY APPLICATION AND AN EXAMPLE CASE STUDY ON CHROMOGRANINS

Tuğberk Kaya
M.Sc. in Neuroscience
Advisor: Özlen Konu Karakayalı
September 2018

Gene expression signatures have been proved to be effective biomarkers of tumorigenesis and metastasis especially when alternative methods are inconvenient or ineffective. Nevertheless, handling very large datasets obtained via high-throughput protocols to extract gene expression signatures may prove challenging. A great number of software packages that facilitate such analyses have been written in R programming language are publicly available and free. However, the relatively steep learning curve that is required to use R proficiently prevents the utilization of these packages. I have developed the Shiny application Co-expression Modules and Pairs (CoEX-PM) using R programming language and the R package *shiny*. The CoEX-PM application handles human Affymetrix microarray data and enables users to generate pairwise correlation plots, conduct meta-correlation analysis with user-selected GEO datasets along with co-expression module generation by WGCNA program for genes of interest. The CoEX-PM application provides the user with a GUI, therefore, does not require any coding knowledge to perform the analyses.

Pheochromocytoma (PCC) and neuroblastoma (NB) are neural-crest derived tumors, common in adults and children, respectively and are both associated with high-rate of morbidity and mortality. In addition, both tumor types display neuroendocrine tumor (NET) characteristics. Chromogranin A (CgA) has been linked with NETs as a moderately sensitive and non-specific tumor marker. The chromogranin family consists of up to seven members, three of which are

chromogranin (CgA), chromogranin B (CgB) and secretogranin II (SgII) or occasionally named as chromogranin C (CgC). However, it is not known whether chromogranin/secretogranin family members are differentially co-expressed in PCC and NB. Here, I investigate the degree of co-expression in gene networks by analyzing gene expression signatures of the chromogranin/secretogranin paralogous gene family using CoEX-PM application on neuroendocrine tumor datasets. The findings indicate presence of concise and highly co-expressed functional components in PCC and NB driven by chromogranin expression signatures.

ÖZET

KO-EKSPRESYON ÇİFTLERİ VE MODÜLLERİ (CoEX-PM): BİR SHINY APLİKASYONU VE KROMOGRANİNLER ÜZERİNDE BİR İNCELEME

Tuğberk Kaya
Nörobilim, Yüksek Lisans
Tez Danışmanı: Özlen Konu Karakayalı
Eylül 2018

Gen ekspresyon profillerinin, özellikle alternatif metodlar yetersiz ve etkisiz kaldığında, etkin tümörijenez ve metastaz bio-belirteçleri oldukları kanıtlanmıştır. Bununla birlikte, yüksek verimli protokoller aracılığıyla elde edilen büyük veri kümelerini analiz ederek gen ekspresyon profilleri bulmaya çalışmak zorlu olabilir. R programlama dilinde bu tip analizleri kolaylaştıran çok sayıda yazılım paketi ücretsiz kullanıma açık şekilde yer almaktadır. Fakat, R'yi etkili bir şekilde kullanmak için gerekli olan nispeten sarp öğrenme eğrisi, bu paketlerin kullanılmasını kimi zaman engellemektedir. Bu tez kapsamında R programlama dili ve *shiny* paketini kullanarak CoEX-PM aplikasyonunu geliştirdim. CoEX-PM uygulaması, insan Affymetrix mikrodizi verilerini kullanır ve kullanıcıların çift yönlü korelasyon grafikleri oluşturmasına, kullanıcı tarafından seçilen GEO veri kümeleriyle meta-korelasyon analizi gerçekleştirmesine ve ilgili genler için WGCNA programı ile birlikte ko-ekpresyon gen modülleri oluşturmasına olanak tanır. CoEX-PM tüm bu analizleri gerçekleştirmesi için kullanıcıya bir arayüz sağlar, bu nedenle herhangi bir kodlama bilgisi veya tecrübesi gerektirmemektedir.
Pheochromocytoma (PCC) ve nöroblastoma (NB), sırasıyla yetişkinlerde ve çocuklarda sık görülen nöral-krest kaynaklı tümörlerdir ve her ikisi de yüksek oranda morbidite ve mortalite ile ilişkilidir. Ek olarak, her iki tümör tipi de nöroendokrin tümör (NET) özelliklerini gösterir. Kromogranin A (CgA), orta derecede hassas ve spesifik olmayan bir nöroendokrin tümör markörü olarak rapor edilmiştir. Kromogranin ailesinin 7 üyesi vardır, bunlardan üçü

kromogranin (CgA), kromogranin B (CgB) ve sekretogranin II (SgII) veya kimi zaman kromogranin C (CgC) olarak adlandırılır. Kromogranin / secretogranin aile üyelerinin PCC ve NB'de farklı ko-ekspresyon şekilleri gösterip göstermedikleri bilinmemektedir. Bu tezde, nöroendokrin tümör veri kümeleri üzerinde CoEX-PM uygulaması kullanılarak kromogranin / secretogranin paralog gen ailesinin gen ekspresyon imzaları analiz edilmiş, gen ağlarındaki ko-ekspresyon derecesi araştırılmıştır. Bulgular, kromogranin ekspresyon seviyesi ile bağlantılı, PCC ve NB'de özlü ve yüksek düzeyde birlikte ifade edilen fonksiyonel bileşenlerin varlığını göstermektedir.

Acknowledgements

I acknowledge that I was financially supported in the form of monthly stipends by the Neuroscience Department affiliated with the Graduate School of Engineering and Science, Bilkent University.

I would like to express my deepest gratitude to my thesis advisor Dr. Özlen Konu for always supporting me even through the darkest times, pointing at the right direction and keeping me on track. I thank Drs. Michelle M. Adams and Aybar C. Acar for being members of my thesis committee and helpful comments they provided. I also thank Dr. Huma Shehwana for sharing her R codes when needed and Alperen Taciroğlu for valuable discussions on shiny applications. I thank all the members of Konu lab for their constant positivity and willingness to help each other without hesitation. Lastly, I'd like to thank my friends and family who are the ones that make it all worth it.

Contents

1. Introduction 1

Web tools using Shiny for gene expression analysis 1

Neuroendocrine tumors and their classification 2

Incidence and prevalence of NETs 3

Pheochromocytomas/Paragangliomas 4

Neuroblastoma 4

Microarray Technology 5

Meta-Analysis 6

Previous microarray studies focusing on Pheochromocytoma 8

Neuroendocrine Tumor (NET) Biomarkers 9

Biomarker limitations and issues 10

Chromogranins: structure and function and evolution 12

2. Aims 13

3. Methods 15

3.1. Data Acquisition and Normalization 15

3.2. Shiny Application Design 16

3.2.1. Tab 1: Pairwise correlation analysis. 17

3.2.2. Tab 2: Meta-analysis Tab 18

Metacor package 18

3.2.3. Tab 3: WGCNA Tab 18

4. Results 20

4.1. COEX-PM Application 20

4.1.1 Gene-Correlation Tab 20

4.1.2. Meta-correlation Tab 24

4.1.3. WGCNA Tab 25

4.2. Case-study: Chromogranins in NETs 29

4.2.1. Pairwise correlations between secretogranin-chromogranin genes 29

4.2.2. Meta-correlation of CHGA and CHGB across transcriptome 39
4.2.3 Chromogranin and Secretogranin Co-expression modules using WGCNA Tab 61
Conclusions and Discussion.....79
Future Prospects.....85
References 86

Table of Figures

Figure 1:The screenshot of Sample Clustering part of the Two Gene Correlation Tab. 21
Figure 2: The screenshot of Plot part of the Two Gene Correlation Tab. 23
Figure 3: The screenshot of Meta-correlation Tab 24
Figure 4: The screenshot of WCGNA Tab: Limma Differential Gene Expression Analysis – Phenodata table. 25
Figure 5: The screenshot of WCGNA Tab: Limma Differential Gene Expression Analysis – limma table..... 26
Figure 6: WGCNA Tab: Sample Clustering and Power Analysis – Sample clustering scheme ... 27
Figure 7: Network topology analysis plots for various soft-thresholding powers. 28
Figure 8: WGCNA Tab: Module Construction – Module dendrogram of genes, each color represents a different module, grey color is reserved for genes that are not a member of any module..... 28
Figure 9: CHGA-CHGB correlation plots for nine distinct datasets. GSE67066, GSE2841 – PCC; GSE13136, GSE16476, GSE16237, GSE12460 – neuroblastoma; GSE39612 – Merkel cell carcinoma; GSE35679 – lung carcinoid, GSE12667 – lung adenocarcinoma. 30
Figure 10: Correlation plots of chromogranins of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE67066 dataset (PCC, pheochromocytoma). The color gradient implies the expression level. 31
Figure 11: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE2841 dataset. The color gradient implies the expression level. GSE2841 contains 76 pheochromocytoma samples of various genetic origin. 32
Figure 12: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE16476 dataset. The color gradient implies the expression level. GSE16476 contains 88 human neuroblastoma samples. 33
Figure 13: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE13136 dataset. The color gradient implies the expression level. GSE13136 contains 30 primary neuroblastoma samples. 34
Figure 14: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE12460 dataset. The color gradient implies the expression level. GSE12460 contains 64 neuroblastic tumors (mainly neuroblastoma). 35
Figure 15: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE35679 dataset. The color gradient implies the expression level. GSE35679 contains 13 primary lung carcinoid samples. 36
Figure 16: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE12667 dataset. The color gradient implies the expression level. GSE12667 contains 75 lung adenocarcinoma samples..... 37
Figure 17:Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE39612 dataset. The color gradient implies the expression level..... 38
Figure 18: Histogram of r.mean values for CHGA (Left) and CHGB (right). The r.mean values was generated a result of the meta-correlation analysis using the GSE67066 and GSE2841 (PCC) datasets. 39
Figure 19:Correlation plot of the r.mean values for CHGA and CHGB. The r.mean values was generated a result of the meta-correlation analysis using the GSE67066 and GSE2841 (PCC)

datasets. The color gradient represents the chromogranin divergence (chga - chgb). 40
Figure 20: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the CHGB gene. Genes with r.mean>0.55 and p value < 0.003 were selected for the STRING mapping. ... 43
Figure 21: Functional enrichment results for the selected genes mentioned in Figure 20. 44
Figure 22: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the CHGA gene. Genes with r.mean>0.48 were selected for the STRING mapping 45
Figure 23: The functional annotation of CHGA PPI network in Figure 22. 46
Figure 24: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the SCG2 gene. Genes with r.mean>0.65 were selected for the STRING mapping 47
Figure 25: Functional annotation of Protein-Protein interaction network in Figure 24. 48
Figure 26: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the SCG3 gene. Genes with r.mean>0.6 were selected for the STRING mapping. 49
Figure 27: Functional annotation of PPI network in Figure 26. 50
Figure 28: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the CHGA gene. Genes with r.mean>0.6 were selected for the STRING mapping 51
Figure 29: Functional annotation of PPI network in Figure 28. 52
Figure 30: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the CHGB gene. Genes with r.mean>0.6 were selected for the STRING mapping. ... 53
Figure 31: Functional annotation of PPI network in Figure 30. 54
Figure 32: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG2 gene. Genes with r.mean>0.65 were selected for the STRING mapping.... 55
Figure 33: Functional annotation of PPI network in Figure 32. 56
Figure 34: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG3 gene. Genes with r.mean>0.58 were selected for the STRING mapping.... 57
Figure 35: Functional annotation of PPI network in Figure 34. 58
Figure 36: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG5 gene. Genes with r.mean>0.65 were selected for the STRING mapping.... 59
Figure 37: Functional annotation of PPI network in Figure 36. 60
Figure 38: Sample clustering for GSE67066 with a cut-off line at 140. 61
Figure 39: Network topology analysis for filtered GSE67066. 62
Figure 40: Clustering dendrogram of probes, with dissimilarity based on topological overlap. Module colors are shown for their assigned probes..... 64
Figure 41: Module-gene associations, first part. Rows correspond to module eigengenes, columns to genes of interest. Each cell displays the corresponding correlation and p-value. Color coding is established by correlation according to the legend. 65
Figure 42: Module-gene associations, second part. Rows correspond to module eigengenes, columns to genes of interest. Each cell displays the corresponding correlation and p-value. Color coding is established by correlation according to the legend. 66
Figure 43: Scatterplot of Gene Significance (GS) for SCG2 and CHGA for Module Membership (MM) in the black and red modules, respectively..... 68
Figure 44: Hierarchical clustering dendrogram of the eigengenes and the genes of interest. 69
Figure 45: The heatmap of eigengene adjacency with gene of interests added. 69

Figure 46: PPI network of black module genes with top 200 Gene Significance. 72
Figure 47: PPI network of red module genes with top 200 GS. 74
Figure 48: PPI network of salmon module genes. 75
Figure 49: PPI network of blue module genes with top ~200 GS. 77

List of Tables

Table 1: The datasets that were acquired from the GEO database and used in further analyses. . 16
Table 2 : Functional annotation table for genes with high chga-chgb r.mean 41
Table 3: Functional annotation table for genes with low chga-chgb r.mean..... 41
Table 4: Most correlated MEs for each gene of interest. 67

1. Introduction

Accumulation of massive amounts of high-throughput data in publicly open repositories such as Gene Expression Omnibus (GEO) and ArrayExpress have been crucial in enabling researchers to access and share data online for free. Over the last decades, microarray studies have provided great insight about the disease state with regards to tumorigenesis and metastasis; microarrays also provide researchers with ways to determine significant patterns of gene expression that characterize the tumor state. However, there is still a need to integrate different data types to analyze co-expression of pairs of genes from high throughput expression datasets effectively yet tools that contain graphical user interfaces (GUI) on the WWW are scarce.

The characterization of elements underlying neuroendocrine tumors are of particular interest for this thesis. The origin of neuroendocrine tumors vary greatly and so does their physiological features, which make their prognosis and intervention dramatically difficult. Gene expression signatures have proved to be effective biomarkers of tumorigenesis and metastasis when other methods of diagnosis are inadequate. On the other hand, analyzing and manipulation of very large datasets obtained via high-throughput protocols poses multiple challenges. Nevertheless, a great number of packages in R programming language have been written by expert programmers to assist researchers in handling very large biological datasets. However, there is a steep learning curve to overcome to use R effectively and this might make it difficult to utilize these valuable packages. The R package *Shiny* is a web application framework for R and lets the users to develop applications that conduct simple or complex analyses written in R.

Web tools using Shiny for gene expression analysis

There is a need in the literature for efficient tools that conduct gene expression analysis as well as bypass the coding process and convert the process to a customizable interface. There have been multiple studies addressing this problem; using Shiny to develop applications that facilitate high-throughput expression data analysis.

ScanGEO [1] is a Shiny application that identifies differentially expressed genes across

multiple GEO datasets. This application enables the parallel analysis of GEO datasets. It retrieves relevant datasets using the selected organism and an optional search term. For the differential gene expression part, a custom list of genes or a KEGG pathway is selected. The results consist of a list of most significant genes and a list of studies with the most number of significant genes.

Another Shiny application that handles GEO datasets is the *shinyGEO* [2] application. When an accession number and platform name is entered, a table is obtained from the GEO, displaying a summary of the clinical data for the dataset. It is possible to conduct differential expression analysis or, if relevant to the dataset, survival analysis by inputting a probe ID of interest and a phenodata column to determine the groups for comparison. The output is the visualization of the log2 expression of the selected gene.

Differential gene expression analysis is not the only method implemented in Shiny applications. NAP: The network analysis profiler [3] is an online tool that automates network profiling and topology comparison. Methylation plotter [4] enables dynamic visualization of user-provided DNA methylation data. In the literature, there are various operations built-in Shiny applications, ranging from categorization to interactive plotting of expression data.

Neuroendocrine tumors and their classification

Neuroendocrine tumors (NETs) are a diverse group of neoplasms that involve various organs and display diverse biological characteristics. Due to the diffuse nature of the neuroendocrine system, NETs vary greatly in their tissue of origin. Nonetheless, they share common pathological features regardless of anatomical site thus have been primarily classified accordingly and further specified with regards to tumor differentiation, hormones or amines secreted, and tumor grade and stage [5].

In some instances, excessive hormone secretion due to neuroendocrine nature of the tumor leads to clinical symptoms and syndromes; this phenomenon is referred to as tumor functionality and has been used in the literature for NETs secreting active bio-compounds [6]. Nevertheless, despite of the fact that functionality is an important factor in prognosis of certain

types of NETs, the biological tendencies of a neuroendocrine tumor is mainly influenced by how far the tumor has progressed in terms of stage and grade. Thus, the diagnosis process for both non-functional and functional NETs that originated from the same organ is identical [5].

In general, NETs are slow-growing tumors. They can originate in various parts of the human body but they are more commonly seen in pancreas, GI tract, lung and some other organs that display endocrine characteristics. As mentioned earlier, it has been discovered that NETs are able to synthesize and secrete certain products; the peptides or amines secreted by the NETs often lead to pathological symptoms and are utilized as tumor markers [7].

Incidence and prevalence of NETs

NETs are seen in every five in a thousand of all cancers [8]. The incidence rate is about two in a million. This number has slowly increased from 19 to 52/100,0000 people per year throughout the last 30 years [910]. Among the tumors of the same organ, incidence of NETs have a higher rate of increase [11]. NETs increase with age and peak between 50-70 years. Modernization of the diagnostic tools and accumulation of empirical data about the disease are implied to be the major factors driving the increase for incidence rate of NETs [12]. Because NETs progress slowly, their incidence correlate positively with their prevalence which has been estimated to be 35:100,000 a year.

NETs are observed with some differences but mostly similarities among different demographic groups. Incidence rates of NETs between men and women are very similar in the U.S. [13], however, there are exceptions to the rule, NETs that arise from rectal sites have a higher rate of occurrence in men than women [14]. Pancreatic NETs and the small intestinal NETs have also slightly higher incidence rates in males but low-grade NETs and lung NETs appear with a higher rate of incidence in women [15].

Pheochromocytomas/Paragangliomas

Pheochromocytomas (PCCs) are rare types of tumors arising in neural crest tissue. When located at extra-adrenal tissue, they are designated as paragangliomas (PGLs). PCCs and PGLs share a common site of origin; they are both derived from neuro-ectoderm. In addition, PCCs are more correctly referred to as intra-adrenal paragangliomas [5]. Pheochromocytomas and paragangliomas show very similar fundamental histological characteristics. Parasympathetic PGLs more often arise from the head and neck tissues whereas sympathetic PGLs could be found in either adrenal medulla (namely PCCs) or at the thoracic, abdominal or pelvic area [16].

.

Neuroblastoma

Neuroblastoma is an early (childhood or embryonal) tumor that arises from the autonomic nervous system, mainly the peripheral sympathetic nervous system. Therefore, the tumor is derived from the neural-crest tissues while its site of origin is a precursor cell that is still developing [17]. Since it is a disease of developing tissues, the affected demographics consist of mostly children as young as 17 months [18]. Typically, sympathetic nervous tissue is the site of origin for these tumors, arising from the paraspinal ganglia and adrenal medulla; the tumors result in mass lesions that could be observed in chest, abdomen, pelvis and neck area. The clinical manifestation of the disease is highly prone to variability. The disease is highly common in very young children, in fact it is the most common cancer type that manifests during the first year of life [19].

Throughout many decades of clinical evidence and research, it has been noted that neuroblastomas are associated with unpredictable and dramatic clinical symptoms [20]. In parallel with its unusual nature, neuroblastoma is the leading cause for mortality among the childhood cancers, but at the same time, it is one of the cancers that has received significant improvement in survival rates in recent decades [21] Data collected by the Surveillance, Epidemiology, and End Results databases report that 5-year survival rates during the period from 1975-1977 to 1999-

2005 increased from 52% to 74%, respectively [20]. However, this improvement is mainly due to better outcome in patients with more benign tumors; the survival or cure rates among the children with aggressive neuroblastoma have seen only a moderate increase, despite the dramatic improvements in therapy technologies and methods [22].

Neuroblastoma has been widely considered to be an aggressive and abnormal representation of the developing sympathetic nervous system tissue, however, until the last decade, there has been a lack of knowledge in terms of the genetic basis of the condition. Similar to other cancer types, a group of cases has been identified to exhibit autosomal dominant inheritance. The results of the study conducted by Mosse et al. showed that in most cases of hereditary neuroblastoma, activation of mutations are observed in the tyrosine kinase domain of anaplastic lymphoma kinase (*ALK*) [23]. In addition, loss-of-function mutations in *PHOX2B* gene is present in children with sporadic or familial neuroblastoma who are also diagnosed with Hirschsprung's disease, congenital central hypoventilation syndrome, or both [24]. Therefore, if a patient has a family history of neuroblastoma or other clinical complications with a high-risk mutation, genetic screening for mutations in both of these risk genes, ALK and PHOX2B, should be considered. In addition, a recent study has implicated chromogranins in blood of neuroblastoma patients with prognosis of the disease suggesting further research into chromogranin family is needed in this cancer type [25].

Microarray Technology

For a long time, molecular biologists could analyze one of few genes at a time since the techniques depended on methods using nucleic acid probes (*in situ* hybridization, Norther blotting, etc.) or antibodies (Western blotting, etc.). The emergence of genomic information and advancement of array-based profiling technologies have made it possible to obtain information about the state of thousands of genes in a single assay. The basis for the microarray concept is as follows; RNA isolation is performed using samples of interest to process the isolated RNA and produce the target sequence which is being tested in terms of abundance or identity. Next, the target binds to the probes that are in a pre-determined formation on the array and consist of sequences that correspond to certain genes. The hybridization taking place between the target

and the tethered probe enables a measurement by quantifying the hybridization affinity for each target. This quantitative data characteristic is obtained via software and processed through many pipelines in order to extract biological information [26]. Microarrays have been primarily used to measure gene expression levels.

An array is sensitive only to the sequences it was design to detect. In other words, if the solution that is to be hybridized contains sequences with no complementary sequence on the platform matrix, it is not possible to detect those sequences using an array platform with that particular set of probes. As far as gene expression analysis purposes are concerned, this basically implies that only genes who are members of previously annotated reference genomes will be present on the array. This can prove especially problematic for genomes with high variability such as bacteria [27]. On the other hand, one critical advantage of microarray datasets is their widespread and frequent usage allowing conductance of meta-analysis to obtain reliable conclusions based on transcriptomics in different species.

Meta-Analysis

Due to the advancement of high-throughput omics data generation technologies and their increasing availability, researchers from almost every medical field are able to utilize these methods. Nevertheless, robust data analysis methods are required to be able to wholly interpret the biological meaning in the massive data. The availability of the microarray technology contributed to the cumulative abundance of transcriptomic data as well as RNA-seq experiments which have been especially significant in the last decade. In order to collect, store and share the growing body of large data publicly, online repositories Gene Expression Omnibus (GEO)[28]and ArrayExpress [29], were established and proved tremendously useful in terms of availability and accessibility of raw data.

More often than not, multiple transcriptomic studies that focus on a similar pathological state are available although, even more often, it is the case that each study is based on a small number of samples, therefore lacking significance for their findings. If the data used in each study were merged, the statistical power of the analysis and the significance of the findings would increase [30]. Nevertheless, it is problematic to compare heterogeneous datasets directly

due to inherent complications in biological and experimental parameters [31].

Meta-analysis, which is composed of statistical tools and approaches to merge data from multiple studies is a practical solution to overcome the abovementioned challenges. Overall, meta-analysis is an invaluable tool for researchers who are interested in investigating existing datasets to identify potential biomarkers and pathways. Arguably, one of the biggest problems of microarray studies is that since there are thousands of probes, each representing a different hypothesis with relatively small number of samples, thus the statistical power of the analysis dramatically drops. Accepting a very conventional positive rate of 0.05 in array platform with over 50,000 probes would mean over 2500 genes are seen as a differential element due to mere chance only [32]. Merging the data from multiple studies provides more robust results so that the false positively significant genes will not be present whereas the true positives can still be observed.

In addition to random noise, another factor that contributes to false discovery is biological and experimental differences in the study design [30]. There is no doubt that some variation in gene expression levels identified by different studies are the result of true biological variation in samples. However, distinct experimental parameters such as use of different media at different concentrations could alter the results even when using identical cell lines. This type of experimental variation holds "true" in the context of those particular experimental conditions but still will not be reproducible in other studies with slightly altered conditions [30]. Meta-analysis helps prevent such experimental factors and presence of outliers from affecting the results and helps unmask the true and consistent significant findings.

While conducting a meta-analysis, it is required to control for additional systematic biases or effects other than biological and experimental factors – sometimes referred to as "lab effect" to emphasize the potential variability stemming from conditions in different laboratories – such as using data from different platforms (platform-effect) and even using samples from various species (species-effect) [33]. Comparison or integration of data across platforms requires an extra step. Due to the many-to-many relationship between genes and probes, in varying platforms the same gene may map to a different set of probes, therefore a common set of

identifiers should be constructed when mapping across platforms. [34]. A robust cross-species analysis is not among the purposes of this article, thus, no further details will be explained on the topic other than that an accurate knowledge of orthology is required and HomoloGene in NCBI and the Inparanoid [35] eukaryotic ortholog databases could be most useful [36].

Previous microarray studies focusing on Pheochromocytoma

There is a number of studies investigating the landscape of transcription profiling in PCC as well as NB. In this section, studies of significance to this article will be explained in brief fashion in terms of their methodology, hypotheses and findings. In addition, raw microarray data are available for almost all of the following studies and, those available, have been utilized for the purposes of this thesis as part of the datasets used for the meta-analysis (see methods for details).

In the study conducted by Dahia et al. [37],the researchers showed that there was a common function shared by three genes thought to be responsible for hereditary PCCs. The authors of the study performed comparative analysis of expression profiling of 76 PCCs and PGLs with a variety of mutations that are representatives of distinct PCC syndromes, sporadic tumors and familial tumors without an identifiable mutation.

Until 2005, the year Dahlia et al. published their results, six independent genes were identified as inherited or sporadic mutation sites: *RET, VHL, NF1* and subunits B, C and D of succinate dehydrogenase (*SDH*). To investigate the molecular interactions between the pathways that these aforementioned genes facilitate, Dahlia et al. conducted a microarray experiment using a large number of PCC samples [37]. The results showed that, proteins encoded by specifically *VHL, SDHB* and *SDHD* were important for the function of the transcription factor, hypoxia-inducible factor 1 subunit alpha (*HIF1*α) playing a facilitative role in adapting cells to low levels of oxygen or hypoxia. *VHL* has been linked to von Hippel-Lindau disease (VHL), the disease it is named after. VHL is an inherited disorder that makes individuals more susceptible to PCCs. It was reported in earlier studies that the absence of *VHL* leads to accumulation *HIF1*α, and the lack of *HIF1*α degradation mimic a state of hypoxia [38].

Neuroendocrine Tumor (NET) Biomarkers

A fundamental and thorough knowledge of the nature of neoplasia structure is required to make further advancements in NET medical care [39]. Nevertheless, another crucial unmet requirement to investigate this class of tumors is the lack of robust biomarkers with proven statistical significance and biological convenience [40]. Even with the diverse spectrum of advancing technologies that contribute to modern medicine and therapeutic tools, detection of a disease state in the very early stages remains crucial. In addition, multi-utility tools that are able to track therapeutic efficacy and assess prognosis are needed to reveal robust biomarkers [41].

Single-analyte biomarkers have proved useful in the identification of distinct features of different disease states. The amount of information that single-analytes convey is not comprehensive enough, failing to explain the diverse nature of NETs [41]. It is required to acquire and accumulate large omics datasets with a wide range of samples to create an understanding of NET diseases. Therefore, single-analyte measurements will become less reliable and the research paradigm is likely to shift towards the use of more comprehensive biomarkers such as multi-analytes. This turn of events represents an innovative step in establishing a system of personalized medicine with the accumulation of multi-dimensional biological data. Similarly, using blood samples to acquire biological information compared to biopsy will enhance the robustness of the information and help avoid the potential risk factors involved in such invasive protocols [41].

Advancements in discovery-based approaches, such as microarray expression studies and proteomic approaches using mass spectrometry, have been critical in the identification of a number of candidate cancer biomarkers [42]. The advent of microarray-based profiling along with its rising availability has tremendously boosted the power to categorize cancers, notably breast tumors [43]. But beyond the improvements on classification, these advances could be utilized for solving fundamental problems in classification, such as presence of metastasis. A problem that often arises in clinical classification of metastatic tumors is determining the primary site of tumor origin. This problem, for example, could be solved by using classification techniques based on

gene expression signatures to predict the origin of the metastatic lesion based on its highest similarity to a tissue type. Accordingly, gene expression can provide more precise and extensive information about a tumor than histology could provide [42].

Scientific consensus on the validation of single-analyte biomarkers, for diagnosis and prognosis of NET, have shifted towards utilization of multi-analyte biomarkers based on the idea that a single-analyte test cannot account for the multidimensional pathophysiology of a condition. In addition, the growing body of interdisciplinary research has contributed greatly to developing multi-analyte biomarkers by providing the necessary statistical power to analyze multi-variable complexes. NETs require secretion of neuroendocrine factors from vesicles and are associated with chromogranins [44]. Accordingly, the multivariate nature of chromogranin assemblies and the degree of divergence in these assemblies in between different cancers could be valuable to discover in the current biomarker field. However, there is no comprehensive study in the literature on this for PCCs and NBs or any other NETs. CgA, the protein encoded by the *CHGA* gene is the major non-specific marker of NETs [44]. CgA levels correlate with malignancy and it has been reported that non-elevated levels of CgA could be an evidence leading to exclusion of PCC among the candidate for the disease state [45].

Biomarker limitations and issues

Ideally, a biomarker should be exclusively related to and representative of the particular disease state of interest while exhibiting an observable signal without background noise from irrelevant disease states. Gene expression profiling has been seen as a tool with great utility but its shortcomings should be evaluated as well. Although a vast number of genes, in scales of thousands, are detected to be highly expressed in malignant tissue compared to benign, a rather limited number of proteins or transcripts have been detected to have elevated states in tumor pathology [46]. This problem does not necessarily make gene expression profiling a less significant tool or obsolete but it shows that enhancements are needed. Development of analytically robust algorithms based on advanced statistics to detect abnormalities that are capable of discriminating among specific types of neoplasia as well as providing information about its pathological tendencies is crucial for this type of optimization [41].

To some extent, economic constraints have also been a discouraging factor in the innovation of reliable markers. Most of the clinically established markers that have been studied are in the public domain and free of intellectual property protection. The absence of intellectual property protection discourages companies to pick up well-studied candidate biomarkers or assays to run clinical tests for further validation. Even protected assays often prove problematic for the company since it is a daunting challenge to develop a high grade, reproducible assay that is well beyond the conventional academical research ambitions to get FDA approval [41]. FDA has given approval for a very small number of biomarkers in the last ten years, and none were NET biomarkers, therefore, currently there are no NET biomarkers for the disease that are also recognized and approved by the FDA[47].

Fundamentally, the robustness of a biomarker could be quantified in terms of its sensitivity as well as specificity. Although, these two terms are sometimes referred to as "accuracy" melding both sensitivity and specificity together; the usage of the term accuracy fails to convey the effect size when sensitivity or sensitivity of a biomarker is lacking. Sensitivity measures the proportion of actual (true) positive cases, whereas, specificity measures the proportion of actual negative cases. Since a theoretically infinite sensitivity measure never fails to spot a false negative and an equally robust specificity measure never fails to classify a false positive, a test with such statistical power that is trying to distinguish between malignant and benign tumors would identify all malignant cases and also never classify an actual benign sample as malignant [31]. However, this type of scenario of perfect sensitivity and specificity is not possible in reality as the Bayes error rate dictates [48].

Chromogranins: structure and function and evolution

Chromogranins, sometimes referred to as secretogranins or granins, are a class of acidic, secretory proteins. The granin family is composed by seven members; the three main members of the family are chromogranin (CgA), chromogranin B (CgB) and secretogranin II (SgII) or otherwise named as chromogranin C (CgC). In addition, there are SgIII, VGF, 7B2 (also referred to as secretogranin V), NESP55, and proSAAS which are related proteins and together with the main proteins they form the granin family. However, the members and the exact function of the granin protein family is still up to debate and also is of interest for the aims of this thesis. Granins are found in a diverse range of neuroendocrine cells, inside the secretory granules [49] [50]. The granin family proteins are a major part of the regulatory scheme underlying the secretory pathway that facilitates neurotransmitters, hormones, growth factors and regulated delivery of peptides [51].

Phylogenetic analyses of CgA, CgB and SgII sequences showed that monophyletic groups were generated when using different sequences for alignment; and these groups were consistent and displayed precise classification for each granin [49]. These results also identified that CgA and CgB precursors together compose the two monophyletic groups that are related, implying that it is most likely the case that they have diverged from an ancestral common precursor , on the other hand, SCG2 sequences represent a distinct monophyletic group [52].

The human CgA, Cgb and SgII are cleaved at multiple sites by a spectrum of proteases, at 12, 18 and 9 pairs of basic aminoacids in CgA, CgB and SgII, respectively [53]. Granins play a functional role in various processes. In addition to their most abundant role in the formation of granules , they facilitate important functions in costorage and corelease of catechloamines [50], regulating mechanisms of response to microbes [54], providing vasculature integrity [55], forming new vessels in wound healing [56], and inflammatory diseases [57].

2. Aims

Techniques that generate high-throughput data have become mainstream in biological sciences. Since the emergence of these techniques, it has become very difficult to conduct biological research without utilizing massive datasets. However, handling such magnitudes of biological data can be very computationally demanding, especially for biology-oriented researchers. It could be unfair and unrealistic to expect all biologists to be excellent bench researchers and expert programmers at the same time. Therefore, computational assistance in some form is needed to address the computational demands that arise from analyzing high-throughput biological data.

A common response to this problem has been the assembly of interdisciplinary research teams and collaborations. The idea is that, the complementary skill-sets that the team members possess would combine in such a way that all aspects of research roles are covered. In reality, this theory does not always hold true; some computational analyses are needed to be conducted with such frequency and emergency enabling all or most members to conduct them is essential to increasing efficiency. Therefore, setting up an internal infrastructure that makes it possible to conduct multiple analyses for multiple users makes a lot of sense. In addition, it is always preferable to depend on a program rather than a person, to perform complex repeated tasks to eliminate errors and elevate precision.

Previously in our lab, Huma Shehwana compared expression network of two paralogous genes, mineralocorticoid receptor (MR) and glucocorticoid receptor (GR), using a *Shiny* application that she has develop focusing on multiple breast cancer microarray datasets [58]. In addition, she has performed meta-analysis using *metacor* program in R to obtain a more reliable measure of MR and GR correlated genes. However, the *Shiny* app was designed to accommodate breast cancer datasets and it did not allow analysis of any other collection of expression datasets and in a semiautomated fashion. In addition, there is still a need in the literature to automatize the widely used WGCNA program [59], explained in more detail in methods section, for network module analysis using a *Shiny* app.

Accordingly, the specific purposes of this thesis are as follows:

1. Design and generate a *Shiny* application called CoEX-PM using the R programming language that facilitates analyses using available microarray data (GPL96 and GPL570 platforms) from the GEO database to analyze expression of gene pairs or modules. To accomplish this task, the application will be designed so that the users are able to construct correlation plots of co-expression vectors and meta-correlation analysis of co-expression correlation for selected pairs of genes as well as perform network module analysis using WGCNA with user-selected genes, based on GEO datasets and other changeable parameters on a reactive user interface using *Shiny*.
2. Demonstrate a case-study utilizing CoEX-PM for NET datasets acquired from the GEO database, mainly focusing on PCCs and NB, to identify chromogranin driven gene networks and functional modules underlying the disease state. Chromo/secretogranins have been used as biomarkers for neuroendocrine tumors. However, their co-expression patterns in these cancers and how they diverge from each other is not known well. The outcome of this thesis will help clarify which chromo/secretogranins are likely to be expressed similarly or divergently and whether there are differences in between PCC and NB. Assessment in lung and Merkel's cancers will also be performed. In addition, in the case study I will use STRING protein-protein interaction database and association functional analysis tools to identify the functions of the chromo/secretogranin associated co-expression networks in different cancers.

3. Methods

3.1. Data Acquisition and Normalization

The datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) repository [28]. The datasets and their contents were explained in Table 1. For the purposes of this study, raw data files were required to conduct the analyses within the *CoEX-PM* Shiny application. The acquisition of the raw data was accomplished by using the *GEOquery* package and more specifically by the *getGEOSuppFiles()* function. Given the dataset ID (e.g. GSE2841 in platform GPL570), the function simply downloaded the supplementary files provided by the author. In the case that the raw data were available in GEO, they were downloaded in. CEL format, usually archived as a .tar file.

The normalization step is crucial to account for the technical variation between arrays. In this thesis, I have used the Robust Multi-array Average (RMA) method to normalize the raw data, specifically, the *rma* function from the *affy* package. There are different rma functions in several different R packages, the user should pick the most convenient one to use with regards to the data or R object type the user is operating on. The *affy* package enables the manipulation and analysis of Affymetrix data at the probe level [60]. RMA is a normalization method widely used for microarray data pre-processing. an algorithm utilized to construct an expression matrix from Affymetrix data. The steps in RMA normalization are as follows in order; background correction, log transformation and quantile normalization (across arrays), probe level intensity calculation by fitting a linear model to the processed data and probe set summarization. Since it works on all arrays simultaneously, RMA uses a lot of memory. However, this relatively high-memory usage is no longer considered a downside of the method because modern computers possess high enough computational power.

Table 1: The datasets that were acquired from the GEO database and used in further analyses.

Pheochromocytoma	**Study**	**Platform**	**Sample Number**	**Sample Type**
GSE2814	Dahia et al. [37]	GPL96	76	PCC and paragangliomas
GSE67066	Unpublished	GPL570	40 B – 11 M	Benign and malignant PCC
E-MTAB-733(ArrayExpress)	Burnichon [61]	GPL570	188	PCC
Neuroblastoma	**Study**	**Platform**	**Sample Number**	**Sample Type**
GSE13136	Lastowska et al. [62]	GPL570	30	Neuroblastoma
GSE16476	Molenaar et al. [63]	GPL570	88	Neuroblastoma
GSE73537	Valentijn et al. [64]	GPL570	34	Neuroblastoma
GSE16237	Ohtaki et al. [65]	GPL570	51	Neuroblastoma
GSE12460	Janoueix-Lerosey et al. [66]	GPL570	52 neuroblastomas, 8 ganglioneuroblastomas, 3 ganglioneuroma, 1 unknown	Neuroblastoma and other
WGCNA	**Study**	**Platform**	**Sample Number**	**Sample Type**
GSE67066	Unpublished	GPL570	40 B – 11 M	Benign and malignant PCC
GSE16476	Molenaar et al. [63]	GPL570	88	Neuroblastoma

3.2. *Shiny* Application Design

There are multiple ways to define a *Shiny* application, meaning that it could be formatted in multiple ways. A conventional *Shiny* application is made up of two parts; a server side and a user interface (UI) side [67]. The UI part contains the information to be shown and parts for user interaction, such as a numeric input box or a boxplot. The server part contains the code that does the computations needed to construct the outputs of the application. Depending on the type of the input/output, varying UI elements are available for use, for instance, using the *renderUI()*

command in the server side enables the construction of reactive inputs, according to the previous input given by the user.

There are also numerous ways to structure the layout of a Shiny application. A fairly simple option could be to have a sidebar for inputs and main area for output or a more complex and customizable layout could be accomplished by using the *Shiny* grid layout system. Another option is to build a segmented structure using the *tabsetPanel()* and *navlistPanel()* functions. In addition to the tab-based segmentation method, applications with multiple top-level components can be accomplished with the *navbarPage()* function. The *CoEX-PM* application in this thesis uses a hybrid of the aforementioned methods. There are three top-level components that are segmented into multiple tabs individually as described in the following sections.

3.2.1. Tab 1: Pairwise correlation analysis.

Correlation between any two genes can be calculated based on expression of mRNA levels obtained from a high-throughput study. This correlation can be obtained either in the form of parametric or nonparametric formulas, Pearson's or Spearman's correlations respectively [68]. The correlation coefficient reflects the degree of similarity in direction and/or relative magnitude of change from point to point (e.g., time to time, tissue to tissue). A positive correlation between a pair of two genes indicates the degree of similarity in co-expression whereas a negative correlation may indicate opposing actions taking place on the expression of the genes.

In this tab, the aim has been to take as input the names of the genes of interest along with the GEO ID of a dataset before plotting the correlation between expressions of genes using the *cor()* and *cor.test()* functions in tandem in R. The *cor()* function has been used to calculate the correlation of a particular probe to all the other probes on the array, whereas the *cor.test()* function was utilized to establish an association between paired samples of the correlation coefficient vectors of the two genes of interest. The *cor()* function was the core element of a custom function that takes a gene symbol and a data frame as input returning a two-column data frame with the correlation co-efficient *r* and p-value as columns, respectively. For plotting purposes, generic plotting functions of R and relevant *ggplot2* functions were used [69]. In addition, *ggsave()* function has been used to save plots in a preferred directory on the local machine.

3.2.2. Tab 2: Meta-analysis Tab

In *CoEX-PM* I have made available meta-analysis of a correlation coefficient between expression of a gene of interest with that of each of the remaining genes in the expression analysis platform from a selected set of GEO datasets. In the previous tabs the required datasets must have been acquired into the *Shiny* work environment so that the Meta-analysis tab can work.

The main R function used in this module/tab is *metacor*. This function has been deposited in the CRAN repository [70].

Metacor package

There are two core functions of the *metacor* package, *metacor.DSL()* and *metacor.OP()*. For this study, I have used the *metacor.DSL()* function which implements the DerSimonian-Laird (DSL) random-effect meta analytical approach where correlation coefficients are treated as effect sizes [71]. The parameters of the function have been designed as follows:

metacor.DSL(r, n, labels, alpha = 0.05, plot = TRUE, xlim = c(-1, 1), transform = TRUE)

where *r* is a vector correlation and *n* is a vector of sample sizes. The function returns a number of values (mainly the mean effect sizes z.mean/r.mean and their confidence intervals) which are then structured into a data frame object to make it easier to manipulate and visualize the data. The construction of the data frame object as well as some other pre-processing and data acquisition steps have been done by a custom function written by Dr. Huma Shehwana, an alumna of our lab [58].

3.2.3. Tab 3: WGCNA Tab

Weighted gene co-expression network analysis (WGCNA) is a robust method for identifying biologically relevant expression patterns in highly multivariate and complex data, such as microarray datasets [59]. Conventionally, network building methods quantify the pairwise correlations for each gene individually but WGCNA also quantifies the extent to which these pairs of genes share the same neighbors. As a result, a topological overlap matrix is created. The overlap matrix is converted to a dissimilarity matrix which is then subjected to hierarchical clustering [59].

This line of computation results in a dendrogram which clusters genes with similar expression patterns into distinct branches and most highly connected nodes or “hubs” are located the branch tips. Furthermore, WGCNA provides ways to cluster these branches into separate “modules” that contain a group of genes with highly covarying expression patterns across the dataset. Each module contains an “eigengene” element corresponding to its expression pattern and genes that are highly correlated with the module eigengene are referred to as “hub” genes. Importantly, by enabling additional genetic, developmental or phenotypic traits to be associated with tens of modules, rather than tens of thousands of variables, WGCNA resolves the multiple testing problem as well as providing a direct way to detect experimental treatment effects in structured networks.

In *CoEX-PM* I have made available WGCNA analysis to be performed on the selected GEO datasets and identified the modules and found the co-expression clusters associated with the selected genes before performing the relevant plotting. In addition, the main R package to be used is the *WGCNA* package coupled with the *limma* package for the pre-processing step of identifying differentially expressed genes [72] [59].

I have incorporated WGCNA module into the *CoEX-PM* using a modified version of a conventional WGCNA workflow documented by the authors of the package [59]. First of all, to select the genes to be included in the WGCNA, a differential expression analysis has been conducted. As a result of this analysis, top 10K genes was selected to be included in the WGCNA analysis. This type of gene filtering is required since WGCNA operates optimally working with 10K-20K genes, also any number of genes higher than this band requires a lot of computational power. Therefore, it is not advisable and, in most circumstances, impossible to conduct WGCNA with the whole probe set of a microarray platform.

4. Results

4.1. *COEX-PM* Application

4.1.1 Gene-Correlation Tab

The *CoEX-PM* application has been uploaded online in the GitHub repository hosting service. It is freely available in GitHub where there are user's guide documentations as well, explaining the work flow in detail.

The main screen of the *CoEX-PM* is provided in Figure 1. The three tabs in this module can be accessed by clicking on the tabs. For exemplary purposes the dataset GSE67066 was selected and uploaded for analyzing one of the NET datasets already have been deposited in the application. As it is possible to select an existing dataset (PCC or Neuroblastoma options) it is also possible to upload a new dataset to analyze using GSE ID number from GEO [28]. Before analysis of gene pairs this module performs a hierarchical clustering based on *hclust* function using the sample files starting with GSM allowing the user to select and filter out outliers, if any, from the whole dataset. This can be done by selecting a cut-off point on the branch lengths using *cutreeStatic* function in R. In addition, the boxplots of normalized expression values can be visualized for quality control of the normalization procedure applied (Figure 1).

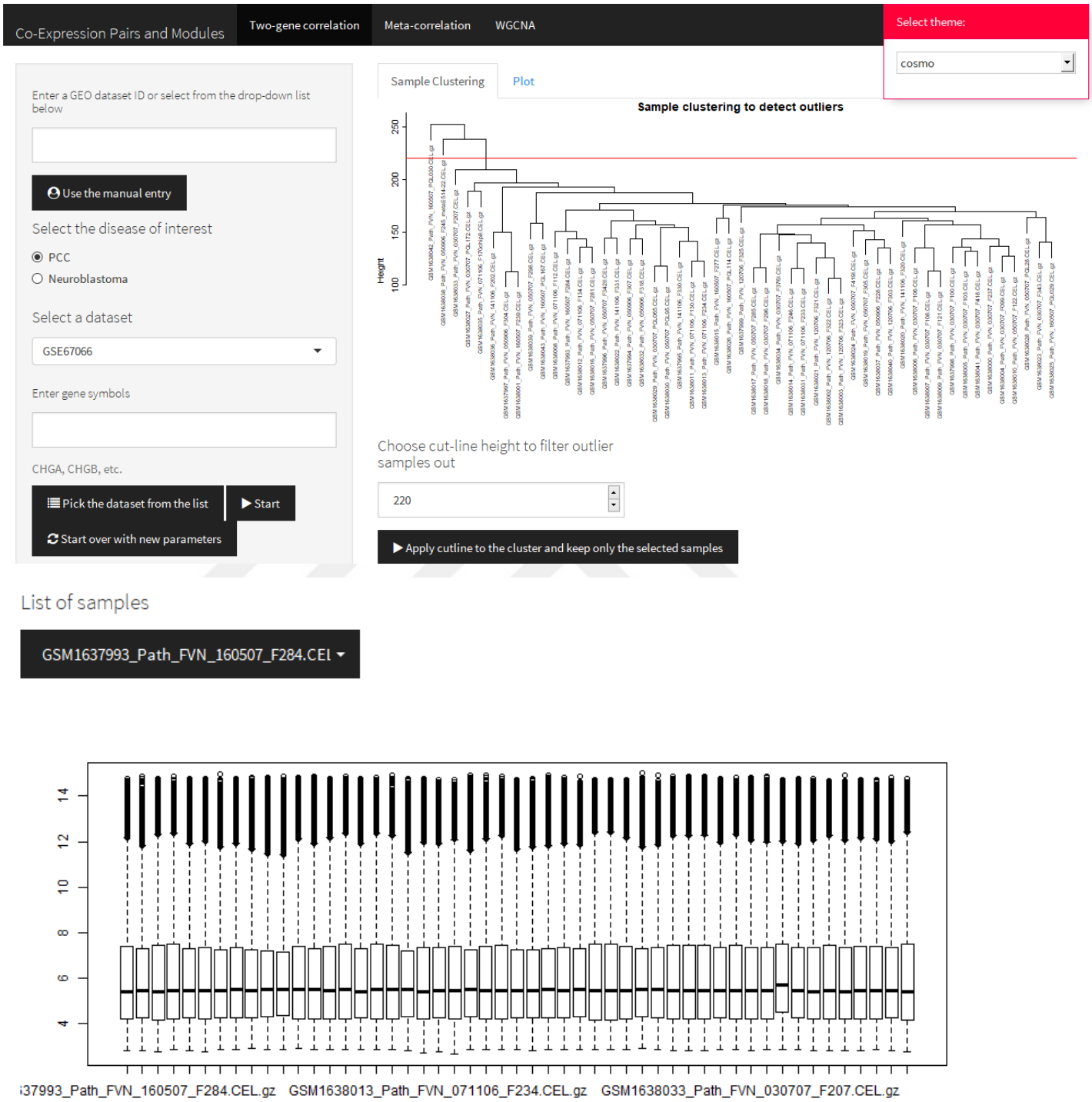


Figure 1:The screenshot of Sample Clustering part of the Two Gene Correlation Tab.

Once the clustering has been done and samples selected the user can perform the plotting of the genes (Figure 2). This can be done by clicking on the plot tab. On the left panel the gene names to be entered into the plot tab are provided with a '-' sign in between. As many genes as the user wants can be entered (Figure 2). Once the gene names are uploaded the program allows selection of two of them for doing a scatterplot between the correlation coefficients obtained from the expression values for the selected genes. This page provides relevant information about the procedures, and data used for the plotting, i.e., the probeset IDs, the GSE number, number of samples the dataset contains. In addition, there is an option for coloring using either the p-values or the expression values (mean expression for each probe across all selected samples) based on significance or intensity, respectively. This coloring has been done using the function *scale_colour_gradient2()* which creates a diverging color gradient (low-mid-high). When the *"Construct the correlation plot"* button is pushed a scatter plot is generated by using the *ggplot()* function in R. At this point it is important to emphasize that the points on the plot represent the correlation coefficients of selected genes with all the other probes on the array platform, not the signal intensity levels of selected genes. A vector for each inputted gene is created, containing the correlation values with the other genes and these two vectors are visualized on the scatterplot. Overall, the first tab investigates user-inputted gene pairs in terms of how correlated their correlations with other genes are. A high r value would mean that the genes have a similar correlation pattern in terms of their correlation with other probes and vice versa.

The plot is interactive and this is established by the use of two Shiny functions; *nearPoints()* and *brushedPoints()*. By adding click and brush input objects to a specific plot's output command, it is possible to print out the clicked and brushed regions on a plot. When clicked on a point on the plot, *nearPoints()* function searches the nearby pixels for points on the clicked spot, then returns a list of the relevant points. Similarly, *brushedPoints()* function searches inside the brushed region for points on plot and prints out the relevant points. For the *CoEX-PM* application, r and p values of each clicked-on and/or brushed points are printed out (Figure 2: Plot part of the Two-Gene Tab). Lastly, "*the brushed points table*" can be downloaded by pushing the download button at the bottom of the screen.

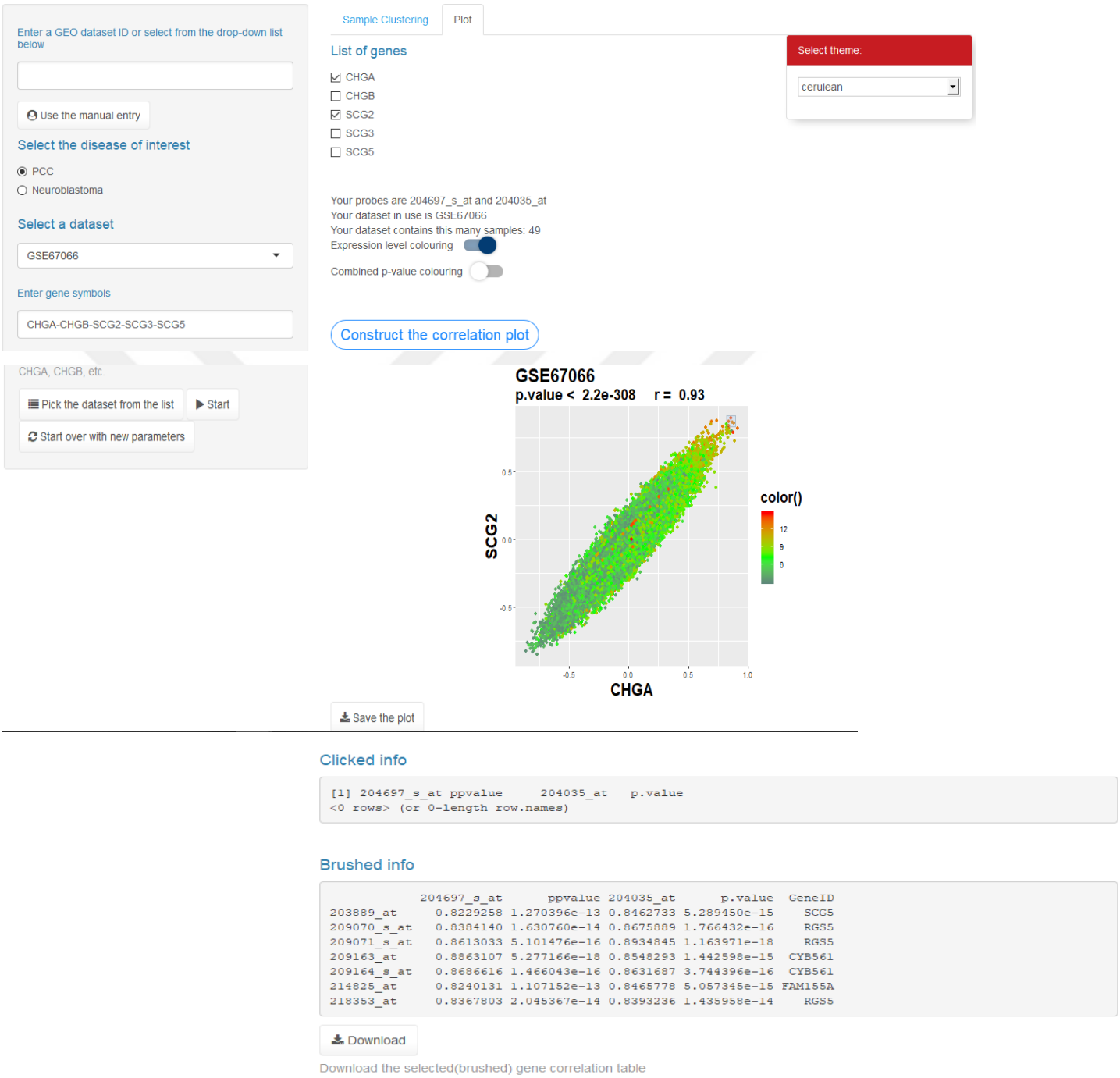
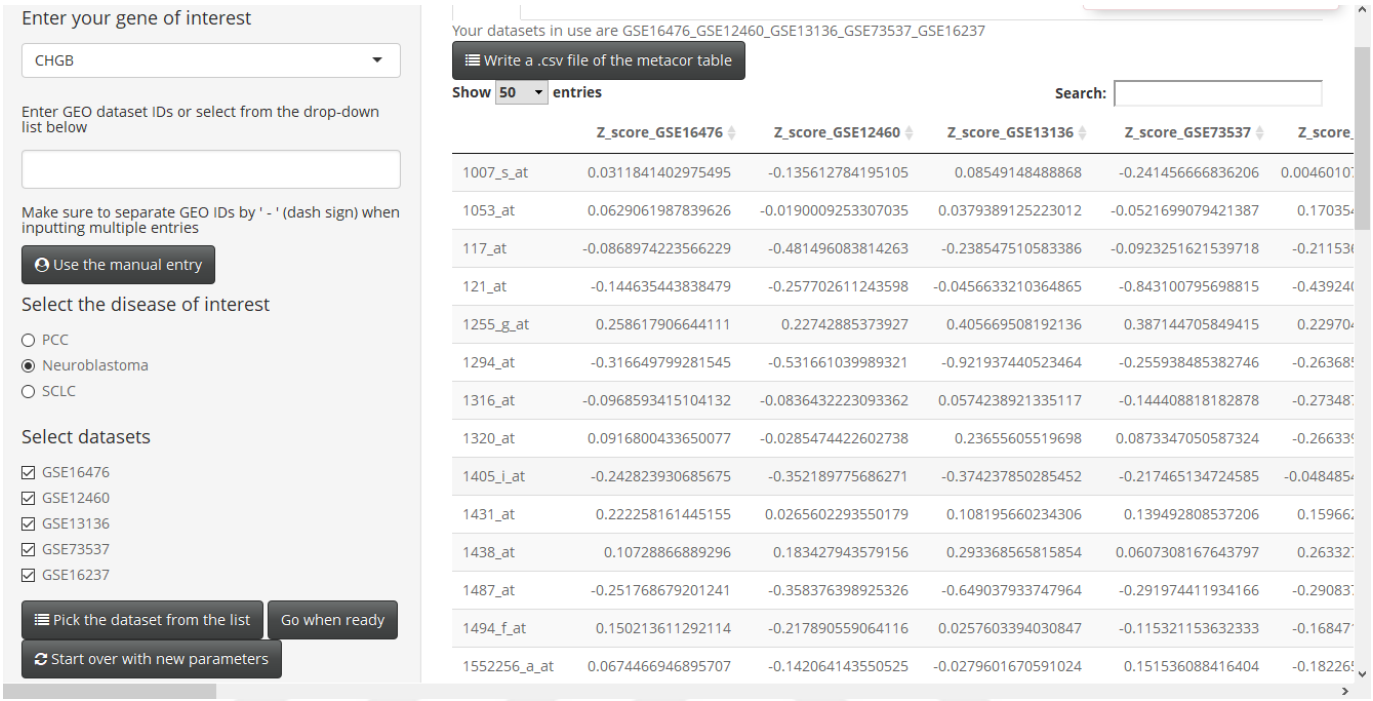


Figure 2: The screenshot of Plot part of the Two Gene Correlation Tab.

4.1.2. Meta-correlation Tab



	Z_score_GSE16476	Z_score_GSE12460	Z_score_GSE13136	Z_score_GSE73537	Z_score_
1007_s_at	0.0311841402975495	-0.135612784195105	0.08549148488868	-0.241456666836206	0.0046010
1053_at	0.0629061987839626	-0.0190009253307035	0.0379389125223012	-0.0521699079421387	0.17035
117_at	-0.0868974223566229	-0.481496083814263	-0.238547510583386	-0.0923251621539718	-0.21153
121_at	-0.144635443838479	-0.257702611243598	-0.0456633210364865	-0.843100795698815	-0.43924
1255_g_at	0.258617906644111	0.22742885373927	0.405669508192136	0.387144705849415	0.22970
1294_at	-0.316649799281545	-0.531661039989321	-0.921937440523464	-0.255938485382746	-0.26368
1316_at	-0.0968593415104132	-0.0836432223093362	0.0574238921335117	-0.144408818182878	-0.27348
1320_at	0.0916800433650077	-0.0285474422602738	0.23655605519698	0.0873347050587324	-0.26633
1405_i_at	-0.242823930685675	-0.352189775686271	-0.374237850285452	-0.217465134724585	-0.04848
1431_at	0.222258161445155	0.0265602293550179	0.108195660234306	0.139492808537206	0.15966
1438_at	0.10728866889296	0.183427943579156	0.293368565815854	0.0607308167643797	0.26332
1487_at	-0.251768679201241	-0.358376398925326	-0.649037933747964	-0.291974411934166	-0.29083
1494_f_at	0.150213611292114	-0.217890559064116	0.0257603394030847	-0.115321153632333	-0.16847
1552256_a_at	0.0674466946895707	-0.142064143550525	-0.0279601670591024	0.151536088416404	-0.18226

Figure 3: The screenshot of Meta-correlation Tab

The meta-correlation tab does not have sub-tabs unlike the other two main tabs of the application; the two-gene correlation tab and the WGCNA tab. The core output of this tab is the meta-correlation analysis table which is interactable in RStudio or a web browser (Figure 3). The table can be sorted by any selected column by clicking on the column name. In addition, there is a search box that enables quick filtering of the table. Especially for the content of the "SYMBOL" column, the search box proves useful. The table result can be saved locally by clicking the button on top of table "*Write a .csv file of the metacor table*". Consequently, the table is saved in .csv format which is a flexible format to investigate in multiple platforms.

There is a preset list of GEO datasets that the user can choose from or the user could enter a dataset ID manually. The user pushes either the "*Use the manual entry*" action button to make the manual entry the working dataset list or the "*Pick the datasets from the list*" action button to designate the preset list of datasets as the working dataset list. Once the gene of interest is also entered, the analysis can be started by clicking on the "*Go when ready*" button. An initial meta-correlation analysis takes a certain amount of time due to downloading, pre-processing and meta-

analysis of fairly large datasets. Once a dataset is used, it is saved locally in a format that makes it faster to use the same dataset in another analysis. Overall, the meta-analysis results are presented in a tabulated form in which the first column refers to the probeset ID and the *metacor* program results are provided for each probeset (Figure 3).

4.1.3. WGCNA Tab

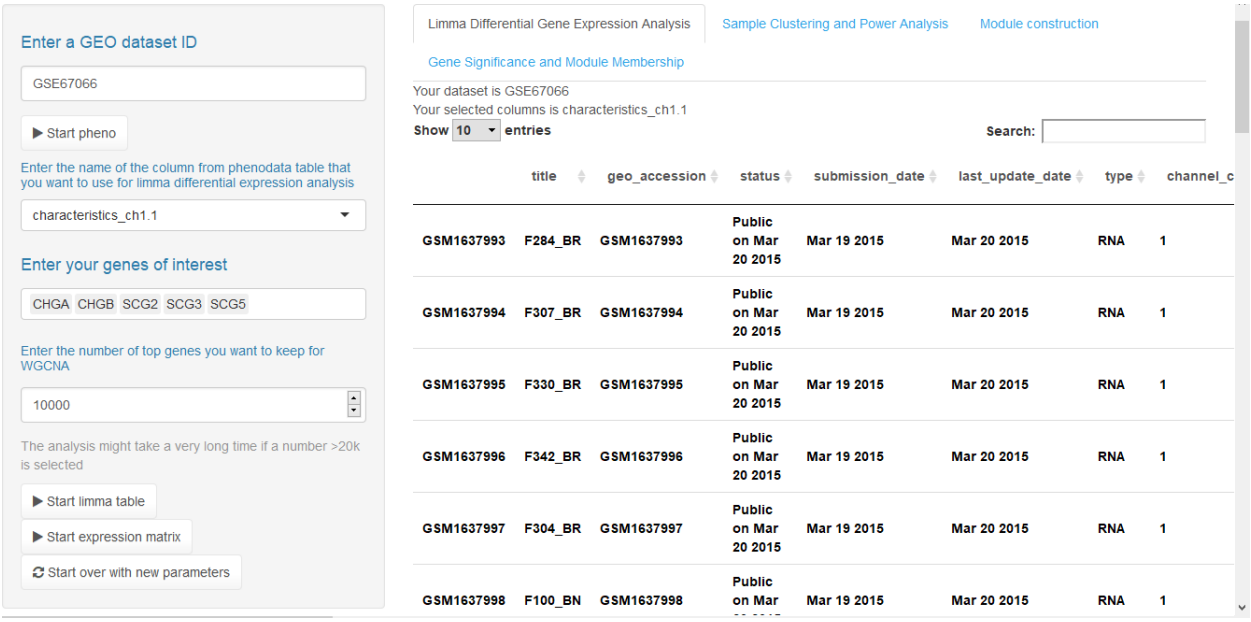


Figure 4: The screenshot of WCGNA Tab: Limma Differential Gene Expression Analysis – Phenodata table.

Show 25 entries Search:

	logFC	AveExpr	t	P.Value	adj.P.Val	B	SYMBOL
242186_x_at	-1.5061056070961	3.76106376355507	-6.35642111894037	5.55798293960962e-8	0.00303882717223156	7.82419712982871	ADGRL3
229638_at	-1.89489779556868	6.56861448098528	-5.96528828869236	2.29332286055843e-7	0.00544944586877192	6.58259535349604	IRX3
236264_at	-2.09018002253594	4.55926002182471	-5.8807465205321	3.11102244553716e-7	0.00544944586877192	6.31490979526664	ADGRL3
1552897_a_at	-1.61745652624853	4.54462985704158	-5.81185427262148	3.98679167354142e-7	0.00544944586877192	6.0970580822381	KCNG3
219908_at	-0.905201862028775	3.64729224472608	-5.61790184112745	7.994269877882e-7	0.00771125583441195	5.48544248182425	DKK2
218309_at	1.4309678375904	11.5301323258558	5.55514937157886	0.00000100033944350331	0.00771125583441195	5.28819912521105	CAMK2N1
235342_at	2.65264932671139	10.7819742033321	5.5412600408286	0.00000105116334076147	0.00771125583441195	5.24459055512948	SPOCK3
209866_s_at	-2.36122246658584	5.11080410371967	-5.48997592555505	0.00000126196509551744	0.00771125583441195	5.08373366028155	ADGRL3
231808_at	-1.1347329917459	6.4689412862954	-5.45984929326143	0.00000140477681321807	0.00771125583441195	4.98936188034777	OIP5-AS1
201131_s_at	-1.13108375026427	3.80927575440009	-5.45649321040685	0.00000142164380862694	0.00771125583441195	4.9788547746433	CDH1
214204_at	1.76661535245799	7.54268540072036	5.3994473061921	0.00000174098471863974	0.00771125583441195	4.80044317113782	PACRG
209867_s_at	-1.76336027002019	5.65264270628332	-5.37010996566858	0.00000193185937072918	0.00771125583441195	4.70883191944637	ADGRL3
230560_at	-2.65422608587618	4.43350518338117	-5.34094781117887	0.00000214206236044956	0.00771125583441195	4.61786736548647	STXBP6
235182_at	-1.42491786965793	4.48800920547351	-5.33963243366612	0.00000215205864347473	0.00771125583441195	4.61376674659262	ISM1
63305_at	-0.824544500310589	4.99665656475174	-5.3378908831698	0.00000216536459189903	0.00771125583441195	4.60833787817556	PKNOX2

Figure 5: The screenshot of WCGNA Tab: Limma Differential Gene Expression Analysis – limma table.

Array platforms contain tens of thousands of probes and, more often than not, a great number of genes that correspond to these probes might be irrelevant to the researchers. In addition, WGCNA pipeline computationally struggles when given a number of genes higher than 10-20K. Therefore, I have implemented a proven method of filtering genes by using the *limma* package. A straight-forward differential gene expression analysis has been conducted prior to WGCNA. Once a dataset ID has been entered, upon the activation the of the "*Start pheno*" action button, an interactive data table containing phenotypic data elements of the inputted dataset appears on the main panel. This table enables a quick examination of the phenodata to pick a column that is to be used as a grouping factor element for the *limma* analysis (Figure 4). After selecting a column and clicking on the "*Start limma table*" button on the side panel, *limma* analysis results are shown on the main panel (Figure 4). Similar to the phenodata table, the *limma* table is also interactable and searchable with a search box (Figure 5). The user is able to keep the preferred number of top genes (in terms of p-value) for the WGCNA. Lastly, a list of genes of interest are entered which are to be related to the modules in the following analyses. Once the aforementioned processes have been completed, clicking the '*start expression matrix*'

action button, the user can move on to the next tab. At this point, the raw data have been normalized, filtered semi-automatically by differential gene expression analysis and user input and a list of genes of interest have been obtained for later steps.

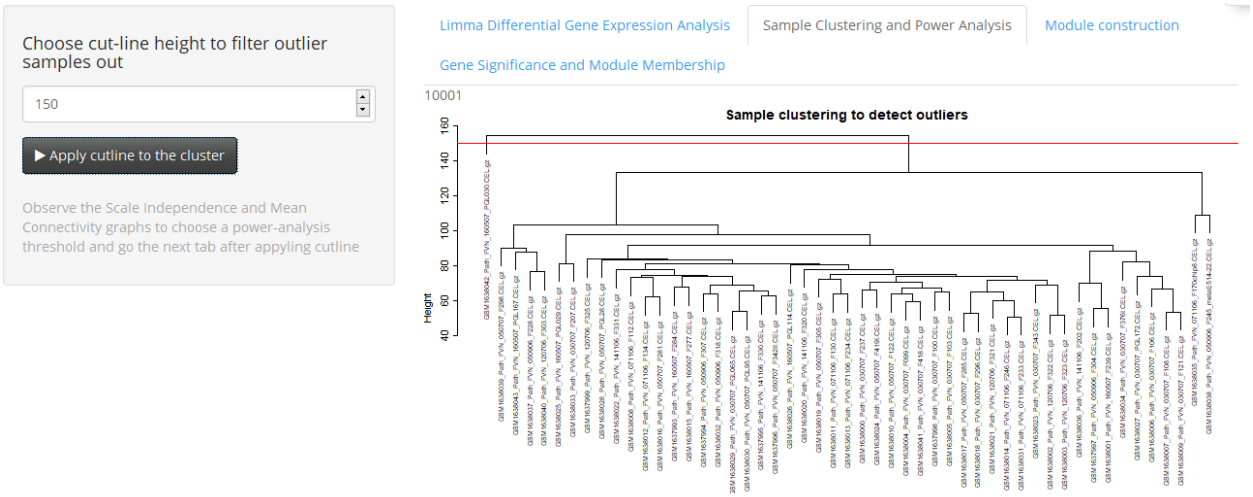


Figure 6: WGCNA Tab: Sample Clustering and Power Analysis – Sample clustering scheme

A filtering at the probe level has taken place up to this point but now the samples are clustered to detect any outliers. A cut-off line height can be used to omit the outlier samples (Figure 6). After the numeric input is given and "*Apply cutline to the cluster*" action button is activated, network topology analysis plots are created. One of the parameters needed to be determined is the soft thresholding power β to which co-expression similarity is raised to calculate adjacency [59]. WGCNA package documents suggest that the soft thresholding power could be selected based on the approximate scale free topology. For this, two plots, scale independence and mean connectivity, can be used to decide on the threshold (Figure 7). This is important since selecting a precise power threshold is required to optimize the sensitivity and specificity of this analysis (Figure 8).

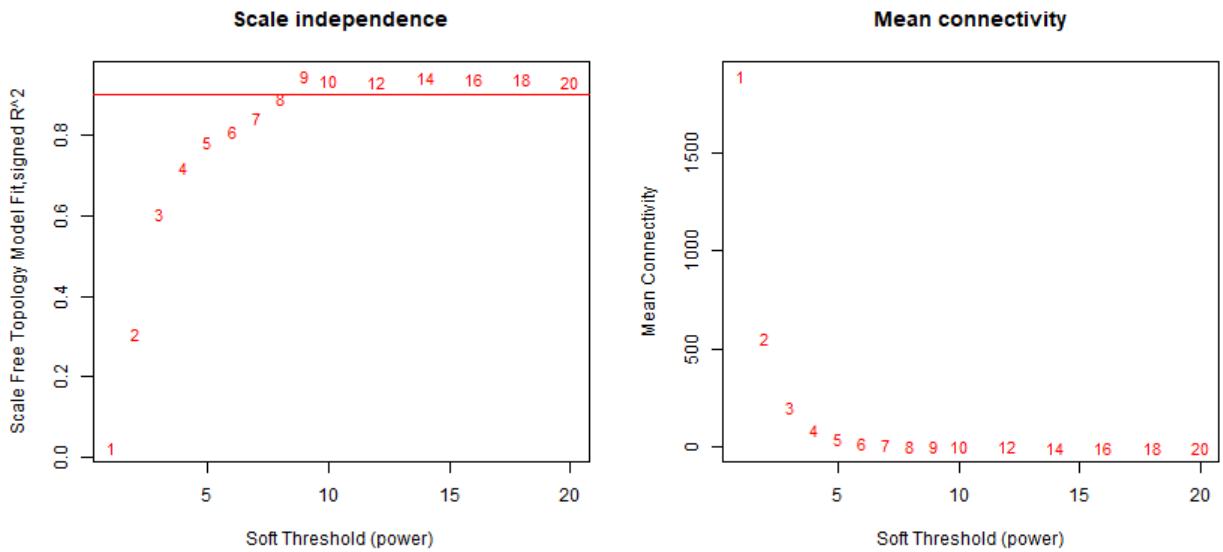


Figure 7: Network topology analysis plots for various soft-thresholding powers.

Modules are generated by WGCNA and module colors are shown below the clustergram (Figure 8). The tabs in the WGCNA module also include graphics for gene significance and module membership (Figure 8).

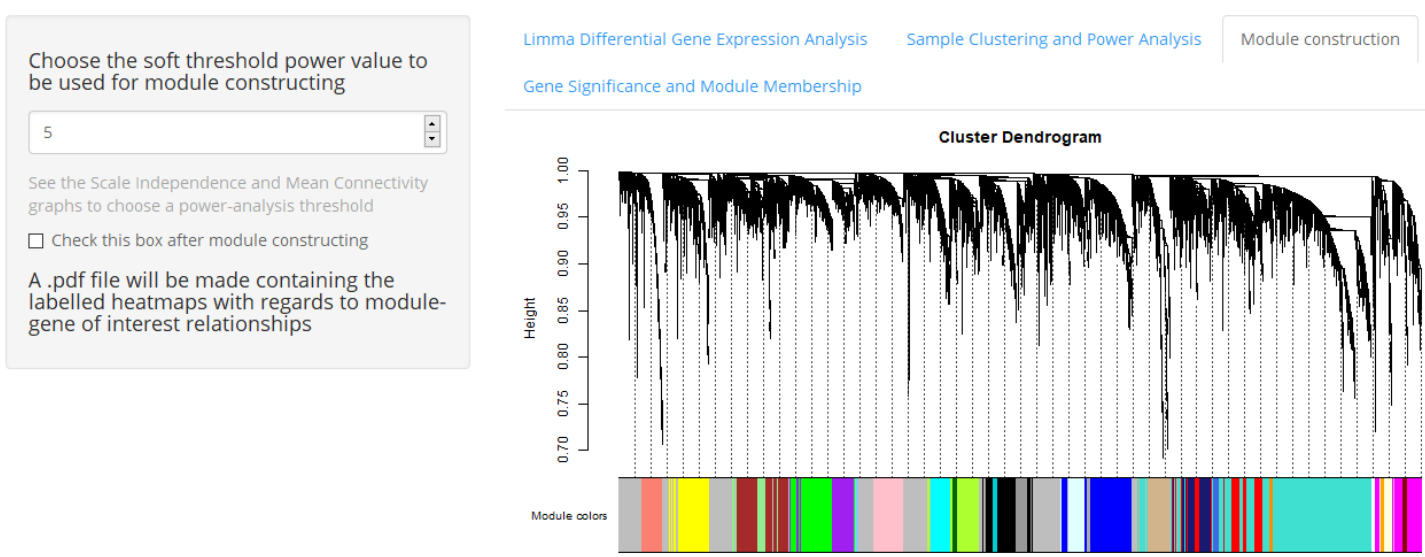


Figure 8: WGCNA Tab: Module Construction – Module dendrogram of genes, each color represents a different module, grey color is reserved for genes that are not a member of any module.

4.2. Case-study: Chromogranins in NETs

As explained in detail in the introduction there are up to seven related chromogranins/secretogranins, i.e., CHGA, CHGB, SCG2, SCG3, and SCG5, in the human genome. These proteins are transcribed from different loci and can form more complex structures post-translationally and take roles in secretion of molecules, including neurotransmitters. One of the primary aims I have focused in this thesis has been to find out whether the chromogranin family members are co-expressed differentially in different cancers in which there are neuroendocrine subtypes.

4.2.1. Pairwise correlations between secretogranin-chromogranin genes

These five secretory components, SCG2, SCG3, SCG5, CHGA and CHGB, have been studied in different GSEs. The expressions of CHGA and CHGB with each other (Figure 9) and with the expression of each of SCG2, SCG3, SCG5 have been identified using *CoEX-PM* (Figure 10 for GSE67066; Figure 12 for GSE16476; Figure 14 GSE12667; Figure 13 GSE13136; Figure 11 GSE2841).

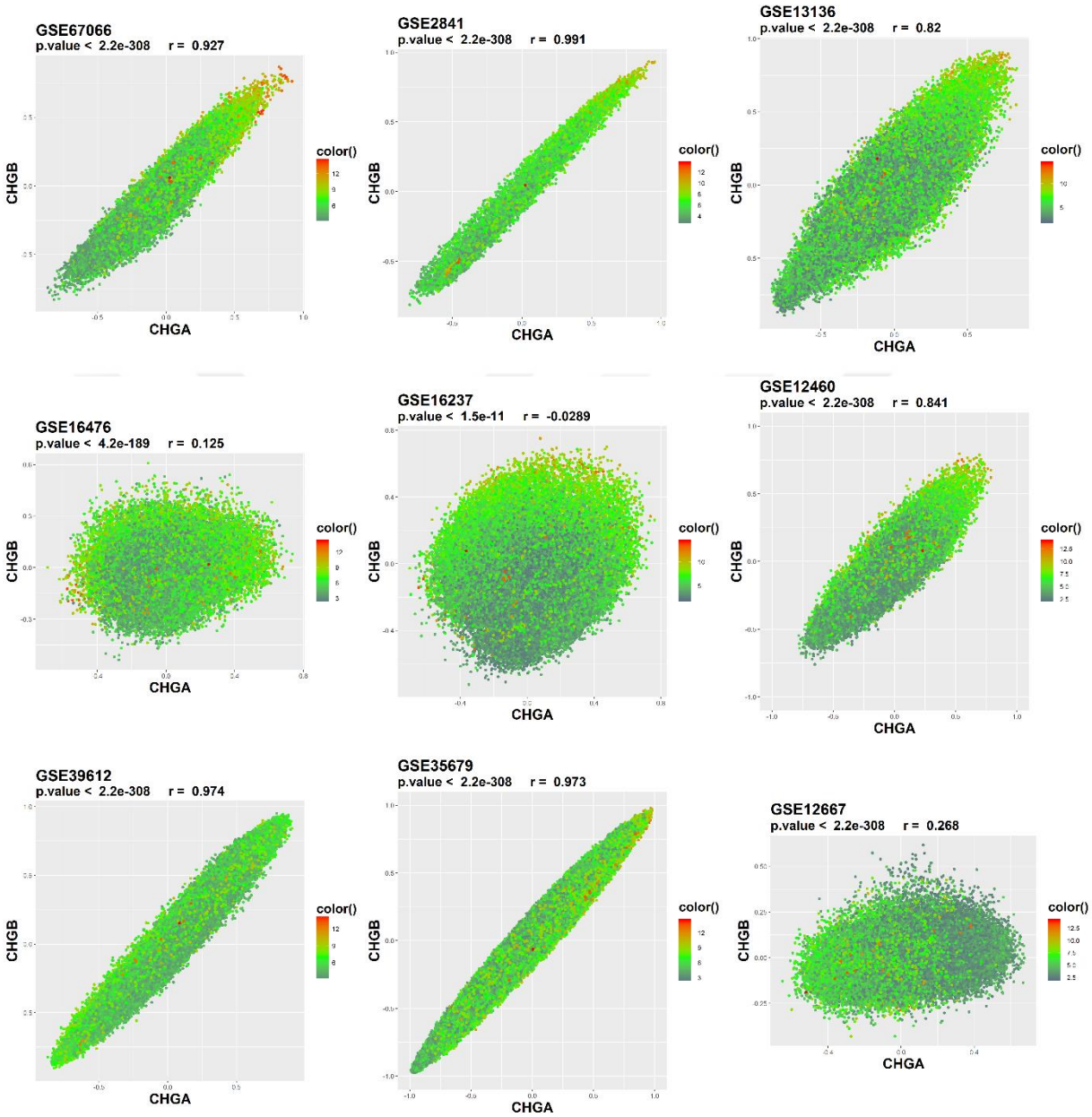


Figure 9: CHGA-CHGB correlation plots for nine distinct datasets. GSE67066, GSE2841 – PCC; GSE13136, GSE16476, GSE16237, GSE12460 – neuroblastoma; GSE39612 – Merkel cell carcinoma; GSE35679 – lung carcinoid, GSE12667 – lung adenocarcinoma.

CHGA and CHGB genes were highly correlated in most of the datasets examined except GSE16476 and GSE16237, two of the neuroblastoma datasets (Figure 9). Interestingly neuroblastoma dataset GSE16476 was one of the less correlated datasets suggesting that these

two highly paralogous genes might have diverged in their function based on their co-expression with different sets of genes (Figure 9). In the other datasets in which CHGA and CHGB were expressed the correlation between expression patterns of these genes were sustained.

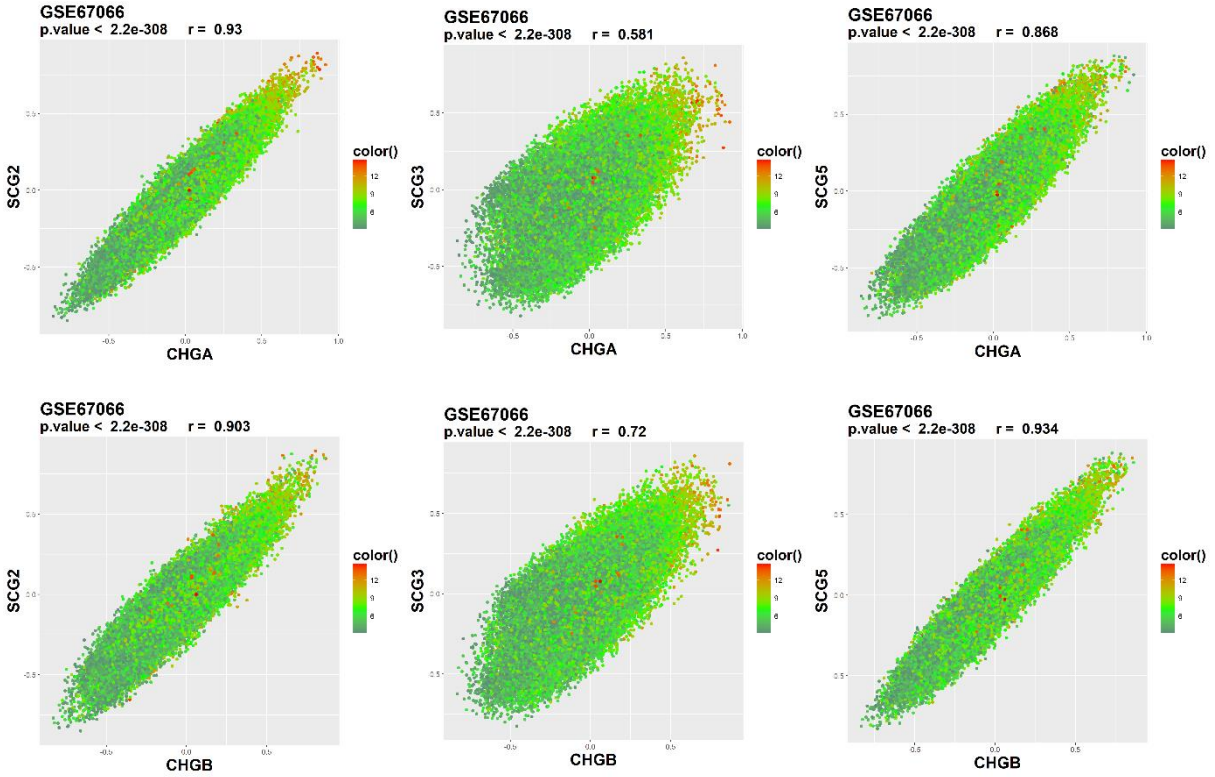


Figure 10: Correlation plots of chromogranins of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE67066 dataset (PCC, pheochromocytoma). The color gradient implies the expression level.

To examine the correlation between the other three secretogranins in a pairwise fashion, the correlations coefficients of CHGA and CHGB with secretogranins was plotted against those of SCG2, SCG3, and SCG5, respectively. In GSE67066 PCC dataset SCG3 was not as highly correlated with CHGA and CHGB while SCG2 and SCG5 were relatively more correlated with either CHGA or CHGB. Similarly, SCG3 was slightly less correlated to CHGA and CHGB in the GSE2841 PCC dataset (Figure 10). This has suggested that the co-expression network of SCG3

could have diverged from the rest of the chromogranin/secretogranin family in PCC.

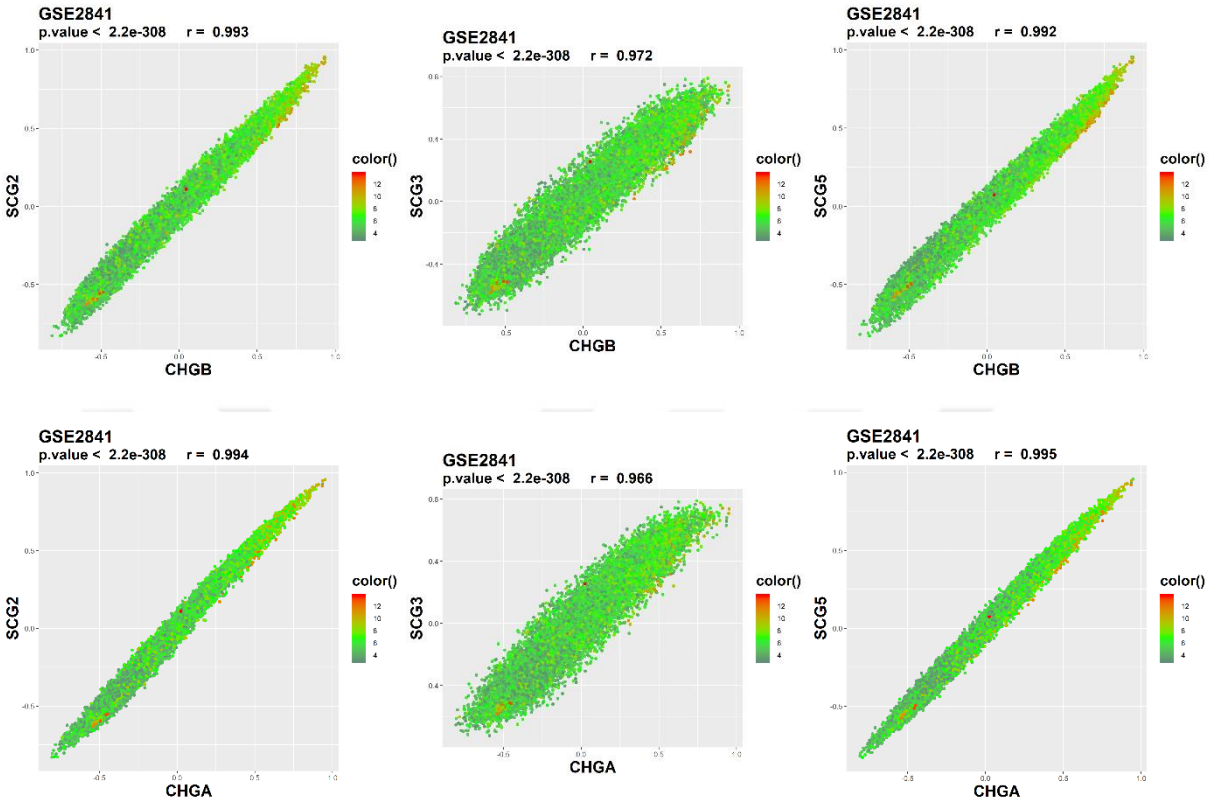


Figure 11: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE2841 dataset. The color gradient implies the expression level. GSE2841 contains 76 pheochromocytoma samples of various genetic origin.

Similarly, as in the other PCC dataset SCG3 behaves relatively less correlated when compared with the other two secretogranins suggesting divergence of expression between secretogranin 3 and the others (Figure 11).

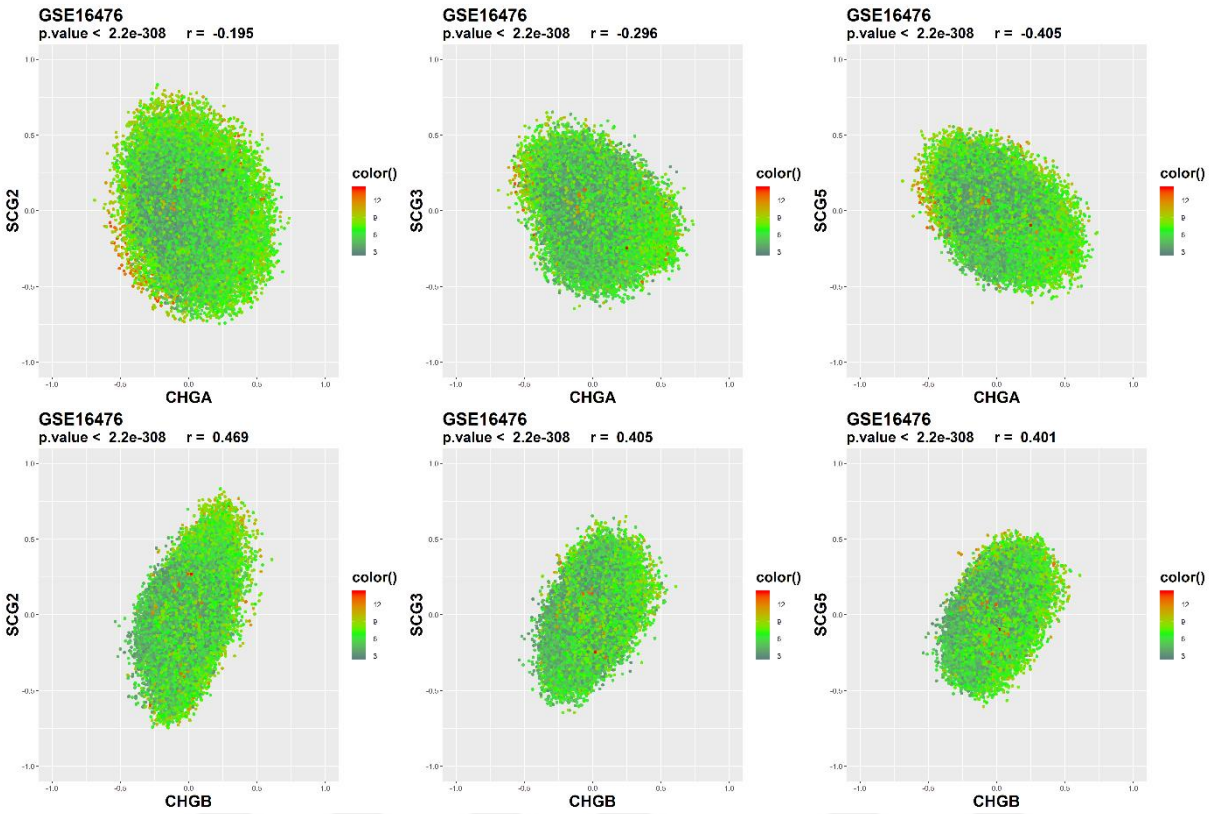


Figure 12: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE16476 dataset. The color gradient implies the expression level. GSE16476 contains 88 human neuroblastoma samples.

Interestingly, as I have seen the dissociation between CHGA and CHGB co-expression scatterplots SCG2, SCG3, and SCG5 all were negatively correlated with CHGA co-expression values while positively correlated with CHGB co-expression values (Figure 12). This suggests that expression divergence taking place between CHGA and CHGB was mostly driven by the distinctly opposite behavior of CHGA in neuroblastoma while CHGB, SCG2, SCG3, and SCG5 were relatively similar in their co-expression networks.

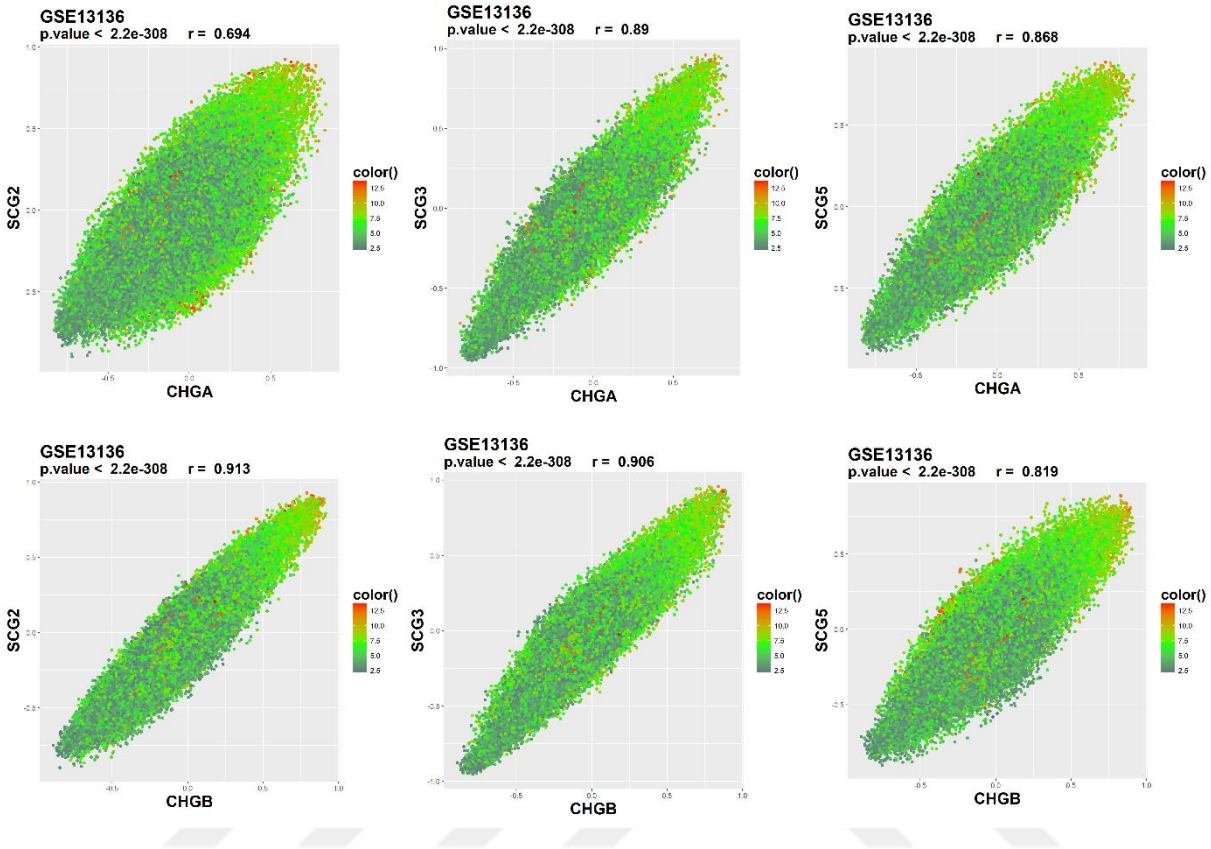


Figure 13: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE13136 dataset. The color gradient implies the expression level. GSE13136 contains 30 primary neuroblastoma samples.

The observation exhibiting opposing expression signatures in the other neuroblastoma dataset was not seen in this one (Figure 13).

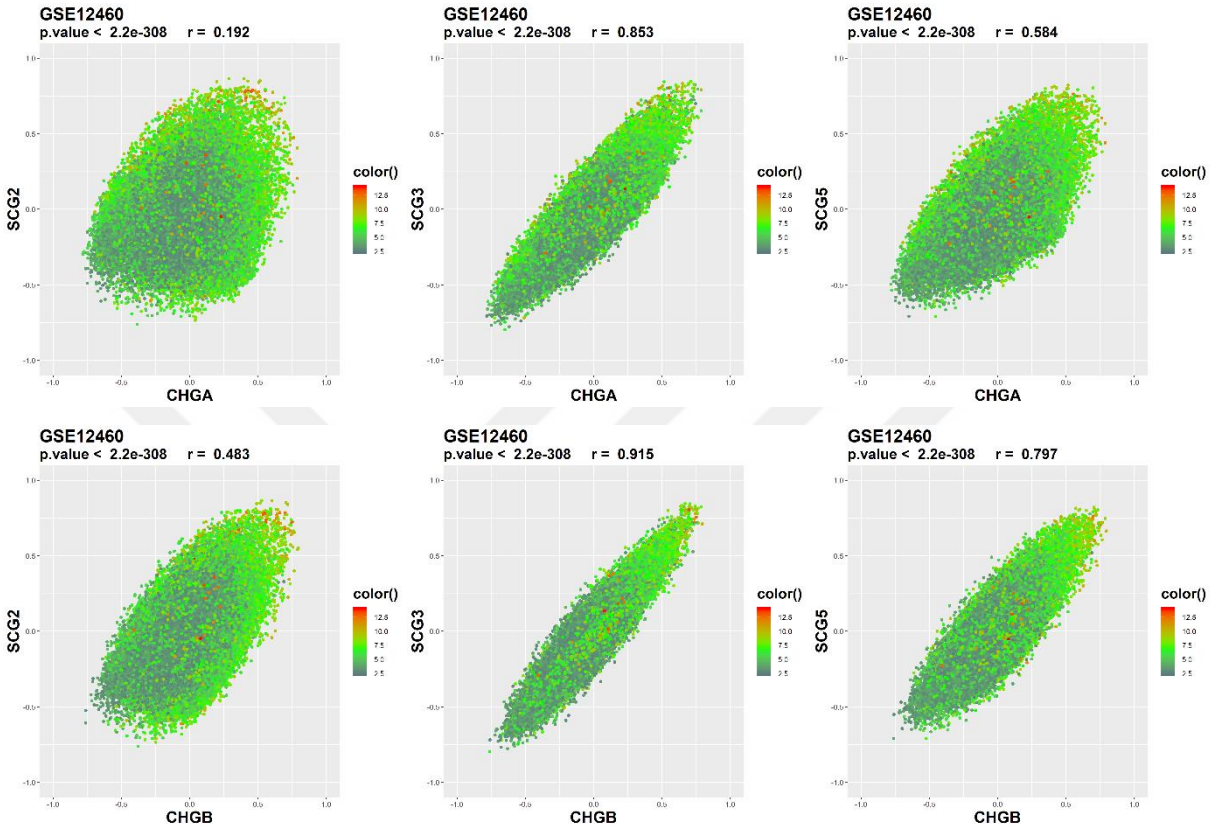


Figure 14: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE12460 dataset. The color gradient implies the expression level. GSE12460 contains 64 neuroblastic tumors (mainly neuroblastoma).

Although SCG2 was relatively less correlated with either CHGA and CHGB, the other secretogranins exhibited high similarity with chromogranins (Figure 14). After analyzing the pairwise co-expression pattern of CHGA and CHGB in PCC and neuroblastoma datasets, I have conducted the same analyses using datasets of several other cancer types to see whether the chromogranin pairwise co-expression signature could be replicated in non-neuronal cancers.

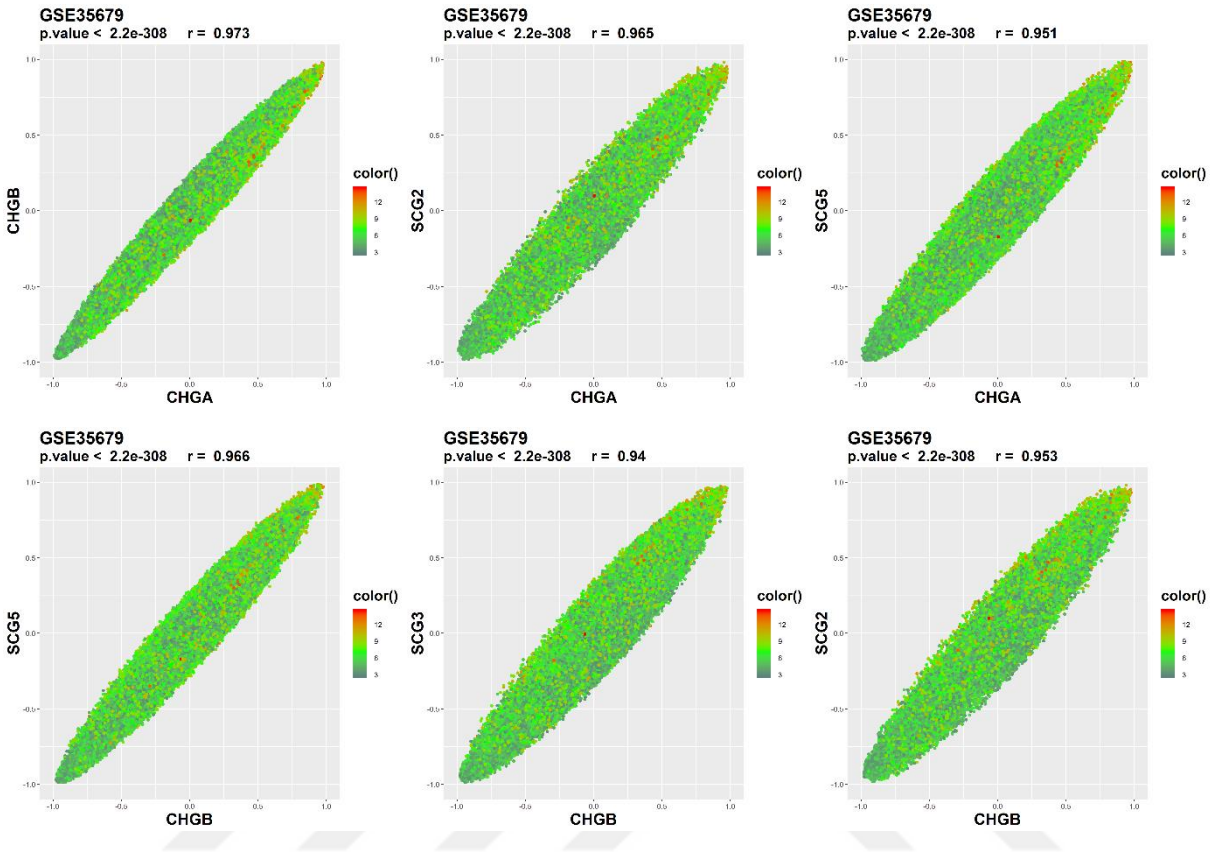


Figure 15: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE35679 dataset. The color gradient implies the expression level. GSE35679 contains 13 primary lung carcinoid samples.

NETs also occur in cancers of non-neuronal origins, such as lung cancer. All secretogranins and chromogranins were highly significantly positively correlated in the lung carcinoid samples (Figure 15).

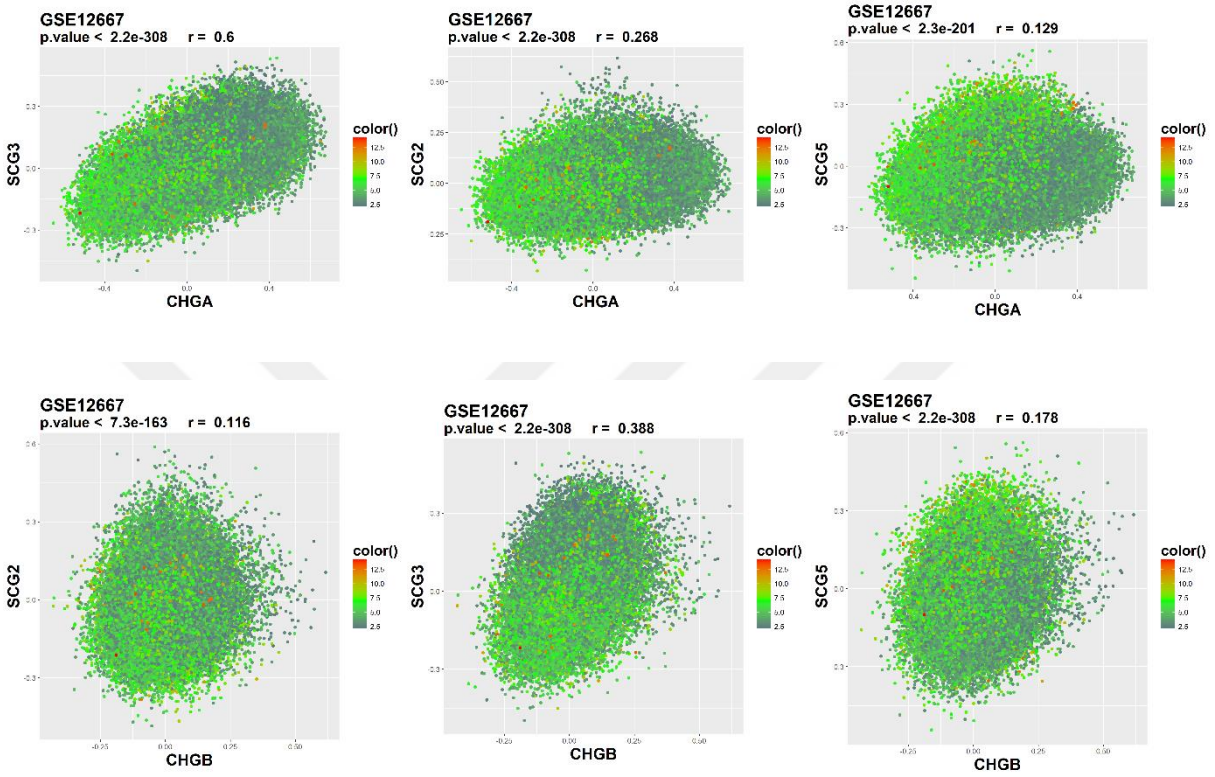


Figure 16: Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE12667 dataset. The color gradient implies the expression level. GSE12667 contains 75 lung adenocarcinoma samples.

On the other hand, in lung adenocarcinoma dataset, the correlation networks among the selected genes were relatively dissimilar. A similar finding to the neuroblastoma existed in lung adenocarcinoma samples. Although CHGA and CHGB (Figure 16) were highly correlated with each other the co-expression correlation of CHGA with other secretogranin (and the same for CHGB) were weaker.

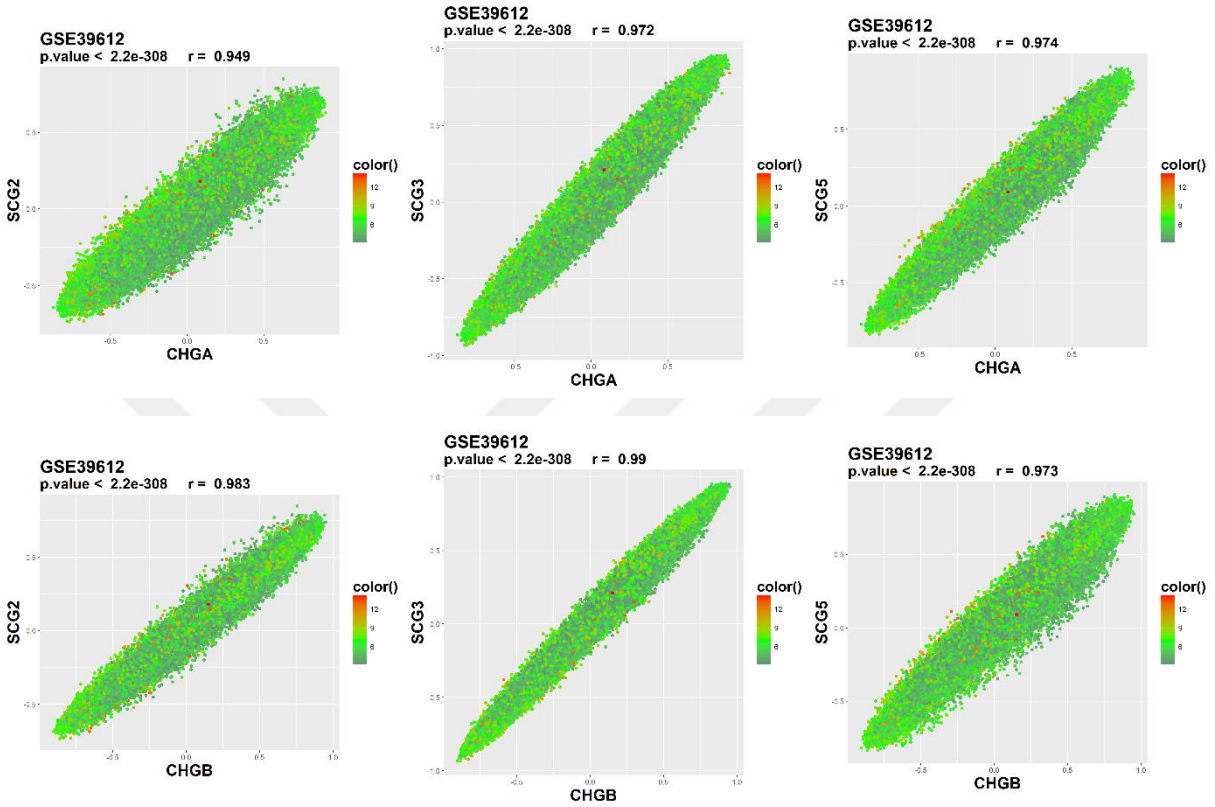


Figure 17:Correlation plots of chromogranin of interest (CHGA-CHGB) with secretogranin genes of interest (SCG2-3-5) for the GSE39612 dataset. The color gradient implies the expression level.

GSE39612 contains primary Merkel cell carcinomas, metastatic Merkel cell carcinomas, primary cutaneous squamous cell carcinomas, and basal cell carcinomas. Only the primary and metastatic Merkel cell carcinoma samples were used in this analysis (Figure 17). In this dataset the CHGA and CHGB were correlated with all of the secretogranins highly yet the order of correlation was different for each, i.e., for CHGA: SCG3>SCG5>SCG2; for CHGB: SCG3>SCG2>SCG5. This suggested also a divergence between the co-expression between CHGA and CHGB in terms of their association with other secretogranins. Yet all of the correlation plots were highly significant as in the case of lung carcinoid dataset (Figures 15 and 17).

4.2.2. Meta-correlation of CHGA and CHGB across transcriptome

The meta-correlation tab has enabled the implementation of the DerSimonian-Laird (DSL) meta-analytical approach with correlation coefficients as effect sizes and was tested for CHGA and CHGB genes and their correlated transcriptomic partners. To better explain the meta-correlation coefficient distribution of each gene a histogram of *meta-cor* r values have been plotted (Figure 18). In addition, I have plotted the meta-correlation r values from CHGA and CHGB across each other using a scatter plot to test whether they were correlated with each other (Figure 19).

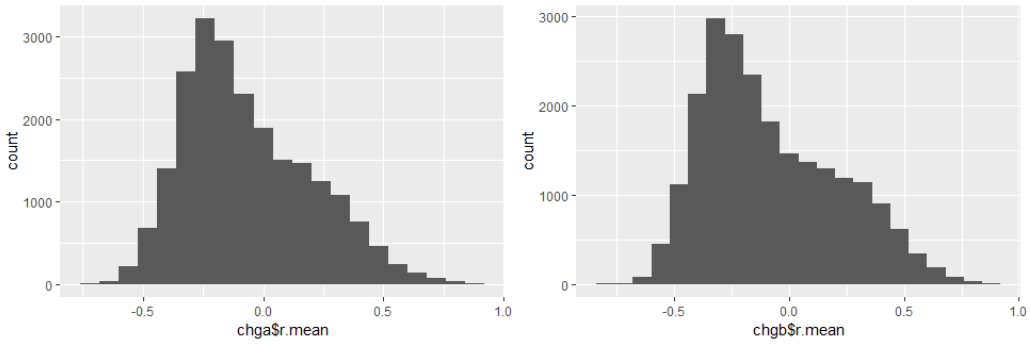


Figure 18: Histogram of r.mean values for CHGA (Left) and CHGB (right). The r.mean values was generated a result of the meta-correlation analysis using the GSE67066 and GSE2841 (PCC) datasets.

As we have seen above in the pairwise correlations the chromogranin network based on the two chromogranins was relatively concise and similar between datasets. In an attempt to increase the confidence for the correlation between co-expression networks of CHGA and CHGB, I have performed a meta-analysis for each gene separately and plotted the correlation values with each other (Figure 19). Interestingly a high correlation existed between meta-gene correlations of these two genes yet there was a considerable number of pairwise correlations exhibiting divergence between the chromogranins. A filtering threshold of 0.5 was used to identify (r_chga-r_chgb<-0.5 or >0.5) divergent gene pairs that were differentially correlated between CHGA and CHGB networks (Figure 19). The functional analysis of these genes indicated divergent pathways between CHGA and CHGB co-expression networks (Tables 2 and

3).

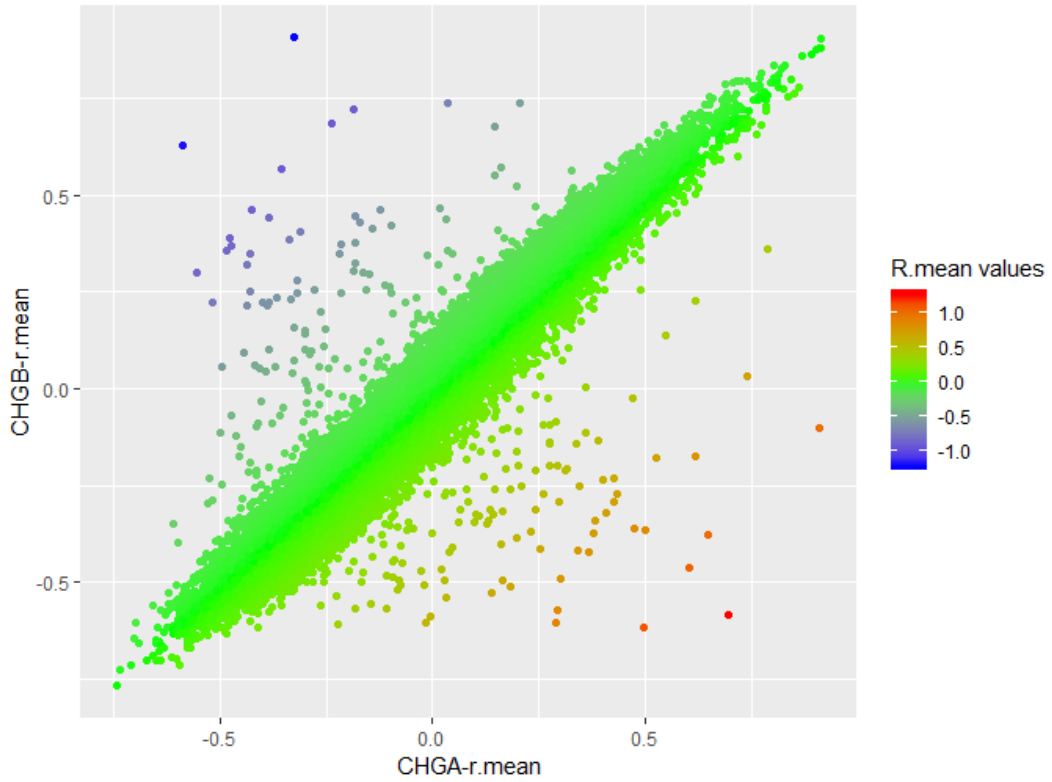


Figure 19:Correlation plot of the r.mean values for CHGA and CHGB. The r.mean values was generated a result of the meta-correlation analysis using the GSE67066 and GSE2841 (PCC) datasets. The color gradient represents the chromogranin divergence (chga - chgb).

Table 2: Functional annotation table for genes with high chga-chgb r.mean Value (chga – chgb > 0.5)

Annotation Cluster 1	Enrichment Score: 1.04		Count	P_Value	Benjamini
GOTERM_BP_DIRECT	regulation of phosphatidylinositol 3-kinase signaling	RT	4	1.6E-2	1.0E0
GOTERM_BP_DIRECT	positive regulation of phosphatidylinositol 3-kinase activity	RT	3	1.9E-2	9.7E-1
GOTERM_MF_DIRECT	1-phosphatidylinositol-3-kinase activity	RT	3	3.1E-2	1.0E0
GOTERM_BP_DIRECT	phosphatidylinositol-3-phosphate biosynthetic process	RT	3	4.2E-2	9.8E-1
GOTERM_MF_DIRECT	phosphatidylinositol-4,5-bisphosphate 3-kinase activity	RT	3	6.2E-2	1.0E0

Annotation Cluster 5	Enrichment Score: 0.8		Count	P_Value	Benjamini
UP_KEYWORDS	Cell membrane	RT	21	3.2E-2	9.0E-1
GOTERM_CC_DIRECT	plasma membrane	RT	23	2.9E-1	1.0E0
UP_SEQ_FEATURE	topological domain:Extracellular	RT	13	4.2E-1	1.0E0

CHGA was more highly associated with PI3K signaling and with proteins found on cell membrane (Table 2) while CHGB was found to have associations with cGMP, cAMP and andrenergic signaling and lipid metabolism (Table 3).

Table 3: Functional annotation table for genes with low chga-chgb r.mean value (chga - chgb < -0.5)

Annotation Cluster 1	Enrichment Score: 1.73		Count	P_Value	Benjamini
UP_SEQ_FEATURE	domain:RGS	RT	3	1.7E-2	1.0E0
SMART	RGS	RT	3	1.8E-2	6.9E-1
INTERPRO	Regulator of G protein signalling superfamily	RT	3	2.2E-2	1.0E0
Annotation Cluster 2	**Enrichment Score: 1.08**		**Count**	**P_Value**	**Benjamini**
UP_SEQ_FEATURE	short sequence motif:SH3-binding	RT	3	5.4E-2	1.0E0
UP_KEYWORDS	SH3-binding	RT	3	6.3E-2	7.9E-1
GOTERM_MF_DIRECT	SH3 domain binding	RT	3	1.7E-1	1.0E0
Annotation Cluster 3	**Enrichment Score: 1.05**		**Count**	**P_Value**	**Benjamini**
KEGG_PATHWAY	cGMP-PKG signaling pathway	RT	5	3.8E-2	1.0E0
KEGG_PATHWAY	Adrenergic signaling in cardiomyocytes	RT	4	1.0E-1	9.7E-1
KEGG_PATHWAY	cAMP signaling pathway	RT	4	1.8E-1	9.5E-1
Annotation Cluster 4	**Enrichment Score: 0.87**		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	lipid metabolic process	RT	4	8.0E-2	1.0E0
KEGG_PATHWAY	Glycerophospholipid metabolism	RT	3	1.2E-1	9.5E-1
UP_KEYWORDS	Lipid metabolism	RT	5	2.5E-1	9.5E-1
Annotation Cluster 5	**Enrichment Score: 0.81**		**Count**	**P_Value**	**Benjamini**
UP_SEQ_FEATURE	lipid moiety-binding region:S-palmitoyl cysteine	RT	4	1.3E-1	1.0E0
UP_KEYWORDS	Palmitate	RT	5	1.4E-1	9.3E-1
UP_KEYWORDS	Lipoprotein	RT	8	2.1E-1	9.6E-1

Next, I have taken a protein-protein interaction (PPI) network-based annotation approach to identify the functional components of correlated gene sets obtained from meta-analyses using STRING database. In this I have taken the meta-analysis correlation thresholds as indicated in the figure legends and obtained the gene list to be entered into STRING from the meta-correlation results .csv file generated using the *metacor* tab of the *CoEX-PM* application. In Figure 20 and Figure 22 CHGB co-expressed and CHGA co-expressed genes, were shown respectively, based on meta-analysis results of NB.

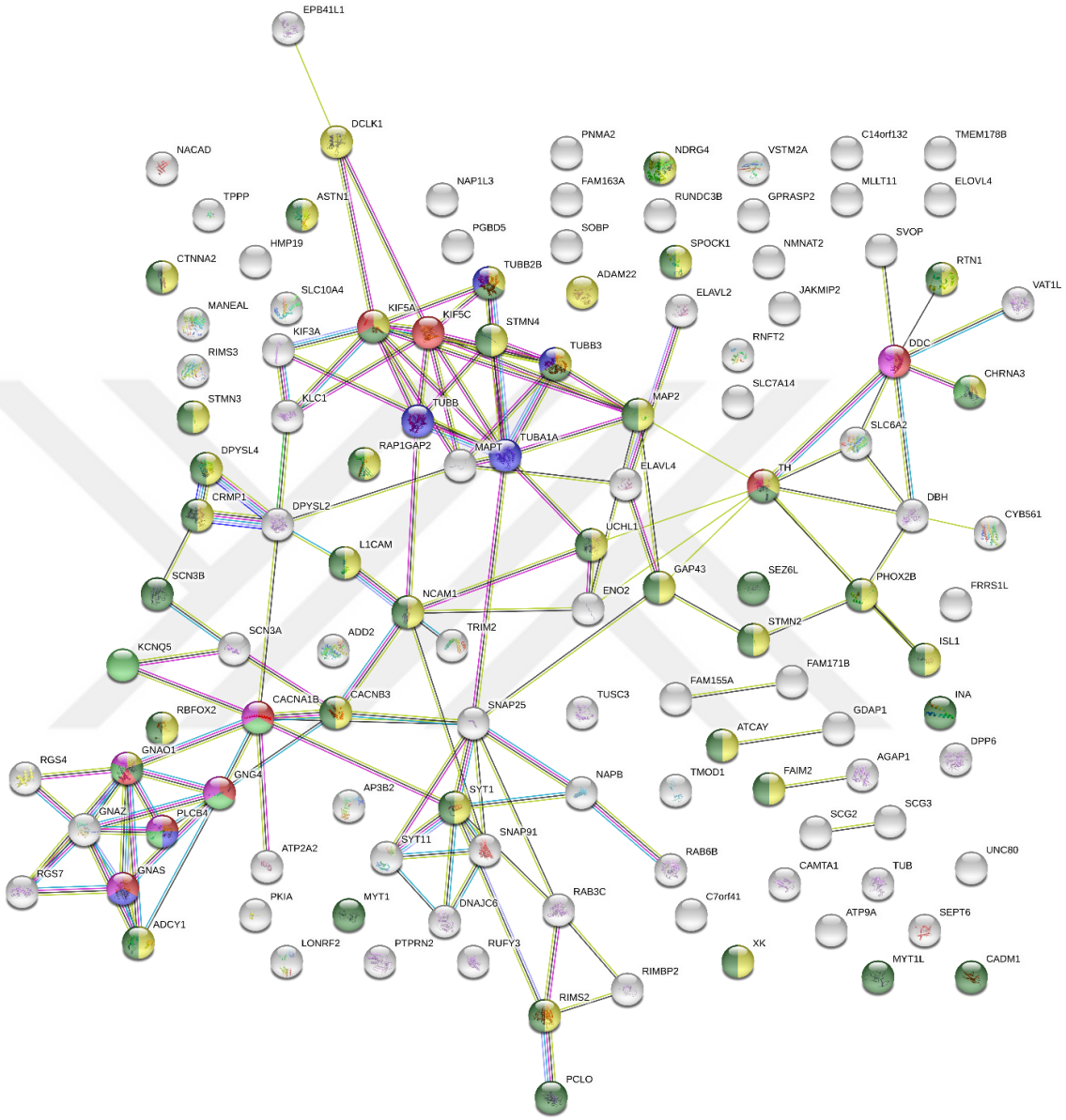


Figure 20: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the CHGB gene. Genes with r.mean>0.55 and p value < 0.003 were selected for the STRING mapping.

Based on Figure 20-21, the meta-co-expression network of CHGB in NB had enriched GO and KEGG groups related with neurogenesis, dopaminergic and cholinergic synapses. There were at least three subnetworks in the String PPI network based on CHGB. Two of these sub-networks that were consistently present in distinct meta co-expression networks were the tubulin network and the CHRNA3-DDC-TH-DBH group of synaptic components.

Network Stats

number of nodes:	121	expected number of edges:	36
number of edges:	130	PPI enrichment p-value:	< 1.0e-16
average node degree:	2.15	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.343		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0022008	neurogenesis	34	5.19e-10
GO:0048699	generation of neurons	33	5.19e-10
GO:0031175	neuron projection development	24	8.32e-10
GO:0007399	nervous system development	39	8.97e-10
GO:0030182	neuron differentiation	28	8.97e-10
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0005200	structural constituent of cytoskeleton	6	0.0238
GO:0008092	cytoskeletal protein binding	12	0.0238
GO:0031852	mu-type opioid receptor binding	2	0.0238
GO:0051430	corticotropin-releasing hormone receptor 1 binding	2	0.0238
GO:0005525	GTP binding	10	0.0347
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0097458	neuron part	38	2.46e-17
GO:0045202	synapse	29	8.45e-16
GO:0044456	synapse part	24	5.73e-13
GO:0030424	axon	21	6.03e-13
GO:0043005	neuron projection	28	6.28e-12
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04728	Dopaminergic synapse	9	1.51e-05
04540	Gap junction	6	0.00141
04725	Cholinergic synapse	6	0.00334
04726	Serotonergic synapse	6	0.00334
04911	Insulin secretion	5	0.00684
			(more ...)

PFAM Protein Domains

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
PF00836	Stathmin family	3	0.000604
PF00091	Tubulin/FtsZ family, GTPase domain	4	0.0026
PF03953	Tubulin C-terminal domain	4	0.0026

INTERPRO Protein Domains and Features

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR000956	Stathmin family	3	0.000645
IPR030514	Stathmin, conserved site	3	0.000645
IPR000217	Tubulin	4	0.00209
IPR003008	Tubulin/FtsZ, GTPase domain	4	0.00209
IPR008280	Tubulin/FtsZ, C-terminal	4	0.00209
			(more ...)

Figure 21: Functional enrichment results for the selected genes mentioned in Figure 20.

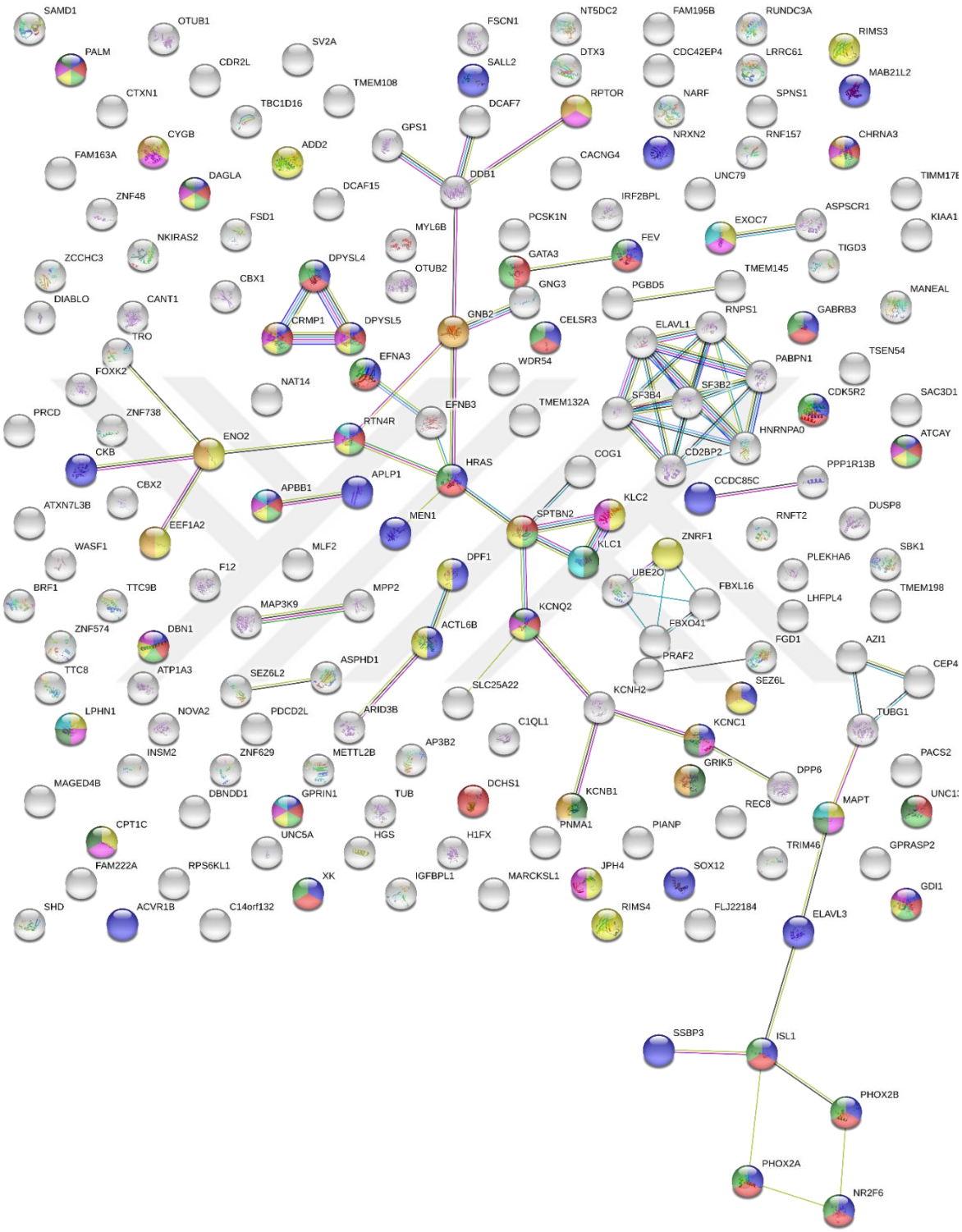


Figure 22: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the CHGA gene. Genes with r.mean>0.48 were selected for the STRING mapping .

Network Stats

number of nodes:	181	expected number of edges:	63
number of edges:	77	PPI enrichment p-value:	0.0472
average node degree:	0.851	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.282		

Functional enrichments in your network

Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0007399	nervous system development	39	0.00182
GO:0022008	neurogenesis	28	0.0439
GO:0048699	generation of neurons	27	0.0439

Cellular Component (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0097458	neuron part	30	1.2e-05
GO:0043005	neuron projection	21	0.00754
GO:0030424	axon	12	0.0285
GO:0030426	growth cone	7	0.0285
GO:0044297	cell body	13	0.0387
			(more ...)

Figure 23: The functional annotation of CHGA PPI network in Figure 22.

The NB meta co-expression network for CHGA does not have as many interactions as CHGB. This is in accord to the findings shown in the pairwise correlation analysis, where the correlation coefficient was consistently higher between CHGB-secretogranin pairs than it was for CHGA. Nevertheless, there are common components of the two network such as sets of genes that play a role in nervous system development or neurogenesis (Figure 22).

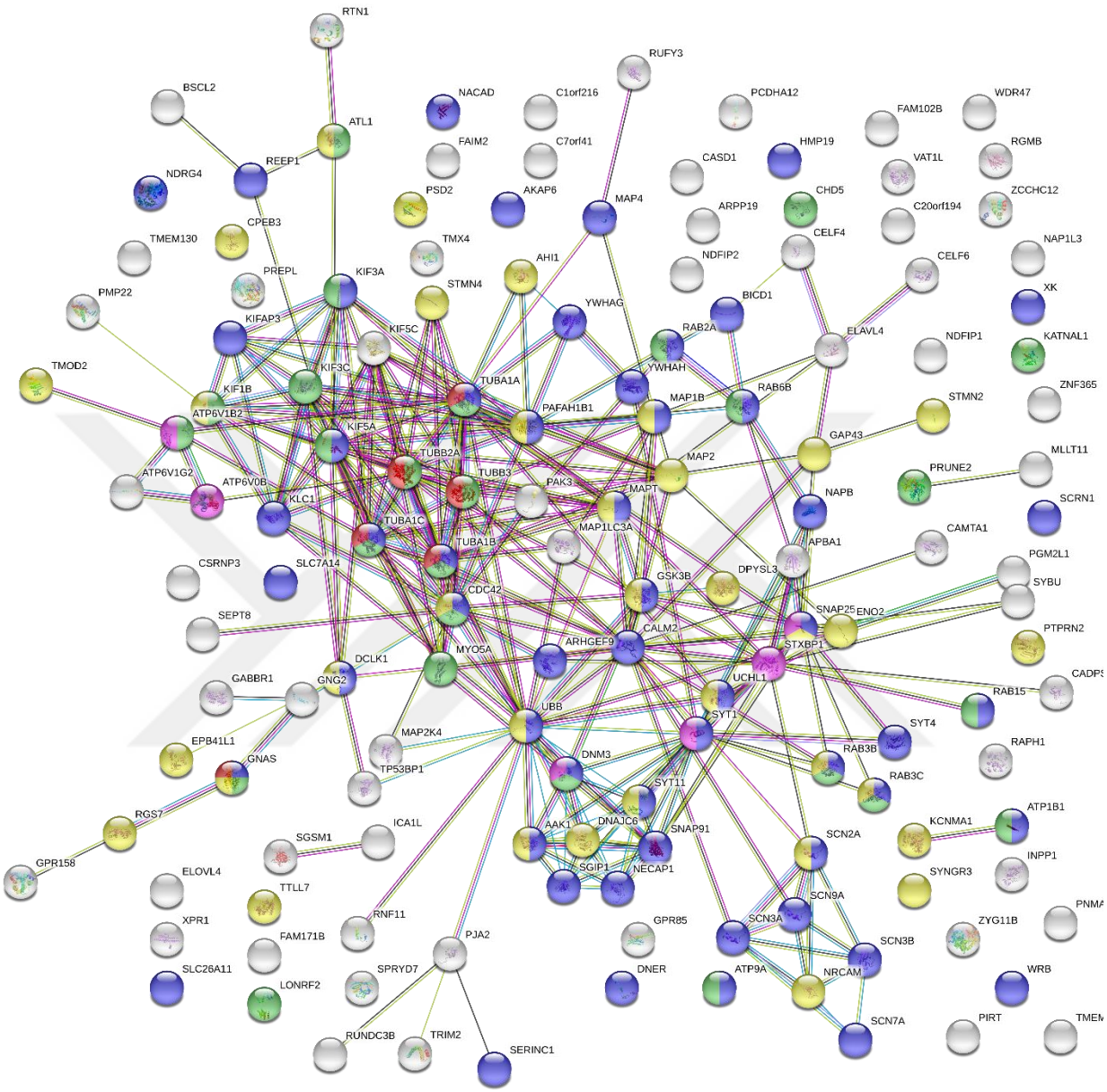


Figure 24: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the SCG2 gene. Genes with r.mean>0.65 were selected for the STRING mapping

Network Stats

number of nodes:	147	expected number of edges:	88
number of edges:	279	PPI enrichment p-value:	< 1.0e-16
average node degree:	3.8	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.408		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0030705	cytoskeleton-dependent intracellular transport	14	6.56e-11
GO:0072384	organelle transport along microtubule	10	2.92e-10
GO:0051656	establishment of organelle localization	17	4.48e-08
GO:0051234	establishment of localization	56	7.16e-08
GO:0010970	microtubule-based transport	11	8.75e-08
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0016462	pyrophosphatase activity	26	1.31e-08
GO:0017111	nucleoside-triphosphatase activity	25	1.76e-08
GO:0008092	cytoskeletal protein binding	20	2.48e-07
GO:0003924	GTPase activity	14	1.92e-06
GO:0005525	GTP binding	15	3.28e-05
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0097458	neuron part	35	1.14e-11
GO:0043005	neuron projection	30	6.59e-11
GO:0005874	microtubule	19	3.93e-09
GO:0030426	growth cone	12	1.29e-08
GO:0042995	cell projection	37	1.29e-08
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
05130	Pathogenic Escherichia coli infection	6	0.000553
04721	Synaptic vesicle cycle	6	0.000649
04540	Gap junction	6	0.00285
04145	Phagosome	7	0.00545

PFAM Protein Domains

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
PF00091	Tubulin/FtsZ family, GTPase domain	5	0.000154
PF03953	Tubulin C-terminal domain	5	0.000154

INTERPRO Protein Domains and Features

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR000217	Tubulin	5	0.00011
IPR003008	Tubulin/FtsZ, GTPase domain	5	0.00011
IPR008280	Tubulin/FtsZ, C-terminal	5	0.00011
IPR017975	Tubulin, conserved site	5	0.00011
IPR018316	Tubulin/FtsZ, 2-layer sandwich domain	5	0.00011
			(more ...)

Figure 25: Functional annotation of Protein-Protein interaction network in Figure 24.

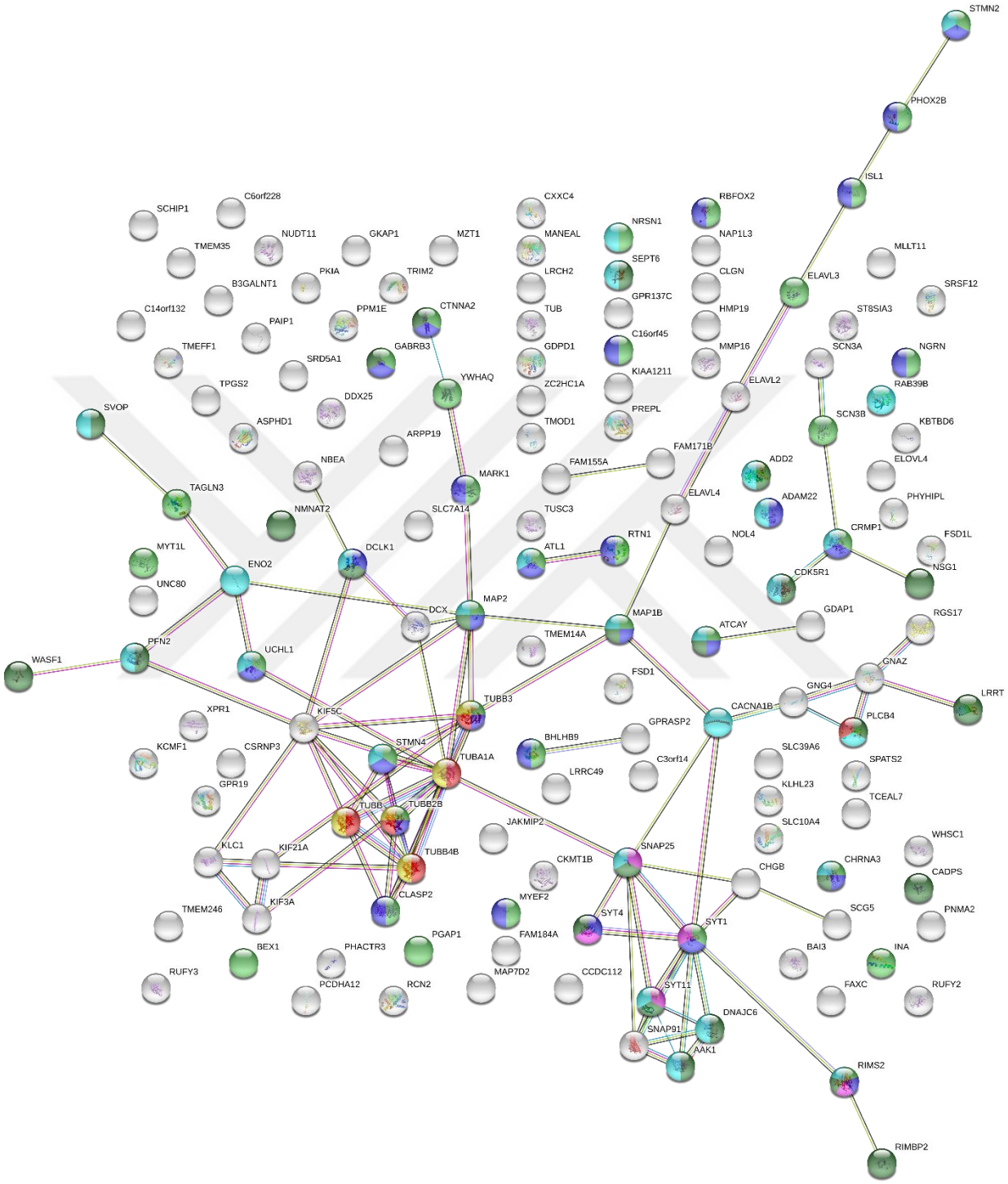


Figure 26: Protein-Protein interaction network of genes selected after meta-correlation analysis of GSE16476_GSE12460_GSE13136_GSE73537_GSE16237 (neuroblastoma) for the SCG3 gene. Genes with r.mean>0.6 were selected for the STRING mapping.

Network Stats

number of nodes:	149	expected number of edges:	33
number of edges:	93	PPI enrichment p-value:	< 1.0e-16
average node degree:	1.25	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.219		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0007399	nervous system development	34	0.000906
GO:0022008	neurogenesis	28	0.000906
GO:0048699	generation of neurons	27	0.000906
GO:0043623	cellular protein complex assembly	12	0.00375
GO:0048791	calcium ion-dependent exocytosis of neurotransmitter	5	0.00375
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0045202	synapse	25	2.52e-09
GO:0044456	synapse part	20	3.63e-07
GO:0097458	neuron part	26	1.35e-05
GO:0030424	axon	15	1.69e-05
GO:0043005	neuron projection	22	2.89e-05
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
05130	Pathogenic Escherichia coli infection	6	0.000598
04540	Gap junction	6	0.00461

PFAM Protein Domains

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
PF00091	Tubulin/FtsZ family, GTPase domain	5	0.000165
PF03953	Tubulin C-terminal domain	5	0.000165

INTERPRO Protein Domains and Features

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR000217	Tubulin	5	0.000118
IPR003008	Tubulin/FtsZ, GTPase domain	5	0.000118
IPR008280	Tubulin/FtsZ, C-terminal	5	0.000118
IPR017975	Tubulin, conserved site	5	0.000118
IPR018316	Tubulin/FtsZ, 2-layer sandwich domain	5	0.000118

Figure 27: Functional annotation of PPI network in Figure 26.

The neuroblastoma meta co-expression network of SCG3 displayed a particular fashion. There were many isolated nodes with no connections, however, the sub-networks within the main network were very well-connected. In figure 26, it is possible to identify at least 4 such sub-networks. The tubulin network which was present in CHGB and SCG3 meta co-expression networks as well, seems to be the most prominent network. Interestingly, neuroblastoma meta co-expression networks for both SCG2 and CHGB had more interactions than PPI's of other

genes of interest. Both networks included the reoccurring concise hubs, especially the tubulin and the CHRNA3-DDC-TH-DBH groups, however, the tubulin hub members were represented better in the SCG2 PPI network whereas the dopaminergic signaling group was more prominent in the CHGB PPI network. This showed the divergence of functional pathways and processes that are regulated by chromogranin and secretogranin co-expressed genes.

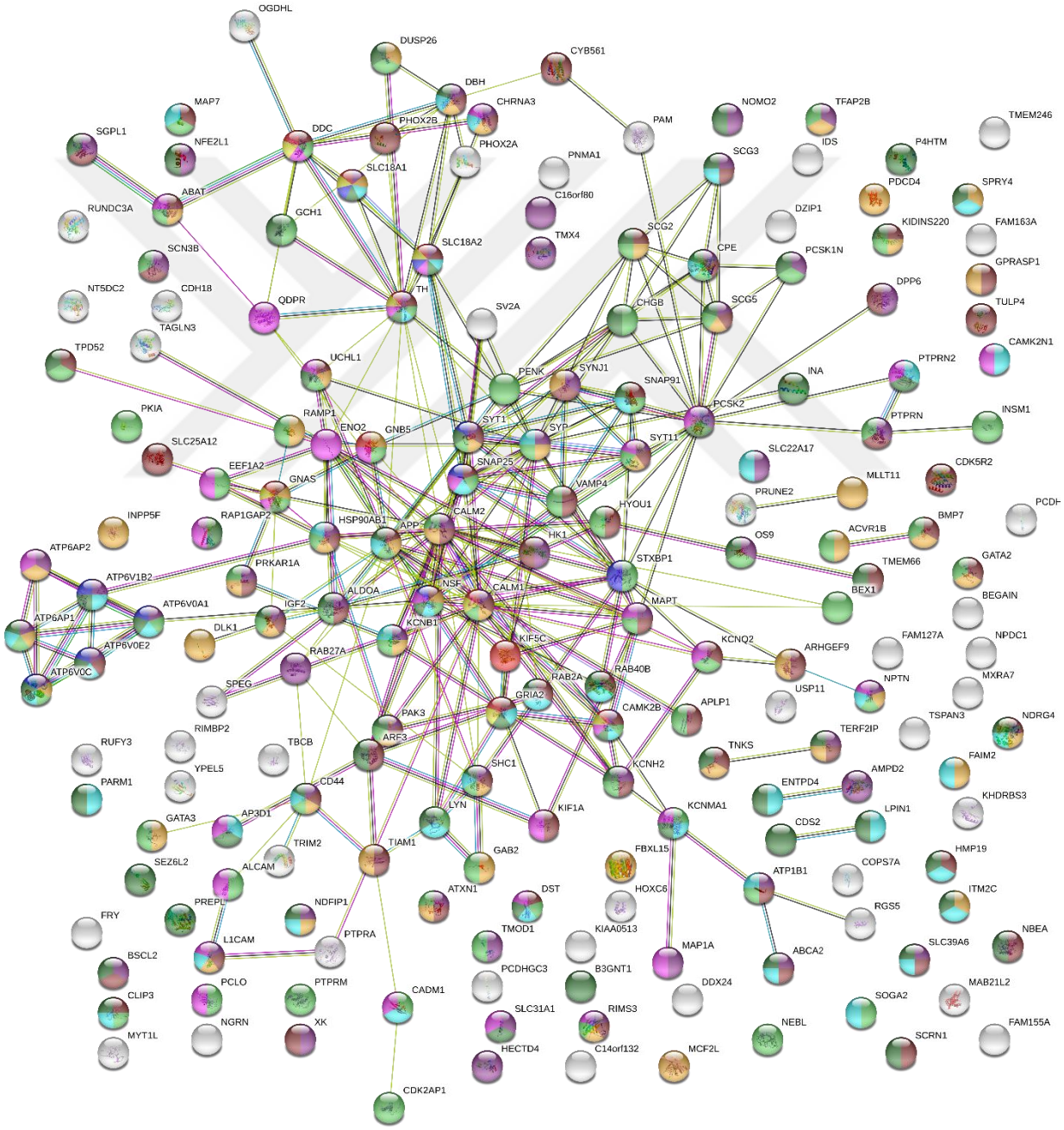


Figure 28: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the CHGA gene. Genes with r.mean>0.6 were selected for the

STRING mapping

Network Stats

number of nodes:	190	expected number of edges:	109
number of edges:	268	PPI enrichment p-value:	< 1.0e-16
average node degree:	2.82	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.352		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0065008	regulation of biological quality	62	4.99e-08
GO:0051179	localization	76	5.74e-08
GO:0023051	regulation of signaling	56	3.01e-07
GO:0010646	regulation of cell communication	57	5.35e-07
GO:0032940	secretion by cell	22	6.14e-07
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0005515	protein binding	75	1.55e-05
GO:0046961	proton-transporting ATPase activity, rotational mechanism	5	0.00208
GO:0017075	syntaxin-1 binding	4	0.00233
GO:0019829	cation-transporting ATPase activity	6	0.00239
GO:0043168	anion binding	46	0.00239
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0098805	whole membrane	50	4.08e-10
GO:0012505	endomembrane system	71	5.44e-10
GO:0097458	neuron part	34	4.86e-08
GO:0031410	cytoplasmic vesicle	34	5.64e-08
GO:0016023	cytoplasmic membrane-bounded vesicle	32	7.03e-08
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04721	Synaptic vesicle cycle	10	5.45e-08
05031	Amphetamine addiction	8	2.2e-05
04728	Dopaminergic synapse	10	2.35e-05
05030	Cocaine addiction	6	0.000342
05034	Alcoholism	9	0.000342
			(more ...)

Figure 29: Functional annotation of PPI network in Figure 28.

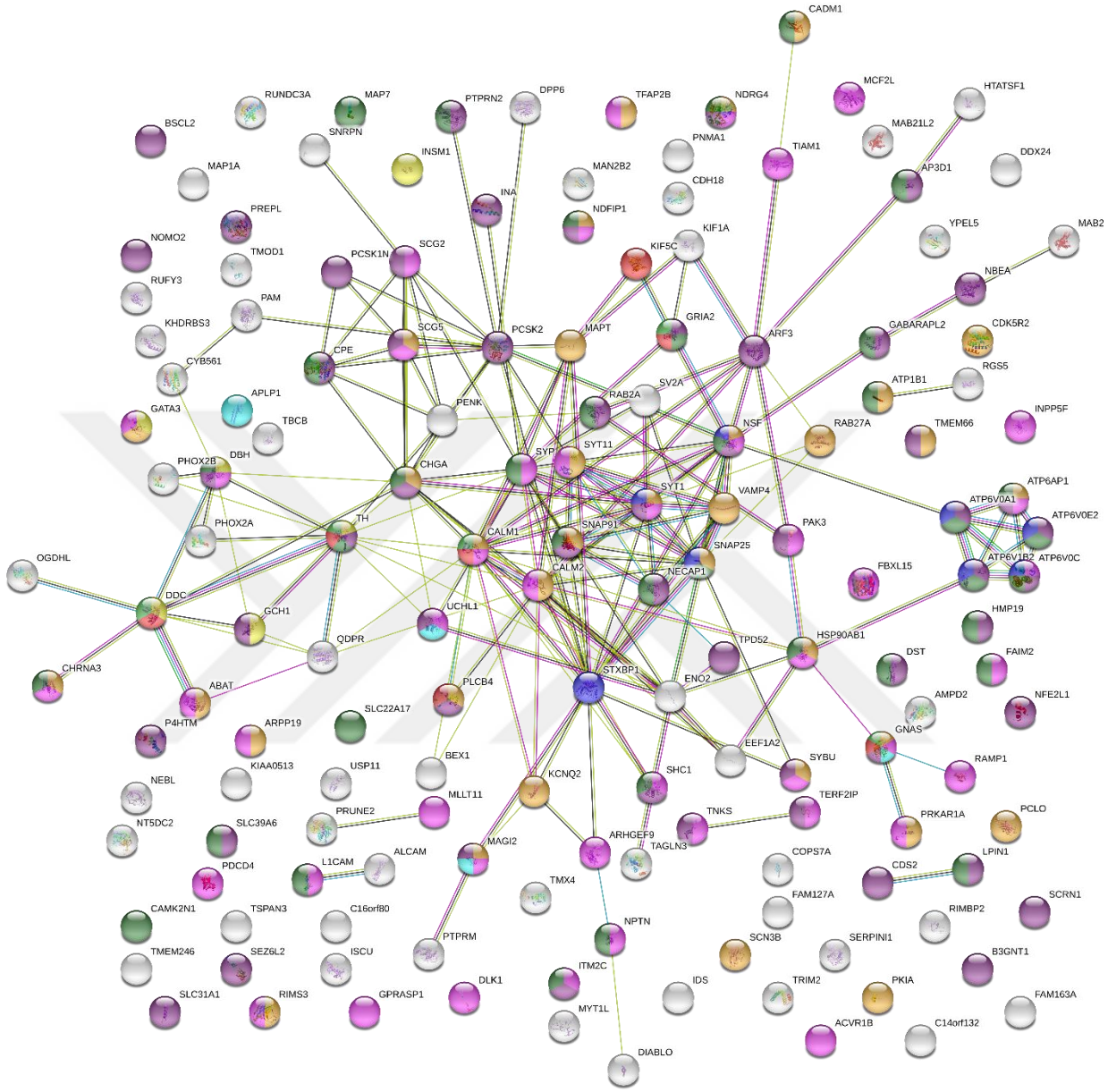


Figure 30: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the CHGB gene. Genes with r.mean>0.6 were selected for the STRING mapping.

Network Stats

number of nodes:	152	expected number of edges:	64
number of edges:	189	PPI enrichment p-value:	< 1.0e-16
average node degree:	2.49	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.349		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0042423	catecholamine biosynthetic process	6	1.32e-05
GO:0046916	cellular transition metal ion homeostasis	10	1.32e-05
GO:0051049	regulation of transport	35	1.32e-05
GO:0023051	regulation of signaling	44	3.55e-05
GO:0006879	cellular iron ion homeostasis	8	4.36e-05
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0017075	syntaxin-1 binding	4	0.000873
GO:0019829	cation-transporting ATPase activity	6	0.000873
GO:0046961	proton-transporting ATPase activity, rotational mechanism	5	0.000873
GO:0031690	adrenergic receptor binding	4	0.00221
GO:0005515	protein binding	56	0.00225
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0097458	neuron part	30	1.54e-07
GO:0043005	neuron projection	25	9.89e-07
GO:0012505	endomembrane system	54	1.07e-06
GO:0031410	cytoplasmic vesicle	28	1.07e-06
GO:0098805	whole membrane	37	1.07e-06
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04721	Synaptic vesicle cycle	8	4.09e-06
05110	Vibrio cholerae infection	6	0.000299
04728	Dopaminergic synapse	7	0.00284
04966	Collecting duct acid secretion	4	0.00284
05031	Amphetamine addiction	5	0.00585
			(more ...)

Figure 31: Functional annotation of PPI network in Figure 30.

The meta co-expression networks for CHGA (Figure 28) and CHGB (Figure 30) contained common gene clusters, mostly genes involved in dopaminergic pathways and signaling processes. However, the number of highly meta-correlated genes for CHGA was slightly higher at the same threshold. Based on the previous findings from pairwise correlation analysis, CHGA-CHGB correlation was stronger in PCC than it was in NB, therefore it is feasible to anticipate common nodes and sub-networks in meta co-expression networks of these two genes. The

CHRNA3-DDC-DBH-TH that was re-occurring in neuroblastoma networks is again present here. In addition, the ATPase network was sustained between CHGA and CHGB networks implying its importance for both proteins.

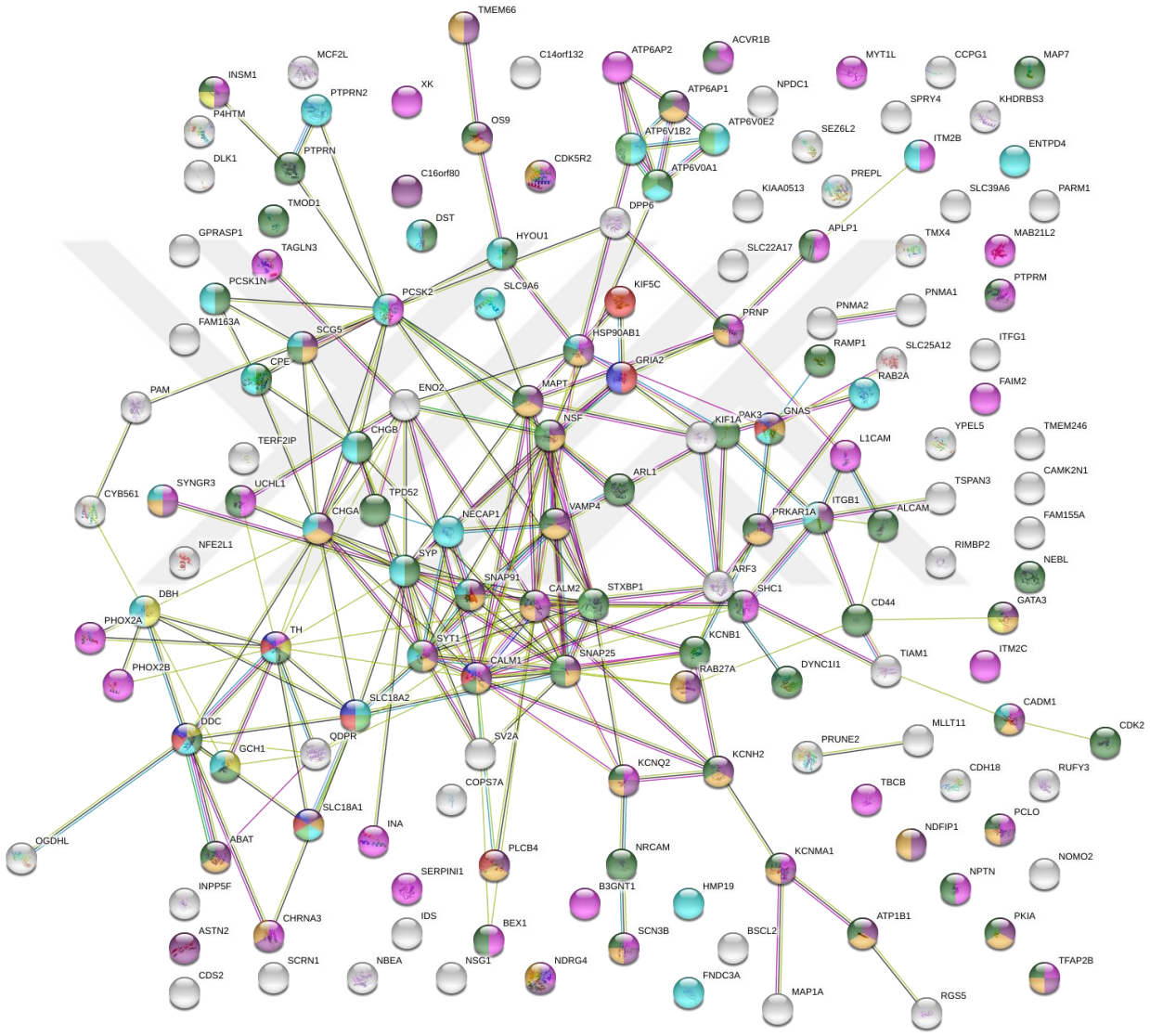


Figure 32: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG2 gene. Genes with r.mean>0.65 were selected for the STRING mapping.

Network Stats

number of nodes:	155	expected number of edges:	61
number of edges:	213	PPI enrichment p-value:	< 1.0e-16
average node degree:	2.75	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.332		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0065008	regulation of biological quality	52	8.5e-07
GO:0007399	nervous system development	40	2.54e-06
GO:0032879	regulation of localization	42	5.74e-06
GO:0051049	regulation of transport	36	5.74e-06
GO:0042423	catecholamine biosynthetic process	6	9.87e-06
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0005515	protein binding	60	0.0012
GO:0017075	syntaxin-1 binding	4	0.00206
GO:0015077	monovalent inorganic cation transmembrane transporter activity	12	0.0123
GO:0008092	cytoskeletal protein binding	14	0.0143
GO:0019829	cation-transporting ATPase activity	5	0.0143
			(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0098805	whole membrane	46	3.04e-11
GO:0012505	endomembrane system	62	4.76e-10
GO:0031410	cytoplasmic vesicle	33	1.03e-09
GO:0043005	neuron projection	28	5.02e-09
GO:0016020	membrane	100	8.14e-09
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04721	Synaptic vesicle cycle	9	1.99e-07
04728	Dopaminergic synapse	9	5.77e-05
05031	Amphetamine addiction	7	5.77e-05
05030	Cocaine addiction	6	0.000132
05034	Alcoholism	8	0.000569

Figure 33: Functional annotation of PPI network in Figure 32.

The PCC meta co-expression network for SCG2 (Figure 32) contained the previously consistent sub-networks as in Figure 28 and 30). The CHRNA3-DDC-DBH-TH group, the central SYP-Calmodulin sub-network, as well as the ATPase group were all retained. CHGB and SCG2 co-expression pattern in PCC seem to be very intertwined, having many common components.

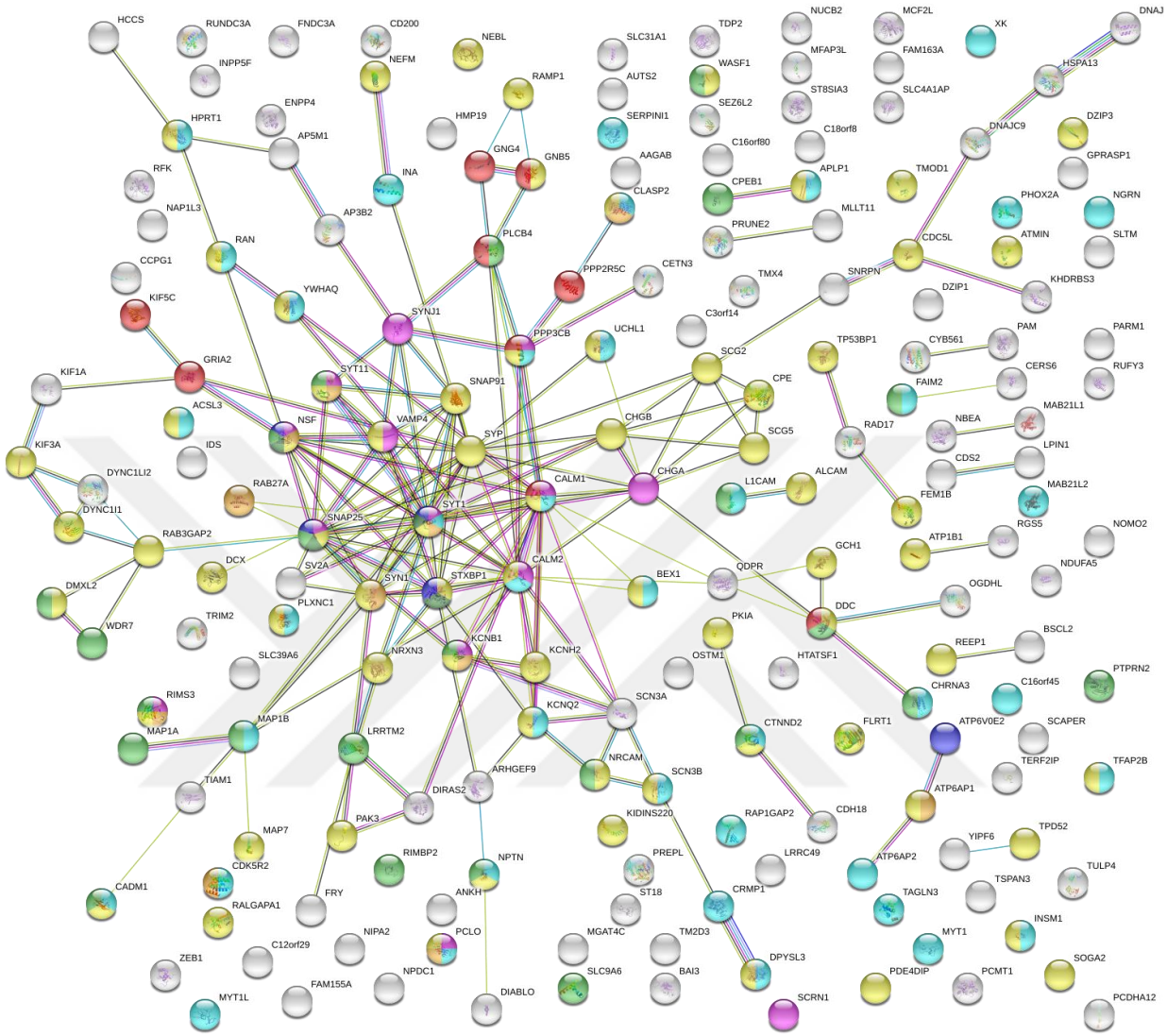


Figure 34: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG3 gene. Genes with r.mean>0.58 were selected for the STRING mapping.

Network Stats

number of nodes:	193	expected number of edges:	84
number of edges:	184	PPI enrichment p-value:	< 1.0e-16
average node degree:	1.91	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.354		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0006887	exocytosis	14	0.000933
GO:0007399	nervous system development	40	0.000933
GO:0017157	regulation of exocytosis	11	0.000933
GO:0051640	organelle localization	16	0.000933
GO:0051656	establishment of organelle localization	14	0.000933
			(more ...)

Molecular Function (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0005515	protein binding	70	0.00243
GO:0017075	syntaxin-1 binding	4	0.00495

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0045202	synapse	24	2.08e-06
GO:0097458	neuron part	32	2.08e-06
GO:0043005	neuron projection	27	5.54e-06
GO:0044456	synapse part	20	1.67e-05
GO:0042995	cell projection	37	3.48e-05
			(more ...)

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04728	Dopaminergic synapse	9	0.000786
04721	Synaptic vesicle cycle	5	0.0365

Figure 35: Functional annotation of PPI network in Figure 34.

Most of the components of the consistent sub-networks mentioned previously were present in the meta co-expression network for SCG3 (Figure 34). However, the r.mean cut-off was required to be lower for these sub-networks to appear and still some of their components were still missing. The sustained sub-networks were not represented as wholly for SCG3 (Figure 34). When the cut-off value was taken as r.mean>0.6, the sub-networks were represented by even less components and when the cut-off was r.mean>0.55, they were present fully as seen in earlier networks (data not shown).

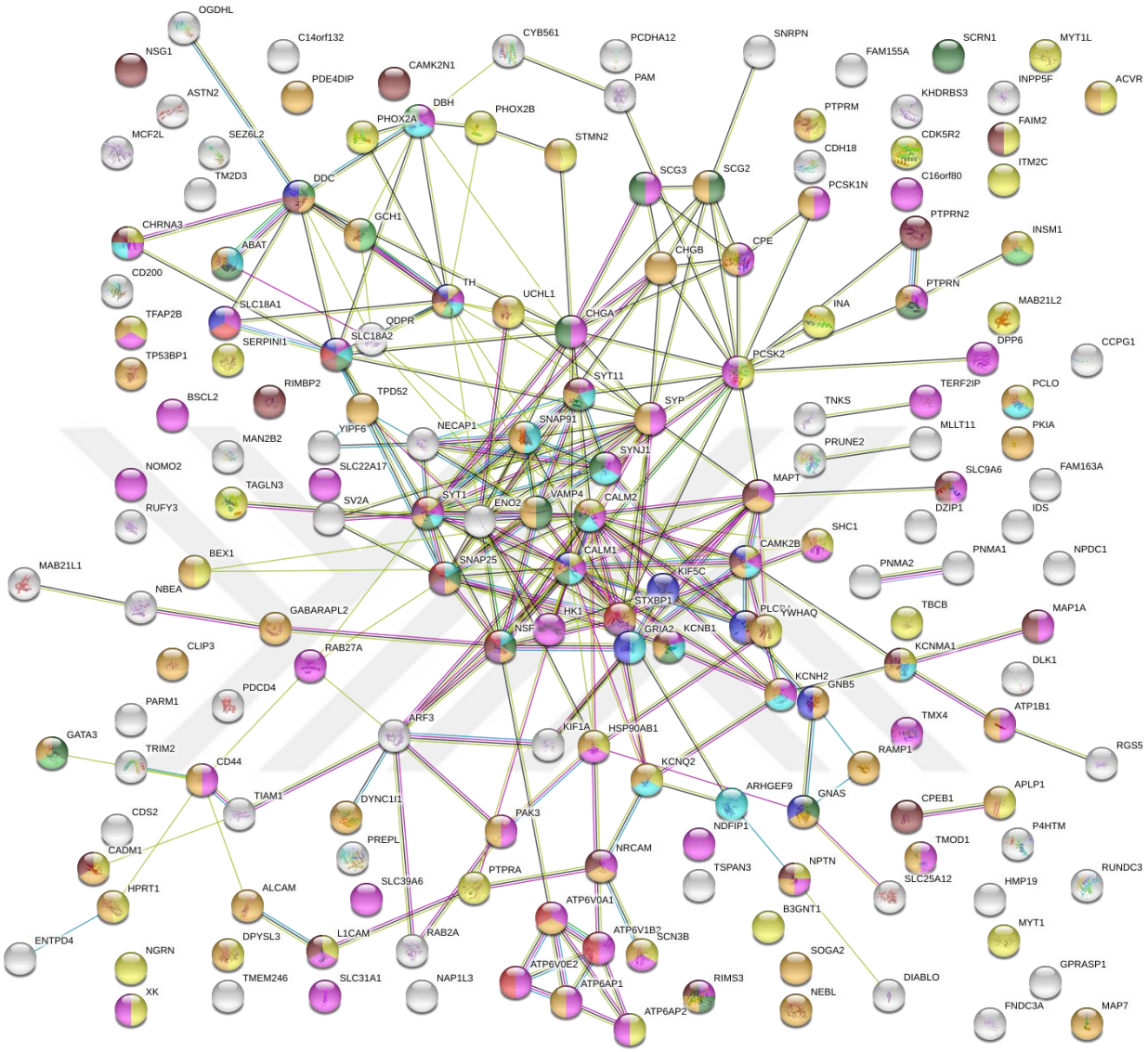


Figure 36: PPI network of genes selected after meta-correlation analysis of GSE67066_GSE2841 (PCC) for the SCG5 gene. Genes with r.mean>0.65 were selected for the STRING mapping.

Biological Process (GO)

pathway ID	pathway description	count in gene set	false discovery rate
GO:0007399	nervous system development	44	1.14e-06
GO:0065008	regulation of biological quality	53	1.17e-05
GO:0007268	synaptic transmission	21	1.59e-05
GO:0032940	secretion by cell	19	1.59e-05
GO:0042423	catecholamine biosynthetic process	6	1.59e-05
			(more ...)

Molecular Function (GO)

pathway ID	pathway description	count in gene set	false discovery rate
GO:0005515	protein binding	65	0.00142
GO:0017075	syntaxin-1 binding	4	0.0032
GO:0000149	SNARE binding	7	0.0168
GO:0008092	cytoskeletal protein binding	15	0.0168
GO:0015077	monovalent inorganic cation transmembrane transporter activity	12	0.0185
			(more ...)

Cellular Component (GO)

pathway ID	pathway description	count in gene set	false discovery rate
GO:0097458	neuron part	36	3.88e-10
GO:0031410	cytoplasmic vesicle	35	1.06e-09
GO:0098805	whole membrane	46	1.06e-09
GO:0043005	neuron projection	29	1.46e-08
GO:0045202	synapse	25	1.49e-08
			(more ...)

KEGG Pathways

pathway ID	pathway description	count in gene set	false discovery rate
04721	Synaptic vesicle cycle	9	5.26e-07
04728	Dopaminergic synapse	11	1.22e-06
05031	Amphetamine addiction	8	7.12e-06
05034	Alcoholism	9	0.000183
05030	Cocaine addiction	6	0.000199
			(more ...)

Figure 37: Functional annotation of PPI network in Figure 36.

Unlike SCG3, all of the consistent meta-co-expression network hubs were present in the SCG5 related PPI network. The main hubs were represented with no missing members, despite the high threshold of r.mean value (Figure 36).

4.2.3 Chromogranin and Secretogranin Co-expression modules using WGCNA Tab

In this part of the analysis I have implemented a customized pipeline for WCGNA, weighted gene co-expression network analysis. I have used the PCC dataset GSE67066, mentioned in detail in the earlier chapters. In order to preprocess and filter the data initially, a *limma* differential expression analysis pipeline was integrated to the main workflow of the WGCNA tab. The tab was used in accord to the guidelines explained in the methods section.

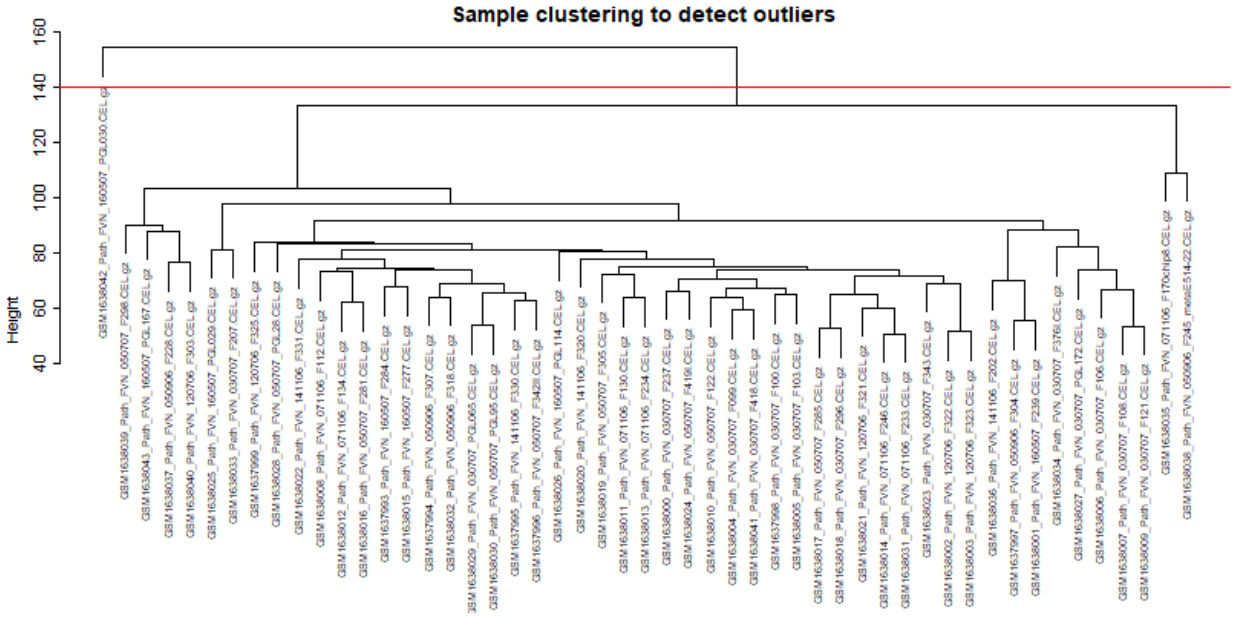


Figure 38: Sample clustering for GSE67066 with a cut-off line at 140.

After filtering the probes of the expression matrix – keeping only the most significant 10K probes - by *limma* differential expression analysis, the samples were hierarchically clustered to detect any outliers. In this case, a single sample was omitted since it fell out of the main group (Figure 38). Note that, the initial filtering analysis depends on many such as the factor vector for grouping or the number of probes to be included in module construction. For example, a researcher interested in gender-related differential expression signatures does not choose a factor containing malignancy state information for each sample. They would rather choose a factor displaying gender data.

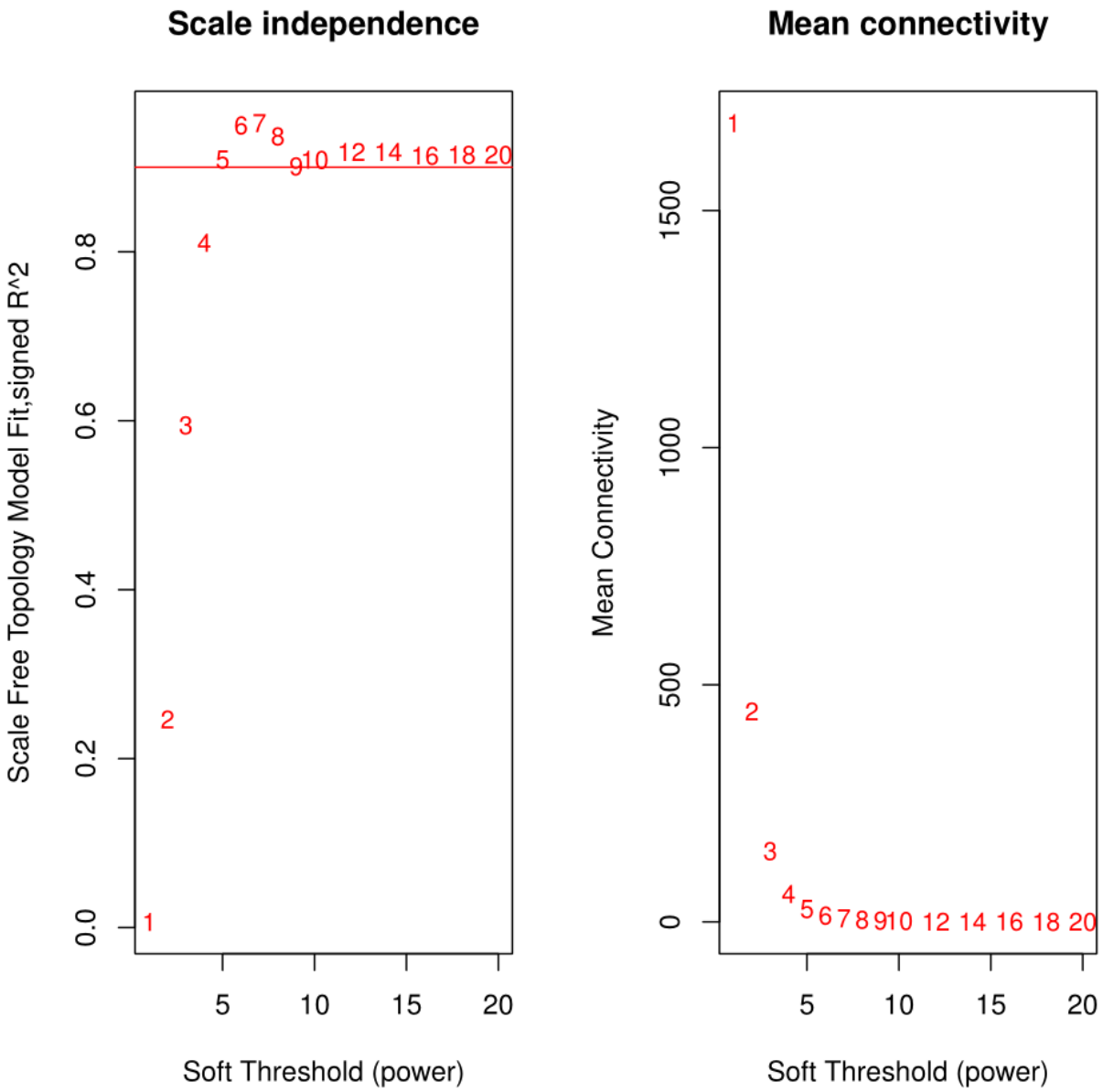


Figure 39: Network topology analysis for filtered GSE67066.

As mentioned earlier in the methods section, the authors of the WGCNA package proposes the Scale Free Topology model to determine the power value for the module construction step. Since, it is at 5 that, SFT Model reaches the threshold first and the mean connectivity pattern starts to flatten (Figure 39), thus a power threshold value of 5 was selected [59]. The power threshold value involves a certain number (5 in this case) to which the correlation is raised. In other words, if the power threshold is too low, there would be too many connections

in the co-expression network than it is appropriate to convey meaningful information, thus the high mean connectivity at 1 as shown in Figure 39. In contrast, if it is picked too high, hypothetically there would be no connections between the nodes in the co-expression network and genes would not be able to classify into modules. There is a trade-off between specificity and connectivity needs to be taken into consideration when deciding the power threshold value. However, there is no consensus on what the best method of picking the soft threshold and it greatly depends on the structure of the data that is being operated on.

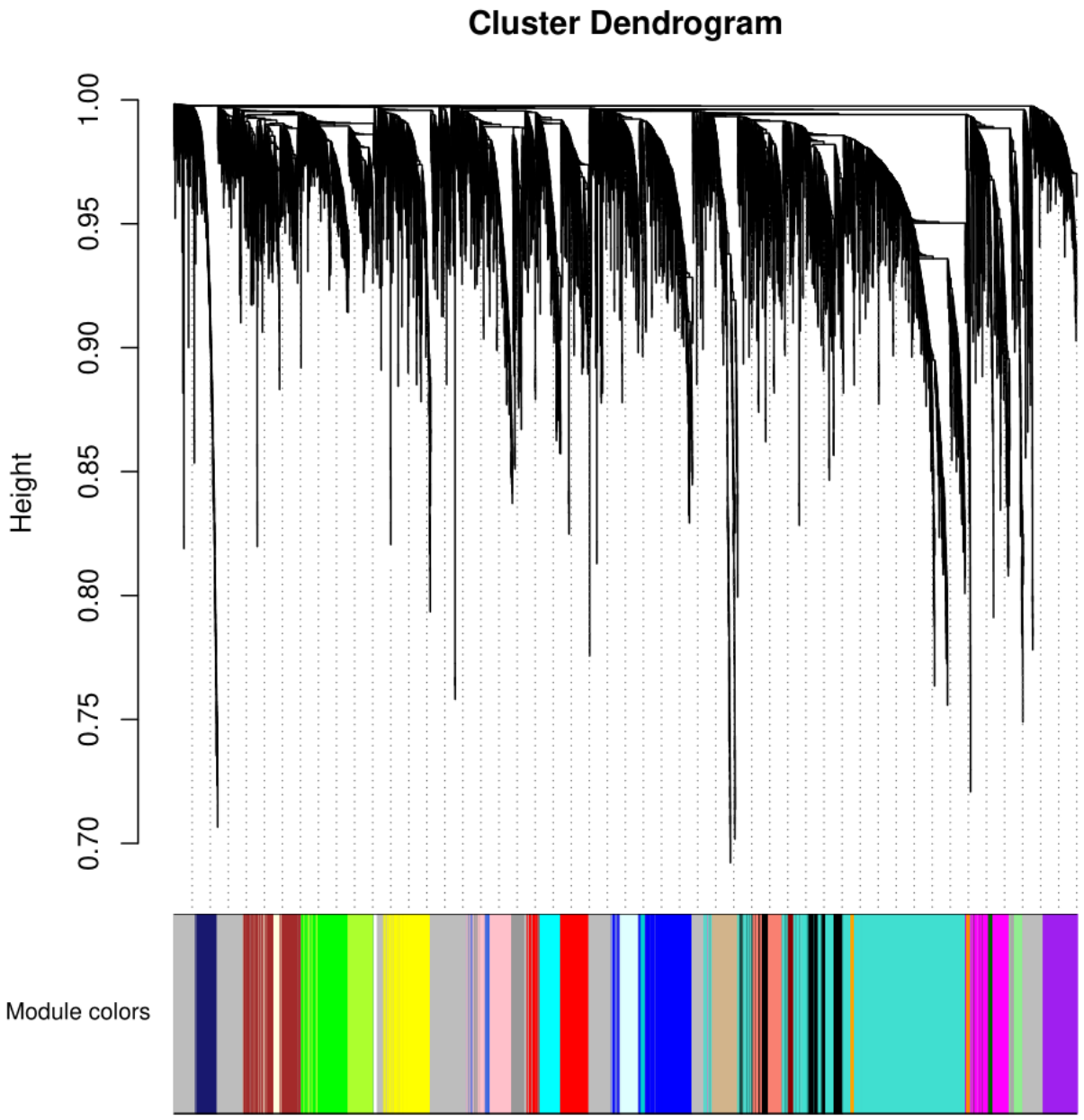


Figure 40: Clustering dendrogram of probes, with dissimilarity based on topological overlap. Module colors are shown for their assigned probes.

The modules were constructed as shown in the Figure 40, each module was represented by a different color and the grey parts indicated probes that could not be assigned to any module. In addition to visualizing the module-gene relationship, module eigengene (summary profile) of each module was kept for further analyses. The next step was relating these eigengenes to external traits or, in this case, expression vectors of our genes of interest.

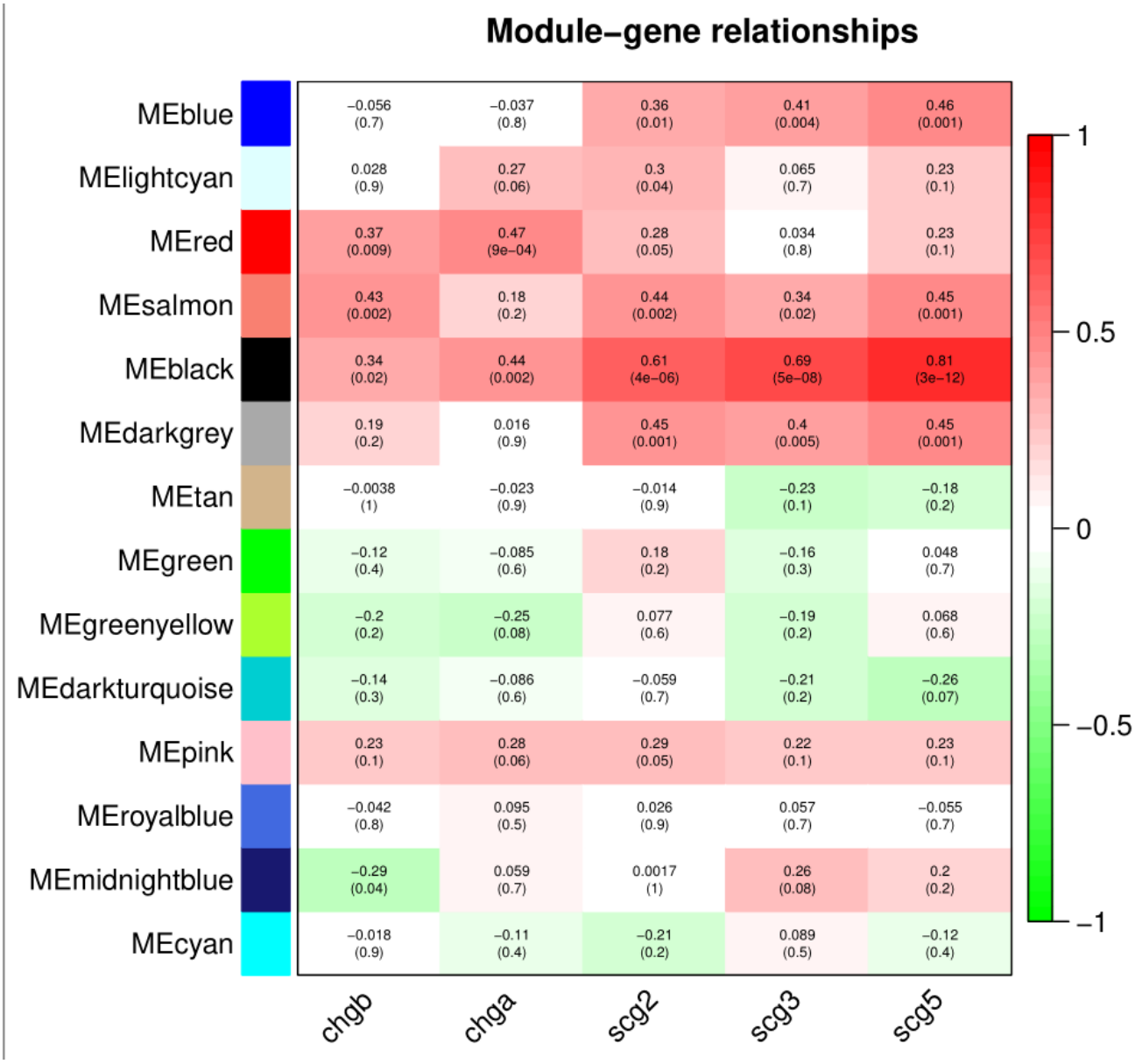


Figure 41: Module-gene associations, first part. Rows correspond to module eigengenes, columns to genes of interest. Each cell displays the corresponding correlation and p-value. Color coding is established by correlation according to the legend.

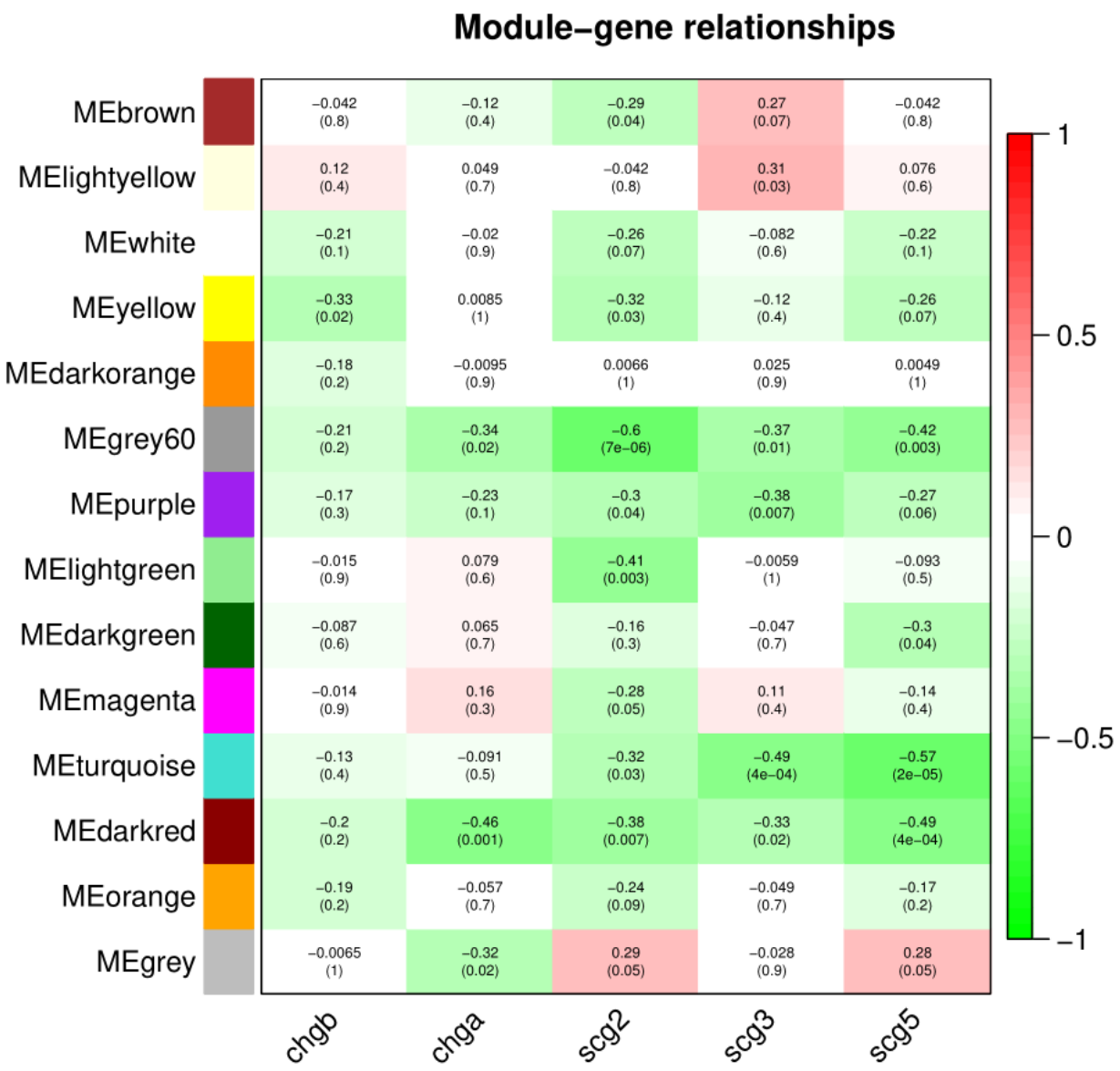


Figure 42: Module-gene associations, second part. Rows correspond to module eigengenes, columns to genes of interest. Each cell displays the corresponding correlation and p-value. Color coding is established by correlation according to the legend.

The black module eigengene (ME) was highly positively correlated with the genes of interest, especially the secretogranins. In contrast, dark red and grey60 MEs were highly negatively correlated with the genes of interest. Interestingly, we can observe that a number of MEs tended to be highly correlated with SCG2-3-5 but less correlated with CHGA and CHGB; namely, darkgrey, blue and black MEs. A similar pattern was evident for turquoise ME but the direction of the correlation was the opposite (Figure 42). The most correlated Mes have been indicated in Table 4.

Table 4: Most correlated MEs for each gene of interest.

CHGB	Red, Salmon, Black
CHGA	Red, Black
SCG2	Blue, Salmon, Black, Dark grey
SCG3	Blue, Salmon, Black, Dark grey, Light yellow
SCG5	Blue, Salmon, Black, Dark grey

SCG2-3-5 genes were highly correlated with the same modules, with the exception of light-yellow ME which was only related to SCG3. In parallel, CHGA and CHGB were most correlated to almost exactly the same set of MEs, however, CHGA is the only gene of interest that does not have >0.3 correlation with the salmon ME.

The authors of the WGCNA package propose that we can quantify associations of individual genes with our genes of interest by defining Gene Significance GS as (the absolute value of) the correlation between the probe on the array and the genes of interest. For each module, we also can define a quantitative measure of module membership MM as the correlation of the module eigengene and the gene expression profile [59]. By utilizing GS and MM measures, it is possible to identify genes that are highly significantly related to our genes of interest as well as high module membership in interesting modules.

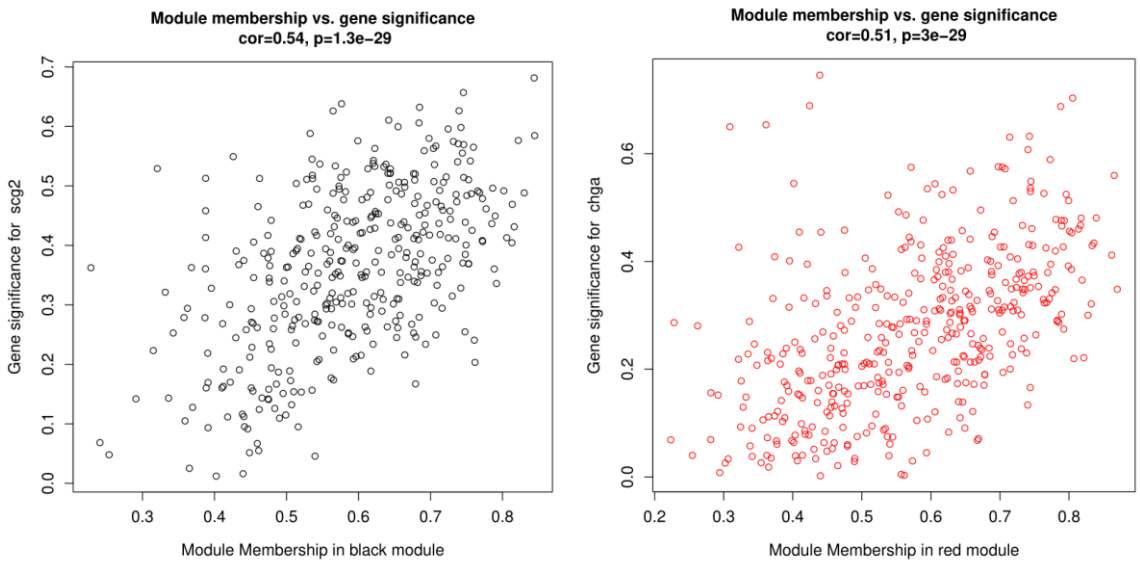


Figure 43: Scatterplot of Gene Significance (GS) for SCG2 and CHGA for Module Membership (MM) in the black and red modules, respectively.

SCG2 was most correlated to black ME and CHGA was most correlated to red ME. There was a highly significant correlation between GS and MM in these respective modules, suggesting that genes highly significantly associated with a gene of interest were often the ones that were the central elements of modules associated with that gene of interest (Figure 43).

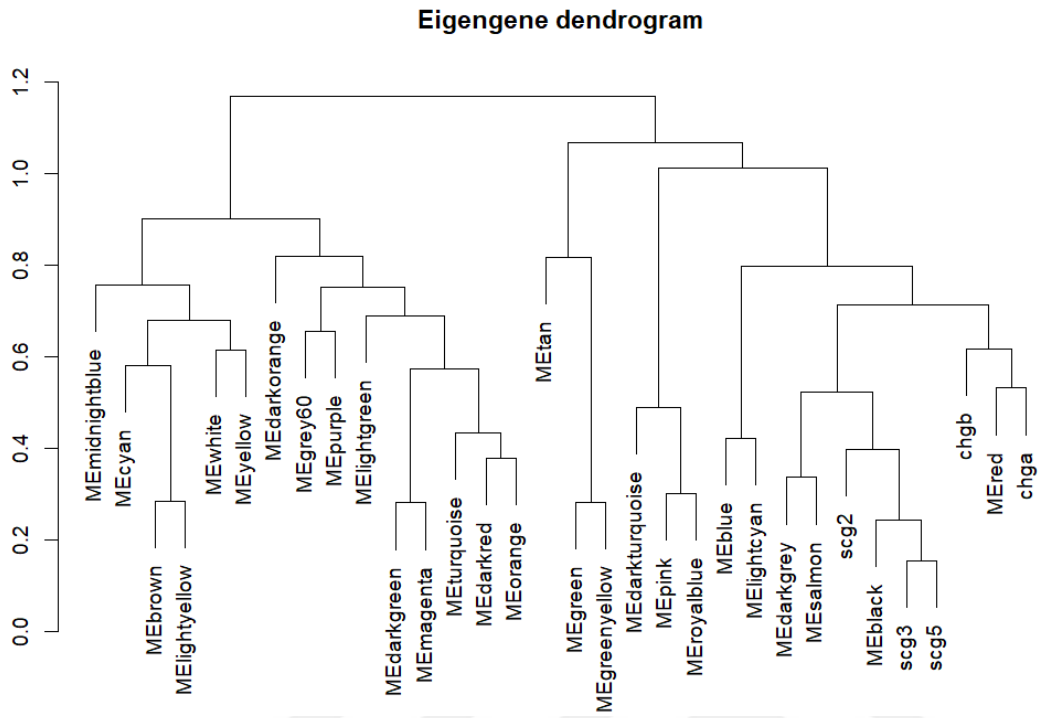


Figure 44: Hierarchical clustering dendrogram of the eigengenes and the genes of interest.

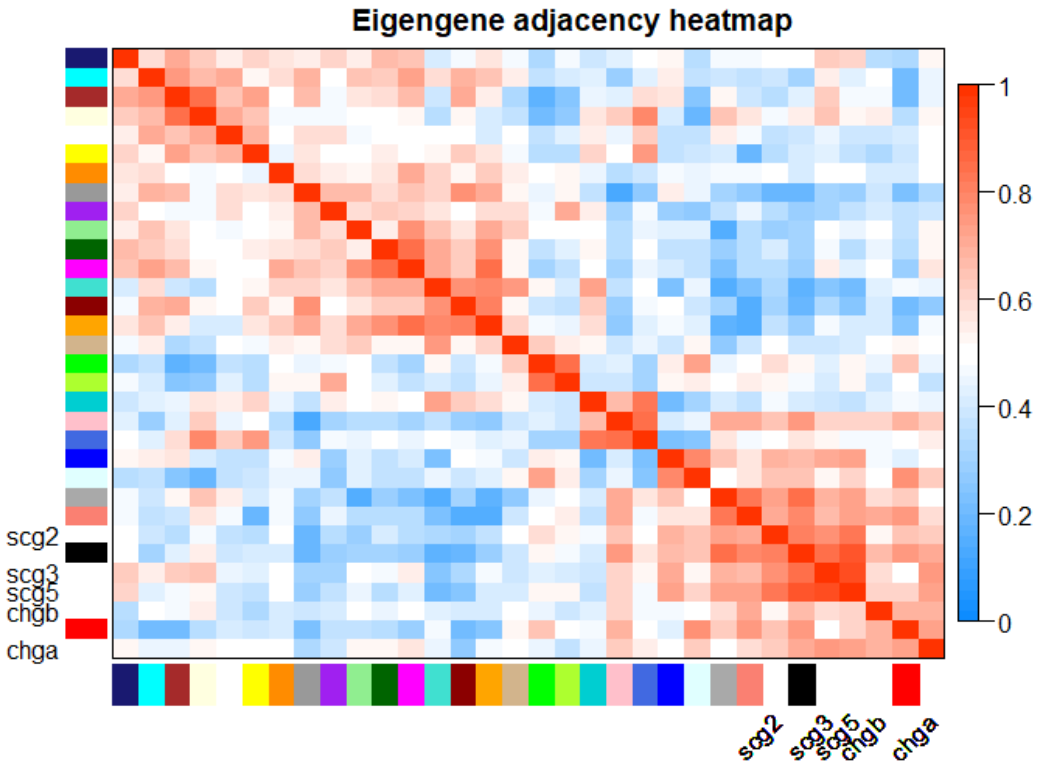
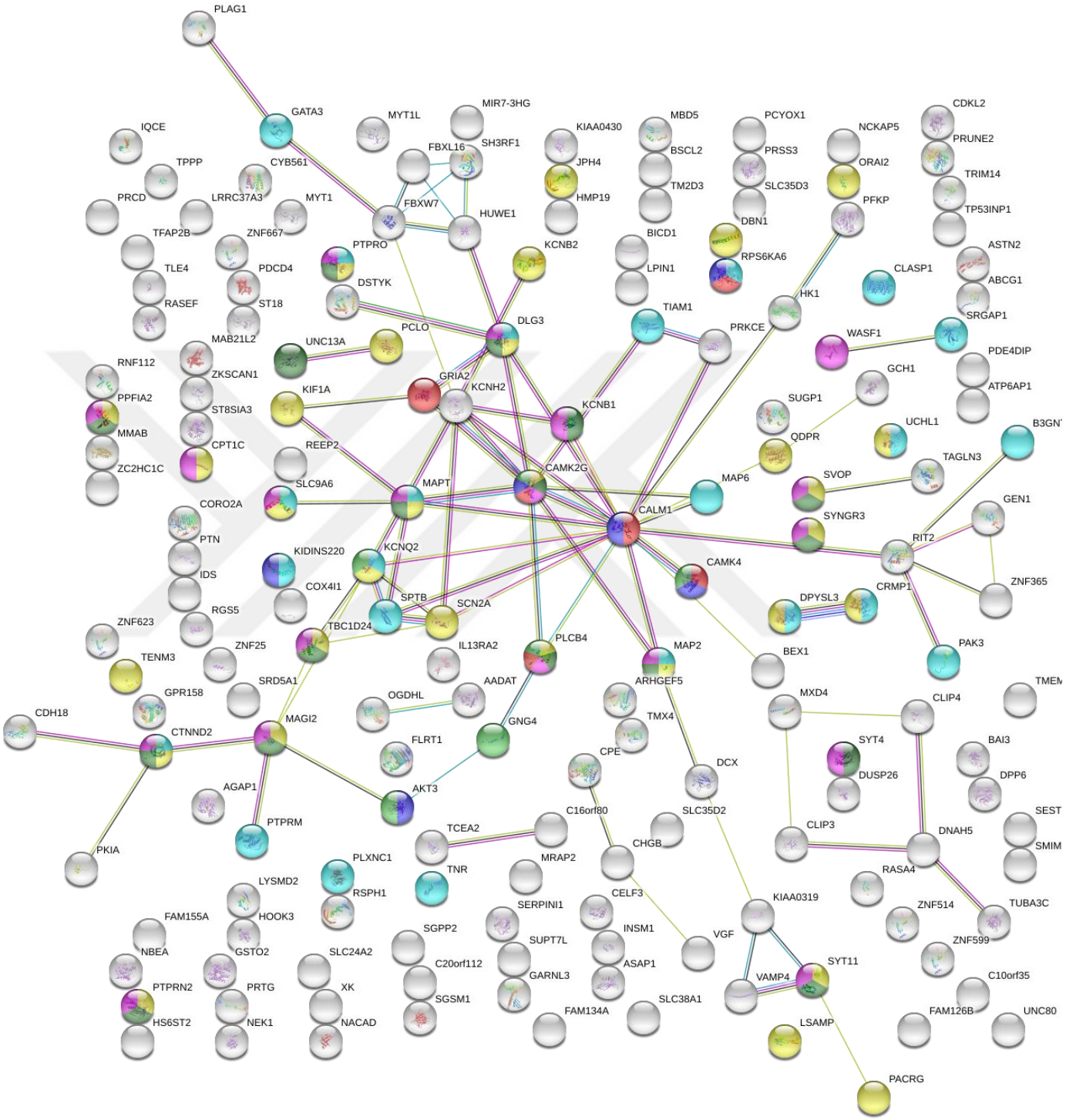


Figure 45: The heatmap of eigengene adjacency with gene of interests added.

The eigengene dendrogram and heatmap can identify clusters of correlated eigengenes also defined as *meta-modules* (Figure 45). For instance, the dendrogram shows that dark grey and salmon modules were highly related; their mutual correlations were stronger than their correlation with SCG2 (Figure 44). Nevertheless, dark grey, salmon, black and SCG2-3-5 genes could be classified in the same meta-module. However, the blue module which was also significantly correlated to SCG2-3-5 was not in the same meta-module as black, salmon and dark grey. On the other hand, both CHGA and CHGB were associated with red ME (Figure 44). Adjacency map also puts all genes of interest in the same cluster while SCG3-5 were more associated with black and CHGA-B were more associated with red modules. At this point, a data frame containing the following information for all probes was obtained; probe ID, gene symbol, module color, gene significance for genes of interests and p-values in all modules. The data frame is also suitable for use in .csv file format outside of R, for example, to export a list of genes from interesting modules with high gene significance values.


PLAG1
GATA3
MYT1L
MIR7-3HG
IQCE
FBXL16
SH3RF1
KIAA0430
MBD5
PCYOX1
CDKL2
NCKAP5
PRUNE2
TPPP
CYB561
JPH4
BSCL2
PRSS3
ORAI2
TRIM14
PRCD
LRRC37A3
MYT1
FBXW7
HMP19
TM2D3
SLC35D3
PFKP
TP53INP1
TFAP2B
ZNF667
HUWE1
DBN1
BICD1
PTPRO
KCNB2
RPS6KA6
ASTN2
TLE4
PDCD4
LPIN1
CLASP1
ABCG1
DSTYK
HK1
RASEF
ST18
TIAM1
SRGAP1
DLG3
PRKCE
MAB21L2
PCLO
WASF1
UNC13A
PDE4DIP
RNF112
ZKSCAN1
GCH1
GRIA2
KCNH2
ATP6AP1
PPFIA2
KIF1A
KCNB1
SUGP1
ST8SIA3
UCHL1
B3GN
MMAB
QDPR
CPT1C
REEP2
TAGLN3
ZC2HC1C
CAMK2G
MAP6
SVOP
SLC9A6
MAPT
GEN1
CORO2A
CALM1
SYNGR3
PTN
KIDINS220
KCNQ2
RIT2
IDS
CAMK4
COX4I1
ZNF365
SPTB
SCN2A
CRMP1
DPYSL3
RGS5
ZNF623
TBC1D24
PLCB4
MAP2
PAK3
ZNF25
IL13RA2
BEX1
TENM3
SRD5A1
AADAT
ARHGEF5
TMEM
GPR158
OGDHL
MXD4
CLIP4
CDH18
MAGI2
GNG4
TMX4
CTNND2
FLRT1
SYT4
BAI3
CPE
DCX
DUSP26
DPP6
AGAP1
AKT3
SLC35D2
C16orf80
SEST
PTPRM
TCEA2
CLIP3
DNAH5
CHGB
SMIM
PKIA
PLXNC1
MRAP2
RASA4
TNR
LYSMD2
RSPH1
CELF3
KIAA0319
ZNF514
TUBA3C
FAM155A
HOOK3
SERPINI1
INSM1
SGPP2
VGF
ZNF599
NBEA
GSTO2
SUPT7L
SLC24A2
ASAP1
SYT11
C20orf112
C10orf35
PTPRN2
PRTG
XK
GARNL3
VAMP4
SGSM1
SLC38A1
FAM126B
UNC80
HS6ST2
NEK1
NACAD
FAM134A
LSAMP
PACRG

Network Stats

number of nodes:	178	expected number of edges:	53
number of edges:	86	PPI enrichment p-value:	1.45e-05
average node degree:	0.966	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.236		

Functional enrichments in your network

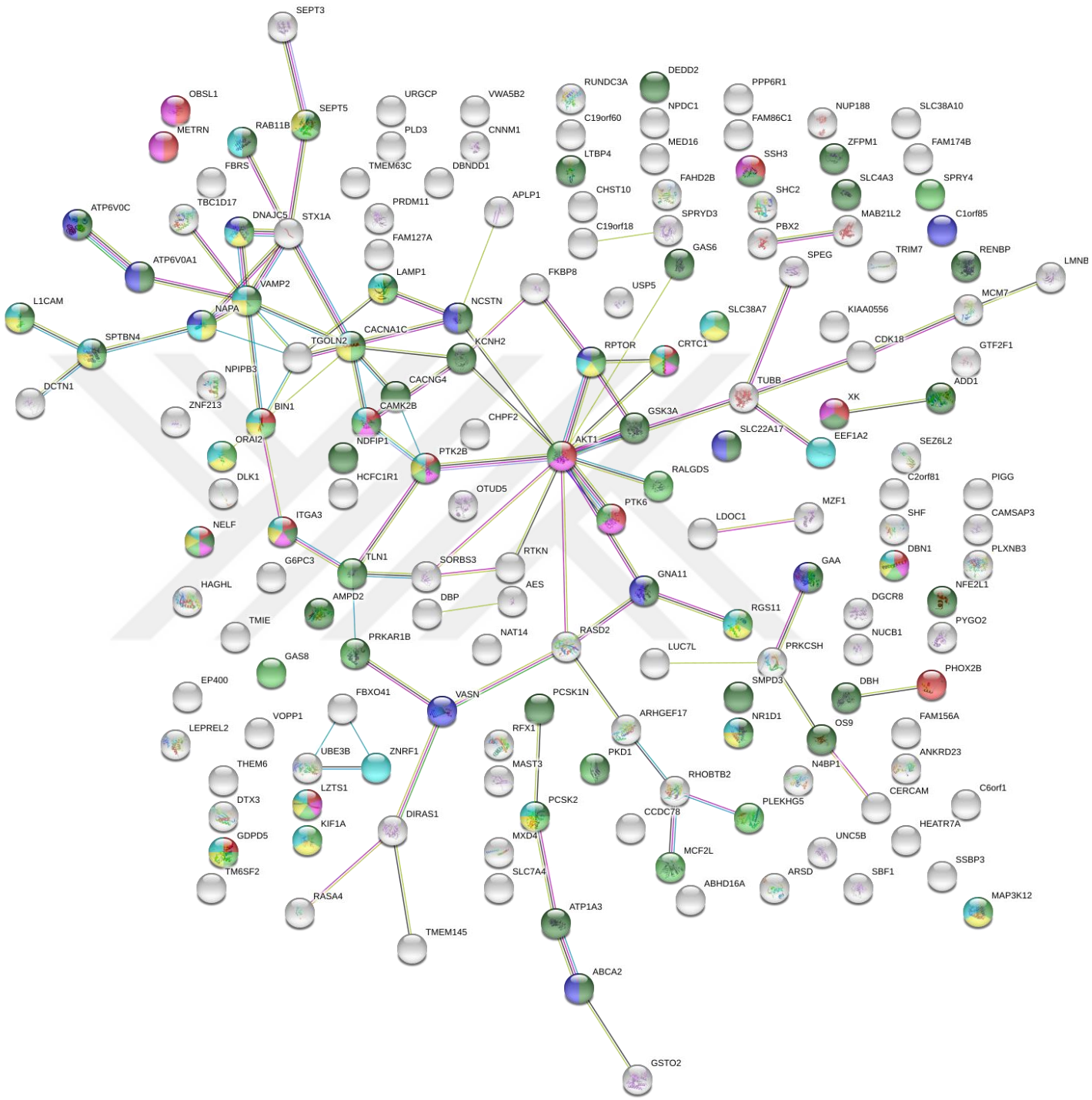
Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0048858	cell projection morphogenesis	29	2.69e-08
GO:0048667	cell morphogenesis involved in neuron differentiation	23	3.28e-07
GO:0048812	neuron projection morphogenesis	23	3.94e-07
GO:0030030	cell projection organization	29	3.67e-06
GO:0000902	cell morphogenesis	28	4.8e-06
			(more ...)

Cellular Component (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0043005	neuron projection	27	1.38e-06
GO:0097458	neuron part	31	1.38e-06
GO:0042995	cell projection	34	0.000206
GO:0044456	synapse part	17	0.00044
GO:0045202	synapse	19	0.00044
			(more ...)

KEGG Pathways			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04720	Long-term potentiation	6	0.00516
04722	Neurotrophin signaling pathway	6	0.049
04725	Cholinergic synapse	6	0.049

Figure 46: PPI network of black module genes with top 200 Gene Significance.

Black module genes with top 200 GS were filtered and obtained using the aforementioned data frame table. As seen in Figure 31, the black ME had a high correlation with all of the genes of interest but mainly SCG2-3-5. Black ME KEGG pathway enrichment was specific to long term potentiation, neurotrophin signaling and cholinergic signaling while CALM1 was at the center of the highly connected PPI (Figure 46).


SEPT3
OBSL1
METRN
SEPT5
RAB11B
FBRS
URGCP
PLD3
TMEM63C
VWA5B2
CNNM1
DBNDD1
RUNDC3A
C19orf60
LTBP4
DEDD2
NPDC1
MED16
PPP6R1
FAM86C1
SSH3
NUP188
SLC38A10
ZFPM1
FAM174B
SLC4A3
SPRY4
TBC1D17
ATP6V0C
DNAJC5
STX1A
PRDM11
FAM127A
APLP1
CHST10
FAHD2B
SHC2
SPRYD3
C19orf18
PBX2
MAB21L2
C1orf85
ATP6V0A1
GAS6
SPEG
TRIM7
RENBP
LMNB
VAMP2
LAMP1
FKBP8
USP5
L1CAM
NAPA
NCSTN
SLC38A7
KIAA0556
MCM7
SPTBN4
TGOLN2
CACNA1C
KCNH2
RPTOR
CRTC1
CDK18
GTF2F1
DCTN1
NPIPB3
CACNG4
TUBB
ADD1
ZNF213
BIN1
CAMK2B
XK
ORAI2
NDFIP1
CHPF2
GSK3A
SLC22A17
EEF1A2
SEZ6L2
PTK2B
AKT1
HCFC1R1
RALGDS
C2orf81
PIGG
DLK1
OTUD5
PTK6
MZF1
LDOC1
SHF
CAMSAP3
NELF
ITGA3
DBN1
PLXNB3
G6PC3
TLN1
SORBS3
RTKN
GAA
GNA11
NFE2L1
HAGHL
AMPD2
DBP
AES
DGCR8
RGS11
PYGO2
TMIE
NAT14
RASD2
NUCB1
PRKAR1B
LUC7L
PRKCSH
GAS8
PHOX2B
EP400
SMPD3
DBH
FBXO41
VASN
PCSK1N
NR1D1
FAM156A
VOPP1
ARHGEF17
OS9
LEPREL2
RFX1
UBE3B
ZNRF1
PKD1
ANKRD23
THEM6
MAST3
N4BP1
RHOBTB2
LZTS1
CERCAM
C6orf1
DTX3
PCSK2
PLEKHG5
CCDC78
KIF1A
DIRAS1
HEATR7A
GDPD5
MXD4
UNC5B
MCF2L
TM6SF2
SLC7A4
SSBP3
ARSD
SBF1
ABHD16A
RASA4
ATP1A3
MAP3K12
TMEM145
ABCA2
GSTO2

Network Stats

number of nodes:	169	expected number of edges:	64
number of edges:	89	PPI enrichment p-value:	0.00197
average node degree:	1.05	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.287		

Functional enrichments in your network

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0045664	regulation of neuron differentiation	16	0.00681
GO:0065008	regulation of biological quality	46	0.00681
GO:0010975	regulation of neuron projection development	13	0.0105
GO:0031344	regulation of cell projection organization	14	0.0347

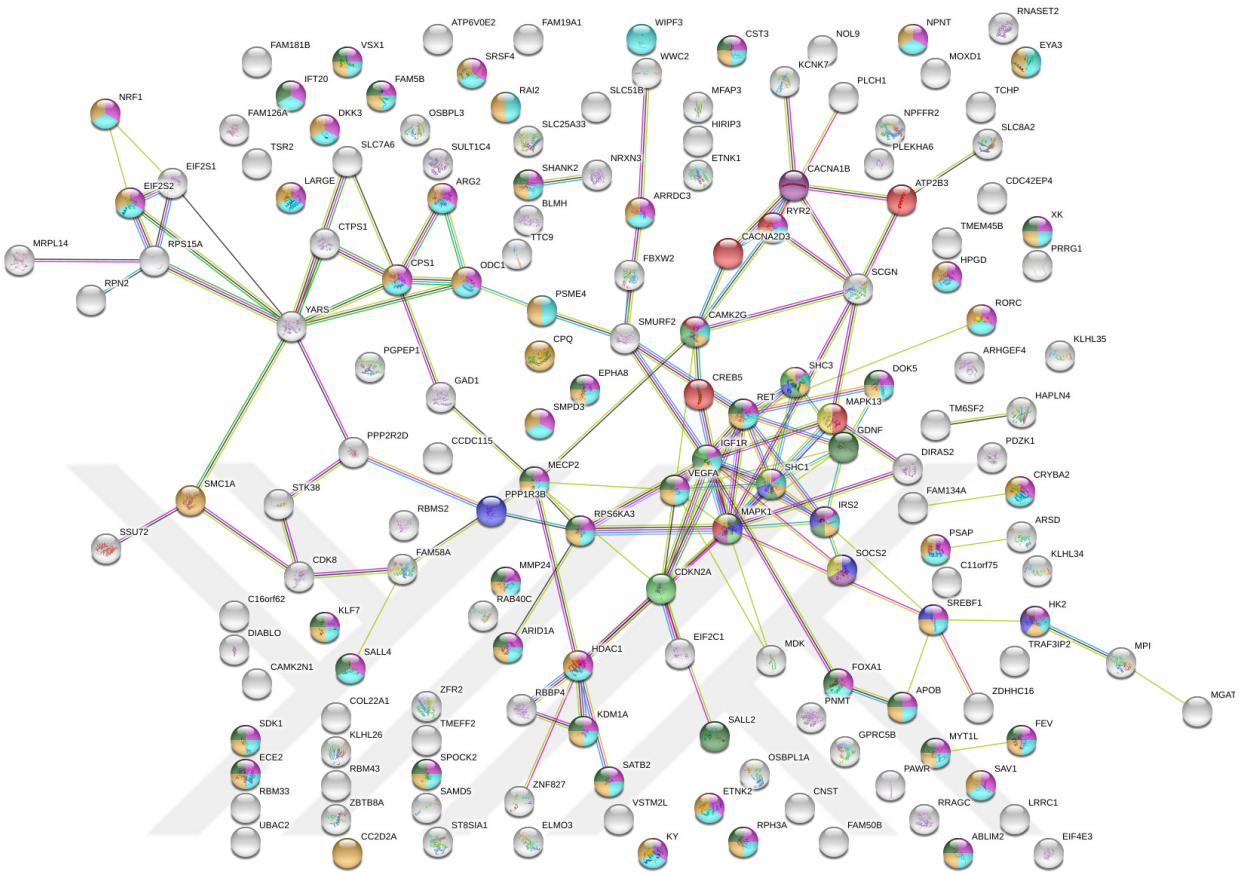
Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0042995	cell projection	34	5.64e-05
GO:0043005	neuron projection	24	5.64e-05
GO:0097458	neuron part	27	5.64e-05
GO:0005774	vacuolar membrane	12	0.00109
GO:0044463	cell projection part	21	0.00109

(more ...)

Figure 47: PPI network of red module genes with top 200 GS.

The red ME was correlated more with CHGA and CHGB than the other secretogranins, rather opposite of the trend that was observed with the black ME. However, since both MEs were positively correlated with the genes of interest, it is expected to see similar functional enrichment elements in PPI networks of genes with high GS from these two modules (Figure 44-45). No KEGG pathway was significantly enriched for the red ME yet biological processes (GO) included regulation of neuron differentiation and neuron projection development indicating the role of chromogranins in these two processes (Figure 47). AKT1 and VAMP2 were among the highly connected nodes.



Network Stats

number of nodes:	162	expected number of edges:	99
number of edges:	125	PPI enrichment p-value:	0.00604
average node degree:	1.54	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.286		

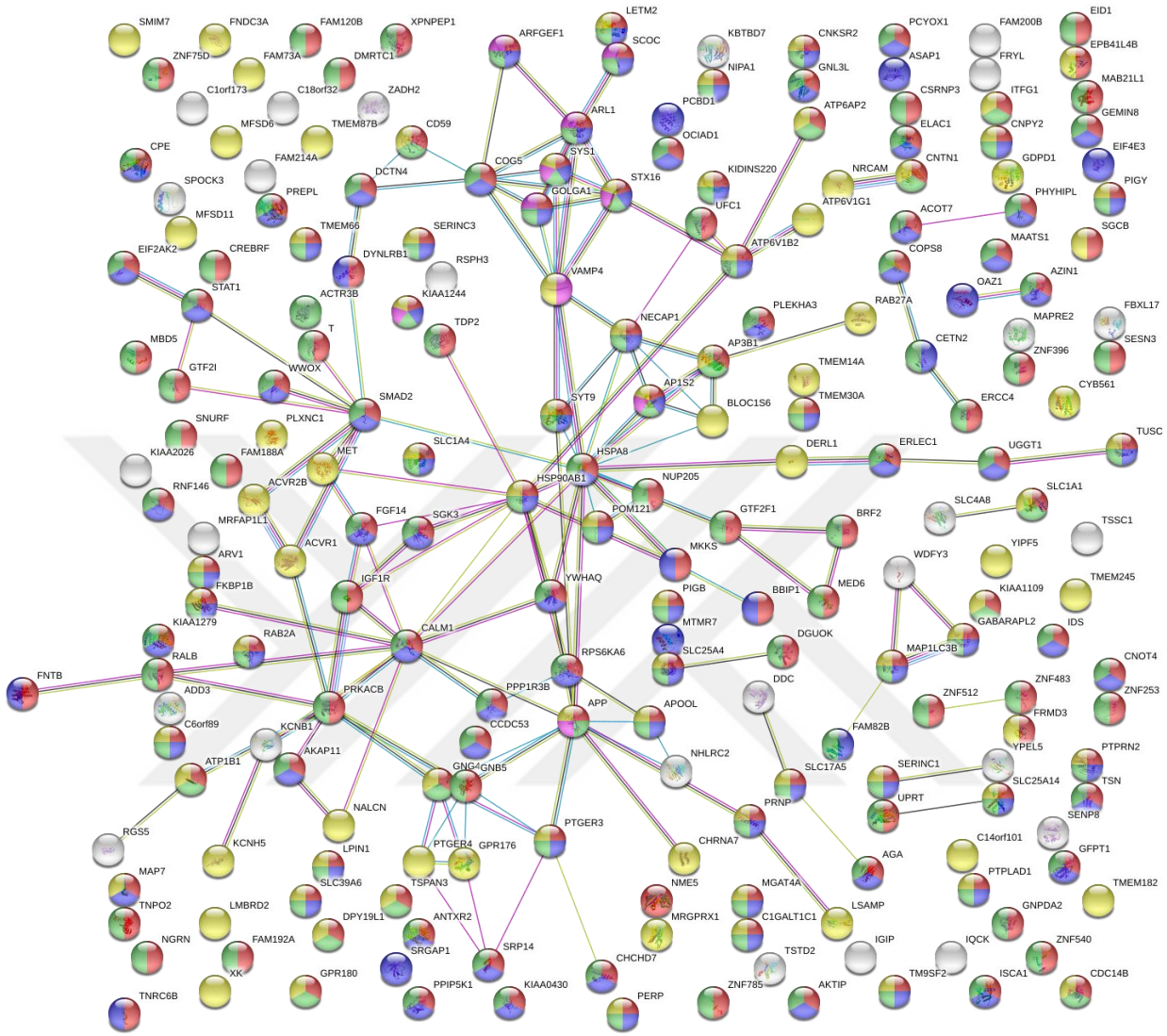
Functional enrichments in your network

Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0048731	system development	54	0.000248
GO:0007275	multicellular organismal development	58	0.000319
GO:0048856	anatomical structure development	57	0.000666
GO:0007399	nervous system development	35	0.00159
GO:0044767	single-organism developmental process	60	0.00159
			(more ...)

KEGG Pathways			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
04910	Insulin signaling pathway	8	0.0015
05214	Glioma	6	0.0015
04930	Type II diabetes mellitus	5	0.00265
04261	Adrenergic signaling in cardiomyocytes	7	0.00957
04917	Prolactin signaling pathway	5	0.0105
			(more ...)

Figure 48: PPI network of salmon module genes.

Salmon module was one of the less crowded modules, therefore, I included all the genes that were a member of the salmon module for PPI network construction (Figure 48). The salmon ME was less related to SCG2-3-5 than the black ME was, as it was evident in Figure 40 and Figure 42 - module-gene relationships and eigengene dendrogram, respectively - but it was still one of the most related MEs to the genes of interest. The dark grey ME seems to be very closely related to the salmon ME but there are simply not enough members in the dark grey module to consider it a significant group of genes. Salmon ME identified a distinct group of genes involved in pathways such as insulin signaling and adrenergic and prolactin signaling in the KEGG enrichment analysis (Figure 48). These indicate that SCG2-3-5 could be involved in diabetes physiology and pathology and its involvement in neurodegenerative diseases.



Network Stats

number of nodes:	218	expected number of edges:	128
number of edges:	148	PPI enrichment p-value:	0.0411
average node degree:	1.36	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.316		

Functional enrichments in your network

Cellular Component (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0043226	organelle	151	0.00236
GO:0044444	cytoplasmic part	104	0.00236
GO:0043227	membrane-bounded organelle	144	0.00254
GO:0005802	trans-Golgi network	10	0.00391
GO:0044425	membrane part	92	0.00391
			(more ...)

Figure 49: PPI network of blue module genes with top ~200 GS.

The blue module has been interesting because it was highly correlated with SCG2-3-5 but it was neither highly positively nor negatively correlated with CHGA or CHGB (Figure 42). This indicated that members of the blue module contribute to the divergence between co-expression networks of CHGA-CHGB and SCG2-3-5 in PCC. The blue ME included organelle enriched genes some of which belonged to trans-Golgi network and other membrane bound organelles.

Conclusions and Discussion

In the present thesis I have developed a streamlined *Shiny* tool that a) takes a GEO dataset, downloads *.cel* files and normalizes the data; b) allows selection of expression matrix for a given set of genes; c) enables scatterplot of correlation coefficients of each gene with any other gene against each other; d) makes possible meta-analysis of selected GEO datasets; e) integrates WCGNA protocol for analysis of modules associated with the selected set of genes. The name of this Shiny application is CoEX-PM and it represents a unique integrated approach to analyze a group of related (either evolutionarily or functionally or structurally) genes in their co-expression divergence. In the literature there is no online application that can do all the steps (a-e) and thus this thesis is an important addition to the relevant gene expression analysis literature. The CoEX-PM application also bypasses the need for coding; bridging the gap between researchers with no programming knowledge and very powerful yet demanding packages written in R language. This tool provides alternative methods of analyzing the same data, as well as enabling the user to perform the same analyses with a wide variety of data, therefore it is unique in terms of flexibility of use.

In addition, I have used CoEX-PM to analyze 5 related chromogranins/secretogranins in terms of their co-expression networks in two separate neuronal cancers, pheochromocytoma (PCC) and neuroblastoma (NB). I have selected multiple datasets from GEO and demonstrated the use of CoEX-PM. One of the most important findings of the study was clustering of SCG2-5 and CHGA-B into separate clusters although all of the genes of interest were relatively similar in behavior. I have obtained PPI networks for different modules of WCGNA indicating the functional contributions of each module into the function of selected secretogranins/chromogranins. This is the first study in the literature that analyzed selected chromogranins and their co-expression divergence in these two cancers. For validation purposes, all the analyses that have taken place for the case study part were also done outside of the *Shiny* platform, on the R console using relevant data and the results were consistent.

I have identified multiple hubs (highly connected genes) which might lead to further studies with respect to chromogranin/secretogranin family. Among these is a group of genes that

were consistently reoccurring in PPI networks for both CHGA and CHGB in both PCC and neuroblastoma: DBH-TH-DDC-CHRNA3 gene cluster. These genes are elements of the dopamine-norepinephrine biosynthesis pathway, each gene encoding an enzyme that regulates the production of important neurotransmitters [73]. Since adrenal gland origin PCCs tend to be catecholamine secreting tumors, expression of genes involved in catecholamine biosynthesis is expected to be altered. However, the aforementioned hub was present across PCCs and neuroblastoma samples in CHGA and CHGB correlated PPI networks and partially conserved in SCG2-3-5 correlated PPI networks. Therefore, the dopaminergic hub indicates a common co-expression pattern for neural crest-derived cancer types. On a related note, metanephrine – a metabolite that is produced as a result of inactivation of epinephrine via methylation – has been reported to be a marker for PCCs and PGLs. Among several alternatives, urine or plasma levels of metanephrines was proved to be the most accurate reporter [74]. In addition, a study testing multiple catecholamines in terms of diagnostic sensitivity in neuroblastoma, identified that normetanephrine (synthesized via inactivation of norepinephrine) was the most sensitive single marker for neuroblastoma at a rate of 89% [75]. The authors of the study also reported that their urinary catecholamine metabolite panel consisting of 8 metabolites had the highest diagnostic sensitivity at 95%. My results suggest that dopaminergic markers can be important in NET since their synthesis are coupled with chromo/secretogranin expression suggesting presence of a tight co-regulation. Another consistent sub-group across most PPIs was the synaptophysin (*SYP*)-calmodulin (*CALM1*) hub which consisted of genes that have a role in synaptic transmission. *SYP* gene is present in all neuroendocrine cells [76]. This hub was reoccurring across all genes of interest in PCC, however, it was partially present in neuroblastoma. Instead of the whole re-occurring group, a set of genes within the hub was retained. Genes that encode microtubules were re-occurring more in neuroblastoma. Namely, some of the genes in the microtubule hub were *TUBB2A, MAPT1, TUBBA1A, MAP1B*. This particular hub was especially enriched for SCG2 even though it was consistent for other genes of interest, except CHGA where it was absent. Microtubules are an integral component of mitosis and intra-cellular transportation; therefore, it is vital to understand their relation to cancer [77].

Looking at the data generated from pairwise co-expression vector correlation plots and meta-correlation analyses, CHGA-CHGB co-expression is less evident in neuroblastoma when

compared to PCC. The correlation between the two genes was not very high in neuroblastoma, each secretogranin-CHGB pair had a higher correlation than the CHGA counterpart. CHGB meta co-expression PPI networks were more enriched in neuroblastoma and had more common elements with the PPI networks of other secretogranins of interest. Therefore, the pairwise correlation and meta-correlation results were in parallel, indicating that CHGB expression pattern has diverged from CHGA more in neuroblastoma dramatically than it is the case in PCC. In addition, there was a sharp contrast in the pairwise correlation plots of neuroblastoma datasets GSE16476 and GSE13136. After reproducing the same results outside of Shiny on the R console and double-checking the parameters, the most feasible explanation for such a distinct pattern could be the composition of the samples. Unfortunately, for most of the neuroblastoma datasets I have used, there is a lack of detailed phenotypic data. I can only speculate that disease stage could be different between samples from different studies and since neuroblastoma is a very aggressive cancer affecting very young individuals, the expression levels were altered significantly. Another possibility is that due to the small sample size of GSE13136 (30 samples), the false positives were unidentified and contributed to the apparent high correlation.

Lung carcinoids and Merkel cell carcinomas conventionally display neuroendocrine characteristics. In parallel, the pairwise co-expression signature was present in the GSE39612 (Merkel cell, Figure 17) and GSE35679 (lung carcinoid, Figure 15) datasets whereas a distinct pattern of correlation was observed in the GSE12667 (lung adenocarcinoma, Figure 16). Carcinoids are a type of tumor originating in neuroendocrine cells whereas carcinoma arises from epithelial tissue [78] [79]. This therefore nicely explains the higher correlation between the chromo/secretogranin co-expression profiles in carcinoids than in carcinomas in CoEX-PM analyses. Therefore, tumors with same site of origin displays distinct characteristics in terms of secretogranin co-expression profiling. Interestingly, a distinct yet highly correlated co-expression pattern with the granin family was observed in pairwise plots of Merkel cell carcinoma samples. However, the clinical utility of blood level of CgA has been tested and found to be ineffective as a diagnostic tool for Merkel cell carcinomas. The CgA blood levels did not correlate with survival or recurrence [80].

One of the most prominent problems of analyzing microarray data is that there are

thousands of probes and only tens of samples, in other words, there are too many variables and very few data points. For instance; accepting a common false positive rate of 0.05 in array platform with over 50,000 probes would mean over 2500 genes are seen as a differential element due to chance. In order to identify false positives at a significantly high success rate, i.e. to increase the specificity of a test, I have utilized meta-correlation analyses and WGCNA. Both operations fundamentally possess high sensitivity, albeit the sensitivity is established via distinct methods. In a meta-correlation analysis, or any type of meta-analysis, since multiple datasets are merged, the number of variables does not change but the number of samples or data points increase. Therefore, the test becomes more statistically robust in terms of sensitivity or identifying false positives.

The meta-correlation pipeline I have implemented in the CoEX-PM application involves the *metacor* package that uses a random-effect meta-analytical approach with correlation coefficients as effect sizes. In a fixed-effect meta-analysis, it is assumed that observed differences among study results are due merely to chance and there is no statistical heterogeneity. In contrast, in a random-effects model, the assumption is made that the observed differences in different studies follow some distribution rather than being identical [81]. Most commonly preferred distribution is a normal distribution. Since the differences are considered as random, the issue of heterogeneity becomes obsolete.

WGCNA method ultimately decreases the dimensional variability of the data by constructing tens of modules from thousands of genes. Therefore, it is easier to identify elements that are affecting distinct expression patterns. The modules consist of highly inter-connected genes and the expression signature of each module could be summarized by a single module eigengene [82]. The module eigengene is the vector that is the best representative of the expression level of a module. It is calculated as the first principal component (PCA) or the first right-singular vector (SVD) of the module expression data [83]. The module eigengenes are used to relate modules to each other and external information. Furthermore, each module can be treated as a node - represented by its eigengene – and their connections can be identified using similar network construction operations. The module eigengenes can also be clustered within a biologically meaningful meta-network of modules as shown in Figure 44. A higher-order

organization of the transcriptome is revealed when investigating clusters of related modules rather than modules of inter-connected genes. Therefore, it might be significant to compare eigengene networks between different states, as well as gene co-expression networks [59]. In the case study, I have observed the modules that were highly correlated with the genes of interest contained members of hubs that were consistently observed across meta co-expression PPI networks in PCC. However, the hub genes were not all classified in a single module, rather they were partially present in each module that was highly correlated to the genes of interest. Interestingly, CHGA and CHGA were observed in the same meta-module whereas, SCG2-3-5 were clustered together in a distinct meta-module.

There have been several publications on Shiny applications dealing with meta-analysis; *MetaGenyo* is an application that facilitates the meta-analysis of genetic association studies [84]. The application satisfies a need in the literature for a tool that can be used to investigate genetic variants and phenotype associations. Another application dealing with meta-analyses is *MareyMap* Online. The application offers tools to estimate meiotic recombination rates [85]. Both applications provide users with useful tools with the extra achievement of bypassing the need for coding. However, neither of these applications include calculations of meta-correlation analysis of microarray data. As far as we know, there have been no Shiny application published that offers the range of customization and the detail of visualization that the CoEX-PM does for conducting a weighted gene co-expression network analysis.

The case study findings are summarized as follows:

- The pairwise correlation plots showed that chromogranin/secretogranin family members have common correlation partners and similar correlation patterns in PCC and other tumors that have been known to show NET characteristics.
- Meta-correlation results identified consistently reoccurring distinct hubs of interconnected genes in PCC and neuroblastoma. The meta-correlation results table was generated and saved locally using the meta-correlation tab, then filtered by relevant *r.mean* values to get lists of significant genes. The lists were separately uploaded to STRING in order to obtain PPI networks. This was the only step that was done exclusively outside of Shiny or R. Several common hubs in neuroendocrine tumor types

were identified, implying their importance underlying the disease state. Interestingly, the CHRNA3-DDC-TH-DBH group were present in some meta co-expression PPI networks in both PCC and neuroblastoma. In PCC, CHGA-CHGB meta co-expression networks tended to be similar whereas SCG2-3-5 meta co-expression networks were contained more common elements with each other.

- In the WGCNA part, I have used the GSE67006 PCC dataset and constructed modules with a soft-thresholding power of 5. A number of modules were related to the genes of interest. The related modules contained the hubs that were observed in the meta co-expression PPI networks in PCC. Furthermore, clustering the module eigengenes and the genes of interest showed that CHGA-CHGB pair and SCG2-3-5 trio were observed in tandem but in distinct meta modules (eigengene clusters), validating the findings in the meta-correlation analysis.

Future Prospects

Data intensive research techniques have become common place in every aspect of biological research, the need for tools to analyze and manipulate the data has been increasing correspondingly. The R programming language and the R community provided researchers with a wide range of robust and versatile tools, however, the steep learning curve to be able to utilize R has limited its usefulness for many researchers. The CoEX-PM application presented in this article aims to remove this limitation and enable researchers with no programming background to interactively visualize data and conduct analyses. There is a massive collection of packages developed for biological data analysis; designing interactive and user-friendly applications that utilize these packages would increase their accessibility to a wider pool of users.

For future studies, an application that conducts PCA with a given set of genes and an expression matrix could be designed to see which inputted genes group together. Also, an enrichment protocol could be implemented to determine whether the genes that are grouped together share complementary ontology. Another potentially useful application would be one that facilitates similar operations that the CoEX-PM does but analyses RNA-seq data. Furthermore, since the CoEX-PM application demands high computational power in terms of memory and processing, it could be potentially modified and broken into parts to make it possible to run online. In addition, implementing a gene network construction feature – similar to Cytoscape - to the current CoEX-PM application could prove beneficial. The coloring scheme used in the pairwise correlation plots could be improved in order to convey more meaningful information about the probes; rather than using the mean expression level of each probe across all samples, variance or range of each probe across all samples could be utilized as a metric and coloring scale. Using CoEX-PM, other sets of paralogous genes can be analyzed in different contexts. The ability of CoEX-PM to download data directly from GEO in this stand-alone application allows the researcher to diversify the datasets and generalize conclusions. In addition, a web-based application of CoEX-PM is under development which will take the user data in the form of series expression matrix by manual upload function. All of the tasks and protocols mentioned above as potential improvements could be done directly through R directly but the process to achieve the final result starting with raw data is not very simple. Co-ex-PM *Shiny* package changes this process from coding tasks to interacting with a GUI to accomplish the same task.

References

1. Koeppen, K., Stanton, B. A. & Hampton, T. H. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx452
2. Dumas, J., Gargano, M. A. & Dancik, G. M. ShinyGEO: A web-based application for analyzing gene expression omnibus datasets. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw519
3. Theodosiou, T. *et al.* NAP: The Network Analysis Profiler, a web tool for easier topological analysis and comparison of medium-scale biological networks. *BMC Res. Notes* (2017). doi:10.1186/s13104-017-2607-8
4. Mallona, I., Díez-Villanueva, A. & Peinado, M. A. Methylation plotter: A web tool for dynamic visualization of DNA methylation data. *Source Code for Biology and Medicine* (2014). doi:10.1186/1751-0473-9-11
5. Nasir, A. & Coppola, D. *Neuroendocrine tumors: Review of pathology, molecular and therapeutic advances*. *Neuroendocrine Tumors: Review of Pathology, Molecular and Therapeutic Advances* (2016). doi:10.1007/978-1-4939-3426-3
6. Comprehensive, N. & Network, C. Neuroendocrine Tumors. (2014).
7. Barakat, M., Yalcin, S. & Oberg, K. *Neuroendocrine tumours*. *Springer Heidelberg New York Dordrecht London* **1,** (2015).
8. Fraenkel, M., Kim, M. K., Faggiano, A. & Valk, G. D. Epidemiology of gastroenteropancreatic neuroendocrine tumours. *Best Pract. Res. Clin. Gastroenterol.* **26,** 691–703 (2012).
9. Modlin, I. M., Moss, S. F., Chung, D. C., Jensen, R. T. & Snyderwine, E. Priorities for improving the management of gastroenteropancreatic neuroendocrine tumors. *Journal of the National Cancer Institute* **100,** 1282–1289 (2008).
10. Yao, J. C. *et al.* One hundred years after 'carcinoid': Epidemiology of and prognostic factors for neuroendocrine tumors in 35,825 cases in the United States. *J. Clin. Oncol.* **26,** 3063–3072 (2008).
11. Modlin, I. M. *et al.* Gastroenteropancreatic neuroendocrine tumours. *The Lancet Oncology* **9,** 61–72 (2008).
12. Barakat, M. T., Meeran, K. & Bloom, S. R. Neuroendocrine tumours. *Endocrine-Related Cancer* **11,** 1–18 (2004).
13. Modlin, I. M., Lye, K. D. & Kidd, M. A 5-decade analysis of 13,715 carcinoid tumors. *Cancer* **97,** 934–959 (2003).
14. Soga, J. Early-stage carcinoids of the gastrointestinal tract: An analysis of 1914 reported cases. *Cancer* **103,** 1587–1595 (2005).
15. Korse, C. M., Taal, B. G., Van Velthuysen, M. L. F. & Visser, O. Incidence and survival of neuroendocrine tumours in the Netherlands according to histological grade: Experience of two decades of cancer registry. *Eur. J. Cancer* **49,** 1975–1983 (2013).
16. Castro-Vega, L. J., Lepoutre-Lussey, C., Gimenez-Roqueplo, A.-P. & Favier, J. Rethinking pheochromocytomas and paragangliomas from a genomic perspective. *Oncogene* **35,** 1080–1089 (2016).
17. Hoehner, J. C. *et al.* A developmental model of neuroblastoma: Differentiating stroma-poor tumors' progress along an extra-adrenal chromaffin lineage. *Lab. Investig.* **75,** 659–675 (1996).
18. London, W. B. *et al.* Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the Children's Oncology Group. *J. Clin. Oncol.* **23,** 6459–6465 (2005).
19. Ries, L. a. G. *et al.* Cancer incidence and survival among children and adolescents: United States SEER Program 1975-1995. *NIH Pub. No. 99-4649* 179 pp. (1999).
20. Maris, J. M. 2010. Maris. Recent Advances in Neuroblastoma. *N. Engl. J. Med.* **362,** 2202–2211 (2010).
21. Yamamoto, K. *et al.* Spontaneous regression of localized neuroblastoma detected by mass screening. *J. Clin. Oncol.* **16,** 1265–1269 (1998).

22. Maris, J. M., Hogarty, M. D., Bagatell, R. & Cohn, S. L. Neuroblastoma. *Lancet* **369,** 2106–2120 (2007).
23. Mossé, Y. P. *et al.* Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455,** 930–935 (2008).
24. Bourdeaut, F. *et al.* Germline mutations of the paired-like homeobox 2B (PHOX2B) gene in neuroblastoma. *Cancer Letters* **228,** 51–58 (2005).
25. Yao, J. C. *et al.* Chromogranin A and neuron-specific enolase as prognostic markers in patients with advanced pNET treated with everolimus. *J. Clin. Endocrinol. Metab.* (2011). doi:10.1210/jc.2011-0666
26. Murphy, D. GENE EXPRESSION STUDIES USING MICROARRAYS: PRINCIPLES, PROBLEMS, AND PROSPECTS. *Adv. Physiol. Educ.* (2002). doi:10.1152/advan.00043.2002
27. Bumgarner, R. Overview of dna microarrays: Types, applications, and their future. *Curr. Protoc. Mol. Biol.* (2013). doi:10.1002/0471142727.mb2201s101
28. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41,** (2013).
29. Kolesnikov, N. *et al.* ArrayExpress update-simplifying data submissions. *Nucleic Acids Res.* **43,** D1113–D1116 (2015).
30. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40,** 3785–3799 (2012).
31. Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24,** 374–382 (2008).
32. Cahan, P. *et al.* Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene* **401,** 12–18 (2007).
33. Larsson, O., Wennmalm, K. & Sandberg, R. Comparative Microarray Analysis. *Omi. A J. Integr. Biol.* **10,** 381–397 (2006).
34. Carter, S. L., Eklund, A. C., Mecham, B. H., Kohane, I. S. & Szallasi, Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* **6,** 107 (2005).
35. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43,** D234–D239 (2015).
36. Hulsen, T., Huynen, M., de Vlieg, J. & Groenen, P. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* **7,** R31 (2006).
37. Dahia, P. L. M. *et al.* A HIf1α regulatory loop links hypoxia and mitochondrial signals in pheochromocytomas. *PLoS Genet.* **1,** 0072–0080 (2005).
38. Kaelin, W. G. The von Hippel-Lindau gene, kidney cancer, and oxygen sensing. *J. Am. Soc. Nephrol.* **14,** 2703–2711 (2003).
39. Brooks, J. D., Levy, M. A., Lovly, C. M. & Pao, W. Translational genomics : The challenge of developing cancer biomarkers paradigm Translational genomics : The challenge of developing cancer biomarkers. 183–187 (2012). doi:10.1101/gr.124347.111
40. Modlin, I. M., Drozdov, I. & Kidd, M. Gut neuroendocrine tumor blood qpcr fingerprint assay: Characteristics and reproducibility. *Clin. Chem. Lab. Med.* **52,** 419–429 (2014).
41. Modlin, I. M., Bodei, L. & Kidd, M. Neuroendocrine tumor biomarkers: From monoanalytes to transcripts and algorithms. *Best Pract. Res. Clin. Endocrinol. Metab.* **30,** 59–77 (2016).
42. Feero, W. G., Guttmacher, A. E., McDermott, U., Downing, J. R. & Stratton, M. R. Genomics and the Continuum of Cancer Care. *N. Engl. J. Med.* **364,** 340–350 (2011).
43. Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *N. Engl. J. Med.* **360,** 790–800 (2009).

44. Annaratone, L. *et al.* Search for Neuro-Endocrine Markers (Chromogranin A, Synaptophysin and VGF) in Breast Cancers. An integrated Approach Using Immunohistochemistry and Gene Expression Profiling. *Endocr. Pathol.* **25,** 219–228 (2013).
45. Prestifilippo, a & Blanco, G. Chromogranin A and Neuroendocrine Tumors. *Intechopen* (2012). doi:10.1016/j.endonu.2012.10.003
46. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).
47. Modlin, I. M. *et al.* Neuroendocrine tumor biomarkers: Current status and perspectives. *Neuroendocrinology* **100,** 265–277 (2014).
48. Tumer, K. & Ghosh, J. Estimating the Bayes error rate through classifier combining. in *Proceedings - International Conference on Pattern Recognition* (1996). doi:10.1109/ICPR.1996.546912
49. Montero-Hadjadje, M. *et al.* Chromogranins A and B and secretogranin II: Evolutionary and functional aspects. *Acta Physiol.* **192,** 309–324 (2008).
50. Borges, R. *et al.* Granins and catecholamines. Functional interaction in chromaffin cells and adipose tissue. *Adv. Pharmacol.* **68,** 93–113 (2013).
51. Bartolomucci, A. *et al.* The extended granin family: Structure, function, and biomedical implications. *Endocrine Reviews* (2011). doi:10.1210/er.2010-0027
52. A??t-Ali, D. *et al.* Molecular characterization of frog chromogranin B reveals conservation of selective sequences encoding potential novel regulatory peptides. *FEBS Lett.* **511,** 127–132 (2002).
53. Portela-Gomes, G. M. & Stridsberg, M. Selective processing of chromogranin A in the different islet cells in human pancreas. *J. Histochem. Cytochem.* **49,** 483–490 (2001).
54. Metz-Boutigue, M. H., Goumon, Y., Lugardon, K., Strub, J. M. & Aunis, D. Antibacterial peptides are present in chromaffin cell secretory granules. *Cellular and Molecular Neurobiology* **18,** 249–266 (1998).
55. Ferrero, E. *et al.* Chromogranin A protects vessels against tumor necrosis factor alpha-induced vascular leakage. *FASEB J.* **18,** 554–556 (2004).
56. Crippa, L. *et al.* A new chromogranin A-dependent angiogenic switch activated by thrombin. *Blood* **121,** 392–402 (2013).
57. Zhang, D. *et al.* Two chromogranin a-derived peptides induce calcium entry in human neutrophils by calmodulin-regulated calcium independent phospholipase A2. *PLoS One* **4,** (2009).
58. Shehwana, H. Ph.D. Thesis. (2017).
59. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** (2005).
60. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* (2004). doi:10.1093/bioinformatics/btg405
61. Loriot, C. *et al.* Epithelial to mesenchymal transition is activated in metastatic pheochromocytomas and paragangliomas caused by SDHB gene mutations. *J. Clin. Endocrinol. Metab.* **97,** 954–962 (2012).
62. Łastowska, M. *et al.* Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. *Oncogene* **26,** 7432–7444 (2007).
63. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).
64. Valentijn, L. J. *et al.* TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47,** 1411–1414 (2015).
65. Ohtaki, M. *et al.* A robust method for estimating gene expression states using Affymetrix microarray probe level data.

BMC Bioinformatics **11,** 183 (2010).

66. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455,** 967–970 (2008).
67. Winston Chang, Joe Cheng, JJ Allaire, Y. X. and J. M. shiny: Web Application Framework for R. R package version 1.1.0. (2018). Available at: https://cran.r-project.org/package=shiny.
68. Hauke, J. & Kossowski, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* (2011). doi:10.2478/v10117-011-0021-1
69. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* (2011). doi:10.1111/j.1541-0420.2011.01616.x
70. Laliberté, E. metacor: Meta-analysis of correlation coefficients. *R package version 1.0-2.* (2011). Available at: https://cran.r-project.org/package=metacor.
71. Schulze, R. *Meta-analysis: a comparison of approaches.* (Hogrefe & Huber, Gottingen, Germany., 2004).
72. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv007
73. Daubner, S. C., Le, T. & Wang, S. Tyrosine hydroxylase and regulation of dopamine synthesis. *Archives of Biochemistry and Biophysics* (2011). doi:10.1016/j.abb.2010.12.017
74. Scholz, T., Schulz, C., Klose, S. & Lehnert, H. Diagnostic management of benign and malignant pheochromocytoma. *Exp. Clin. Endocrinol. Diabetes* **115,** 155–9 (2007).
75. Verly, I. R. N. *et al.* Catecholamines profiles at diagnosis: Increased diagnostic sensitivity and correlation with biological and clinical features in neuroblastoma patients. *Eur. J. Cancer* (2017). doi:10.1016/j.ejca.2016.12.002
76. Wiedenmann, B., Franke, W. W., Kuhn, C., Moll, R. & Gould, V. E. Synaptophysin: a marker protein for neuroendocrine cells and neoplasms. *Proc. Natl. Acad. Sci. U. S. A.* (1986). doi:10.1073/pnas.83.10.3500
77. Kavallaris, M. Microtubules and resistance to tubulin-binding agents. *Nature Reviews Cancer* (2010). doi:10.1038/nrc2803
78. Swarts, D. R. A., Ramaekers, F. C. S. & Speel, E. J. M. Molecular and cellular biology of neuroendocrine lung tumors: Evidence for separate biological entities. *Biochimica et Biophysica Acta - Reviews on Cancer* (2012). doi:10.1016/j.bbcan.2012.05.001
79. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* (2014). doi:10.1038/nature13385
80. Gaiser, M. R. *et al.* Evaluating blood levels of neuron specific enolase, chromogranin A, and circulating tumor cells as Merkel cell carcinoma biomarkers. *Oncotarget* (2015). doi:10.1038/jid.2015.69
81. Higgins JPT, G. S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. *The Cochrane Collaboration* (2011). Available at: www.handbook.cochrane.or.
82. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* (2007). doi:10.1186/1752-0509-1-54
83. Wall, M., Rechtsteiner, A. & Rocha, L. Singular value decomposition and principal component analysis. *A Pract. Approach to Microarray Data Anal.* (2003). doi:10.1007/0-306-47815-3_5
84. Martorell-Marugan, J., Toro-Dominguez, D., Alarcon-Riquelme, M. E. & Carmona-Saez, P. MetaGenyo: A web tool for meta-analysis of genetic association studies. *BMC Bioinformatics* (2017). doi:10.1186/s12859-017-1990-4
85. Siberchicot, A., Bessy, A., Guéguen, L. & Marais, G. A. B. Marey map online: A user-friendly web application and database service for estimating recombination rates using physical and geneticmaps. *Genome Biol. Evol.* (2017). doi:10.1093/gbe/evx178