Development and exploitation of GeneFriends: An online database for gene and transcript co-expression analysis

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

By

Sipko van Dam

February 2017

Abstract

Although many diseases have been well characterized at the molecular level, the underlying mechanisms often remain unclear. This may be attributed to the large number of genes for which it remains unknown in which biological processes and diseases they play a role. Genes involved in the same biological processes and diseases are often co-expressed, which information can be used to predict the biological process a poorly annotated gene likely plays its primary role in. With this purpose, we constructed a co-expression network from a large number of microarray and RNA-seq samples. We conclude that co-expression analysis can be used to postulate the functions of both coding and non-coding genes. Additionally, it can be used to predict diseases they likely play an important role in. It is also shown that gene-function predictions based on a co-expression network that is constructed on a transcript rather than gene level can differentiate between different functions of transcripts originating from the same gene. We have created an online resource, GeneFriends, the first online resource that utilizes a co-expression network constructed from RNA-seq data, also allowing users to query for co-expression at the transcript rather than gene level. This allows researchers to identify and prioritize novel candidate genes and transcripts involved in biological processes and complex diseases. This is a valuable resource to the research community as supported by usage of GeneFriends in a number of independent publications. GeneFriends is available online at: http://GeneFriends.org/.

To validate the ability of our tool to identify genes that are relevant to diseases, we tested GeneFriends by conducting a co-expression analysis with seed lists for aging, cancer, and mitochondrial complex I disease. We identified several candidate genes that have previously been predicted as relevant targets for each of these diseases. Some of the identified genes

were already being tested in clinical trials supporting the effectiveness of this approach. Furthermore, two of the novel candidates of unknown function that were identified by GeneFriends as co-expressed with cancer genes were selected for experimental validation. Knock-down of the human homologs (*C1ORF112* and *C12ORF48*) of these two candidate genes in HeLa cells slowed proliferation suggesting that these genes indeed play a role in cancer growth.

Co-expression analyses often lead to large lists of gene-disease associations without a clear indication which genes are most relevant for follow up studies. To select such relevant genes, those that are important nodes in a co-expression network are often identified under the notion that these are of higher biological relevance than the others. To validate if this method selects the most relevant genes for aging, we conduct a co-expression analysis on a rat thymus dataset and identified transcription factors that are important network nodes. Whilst literature supports that some of these transcription factors may be important regulators of the aging process, this method can also miss some of the most interesting intervention targets.

Lastly, in a rat brain aging RNA-seq dataset, generated in our lab, we tested if we could identify co-expression modules for which the expression correlates with aging and investigate if we can identify dietary interventions that potentially affected this correlation. Although modules were identified that correlated with aging, no significant effect of the dietary interventions for any of these modules was detected. Additionally, this dataset contained detailed information about the expression of microRNAs in addition to the whole transcriptome data. This was utilized to investigate if expression of microRNAs and their targets are negatively correlated, which we did not observe.

Acknowledgements

First I would like to thank my supervisor, João Pedro de Magalhães for his guidance during this project. I am very grateful for the freedom with which I was allowed to conduct my research and the guidance with the writing and publishing of my papers and the opportunity to draft a grant proposal. Furthermore, I would like to thank Rui Cordeiro for conducting the wet-lab work side of the project validating the effectiveness of the tool. I would like to thank Thomas Craig for doing the design of the GeneFriends website and supplying me with the .css file entailing this design, as well as maintaining the servers on which most of the GeneFriends work was conducted and the hosting of the web tool. I would like to thank my brother for many fruitful discussions and support during my project. Additionally, I would like to thank him for my initial programming knowledge which I owe to his undeniable devotion to the art of programming. I would like to thank Shona for helping me draft my first manuscript and teach me the dos and don'ts of publishing. I would also like to thank Daniel Wuttke for his support and assistance with the microarray based co-expression section of this work and Brad T. I am grateful to Sherman from the DAVID support team for the analysis on the Inferred from Expression Pattern (IEP) code usage in categories used in the DAVID enrichment tool's default settings. I would like to thank Michael Keane, Michael Stevens, Gianni Monaco and Robi Tacutu for fruitful discussions and assistance during this project. I would like to thank Aoife Doherty for her assistance with the writing of the manuscript, describing the RNA-seq based co-expression network, and the writing of the co-expression review. I would like to thank John Herbert, Jay Hinton and Karsten Hokamp for testing and fruitful suggestions on the ReadCounter tool. I would like to thank everyone who has made my stay in Liverpool an enjoyable time. I have experienced some challenges with the language editing of this thesis

and would like to thank Nilouq Stoker, Shona Wood, Aoife Doherty, Kate McIntryre, Urmo Võsa, Monique van der Wijst and João Pedro de Magalhães for helping me with the language editing of parts of this thesis. Finally I would like to thank the Biotechnology and Biological Sciences Research Council for funding this GeneFriends, and I would like to thank the University of Liverpool, the Institute of Integrative Biology specifically, for covering the fees for my PhD project.

Table of contents

List of Figures

List of Tables

List of abbreviations

ANOVA – Analysis of Variance

AUC – Area Under the Curve

AUROC – Area Under the Receiver Operator Curve

BLAST – Basic Local Alignment Search Tool

BSN – Biological Scaling Normalization

BWA – Burrows-Wheeler Aligner

ChIP – Chromatin Immunoprecipitation

CI – Mitochondrial Complex I

CDNA – Complementary DNA

DAVID – Database for Annotation and Visualization and Integrated Discovery

DICER – Differential Correlation in Expression for meta-module Recovery

DREAM – Dialogue for Reverse Engineering Assessments and Methods

ENCODE – Encyclopedia of DNA Elements

EQTL – Expression Quantitative Trait Loci

ES – Enrichment Score

FDR – False Discovery Rate

FPKM – Fragments Per Kilobase of transcript per Million mapped reads

GBA – Guilt-By-Association

GB – Gigabyte

GEO – Gene Expression Omnibus

GFF – General Feature Format

GO – Gene Ontology

GTEx – Genotype-Tissue Expression

GTF – General Transcript Format

GWAS – Genome Wide Association Study

HGNC – HUGO Gene Nomenclature Committee

HISAT – Hierarchical Indexing for Spliced Alignment of Transcripts

HUGO – Human Genome Organization

ID – Identifier

IEP – Inferred from Expression Pattern

LAP – Liver Activating Protein

LIP – Liver Inhibiting Protein

LncRNA – Long ncRNA

MCL – Markov Clustering Algorithm

MCODE – Molecular Complex Detection

ncRNA – non-coding RNA

NGS – Next Generation Sequencing

OMIM – Online Mendelian Inheritance in Man

PC – Principal Component

PCA – Principal Component Analysis

PDAC – Pancreatic Ductal Adenocarcinoma

PSI-BLAST – Position-Specific Iterative BLAST

RefSeq – Reference Sequences

RISC – RNA-Induced Silencing Complex

RMA – Robust Multi-array Average

RNA-seq – RNA sequencing

siRNA – small Interfering RNA

snRNA – small nuclear RNA

snoRNA – small nucleolar RNA

SNP – Single Nucleotide Polymorphisms

SRA – Sequence Read Archive

STAR – Spliced Transcripts Alignment to a Reference

TF – Transcription Factor

TFBS – Transcription Factor Binding Site

TMM – The trimmed Mean of M-values normalization Method

TPM – Transcripts Per Million

WGCNA – Weighted Gene Co-expression Network Analysis

Chapter 1: Introduction

The number of identified genes has tripled over the last few years. Even though most of these new genes do not encode proteins they can play an important role in gene regulation and disease. For most of these non-coding genes no information is available about their function. Genome wide analyses, such as RNA sequencing (RNA-seq) experiments or genome sequencing project, may find such genes differentially expressed or mutated. The lack of annotation makes interpretation of any results including such genes difficult. Function predictions for such genes facilitate the interpretation of such results and the design for potential follow-up studies.

The functions of many genes are still not completely understood, an issue that has vastly expanded with the recent identification of many novel non-coding genes [1]. For decades, studying individual genes and their products has provided a wealth of knowledge about the functions and regulatory mechanisms of a wide range of genes [2]. However, it is clear that this reductionist method is inappropriate to fully understand gene functions and regulation of whole systems. Therefore, scientists have favored the development of high-throughput technologies and data-analysis methods to identify the functional and regulatory status of a gene from a systematic perspective [3]. One of these methods is co-expression network analysis, an approach that emerged shortly after the introduction of microarrays to assess genome-wide gene expression. Gene co-expression networks can be used for multiple purposes among which candidate gene prioritization and functional annotation.

Gene co-expression networks can be used to predict in which biological processes a gene likely plays its primary role [4-6], to prioritize candidate disease genes [4, 7-20] or to discern transcriptional regulatory programs [21]. New gene- or disease-function associations in this

thesis are defined as the function or disease annotation of the top 5 percentile co-expression partners of a particular gene being enriched for a particular function or disease based on well-defined annotations such as GO for function predictions and OMIM for disease gene predictions. The expectation is that a gene plays its most prominent role in the biological processes or diseases for which the enrichment is found. With recent advances in transcriptomics, co-expression networks constructed from RNA-seq data should also enable the inference of functions and disease associations for non-coding genes and splice variants, which is the main aim of this thesis. Previous studies have already shown that this is possible for coding RNAs, which we review in this chapter. We discuss different types of co-expression networks, metrics and how co-expression can be used to identify regulators of a network. Lastly, we discuss some of the advantages and controversies associated with co-expression methods.

1.1. Microarrays and RNA-seq data for co-expression analysis

Co-expression information is obtained from large numbers of gene expression snapshots, such as microarray or RNA-seq data from humans and model systems. These platforms describe the activity/expression of each individual gene at a given time point and when many are combined in a co-expression analysis, create a picture of which genes have a tendency to be co-activated. This picture then represents a co-expression network. Since expression data is a prerequisite to co-expression analysis it is not surprising that this type of analysis emerged after the rapid evolution of microarrays in the beginning of this millennium. The rapid growth of expression data has allowed researchers to combine data from different experiments for co-expression network analyses. Different approaches for co-expression analyses were suggested to identify gene functions and causative relationships with important phenotypic parameters [22], soon

after followed by the first co-expression network analyses. To date co-expression analyses has facilitated functional classification of genes [16, 23, 24] as well as identification of genes associated to diseases [8-20] using a Guilt-By Association (GBA) approach, described in Section 1.5. Furthermore, co-expression network analysis has allowed researchers to separate driver from passenger genes, helping prioritize targets for intervention studies [25-28]. Recent developments have allowed co-expression to give more detailed insights into genetic variation that is causative to diseases, which help understanding the mechanisms underlying the diseases and aid the design of intervention studies [5].

1.1.1. Non-coding RNAs: Definition, functions and mechanisms

Co-expression analyses have, up to now, focused on coding genes, due to the limited ability of microarrays to measure expression of non-coding RNAs (ncRNAs). As the name indicates these ncRNAs do not encode proteins, yet many of them are thought to have regulatory roles [29] and to play a role in disease [30, 31]. Unfortunately for most no information is available about the biological process they most likely play a role in, an issue we aim to tackle in this thesis using co-expression analysis. Not all mechanisms through which ncRNAs exert their regulatory role are clear yet, but it is known they can epigenetically modify the DNA by recruiting chromatin remodeling complexes to specific loci [32, 33] or recruit transcription factors directly. Conversely they can bind to promotors to prevent transcription initiation. Additionally, they can modify proteins post transcriptionally by preventing splicing events from occurring through various mechanisms. These mechanisms and more are more elaborately discussed in [34]. There are different subclasses of ncRNAs, such as long non-coding RNAs (lncRNA), microRNAs (miRNA), antisense RNAs, small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). LncRNA are those ncRNAs longer than 200 basepairs. LncRNAs are essential in many physiological processes, such as X-inactivation specific transcript (*Xist*), which

is important for X-chromosome inactivation in mammals [35]. *Xist* was used in this thesis to validate our method. MicroRNAs and antisense RNAs are RNAs that function by binding to mature RNAs of their target genes to degrade these in conjunction with a RNA-induced silencing complex (RISC) complex [36, 37] or simply inhibit their expression effectively downregulating the target gene. SnRNAs and snoRNAs are small RNAs (<50bp) that guide the chemical modifications of other RNAs, such as methylation, and can play a role in splicing [38, 39]. A more detailed distinction between ncRNA types is described in [40] table 1. Although many ncRNAs have been identified it remains unclear how many are functional. GeneFriends helps researchers identify the co-expression partners of such genes and which functions these are enriched for, if any. If there are no such functions we argue it suggests these are less likely to be functional ncRNAs.

Although co-expression networks can be constructed from both microarrays and RNA-seq data, a major benefit of RNA-seq is that it quantifies the expression of the over 70,000 ncRNAs not usually measured with microarrays [1], including many recently annotated lncRNAs. Microarrays nowadays also include probes for ncRNAs, but the number of ncRNAs that is discovered is rapidly expanding and therefore these arrays become quickly outdated. It is also possible to use data from tiling arrays [41], which also measure gene expression of ncRNAs. A study comparing tiling data and RNA-seq data reports that RNA-seq data is more reliable, having a higher dynamic range if more than 4 million reads are sequenced, and should be used as a gold standard [42]. In our work we only use samples in which more than 10 million reads were sequenced and results obtained from these samples are thus likely of higher quality than those that would be obtained from tiling arrays. We also are under the impression that RNA-seq will in the future be more commonly used than microarrays and tiling arrays adding more

power of our future analyses than these two alternatives would. For these reasons we have used RNA-seq data for the identification of co-expression of ncRNAs with coding genes.

When RNA-seq experiments are performed, one of the library preparation steps that is often used is a ribosomal depletion step. This is used to remove the highly abundant ribosomal RNA, which is not of interest. This step achieves its goal by removing short RNAs. However, this also means other short RNAs are removed, such as miRNAs [43]. Thus, to effectively measure the expression of miRNAs, a different protocol should be used, which is not the case for most of the experiments we used to construct our co-expression network. The undetected miRNAs are excluded from our database and analyses conducted in Chapter 3. To assess whether co-expression can be used to detect targets of miRNAs on a genome wide scale, we conducted a separate analysis on an in-house generated rat brain aging dataset. In this dataset miRNAs were isolated using a protocol specifically tailored to this purpose (Chapter 4).

Apart from its increased accuracy when measuring low-abundance transcripts [14], RNA-seq also has other benefits [44]. It distinguishes expression profiles of closely-related paralogues better than microarray-derived profiles [45]. RNA-seq can also distinguish between the expression of different splice variants [46, 47], which can have distinct interaction partners [48] and biological functions [49]. This utility was used in Chapter 3 to construct a transcript level co-expression network. Co-expression analysis on RNA-seq data can assign putative roles to different splice variants and lncRNAs [50], and identify diseases in which they might play a part [50].

1.2. From sample to gene expression

To determine expression levels using microarrays many well established tools are available. Microarrays contain many microscopic spots, each containing probes that are specific for a

gene. Messenger RNAs (mRNAs) of genes are commonly labeled/dyed with a fluorescent label and will bind to the probes with their complementary sequence (provided the gene is expressed). Each microscopic spot can then be read out by a machine that translates the amount of label in the well to a signal representing the expression of the gene for which this spot contains specific probes. [51]

RNA-seq relies on a different strategy to determine the expression of genes. It revolves around the sequencing of the mRNAs present in a sample. Prior to the sequencing, a complementary DNA (cDNA) library is created and amplified to increase the abundance of each mRNA. This is then fragmented and for each fragment a short region is sequenced. This can either be done single end (only sequencing one end of the fragment) or paired end (sequencing both ends), the latter allowing more accurate mapping of the reads [52]. The reads are typically mapped to the respective genome and the number of reads per gene are counted. These counts then represent the expression level of the gene. We have computed a list of tools that can be used by readers that are interested in conducting co-expression analysis with RNA-seq data (Table 1.1), most of which are discussed in the sections below.

Tool/method	Description
Quality control	
FastQC (see further information)	A tool that can be used both with and without a user interface. Uses .fastq .bam or .sam files to identify and highlight potential issues in the data, such as low base quality scores, low sequence quality, GC content biases.
Mappers	
Kallisto [53]	A tool that uses a pseudo-alignment strategy to assigns expression values to transcripts/genes, to achieve optimal speed whilst maintaining comparable accuracy to other tools. Limitation is that it maps to a transcriptome/gene annotation file and does identify new genes that are not annotated in this file. Uses little memory and can be run on a regular desktop computer.
Salmon [54]	Another pseudo-alignment tool that performs comparably to Kallisto.
STAR [55]	A read aligner that maps reads to a genome. Detects splice variants and novel genes. An example shows that this tool maps approximately 50 times faster than Tophat and Tophat2. The tool uses a large amount of memory (approximately 27 GB when mapping to the human genome).
HISAT [56]	A read aligner that maps reads with slightly faster speed with comparable accuracy [57]. The newer HISAT2 version aligns to genotype variants which will likely result in a higher accuracy. HISAT2 will be the core of the next version of Tophat (Tophat3).
Bowtie/Tophat/Tophat2 [58]	The first widely used mapping tool. Detects splice variants and novel genes. Although much slower than most other mappers whilst requiring a relatively large amount of memory and a number of reports stating it maps with a relatively low accuracy, still widely used.
Read count tools	
HTseq [59]	Assigns expression values to genes based on reads that have been aligned with e.g. STAR or HISAT. Well supported by the author.
FeatureCounts [60]	Similar to HTseq, but much faster. Results are slightly different due to slightly different read-to-gene assignment strategies.
Normalization	
FPKM/RPKM [61, 62]	Widely used normalization methods that correct for the total number of reads in a sample whilst also accounting for gene length. TMM has been suggested as a better alternative.
TPM [63]	Similar to FPKM, but normalizes the total expression to a total of 1 million. The summed expression of a TPM normalized samples is thus always 1 million.
TMM [64]	Similar to FPKM/RPKM, but puts these expression measures onto a common scale across different samples
RAIDA [65]	Utilizes ratios between counts between genes in each sample for normalization to avoid problems caused by differential transcript abundance between samples (resulting from differential expression of highly abundant genes transcripts).
Module detection	

WGCNA [66]	Constructs a co-expression network using a user selected method; Pearson correlation by default. Uses hierarchical clustering and has varying tree cutting options to identify modules.
DiffCoEx [67]	Uses a similar approach to WGCNA, but to identify and group similarly differentially co-expressed genes instead of co-expression modules altering in co-expression strength as a whole, creating modules of genes that have the same different partners between different samples.
CoXpress [68]	Identifies co-expression modules, similar go WGCNA, in each sample group and tests if these modules are also co-expressed in other groups.
Biclustering [69]	Identifies modules that are unique to a subset of samples without the need of prior grouping of samples.
GSVD [70]	Identifies "genelets", which can be interpreted as modules representing partial co-expression signal from multiple genes. These signals are then compared between two groups to identify genelets unique to samples and those that are shared with the two groups.
HO-GSVD [71]	Similar to GSVD, but across multiple groups rather than only two.
Functional enrichment	
DAVID [72]	A widely used tool, with an online web interface. Users supply a list of genes and select the annotation categories from various sources to identify enrichment for.
PANTHER[73]	Uses a comprehensive protein library combined with human curated pathways and evolutionary ontology. If a gene is not in the library, a gene is classified based on proteins with conserved sequences for which the function is known.
g:Profiler	Functional enrichment tool, enabling users to perform enrichment analyses for gene ontologies, KEGG pathways, protein-protein interactions, transcription factor- and miRNA binding sites. Also available as R package.
ClusterProfiler Enrichr	R package for overrepresentation and gene set enrichment analyses for several curated gene sets. Allows users to compare the results of analyses performed on several gene sets. Has an easy-to-use web tool for performing gene overrepresentation analyses and using comprehensive set of up-to-date functional annotations.
GSEA [74]	Another widely used tool that can optionally be used with a desktop interface. Uses an extensive collection of geneset annotations and has documentation of the different features available with this tool.
Visualization	
Cystoscape [75]	A widely used tool for visualization of networks that has many plug-ins available that can help further analyze the network.
Biolayout [76]	Similar to Cystoscape, but less widely used and does not have the Cytoscape plug-ins. Can load and visualize much larger networks than Cytoscape.

Table 1.1. Tools for RNA-seq data based network analysis
In this table we describe a number of tools required for the different steps of an RNA-seq co-expression network analysis. This list includes the tools we recommend at the time of writing,

but many others are available and continuously released. As such, we recommend to consult literature to select the most appropriate tools and methods for this type of analyses.

To determine the gene expression levels from RNA-seq sequences, obtained from RNA-seq machines, several tools are available. For quality control of the samples we would advise FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Unless there is a specific interest in unannotated genes, we would recommend Kallisto [53], which can map the reads to the transcriptome on a desktop computer. Otherwise we would recommend tools such as Spliced Transcripts Alignment to a Reference (STAR) [55] or Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT) [56], which map at a much faster rate than TopHat [77] whilst achieving a similar accuracy if not better [55]. Kallisto reports expression levels per transcript, whereas STAR and HISAT report the genomic location of each read. To convert these to expression levels per gene (or transcript/exon) we advise to use FeatureCounts [60] or the much slower, but well supported, HTseq tool [59]. The evolution of these tools is rapid and the tools described here may have been superseded by better tools at the time of reading. As such, we would advise to follow forums such as SEQanswers (http://seqanswers.com/forums/forumdisplay.php?f=26) for the latest developments. In this thesis, we have used STAR in conjunction with FeatureCounts as Kallisto was not readily available at the time we started this project.

Although RNA-seq has many benefits over microarrays, it still has limitations. RNA-seq struggles to determine which splice variant is expressed if multiple splice variants share the same expressed exon. This can be circumvented using knowledge acquired from the mapping of other reads in the same region that do not map to shared exons. For example, if “transcript a” and “transcript b” are partly overlapping and there are e.g. 90 reads ambiguously mapping to both transcripts and simultaneously there are 100 reads unambiguously mapping to “transcript a” and 0 reads to “transcript b”, it is likely the 90 ambiguously mapping reads are originating from “transcript a”. This method is, for example, utilized by Bitseq [78], which uses

a Bayesian approach to estimate the origin of an ambiguously overlapping read based on how many reads have been mapped to each of these transcripts. This Bayesian approach calculates the probability that each read originate from a given transcript based on which transcripts the other reads were assigned to. These probabilities are then used to determine if the transcript is differentially expressed between different samples. SpliceNet is another tool that uses a similar method. SpliceNet effectively divides the reads, mapping to an exon shared with 2 splice-variants, proportionally to the total expression of each of the two whole isoforms [79]. This thus means that a particular isoform is considered to have no expression if it has no expression in any exons that are not shared with any other isoforms, even if the shared exons do have expression. This method may, however, introduce different biases. For example, if these ambiguously overlapping reads are in reality originating from a different transcript than the other reads in their vicinity, which could be biologically relevant. For example, in a hypothetical situation where a large transcript largely overlaps a smaller transcript and some reads overlapped the larger transcript non-ambiguously, the ambiguously overlapping reads would be assigned to the larger gene. Then if the smaller gene has an inhibitory function on the larger gene, which could hypothetically be functioning by binding to the larger gene's transcripts, the incorrect assignment of reads could lead to incorrect biological conclusions (i.e. the large gene's proteins quantity or activity is increased instead of decreased). In this thesis, we have assigned ambiguously mapping reads to each transcript, meaning strong co-expression between transcripts that share the same exons is likely to be reported, but not necessarily biologically meaningful. This is a bias that should be considered when interpreting the results obtained from our web-tool.

1.3. Expression data normalization

Normalization of expression data is necessary for the removal of non-biological variance (introduced by, e.g., different read depths, the use of different preparation protocols, machines and varying environmental variables, such as temperature and humidity), which can introduce biases when attempting to conduct a biologically meaningful comparison between different samples especially those generated in separate batches [80]. Microarray data is commonly normalized using MAS5 [81] or Robust Multi-array Average (RMA) [82]. MAS5 normalizes each array individually based on the average of the perfect-match/mismatch values. RMA, on the other hand, uses information from other arrays in the dataset to normalize each array. MAS5 and RMA have been compared in more detail in [83].

Normalization methods for RNA-seq data are different from those used for microarray data. Also for RNA-seq data there are several different methods that are being used, but not without debate. Fragments/Reads Per Kilobase of transcript per Million mapped reads (FPKM/RPKM) [61, 62] and Transcripts Per Million (TPM) [63] values, are normalization methods that correct for the total number of reads in a sample, as well as gene length [61]. One issue with these methods is that if a very highly expressed gene increases in expression, these normalization methods will make it appear as if all other genes decrease in expression. This is particularly important for co-expression analysis as it will create a false impression of positive co-expression between all these other genes. The commonly used Trimmed Mean of M-values (TMM) normalization of FPKM values and a more recent method [64], Ratio Approach for Identifying Differential Abundance (RAIDA) [65], resolve this issue and are preferable over other normalization methods [80]. Although, unlike TMM normalization, RAIDA can cope with differences in total expression levels of RNA between samples. Both TMM and RAIDA rely on the assumption that the expression of majority of genes is stable across the samples [80],

which may not be the case. For example, in cancer samples this is commonly not the case. Some of these methods were not available at the time we started this project. We corrected for the total number of reads in the sample, but not for gene length as this makes no difference for the resulting Pearson correlation values calculated using our method.

Minimum read depth and sample size required for co-expression analyses. To create co-expression networks from RNA-seq data, a 20-sample minimum has been suggested [66, 84].Increased sample sizes produce networks with a higher connectivity, a term explained in [84]. Not surprisingly, higher quality data tend to result in more accurate co-expression networks [84, 85]. It is therefore essential to select a number of criteria for data quality control. A higher total read depth for RNA-seq samples increases the accuracy of the data, especially for genes with low expression [84, 85]. For RNA-seq data, sequencing depth cut-off thresholds are usually selected arbitrarily. Several co-expression studies have used a cut-off of 10 million reads per sample [50, 84, 86], which was also used in the project described in this thesis. Co-expression networks constructed using a 10 million reads per sample cut-off have been suggested to have a similar quality to microarray-based co-expression networks if constructed from the same number of samples [84], but with decreasing quality with fewer reads. The percentage of mapped reads is another frequently considered cut-off in which samples with less than 70 or 80% of the reads mapping to the genome are removed. Giorgi et al. demonstrated, using 65 *Arabidopsis thaliana* samples with 12 million reads but only a 30% mapping cut-off threshold, that the resulting RNA-seq-based co-expression network had a lower similarity to biological networks than microarray networks [87]. Cut-off thresholds may vary per species, based on, among other factors, the quality of the genome annotation. As more and higher quality data becomes available, higher cut-off thresholds may be preferable. In this thesis, we used a cutoff of 10 million reads and over 4000 samples.

To ensure that a network is robust, bootstrapping can be used [88]. This is the repetitive construction of networks from subsets of the data, which are then used to assess the reproducibility of the network created from the entire data set. Randomizing the dataset (e.g. by randomly reassigning expression values to their gene/transcript identifiers and reconstructing the network) can also help identify correlations that occur stochastically because of specific biases rather than as a result of biologically-relevant interactions [50], which is a method we applied in 99.

1.4. Co-expression networks

A co-expression network identifies which genes have a tendency to show a coordinated expression pattern across a group of samples. This co-expression network can be represented as a gene–gene similarity matrix, which can be used in downstream analyses (Figure 1.1). Canonical co-expression network construction and analyses can be described with the following three steps.

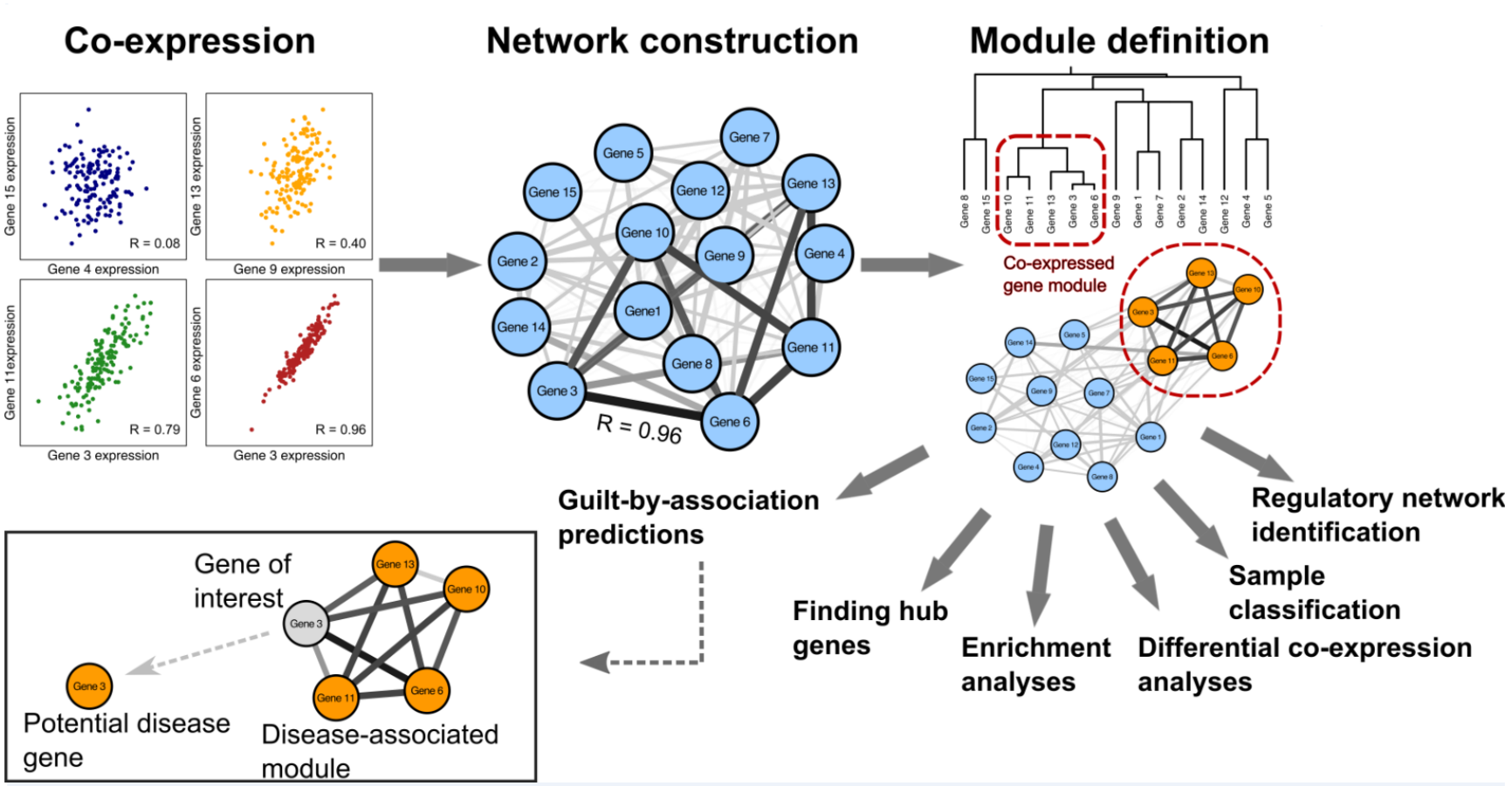

Figure 1.1: Example co-expression network analysis
First, pairwise correlation is determined for each possible gene pair in the expression data. These pairwise correlations can then be represented as a network. Modules within these networks are defined using clustering analysis. The network and modules can be interrogated to identify regulators, functional enrichment and hub genes. Differential co-expression analysis can be used to identify modules that behave differently under different conditions. Potential disease genes can be identified using a guilt-by-association (GBA) approach that highlights genes that are co-expressed with multiple disease genes. This figure was created by Urmo Võsa.

In the first step, individual relationships between genes are commonly defined based on correlation measures or mutual information [89-91] between each pair of genes. These relationships describe the similarity between expression patterns of each possible gene pair across all the samples. Different measures of correlation have been used to construct networks, including Pearson's or Spearman's correlations [19, 92]. Alternatively, least absolute error regression [93] or a Bayesian approach [94] can be used to construct a co-expression network. The latter two have the added benefit that they can be used to identify causal links and have been explained elsewhere [22]. The largest difference between correlation and mutual information is that mutual information can also measure non-monotonic relationships, whereas correlation cannot. In a monotonic relationship one variable increases as the other variable increases, or alternatively one variable decreases as the other increases, whereas this is not the case for non-monotonic relationships (Figure 1.2).

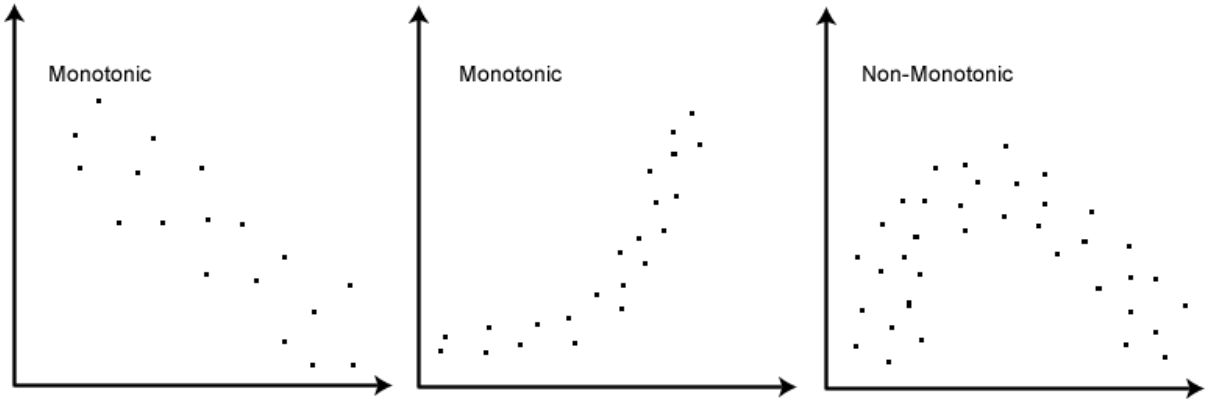

Figure 1.2: Monotonic versus non-monotonic relationships
In a monotonic relationship one variable decreases as the other variable increases or alternatively one variable increases as the other variable increases (although not necessarily in a linear fashion). In non-monotonic relationships this is not the case.

Where Pearson correlation is sensitive to outliers Spearman correlation is not. Pearson correlation is parametric, assuming the data is normally distributed, whereas Spearman

correlation is not. Since Pearson correlation is sensitive to outliers, Spearman correlation is preferable for co-expression analysis with smaller sample sizes and/or noisier data [95]. In stationary data (as opposed to time series data) mutual information performs similar to Pearson or Spearman correlation, which can be attributed to the rare occurrence of non-linear relationships in these networks [96]. Another correlation method, biweight midcorrelation is also less sensitive to outliers and was shown to outperform Pearson- and Spearman correlation in stationary data [96], yet is not as commonly used. For a discussion of other types of similarity measures we refer to [95]. To construct our co-expression networks we used Pearson correlation. We opted to use this metric to allow users to compare our RNA-seq based co-expression networks with previously published microarray based co-expression networks, which often use Pearson correlation, without having to worry about a potential bias introduced by the use of a different correlation metric.

In the second step, co-expression associations are used to construct a network. Such a network consists of nodes representing genes and edges representing the presence and the strength of the co-expression relationship (Figure 1.1) [97]. A co-expression network can be either weighted or unweighted. In a weighted network all nodes are connected to each other. These connections have continuous weight values between 0 and 1 that indicate the strength of co-regulation between the genes. In an un-weighted network the interaction between gene pairs is binary, i.e. either 0 or 1, indicating genes are either connected or unconnected. An un-weighted network can be created from a weighted network by, for example, considering all genes with a correlation above a certain threshold to be connected and all others unconnected. Weighted networks have produced more robust results than un-weighted networks [98].

In the third step, modules (in this thesis defined as: groups of strongly co-expressed genes) are identified using one of several available clustering techniques. Clustering, in co-expression analyses, is a method that can be used to identify groups of genes that have a similar expression pattern across multiple samples, to produce groups of co-expressed genes rather than only pairs. Many clustering methods are available, including k-means clustering and hierarchal clustering, and are discussed in detail in [99]. Modules can subsequently be interpreted by functional enrichment analyses**,** a method that can be used to identify and rank overrepresented functional categories in a list of genes [44, 100, 101]. In this thesis, we have used the 5 percentile closest genes in the network, for a gene of interest, as a module, which were subsequently used for the functional enrichment analysis.

With co-expression analysis, it is important to consider the heterogeneity of the samples. Tissue-specific or condition-specific co-expression modules may not be detectable in a co-expression network constructed from multiple tissues or conditions, because the correlation signal of the tissue/condition-specific modules is diluted by the lack of correlation in other tissues/conditions. However, limiting co-expression analysis to a specific tissue or condition also reduces the sample size, thereby decreasing the statistical power to detect shared co-expression modules. An over representation of samples from a specific origin may bias the co-expression network for processes described by these samples. To ensure this was not the case for our samples, we have analyzed the cell and tissue types from which these originated in Section 3.3.4. Data describing multiple tissues and conditions should be used for identification of common co-expression modules, while differential co-expression comparing different conditions or tissues will be better suited to identify modules unique to a specific condition or tissue. Tissue specific networks and differential co-expression are further discussed in the discussion of this thesis. In this thesis, we aim to assign potential functions to ncRNAs and

splice variants in a non-tissue-specific manner. Motivation for this choice is that we aimed to identify functions that are not specific to a specific tissue or cell type, so they would be of use to a broad range of researchers, which work with varying tissue and cell types. Additionally, at the start of this project the amount of RNA-seq data was limited and isolating a specific tissue would have further decreased the sample size.

1.5. Guilt-by-association based on gene co-expression

The next step in a co-expression analysis is to derive biological meaning from the constructed co-expression network. A widely used approach to attach biological meaning to co-expression modules is to determine functional enrichment among the genes within a module. Assuming that co-expressed genes are functionally related, enriched functions can be assigned to poorly annotated genes within the same co-expression module, an approach commonly referred to as 'Guilt-By-Association' (GBA) [7, 16]. Poorly annotated genes, in our definition, are those genes that have been annotated as genes, but which have not been annotated to any ontologies or pathways yet. GBA approaches are also widely used to identify new potential disease genes. If a substantial proportion of the genes within a module is associated with a particular disease, it suggests that other genes within this module play a more prominent role in this disease than those genes that are not within this module [11, 13, 16-20]. In this thesis, we have used this GBA concept to assign putative functions to genes and to predict new disease genes.

In 1998 the first microarray based co-expression work was published using *S. cerevisiae* data [102], later followed by a study conducting a co-expression analysis on human fibroblast data [103]. In these studies the correlation of the expression between all included genes was calculated based on data obtained from different conditions and time points. Even though utilizing these early low quality versions of micro-arrays, GBA proved strikingly effective at

grouping functionally related genes based on co-expression. It was later shown that genes encoding physically interacting proteins are likely to be co-expressed [103-106]. These observations underpin the assumption that co-expressed genes are indeed more likely to be functionally related. Using a GBA approach on co-expression data allowed poorly annotated genes to be related to well annotated genes and predict the function the poorly annotated gene(s) most likely play their primary role in [16, 107]. GBA has been widely used in co-expression analyses for gene function association/annotation to processes [108-113] and for the prediction of numerous disease genes involved in diseases such as cancer [8-12, 114], schizophrenia [13], diabetes [14, 15] and others [16-19]. Due to the successes obtained from early microarray versions and correspondingly smaller sample sizes, we felt that it should also be possible to use current RNA-seq data, even though limitedly available at the time, to assign putative functions to ncRNAs in a similar fashion.

1.5.1. Gene associations

In this thesis, GBA is widely applied to predict the biological process in which poorly annotated genes likely play their primary role. We use three different methods to achieve this.

One, we define new gene-function "associations" as: A gene being co-expressed (arbitrarily defined as the top 5 percentile co-expressed genes) with a list of genes enriched for that function. The functions can be defined in, for example, Gene Ontology (GO) [115] or Reactome terms [116]. Disease annotations are often obtained from specific studies and will have been annotated to that disease based on different types of evidence, which can be found in the corresponding referenced papers. Whether this association is significant is determined by the False Discovery Rate (FDR) value as calculated and obtained from DAVID [72]. Thus we define the categories for which the FDR is smaller than 0.01 as "associated" and those that have

larger FDR values as “not associated”. When we refer to “previous associations obtained from the literature”, we refer to genes that have been annotated to a specific category such as an ontology or disease. Again, the evidence for these associations can vary between the associations and can be found in the corresponding references.

Two, we obtain the ranking of all genes with a gene of interest and split these into a group of genes that are annotated to a category and those that are not. Consecutively, we calculate the difference between the rankings of the genes in these two groups, using a Mann-Whitney U test. This is a non-parametric test, therefore not requiring the data to be normally distributed. Additionally, the Area Under the Receiver Operator Curve (AUROC) or Area Under the Curve (AUC) in short can be derived from the results of this statistical test. The AUC is an indicator of how well false positives can be separated from true positives. An AUC of 1 or 0 indicates perfect separation where an AUC of 0.5 indicates the opposite. The significance value is calculated and if below 0.01 we define this gene as associated to that and otherwise not. This is used in Section 3.3.3 to estimate how often associations are found for the functions each gene has previously been annotated to.

Three, to associate a gene to a disease, we calculate how often a gene is partners (in the top 5 percentile co-expressed genes of the disease genes) with any of the genes in the disease geneset. Then we calculate it this significantly higher than expected by random chance described in Section 2.6.4. If this is the case we define this gene as associated to that disease and otherwise not.

1.6. Transcriptional binding site analysis

Transcription factor activity is often controlled by factors other than expression, such as a wide range of post-translational modifications like phosphorylation [117], acetylation [118] and

methylation [119], as well as ligand binding [120], and interaction with other transcription factors [121, 122]. As a result, transcription factors are often not co-expressed with their targets. This makes it difficult to identify the targets of transcription factors by simply investigating co-expressed genes with a particular transcription factor. To predict which regulatory elements are controlling the expression of functional groups in different phenotypes, we analyzed the enrichment for specific transcription factor binding sites (TFBS) among co-expressed genes with the aging seed list, as depicted in figure 1.3. Although highly co-expressed genes share TFBSs in yeast [123], this approach appears to be less successful in more complex organisms. In human, mouse, and fruit flies, it was shown that genes sharing TFBSs are not more likely to be co-expressed [124] unless the study is tissue specific [125], which is also supported by other studies [126-128]. On the other hand, in *Arabidopsis thaliana,* it was shown that co-expressed genes that share the same motifs can be assigned to modules [129] for several well-known motifs such as G-box, MYB, W-box, and site II element and associate them with specific immune and metabolic pathways [130]. Using a similar approach transcription factors *E2F1*, *GATA2,* and *NFKB1* were associated with cell cycle, fat, and muscle/glycolysis, respectively, using bovine muscle data [131]. The GATA TFBS was associated with aging in worms with the *Elt-3* GATA transcription factor being co-expressed in some tissues, but not in others [132]. Lastly, it was shown that promoter regions, which usually contain multiple TFBS, better correlate to co-expressed modules than single TFBSs [126, 133]. As there are cases where this type of analysis has yielded biologically interesting results, we opted to conduct a similar analysis on our list of genes co-expressed with aging genes (Section 2.3.4).

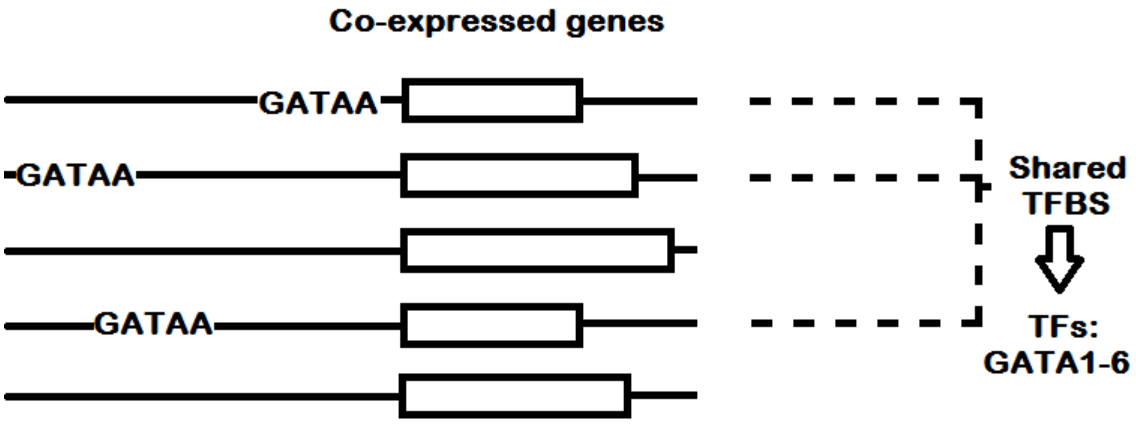

Figure 1.3: Identification of transcription factors potentially regulating co-expression modules After modules have been identified, the genes within a module can be investigated for enrichment of TFBSs. If there is an enrichment for a particular TFBS, the TF that has been annotated to act on this TFBS can be identified. This TF may be (partially) regulating the expression of this module.

1.6.1. MicroRNAs and GBA

As mentioned in Section 1.1.1, our co-expression analysis and the database we created are not tailored to the analysis of miRNAs. These tend to function by deactivating the gene they target. In Chapter 4 of this thesis, we set out to test if we can identify those targets by identifying genes that show a negatively correlated expression pattern with the miRNA. The expectation is that, for those miRNAs that degrade their targets in conjunction with a RISC complex [36], such a negative correlation is detectable. The proportion of miRNAs that function by degrading their targets rather than just deactivating their target through binding of the target's mRNA is not known, but we expected that a significant proportion of the miRNAs indeed do degrade their targets. To conduct this analysis, we used an in-house generated rat aging brain dataset. Additional to our miRNA analysis, we aimed to identify key regulators in the rat brain aging process. To achieve this, we conducted a co-expression analysis in which we identified a number of hub genes that could potentially be key players in the aging process of the rat brain (Section 4.4.4).

1.7. Hub genes

GBA, as described earlier, can be used to associate genes to diseases, but this often leads to large lists of new gene-disease associations. It often remains unclear which of these genes is most likely the causal factor. A widely-employed approach to identify genes that are key players (defined as the gene playing a much more prominent role in that particular biological process or disease than most others), is to identify highly connected genes in a co-expression network, commonly referred to as hub genes. Network hubs have been shown to generally be more relevant to the functionality of networks [134]. This is also the case in biological networks [97], although mathematical derivations show that this is only the case for intra-modular hubs, as opposed to inter-modular hubs [135, 136] (Figure 1.4).

1.7.1. Centrality and connectivity

To identify hub genes "betweenness centrality" is often used. Betweenness centrality, which can be interpreted as the relevance of a node in the network, is determined by counting the number of shortest paths between any other pair of nodes going through this node [137]. To measure the robustness of a network, network connectivity is often measured. Connectivity indicates how many genes have to be removed from the network to disconnect the remaining genes in the network. Identifying hub genes in co-expression networks has led to the identification of several genes essential in diseases such as cancer [25], type 2 diabetes [26], chronic fatigue [27] and others [28, 138].

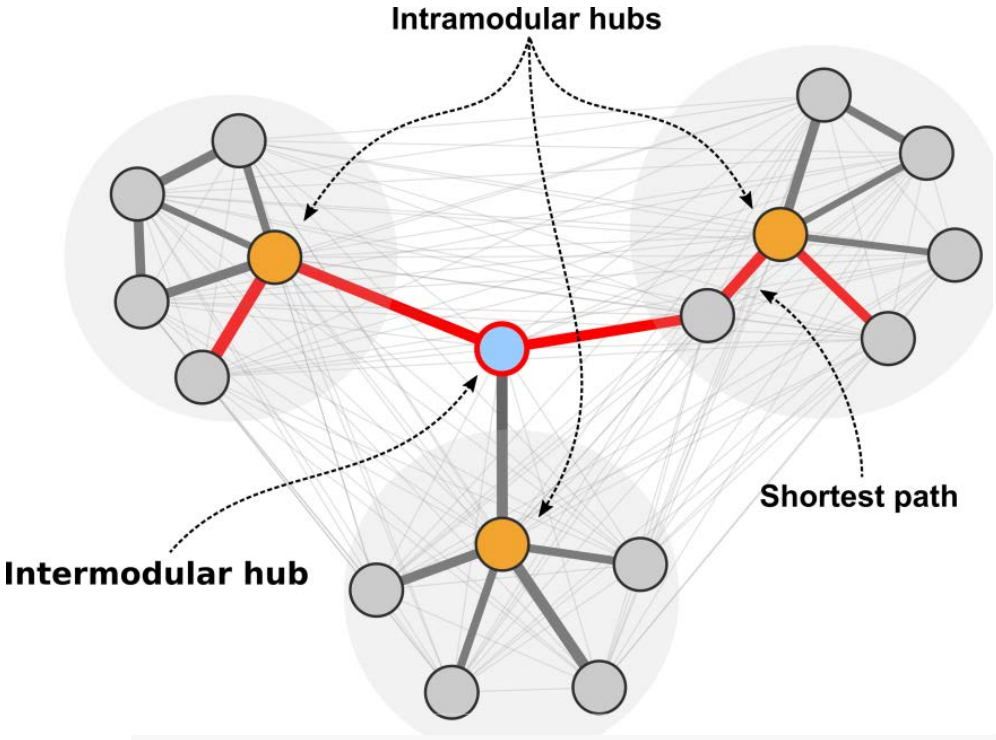

Figure 1.4: Hypothetical network explaining inter-, intra-modular hubs and network centrality The blue node is an inter-modular hub having a high network centrality as the highest number of shortest paths, from one node to another node in the entire network, go through this node (Red line is one example of a shortest path through the network between two nodes). Although this inter-modular hub is important for the network's structure it does not have a higher biological relevance as opposed to the intra-modular hubs. Intra-modular hubs (marked with orange) are central to individual modules, defined by the highest number of shortest paths from one node to another within the module going through this node. These intra-modular hubs have a higher biological relevance than the nodes that are not intra-modular hubs.

1.8. Weighted Gene Correlation Network Analysis (WGCNA)

In Section 4.4.2 of this thesis, we set out to identify the key regulators in a rat brain aging dataset, generated in our lab. These key regulators were identified by identifying hub genes in a tissue specific co-expression analysis. To identify these hub genes, we used WGCNA, a tool that can be employed for this purpose. This tool is easy to implement and works by first creating co-expression modules using hierarchical clustering based on a correlation network created from expression data [66]. Hierarchical clustering iteratively divides each co-expression cluster into sub clusters based on how strongly genes are co-expressed with each other. This creates a tree in which the branches represent the co-expression modules. By cutting the branches at a certain height, the modules are defined (Figure 1.1). These modules are often large and it is important to identify the genes in the module that best explain the behavior of the module, which are the intra-modular hub genes. This is why the centrality of each gene within a module and each gene is determined; those having a high centrality being intra-modular hub genes.

1.8.1. Eigengenes

Additionally, WGCNA determines the genes behaving similar to the eigengene of a module, as well as the intra-modular hub genes, which tend to coincide with each other in our experience. An eigengene is a hypothetical gene that best describes the average expression changes of the relative module between different samples. This eigengene is a vector that represents the partial expression of each gene (to a different extent per gene). This is calculated using Principal Component Analysis (PCA) on the expression of the genes within the module across all samples, where the first Principal Component (PC) is defined as the eigengene. PCA is a method to summarize variance of multiple variables into linear vectors. PCA first summarizes all variables into a linear variable describing as much variation as possible, becoming the first

PC. Then this process is iterated over the remaining variation (thus the variation that is not described by any of the lower PCs), until all variance is explained. The first PC is always the vector that best describes the variance of all variables in a module and is used as the vector that best represents the module, being the eigengene.

1.9. Differential co-expression analysis

WGCNA can also be used for differential co-expression analysis. Differential co-expression analysis can identify biologically-important differential co-expression modules that would not be detected using regular co-expression analyses. Identification of differentially connected hub genes can also help to identify potential regulatory genes and thus explain phenotypic differences that would not be uncovered in a standard differential expression or co-expression analysis [139-142]. Differential co-expression analysis has been used to identify genes underlying differences between healthy and disease samples [139-142] or between different tissues [143], cell types [144] or species [145].

Most differential co-expression analyses rely on differential clustering; they identify clusters that contain different genes or behave differently under changing conditions or phenotypes. With “behaving differently”, we are referring to altering activity of a module in different sample groups, where activity of a module can, for example, be represented by an eigengene, as described in Section 1.8.1. There is a number of tools which identify co-expression modules in the study samples and then correlate these to predefined sample subpopulations representing, for example, disease status or tissue type. Three such tools are: WGCNA [66], Differential Correlation in Expression for meta-module Recovery (DICER) [141] and DiffCoEx [67], which have been compared in [141] [146].

These different tools have slightly different interpretations of differential expression which is explained in this paragraph. WGCNA determines the activity and importance of each module in each subpopulation of samples (Figure 1.5a and c). It then prioritizes which genes in these modules are likely to underlie the phenotype associated with the module by identifying either genes behaving similarly to the eigengene of the module or those genes that are intra-modular hub genes (these tend to coincide), as described in [66]. By design, DICER is tailored to identify module pairs that correlate differently between sample groups, e.g. modules that form one large interconnected module in one group compared to several smaller modules in another (Figure 1.5d). DICER may be particularly useful for time series experiments in which co-expression changes are gradual, e.g., cell cycle series experiments, where modules are specific to a particular phase and co-expressed in transitions between phases. DiffCoEx focuses on modules that are differentially co-expressed with the same sets of genes. The most extreme case of this behavior is sets of genes which 'hop' from one set of correlated genes to another in a coordinated manner (Figure 1.5e). In this case, DiffCoEx would cluster 'hopping' genes in a similar manner. These are the most likely genes to explain different phenotypes that are associated with the two different networks. Each of the methods detects specific module changes by design, but they can also detect modular changes that they were not specifically designed for, and may outperform other tools in the identification of these changes [21].

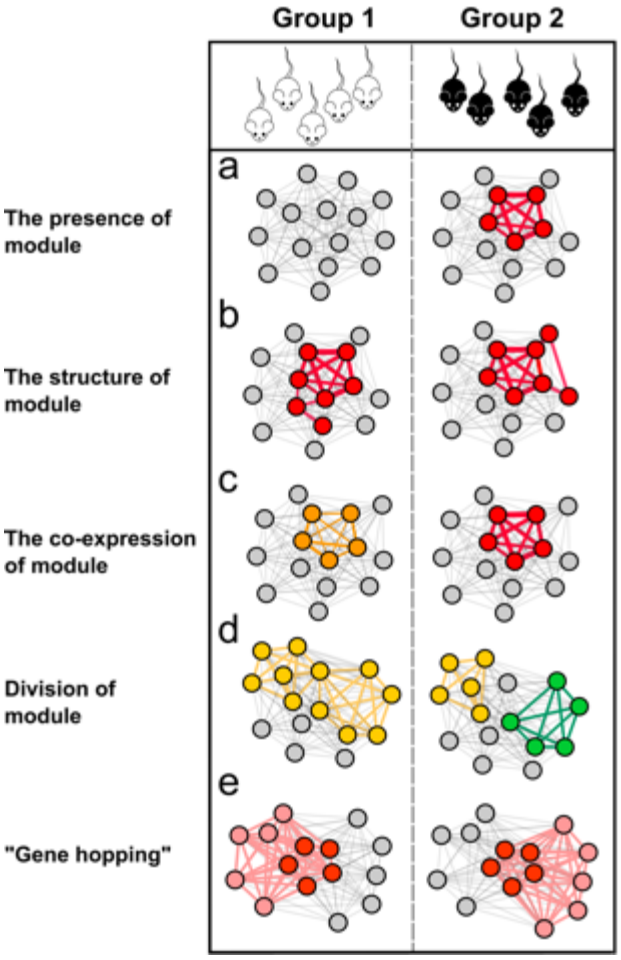

Figure 1.5: Differences in gene co-expression pattern changes that can occur between samples
Differential co-expression can manifest as the presence of the module in only one of the sample groups (a), differences in the structure of the module (b) or the differences in the correlation strength between the members of the modules (c). Additionally, the differential co-expression can be detected if some of the genes switch to another module ("gene hopping", d) or if one larger interconnected module splits into several smaller ones (e). This figure was created by Urmo Võsa.

A number of studies have used differential co-expression network analyses to identify networks unique to specific tissues [143] or disease states [147]. The rapid increase in publicly available RNA-seq data and projects such as Genotype-Tissue Expression (GTEx) and Encyclopedia of DNA Elements (ENCODE), which generate RNA-seq profiles on a large scale, has enabled co-expression analysis within and across different tissues [143, 148]. The GTEx project collects and provides expression data from multiple human tissues for the study of gene expression, regulation and their relationship to genetic variation [149]. In a study comparing RNA-seq data from 35 tissues from the GTEx dataset, a tissue hierarchy was constructed based on the average gene expression in each tissue. Related tissues, such as those from different brain regions, clustered together. This hierarchy was used to construct a single combined co-expression network derived from the tissue-specific co-expression networks; a meta-network. It was then shown that in tissue-specific networks TFs with functions specific to that tissue tend to be highly expressed together with tissue-specific genes. These genes tend to form a stronger connection with each other than with other genes, but remain at the periphery of the network (thus having low centrality), whilst the tissue-specific TFs become more central to that module [143]. Thus, tissue-specific TFs could be uncovered by identifying modules with increased co-expression strength in tissue-specific networks (Figure 1.5a and c) and by pinpointing the central hubs of these modules. In contrast, genes that are not TFs but are tissue-specific should be detectable by identifying those genes that are at the periphery in these modules (Figure 1.5b). Moreover, some TFs have different roles in different tissues. These TFs would be expected to be hub genes that are central to one module under one condition and central to another module in another condition.

In our analysis, we used WGCNA to determine which modules are differentially expressed at different ages in rats (Sections 4.3.2, 4.3.3 and 4.4.3). The activity of a module in a subgroup is

indicated by the eigenvalues of the eigengene (of a particular module) for the different samples in that group as explained in Section 1.8.1. The correlation between the module eigenvalues and the phenotype of interest, represented by that subgroup or multiple subgroups, then suggests the importance of this module for that particular phenotype. For example, in a dataset where different samples are derived from individuals with a different age, the age can be correlated with the eigengene eigenvalues of each module to find the module that has the strongest correlation. It then prioritizes genes in these modules identifying the genes that behave most similar to the eigengene of that module (by simply correlating the eigengenes' eigenvalues with the genes expression values across the samples). WGCNA has widely been shown to perform well under many different circumstances and for different purposes [66]. A comparison between these tools and others, including WGCNA, show that WGCNA and ARACNE perform best at defining the network structure of *E. coli* [21], for which a well-defined regulatory network is available and was used as a golden standard [150].

We used WGCNA on a rat brain dataset created in our own group, to identify modules that may or may not be important in different feeding regimes and aging (Section 4.4.3). After our analysis on our in-house generated rat brain aging dataset (Sections 4.4.2 and 4.4.3), we tested, in a rat aging thymus dataset (Section 4.4.4), if WGCNA would identify a gene, *Foxn1*, proven to be able to regenerate the thymus of old rats seemingly reverting aging [151], as a hub gene in any of the modules of which the activity correlates with aging. The purpose of this exercise was to test if WGCNA would identify a gene with a seemingly high aging intervention potential as a key regulator, and to which extent this factor would be prioritized. To achieve this, we use WGCNA to conduct a co-expression analysis on this dataset containing gene expression data from rat thymi at different ages.

1.10. Contributions

In general all work described in this thesis was conducted by Sipko van Dam, unless stated otherwise. In particular the co-expression analysis and most of the construction of the website, including the corresponding databases, were conducted by Sipko van Dam. The exemptions on the construction of the website are the following: 1. The .css file describing the layout and design of the web page. 2. The javascript that loads the DAVID table dynamically. 3. The symbol visible on the front page. These were all created by Thomas Craig. 4. The experimental work conducted on the *C1ORF112* and *C12ORF48* gene knockouts described in Chapter 2, which was carried out by Rui Cordeiro. Additionally, several people have proofread the thesis to aid with the structure and the language of this thesis.

1.11. Aims

A key objective in the post-genomic era is to systematically identify all gene products as well as their functions and interactions within a living cell. Genes involved in common biological processes and diseases are often co-expressed. We use these co-expression associations to predict the function of poorly annotated genes and associate them to diseases. Although this has previously been achieved for coding genes, no tools were available that allowed users to retrieve co-expression associations for non-coding genes or on a transcript rather than gene level. In this project, we created a tool that allows users to obtain new gene/transcript-function or disease associations for those that are poorly annotated. This includes ncRNAs, for which no tool was available at the time of this project. This is of great importance to enable prediction for these ncRNAs as well, since these poorly annotated genes and RNAs may play key roles in diseases, for which, in most cases, the molecular mechanisms often remain unclear. We attribute this partly to the lack of information for some elements, such as these ncRNAs. The tool we have created allows users not only to identify those coding and non-coding genes that are co-expressed with groups of disease genes, but also predict the biological process/pathway they likely play a role in. Furthermore, diseases can be caused by deviations in isoform expression patterns of a gene. At the start of this project, there were no tools available that allowed users to quickly identify the biological processes different transcripts, originating from the same gene, associate with. GeneFriends enables users to obtain co-expression based transcript-function predictions, thereby enabling them to potentially better understand which deviations in isoform expression patterns lead to particular phenotypes.

Here we list the goals of this project:

1. Construct a co-expression network from a large number of microarray samples and create a resource that is useful to the research community, in particular a website that can be queried with a single gene or multiple genes simultaneously. The purpose of this tool is to allow users to query sets of genes representing a certain disease or function to identify new gene associations. Additionally, we included the option to query single genes, to allow users to predict the biological process in which they most likely play their primary role based on enrichment among co-expression partners (Chapter 2). Similar works are readily available (Table 5.1) and this part of this project is most comparable to COXPRESdb [6]. A limitation of COXPRESdb is that the user interface, in our opinion, is not user friendly. For example, it is not possible to download the full co-expression networks and the results for queries are limited to the first 300 co-expressed genes. We created a similar database and conducted a similar analysis to acquire experience and to establish a user friendly website that could be used also for our second aim. Unique to our website is that we allow users to download the entire co-expression network as well as full lists of co-expressed genes with the query gene(s). The latter allows users to also investigate the least co-expressed genes, which can potentially also be biologically relevant.

2. Construct a co-expression network from RNA-seq data to include ncRNAs and add this to our existing website. As part of this aim we tested if the GBA approach, know to perform well at associating coding RNAs to biological processes and diseases, also performs well on ncRNAs. To our knowledge this had not previously been attempted on a genome wide scale and we were the first to publish a database that allowed this. (Chapter 3).

3. Construct a co-expression network from RNA-seq data for transcripts rather than genes and add this to our existing website. As part of this aim we identified transcripts originating from

the same gene that had different co-expression partners. The purpose was to elucidate if it is possible to find genes for which different transcripts associate with different functions using our co-expression database. Although transcript specific expression is gaining more attention, this is progressing slowly and our publication on our database, which allowed identification of transcript-function associations using co-expression on a genome wide scale, was the first of its kind [50]. (Chapter 3).

4. Conduct a co-expression analysis on our in-house generated rat aging RNA-seq dataset, to identify if the expression of miRNAs and their targets is negatively correlated. Although research has shown that this is the case, we aimed to test if this relationship would also be detectable in co-expression data, which had not been previously reported to our knowledge. (Section 4.4.1)

5. Conduct a co-expression analysis on our in-house generated rat aging RNA-seq dataset to identify modules that are altering in activity during aging (Section 4.4.2), as well as to identify modules that counter this effect through dietary interventions (Section 4.4.3).

6. Conduct an analysis to test if hub gene selection would identify *Foxn1* as one of the most important genes in the aging process in an aging dataset. This gene has proven to play a major role in thymus regeneration [151] and if it would not be identified it would indicate that selection of hub genes can miss important targets that would potentially be very targets for intervention studies. Motivation for this in silico experiment was that the importance of hub genes has been widely supported, but debated as well. We were interested to know how well this gene of great relevance, would be prioritized by this method. (Section 4.4.4)

These aims are more elaborately described and discussed in the following three chapters. In Chapter 5, the results of the research conducted to achieve these aims are further discussed and summarized in Chapter 6.

Chapter 2: GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases.

As described in the aims in the previous chapter, the aim of the work in this chapter is to associate new genes to diseases based on a co-expression network. The network was constructed using microarray data from the Gene Expression Omnibus (GEO) database [152]. Initially, the assumption was that only high quality data would be submitted to this database and therefore each included dataset would increase the reliability of the network. We have learned that poor quality data is also present in such databases, which may negatively impact the quality of the network. Since we used a number of quality control cutoffs, the samples with the poorest quality will have been excluded. We observe that genes co-expressed with genes of known function are functionally enriched for their annotated functions, supporting the notion this network can be used for this purpose. Similarly, when querying the co-expression network with sets of disease genes, there is a statistically significant enrichment for other genes annotated to the same disease but not present in the initial geneset. This suggests that the network can also be used to prioritize genes that are more likely to play an important role in a particular disease. It is likely possible to improve the network's ability to associate new genes to biological functions and disease by pruning the data for high quality data only and correcting for biases that may exist within these datasets. Although similar works have been previously published this database comes with an interface that is easy to use as supported by published works that have utilized our database over others [153-157]. Additionally, it laid the groundwork for the work conducted in Chapter 3. Further to the construction of our database, we have predicted that C/ebp transcription factors play a role in aging, supporting suggestive evidence that is readily available [158, 159]. Additionally, we have

made novel associations between previously poorly annotated genes and the cancer/cell cycle ontologies, which were experimentally validated. This work was published in *BMC Genomics* [20]. Thomas Craig designed the website and supplied the corresponding css file, Rui Cordeiro conducted the experiment described in Section 2.3.6, Shona Wood guided Rui with the design and execution of this experiment and significantly edited the manuscript to improve its readability. João Pedro de Magalhães provided guidance throughout the project and helped drafting and editing the manuscript.

2.1. Abstract

Although many diseases have been well characterized at the molecular level, the underlying mechanisms are often unknown. Nearly half of all human genes remain poorly studied, yet these genes may contribute to a number of disease processes. Genes involved in common biological processes and diseases are often co-expressed. Using known disease-associated genes in a co-expression analysis may help identify and prioritize novel candidate genes for further study. We have created an online tool, called GeneFriends, which identifies co-expressed genes in over 1,000 mouse microarray datasets. GeneFriends can be used to assign putative functions to poorly studied genes. Using a seed list of disease-associated genes and a guilt-by-association method, GeneFriends allows users to quickly identify novel genes and transcription factors associated with a disease or process. We tested GeneFriends using seed lists for aging, cancer, and mitochondrial complex I disease. We identified several candidate genes that have previously been predicted as relevant targets. Some of the genes identified are already being tested in clinical trials, indicating the effectiveness of this approach. Co-expressed transcription factors were investigated, identifying C/ebp genes as candidate regulators of aging. Furthermore, several novel candidate genes that may be suitable for

experimental or clinical follow-up, were identified. Two of the novel candidates of unknown function that were co-expressed with cancer-associated genes were selected for experimental validation. Knock-down of their human homologs (*C1ORF112* and *C12ORF48*) in HeLa cells slowed proliferation, indicating that these genes of unknown function, identified by GeneFriends, may be involved in cancer. GeneFriends is a resource for biologists to identify and prioritize novel candidate genes involved in biological processes and complex diseases. It is an intuitive online resource that will help drive experimentation. GeneFriends is available online at: http://GeneFriends.org/.

2.2. Background

Over the last decade, microarray technology has allowed researchers to measure gene expression levels across large numbers of genes simultaneously identifying genes and biological processes that are activated or impaired under different conditions. Potential biomarkers [160-163] and genes involved in a number of diseases, such as cancer, have been identified by microarray analyses [164, 165]. By combining gene expression data in a meta-analysis, greater power and more information can be gained from existing data. Meta-analyses have been successfully used to identify new relationships between genes and new candidate disease-associated genes [4, 166]. Microarrays provide large-scale, genome-wide data, from which coordinated changes in gene expression can be inferred. Information about these coordinated changes is valuable as they can be harnessed to identify the factors involved in disease and the functions of many poorly studied genes. One of the issues that arises with these large-scale datasets, however, is that it becomes harder to interpret the data and identify key players. To facilitate the attainment of biological understanding of results acquired

from large-scale dataset analysis in which poorly annotated genes are identified as key players, we have created a tool to assign putative functions to such genes: GeneFriends.

GeneFriends is based on a co-expression analysis, in which the general behavior of genes, relative to each other, is studied. This makes it possible to uncover genetic modules that are functionally related [6]under the assumption that those genes active in the same biological processes are co-expressed. The main theory behind this approach is that functionally related genes are more likely to be co-expressed [102, 167, 168]. This “guilt-by-association” concept has already been used to relate hundreds of unidentified genes to inflammation, steroid-synthesis, insulin-synthesis, neurotransmitter processing, matrix remodeling and other processes [4, 7]. Some of the predicted results have been experimentally validated demonstrating the effectiveness of this approach [4]. Candidate genes for cancer, Parkinson’s and Schizophrenia have also been identified using this approach [7, 169]. Furthermore, it is possible to identify transcriptional modules that may play causative roles in the disease or process under study [4, 166].

The aim of this work was to construct an online tool that can be used to derive novel candidate genes for further studies in aging and complex diseases, in a quick and intuitive manner. Aging is not considered a disease, yet older individuals are more susceptible to several diseases such as Alzheimer’s, Parkinson’s and cancer. This is one of the reasons why research in this field is rapidly expanding and several hundreds of genes have been linked to aging [170]. A major bottleneck in aging/complex disease research is that it is difficult to determine the causality of transcriptional alterations. It is also unclear if the altered expression profile observed with aging/complex disease consists of one particular biological module or whether it consists of

genes that act separately from each other. To this end, GeneFriends outputs transcription factors co-expressed with the genes supplied by the user.

Underlying GeneFriends is a genome-wide co-expression map created using over 1,000 mouse microarray datasets. We validated our co-expression map by showing that functionally related genes are more likely to be co-expressed. We then used GeneFriends to study transcriptional changes with aging, cancer and mitochondrial disease. Multiple candidate genes associated with cancer and mitochondrial diseases, including poorly annotated, were identified. Two of the novel candidates of unknown function that were co-expressed with cancer-associated genes were experimentally validated by knock-down in HeLa cells this slowed growth, supporting our predictions. This demonstrates that GeneFriends is a useful resource for studying complex diseases/processes and can infer function of poorly studied genes. GeneFriends is freely available online to allow researchers to quickly identify candidate genes co-expressed with their genes of interest (http://GeneFriends.org/).

2.3. Results

2.3.1. GeneFriends: An online tool for the research community

The aim of the project was to create a user-friendly tool, which can take a list of genes related to a given disease or process and quickly identify new candidate genes. Using co-expression profiling, the genes are given a rank reflecting which genes tend to co-activate with the list of input genes the most and which the least. This ranked list of genes then helps prioritize candidates for experimental follow up. Underlying GeneFriends is a *Mus musculus* co-expression map created from 1,678 microarray datasets, comprising over 20,000 individual samples from previously published experiments. To create the co-expression map, we employed a vote counting method. The co-expression map contains ≈ 427.5 million gene pairs

(20,676 x 20,676) arranged in a matrix and are given a scored based on how often they are co-expressed across all microarrays (see Materials and Methods).

The input for GeneFriends is either a single gene or a list of Entrez or gene symbol identifiers. The output is a list of co-expressed genes, which can be downloaded or viewed online. GeneFriends has the following functionalities:

1. It searches for co-expressed genes based on a seed list or a single gene, and provides a ranked list of significantly co-expressed genes.
2. It identifies the GO terms and enrichment for the significantly co-expressed genes.
3. It returns a ranked list of significantly co-expressed transcription factors.

We feel this output will help researchers in various fields identify interesting genes for follow up studies in a quick and intuitive manner. To test if this novel tool can be used to derive biologically-relevant predictions, we tested gene sets related to aging, cancer or mitochondrial complex I disease. Furthermore, we tested two predicted novel cancer candidates experimentally, as detailed below.

2.3.2. Testing the co-expression map

The biological significance of the co-expression map was verified using nine well annotated genes known to be involved in three biological processes: three genes annotated to cell division cycle, three genes known to be important in the immune system and three genes annotated to fatty acid metabolism. For each gene the 5% most strongly co-expressed genes were selected and DAVID [171] was employed to detect enriched biological processes and functions. This resulted in the functional enrichment categories of the co-expressed genes for each of the nine genes. For each of the nine genes, the functional enrichment of their co-

expressed genes was consistent with its known annotation: for the three cell cycle genes: cell cycle (GO:0007049) FDR< 10^{-50}, the three immune genes: inflammatory response (GO:0006954) FDR<10^{-5}, the three fatty acid metabolism genes: fatty acid metabolic process (GO:0006631) FDR< 10^{-15}. Detailed results are included in the supplementary data (Supplement 1). Additionally, figure 2.1 shows the clustering of the co-expression map's network, demonstrating that co-expressed genes tend to be involved in the same biological processes.

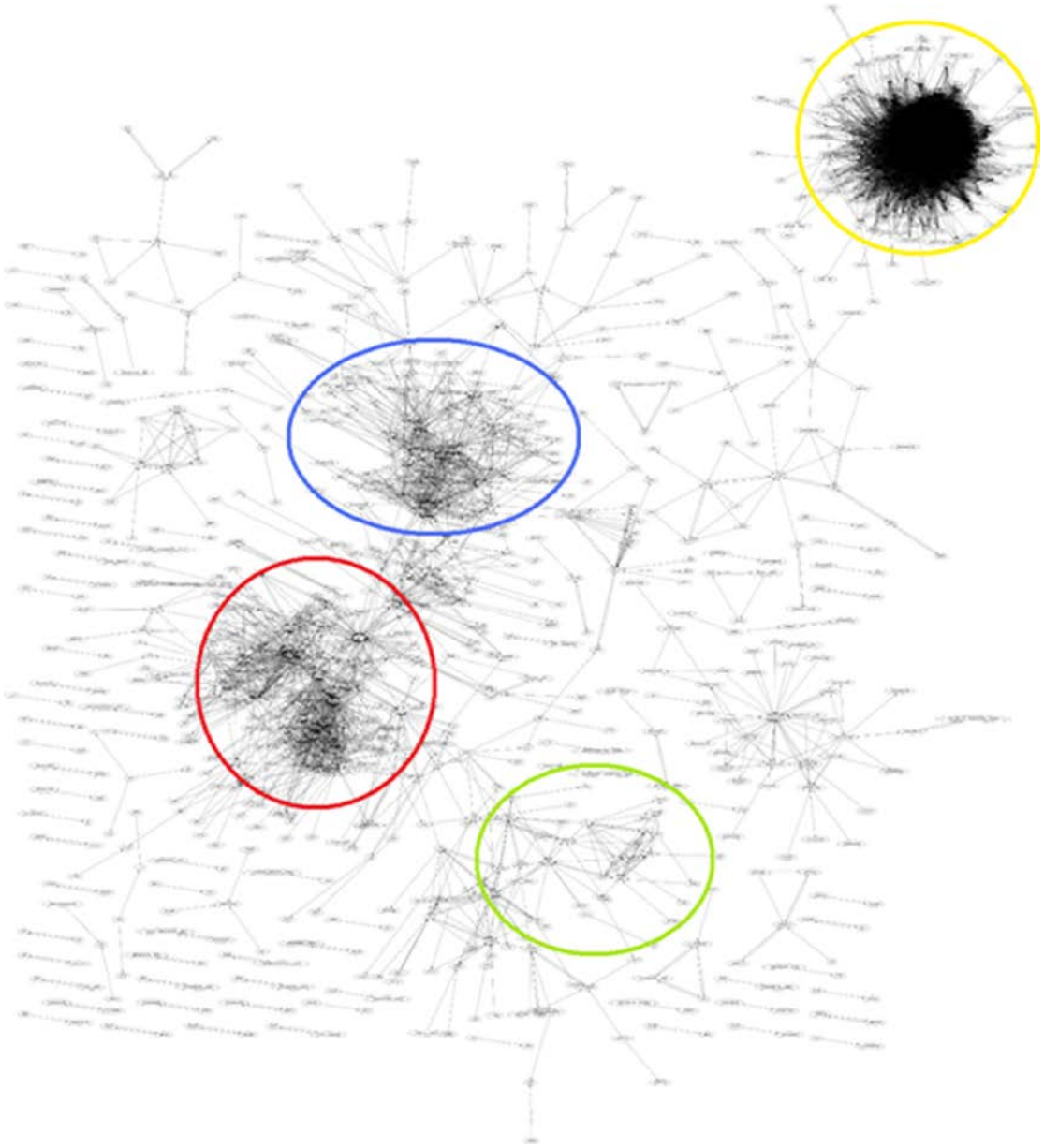

Figure 2.1: Gene clustering in the network of co-expressed genes
Illustration to visualize the structure of the co-expression network. Each node represents a gene (gene names are not readable to avoid clotting the figure), and the connections indicate a co-expression ratio of at least 0.8 between gene pairs. Genes without any edges are not plotted. Circles were drawn arbitrarily around densely connected gene clusters. Functional enrichment among the genes within these circles was identified to investigate if co-expressed genes are functionally related. Yellow: Cell cycle (Enrichment score: 66, Benjamini: $5.2x10^{-83}$); Blue: Extra-cellular matrix, collagen (Enrichment score: 16, Benjamini: $3.0x10^{-22}$); Red: Immune system (Enrichment score: 6, Benjamini: $1.3x10^{-16}$); Green: Fatty acid metabolism (Enrichment score: 16, Benjamini: $1.6x10^{-23}$). Benjamini corrected p-values < 0.01 indicate that the co-expressed genes are enriched for the respective function and thus tend to be functionally related.

GeneFriends uses a vote-counting method to rank co-expression. We compared GeneFriends to COXPRESdb [6], which utilizes the more commonly used correlation value (Pearson). To do so, we selected 3 genes with known functions and retrieved output from both tools and used DAVID to determine enriched categories. We identified the functional enrichment among the top 300 strongest co-expressed genes retrieved from each of the tools. The results show the same categories, with slightly different enrichment scores although the overlap in the specific genes among the top 300 strongest co-expressed genes can vary (table 2.1). These results show that functional enrichment of the top 300 genes, for these 3 genens, between GeneFriends and COXPRESdb are similar.

When comparing the numbers of transcription factors present in the top 1000 co-expressed genes from GeneFriends and COXPRESdb, the results are similar. This demonstrates that our approach performs similar to using Pearson correlation to create a co-expression map, in the sense that it leads to the same functional enrichment of gene co-expression partners (Table 2.1). In Chapter 3, we switched to the use of Pearson correlation as it is more commonly used/accepted and does not have a bias toward datasets that contain more conditions, as described in Section 2.6.2.

Gene	Category	COXPRESdb	GeneFriends	Overlap (out of top 300 genes)
Brca1				197
	Cell Cycle	66.35	75.09	
	Chromosome	54.62	50.5	
H2-Aa				94
	Disulfide bond	30.09	11.27	
	Immune response	19.5	14.52	
Ppara				100
	Fatty Acid metabolism	8.75	19.1	
	Peroxisome	14.37	11.12	

Table 2.1: Comparison GeneFriends and COXPRESdb co-expression analysis results
To identify the differences in the results between two different co-expression databases, using two different approaches, we selected the top 300 strongest co-expressed genes for 3 annotated genes, retrieved from each database. Functional enrichment scores for these 300 genes were retrieved from DAVID. The last column indicates the overlap between the genes in the top 300 genes retrieved from both databases. Although the overlap of co-expressed genes is not as large as we expected the functional annotation of the enriched genes is similar, leading to the same predicted biological processes for the query genes.

2.3.3. Candidate gene prediction from process/disease gene lists

We used GeneFriends to identify novel candidate genes associated with specific processes or diseases. The results show the number of times each of the 20,676 genes in the co-expression map were "friends" with genes in the disease gene seed-list and corresponding p-values indicate the statistical significance of the co-expression (see Materials and Methods). The p-value is calculated based on the number of input genes a given gene is co-expressed with and the total number of genes it is co-expressed with (Materials and Methods). DAVID was used to interpret the broader biological significance of the results. Functional enrichment analysis was conducted on all genes with a co-expression p-value $<10^{-6}$ using the default settings in DAVID. The stringent cutoff of 10^{-6} is based on a Bonferroni correction for 20,677 genes on a 0.05 p-value significance cutoff.

2.3.4. Aging-related gene prediction and putative transcriptional mechanisms

GeneFriends was used to identify genes related to aging. A seed list of genes known to be consistently over-expressed with age in mammals was used [170]. In total, 1119 genes were co-expressed with the aging seed list at $p <10^{-6}$; table 2.2 shows the top 25 genes.

Gene	Previous association evidence	Reference
Thbs1	Plays a role in platelet aggregation, angiogenesis and tumorigenesis	[172, 173]
Ctsh	No previous associations	
2310043n10rik	No previous associations	
Sat1	Induction has been suggested as a therapeutic strategy for treating colorectal cancer	[174]
Tcn2	No previous associations	
Pgcp	No previous associations	
D12ertd647e	No previous associations	
Cd74	Initiates signaling leading to cell proliferation and survival	[175]
B2m	B2m deficient mice suffer from tissue iron overload	[176]
Tgm2	Overexpression increases apoptosis in neuroblastoma cells. Implicated in fibrosis, neurodegenerative and celiac disease	[177] [178]
Rarres2	No previous associations	
Anxa1	Plays an important role in anti-inflammatory signaling, apoptosis and proliferation	[179, 180]
Il10rb	No previous associations	
Ctsc	Mutations cause Papillon-Lefevre Disease	[181, 182]
Lipa	Mutations can cause Cholesteryl ester storage disease and Wolman disease	[183]
IL3ra1	No previous associations	
Lgals3bp	Associated with cancer and metastasis	[184]
Pros1	Associated with Thrombosis	[185, 186]
Fcgr2b	No previous associations	
Scd1	Plays an important role in body weight regulation and development of obesity	[187]
Ifi35	No previous associations	
Ctla2b	No previous associations	
Cebpd	Implicated in adipocyte differentiation, learning and memory, mammary epithelial cell growth control. Loss of *Cebpd* leads to chromosome instability	[188-190] [191]
Fcgrt	No previous associations	
H2-t23	No previous associations	

Table 2.2: Top 25 genes co-expressed with aging related genes
For each gene, the number of times it is in the top 5 percentile co-expressed genes of the genes in the aging set, was counted. Then the overall occurrence of this gene in the top 5 percentile of any gene was counted. Using this overall occurrence, the chance this gene would occur this many times in the top 5 percentile co-expressed genes of only the aging related genes, was calculated using the binomial. The results were ranked on the corresponding p-values leading to the ranking as shown in the table. For a more detailed description, we refer to the methods in Section 2.6.4. For a full list and the corresponding p-values, we refer to supplement 2.

Next, we tested if there was an enrichment for aging associated genes using a curated list of aging associated genes from a separate source [192] (Supplement 3). Genes in the initial seed list were removed from the list of co-expressed genes and a Fishers exact test proved there are significantly more aging related genes in the list of co-expressed genes (top 5 percentile strongest co-expressed genes) as compared to those that are not co-expressed ($p<0.01$).

Many of the co-expressed genes have been associated with age-related diseases, such as Alzheimer, Parkinson and cancer. Several other genes that have been shown to play a role in aging, such as lysosomal-associated membrane protein-2 *Lamp2* [193] ($p < 5.68^{-30}$), *Fas* [194] ($p<2.70^{-31}$) and growth hormone receptor *Ghr* [195] ($p<1.34^{-19}$), also showed significant co-expression with the aging seed list genes. *Anxa2*, *Anxa3* and *Anxa4* also show a low p-value ($p<10^{-25}$), as well as several S100 calcium binding proteins, which have been shown to interact with annexins [196].

The most significantly over-represented functional clusters were inflammatory response (enrichment score (ES) = 24.13, FDR = $1.97x10^{-18}$), vasculature development (ES = 10.18, FDR = $2.31x10^{-8}$) and lysosome (ES = 9.00, FDR = $2.25x10^{-8}$). Since most of the genes in the seed list were classified in the categories related to the immune system, it was unsurprising to find similar results for the co-expressed genes.

Eighty genes, from the initial 181 genes in the aging seed list, showed a co-expression p-value $<10^{-6}$, suggesting the presence of shared transcriptional modules. In order to investigate the underlying transcriptional mechanisms that may induce this expression profile, we used the co-expressed transcription factor (TFs) results obtained from GeneFriends. Table 2.3 shows the 10 most significantly co-expressed TFs with aging.

Transcription factors	p-value	Gene Name
C/ebpδ	$7.90X10^{-34}$	CCAAT/enhancer binding protein (C/EBP), delta
C/ebpα	$1.19X10^{-30}$	CCAAT/enhancer binding protein (C/EBP), alpha
C/ebpβ	$3.78X10^{-30}$	CCAAT/enhancer binding protein (C/EBP), beta
Creg1	$1.70X10^{-29}$	cellular repressor of E1A-stimulated genes 1
Nfe2l2	$1.17X10^{-28}$	nuclear factor, erythroid derived 2, like 2
Irf7	$8.04X10^{-26}$	interferon regulatory factor 7
Klf2	$1.86X10^{-23}$	Kruppel-like factor 2 (lung)
Irf1	$8.17X10^{-23}$	interferon regulatory factor 1
Ostf1	$1.96X10^{-22}$	osteoclast stimulating factor 1
Atf3	$2.09X10^{-22}$	activating transcription factor 3

Table 2.3: Ten most significantly co-expressed transcription factors with genes increased in expression with aging

We selected the transcription factors that are most strongly co-expressed with the aging seed list, as these may be important regulators in the aging process. We identified several cebp genes which have also been reported to extend lifespan in mice if their expression levels are altered [197, 198]. Since transcription factors are not always co-expressed with their targets, other transcription factors that play an important role in the aging process may be missing from this list.

The most TFs with the most significant p-values were *C/ebpα*, *C/ebpβ* and *C/ebpδ* (Table 2.3). Interestingly, these TFs show co-expression (i.e., in top 5% of co-expressed genes) with a significant proportion of the genes co-expressed with the aging seed list: 477 out of 1119 genes (p-value < 10^{-100}) for all 3 TFs and 730 out of 1119 (p-value < 10^{-100}) were co-expressed with at least 2 out of 3 C/epβ genes.

Since these TFs are co-expressed with the aging-related genes it was expected that these genes, at least in part, would be regulated by the co-expressed TFs. Therefore, they would share TFBSs for these TFs. To identify over-represented binding motifs in the genes co-expressed with the aging genes (p-value<10^{-6}), we employed FactorY [199]. For the aging gene set, this revealed *Nfkb* as the most significant result (Table 2.4). Some of the TFBSs identified are targeted by the co-expressed TFs with the aging seed list such as; *Nfkb1* (p_{TFBS} <1.48 x10^{-5}, $p_{Coexpress}$ = 4.44 x10^{-9}), the *C/ebp* (p_{TFBS}< 6.95 x 10^{-3}, $p_{Coexpress}$ =7.9 x 10^{-34}) genes and *Irf1* (p_{TFBS} < 5.8 x 10^{-4}, $p_{Coexpress}$ <8.17x10^{-23}) genes (Table 2.4). However, TFBSs for *Isre*, *Nfkb2* (p65) and *Sp1* were identified as over-represented but not co-expressed and many co-expressed TFs did not have over-represented binding sites.

	N Total	N Select	p-value (corrected)
MA0061 (*NF-kappaB*)	2132	172	1.48E-05
V$ISRE_01 (*ISRE*)	2411	185	7.67E-05
MA0107 (*p65*)	2174	167	1.92E-04
V$GC_01 (*GC*)	9594	596	1.65E-04
V$NFKAPPAB_01 (*NFKAPPAB*)	2085	161	1.34E-04
MA0051 (*Irf-2*)	1440	117	3.13E-04
V$SP1_01 (*SP1*)	7474	475	3.30E-04
V$NFKB_Q6 (*NFKB*)	2433	179	4.39E-04
MA0056 (*MZF_1-4*)	4521	304	4.90E-04
V$NFKAPPAB65_01 (*NFKAPPAB65*)	2204	164	4.66E-04
V$IRF1_01 (*IRF1*)	1851	141	5.72E-04
V$NFKB_C (*NFKB*)	1635	123	3.18E-03
V$MZF1_01 (*MZF1*)	5057	327	3.76E-03
MA0050 (*Irf-1*)	3187	216	5.54E-03
V$CREL_01 (*CREL*)	3152	213	6.85E-03
MA0102 (*cEBP*)	3106	210	6.95E-03
V$SREBP1_02 (*SREBP1*)	1707	123	1.30E-02
MA0073 (*RREB-1*)	5643	353	1.80E-02
V$PAX4_03 (*PAX4*)	3765	245	1.73E-02
V$SP1_Q6 (*SP1*)	9475	565	1.96E-02
MA0079 (*SP1*)	7190	438	2.38E-02
V$MZF1_02 (*MZF1*)	5883	363	3.30E-02
MA0101 (*c-REL*)	3146	204	4.96E-02
V$IRF2_01 (*IRF2*)	1412	100	4.94E-02
V$GKLF_01 (*GKLF*)	7519	452	4.82E-02
V$STAT_01 (*STAT*)	3411	219	4.93E-02

Table 2.4: TFBS enrichment analysis
These are the results retrieved from FactorY [199], using the list of genes co-expressed with our aging seed list as input for the TFBS enrichment analysis. “N Total” indicates the number of genes that this transcription factor targets genome wide. “N Select” is the number of targets that have a transcription factor binding site in the input set of genes. The p-value indicates the significance, which has been corrected for multiple testing (Bonferroni). *cEBP*, for which several genes are also found to be strongly co-expressed with the aging associated genes, is also among this list further supporting the notion that this gene may be a regulator of the aging associated genes.

2.3.5. Cancer-related gene prediction

A seed list of 45 cancer-related genes was used as input for GeneFriends. DAVID analysis identified Cell cycle (ES = 58.84, FDR =$2.9x10^{-77}$) and DNA replication/repair (ES = 34.99, FDR = $6.0x10^{-51}$) as the most significant over-represented categories for cancer-related co-expressed genes. This is expected given the fact that cancer arises from the uncontrolled division of cells. Table 2.5 shows the top 10 co-expressed genes.

Gene	Previous association evidence	Reference
Nfkbil2	Confers resistance to DNA damaging agents and is a component of the replication stress control pathway	[200]
Chtf18	Involved in checkpoint response and chromosome cohesion	[201]
Cdc25c	Over expression associated with poor prognosis of cancer	[202]
Cdc7	Effective in inhibition of cancer growth	[203]
E130303b06rik	No previous associations	
Cep152	Involved in centriole duplication	[204]
Bc055324	No previous associations	
Cenpp	Required for proper kinetochore function and mitotic progression	[205]
Anln	Increased in expression in lung carcinogenesis and suggested as target	[206]
Hirip3	No previous associations	

Table 2.5: Top 10 genes co-expressed with cancer-related genes
Most of these genes have been previously reported as genes that play a role in cancer development. Others have been associated to biological processes that can cause cancer if they are disrupted. For a full list refer to supplement 4.

From the original seed list, only 6 genes were co-expressed with the entire set of cancer genes (p-value $<10^{-6}$), which could be due to the heterogeneity of cancer etiology. However, there were several significantly co-expressed genes, not included in the seed list, that have previously been associated with cancer. For example, *Cdc25A*, *Cdc25B* and *Cdc25C*, members of the Cdc25 family, are significantly co-expressed ($p < 10^{-6}$) with the cancer genes. There were a high number of significantly co-expressed centromere proteins co-expressed with the cancer seed list. These proteins play a role in chromosome segregation, and incorrect segregation of chromosomes during the cell cycle can lead to cancer [207]. *Cep152* is involved in centriole duplication [204]. *Cenpp*, as well as *Cenpn, Cenpf, Cenph, Cenpj, Cenpl, Cenpc1, Cenpt, Cenpk, Cenpm, Cenpe, Cenpq, Cenpa and Cenpl* are all co-expressed significantly with the cancer seed list and are part of the CENP-A NAC complex (Table 2.6). This complex is required for proper kinetochore function and mitotic progression and its disruption can lead to incorrect chromosome alignment and segregation that preclude cell survival despite continued centromere-derived mitotic checkpoint signaling [208, 209]. *Plk1*, *Aurka*, *AurkB* and *Cdca8* are in the top 50 co-expressed genes, these play an important role in cancer formation [210, 211].

Gene symbol	Seed list co-expression partners	Total co-expression partners	Seed list gene number	p-value
Cenpp	23	2247	45	2.54E-11
Cenpn	22	2216	45	1.61E-10
Cenpf	18	1365	45	1.74E-10
Incenp	20	2062	45	2.54E-09
Cenph	25	3505	45	5.01E-09
Cenpj	20	2291	45	1.56E-08
Cenpi	24	3512	45	3.09E-08
Cenpc1	14	1088	45	4.45E-08
Cenpt	18	1983	45	6.33E-08
Cenpk	20	2515	45	7.55E-08
Cenpm	12	1019	45	1.29E-06
Cenpe	17	2169	45	1.38E-06
Cenpq	21	3495	45	3.52E-06
Cenpa	19	3090	45	9.78E-06
Cenpl	23	4398	45	9.90E-06

Table 2.6: List of CENP-A NAC complex related genes co-expressed with the list of cancer associated genes

P-values are calculated using a cumulative binomial test, determining the significance of observing this number of seed list co-expression partners based on the total number of partners in the entire network and the number of seed list genes. It is important to consider that the p-value is dependent on the number of genes in the network, meaning if more genes are added to the network it is more likely to observe more significantly co-expressed genes (100 out of 1000 is more significant than 10 out of 100 according to a binomial test).

Several poorly annotated genes (*Bc055324, E130303B06Rik, 4930547N16Rik, F730047E07Rik*, *1110034A24Rik*, and *4632434I11Rik*) were co-expressed with the cancer-related genes suggesting these genes might play a role in occurrence or pathophysiology of cancer. One of these poorly annotated genes, *Bc055324*, is a predicted protein coding gene, which has a high co-expression ratio of more than 0.7 with the cancer genes *Rad51* and Ccdc6 [212], indicating this gene is increased in expression in >70% of the cases that *Rad51* is increased in expression. Many other cancer-related genes, such as Brca1 and Brca2, also show a strong co-expression with the *Bc055324* gene (Supplement 4). The genes co-expressed with *Bc055324* show an enrichment genes annotated to the cell cycle ontology (ES = 52, FDR = $1.7*10^{-74}$). A Basic Local Alignment Search Tool (BLAST) analysis of the protein sequence shows no there is no significant homology to other *Mus musculus* proteins. Similar sequences, however, are found in a large number of different multi-cellular species such as *Gallus gallus, Bos taurus* and *Homo sapiens* and there also is a significantly similar gene present in *Arabidopsis thaliana*, suggesting it is conserved in plants as well. Since this gene is well conserved it is likely a functional gene, as opposed to being a pseudogene.

2.3.6. Validating the role of *C1ORF112* and *C12ORF48* in growth of cancer cells

To test our predictions, we employed small interfering RNA (siRNA) to knock down the human homologs of *Bc055324* (*C1ORF112*) and *4930547N16Rik* (*C12ORF48*) in the widely-used HeLa cancer cell line. These two genes were selected for validation because they are co-expressed with genes involved in the cell cycle (DAVID Enrichment score: 56, FDR<1.0E-10), thus knockdown of these genes should lead to a measurable phenotype. The expectation is that, if these genes play a role in cancer, they play a crucial role in proliferation. As such, knocking these genes down should lead to decreased growth rate of the cells. Furthermore, we selected these two genes because validated siRNAs were available (see Materials and

Methods) for these genes. The results show that the growth rate of the cancer cells is significantly lower when either *C1ORF112* or *C12ORF48* are knocked down (Figure 2.2). These results support our predictions and demonstrate that *C1ORF112* and *C12ORF48* are important for cell growth. We do note that this phenotype is likely to occur when knocking down any gene. In yeast it was shown that 15% of all homozygous diploid disruptions cause reduction in growth rate [213]. It is imaginable this is also the case for mouse cell line knockouts, which would mean this phenotype would be observed for a relatively large proportion of gene knockdowns, i.e. 15%. Nonetheless, the fact that we observe this reduced growth rate supports the notion that these genes could be important for cancer growth.

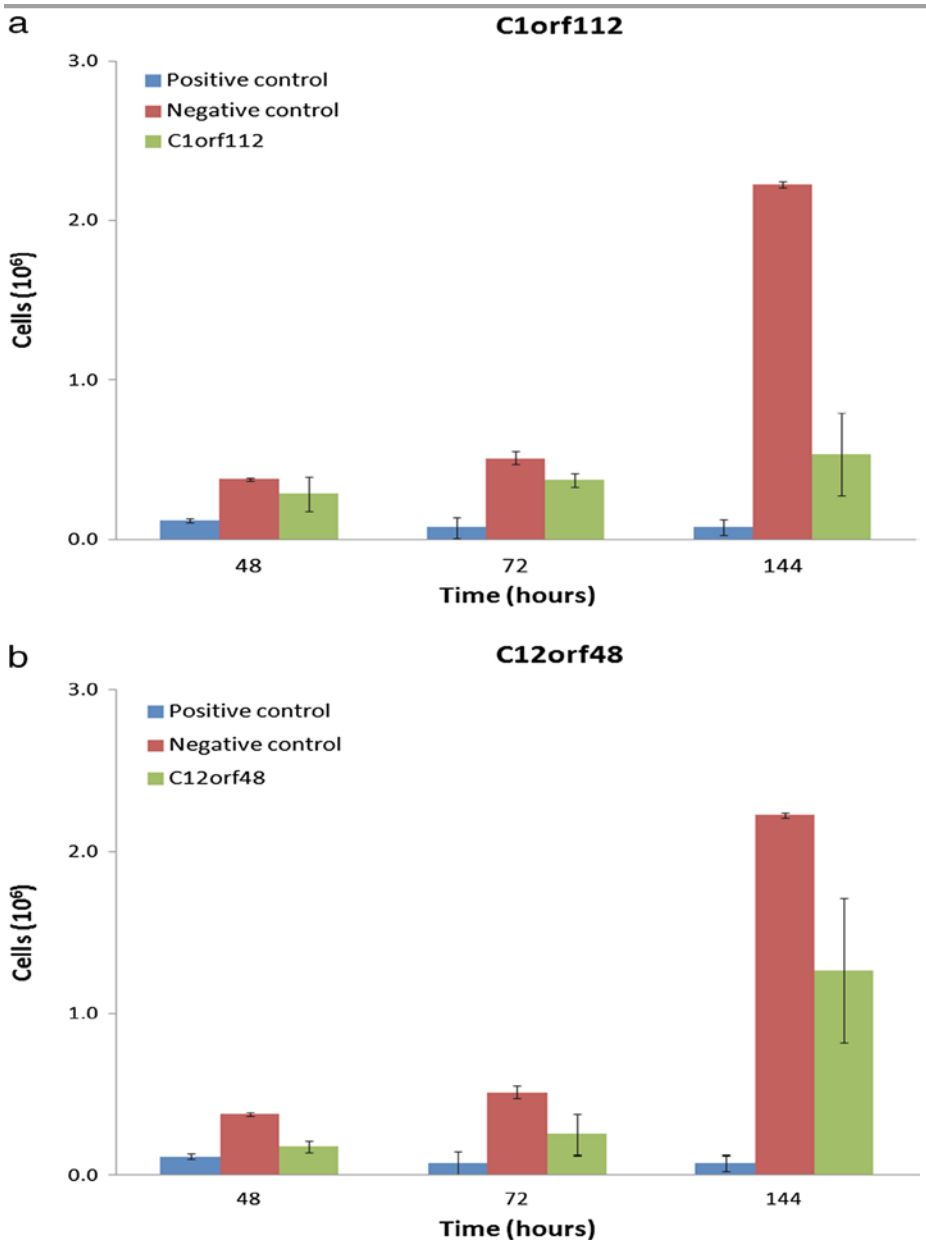

Figure 2.2: Knock-down of candidate cancer related genes slows growth of HeLa cells
a. Cell counting assay for the knock down of the human homolog gene of *Bc055324* (*C1ORF112*). b. Cell counting assay for the knock down of the human homolog gene of *4930547N16Rik* (*C12ORF48*). Error bars indicate the standard deviation. Negative control contains siRNA's targeting non-mammalian genes. Positive control contains siRNA's inducing apoptosis. The knock-down of either of these genes appear to reduce the proliferation rate of the cells.

2.3.7. Mitochondrial complex I disease-related gene prediction

All 10 genes in the seed list of mitochondrial complex I disease genes were significantly co-expressed with each other. This functional enrichment for the co-expressed genes with this seed list was the strongest amongst all disease gene seed lists tested, indicating these co-expressed genes are involved in the same process and are tightly regulated (Table 2.7). Table 2.8 shows the top 10 co-expressed genes with the seed list.

Annotation Cluster	Enrichment Score	Category	Term	Count	%	List Total	FDR
1	210.25	GOTERM_CC_FAT	GO:0005739~mitochondrion	350	61.73	466	3.62E-250
		SP_PIR_KEYWORDS	mitochondrion	275	48.5	539	3.75E-234
		SP_PIR_KEYWORDS	transit peptide	197	34.74	539	4.56E-181
2	64.54	GOTERM_CC_FAT	GO:0044429~mitochondrial part	193	34.04	466	8.87E-147
		GOTERM_CC_FAT	GO:0005743~mitochondrial inner membrane	124	21.87	466	4.32E-97
		GOTERM_CC_FAT	GO:0031966~mitochondrial membrane	134	23.63	466	3.70E-96
3	32.71	GOTERM_CC_FAT	GO:0005759~mitochondrial matrix	77	13.58	466	1.77E-62
		GOTERM_CC_FAT	GO:0031980~mitochondrial lumen	77	13.58	466	1.77E-62
		GOTERM_CC_FAT	GO:0031974~membrane-enclosed lumen	95	16.75	466	4.55E-10
4	18.25	GOTERM_BP_FAT	GO:0045333~cellular respiration	28	4.94	400	5.90E-23
		GOTERM_BP_FAT	GO:0015980~energy derivation by oxidation of organic compounds	32	5.64	400	8.20E-21
		GOTERM_BP_FAT	GO:0051186~cofactor metabolic process	39	6.88	400	7.78E-19
5	14.49	GOTERM_MF_FAT	GO:0016651~oxidoreductase activity, acting on NADH or NADPH	26	4.59	402	8.12E-22
		GOTERM_MF_FAT	GO:0016655~oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	18	3.17	402	9.52E-17
		GOTERM_MF_FAT	GO:0050136~NADH dehydrogenase (quinone) activity	17	3	402	3.03E-16
6	14.06	SP_PIR_KEYWORDS	ribosomal protein	49	8.64	539	2.52E-29
		GOTERM_CC_FAT	GO:0005840~ribosome	49	8.64	466	1.04E-23
		SP_PIR_KEYWORDS	ribonucleoprotein	50	8.82	539	1.00E-21
7	10.35	GOTERM_BP_FAT	GO:0006006~glucose metabolic process	26	4.59	400	5.26E-10
		GOTERM_BP_FAT	GO:0019318~hexose metabolic process	27	4.76	400	6.34E-09

	GOTERM_BP_FAT	GO:0044275~cellular carbohydrate catabolic process	17	3	400	1.73E-08

Table 2.7: Enrichment of genes co-expressed with mitochondrial complex I disease genes
”List total” indicates the total number of genes annotated to this term, where “count” indicates the number of these genes in the list of co-expressed genes. The “FDR” indicates the False Discover Rate (FDR) based significance of this enrichment. There is a strong enrichment for mitochondrial processes. The strong enrichment for these specific mitochondrial processes and strong co-expression among these genes suggests they are tightly regulated.

Gene	Previous disease association evidence	Reference
Atp5j	Risk factor for ischemic heart disease end-stage renal disease	[214]
Cox7a2	No previous association evidence	
Ndufa1	No previous association evidence	
Ndufb7	No previous association evidence	
Cox7c	No previous association evidence	
Cox5b	Interacts with the human androgen receptor	[215]
Atp5f1	No previous association evidence	
D830035I06/Atp5k	Atp5k has been associated with atherosclerosis	[216]
Deb1	*C.elegans* mutants were paralyzed and had disorganized muscle	[217]
Ndufb6	No previous association evidence	

Table 2.8: Top 10 genes co-expressed with mitochondrial complex I disease related genes
References for all genes previously associated with disease have been supplied. Since multiple genes associated with disease could be recovered, the other genes in this list may also cause a mitochondrial complex I related disease if mutated or ablated.

The results included a number of genes that have been associated with several diseases amongst which Alzheimer's and Parkinson's disease. Not surprisingly, DAVID analysis identified Mitochondrion (ES = 210.25, FDR= $3.6x10^{-250}$), Cellular respiration (ES = 18.25, FDR = $5.9x10^{-23}$) and Oxidoreductase activity, acting on NADH (ES = 14.49, FDR = $2.3x10^{-22}$), as the most significant functional clusters.

The co-expressed genes include several mitochondrial complex I genes (not in seed list), multiple cytochrome c proteins, and genes involved in the ATP synthase complex. Furthermore, there are approximately 50 poorly annotated genes co-expressed. A pseudogene, 3000002C10Rik, shows a co-expression ratio of >0.50 with 512 genes. Classification of these 512 genes using DAVID results in an enrichment score of 53.7 (FDR = $2.9x10^{-70}$) for mitochondrial genes. Therefore, *3000002C10Rik* may play a biologically relevant role in mitochondrial processes.

2.3.8. Predicting functions of poorly annotated genes

To investigate if it is possible to predict or estimate a given gene's function based on its co-expression pattern, we inspected a selection of poorly annotated genes. Using DAVID, the functional categories for the top 5% co-expressed genes were obtained. Table 2.9 shows the functional categories for the poorly annotated genes with the highest significance value.

Un-annotated Gene	DAVID Functional Annotation	ES	FDR
0610006I08Rik	Mitochondrion	32.75	$1.1x10^{-40}$
0610006L08Rik	Disulfide bond/secreted	33.14	$3.1X10^{-38}$
	PeptidaseS1/Chymotrypsin	16.25	$1.4X10^{-20}$
0610010D20Rik	Peroxisome	21.8	$3.8X10^{-22}$
	Fatty acid metabolism	21.78	$1.7X10^{-22}$
	Drug metabolism/CytochromeP450	17.94	$6.1X10^{-20}$
0610031J06Rik	Lysosome	18.59	$2.4x10^{-18}$
0610037L13Rik	Ribosomal protein	16.97	$1.2x10^{-22}$
0610037M15Rik	Immune response	24.02	$2.0\ x10^{-28}$
0710008K08Rik	Vasculature development	13.26	$1.9x10^{-12}$
	Lung development	12.36	$2.4x10^{-10}$

Table 2.9: Top functional annotation clusters of the 5% genes with the strongest co-expression with the poorly annotated genes

Term clusters as defined by DAVID, (summarizing similar term definitions originating from different sources) with an enrichment score (ES) above 5 are displayed (10 for *0610010D20Rik*). Cluster titles and FDR were selected based on the most significant annotation within the cluster. For each gene a different enrichment among their co-expressed genes is uncovered. The fact that these results contain similar enrichment scores as the annotated genes used to validate our approach, is suggestive that these genes play a role in the processes noted in the table. For full lists refer to supplement 5.

While some of the categories identified are broad, others are more specific. These results show that it is possible to use GeneFriends to infer gene functions for unannotated genes. The fact that for 9 annotated genes the inferred functions are coherent with the known functions supports the notion that these inference are reliable, as shown in Section 2.3.2. This is further supported by the experimentally validated results for *C1ORF112* or *C12ORF48*, showing that genes co-expressed with cancer genes are important for proliferation speed.

2.4. Discussion

2.4.1. GeneFriends: A genetics and genomics tool for the research community

GeneFriends is freely available online (http://GeneFriends.org/) and is an intuitive tool, which can be used to identify the genes co-expressed from a user supplied gene list. This simple, yet powerful new tool can be a valuable resource for genome interpretation, annotation, mouse genetics, functional genomics, and transcriptional regulation. It may also be useful to develop network analyses of mouse genes in a variety of studies.

We tested GeneFriends to determine whether it can give biologically relevant data. We also demonstrated how GeneFriends can be used to quickly identify interesting gene targets for follow-up studies. Furthermore, we experimentally validated that two poorly annotated genes co-expressed with a cancer seed list are important for cell proliferation. Below, we discuss the findings we have obtained from our example analyses and the biological relevance of our results.

2.4.2. Validation of the co-expression map

Our analyses rely on the assumption that co-expressed genes tend to be involved in the same biological processes. Our results clearly support this, as co-expressed clusters of genes show strong enrichment for functional categories and specific processes, indicating a significant number of genes within these clusters play a role in the same process (Figure 2.1). This is further supported by the fact that co-expression partners of genes for which the function is well established show enrichment for the known functions of these tested genes (Supplement 1). There is a high degree of functional coherence between co-expressed genes. This supports the notion that our co-expression map can be used to obtain biologically-relevant information.

Given the intrinsic noisy nature of microarray data, we used a vote counting approach, which is a standard meta-analysis technique, to build our co-expression map [218]. Dealing with the noisy nature of the data is particularly important when combining large and diverse datasets and meta-analysis has been shown to increase sensitivity when studying aging, for example [170]. Using this vote counting method, single gene expression or sample outliers do not heavily impact on the overall gene correlation scores. We assumed the vast majority of samples uploaded to the public database are of high quality and therefore aimed to include as much of this data as possible. Experience has taught this is not necessarily the case and that refined quality control will likely improve the reliability of the results obtained from the network. Nonetheless, we feel our network can be reliably used to assign putative functions to unknown genes and identify possible disease gene targets. This is based on the observation that co-expression analysis for 9 well annotated genes identified the correct role and that genes not present in a seed list of disease genes are often also annotated in the context of the corresponding disease according to the literature (Table 2.2, table 2.5 and table 2.8).

2.4.3. Co-expression analysis of genes associated with aging

Using a guilt-by-association method, we identified candidate genes related to seed lists of genes previously associated with diseases or processes. Our study not only identified genes that are relevant to aging according to current theories of aging, e.g. inflammation, but it also identified novel candidates for further research. *C/ebp* transcription factors showed the strongest co-expression and are therefore candidate activators of the altered expression patterns with age. TFBS for *C/ebp* genes were identified in the aging genes and there is some evidence of a transcriptional cascade via SP1 [219]. The two proteins encoded by the *C/ebpβ* gene are liver activating protein (LAP) and liver inhibiting protein (LIP), which have opposing effects [220, 221]. The LIP protein is also capable of inhibiting other *C/ebp* proteins. This could explain why *C/ebp* transcription factors themselves are not found to be increased/decreased in expression with age. This could also be due to the fact that TFs are sometimes expressed at low levels causing the expression not to be detected by microarrays.

The fact that replacement of the *C/ebpα* gene with *C/ebpβ* increases lifespan by 20% supports the case that these *C/ebp* genes play an important role in aging [197, 198]. These different isoforms may alter the rate of aging [222], indicating that altering the isoform expression of these genes can affect lifespan. Moreover, the life-extending drug rapamycin may affect isoform ratios of *C/ebpβ*. Rapamycin has been shown to increase lifespan via the suppression of *Mtor* [223] which in turn controls the isoform ratios of *C/ebpβ* [224]. Therefore, we speculate that rapamycin may, in part, exert its life extending effect through *C/ebpβ*.

2.4.4. Co-expression analysis of cancer genes and experimental validation of candidates

We used GeneFriends to identify new candidate cancer genes. Many of the cancer genes in the initial seed list were not present in the results indicating they are not co-expressed with each other. This may be due to the fact that this set of cancer genes includes both oncogenes and tumor suppressor genes, which are not expected to be co-expressed. Also cancer can arise through different mechanisms. Therefore, the genes identified as co-expressed in this study are likely involved in common pathways leading to cancer, or are at least triggered by transformation.

Genes that are co-expressed with several oncogenes may prove to be useful targets in countering the proliferative effect of these genes in tumors. Examples of such genes are *Cdc7* and *Cdc25*. These genes were both identified as co-expressed in our study and were readily being studied in cancer context. *Cdc25* has been suggested as a therapeutic cancer target, and on-going studies in this direction have shown some level of success [202, 225-227]. Two compounds that target *Cdc7* are currently in phase I clinical trials [228]. The fact that candidate genes identified by our method have already been suggested as potential drug targets shows that GeneFriends can be useful for the identification of candidate targets for cancer studies.

Bc055324 is one of the poorly annotated genes that is strongly co-expressed with a large number of cancer genes. Knock-down of the human homolog, *C1ORF112*, in HeLa cells diminishes cell growth, which, adding the fact that *Bc055324* knockout mice are not viable [229] (http://www.europhenome.org/), demonstrates that this gene is functional, as opposed to being a pseudogene. Further studies of this gene in the context of cell cycle regulation, development, and cancer are warranted. These results show that GeneFriends can indeed be

used to identify novel targets for particular diseases. In addition, it confirms that the functional enrichment of co-expressed genes can give indications about a poorly annotated gene's function. The other poorly annotated gene co-expressed with cancer we tested was *4930547NRik* (*C12ORF48*). *C12ORF48* was recently shown to be over-expressed in pancreatic ductal adenocarcinoma (PDAC) cells [230] and in other aggressive and therapy-resistant malignancies [230]. In line with our findings in HeLa cells, knock down of the *C12ORF48* significantly suppressed PDAC cell growth [230].

2.4.5. Co-expression analysis of mitochondrial I complex disease genes

Mitochondrial complex I diseases include isolated complex I deficiency, which is the most common enzymatic defect of the oxidative phosphorylation disorders and can cause a wide range of clinical disorders [231, 232]. These include macrocephaly with progressive leukodystrophy, nonspecific encephalopathy, cardiomyopathy, myopathy, liver disease, Leigh syndrome, Leber hereditary optic neuropathy, and some forms of Parkinson's disease [233-235]. Mutations in the nuclear encoded mitochondrial genes have been previously associated with several pathologies [236, 237]. However, half of the patients with mitochondrial complex I (CI) deficiencies lack mutations in any known CI subunit. This suggests that yet unidentified genes crucial for maturation, assembly, or stability of CI may be involved in these diseases [237]. We identified several poorly annotated genes that show a strong co-expression with the mitochondrial disease gene set. As most of the other co-expressed genes encode mitochondrial proteins, these poorly annotated genes most likely also encode mitochondrial proteins. This is further supported by the fact that a number of these poorly annotated genes have been shown to be active in the mitochondria in another large-scale study [238]. Some of these genes could be responsible for the CI deficiency phenotype and are therefore promising candidates for further studies.

2.5. Conclusion

In this study we created a tool that identifies co-expressed genes from a user's seed list. Moreover, it returns the GO term enrichment of this list as well as a separate list of the co-expressed transcription factors. This allows novel candidate genes to be quickly identified for follow up studies. GeneFriends employs a biologically-relevant co-expression map and a guilt-by-association method to identify novel candidate genes for complex diseases. We demonstrated the biological relevance of this tool by analyzing aging, cancer and mitochondrial I complex deficiency seed lists. Furthermore, we experimentally validated two poorly annotated candidate genes co-expressed with cancer-related genes. We also demonstrated how GeneFriends can be used to investigate transcription factors that are co-expressed with seed genes of interest, helping to elucidate the regulatory mechanisms. GeneFriends is freely available online (http://GeneFriends.org) to other researchers allowing the identification and prioritization of candidate genes to study other complex diseases and processes.

2.6. Materials and Methods

2.6.1. Data selection

To create the co-expression map, normalized microarray datasets, obtained from the GEO database, were used [152]. GEO files GSE1 to GSE18120 were downloaded containing 16,916 datasets in total. From these, 3,850 *Mus musculus* datasets, containing 64,849 microarrays and the corresponding annotation files, were extracted. Mouse experiments are generally better controlled than human studies and there is less variation caused by genotypic factors in mice. To reduce the effect of genotypic differences between individuals on co-expression patterns, *Mus musculus* data was used instead of *Homo sapiens data*. Using mouse data also allows more datasets to be included and they originate from a more diverse set of experiments [170]. This potentially allows for the investigation of target genes in the different mouse models of aging and complex diseases.

All datasets containing annotation files that did not include gene symbols for at least 90% of the probes present in the data, were removed. All microarray datasets containing values higher than 25 were log transformed, under the assumption this data was non-log-transformed data. To remove poor signal, low quality data or data containing nonsense values up to 10^{99}, datasets containing no values above 2log(5,000) or one or more values over 2log(20,000,000), were removed . Datasets with no reference to any annotation file were removed as well. After these steps, 1,678 datasets representing 8,417 different conditions and 21,744 individual samples remained. The probe IDs were converted into gene symbols. If multiple probe IDs mapped to the same gene symbol, they were averaged. Within each dataset the experimental conditions were manually determined. Microarrays from individuals under the same conditions were averaged; in other words, replicates were averaged. Missing

gene expression values were replaced by the average expression value of the replicates. If these were also missing or not available, the gene was removed.

2.6.2. Constructing the co-expression map

To create GeneFriends, we first constructed a genome-wide co-expression map, using normalized *Mus musculus* microarray data from the GEO database. This describes which genes are related based on how often they are co-expressed. In total, 1,678 mouse datasets containing 8,417 different conditions and 21,744 individual samples met our data selection criteria. To construct our expression map, the different conditions within each dataset were compared to each other. Since different datasets contain different probes mapping to different gene symbols, a selection was made. Only those gene symbols that are present in gene platform file GPL1261 (Affymetrix GeneChip Mouse Genome 430 2.0 Array) were used. This platform contains 20,676 gene symbols and is the most common platform used for microarrays amongst those included in this work. All of these gene symbols were present in over 850 datasets.

In this work, we have used a vote counting approach to quantify co-expression for approximately 400 million (20,676*20,676) gene pairs. We used these pairs to establish if genes were co-regulated; co-regulation being defined as both genes increasing or decreasing in expression at least two-fold simultaneously, a standard (even if arbitrary) measure of differential expression, between any pair of conditions within each dataset. By only comparing conditions within the same datasets, we avoid inducing non-biological correlations resulting from inter-dataset technical biases. Then based on how often gene pairs were co-regulated compared to how often the single genes showed a two-fold increase or decrease in expression, we calculated a co-expression ratio, which quantifies how strongly two genes are

co-expressed, for all 20,676*20,676 gene pairs. The number of times two genes were simultaneously differentially expressed in the same direction (i.e. relative up or down regulated), was calculated using the equation:

$$N_{gene1,gene2} = \sum_{i=1}^{s-1}\sum_{j=i+1}^{s}\begin{cases}1, & \text{if } \left(\left(\frac{gene_{1,i}}{gene_{1,j}} > 2\right) \ \& \ \left(\frac{gene_{2,i}}{gene_{2,j}} > 2\right)\right)_+ \\ 0, & \text{otherwise}\end{cases}$$
$$\sum_{i=1}^{s-1}\sum_{j=i+1}^{s}\begin{cases}1, & \text{if } \left(\left(\frac{gene_{1,j}}{gene_{1,i}} > 2\right) \ \& \ \left(\frac{gene_{2,j}}{gene_{2,i}} > 2\right)\right) \\ 0, & \text{otherwise}\end{cases}$$

Where *s* is the total number experimental conditions (or samples, if no replicates are used), "gene" represents the expression of the gene (in sample i or j), and N is the number of times two genes are differentially expressed (in the same direction) simultaneously. The total number of times each gene was relatively up or down regulated (Q_{gene}) (i.e., >2 fold) was calculated using the following equation:

$$Q_{gene} = \sum_{i=1}^{s-1}\sum_{j=i+1}^{s}\begin{cases}1, & \text{if } \left(\frac{gene_i}{gene_j} > 2 \mid \frac{gene_j}{gene_i} > 2\right) \\ 0, & \text{otherwise}\end{cases}$$

Where *s* is the total number of experimental conditions.

From the values *N* and *Q,* the co-expression ratio was deducted. The genes were then ranked based on their *N*/*Q* ratio. A ratio of 0.50 would indicate that, if gene 1 is increased or decreased in expression in 50% of the cases, gene 2 is also increased or decreased in expression. Each gene pair is present in at least 850 datasets, thus the ratio is based on a large number of measurements. We note that datasets that the number of comparisons that can be made within 1 datasets is increasing rapidly (n!) with the number of conditions present within a dataset, causing these datasets to weigh more heavily and introducing a strong bias toward

datasets describing more conditions. To construct the RNA-seq data based co-expression network we used Pearson correlation, a more commonly applied method to assess gene co-expression (Chapter 3).

2.6.3. Testing the co-expression map

One of the objectives of the co-expression network is to assign putative functions to poorly annotated genes, based on the functional enrichment among co-expressed genes. To test whether this is possible and likely identifies the correct function, a set of 9 well annotated genes were selected and their function was predicted based on their co-expression partners: Three genes that are known to be active in fatty acid metabolism: *Ppara, Acaa2* and *Acadm*; three genes known to be involved in immune response: *Cd4*, *Cd8* and *Il10*; and three cell cycle genes: *Cdc6*, *Cdc7* and *Cdc8*. For each of these genes, the top 5 percentile co-expressed genes were selected for functional enrichment analysis, which was conducted using DAVID [72] (see Section 2.6.6 for functional enrichment analysis).

2.6.4. Prediction of novel candidate genes in aging and complex diseases

To identify genes co-expressed with known disease genes, three disease-related gene sets were included. The first of these was an aging gene set. It consisted of genes over-expressed with age, obtained from a meta-analysis of aging microarray studies in mice, rats and humans that revealed several conserved genes increasing or decreasing in expression with age [170] (Supplement 6). The second gene set included was a set of cancer-related genes [239] (Supplement 6). This is a manually curated cancer set that includes only heritable cancer genes with strong evidence that mutations in these genes are causative for cancer. The third gene set added included genes known to cause diseases through mitochondrial complex I deficiencies. The genes in this set contain the nuclear mitochondrial complex I deficiency genes in the

Online Mendelian Inheritance in Man (OMIM) database (Supplement 6). Gene symbols that were not present in the co-expression map were not included in the analysis.

Using the seed lists described in the previous paragraph, a "guilt-by-association" approach was employed with the aim of finding new potential disease-related gene targets. In this approach, the top 5% most co-expressed genes with each gene were considered “friends” of that particular gene. For each of the 20,676 genes, we calculated how many times it was “friends” with the disease related genes. Next, the probability that a gene was “friends” with this number of disease genes was calculated, as follows: The number of times each gene was “friends” with any other gene was counted and consecutively the chance a gene is "friends" with another gene was calculated.

p = total number of times this gene is friends with other genes/total number of genes

Where *p* thus is the chance that a particular gene occurs in the top 5% of a random gene.

We assume the following null hypothesis: The probability of a gene being a “friend” with one of the *n* disease genes equals the probability *p* of being a “friend” with a random gene. Then the probability of a gene being a “friend” with *k* or more genes from the disease list can be calculated by using the right-tail of the binomial distribution.

$$\Pr(K \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Where Pr(K>= k) is the probability that a gene would be “friends” with *k* or more genes in the disease gene set; *k* is the number disease gene “friends”; *n* is the number of genes in the gene set. When calculating *p* the number of occurrences of a gene in the top 5% of all genes was

included. This is necessary since some genes tend to be co-expressed more often, in general, than other genes.

To test whether there was a significantly larger number of aging genes among the co-expressed genes versus those that are not, we also used a Fisher exact test. A curated list of aging genes was obtained from GenAge [192]. We selected all *Mus musculus* aging related genes from build 18 (11/10/2015).

2.6.5. Experimental validation of cancer-predicted genes *Bc055324* and *4930547N16Rik*

To test the predictions from the analyses using GeneFriends, we took poorly annotated genes that had the strongest co-expression with the cancer disease gene list. Validated siRNAs were available from Qiagen for two the human homologs of the top poorly annotated genes: *Bc055324* (*C1ORF112*) and *4930547N16Rik* (*C12ORF48*). The experiment was conducted in human HeLa cells using standard culture conditions. A negative and a positive control were also included (Qiagen). The positive control contained a mix of several apoptosis-inducing siRNAs, demonstrating that the transfection was successful through the observed elevated cell death. The negative control consisted of siRNAs targeting non-mammalian genes. The full protocol followed for this experiment is described in supplement 7.

2.6.6. Gene set function enrichment analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) [171] was used to identify enriched functional groups within these genes. The default settings were used in this analysis. The results were ranked based on p-value and genes with a p-value $<10^{-6}$ were selected. We adapted our significance 0.05 p-value cutoff for multiple testing based on a Bonferroni correction for 20,677 genes resulting in a stringent cutoff of 10^{-6}. In addition,

several geneset, containing random sets of genes, were created and the co-expression network was used to identify functional enrichment among co-expressed genes with these random sets of genes. The significance p-values found for all results were $>10^{-5}$, indicating no significant results are found using these random sets of genes with p-value a cutoff of 10^{-6}.

Next, to understand the significance of the DAVID enrichment score, 1000 genes were randomly selected and used as an input for DAVID. This resulted in an enrichment score of 2.2 with an FDR score of 0.7 for the most significant category found. The same was done for smaller sets of genes, resulting in similar scores. This indicates that the enrichment scores of >10 and FDR $<10^{-10}$ are unlikely to be reported when there is not an actual significant enrichment among the list of genes used (as opposed to these scores being the result of some type of bias). Benchmarking using our co-expression map and COXPRESdb revealed similar results (Table 2.1), suggesting our co-expression map is not inferior to those built using correlation measures. Therefore, our work demonstrates that vote counting is a viable method to build co-expression maps.

A concern with our analysis is that some of the genes in functional categories used by DAVID were assigned based on their expression pattern; if this would be the case it could lead to circular reasoning. However, an analysis conducted by the DAVID team shows that less than 1% (78/20676) of the genes is grouped based solely on their expression pattern (See table 2.10 for a full list of these genes).

name
5'-3' exoribonuclease 2
acyl-CoA thioesterase 11
ADP-ribosylation factor-like 6
amelogenin X chromosome
antigenic determinant of rec-A protein
aryl hydrocarbon receptor nuclear translocator 2
Bardet-Biedl syndrome 1 (human)
Bardet-Biedl syndrome 4 (human)
Bardet-Biedl syndrome 9 (human)
bestrophin 1; hypothetical protein LOC100046789
bestrophin 2
bone morphogenetic protein 10
bone morphogenetic protein 2
bone morphogenetic protein receptor, type 1B
cartilage associated protein
CCAAT/enhancer binding protein (C/EBP), gamma
cDNA sequence *Bc054059*
centrin 2
centrosomal protein 57
collagen, type I, alpha 1
collagen, type XIII, alpha 1
DiGeorge syndrome critical region gene 14
DNA methyltransferase 3B
EGF-like domain 7
EGF-like domain 8
endoglin
four and a half LIM domains 2
gametogenetin
GATA binding protein 4
GATA binding protein 6
glycerol kinase-like 1
growth arrest specific 8
helicase, lymphoid specific
HERPUD family member 2
insulin-like growth factor 1
insulin-like growth factor 2
interferon regulatory factor 6
interleukin-1 receptor-associated kinase 1
intraflagellar transport 81 homolog (Chlamydomonas)
Iroquois related homeobox 1 (Drosophila)
Iroquois related homeobox 3 (Drosophila)
LIM homeobox transcription factor 1 beta
MAD homolog 1 (Drosophila)
MAD homolog 5 (Drosophila)
McKusick-Kaufman syndrome protein
microtubule-associated protein 1S

misshapen-like kinase 1 (zebrafish)
mohawk homeobox
muscleblind-like 1 (Drosophila)
neurensin 1
NK-3 transcription factor, locus 1 (Drosophila)
NK2 transcription factor related, locus 5 (Drosophila)
NK2 transcription factor related, locus 6 (Drosophila)
nuclear receptor subfamily 6, group A, member 1
peroxisome proliferator activator receptor delta
piwi-like homolog 2 (Drosophila)
placental specific protein 1
proprotein convertase subtilisin/kexin type 2
prospero-related homeobox 1
protocadherin 18
renin 1 structural; similar to renin 2 tandem duplication of Ren1; renin 2 tandem duplication of Ren1
salt inducible kinase 1
similar to iroquois-class homeobox protein IRX2; Iroquois related homeobox 2 (Drosophila)
similar to Nanog homeobox; Nanog homeobox
STEAP family member 4
stromal cell derived factor 4
suppressor of variegation 3-9 homolog 2 (Drosophila)
tescalcin; similar to Tescalcin
testis-specific serine kinase 3
tetratricopeptide repeat domain 8
timeless homolog (Drosophila)
titin-cap
triggering receptor expressed on myeloid cells-like 1
tripartite motif-containing 32; RikEN cDNA 3632413A11 gene
tubulin, gamma 1
tumor necrosis factor, alpha-induced protein 1 (endothelial)
uncoupling protein 1 (mitochondrial, proton carrier)
zinc finger protein 105

Table 2.10: List of 79 genes annotated to functional categories solely based on co-expression

One concern with our co-expression was that the reasoning would be circular if the annotated genes would have been previously annotated to a particular term because they are co-expressed with other genes in that term. To this end, we contacted the DAVID team, which were kind enough to supply a list of genes that were annotated to terms solely based on co-expression, which is listed above. This is a very small number compared to the genes annotated based on at least one other source of evidence. Therefore we are not concerned this bias is an issue in our analyses.

2.6.7. BLAST

A Position-Specific Iterative (PSI-BLAST) search was conducted with the *Bc055324* gene against the non-redundant protein sequence database. The protein sequence of this gene was recovered from GenBank. All sequences recovered in the initial search with a p-value <0.005 were used in the PSI-BLAST search. This last step was iterated twice.

Chapter 3: A human RNA-seq-based gene and transcript co-expression database

In the previous chapter, we have established a web interface and created a database that can be used to query co-expression partners for coding genes. The results described in the previous chapter, support the notion that this database and the corresponding web interface can be used to predict which biological process and well-studied diseases a poorly annotated genes plays a role in . However, this co-expression network does not contain ncRNAs, which are often poorly annotated and may play important roles in diseases. Often these ncRNAs are not studied in the context of disease, because the lack of information about their function makes it hard to interpret the functional meaning of associations between such genes and disease. In this light, the second goal of the project described in this thesis was to create a co-expression database that would also include co-expression information for these ncRNAs. To achieve this, we used RNA-seq data, which includes expression data for ncRNAs. The resulting database could be queried for genes annotated to a particular disease to uncover ncRNAs that are strongly co-expressed with these disease genes suggesting these may also play a role in these disease, as we show in an example. Additionally, our database allows users to query ncRNAs that may appear differentially expressed or mutated in the disease or process they study and acquire potential functions for these genes. This information thereby aids the interpretation of results obtained from such differential expression analyses. The predictions can also be utilized to prioritize the ncRNAs that are most likely of the user's interest and design follow-up experiments to acquire further insights into the function of these genes.

RNA-seq data can also be used to assess the expression on a transcript level to a certain extent. Additional to including ncRNAs in our database, we created a transcript specific

database that allows users to query transcripts rather than genes, adding another layer of detail to our database. This feature allows researchers to query our database for transcript level co-expression. This will aid researchers that identify differential splicing in their study with the interpretation of their results. Researchers may, for example, find that in some cases a different isoform becomes expressed in their samples of interest due to, for example, genetic variation [240]. With our database they can test if these different isoforms tend to have different co-expression partners, which would suggest they play a role in different processes and the enrichment analysis results will show which processes these are. This can help the interpretation of their results and aid the design of follow-up experiments.

In this chapter, we created a co-expression database for transcripts, as well as a gene co-expression database, that includes non-coding genes. Additionally, we investigated if it is possible to predict the process in which ncRNAs and different transcripts exert their function. To do so, a similar method was used as the method used in Chapter 2 for their coding counterparts in addition to a number of additional measures described in Section 3.3.3 to 3.3.6. In this chapter, we opted to use Pearson correlation to construct the co-expression network rather than the vote counting approach used in Chapter 2. Motivation for this choice was that Pearson correlation does not suffer from the bias described in Section 2.6.2. The same web design was used (.css file) as for the original GeneFriends website, which was originally supplied by Thomas Craig. Additionally, Thomas Craig maintained the server on which the GeneFriends website is served. João Pedro de Magalhães provided guidance with the project and helped drafting and editing the manuscript. All other work was conducted by Sipko van Dam. This work was published in *Nucleic Acid Research* in 2015 [50].

3.1. Abstract

Co-expression networks have proven effective at assigning putative functions to genes based on the functional annotations of co-expressed genes, at candidate disease gene prioritization and in improving understanding of regulatory networks. The growing number of genome resequencing efforts and genome-wide association studies often identify loci containing novel genes and there is a need to infer their functions and interaction partners. To facilitate this, we have expanded GeneFriends, an online database that allows users to identify co-expressed genes with one or more user-defined genes. This expansion entails an RNA-seq-based co-expression map that includes genes and transcripts that are not present in the microarray-based co-expression maps, including over 10,000 ncRNAs. The results users obtain from GeneFriends include a co-expression network as well as a summary of the functional enrichment among the co-expressed genes. Novel insights can be gathered from this database for different splice variants and ncRNAs, such as miRNAs and long non-coding (lncRNAs). Furthermore, our updated tool allows candidate transcripts to be linked to diseases and processes using a guilt-by-association approach. GeneFriends is freely available from http://www.GeneFriends.org and can be used to quickly identify and rank candidate targets relevant to the process or disease under study.

3.2. Introduction

The rapid expansion of microarray data over the past decade has resulted in large repositories, which have been employed in various meta-analyses. This has led to a better understanding of many biological processes and the identification of gene functions, biomarkers and targets for several diseases [161, 162, 170]. Co-expression is a type of meta-analysis, which describes the expression of genes relative to each other, and has been used for over a decade [4]. This method has proven effective at assigning putative functions to genes based on the functional

annotations of the genes they are co-expressed with, as well as a better understanding of the underlying regulatory networks [146, 241-243]. Examples of tools utilizing co-expression data derived from public databases are: GeneFriends (see below), COXPRESdb, CORNET, mouseMap, Genevestigator and STARNET2 [20, 244-249]. All of these works have used microarray data to construct co-expression networks, albeit using different metrics and approaches. Co-expression analyses have identified novel genes to be involved in diseases such as cancer [8, 243], schizophrenia [13] and type 2 diabetes [250], or processes such as stem cell regulation [108] and the cell cycle [251].

Transcriptome sequencing (RNA-seq) is a powerful and emerging technology that allows researchers to measure differential expression of genes more accurately than microarrays [252]. Like microarray databases, RNA-seq databases are growing exponentially (Figure 3.1). This creates the opportunity for meta-analyses similar to those conducted using microarrays, such as co-expression analyses.

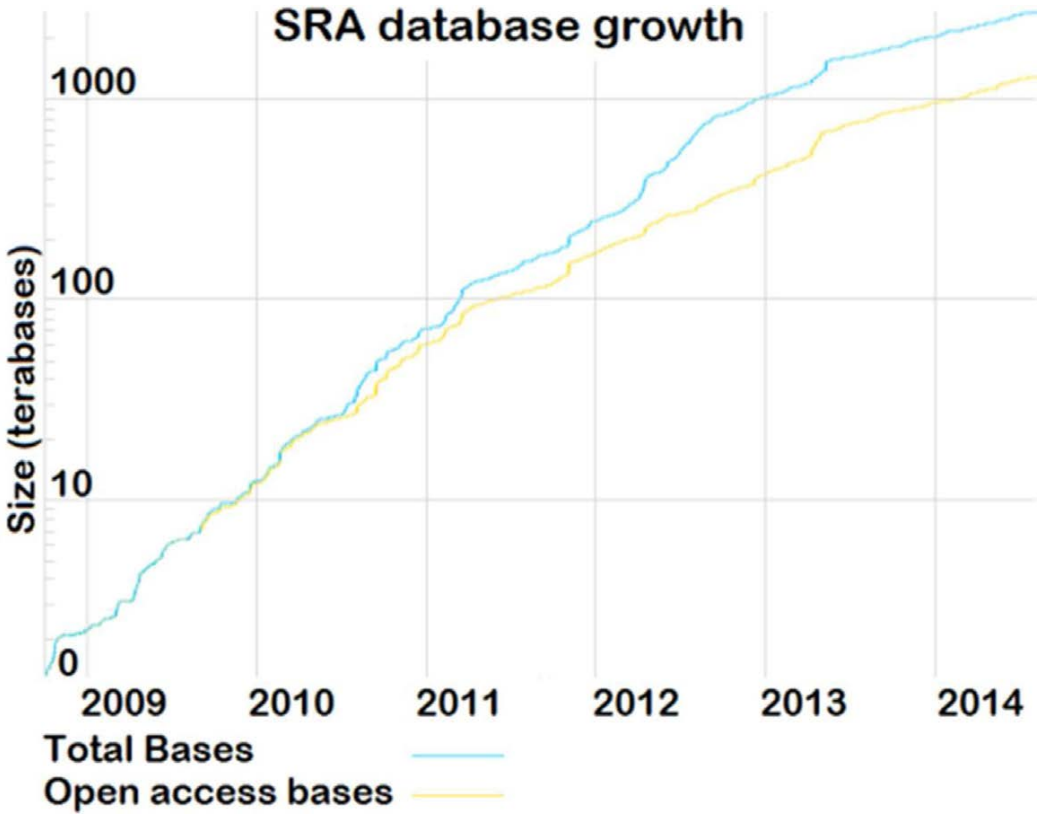

Figure 3.1: Exponential growth curve of RNA-seq data
The graph shows that the amount of available RNA-seq data is increasing exponentially at a rate of approximately 4 fold per year. If this trend is maintained in the next 5 years, 1000 fold more RNA-seq data will be available. Although the graph suggests the rate of this increase will slow down, it is still very likely that a very large amount of RNA-seq data will become available that can be used for co-expression analyses. As such, we felt that even if the data available at the start of this project would have been insufficient, this issue would resolve in time. Taken from the Sequence Read Archive (SRA) [253].

RNA-seq also measures expression of different splice variants and ncRNAs (ncRNAs), which can play important roles in gene expression regulation [29, 254]. The approximately 20,000 human genes only make up a small portion of the over 60,000 coding and ncRNAs [255] that encode the over 200,000 transcripts measured using RNA-seq [256], which greatly increases the challenges faced by researchers when interpreting RNA-seq results. A bottleneck in RNA-seq analyses is that even though a large number of transcripts can be detected as differentially expressed, often many have not been well studied. It is frequently unclear what possible functions poorly studied genes, especially non-coding ones, may have. As such, interpreting results from RNA-seq experiments and understanding the mechanisms involved in the disease or process under study is often impeded. Given the growing community of researchers employing RNA-seq, there is an unmet need for resources that help interpret results from such experiments. Moreover, the growing number of genome resequencing efforts and genome-wide association studies often associate loci containing poorly studied genes, such as ncRNAs, with diseases and traits [31, 257]. To interpret the biological meaning of such identified associations, there is a need to infer putative functions and interaction partners of new candidate genes [31, 258].

As a result of the rapidly evolving sequencing technologies, there are now more RNA-seq samples available than there were microarrays at the time of the construction of the first widely used plant [259] and mammalian [6] co-expression websites. Recently, the first co-expression analysis using RNA-seq data was conducted using 21 striatal samples and showed that co-expression networks created from RNA-seq data are more robust than those created from microarray data [260]. This co-expression map, however, is striatal-specific and is not available online to the research community. No RNA-seq-based co-expression database is currently available for humans or for biomedical models (co-expression tools like CORNET,

Genevestigator, and COXPRESdb are based on microarray data). In this work, we developed the first online RNA-seq co-expression database for the bioscience community.

We had previously created an online co-expression analysis platform using over 3,000 microarray datasets to facilitate the identification of candidate gene targets based on a user-defined list of disease- or process-related genes [20]. This tool, GeneFriends, can be used to assign putative functions to poorly studied genes using a guilt-by-association method (i.e., by investigating which genes a given poorly-studied gene is co-expressed with); it can also identify and prioritize novel candidate genes for further study based on a seed list of genes associated with a given disease or process. This allows allowing researchers to identify novel genes relevant to their study without conducting a microarray or RNA-seq experiment. This tool has been successfully used to identify novel cancer-related genes that were validated experimentally [20]. Whilst many tools are available to identify the function of genes and associate new genes with a seed list, based on different types of interaction data [248, 261-263], information on interaction of ncRNAs is more limited. Therefore, in this work we have created and integrated into GeneFriends a co-expression map, constructed from RNA-seq data, which allows for a better understanding of the regulatory patterns of ncRNAs in relation to mRNAs. Since RNA-seq allows researchers to assess the expression of different transcripts rather than only the gene level expression, we have also constructed a transcript level co-expression map. This is particularly of interest since different transcripts originating from the same gene can have different functions [264] and co-expression is an easy way to detect different co-expression partners, which suggests different functionality.

Understanding the regulated and coordinated changes that occur between ncRNA and coding (including splice variants) expression may reveal novel important players in biological

processes and diseases. Furthermore, RNA-seq has a larger dynamic range and measures expression of more genes, including those previously un-annotated. These include ncRNAs, such as miRNAs and long ncRNAs (lncRNAs), which may be crucial to understand the mechanisms underlying disease and biological pathways. This co-expression map allows these RNAs to be associated with known genes for inferring their function as well as with diseases, processes and pathways, leading to new associations that can be further investigated experimentally. GeneFriends is freely available on http://www.GeneFriends.org.

3.3. Results

3.3.1. Construction of the RNA-seq-based co-expression map

The RNA-seq-based addition to GeneFriends represents 2 co-expression maps: One containing genes (both coding and non-coding) and one containing transcripts. The RNA-seq-based co-expression map was constructed using 4133 quality controlled RNA-seq samples across 240 studies, obtained from the SRA database [265] (Supplement 8). Our aim is to create a co-expression map that captures the behavior of genes under different circumstances. For this reason data describing a range of different cell types was used (Table 3.1).

Number	Cell Types
723	Stem
716	Lymphoblastoid
552	Embryonic
289	Neurons
279	Hesc
268	Extraembryonic
268	Hesc-derived
268	Mesodermal
226	Myeloid
219	Neural
203	Progenitor
170	Pluripotent
158	Fibroblasts
149	Differentiated
124	Blood
117	Breast

Table 3.1: Cell types of the samples included in the construction of the RNA-seq based co-expression network

Determined by counting how many sample descriptions contain each word describing a tissue or cell-type. We calculated the Shannon–Wiener index for the words describing the 16 different tissue types/states in this table resulting in a corresponding Shannon–Wiener index of 2.6 (where the absolute maximum would be ln(16)=2.77, which value indicates maximum possible diversity among 16 description terms).

For condition-specific genes, a co-expression map created from a smaller set of samples may result in a more accurate result [249, 266], but this is not the purpose of this tool, which is aimed at identifying the general role and associations of genes and transcripts. Each included sample complied with the following criteria:

1. Measured using the Illumina HiSeq2000 platform (although in future updates we anticipate also incorporating more recent platforms, like HiSeq2500)

2. Contained at least 10 million reads

3. Used a cDNA library preparation protocol

4. A minimum of 60% of the reads mapped to the Ensembl GRCh37 human genome [255]

The samples were mapped using STAR [55] and read counts per gene were determined with a custom Java program, named ReadCounter (Appendix I). We opted to create our own counting tool since the widely used HTSeq tool [59] was too slow for our purposes. ReadCounter is more efficient, running approximately 3-fold faster on a single core (not shown). Additionally, ReadCounter utilizes multithreaded technology, which, using 8 cores on our system, resulted in a 15 to 20 fold faster runtime. For benchmarking, ReadCounter has extra options that allow results to be identical to those obtained from HTSeq, albeit at a much faster rate. Moreover, ReadCounter can more accurately assess the gene of origin in case a read is overlapping multiple genes on the genome, utilizing the overlap size of the reads with the different genes in a certain region. This advantage has been utilized when constructing our co-expression map. Furthermore, ReadCounter has another advantage. It automatically counts the number of reads mapping to introns as well as reporting ambiguously mapping genes in a separate column. ReadCounter is written in Java and can be run using a command line in the terminal or

command prompt (Mac/Linux/Windows) without the requirement of installation. The tool is free to use and publicly available at http://www.GeneFriends/ReadCounter. A more elaborate description is included on the website. To define the gene regions, the Homo_sapiens.GRCh37.75.GTF annotation file was used. This is based on the human genome assembly 37 [267]. For normalization, the expression per gene/transcript was divided by the combined expression of all genes/transcripts per sample (note that reads that do not map to genes are excluded from the normalization procedure). The resulting data was used to construct the co-expression maps.

To create the co-expression map, we employed the same approach that COXPRESdb used to construct their microarray-based co-expression map [6]. For each possible gene pair combination a weighted Pearson correlation, based on sample redundancy, was calculated. The sample redundancy is calculated based on the number of similar samples in the dataset, and the sample similarity is measured by the correlation between samples (http://coxpresdb.hgc.jp/help/coex_cal.shtml). Next, a mutual rank was calculated based on the ranking of each gene with its partner. The mutual rank is the average rank of two genes relative to each other. For example, we rank all genes based on their expression correlation with gene A and find gene B ranks in e.g. 15^{th} position. Similarly we rank all genes based on their expression correlation with gene B and find that gene A ranks in 100^{th} position. Then the mutual rank of gene A to gene B (and vice versa) is (15+100)/2. This causes genes, such as ribosomal genes that are often strongly co-expressed with many other genes, to have a lower ranking. This is preferred since these genes are often not of interest for functional enrichment analysis or candidate gene prioritization. On the other hand, genes that are more specialized (i.e., playing a role in only a specific biological process) will rank higher.

3.3.2. Database content and user guide

The GeneFriends database, constructed from RNA-seq data, contains co-expression data for 44,248 human genes and for 114,936 transcripts. Transcripts/genes that were not expressed (expression < 10 reads) in at least 10% of the samples were excluded from the co-expression map. As a result, 19,430 out of 63,678 genes and 100,234 out of 215,170 transcripts were excluded. A list of the types of genes found in the co-expression map are shown in table 3.2.

Gene type	Genes	Transcripts
Protein coding	18658	82528
Pseudogene	9483	9888
LncRNA	4997	6221
Antisense	4537	6476
MiRNA	1024	1017
SnRNA	819	814
SnoRNA	444	448

Table 3.2: List of genes and corresponding types present in the co-expression map
A division of different gene types and corresponding transcripts present in the respective databases. Coding genes have far more expressed splice variants than ncRNAs, which are less commonly spliced overall. A more detailed list can be found in supplement 9.

To employ GeneFriends, the user can submit one or multiple gene/transcript IDs. The results then contain the following sections: (i) A list of the 50 strongest co-expressed genes and the corresponding Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) annotation for each gene; (ii) A list of the 25 strongest co-expressed transcription factors; (iii) Top 20 functional enrichment categories of the co-expressed list of genes, including GO [115], Kyoto Encyclopedia of Genes and Genomes (KEGG) [268] and OMIM [269]. To assess functional enrichment among the co-expressed genes, DAVID web services [72] are used, which is a commonly used tool to assess overrepresentation of functional categories among a list of genes. To obtain the DAVID web results the top 1,500 co-expressed genes/transcripts are used (or fewer if there are fewer genes significantly co-expressed (cutoff p-value < 10^{-6}; since correction for multiple testing using the Bonferroni correction: 0.05/44248= $1.12*10^{-6}$ [20]). Additionally, full lists can be downloaded, as well as a network file that can be imported into Biolayout [76] or Cytoscape[75] for visualization and further analyses. Lastly, there is an option to download the functional enrichment of those genes that have an expression pattern which negatively correlates with the expression of the gene(s) of interest, thus those genes with an opposing expression pattern. This is especially interesting for genes/RNAs that downregulate expression of others. Further details can be found on http://www.GeneFriends.org/RNA-seq/about/. A graphical overview of the steps involved in retrieving results from GeneFriends is depicted in figure 3.2.

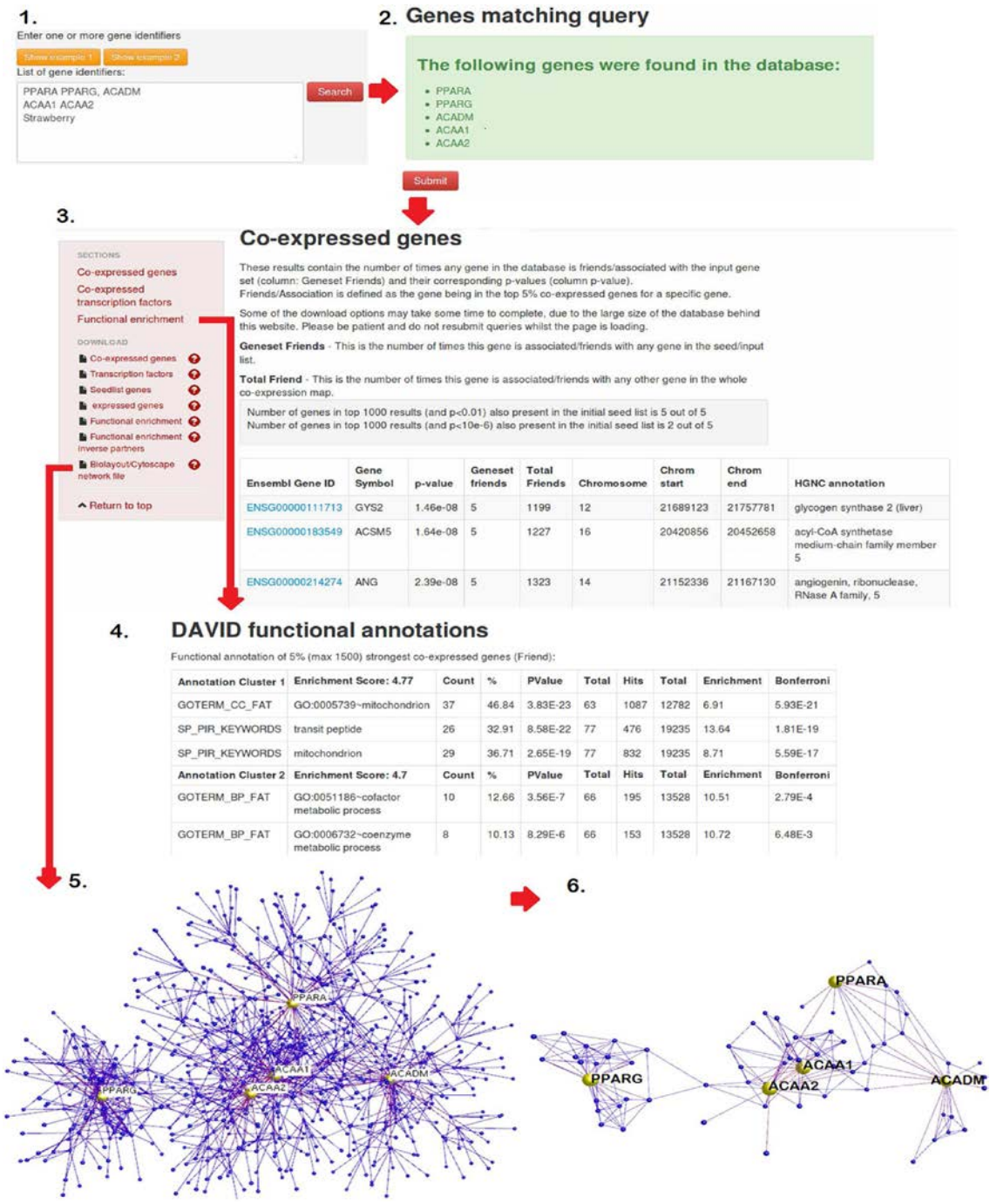

Figure 3.2: A graphical overview of the steps involved in retrieving results from GeneFriends
1. Insert genes 2. Validate input 3. Retrieve co-expressed genes 4. Investigate functional enrichment 5. Visualize the network of co-expressed genes using BioLayout 6. Use BioLayout to select the network of interest by setting different thresholds.

3.3.3. Gene co-expression based function prediction validation

One of the two main purposes of GeneFriends is that users can input a poorly annotated gene or transcript and utilize the functional enrichment of its co-expressed partners to associate it with specific biological processes. To validate this approach, we tested 9 genes for which the functions are well established. We previously used this approach to validate our microarray-based co-expression map [20] and decided to use the same set of genes. We initially picked three categories; cell cycle, immune system and fatty acid metabolism and picked three genes of which we expected co-expressed genes to be functionally enriched for these categories, based on known associations. We expected the following genes to associate with the following categories; the cell cycle: *CDC6*, *CDC7*, *CDCA8*; the immune system: *IL10*, *CD4*, *CD8*; fatty acid metabolism: *ACADM*, *PPARA*, *ACAA2* (Supplement 10). We used DAVID [72] to identify functional enrichment among the top 5% co-expressed genes. For all genes, this showed significant enrichment for the predicted categories, supporting the notion that this approach can be used to elucidate which processes poorly annotated genes play their primary role in. Moreover, for some genes the more specific roles, such as mitochondrial oxidation for *ACADM* and *ACAA2* within these general processes, showed the strongest enrichment, to which these genes are indeed annotated [115]. Others, such as *PPARA*, that are known to be associated with a wider range of processes [270, 271], showed enrichment also for this wider range of processes, underlining the potential of this approach. From these results, we conclude that co-expression results obtained from GeneFriends can be used to predict the processes the genes/transcripts are associated with.

To further support this claim, we estimated the performance of our gene function predictions on a larger scale. We tested if genes that are annotated to the same Reactome term are more strongly co-expressed than those that are not. First, we isolated each gene annotated to one

or more of the 1730 Reactome terms. For each of those 8447 genes, we tested if the genes belonging to the same term ranked significantly higher or lower (based on co-expression with that particular gene) than those that are not part of the term. We conducted a two tailed Mann-Whitney U test between the rankings of the genes belonging to the same Reactome term compared to those not belonging to the that Reactome term. We found for 69.5% of the gene-Reactome term relationships that the genes belonging to the same Reactome term ranked significantly higher (based on co-expression with that gene) than the other genes (p-value <0.01) (Supplement 11). This indicates that genes that are annotated to the same Reactome term tend to be more strongly co-expressed than those that are not. For 10.5% of the gene-Reactome relationships, the genes belonging to the same Reactome term ranked significantly lower than the other genes (Supplement 11) (p-value <0.01). And for 19.9% of the gene-Reactome relationships, there was no significant difference between the ranking of the relative Reactome term genes versus the other genes (Supplement 11) (p-value >0.01).

The fact that for most gene-Reactome term combinations the genes annotated to the pathway indeed rank higher supports the notion functional enrichment analysis of co-expression partners can be used to assign putative functions to the genes in the majority, but not in all cases. To put this in perspective and assure these relationships does not also occur randomly, we reconstructed the co-expression network, but prior to doing so we scrambled the expression of each gene. This scrambling entails the reassignment of each expression value within a sample to another gene randomly. In this way, we randomize our data, but the distribution of the expression values within a sample remains the same. Next, we conducted our analysis, as described in the first paragraph, again. The aim of this exercise is to test if the genes associated with the respective Reactome term also rank higher than those that are not part of the respective Reactome term by random chance. In this randomized case, we found

that for 1.7% of the gene-Reactome relationships the relative Reactome term genes ranked significantly higher than the others (p-value <0.01) (Supplement 12). For 0.5% of the gene-Reactome relationships the relative Reactome term genes ranked significantly lower than the others (Supplement 12) (p-value <0.01). And for 97.8% of the gene-Reactome relationships there was no significant difference between the ranking of the relative Reactome term genes versus the other genes (Supplement 12) (p-value >0.01). This shows that the higher ranking for 70% of the gene-term relationships observed in the previous paragraph are not commonly observed in a network constructed using randomized data that have the same distribution.

The observation that for a minority of genes the genes that do not belong to the same Reactome term rank higher in terms of co-expression, we speculate, may be the result of negative regulators, e.g. inhibitors within a pathway, which are also part of these Reactome terms, but more likely have an inverse expression of the rest of the genes in the pathway. It is difficult to systematically assess if this is indeed the case on a genome wide scale. Another plausible explanation could be that different proteins, translated from the same gene, have opposing functions. This is a phenomenon that is occasionally observed in biology such as with the *C/ebpβ* gene [220]. The fact that for 20% of the genes we do not observe a significantly higher ranking for genes contained within the same term as those that are not, can be explained by the fact transcription is just one layer of genomic regulation and that other regulatory mechanisms likely cause our co-expression method not to be 100% accurate. However, the majority of genes that are annotated to the same Reactome term do indeed rank higher based on their co-expression correlation. This supports the notion that terms for which genes rank higher based on correlation, represent the functional role of those genes.

Next, we tested for each Reactome term if the genes annotated to the term themselves rank higher than those that are not when querying for the co-expression of all genes annotated to the term. The purpose is to test if our geneset (rather than gene specific) based co-expression analysis (as described in Section 2.3.3) is able to retrieve genes that are part of the same geneset. To do so, we used a set of 1730 Reactome terms (Supplement 13) and their associated gene lists (varying from 1 to 2372 genes). These Reactome terms represent biological pathways. Genes annotated to the same term thus are part of the same pathway and thus partake in the same biological functions (e.g. cell growth, immune response or fatty acid metabolism). For each term, we queried GeneFriends for co-expressed genes with the geneset that is annotated to this term. We next applied a one-tailed Mann–Whitney U test to determine if genes that are part of the geneset ranked higher on average than those that are not (Supplement 14). We found that for 1501 (87%) of the Reactome terms there was a significant difference (p-value<0.05) between the ranking of the genes within the seed geneset compared to those not in the geneset, with an average AUC of 0.82 (derived from U statistic from Mann–Whitney U test). The average AUC for genesets containing more than 10 genes was 0.85 and for those containing more than 1000 genes 0.82. The genesets for which no significant difference was observed were mostly either very large (>1000 genes) or very small (<10 genes). The fact that the AUC is not 1 is in accordance with our previously reported result that for a small portion of the genes the ranking of genes in the same Reactome term was actually lower (10.5%), as well as a number of genes that showed no significant difference (19.9%). Therefore these results are in accordance with our expectations.

We also compared the co-expressed gene lists from the RNA-seq-based co-expression map to our previously constructed microarray-based map [20]. Unlike the RNA-seq-based map, the microarray version was created using a vote counting approach and includes a wider range of

data with data from over 4,000 experiments rather than the 240 included in the construction of the RNA-seq version. In an ideal world one would expect this overlap to be 100%, but some differences are to be expected. However, the overlap we observe, as described in table 3.3, is lower than we would have expected. Possible explanations can be the nature and quantity of data that underlies the co-expression network. These observed differences support the notion that the co-expression maps are dependent on the data they are constructed from. Additionally, some biases may exist in RNA-seq data that are not present in microarrays and vice versa. Nonetheless, either co-expression map proves effective at identifying the correct function using the functional enrichment among co-expressed genes for the 9 annotated genes, suggesting that the different co-expressed genes are annotated to similar functional categories.

Gene	Overlap 5 percentile co-expressed genes Microarray and RNA-seq
ACAA2	24%
ACADM	24%
CD4	39%
*CD8*A	34%
CDC6	31%
CDC7	31%
CDCA8	25%
PPARA	9%
IL10	17%

Table 3.3: Overlap of the microarray-based co-expressed gene list with the RNA-seq-based co-expressed gene list

The percentages indicate the percentage of genes that are in the top 5 percentile co-expressed genes in both the RNA-seq and the microarray based co-expression network. These numbers are lower than we expected and support the notion that the data from which the co-expression is constructed lead to different co-expression networks. This is supported by the fact that co-expression networks created from different tissues result in different co-expression networks [143]. The RNA-seq co-expression network is created from a set of samples with a different balance between tissues and cell types than those used for the microarray network, potentially explaining the difference noted in this table. The functional enrichment of co-expressed genes from either network are comparable indicating the genes still associate with the same biological functions. A more elaborate table can be found in supplement 16.

3.3.4. Tissue and cell type diversity of used datasets

Although the 240 studies from which we used the RNA-seq data describe a wide range of conditions, certain conditions may be overrepresented. We counted the prevalence of terms in the summaries of each sample (Supplement 15). The most prevalent tissue and cell type description terms are "stem" and "lymphoblastoid". These were present in 723/4133 (17.6%) and 716/4133 (17.3%) sample summaries respectively. We calculated the Shannon–Wiener index for the words describing different tissue types/states. In total, we detected 16 description terms describing different tissue or cell types with a corresponding Shannon–Wiener index of 2.6 (where the absolute maximum would be ln(16) = 2.77, which value indicates maximum diversity). There was no strong overrepresentation for any disease related terms, "cancer" (259/4133 samples (6.2%)) being the most prevalent. Since co-expression data has been reported to be tissue and condition dependent [249, 272], we anticipate differences between microarray- and RNA-seq-based maps. Although the expression ratios of the microarray version cannot be directly compared to the Pearson correlation, or mutual rank calculated for the RNA-seq version, it is still possible to compare the ranking of each gene to each other. Only genes present in both co-expression maps were included in this analysis. Conducting this analysis, using the 9 genes also used for the validation of the microarray and RNA-seq co-expression network, showed an average overlap of 27% (Standard deviation: 9%) of the top 5% co-expressed genes in the microarray with the top 5% co-expressed genes in the RNA-seq version (Table 3.3).

3.3.5. NcRNA validation

To investigate if it is possible to use GeneFriends to postulate the function of ncRNAs, we investigated the functional enrichment of genes co-expressed with 3 annotated ncRNAs. One ncRNA, *Evf-2*, known to cooperate with *Dlx2,* which plays a critical role in neuronal

differentiation and migration, as well as craniofacial and limb patterning during development [273], and two lncRNAs: *Xist*, a lncRNA active, during embryogenesis, known to trigger X-chromosome inactivation in mice [274, 275], and *HOTAIR*, a lncRNA that is required for silencing of HOXD genes, which, if absent, causes severe limb and genital abnormalities. [276, 277].

We found that genes co-expressed with *EVF-2* (ENSG00000231764) are strongly co-expressed with synaptic transmission (1.61E-50) and neuron projection (1.71E-44) (Supplement 17), which is in accordance with our expectations (Bonferroni corrected p-values are marked in parentheses). *XIST's* co-expressed genes were enriched for embryogenic morphogenesis (1.75E-3) and were most strongly enriched for transcription (9.10E-57), cell cycle (1.70E-18), chromosome organization (2.33E-21) and zinc finger regions (5.72E-35) (Supplement 17). Although there is enrichment for embryogenic morphogenesis, other terms show much stronger enrichment. The enrichment for chromosomal organization is in line with literature, as X-chromosome inactivation involves major reorganization of the X-chromosome and is triggered by *Xist* in mice [275]. We found that the co-expressed genes for *HOTAIR* were enriched for the HOX homeodomain (2.37E-3) and that they are most strongly enriched for the ontologies spermatogenesis (1.72E-13) and reproduction (1.94E-16) (Supplement 17). These results support the notion that GeneFriends can be used to predict functions of ncRNAs.

Since we were curious if functional enrichment could also be detected for genes for which no functional annotation is yet available, we also randomly selected poorly annotated genes until we found 3 with significant functional enrichment. As a result, we tested 4 genes and found significant enrichment for functional categories for 3 of these genes (Table 3.4), supporting the notion that GeneFriends can assign putative roles to poorly studied genes.

Gene	Enrichment Score	Term	Count	List Total	FDR
ENSG00000232862	15.19	GO:0019953~sexual reproduction	66	444	4.71E-21
		GO:0048232~male gamete generation	51	444	3.55E-18
		GO:0007283~spermatogenesis	51	444	3.55E-18
ENSG00000258776	45.52	GO:0050953~sensory perception of light stimulus	97	606	3.40E-69
		GO:0007601~visual perception	97	606	3.40E-69
		SP_PIR_KEYWORDS ~vision	76	845	1.79E-61
ENSG00000271947	15.66	GO:0045202~synapse		102	1005 9.54E-29
		SP_PIR_KEYWORDS ~synapse		74	1446 1.10E-26
		GO:0044456~synapse part		75	1005 4.50E-22

Table 3.4: Top enrichment categories for 3 poorly annotated genes
Terms for which the co-expressed genes are most significantly enriched, are listed. The fact that these results contain similar enrichment scores as the annotated coding genes used to validate our GBA approach (Section 3.3.3) suggests that these genes play a role in the processes noted in this table.

The following functional enrichment was found for these 3 genes, which are all lncRNAs (Bonferroni corrected p-values marked in brackets): ENSG00000271947, synapse (3.23E-29); ENSG00000258776, visual perception (4.22E-69); ENSG00000232862, sexual reproduction (5.19E-21).

We also conducted a systematic analysis on a much larger scale. For each non-coding gene, we assessed if there was an enrichment for any Reactome term annotated genes. For each of the 1730 Reactome terms we determined if the genes that are annotated to the term ranked higher in the co-expression list compared to those annotated to the term, for each non-coding RNA. We used a Bonferroni corrected p-value cutoff of 0.01/1730= 5.78e-6. We found a significant enrichment for at least one of the Reactome terms for 4659 out of 4834 (96%) of the lncRNAs, which is the largest ncRNA category (results for other categories can be found in supplement 18). The average AUC for the most significant term per gene was 0.80. We also conducted this analysis on a network created from scrambled data in which hardly any significant enrichment for these terms is expected to be observed. This was indeed the case. For only 0.2% of the genes a significant enrichment was observed after scrambling (Supplement 19). This further supports the notion that co-expression analysis on non-coding genes can be used to predict functions. However, we do note that in the majority of the cases the category for which the co-expressed genes rank higher most significantly is the olfactory signaling pathway (3127 out of 4659 lncRNAs), whereas this is not the case for the coding genes (1945 out of 18374 coding genes, which number is in accordance with the over 1000 of reported human olfactory receptor genes [278, 279]) (Supplement 20). Additionally, we compared the results obtained for lncRNAs to those obtained for pseudogenes and found that also these had a significantly higher ranking of genes associated with at least one term compared to those not associated for 9006 out of 9314 of the pseudogenes. This was contrary

to our expectation, which was to find little to no significant categories for any of these terms, since pseudogenes by definition are non-functional [280]. Three possible explanations that could potentially explain this observation are: 1. Some of these genes perhaps do have a function and are potentially incorrectly annotated as pseudogenes. 2. These may be genes that have lost their function throughout evolution, but are co-located and co-transcribed on the genome with other genes playing a role in a particular pathway that have not lost their function. This is supported by the observation that intact olfactory receptor genes and pseudogenes tend to be co-located on the genome [278, 281, 282]. This would cause them to remain co-expressed with the genes still participating in that particular pathway even though being non-functional. This would fit with the observation that many of these genes are predicted to be involved in the Olfactory Signaling Pathway (4193 out of 9006) on which the selection pressure is much lower in humans than in other species that much stronger rely on their olfactory sensing pathways to survive [283]. These two potential explanations may also be applicable to the observed overrepresentation of the Olfactory Receptor Reactome term. Since this overrepresentation is not observed for coding genes we do expect that these results are not caused by a technical bias. However, we do think it is important to consider that a significant enrichment for a particular term among co-expressed genes does not imply functionality. Nonetheless, we do suspect it represents the potential function to which these pseudogenes and non-coding genes are, or were earlier in evolution, most likely associated.

3.3.6. Transcript-specific co-expression

Since one of the benefits of the RNA-seq-based co-expression map is that it also contains transcripts, we investigated if it is possible to differentiate between the function of different transcripts originating from the same gene. To this end, we have selected a gene that has multiple transcripts with different co-expression partners, annotated to different processes:

MACF1. This is a protein that binds to actin and microtubules [284] and is important for cell motility [285-287].

We identified the two transcripts with the least overlapping partners, ENST00000360115 and ENST00000482035, which shared only 80 out of their 5747 (top 5%) co-expression partners. Next, we investigated the functional enrichment for the co-expressed transcripts of the two transcripts originating from the same gene. We found that this functional enrichment shows different categories. The top 5% co-expression partners of the ENST00000360115 transcript showed strong enrichment for the GO terms (Bonferroni corrected p-values are marked in parentheses) synapse (2.46 E-27) and neuron projection (3.88E-21), whereas ENST00000482035 partners show strong enrichment for regulation of cell motion (8.19E-12) and extracellular matrix (7.82E-14). The top categories to whichENST00000360115 was associated were not present in the enrichment results for ENST00000482035 and vice versa (Table 3.5).

Transcript: ENST00000482035

Annotation Cluster	Enrichment Score	Category	Term	Count	List Total	FDR
1	12.5	GOTERM_BP_FAT	GO:0051270~regulation of cell motion	70	1811	3.72E-12
		GOTERM_BP_FAT	GO:0030334~regulation of cell migration	63	1811	2.46E-11
		GOTERM_BP_FAT	GO:0051272~positive regulation of cell motion	43	1811	1.11E-09
		SP_PIR_KEYWORDS	extracellular matrix	80	2415	1.67E-13
2	10.92	GOTERM_CC_FAT	GO:0031012~extracellular matrix	104	1787	1.24E-11
		GOTERM_CC_FAT	GO:0005578~proteinaceous extracellular matrix	95	1787	5.62E-10
3	10.89	GOTERM_BP_FAT	GO:0001944~vasculature development	81	1811	3.52E-11
		GOTERM_BP_FAT	GO:0001568~blood vessel development	79	1811	8.03E-11
		GOTERM_BP_FAT	GO:0048514~blood vessel morphogenesis	63	1811	1.32E-06

Transcript: ENST00000360115

Annotation Cluster	Enrichment Score	Category	Term	Count	List Total	FDR
1	18.8	GOTERM_CC_FAT	GO:0045202~synapse	108	1179	6.69E-27
		GOTERM_CC_FAT	GO:0044456~synapse part	83	1179	2.38E-23
		SP_PIR_KEYWORDS	synapse	72	1695	6.55E-21
2	18.26	GOTERM_CC_FAT	GO:0043005~neuron projection	96	1179	1.06E-20
		GOTERM_CC_FAT	GO:0042995~cell projection	140	1179	2.59E-16
		GOTERM_CC_FAT	GO:0030425~dendrite	49	1179	1.86E-10
3	9.74	UP_SEQ_FEATURE	domain:SH3	47	1694	2.82E-08
		SP_PIR_KEYWORDS	sh3 domain	52	1695	2.21E-08
		INTERPRO	IPR001452:Src homology-3 domain	52	1548	2.94E-07

Table 3.5: Top 3 enrichment categories for the following 2 transcripts originating from the same gene: ENST00000360115, ENST00000482035

Co-expression partners from the two different transcripts show different functional enrichment categories. This supports the notion that co-expression analysis can be used to

associate different transcripts, originating from the same gene, to different functions. Our co-expression network could thus be used to identify all genes for which different splice variants associate with different functions. This could be interesting for annotation purposes of these transcripts and corresponding genes. For a full list of enriched categories, we refer to supplement 21.

This shows that there can be a clear distinction between the co-expression results obtained from different transcripts originating from the same gene, and that it is possible to postulate which genes encode transcripts that lead to proteins involved in different processes.

Next, we aimed to identify how often transcripts originating from the same gene are co-expressed with different transcripts. Doing so for each gene resulted in 294,829 comparisons. Of these 294,829 comparisons, 123,650 have less than 10% of overlapping transcripts in the top 5% co-expressed transcripts. This suggests that different transcripts arising from the same gene are often expressed under different conditions and are likely to play roles in different processes, or that some may be non-functional transcripts.

To determine for how many of these transcripts we can assign a putative function, we conducted the same analysis as we did on ncRNAs. We found that for 103639 out of 114933 transcripts there was a significant enrichment for a particular term (Supplement 22). The term which most commonly had the most significantly higher ranking for the genes annotated to the term was “Olfactory Receptor”, as previously observed on a gene level as well, but to a far lesser extent than with the lncRNAs and pseudogenes (13430 out of 103639 transcripts). This indicates that there is no bias toward this Olfactory Receptor pathway for the alternatively spliced variants of the coding genes (which have many more splice variants than most non-coding RNAs). We also conducted this analysis on a matrix constructed from randomized data, where the expression values per transcript within a sample had been randomly reassigned. This causes the data to have the same distribution, but functional enrichment analysis on co-expression partners of a particular gene or list of genes should no longer functional enrichment for any terms. When the same approach was tested on this data, a significantly higher ranking for genes annotated to a Reactome term was observed in only 0.3% of the

cases (Supplement 23), just like the randomized data on a gene level as described earlier. These results suggest it is possible to predict functions on a transcript specific level as well, and that the results we find are not the result of biases in our method.

3.3.7. Gene set co-expression

The second purpose of GeneFriends [20] is that users can submit a list of genes or transcripts associated with a specific disease or biological process to find other genes/transcripts associated with it. This is particularly of interest with the RNA-seq-based co-expression map as it contains non-coding genes, which may play crucial roles in understanding the mechanisms underlying these diseases/processes.

Similar to our previous analysis [20], we used a set of causative cancer genes [239] and identified genes co-expressed with this list of genes. Interestingly, this included a number of genes that one would not find in a microarray-based co-expression map. Using this approach 83 pseudogenes, 1 miRNA (*MIR4444-1*) and 2 antisense RNAs (*EMC3-AS1*, *UBL7-AS1*) were associated with the cancer seed list (Supplement 24). Genes co-expressed with miRNA 4444-1 (Supplement 23) are strongly enriched for genes involved in transcription (Bonferroni corrected p-value: 8.67E-20) and chromatin organization (Bonferroni corrected p-value: 2.58E-14), suggesting this miRNA may exert a role in cancer by affecting the expression profile in cancer cells. This is an example of how GeneFriends can be used to associate non-coding factors with diseases/biological processes and how it can help elucidate possible roles of poorly annotated factors uncovered through this procedure.

3.3.8. RNA-seq-related biases

While GeneFriends provides a unique opportunity to elucidate the roles of unstudied genes, it is important to mention a few possible biases that might be present in the RNA-seq co-

expression map. Since the co-expression map is created from RNA-seq data, any biases existing in this type of data will propagate to the co-expression map, in particular:

(i) In the library preparation of RNA-seq experiments, there is a bias against smaller RNAs [288] for which reason measurements for shorter RNAs, such as miRNAs, may be less accurate.

(ii) One important step in RNA-seq analysis is to assign reads to genes based on their coordinates. However, in some cases genes overlap with each other, making it hard to assess from which gene the read originates. As a result, the read is then ignored. This means that genes that are fully overlapped by other genes can never show expression and this becomes a major issue when mapping to transcripts rather than genes as they commonly overlap each other. For this reason we considered ambiguously overlapping reads to represent the expression of each transcript it overlaps with rather than ignoring it. This will mean that transcripts spawning from the same gene are much more likely to show strong co-expression, which should be considered when retrieving transcript co-expression results from GeneFriends. This can be circumvented using knowledge acquired from the mapping of other reads, e.g. using a Bayesian approach, as described in the introduction Section 1.2, although this will lead to different biases.

(iii) We observed a bias toward positive correlation, as opposed to negative correlation. This may be due to the biological nature of the data, as negative correlation, as a result of negative transcriptional regulation, is expected to be much rarer than positive correlation, as genes involved in the same biological processes more often co-operate rather than inhibit each other. However, it is not unreasonable to state that the normalization procedure has not yet been optimized for RNA-seq data and normalizing by total read counts has been reported to introduce biases [80]. The most commonly applied correction in the past few years calculates

FPKM/RPKM values, which correct for gene length. However, these have been extensively debated [80] and new metrics have been suggested [289], which have also been challenged [290]. Since none of these normalization protocols have been proven to be perfect, we opted to normalize samples by the total expression of all genes (reads that do not map to genes are excluded as these are more likely to introduce biases), until one of these metrics becomes generally accepted, at which point we will reconstruct the co-expression maps. We are, however, confident that these biases minimally affect the effectiveness of our tool, since our aforementioned validation tests have proven consistent with the existing literature.

3.4. Concluding remarks

Over the past century, research has led to a better understanding of many diseases and biological processes. However, the underlying mechanisms often remain unclear. In research there is a tendency to focus on genes that have already been studied to a broader extent and ignore poorly annotated genes. Yet, it is reasonable to assume that some of the unstudied genes play a crucial role and that without studying them, we might never be able to fully understand the mechanisms that underlie these diseases and processes. GeneFriends allows researchers to quickly identify poorly annotated genes that are associated with genes that have readily been associated with the disease/process under study. This unveils new venues for research and helps uncover new findings, as shown, for example, in [20]. This is particularly interesting since GeneFriends also allows association of ncRNAs, such as miRNAs and lncRNAs. These RNAs have been indicated to play crucial regulatory roles in multiple studies [291-293]. Additionally, it is not uncommon that unannotated genes are detected as differentially expressed in a study, yet, since no knowledge is available, they tend to be ignored.

GeneFriends can help identify possible roles of these genes, which will help experimental design.

Since Next-Generation Sequencing (NGS) is an emerging technology, our proposed RNA-seq co-expression tool will be useful for a growing number of researchers to gather clues regarding the many poorly studied transcripts detected by this approach. Unstudied transcripts or genes differentially expressed in a given RNA-seq analysis can be input into GeneFriends to assess the functional enrichment of co-expressed genes, effectively assigning a putative role to the query transcript/gene and identifying possible interaction partners. Knowing the potential roles of these transcripts will allow the assessment of the most interesting transcripts among those differentially expressed in the process under study and generate a hypothesis that can be challenged. This addresses an unmet need for the bioscience community and will help drive post-genome science. GeneFriends is freely available from http://www.GeneFriends.org.

Chapter 4: Correlation of expression of miRNAs with their targets and Weighted Gene Co-expression Network Analysis (WGCNA) of aging rat brain and thymus data

One downside of co-expression analyses is that they often lead to long lists of genes being associated to a process or disease under study and it often remains unclear which genes are the most relevant factors in the phenomena under study. It is desirable to identify those factors that are of greatest relevance in the context of the studied process/disease. In this light, we aimed to use WGCNA to identify genes that are more central to specific networks, under the assumption that central genes play a more important role. We focused on well-connected transcription factors in co-expression networks created from brain-tissue-specific datasets. We used literature to validate their importance to the brain related diseases. Lastly, we investigated the connectivity of a transcription factor that has been shown to have a large regenerative power in thymus tissue [151] in a co-expression network created from thymus samples.

Additional to the connectivity analysis, we aimed to investigate if there is a negative correlation between miRNAs and their annotated targets. These analyses were conducted on data previously created in our lab [294].

4.1. Abstract

Co-expression analyses often lead to long lists of new gene-disease associations and it can be difficult to select the most promising genes for follow-up studies. We tested the importance of transcription factors that act as hub genes (genes that are well connected) in a tissue-specific co-expression analysis on rat brain data created in our lab. The fact that two of the

transcription factors that were hub genes had already been associated with brain diseases supports the notion that this approach helps highlight important genes. We conducted the same analysis on a set of rat thymus samples of different ages to investigate if a transcription factor with highly regenerative power when expressed in an aged thymus, *Foxn1*, would also be hub gene. This was not the case which indicates that, whilst transcription factors that are hub genes in a network can be important factors, not every transcription factor that is an important regulator of a network is necessarily among these hub genes.

In addition to this analysis, we also investigated, in the samples obtained from the rat brain, if annotated targets of miRNAs were down-regulated in expression if the miRNA expression increased and vice versa. We could not find a biologically significant correlation and deem co-expression analyses unfit to uncover miRNA targets by simply identifying genes that are down-regulated when the miRNA is upregulated.

4.2. Introduction

In Chapter 3 of this thesis, we found that it is possible to associate ncRNAs with biological functions using co-expression in the same manner this previously has been done for genes. However, some ncRNAs are known to suppress the expression of other genes. This is in particular true for miRNAs (miRNAs), which inhibit the expression of other genes by, among the other mechanisms described in the introduction of this thesis, binding mRNAs. This binding significantly reduces the efficiency of translation and by cleavage of the mRNA effectively leading to its degradation [295]. As such, we expect that the expression of targets of such miRNAs are negatively correlated with the expression of the miRNA itself. In the RNA-seq based co-expression map we created in Chapter 3, a large number of these miRNAs are not present. This can be explained by the fact that we used data that was created with a cDNA

library preparation involving a ribosomal depletion step. Ribosomal depletion removes smaller RNAs often including miRNAs. As a result, many miRNAs showed no expression in any of the samples and measurements for those that did not get fully removed, are likely biased. Since we were still curious if there indeed is a negative correlation between the expression of miRNAs and their targets, we used a dataset that used a separate isolation protocol to isolate the miRNAs and determine their expression.

Since this dataset contained 39 rat brain samples (unpublished results) from different ages, we were curious if we could utilize co-expression to identify modules that associate with brain aging using the WGCNA package [66]. This is an R package that allows users to construct a co-expression network and has several different functions to analyze this network. The standard WGCNA analysis pipeline suggested by the creator, includes the following steps:

1. Determine correlation between the samples. In this step a dendrogram is constructed indicating which samples resemble each other the most. Replicates are expected to cluster together and if different treatments are applied, it gives an indication of which treatments have the most similar effects, based on how close their branches are in the dendrogram. Additionally, this will allow the user to identify outliers, which are distant from all the other samples, which should be removed from the analysis before proceeding.

2. Additionally, these rats were raised with different diets and we aimed to identify aging associated modules that are affected by these diets. Lastly, we identified the genes that are central to the modules that behave differently at older ages when supplemented with lipoic acid as part of their dietary regime [296].

To establish the relevance of these genes and determine whether it is reasonable to expect that these genes are major players in the aging process, we aimed to elucidate the network behavior of *Foxn1*, a gene that has been shown to regenerate the thymus upon activation in old mice [151]. The expectation was that this gene was central to a module that associates with aging and has a high module membership to this module. We aimed to elucidate whether *Foxn1* is a hub gene and whether a WGCNA analysis, as described in the previous paragraph, will highlight this gene. This gene is one factor that most clearly has shown a role in regulating aging changes by its ability to regenerate an aged thymus, which shrinks significantly with age, in old mice, upon its activation [151].

4.3. Methods

We used 39 RNA-seq samples obtained from rat brain. We used the read counts as determined in [294] by Shona Wood:

"The RNA-seq results from the SOLiD system are output as color space fasta and quality files, files were mapped to the Ensembl release 71 rat reference genome (Rnor_5.0, March 2012 and rn4) using Bowtie (Langmead et al. 2009) and settings appropriate to SOLiD data. For each sample approximately 36 million reads were generated. On average for rn5, 23 million reads per sample were mapped to the reference genome (approximately 63% of reads generated were mapped). "

These samples were obtained from rats with different feeding regimes and measurements were taken at several time points. The different treatments are described by the original study [296], from which table 4.1 was obtained, describing the different regimes and their effects on life span.

Group number	n	Dietary group description	Median survival (95% CI) (days)	Mean survival (days)	S.E. (days)
1	102	Control animals fed *ad libitum* the CRM diet throughout life	926 (909-943)	854	22
2	75	Fed a restricted intake of the CRM diet from 2 months to maintain body weight at 55% age-matched control animals	1047 (930-1163)	1025	25
3	75	Animals fed *ad libitum* the CRM diet supplemented with R/S racemic mixture of α-lipoic acid from 2 months of age	900 (839-961)	858	27
4	24	DR fed the CRM diet until 12 months, then DR fed the α-lipoic acid supplemented diet	1125 (1078-1172)	1068	38
5	25	Ad libitum fed CRM diet, animals switched to DR feeding at 12 months	1031 (1007-1055)	1000	33
6	25	DR fed the CRM diet from 2 to 12 months, then switched to *ad libitum* feeding	975 (935-1015)	914	44
7	25	Ad libitum fed the CRM diet, animals switched to DR feeding at 6 months	1078 (1048-1108)	1021	45
8	25	DR fed the CRM diet 2-6 months, then switched to *ad libitum* feeding	928 (858-998)	909	28
9	25	Animals fed *ad libitum* α-lipoic acid supplemented diet 2-12 months, then switched to DR feeding the CRM diet (no α-lipoic acid supplementation after 12 months)	934 (874-994)	859	57
10	25	Animals fed *ad libitum* lipoic acid supplemented diet 2-6 months, then switched to DR feeding the CRM diet (no α-lipoic acid supplementation after 6 months)	1086 (1059-1113)	1021	51
11	25	DR fed the CRM diet from 2 until 12 months, then switched to *ad libitum* feeding the α-lipoic acid supplement CRM diet	1041 (895-1187)	1009	34
12	24	DR fed the CRM diet 2 until 6 months, then fed *ad libitum* the α-lipoic acid supplement CRM diet	996 (927-1065)	947	43

Table 4.1: Dietary groups with the associated median survival
Brain samples from dietary restricted rats [296] were used for RNA-seq analysis [294]. The resulting data was used in our WGCNA analysis, to test if we can find co-expression modules that have an expression pattern that correlates with age and/or any of these treatments.
CI – Confidence Interval
S.E. – Standard Error

4.3.1. MicroRNA target repression associations

We obtained a list of 5226 targets across 183 annotated miRNAs in miRBase [297]. We next identified how often these targets were negatively co-expressed with their respective miRNA. To do so, we selected the 1% genes with the strongest negative correlation with each miRNA and counted the number of targets that were among this list. We opted to use this vote counting approach since we were interested in the % of targets we could identify using this approach.

4.3.2. WGCNA analysis of rat brain data

Genes that contained no expression in at least 3 out of the 32 samples were removed from the analysis. Samples were clustered by similarity using the hierarchical clustering algorithm, as described in [298] (Figure 4.1). Next, we used this clustering algorithm to identify the modules and calculate the Pearson correlation between the eigengenes of these modules and different treatments. Eigengenes can be viewed as the hypothetical gene that best describes the behavior of the genes in the module, as also described in the introduction of this thesis in Section 1.8.1. This hypothetical gene describes the partial expression of each gene (albeit to a different extent for each gene). This eigengene is the vector that best describes the variation (in a linear fashion) of the genes within this module, therefore being a better descriptor than the average expression of the module genes, which does not capture this variation as well as the eigengene. Apart from the different groups defined by the different treatments, we added a group describing those samples that increase the longevity of the rats. The list of sample groups consists of the following treatment groups: 1. Caloric restriction, 2. *Ad libitum* switch to caloric restriction after 12 months, 3 Caloric restriction switch to *ad libitum* feeding with lipoic acid after 12 months. Lastly, we identified the transcription factors (genes annotated as transcription factors) that are most central and behave most similar to the modules that

behave differently within this group of longer lived rats. To do so, we selected the three transcription factors that behave most similar to the eigengenes for each of these modules. The longevity signature is defined by the expression of the rats that are significantly longer lived than the controls. For a detailed description of the feeding regimes and isolation protocols of the brain tissue, we refer to [296].

4.3.3. WGCNA analysis Rat thymus data

The samples were obtained from [299], retrieved from the SRA database [253] and consist of 32 samples, 8 for each time point at 2,6,21 and 104 days. For each time point there were 8 samples: 4 females and 4 males. We used STAR to map the reads from the 32 samples to the *Rattus norvegicus* genome (Ensembl annotation Rnor_5.0). On average, the samples contained over 20 million reads and 81% of the reads mapped to the genome. To determine whether the expression of *Foxn1* is decreasing with age, we used a one-way Analysis of Variance (ANOVA) test, showing there is a significant yet mild decrease in expression with age ($F(1,30) = 11.08$, $p = 0.002$) (Figure 4.4).

Next, we used WGCNA to do a hierarchical clustering on the samples. This was followed by a hierarchical clustering based on genes. The correlation of these modules with the different traits was determined next. Lastly, we determined which module *Foxn1* is part of and its connectivity and module membership for this module.

4.4. Results

4.4.1. MicroRNA-target repression is not clear from the co-expression network

We observed that 135 out of 5226 miRNA targets were negatively co-expressed with the corresponding miRNA (Cumulative binomial < 5.7e-22). While statistically significant, the percentage was much lower than we expected based on the fact that miRNAs are known to silence their targets (only 2.6% of the targets were negatively co-expressed with the miRNAs). Since we felt these results may be caused by the inclusion of miRNAs that might only show a very marginal differential expression, we decided to conduct the analysis using a different approach. We opted to conduct a differential expression analysis between all the possible combinations of the samples setting a number of criteria that would include only large differential expression of strongly expressed miRNAs. Doing so, we aimed to eliminate any biases caused by miRNAs with marginally altered expression and expected to observe clear negative correlations between targets and the corresponding miRNAs. We used a vote counting approach to determine how often there is a negative effect on the expression of the target when the miRNA is upregulated. We used a 2 fold cut off on the differential expression of the miRNA as a minimum, as well as a minimum expression level of 100 reads in at least one of either samples, for both the miRNA and the target gene expression level. We then identified the extent to which the targets are differentially expressed. It appears that the target is downregulated in 48% and upregulated 52% of the cases a miRNA is upregulated. We did not

use a minimum fold change for the target expression levels as mild changes in expression can be biologically relevant as well [300]. We observed that in more 97% of the cases the fold change of the targets was less than 2, with an average down regulation to 85% of the original expression, if the target gene was down regulated in expression. These results do not seem to indicate any significant increase in down regulation for annotated targets of miRNAs, reinforcing the notion that co-expression networks are not able to identify targets of miRNAs by simply inspecting the strongest negatively co-expressed genes. One observation that may be worth noting is the fact that several targets of different members of the let7 miRNA family consistently are downregulated in expression. However, since the downregulation remains mild and the number of observations is relatively small the significance of this result is debatable. As such, we feel it is not feasible to predict miRNA targets by identifying negatively co-expressed targets.

4.4.2. WGCNA analysis rat brain data

We used WGCNA to cluster the 39 rat brain samples based on correlation between them. These 39 samples consist of 13 sample groups, each containing 3 biological replicates. We expected replicates to cluster together most closely, but in multiple occasions this did not appear to be the case (Figure 4.1).

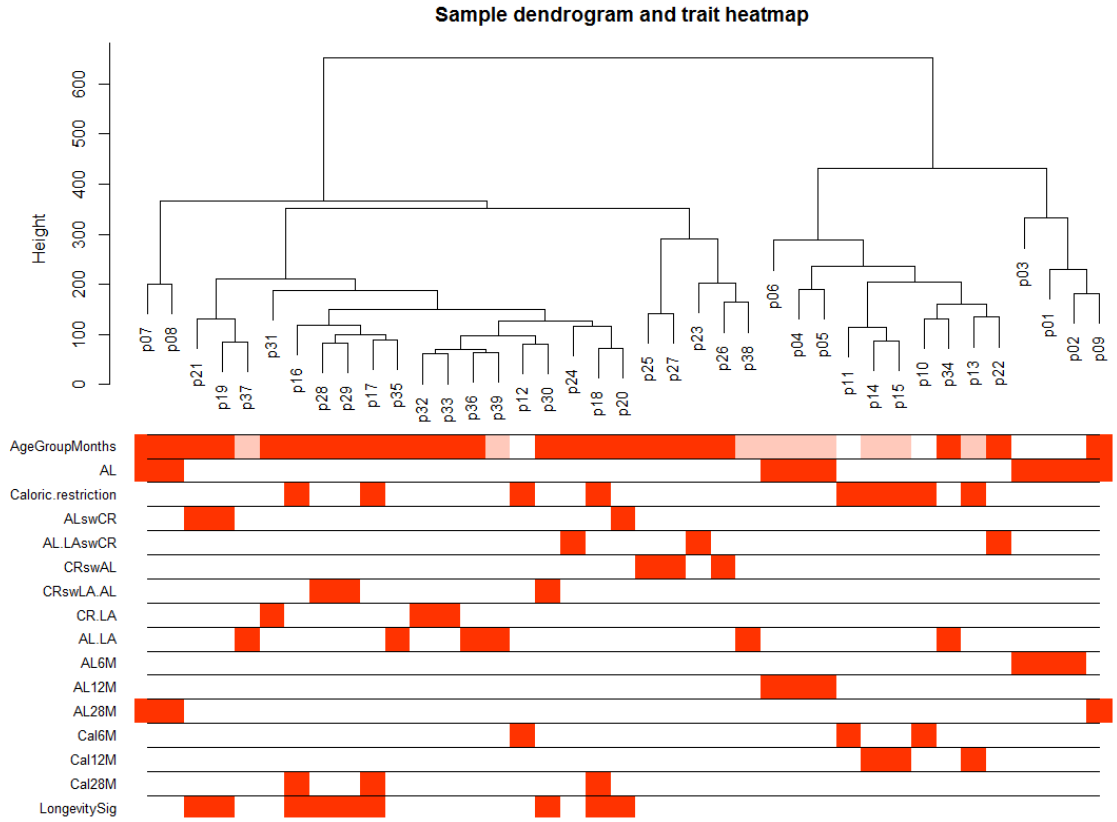

Figure 4.1: Hierarchical clustering of the rat brain samples
Each 3 consecutively numbered samples represent a group defined in table 4.1. For example, p1, p2 and p3 are 3 replicates in the same group and are expected to cluster together. Similarly p4, p5 and p6 are expected to cluster together and so on. Replicates do not always appear to cluster together indicating that individual differences may have a larger effect than certain treatments or that these treatments affect the different individuals differently.

We attribute this unexpected clustering to individual differences, leading to different effects of the treatment and aging in these individuals. Most rats that were fed lipids at any point clustered separately from caloric restricted and *ad libitum* fed rats. The number of genes differentially expressed with age in this tissue (cerebral cortex), was limited (between 8 and 180 in most of the relevant comparisons [47]).

4.4.3. Clustering of modules with traits

Next, we used WGCNA's hierarchical clustering algorithm to determine the different modules present in the co-expression network (Figure 4.2). We observe that several modules behave differently at 28 month old age in those rats that are long lived compared to the normal aging rats (Figure 4.3).

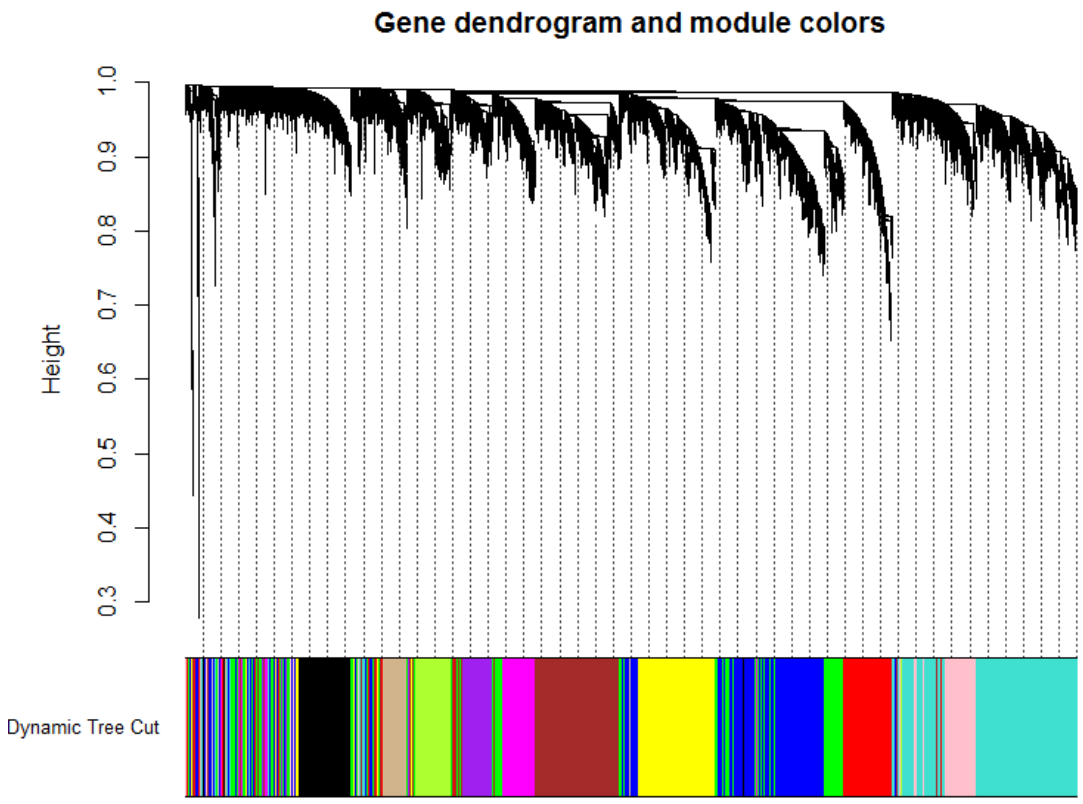

Figure 4.2: Gene clustering dendrogram based on gene expression in 32 rat brain aging samples

Clusters/modules are indicated by the different colors as determined by WGCNA's hierarchical clustering. Each leaf (vertical lines) represents a gene and the y-axis indicates how well the gene is connected to the rest of the genes in the module, as measured by the topological overlap [301], explained in [302]. Branches of the dendrogram represent densely interconnected, highly co-expressed genes. Each of these modules are summarized by the eigengene, a linear variable for which the correlation can be calculated with different traits as shown in figure 4.3.

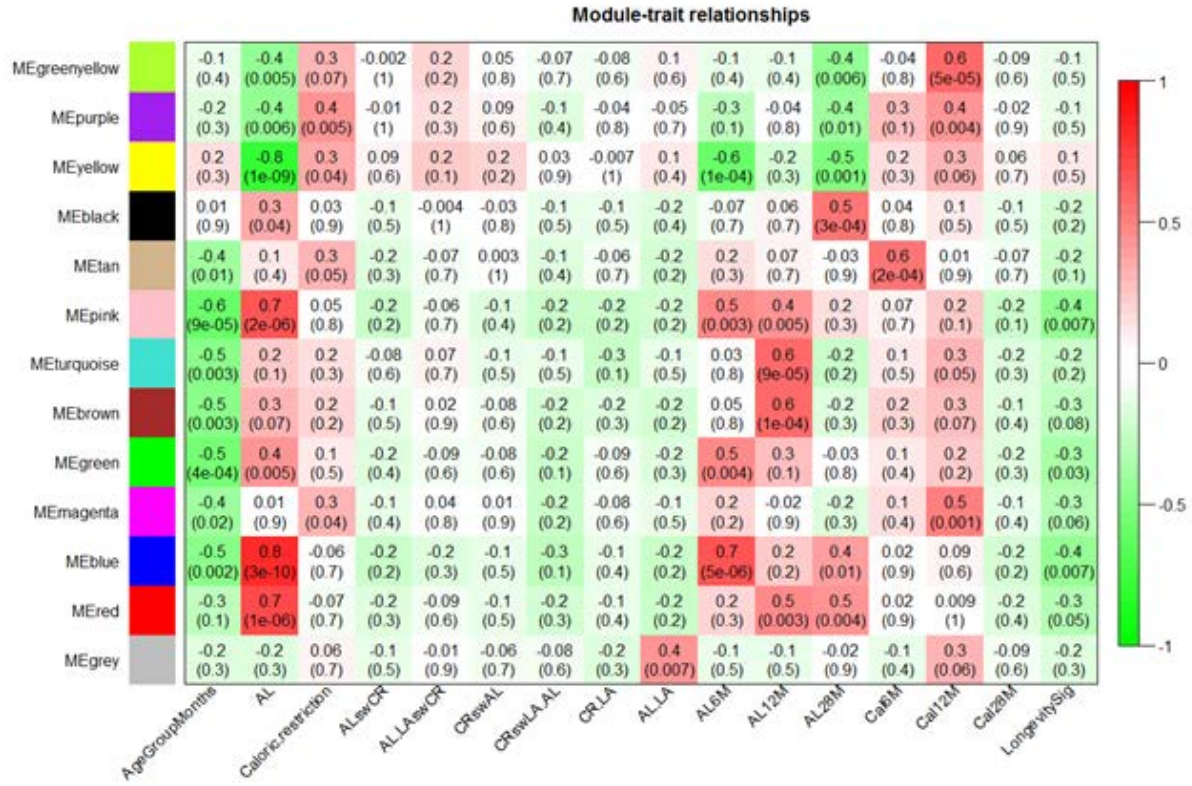

Figure 4.3: Correlation between modules and treatments

The y-axis represents the modules, as derived from the dendrogram depicted in figure 4.2. Treatments are indicated at the bottom. Each cell represents whether the activity of a module is correlated to a particular treatment. Interestingly the blue and red module are increasing in expression with age (AL6M,AL12M,AL28M), but have a lower expression in aged long-lived rats (LongevitySig) (indicated by the green color). Although some of the treatments contribute to the longer lived rats (AL.LA,CR.LA and CRswLA.AL), individually the treatments have no significant correlation with the different modules, suggesting the sample size is too small.

CR - Caloric restriction

Cal - Caloric restriction

AD - *Ad libitum*

LA - Lipoic acid

sw - switch to

M - months

This effect, however, is only observed when all long lived rat groups are combined suggesting that the sample size is to small to draw conclusive results for the individual treatment groups for these modules. Nontheless, it is interesting that the blue and red module have an opposing correlation in the long lived rats compared to the normally aged rats (Figure 4.3: longevitySig versus AL28M). This suggests that these modules are involved in the aging process. Next, we used DAVID [72] to identify functional enrichment within these modules (Table 4.2).

Module color	Functional enrichment	FDR	Enrichment score
Red	GO:0001654~eye development	0.11	3.2
Blue	GO:0043005~neuron projection	2.44E-06	6.2
Yellow	GO:0005739~mitochondrion	5.11E-49	28.9
Pink	GO:0019899~enzyme binding	2.37E-04	4.8

Table 4.2: Functional enrichment of clusters that are differentially expressed with age, but in an opposite manner to rats with extended life spawn through dietary intervention

We were interested in the biological process underlying the modules that appear to behave differently in normally aged rats versus those that are longer lived. We tested this using DAVID functional enrichment analysis [72]. We found the functional enrichment for the yellow module to be most significant for mitochondrion, which are known to play a crucial role in aging [303]. It could be interesting to further investigate the hub genes within this module.

We observe that the functional enrichment for the yellow module is very strong for mitochondrion. It is known that energy metabolism changes with age and that mitochondrial activity decreases with age [304]. Interestingly, we observe that this module is behaving differently in the long lives rats compared to the normally aged rats.

To further narrow down the genes that are interesting in this context, we identified the transcription factors that have a high module membership, indicating they are hub genes. Hubs in networks are known to be more important and this concept also applies to genes in a network [134, 305-307], albeit this only appears to apply to intra-modular hubs, as opposed to inter-modular hubs [98, 135, 136] (Figure 1.4). Transcription factors with a high module membership behave similar to the eigengene of the module. Transcription factors are known to regulate the expression of other genes. As such one would expect to find those that are responsible for the activation/deactivation of these modules to behave similar to the module and thus the eigengene of the module. For each module, we selected the three transcription factors that have the highest module membership and investigated whether they have been previously associated with aging or related pathologies. We found that, except for Nfyc, all the transcription factors with the most similar behavior to the eigengenes of the module were also the most connected transcription factors of the module. One of the transcription factors in the pink module, *Camta1*, has previously been associated with neuropsychological effects in older adults with cardiovascular disease [308]. Another transcription factor, *Atf4,* in the yellow module has been associated with neurodegeneration [309]. These and other transcription factors, *Atf2* [310, 311], *Trerf1* [312, 313], *Tfap2b* [314], have previously been associated with cancer. Furthermore, *Atf2* was previously also associated with osteoarthritis [315]. A large number of these genes have thus been associated with either aging related diseases and/or affect neuropsychology.

4.4.4. WGCNA analysis thymus data

The *Foxn1* gene only had a low expression, averaging 34 reads per sample, which is expected since transcription factors tend to have a lower expression level than other genes [316]. This gene is known to decrease in expression with age, although we only observed a mild decrease in this particular dataset (one-way ANOVA ($F(1,30) = 11$, $p = 0.002$)) (Figure 4.4).

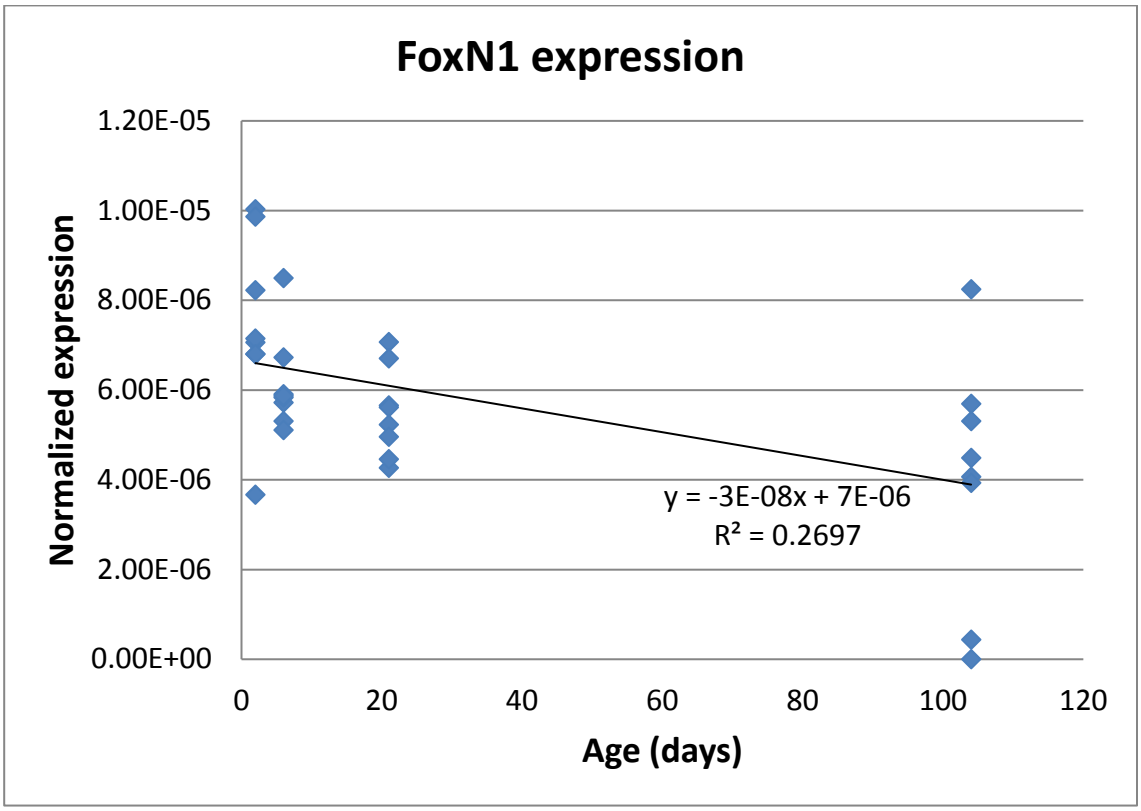

Figure 4.4. Expression of *Foxn1* at different time points
Foxn1 has been previously reported to decrease in expression with age [317]. In this dataset, we also observe this, although the decrease only mild (one-way ANOVA ($F(1,30) = 11$, $p = 0.002$)). It has been reported that the effects of this gene are extremely dose-sensitive [318], which supports the notion that the small observed changes in expression can have a significant impact on the observed phenotype.

It has been reported that the effects of this gene are extremely dose-sensitive [318], which supports the notion that the small observed changes in expression can have a significant impact on the phenotype . The clustering dendrogram indicates that samples largely tend to cluster by age group as expected (Figure 4.5).

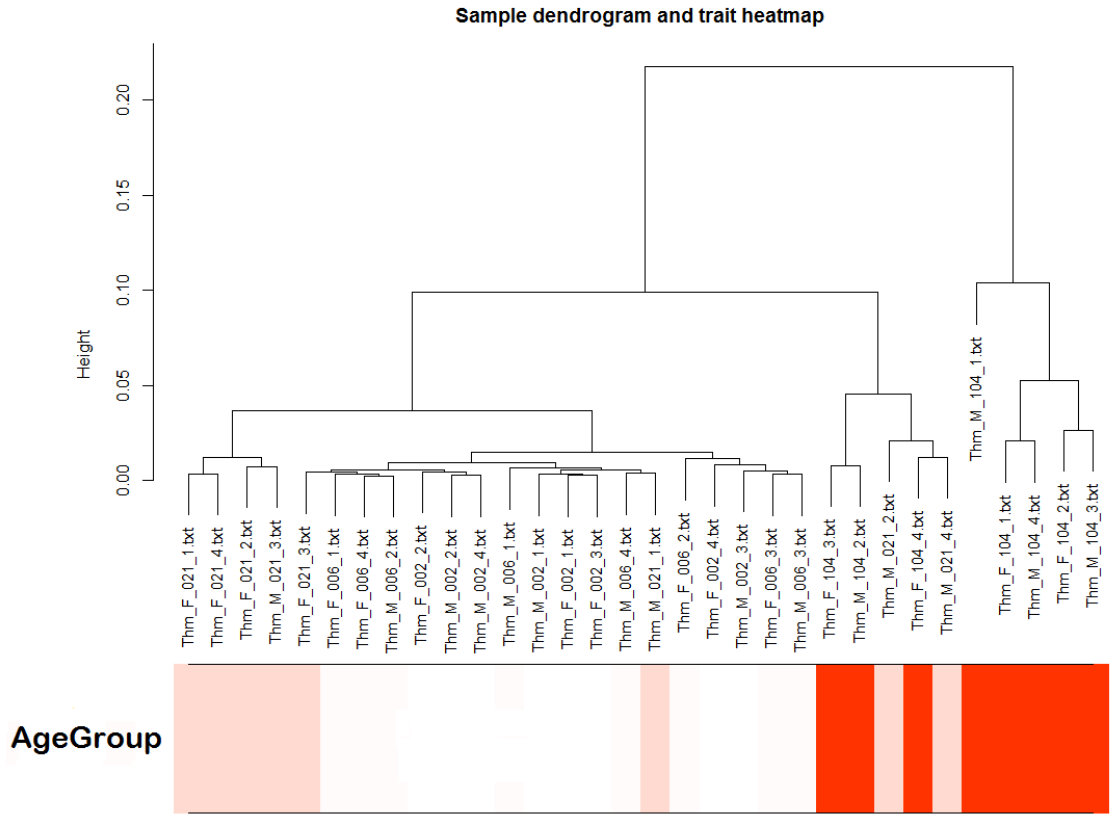

Figure 4.5: Hierarchical clustering of the samples using WGCNA
The first number in each identifier indicates the age of the individual rat (in days) and the second number is the replicate number for that age group. Colors at the bottom also indicate the age group to which the sample belongs. As expected, samples tend to cluster by age group.

Next, we used the same clustering algorithm on genes, rather than the samples. We used the dynamic tree cut algorithm to determine the clusters (Figure 4.6) and observed one very large module.

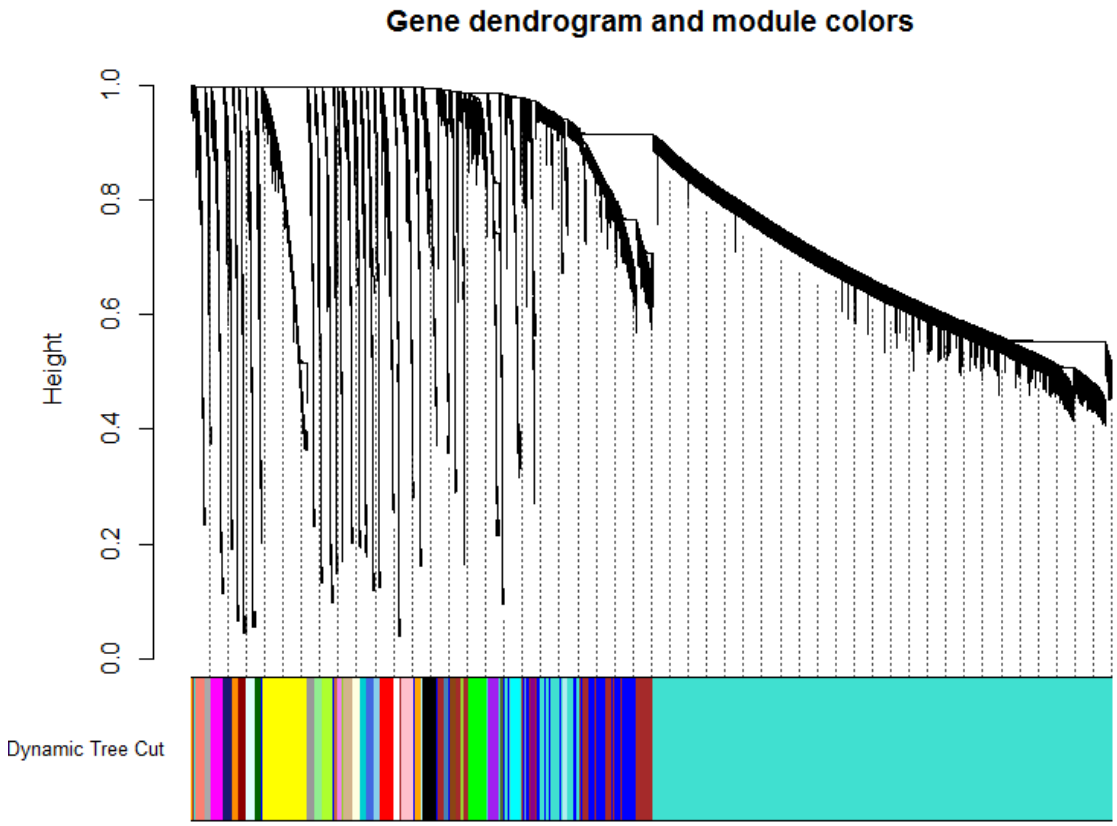

Figure 4.6: Cluster dendrogram indicating the different modules
Colors at the bottom indicate the different clusters, as determined by the dynamic tree cut algorithm. There is one very large module (turqoise), which is enriched for cell cycle and transcription processes. A possible explanation for this module may be a overal decreased gene expression of thymic cells, in particular epithelial cells, which have reduced proliferative capacity, involving many genes [319].

The expression pattern of this large module, as defined by the expression of the eigengene (eigenvalue of the first principal component), has a negative correlation with age (Figure 4.7). There also is a module showing a strong positive correlation with age. Genes in the turqoise module are enriched for the GO terms cell division and transcription (Bonferroni corrected p-value < 0.01), as determined by DAVID's functional enrichment analysis [72].

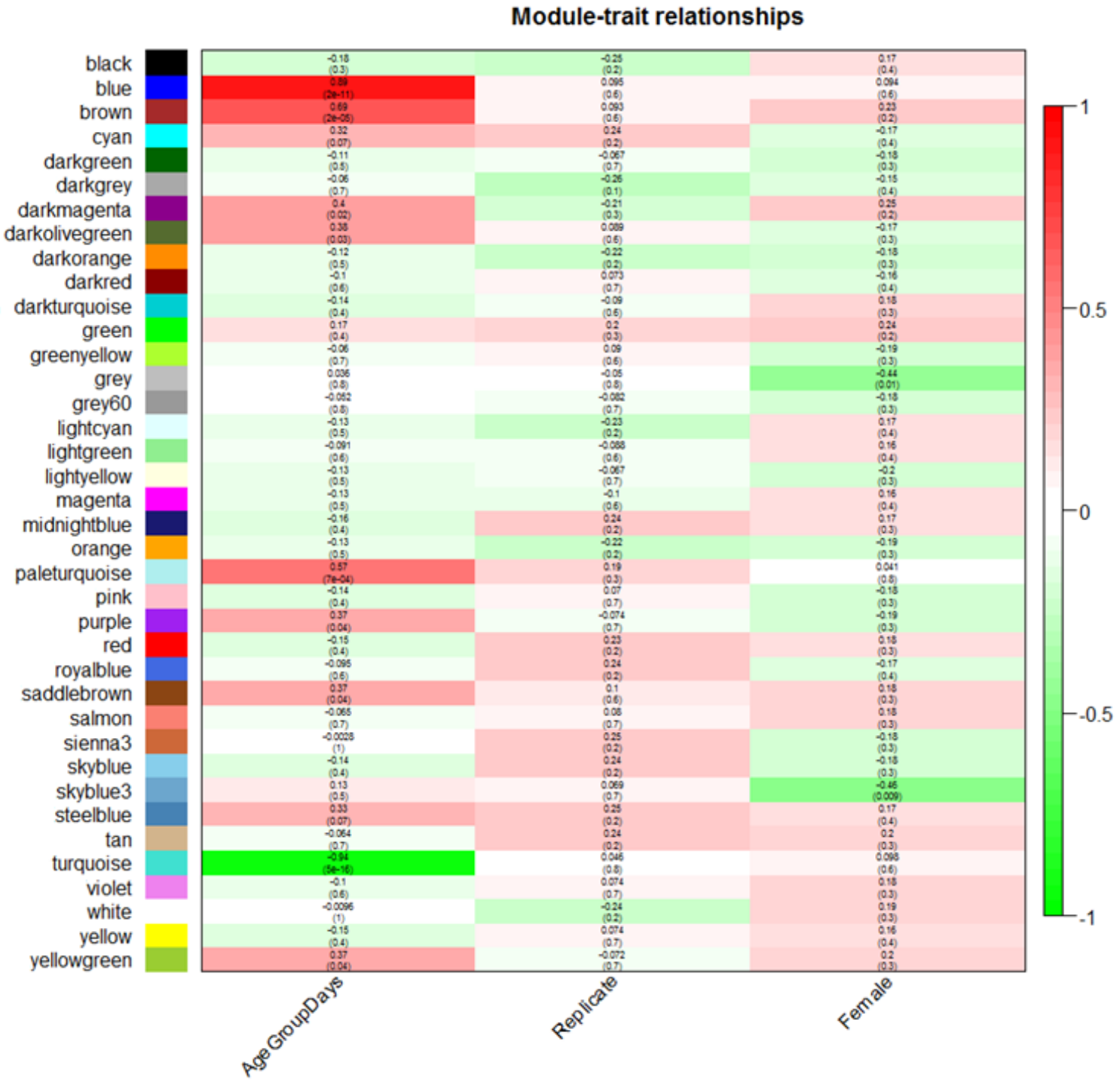

Figure 4.7: Correlation of modules with age, replicate number and sex
The upper value in each box is the correlation of the module (y-axis) with the trait (x-axis). The lower value (in brackets) represent the p-value. Several modules correlate with age and some with sex. To validate that this method does not randomly introduces correlations with uncorrelated variables, we included the correlation between the arbitrary replicate index number and the modules. As expected, there is no correlation between this arbitrary number and any of the modules.

The *Foxn1* gene is part of the turqoise module. This module has a negative correlation with age, which is in accordance with our expectations, since this gene is known to decrease in expression with age [318]. However, we expected this transcription factor to be one of the most strongly connected transcription factors in the module, but this wasnot the case. There are 144 transcription that are more strongly connected within this module and there are 132 other transcription factors with a higher module membership than the *Foxn1* transcription factor. This indicates that this type of analysis does not necessarily highlight all genes that play a large role in the aging process. However, this does not necessarily mean that those genes that do have a high connectivity and module membership are irrelevant. As such, it would be interesting to further investigate the genes that have a high module membership in this module, in other datasets. If these genes behave similar in other datasets, it increases the confidence that these genes are associated with aging and it will be interesting to follow up with experimental validation.

4.5. Discussion

4.5.1. MicroRNA target repression is not clear from the co-expression network

There are several possible explanations for the absence of miRNAs targets in the negatively co-expressed list of miRNAs. An explanation may be that miRNAs behave differently in different tissues. The miRNAs could be targeting different targets [320], such as different isoforms of the same gene that have different functions, due to the presence of different domains whilst maintaining the same miRNA binding domain. As such, the negative correlation in certain tissues may not be observed in brain specific tissue. Also, although speculative, it may be possible that miRNAs work as a negative feedback loop to assure expression of a gene cannot

spiral out of control [321]. In such a case no negative correlation between expression of the miRNA and their targets would be expected.

4.5.2. WGCNA analysis

Eigengenes are defined as the first principal component of a module [66]and by definition explain the largest amount of variation possible, which in our samples is expected to be the variation caused by the different ages. Additionally, we aimed to identify which genes are most central to these networks as central genes are more likely to be important [322].

The goal was to identify the genes that are most central to those networks that are differentially expressed with age, but also behave differently under different feeding regimes. The eigengenes can be used to identify if these modules are differentially expressed in particular samples. In our case, we used an approach to identify which modules behave different under the different treatments. We observed that several modules were significantly differently expressed under *ad libitum* feeding and that some of these modules behave in the opposite manner under caloric restriction. We aimed to identify modules that were behaving in a similar or opposing manner when treated with full caloric restriction as well as when swapped to or from a different diet. However, the results suggest that the sample size is too small to identify significant differential expression of these modules when inspecting groups that consist of only 6 rats (pooled into 3 samples before the RNA-seq step). However, if multiple different treatments are combined we do observe significant effects, which, we feel, is the result of the different lipoic acid and caloric restriction treatment patterns having a similar effect. Since the group size increases, whilst the observed effect on these modules appears to remain similar, the resulting correlation becomes significant. However, it remains impossible to conclude which treatment is most strongly influencing these changes and we can

thus only conclude that lipoic acid and or caloric restriction does appear to have an effect on the expression patterns of these modules.

It would be interesting to further study the *Camta1* and *Atf4* genes in the context of other species, to identify if they are also behaving differently under caloric restriction in mice. Since such experiments have been widely conducted, it should be possible to retrieve this data from RNA-seq databases and conduct such a study.

4.6. Conclusion

Our co-expression networks created from different tissues and conditions, is not able to identify targets of miRNAs based on co-expression partners, as evident from the fact that only 3% of the targets are among the 1% most negatively correlated expression partners.

Although our data suggest that lipoic acid treatment has an opposing effect on the expression of modules that are normally increasing expression at older ages, the sample size of the study appears to be too small to draw significant conclusions as to which treatments best mimic caloric restriction, if at all.

The WGCNA analysis we conducted to identify hub genes in modules differentially expressed with age, can help identify transcription factors relevant to the aging process. This conclusion is based on the observation that several of the most well connected transcription factors are associated with neurodegeneration or other aging related diseases. However, it does not guarantee to identify the most interesting transcription factor in the context of aging, as determined by our analysis of thymus data and the co-expression network behavior of the *Foxn1* gene. This gene is arguably the most crucial single factor to the aging process of the thymus, as evident by its able to regenerate the thymus to its younger state [151]. Although

present in a module of which the activity decreases during the aging process, it is not one of the most well connected hub genes in its module.

Chapter 5: Discussion

In this thesis, we set out to achieve a number of aims, as defined in Section 1.11. Here, we discuss to what extent we have accomplished our aims and put our results in perspective with reports on other co-expression analyses, such as co-expression databases created by other research groups as well as tissue-specific and differential co-expression analyses reported in the literature.

5.1. Co-expression databases

In biological research there is a bias toward well studies genes. Researchers tend to focus more on well annotated genes, which then become better annotated and more focused on. In this vicious circle, new potential targets for follow-up studies are less likely to be discovered. In part, this problem can be attributed to the fact that it is hard to study a gene for which no functional information is available. In Chapter 2 and 3, we describe our database, which can be queried through a web interface by other researchers to quickly identify biological functions to which a poorly annotated gene is associated. Additionally, this database allows for the identification of new genes that may be relevant to a disease or biological process under study.

The fact that the predictions from our co-expression network correspond to the annotation for well annotated genes (Section 2.3.2 and 3.3.3) supports the notion that our tool can be used to predict the biological process the gene plays its primary role in. That co-expression analysis is effective at associating genes with biological functions has been reported by others as well [4-6]. To put our database into perspective with other databases readily available, we have compiled a list of similar databases (Table 5.1). These databases allow users to obtain gene co-

expression partners of seed gene(s) and modules without having to go through the time consuming procedure of constructing a network and conducting clustering analyses. Our co-expression database is not the first, but is novel in the sense that it is the first constructed from RNA-seq data and includes the option to query for co-expression partners on a transcript level. With the emerging of RNA-seq technology, many new genes have been annotated, including many ncRNAs, most of which are not present in existing co-expression databases. For most of these no knowledge is available and co-expression analysis will help predict the biological process it plays its primary role in. We constructed a database and web interface, described in Chapter 2. This allowed us to acquire experience with the construction of a relatively large co-expression database. This was later expanded into a much larger version required for our RNA-seq based co-expression networks, as detailed in Chapter 3.

We aim to aid researchers in the interpretation of their results, for which purpose GeneFriends has already been reportedly used [153-157]. With the wider use of RNA-seq data for differential expression analysis, non-coding genes and alternatively spliced transcripts are more commonly observed as differentially expressed. The interpretation of these observations is hampered by the absence of information on potential functions for such genes and splice variants. Our database will help reduce the bias toward more well studied genes, as it allow researchers to include poorly annotated genes in the interpretation of these results. Additionally, it may aid in the design of validation experiments, by supplying the co-expression based functional predictions. These unidentified genes may be the missing pieces in the puzzle and could potentially serve as targets to, for example, cure disease.

Species	Database name	1. Integrated networks	2. Microarray Based	3. RNA-seq Based	4. Conserved co-expression	5. Tissue-specific	6. TFBS	7. Email address Required	8. Functional enrichment	9. Number of Samples	Citations
Hamster	CGCDB [323]		X			X				295	7
Human	HGCA [324]		X						X	2,000	4
Human	Transitional network [325]		X						X	1,000	1
Human Mouse	TS-CoExp [326]		X		X	X		X	X	7,500	23
Human Mouse	ImmuCo [327]		X			X				12,500	-
Human Mouse rat	dGCR [328]		X		X				X	200,000|5,000	-
Human Mouse Rat	Genenetwork [17, 329]		X	X	X				X	80,000	502
5 Species	GeneFriends [20, 50]		X	X					X	60,000|8,000	17
9 Species	GeneMANIA [330, 331]	X	X			X			X	175*	350
17 Species	GENEVESTIGATOR [245, 332]	X	X	X		X		X	X	130,000|?	2382
12 Species	COXPRESdb [6, 86]		X	X	X		X		X	157,000|10,000	195
11 Species	MaxLink [333, 334]	X	X		X				X	31*	38
10 Species	STARNET [246, 335]		X						X	13,000	67
> 20 Species	MEM [336]	X							X	Many	89
> 20 Species	STRING [261, 337]	X	X						X	Many	4819
A.T.[#]	CressExpress [338]		X			X		X	X	1800	94
A.T.[#]	CORNET [339]	X	X			X			X	3000	77
7 Species	PlaNet [340]	X	X						X	>1400	104
Grapevine	VTCdb [341]		X			X			X	500	6
Rice	RiceFREND [342]		X				X		X	800	14
Rice	RiceArrayDatabase [343]		X						X	1900	34
A.T.[#] Rice Brassica	RiceArrayNet [344]		X							1311	60

A.T.[#] Worm Human Mouse	FunctionalNet [345]	X	X						X	2200	188
7 Species	ATTED-II [346, 347]		X	X	X	X	X		X	12500	682
8 Plant species	PLANEX [348]	X	X						X	12000	9
8 Plant species	CoP [349]	X	X		X				X	10000	47
11 Species	BAR [350]		X			X				>406	477
4 Plant species	GeneCat		X		X					536	96

Table 5.1: Different databases and included features
We have compiled a list of available databases. Although this table includes many databases, more exist. Columns:
1. Combined with other networks (i.e. Protein-protein interaction)
2. Microarray based
3. RNA-seq based
4. Conserved co-expression
5. Tissue-specific
6. TFBS
7. Requires registration with an email address
8. Gene Ontology/Functional enrichment
9. Samples microarrays|RNA-seq (* refers to number of datasets rather than samples)

[#] Arabidopsis thaliana

5.2. RNA-seq co-expression networks

As outlined in this thesis, co-expression can be utilized to predict the biological processes a gene plays its primary role in. These co-expression based predictions have previously commonly been used to predict functions of coding genes, but to a far lesser extent for non-coding genes. In Chapter 3, we have created a tool that allows such co-expression based function predictions on a genome wide scale not only for coding genes, but also non-coding genes, as well as on a transcript level. The latter allows users to query different splice variants, which was not previously possible in any online web-tools and is the main novelty of our work.

The addition of ncRNAs to the database, in conjunction with the ability to query a set of genes allows users to input a list of genes that is annotated to a particular disease or a biological pathway to identify ncRNAs that are co-expressed with this set of genes. This approach can potentially be used to identify ncRNAs that play a role in a particular disease or pathway. Annotating genes as having a disease association may prompt researchers to investigate such genes in more detail, if they, for example, find it differentially expressed or mutated in a disease sample they study.

A benefit of the transcript specific co-expression network is that it allows isoforms with different co-expression partners to be identified relatively easy. This may be used to elucidate potential additional functions that genes may have beside the function to which they have previously been annotated. Genes that have transcripts that are co-expressed with different sets of transcripts may play different roles in different tissues. These different roles can be determined by the functional enrichment of each of the co-expressed gene sets. If a researcher were to find a different isoform expressed in a sample under study he or she could

query our database to test if this transcript likely has a different function from the canonical transcript.

5.3. Tissue specific genes and co-expression

One limitation of the co-expression network we build is that it is tissue-naïve. Some genes play different roles in different tissues. For these genes, co-expression analysis conducted on a wide range of tissues will associate these genes with multiple functions, but losing the information in which tissue it plays either of these roles. This can be deceptive as the most relevant process to the tissue of interest may not be ranked in the top of the enrichment analysis results. This could set a researcher on the wrong track, designing ineffective experiments. To solve this issue, it is possible to conduct a co-expression analysis simply using data originating from one tissue, similar to what we did in Chapter 4. However, this does reduce the number of datasets available and co-expression performs better on larger datasets, provided the same quality control measures are used [326]. Additionally, if more datasets are available, it allows for the luxury of higher quality control standards further improving the accuracy of co-expression analyses [326]. Since the number of publicly available datasets is growing exponentially, the possibility for tissue-specific conserved co-expression analyses will expand to include more tissues and species, increasing both the applicability and the accuracy of such specialized co-expression analyses. We are currently working on building tissue-specific co-expression networks, as well as including information about expression of different transcripts originating from the same gene. For this purpose, we are utilizing RNA-seq data, like in Chapter 3. Below we highlight available literature on tissue specific co-expression networks and note the benefits and drawbacks of such networks.

5.3.1. Whole organism versus tissue-specific co-expression maps

Some types of co-expression networks are more successful at associating genes to certain diseases and biological processes than others. Abnormalities like mental retardation or *Xeroderma Pigmentosum* have tissue-specific phenotypes, even though the mutation is present in the whole organism [351], suggesting that a tissue-specific network is disrupted and that processes active in all other tissues are unaffected. As such, the network module underlying this tissue-specific phenotype is expected to be apparent in a tissue-specific co-expression network, but not necessarily in co-expression network constructed from different types of tissues [249]. In case of a tissue-specific phenotype, it would thus be more appropriate to study tissue-specific co-expression to identify key player(s) in the disease under study.

Tissue-specific co-expression analysis has led to the experimentally validated association of *Mybl1* to spermatogenesis, as well as associations between genes and ataxia [249]. Other examples include the association between decreased expression of brain developmental genes in schizophrenia [13] and the identification of molecular networks underlying other complex phenotypes [136]. Tissue-specific co-expression identifies important genes in these tissue-specific diseases and is a better predictor of functional relatedness between genes [249], as well as it being crucial for identification of regulating genes that control tissue-specific co-expression modules.

Since the aim of our project was to construct a database that would be of use to a wide range of researchers, we opted to construct a co-expression network including data from different tissue and cell types. This allows researchers to query poorly annotated genes and obtain the predicted function in a tissue independent manner. We are currently working on constructing

tissue specific networks, which will allow users to identify the tissues in which different transcripts are expressed and what function they associate with in these tissues.

5.4. Conserved co-expression

Whilst we show that a number of the co-expression derived predictions are in agreement with known annotations for a number of coding and non-coding genes, there are a number of biases that exist, which are discussed in this section. Additionally, we highlight co-expression works focussed on conserved co-expression networks and what the benefits and downsides are of such networks.

5.4.1. Guilt-by-association caveats

The expression of genes that do not have a biologically relevant function, can occur as a result of coincidental co-location of these genes on a chromosome. Because they are co-expressed they are more likely to share *cis*-regulatory DNA motifs, increasing the chances they are co-expressed with a particular module even though they play no role in it. This may lead to the incorrect association of these non-functional genes to a process they play no role in. This is important to consider when using our database and a solution is supplied in the next section, discussing conserved co-expression networks.

When using a GBA approach, it is important to remember that not every gene in a co-expression module necessarily is associated with a function or disease association for which its partner's annotations are enriched. Since co-expression modules often consist of a large number of genes, any overrepresentation of a functional process or group of disease-associated genes quickly becomes statistically significant, as often indicated by deceivingly low *p*-values. Misinterpretation of these low p-values may lead to the incorrect conclusion that all genes in a module play an important part in a particular process or disease. In reality, the

fraction of genes in a module that relate to its main biological function is often under 20% [352], and module-trait correlations can be relatively low ($R < 0.5$) even when statistically significant [353]. To tackle this issue we have therefore also calculated the AUC, which is described in the second paragraph of Section 1.5.1.

Although the GBA approach has been shown to perform well, issues have been raised. In particular biases or deficiencies in gene annotation, like the widely used gene ontologies [354], will be reflected in GBA. Three issues are of particular interest to network analysis:

1. Circular reasoning: Many genes are annotated via associations based on networks. As a result, when doing a GBA analysis, the functional enrichment is strong since there may be, for example, 10 genes annotated to the same category in the network. However, if these 10 genes were annotated based on the fact that they associate with a single gene for which the function has been validated, the functional enrichment effectively originates from only 1 gene in the network. This gene itself may even be an outlier in the network or not present at all [352].

2. Broadly annotated genes: There is a large number of genes that are annotated to a broad range of functions [322], rather than the functions they play their primary role in. As a result, the GBA approach can lead to associations to a large list of functions caused by a number of co-expressed genes associating to many categories. Then the enrichment following from genes assigned to single categories may be underrepresented in the flood of enriched categories identified through genes annotated to many categories. One solution is to weigh gene interactions in functional enrichment analysis by the number of categories they are assigned to [322, 355]. In this thesis, we have not added such weighting to our functional enrichment analysis. Motivation for this choice is that we wish to assign a putative function to every gene even if the assigned functions are very general (genes that are assigned to detailed/specific

terms are usually also annotated to more general) and the more detailed categories do not rank on top. A downside to this method is that the more detailed function a gene is associated with may not rank in the top, but we leave this up to the user's interpretation. Even though these issues exist, in the context of co-expression, GBA still proves effective at correctly associating genes to functional categories and diseases as evident by several experimentally validated results [8, 13, 14, 20, 249].

3. Some genes may be incorrectly annotated to a category. As a result, the GBA approach we use incorrectly associates the gene of interest to a particular biological process. False predictions caused by incorrect annotation of genes will, to an extent, be compensated for by the large number of gene co-expression associations making it more likely to identify functional enrichment for the correct annotation (unless the majority of the hundreds of co-expressed genes is incorrectly annotated, which is unlikely). The issue of incorrect gene function predictions through GBA, due to incorrect annotation of genes, will be bigger in poorly annotated species, where more associations are based on solely computational predictions which are more likely to be inaccurate [352]. Conserved co-expression across related species may help separate true associations from those resulting from annotation biases and identify those genes that likely have a different role in particular species.

5.4.2. Conserved co-expression and species specific differences

One approach to uncover if co-expression likely occurs as a result of functional relatedness, is to test if co-expression with genes annotated to the same biological function is conserved in different species. If the gene is not co-expressed with other genes in a different species, it is likely not functioning in the same biological process. Genes whose function is unrelated to their neighbors are more likely to translocate [16] throughout evolution and are thus less likely

to be co-expressed in multiple species [356]. Identifying conserved co-expression partners to determine functional co-expression is particularly useful for ncRNAs, which tend to be less conserved [357]. Supporting the notion that genes that are co-expressed in multiple species are indeed more likely to be functionally related, is the fact that they are more likely to be physical interaction partners [358]. Uncovering genes that are co-expressed with genes annotated to the same functional category in multiple species can thus reveal functionally related genes with a higher confidence.

One landmark paper in the co-expression field showed that conserved networks are indeed more effective at assigning putative functions to genes than single species networks [4]. Many other papers followed, allowing for a more accurate functional classification of genes [24], as well as associations to diseases [19], in particular brain related diseases [326, 359]. Not only does conserved co-expression analysis help identify disease related genes much more accurately [19], it also identifies only those targets that have a similar co-expression network in both species, making them more likely to function similarly in, for instance, a given model organism and humans. This is particularly interesting in the context of drug target discovery for which extensive and expensive trials are required and the vast majority of these trials report negative results after proving successful in model organisms [360]. Conserved co-expression analysis could help select those targets that are most likely key players in both human and model organisms, effectively reducing the cost and time requirement of experiments and trials by reducing the number of compounds that require testing.

One drawback of conserved co-expression analysis is that those genes that do not exist in both species are automatically excluded from the analysis. For example, if one were to compare conserved co-expression in both human and yeast the vast majority of genes would be

excluded since the yeast genome contains only approximately 6,000 genes compared to the approximate 20,000 protein encoding genes in humans. Furthermore, it has been reported that the number of conserved co-expressed genes between mouse and human is less than 30%, highlighting the limitation of this approach [361] [148, 356]. This low percentage is explained by the differences in regulatory programs in different species, which varies between different biological processes [148]. Moreover, the use of different tissue types from different species to construct the co-expression map may further contribute to this.

Lastly, even though functionally related genes are part of the same co-expression modules in different species, connections between genes of these modules often differ, implying that the regulatory networks are wired differently in different species [24, 356]. This is a likely a contributing factor for the large phenotypic differences observed in different species despite the fact that their gene sequences are largely the same [362].

Conserved co-expression is an avenue that could be explored using the co-expression networks that we have constructed. It is possible to test if a gene's co-expression partner's functional enrichment is similar in our mouse based co-expression network, which we added after this project was finished. Conservation in general is also of interest in context of different splice variants, which may emerge through mutations. If functional, these splice variants will be preserved, but those that are not are less likely to be preserved. Investigating the preservation of splice variants can aid the identification of those splice variants that are more likely to be functional and add or reduce the confidence in the co-expression based predictions based on only one species.

5.4.3. Conserved tissue-specific co-expression

As suggested earlier, the use of different tissue types may affect the accuracy of co-expression networks. As such, the combination of conserved co-expression with the use of tissue-specific data may result in the most accurate networks. By contrast, as this requires tissue-specific data from multiple species, the amount of data available and thereby the accuracy will be significantly lower, reducing the quality of the resulting co-expression network [326]. It appears that tissue-specific conserved co-expression networks identify disease gene relationships that are not found in multi-tissue conserved co-expression networks [326]. This was determined by Piro et al. by removing all edges between genes that are not related to diseases. The remaining edges were considered as disease gene relationships. When the conserved co-expression network was constructed from specific tissues, different disease gene associations surfaced compared to when the network was constructed from multiple tissues. Each of the 13 tissue-specific co-expression network had its own unique disease gene association. Combining these 13 lists of disease associations, obtained from the tissue-specific networks resulted, in 3 fold more disease genes associations as compared to those uncovered in the multi-tissue network [326]. Thus many disease gene associations are found in tissue-specific networks that are not uncovered in multi-tissue networks. This suggests tissue-specific conserved co-expression is a powerful approach to identify disease related genes.

Since there are not many works available in the literature that has compared tissue-specific conserved co-expression to non-tissue-specific conserved co-expression, its potential is not clear yet. This can be ascribed to the fact that this type of analysis requires tissue-specific expression data from multiple species, which is not always be available in large numbers. With the decreasing cost of genome-wide expression analysis and the increasing amount of data publicly available, this issue may resolve, allowing more conserved tissue-specific co-

expression analyses to be conducted. Our view is that, since both conserved and tissue-specific co-expression networks are more accurate than individual species co-expression networks, the most accurate network is most likely obtained from conserved tissue-specific co-expression networks, but at the cost of available samples. In Chapter 4 of this thesis, we have conducted a tissue specific analysis on aging rat brain data and identified 4 gene co-expression modules that are altering in expression with age. Gene annotations within one of these modules were enriched for mitochondrial processes, which have been reported to affect the rate of aging in numerous occasions [303, 363-365]. It would be interesting to test if this same module would be present in other species and if the hub genes within this module are similar. If this is the case it would support the notion that these genes are indeed playing an essential role in the aging process.

As opposed to a conserved co-expression analysis, identifying networks conserved in multiple species, it is also possible to conduct a differential co-expression analysis. This can be used to identify genes that are differentially co-expressed and thus have different co-expression partners between different sample groups, such as two samples derived from different tissues. These genes appear to play a regulatory role in the difference of the phenotype observed between two sample groups [139, 140, 142]. This can be also used to acquire more detailed information about which genes are more likely to be essential in a biological process or disease. For example, if a gene is differentially co-expressed between individuals with a disease and a group of healthy individuals, it may play a regulatory part in the processes leading to the disease. Differential co-expression analyses have been more widely conducted in the past few years and are discussed below. Further to differential co-expression analyses, advances in other genome wide technologies are allowing for integrative analyses, which are, in our opinion, the future of genomics and other omics research.

5.5. Differential co-expression analysis

In Section 4.4.3, we have conducted a differential co-expression analysis, identifying a number of modules that behave differently in long lived rats compared to rats with a normal lifespan and identified the genes that best represent these modules. We wondered if the hub genes we identified in Section 4.4.3 would potentially include some of the most interesting targets for intervention of the rat brain aging phenotype. Unfortunately no targets with great potential for intervention studies in brain tissue had been identified yet, to our knowledge. However, in another tissue it was found that *Foxn1* plays a major role in the aging process. If this gene is activated in an aged thymus, it regenerates this tissue and restores the expression profile to one much more similar to the one observed in a young thymus [151]. We used this information, in conjunction with a rat thymus aging dataset, to test if this TF would be part of an aging related module and if it would be a central gene in this module. This did not appear to be the case showing that this approach does not necessarily detect all of the targets with a great potential for intervention studies (Section 4.4.4).

It may be interesting to use a different method that identifies genes that are co-expressed with different modules in different sample groups, such as DiffCoEx [67]. This method identifies genes that are co-expressed with different modules in different groups, as described in Section 1.9. Such genes may be regulators of the different activity of these modules between the long lived and rats with a normal lifespan. This change is not detected by WGCNA as it represents a different form of differential co-expression; that of the whole co-expression module, rather than genes that change between modules. It first identifies co-expression modules across a number of sample groups and then determines the activity of these modules within each of these groups. If this activity varies between the sample groups, this module is said to be differentially co-expressed.

Differentially connected genes can play a regulatory part in the difference in the observed phenotype between two groups (Figure 1.5d) [139, 140, 142]. For example, one study compared co-expression in mutant cattle with increased muscle growth to co-expression in non-mutants, using a method similar to DiffCoEx. By identifying the most differentially expressed genes and TFs showing the highest differential connection to these genes [142] (Figure 1.5d), the TF containing the causal mutation (myostatin) was identified. Interestingly, the *Mstn* gene, which encodes this TF, hardly changed in expression itself, providing an example of how differential co-expression analysis can uncover biologically important findings that are not revealed by differential expression analysis alone.

The fact that the *Foxn1* gene was not identified as a key player indicates that this analysis can miss targets that are likely to be highly relevant for intervention studies. We showed that this gene is part of a module that is differentially co-expressed between long lived rats and those that have a normal lifespan, although it was not an important hub gene within this module. It may be interesting to use a different method that identifies genes which are co-expressed with different modules such as DiffCoEx [67]. It may be that *Foxn1* plays a role in a particular module in rats with a normal lifespan, but becomes part of a different co-expression module in long lived rats. This change is not detected by WGCNA as it represents a different form of differential co-expression; that of the whole co-expression module, rather than changes within the module.

5.6. Identification of genes associating with disease

Although it is interesting we can use co-expression to associate genes to functional categories, as previously shown in literature [108-113] and in Chapter 2 and 3 of this thesis, we also aimed to identify genes that play an important role in disease. Our results show that it is possible to

uncover genes that are likely to play a role in diseases, such as the *Bc055324* gene in cancer and the Cebp genes in aging. This is in-line with literature [197, 198], which shows that altering the expression of these Cebp genes indeed leads to an extension in lifespan in mice. The *Bc055324* gene experiments conducted in our lab show that this gene is important for the proliferation of HeLa cells (Section 2.3.6). A double knockout of this gene is not viable as determined by a double knockout of this gene in mice (personal communication Paul Potter at MRC Harwell). The fact that conditional knockout mice, initiating knockdown of this gene after birth, show no obvious phenotype after 2 years suggests that this gene is not vitally important for survival (unpublished results) past birth. Although it is not expressed under normal circumstances and proves not to be important for survival, this gene showed to be increased in expression in cancer, which may require this gene for its rapid proliferation, as supported by our knockout experiment in Hela cells [20]. As such, the *Bc055324 gene* would make an interesting target to study further in the context of cancer. Follow up studies are warranted and collaborators are currently investigating the localization of this protein to be able to better determine its role in the cell cycle.

5.7. Transcription factors and co-expression

In Section 2.3.4, we conducted a co-expression analysis and identified co-expressed co-expression partners with a set of aging related genes. It is interesting to identify transcription factors that regulate this set of genes. As TFs are known to activate the expression of a target set of genes, we expected that these are strongly co-expressed with their targets. As such, co-expression would seem like the perfect way to identify these targets, which has proven successful in *E. coli* and *S. cerevisiae* in a number of cases [90, 128, 366, 367]. Although regulatory mechanisms are more complex in higher organisms, co-expression has also been

used to identify regulators in these type of networks [90]. However, transcription factor activity is often controlled by factors other than expression, such as a wide range of post-translational modifications like phosphorylation [117], acetylation [118] and methylation [119] as well as ligand binding [120]. Also, the activation of a gene is often controlled by multiple transcription factors, meaning the activation of a single transcription factor is often not sufficient to increase the expression of its target [368]. Furthermore, there can be competitive binding of other proteins and transcription factors, not allowing a transcription factor to bind and promote transcription [369]. These mechanisms ultimately cause noise in the co-expression of targets with their transcription factors. As such, one would not expect co-expression to be very effective at assigning every target based on the transcription factor's co-expression pattern. Therefore, this approach is, by itself, likely unsuitable to unravel the regulatory network underlying most diseases and biological processes. It also implies that there may be other regulators of the genes that are co-expressed with the aging genes than those reported in table 2.3. The observation that multiple C/ebp genes are co-expressed with the aging related genes does remain interesting. Alteration in isoform expression of these genes has been shown to improve metabolic health in mice [370] and to have positive effects on lifespan (personal communication with Cor Kalkhoven)[197], supporting the notion these genes can be targeted to increase the lifespan of mice. This supports the notion that our co-expression database can indeed be used to acquire supporting evidence for genes that could be interesting targets for intervention studies.

5.8. MicroRNA-target expression correlation.

Most of the work discussed in this thesis has focused on the positive correlation between genes. However, negative correlation between genes can also be biologically meaningful as it

may indicate inhibitory functions of a gene on another gene's expression. This is particularly interesting in the context of miRNAs, which are known to suppress the expression of their targets as also mentioned in the introduction of this thesis (Section 1.6.1). One challenge with miRNAs is that they have a short length, which means they are often removed in the RNA purification step, often utilizing a ribosomal depletion protocol, which removes short RNAs. In Chapter 4, we have therefore used an in-house generated dataset in which the miRNAs were isolated using a protocol specifically tailored to this purpose [294]. Contrary to our expectation, we did not find a significant negative correlation between miRNAs and their annotated targets. This observation supports the notion that the expression of miRNAs and their targets is not necessarily negatively correlated in brain tissue derived datasets and that identifying such correlation relationships will likely not help to identify the targets of miRNAs. We acknowledge that these results have limited meaning as they are based on a single dataset, but unfortunately at the time of this analysis, the availability of datasets with high quality measurements for both total RNA and miRNAs was limited. A different study did find both negative and positive correlation between miRNAs and their targets in human data, but found that neuronal specific miRNAs tend to be co-expressed, rather than negatively correlate, with their targets. Authors suggest that these miRNAs play a role in neuronal homeostasis [321].

The poor annotation of datasets in public databases increases the difficulty of acquiring such datasets. Once a larger number of datasets, including accurate miRNA expression data, become available, it may be worthwhile to conduct a co-expression analysis on such a dataset to more reliably assess the expression correlation patterns between miRNAs and their targets. Although the power of this study was limited, this finding adds support to the notion that

miRNAs and their targets do not necessarily show a negative expression correlation, which is an observation that we feel is valuable to the research community.

5.9. Integrated network analysis

In this thesis, we have focused on co-expression analysis mostly, but we feel a lot of opportunity to better understand mechanisms behind regulation of biological processes and diseases resides in combining co-expression networks with other the types of networks, some of which we highlight below.

Experimental validation often focuses on single genes. As these experiments are costly and time consuming, high confidence predictions of causal genes are of great importance. An analysis based solely on co-expression does not (yet) provide this level of confidence. Incorporation of information from other types of data can help prioritize which genes may underlie a phenotype. This can be achieved, for example, using information describing which genes are TFs, such as for regulatory predictions using GENIE3 [371]. However, a focus on TFs is rarely sufficient, and integration of multiple data types is often required to increase the accuracy and usefulness of the resulting networks [372, 373].

5.9.1. Transcription factor binding site analysis.

Genome-wide TFBS analysis was introduced in the beginning of this millennium using chromatin immunoprecipitation followed by microarray analysis, also known as Chromatin Immunoprecipitation (ChIP)-chip [374], which was later replaced by more accurate ChIP-seq [375], in which the microarray analysis step is replace by RNA-sequencing. This data was used to create a genome-wide integrated regulatory network from gene expression and TFBS data [376]. Combined analysis of ChIP-chip-based TFBSs and expression data initially showed that, in 58% of the cases, the TFs bound to the promoter region of the gene were indeed regulated

by the corresponding TF [377]. A partial least squares approach (a well-known tool for analysis of high-dimensional data with many continuous response variables) was later proposed to identify false positives and distinguish the activation and repression activities of TFs [378]. A more recent method harnesses the rapidly increasing availability of ChIP-seq data in combination with expression data to rank the genes bound by a TF, which can be used to prioritize the most likely TF targets [379]. Tools to conduct similar analyses, integrating expression and ChIP data, have also been published [380].

Multilayer integrated networks. Independent from the approach used to identify them, network modules can be further investigated for shared eQTL gene targets, TF/miRNA targets or enriched binding motifs [148, 381]. Several computational methods and publicly available datasets are available for multi-omics data integration. For example, information about eQTLs can be acquired from recent a large-scale blood-based *trans*-eQTL meta-analysis [382] or eQTL studies conducted in other tissue types [383]. TFBSs can be collected from databases, such as JASPAR and DeepBind [384], which consist of transcription factor binding motifs inferred from experimental data. Binding sites can be further prioritized by investigating tissue-specific ChIP-seq peaks from ENCODE [148]. Finally, miRNA-target interactions can be identified using several *in silico* target prediction tools [385, 386] or utilizing manually-curated databases of experimentally supported target interactions [297, 387, 388].

Combining information from different layers of data may lead to new biologically interpretable associations in a number of ways. If intra-modular hub genes are TFs or targets of a TF, this TF is more likely to have a causal role in the phenotype under study [142]. If multiple Genome Wide Association Study (GWAS) hits exist in the same module, their cumulative presence can explain disease development [381, 389, 390]. Differential methylation states of genes within a

co-expression module can elucidate methylation patterns underlying disease [391]. Finally, if multiple genes are regulated by the same genetic variant (under a *trans*-eQTL effect), it may be possible to identify the gene responsible for the alterations of the network by identifying the *cis*-eQTL gene driving the *trans*-eQTL effects (Figure 4). This is supported by the fact that genes under *trans*-regulation of disease-associated genomic variants are sometimes annotated to the processes or pathways associated with the corresponding disease. Good examples of this are IFN-α and complement pathways in which several genes were under *trans*-regulation of a systemic lupus erythematosus-associated variant, possibly via *cis*-regulation of *IKZF1* [382]. The integration of regulatory genetic variant information into co-expression network analysis with *cis*-eQTLs used as causal anchors, identified *TYROBP* as the most likely causal factor in late onset Alzheimer disease patients, a finding supported by the observation that mutations in this gene are known to cause Nasu-Hakola disease [353].

Overall, integration of multiple data types increases the accuracy of the resulting predictions [372, 373]. For example, modules unique to different subtypes of cancer were identified by integrating tumor genome sequences with gene networks [392], and these modules may be useful for prognosis and identification of putative targets for personalized medicine-based treatments. A recently published tool, CoRegNet, allows the integration of different types of data in a co-expression analysis by identifying co-operative regulators of genes from different data types [393]. Another established approach, cMonkey, achieves similar data integration by calculating the joint bicluster membership probability from different data types by identifying groups of genes that group together in multiple data types [394].

To systematically asses the performance of different tools and methods, projects such as the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges, specifically

DREAM4 and DREAM5 [395], have been invaluable. These challenged researchers to construct regulatory networks from simulated and *in vivo* benchmark datasets. However, these challenges were last posed in 2010 and many new methods and tools have been developed since.

5.10. Future prospects

In recent years, differential co-expression analyses have been increasingly used to analyze large datasets. This may be attributed to decreases in the cost of large-scale gene expression profiling, in particular RNA-seq, leading to increased sample sizes, and to the greater availability of tissue-specific data from perturbation experiments, which are required for fruitful differential co-expression analyses [396, 397].

Large-scale single-cell sequencing technology is increasingly used and the first co-expression studies have uncovered cell-type-specific co-expression modules that would have gone undetected in multi-cell-type co-expression analyses [144, 398]. Since the latter represent the aggregated signals of multiple cell types, they usually cannot detect alterations in cell subpopulations between different experimental groups. This is supported by the observation that the expression of cell cycle genes associated with aging decreased in the analysis of non-cell-type-specific data [399]. However, data from single-cell experiments indicated that this observation was caused by a decreased proportion of the G1/S cells that highly express these genes rather than changing expression within the cells [400].

An additional prospect is the detection of mutations from RNA-seq data [401]. As mutations accumulate with age in different cells, these can be used to identify the origin of the cell.

Mutation accumulation has been used to study cancer development and the origin of metastases [402]. In large-scale single-cell RNA-seq experiments, mutations could be used to separate cells based on their origin, or to group cells based on the mutations they harbour [403]. Cells harbouring the same mutations can be investigated for co-expression patterns, and modules unique to cells with a specific mutation may be detected. This may allow the direct linking of mutations to expression modules, with the limitation that only mutations in coding regions are detectable in RNA-seq data.

Although there are many exciting new possibilities with the increasing availability of single-cell RNA-seq data, many challenges still remain. With single-cell RNA-seq, typically a low number of reads per cell are sequenced and then the signal from multiple cells of the same type is aggregated to acquire a cell-type specific gene expression profile. It is hard to acquire sufficient data for rarer cell populations, such as stem cells, currently limiting specific analyses on these cell types when using these datasets. Additionally, the low number of reads per cell lead to sparse expression matrixes to which normalization methods currently used in canonical RNA-seq analyses are not attuned. These normalization methods often also assume the majority of genes do not alter in expression between different samples, which is not necessarily the case in single-cell RNA-seq, due to the heterogeneous expression across different cells. This is further exacerbated by the low quantity and difficulty of obtaining high quality RNA from single cells. These and other issues are further discussed in [404].

In addition to the normalization issues that exist in single-cell sequencing, the optimal method for normalizing bulk RNA-seq data is also still not clear. The widely used Fragment/Reads Per Kilobase Million (FPKM) normalization has been debated [80] and although alternatives have and are being created, each method has its limitations. Additionally, from our experience, the

use of different mapping tools can in some cases lead to very different results. Although some comparisons between different tools and methods have been made [405], a large-scale comparison, using e.g., public data, would identify such cases and define best practices for pursuing each research question.

With the increased availability of different data types, such as RNA-seq, genome sequences, ChIP-seq, methylome and proteome data, it will become possible to integrate these datasets to more accurately predict regulatory genes. Projects from large consortia, like GTEx [383], the Epigenome Roadmap [406], and ENCODE [148], are already generating data from multiple omics levels that facilitate these integrated analyses. To identify regulatory relationships, perturbation data is preferable, as canonical data cannot distinguish between true and false positive regulatory relationships [396, 397]. Furthermore, regulatory relationships can be highly cell-type, tissue- or developmental stage-specific [397]. Only a handful of tools and methods are currently available, mostly integrating only 2 layers of omics data [407]. Integrated network analyses come with additional mathematical challenges and best practices are far from established. Further research on this topic is of great interest to the research community, as it will allow for a better understanding of regulatory mechanisms explaining co-expression patterns and underlying disease, facilitating the identification of appropriate targets for intervention studies.

5.10.1. Prioritization of causal disease mutations with GeneFriends

Integration of different types of omics data is another interesting avenue we also wish to pursue with GeneFriends. One particular avenue we are interested in is to make GeneFriends more applicable in the clinical setting. It may potentially be interesting to use our co-expression database to prioritize potential disease causing genes by combining disease gene

predictions with information on potential pathogenic mutations harbored by a particular patient. Sequencing is emerging not only in research, but is also slowly starting to get used in the clinic. When a patient has a disease phenotype, in some cases exome sequencing or even whole genome sequencing is used to identify Single Nucleotide Polymorphisms (SNPs) that could possibly explain the phenotype. In most cases the causal mutation is not apparent as multiple risk SNPs are detected and identifying the causal one can be a challenging exercise. In most cases disease annotations have been determined for patients (for example, defined in Human Phenotype Ontology terms further explained in [408]), which have often also been associated with a number of genes. An obvious approach is to first identify mutations in those genes, but if these are not present it may help to prioritize other genes based on their co-expression with the disease term associated genes. These prioritized genes can then be investigated for pathogenic mutations, possibly explaining the observed phenotype in the patient. This prioritization may thereby help identify the causal mutation more rapidly. We could extend our tool to allow researchers to supply 1. Lists of genes containing mutations to our tool, from which we could identify the pathogenic mutations and 2. Phenotypes, which are often associated with lists of genes, from which we could derive a list of predicted disease genes using our already established GeneFriends tool. Identifying those predicted disease genes that also contain a pathogenic mutation, may elucidate potential causal disease genes in patients.

Chapter 6: **Summary**

In Section 1.11 of this thesis, we outlined a number of aims and here we reflect on these. We summarize the achievements described in this thesis and how it advances the state-of-the-art.

Our first aim was to construct a co-expression network from a large number of microarray samples, and to create a user friendly website that can be queried with individual or multiple genes, such as a group of genes previously associated to a disease. We have created a user friendly website that allows users to download full lists of co-expressed genes with their query, as well as the entire co-expression network per species. This also laid the ground work for our second aim. Additionally, in Chapter 2, we showed that our co-expression networks can be used to associate new genes to diseases using a GBA approach on known disease genes. These associations help identify interesting targets for follow up studies, such as the C/ebp transcription factors for aging and the *Bc055324* gene for cancer. GeneFriends has been used in multiple occasions for these purposes, supporting the notion that we have created a resource that is useful to other researchers [153-157].

The second goal achieved, described in Chapter 3, was the expansion of our co-expression database to include ncRNAs and add these to our existing website. Here, we have shown that co-expression analysis can also be used to predict gene functions for ncRNAs. We do note that there appears to be a relatively high abundance of predictions for the Olfactory Receptor category, which warrants further investigation. Accompanied with the web interface, this is the first co-expression database that allows users to query a large number of ncRNAs and obtain predictions on the biological process they play a role in.

Our third aim was to further expand our database to also include predictions for transcripts, which was also achieved, as described in Chapter 3. We have shown that this co-expression network allows identification of splice variants that are strongly co-expressed with different sets of transcripts that are enriched for different functions. The possibility to query the database for transcripts allows researchers to inspect if transcripts, from differentially spliced genes in their dataset of interest, likely have a different function. Identifying such transcripts will allow researchers to identify transcripts potentially explaining phenotypic differences, thereby allowing better interpretation of their results in the context of their study.

In Chapter 4, we tested if it is possible to identify miRNA targets by identifying genes with a negatively correlated expression pattern, the fourth aim of this thesis project. Contrary to our expectations, we showed that in rat brain dataset of rats with different ages miRNAs do not show a significant negative correlation with their targets. Although this is a study on a relatively small dataset, this supports the notion that miRNA targets cannot be identified solely from negative correlation patterns.

Our fifth aim was to identify modules that are altering in activity during aging, as well as identifying modules that counter this effect through dietary interventions. Conducting a tissue specific co-expression analysis on an aging rat brain dataset, we showed that it is possible to identify a module of genes that are altering in expression with age. We could identify a number of transcription factors previously annotated as players in neurodegeneration and other aging related diseases. However, we failed to identify modules that are correlated to specific dietary interventions, possibly due to the small sample size.

Our last aim was to test if hub genes are include all of the most interesting targets for intervention studies. We conducted our analysis, using WGCNA, on an aging thymus dataset to

test if this analysis would identify *Foxn1*, a gene showing massive regenerative potential in aged thymi, as an important gene in the aging module. This appeared not to be the case proving that hub gene selection of modules associated with a trait of interest does not necessarily uncover all promising gene targets.

Published Works

Wood SH, **van Dam S**, Craig T, Tacutu R, O'Toole A, Merry BJ, de Magalhães JP (2015) "Transcriptome analysis in calorie-restricted rats implicates epigenetic and post-translational mechanisms in neuroprotection and aging." *Genome Biology* 16:285.

Monaco G, **van Dam S**, Casal Novo Ribeiro JL, Larbi A, de Magalhães JP (2015) "A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels." *BMC Evolutionary Biology* 2015 15:259.

Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, **van Dam S**, Brawand D, Marques PI, Michalak P, Kang L, Bhak J, Yim HS, Grishin NV, Nielsen NH, Heide-Jørgensen MP, Oziolor EM, Matson CW, Church GM, Stuart GW, Patton JC, George JC, Suydam R, Larsen K, López-Otín C, O'Connell MJ, Bickham JW, Thomsen B, de Magalhães JP (2015) "Insights into the evolution of longevity from the bowhead whale genome." *Cell Reports* 10(1):112-22.

van Dam, S., T. Craig, and J.P. de Magalhaes (2015) "GeneFriends: a human RNA-seq-based gene and transcript co-expression database." *Nucleic Acids Research (Database issue)* D1124-32.

van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhães JP (2012) "GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases." *BMC Genomics* 13:535.

Plank M, Wuttke D, **van Dam S**, Clarke S, de Magalhães JP (2012) "A meta-analysis of caloric restriction gene expression profiles to infer common signatures and regulatory mechanisms." *Molecular BioSystems* 8:1339-1349.

Silva AS, Wood SH, **van Dam S**, Berres S, McArdle A, de Magalhães JP (2011) "Gathering insights on disease etiology from gene expression profiles of healthy tissues." *Bioinformatics* 27:3300-3305.

Posters

van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhães JP. “GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases.”

- PCR & Next-Gen Sequencing, San Francisco. 2014

- Genome Informatics, Cambridge. 2012

- BSRA Annual Conference, Birmingham. 2012

van Dam, S., T. Craig, and J.P. de Magalhaes. “GeneFriends: a human RNA-seq-based gene and transcript co-expression database.”

- Computational RNA Biology, Hinxton. 2015

- $[BC]^2$ Basel Computational Biology Conference, Basel. 2015

Presentations

van Dam, S., T. Craig, and J.P. de Magalhaes. GeneFriends: An RNA-seq based co-expression database.

- $[BC]^2$ Basel Computational Biology Conference, Basel. 2015

Appendix I - ReadCounter: A tool to determine the expression levels of genetic features based on reads mapped to a genome

In the Chapter 3 we constructed a co-expression map from over 4000 samples for which we downloaded the raw data. To convert these samples from sequencing data to read counts per gene it was necessary to count how many reads overlap each gene region. At the time of the construction of our RNA-seq based co-expression network, the fastest tool available to count this was HTSeq, which was too slow and became a bottleneck in the analysis. To solve this issue we aimed to create a tool that supplies the exact same output as HTSeq but in a much faster fashion. Additionally, we aimed to count the reads per exon and intron regions in addition to the entire exonic gene region as determined by HTSeq.

Abstract

With the growing use of RNA-seq to determine gene expression, it is important that the analysis of such data is supported by tools to analyze this type of data. Over the years the sequencing depth in RNA-seq has increased, resulting in a larger amount of data per sample. To make the analysis of such data easier we have written a small walkthrough of the analysis (www.GeneFriends.org/RNA-seqForDummies). Additionally, we wrote a script that will automatically convert raw sequencing data into read counts per gene. As a result of the increase in the number of samples per dataset and the amount of data per sample, analyses can be hampered by slow conversion of sequencing data to read counts per gene. For this reason it is important efficient analysis tools are written. We created a tool that determines the overlap between genomic co-ordinates of reads with genes ultimately allowing to determine the number of read counts per gene from mapped sequencing data. We wrote a tool that is 4 times faster than HTSeq on a single core. Additionally, contrary to HTSeq, our tool

allows for multithreaded processing of data allowing the same analysis taking 15 times less time on whilst producing identical results. Lastly, our tool reports the number of reads per exon and intron. This information can give indications about different transcripts that may be transcribed and if there is possible regulatory components, such as unknown miRNAs, being transcribed from within the introns.

Introduction

The ability to measure gene expression in parallel using microarrays has proven useful in linking genetic factors to specific biological processes [409, 410]. However, the exact mechanisms often remain unclear. This may be due to the exclusion of many ncRNA (ncRNA) and the lack of differentiation between the various transcripts arising from the same genes in microarray analyses. The role of ncRNAs is poorly understood and may be crucial in understanding these mechanisms. It is now possible to investigate the expression of these ncRNAs, thanks to the rapidly expanding sequencing technology [411, 412]. Additionally, RNA-seq technology has the benefit over microarray chips [413] that it allows the measurement of intron and exon specific expression. This allows us to differentiate between expression of different transcripts originating from the same gene [252].

To support the analysis of RNA-seq data a number of tools have been created, but even a simple gene expression analysis often proves challenging. Analyzing expression of non-commonly investigated features such as intron or gene flanking regions would require additional effort. As such, the additional advantages of exon and intron specific read counts are often completely ignored. Moreover, it is common for researchers to opt for microarray technology over RNA-seq technology to avoid of the complexity that comes with it. To address this issue we have created a freely downloadable script that automatically installs and runs the

necessary tools to convert data obtained from the RNA sequencing machine into read counts per gene/transcript (available from: http://www.GeneFriends.org/RNA-seqForDummies/). Additionally, we have created and included a tool, which we coined ReadCounter (available from: http://www.GeneFriends.org/ReadCounter/), This tool reports read counts for whole gene/transcript counts, intron and exon specific counts as well as ambiguous and non-ambiguous counts for every gene simultaneously, requiring no additional effort from the user.

The analyses of RNA-seq data typically requires several steps. The data obtained from the sequencing machine consists of basepair sequences of the genes that were expressed in the samples. These need to be mapped to a genome for which several tools are available such as Tophat [58], Burrows-Wheeler Aligner (BWA) [414] and STAR [55] and others [415], resulting in a .sam or .bam [416] file depending on the user's preference. These files describe the genomic location of the reads on the genome. In many cases researchers want to know how many reads overlap with genes. To determine this it is necessary to compare the read coordinates to the genomic locations of genes. For this purpose several tools are now available. Initially the only widely used tool available was HTSeq [59], but others have become available or have been added to existing tools, such as FeatureCounts [60], *IRanges* [417] and *GenomicRanges*[418]. Using the .sam/.bam file and a General Feature Format (GFF) or General Transcript Format (GTF) file (a file defining the genomic regions of genes/features) [419], these tools determine the number of reads that map to each gene/feature resulting in a file containing the counts/expression per gene. The .GFF or .GTF file contains predefined regions for each feature that can be defined by the user. It would be a tedious procedure to create these files, but fortunately these gene annotation files can be acquired from Reference Sequences (RefSeq) [420] or Ensembl [255]. Current counting tools use these files to count

reads mapping to specific genes or exons. Contrary to ReadCounter these tools do not include read counts for alternative regions such as introns or regions flanking a gene, which can have regulatory roles on gene expression.

After read counts per gene have been determined, normalization is often required to be able to compare the expression of one sample against the expression in other samples. For RNA-seq analysis the appropriate approach to normalize data has been heavily debated. The most widely used approach, calculating FPKM/RPKM values, has been strongly debated and alternatives have been suggested [289]. The use of The trimmed mean of M-values normalization method (TMM) [64] or Biological scaling normalization (BSN) [290] values have been suggested.

An additional issue that arises from the size of RNA-seq data is that the computational step of the analysis becomes more time consuming. In an effort to minimize the time spend, our script utilizes STAR [55] to map the reads, which has been shown to map up to 50 times as fast as Bowtie/Tophat [421, 422] whilst matching results of other tools more consistently [55] . Furthermore, ReadCounter utilizes multi-threaded technology leading to a 10 to 20 fold faster counting rate than HTSeq, whilst producing identical results. To identify the benefits and downsides of ReadCounter, we have compared the performance and results obtained from ReadCounter with those of HTSeq and featureCounts in this paper. ReadCounter is written in Java and is run using a UNIX command, does not require installation and is available from http://www.GeneFriends.org/ReadCounter/.

Methods

To assign reads to genes/features, ReadCounter requires (1) a GTF or GFF file [419], defining the gene exon regions and (2) a sam/bam file [416], containing the mapping information of the

reads. The reads in the sam/bam file are compared to the GFF file and assigned to the genes that contain overlapping exons. ReadCounter has a series of options that are similar to other tools to give researchers the freedom to adjust count settings to their preferences. These options can be found on http://www.GeneFriends.org/ReadCounter/about/. ReadCounter automatically accounts for pair end files as well as the need for sorting.

The output of ReadCounter consists of a number of files:

1. A file containing the same format as the result file that would be obtained from HTSeq, for the purpose of working with currently available downstream tools.

2. An extensive results file containing reads per gene as in the file described above, but additionally, the ambiguous counts per gene, intron counts per gene, ambiguous intron counts per gene. Furthermore, it contains the number of reads mapping to the regions flanking the gene (default = 10,000 base pairs before and after). And lastly, it contains the counts per exon and per intron as well as the ambiguous counts per intron and exon and number of different exons/introns these counts overlap with.

3. A file containing warnings generated by the tool.

4. A file defining all intron and exon regions for each gene obtained from the GTF file. To reduce ambiguity, if the GTF file defines different exons overlapping each other, they are redefined as 1 exon.

5. In case of unsorted paired-end input files, an extra folder is created containing a list of sorted sam lines, eliminating the requirement of re-sorting the file in future runs (using ReadCounter only).

At the end of the script the tool reports the percentage of reads mapping to genes, introns and exons for ambiguous (overlapping gene regions of multiple genes, present at a particular genomic region genes, equally) and non-ambiguously overlapping reads, as well as an estimate of the average fragment size.

ReadCounter follows the list of steps below to count the number of reads mapping to each gene, intron and exon.

Loading GTF/GFF File:

1. Read and reformat the GTF file removing any ambiguity and save the reformatted file for any future use.

2. Split the genome up into bins, where each bin represents a region of the genome. Its index is related to the region it contains. This can then be used to instantly identify the bin(s) that is relevant to any particular read.

3. Read the reformatted file and assign any genes overlapping the regions represented by these bins (Figure 1).

Reading Sam File & assigning reads

4. For each read or read pair, retrieve the genes from the relevant bin(s).

5. Compare for each read/pair the size of the overlap with all retrieved genes.

6. Assign the read to the gene it overlaps with the following priorities if multiple genes are overlapping:

- The gene/feature that overlaps with both sides of a read pair

- The gene/feature with the biggest overlap size

- Exons over introns (unless the exonsOverIntrons option is set to false)

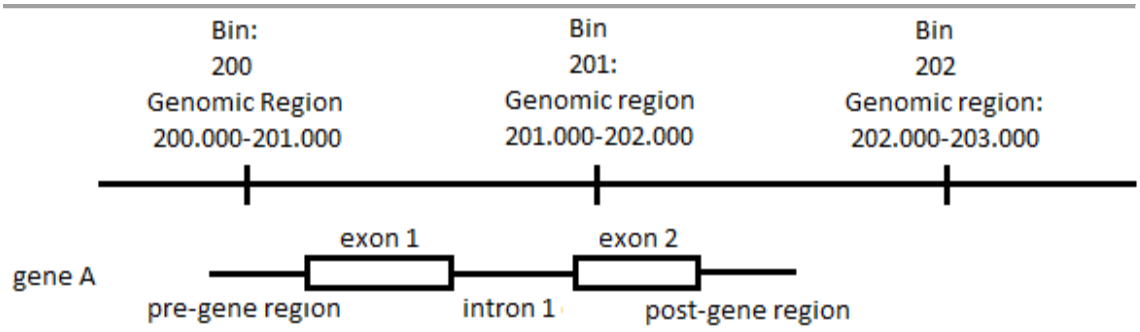

Figure A1: Graphical representation of the bin-system employed by ReadCounter
The following features are in the following bins:
Gene A: Bin 199,200,201,202
Pre-gene-regio: Bin 200,201
Exon 1, Intron 1: Bin 200
Exon 2 Bin: 200,201
Post-gene region: Bin 201

ReadCounter employs multi-threading technology to optimize speed, however at high core numbers hard drive or memory speed may become rate limiting. Additional to determining the gene the read overlaps, the specific exon is also identified. Furthermore, overlap with introns is also identified and reported.

Results

To test and validate our tool we compared it to 2 other tools: FeatureCounts and HTSeq. To be able to compare ReadCounter to HTSeq, ReadCounter has an option to disregard the overlap size when assigning reads to genes/features, as well as subtracting 1 basepair from each exon as HTSeq does not seem to consider the last basepair of each exon as inclusive. Also the "beSmart" option is disabled, which normally counts paired-end reads that map to both introns flanking an exon to the exon rather than the two flanking introns. This allows ReadCounter to obtain identical results to HTSeq, which we consider the current standard for counting reads.

Performance

We tested all three tools using an 83 gigabyte (GB) .sam file containing approximately 222 million read pairs. First we tested the time each tool required to count the reads using the same settings (Table A1). Contrary to HTSeq both ReadCounter and featureCounts have the ability to run multi-threaded, for which reason we compared the multi-threaded performances for both these tools. We think it is reasonable to assume most computers used for RNA-seq analyses have at least 4 and most of the times will have 8 cores and used this number for our tests (Table A1).

Since both ReadCounter and featureCounts can deal with unsorted files, we also tested this. To do so, we scrambled the lines in the initial .sam file and saved the results. This scrambled file was then used with each tool (Table A1).

	1 Core Sorted File	4 Cores Sorted File	8 Core Sorted File	4 Core Unsorted File
ReadCounter	55 Minutes	14 Minutes	10 minutes	29 minutes
FeatureCounts	9 Minutes	9 Minutes	9 Minutes	127 Minutes
HTSeq	161 minutes	-	-	-

Table A1: Runtime comparison between different tools using the same options on an 83 GB file containing 222 million read pairs

We tested our tool counting the number of reads that overlap each gene with each of the 3 tools on 2 samples, one that in which the reads have been sorted and one that is not sorted. HTSeq requires the reads to be sorted, in order to assess expression per gene and was therefore excluded from the comparison for the unsorted files.

ReadCounter unique output

As stated before, ReadCounter obtains the results for regions flanking a gene as well as intron specific read counts. Furthermore, it reports the number of reads mapping ambiguously. To investigate the relevance and occurrence of these additional results, we tested ReadCounter on 2393 mouse samples and identified the number of reads mapping to these regions on average (Table A2). ReadCounter has a series of options that are similar to other tools to allow researchers the freedom to adjust count settings to their likings. This includes a number of options on how and when reads are counted to introns and how ambiguous maps are determined. These options can be found on www.GeneFriends.org/ReadCounter/about/.

	Average read counts	Standard deviation	compared to exonic maps	Standard deviation
Exon	14.199.675	7.193.162	100%	-
Exon Ambiguous	1.254.605	658.260	9.63%*	5.89%
Intron	1.520.837	1.591.286	15.01%	13.33%
Flanking gene regions (10KB)	99777	80213	0.84%	1.24%

Table A2: Number of reads mapping to additional regions

To estimate the percentage of reads that would be ignored if all ambiguously overlapping reads would be ignored, we counted how many reads ambiguously overlapped multiple genes (9.63%). Similarly we counted the number of reads that mapped to introns, which could be biologically relevant (15.01%). Lastly the number of reads mapping in close proximity (1000 basepairs) to a gene (not overlapping other genes), representing reads mapping to regions such as promotors and enhancers in close proximity to the gene, which appeared relatively low at 0.84 %.

Samples with less than 1 million reads overlapping any intron/exon or with more uniquely overlapping reads to introns than exons were excluded (224 samples).

*Outliers, those with more ambiguous reads then specific reads, were excluded (5 samples).

Exon reads

On average 14714 (Standard deviation: 4033) features have more than 10 reads mapping to them. Of those features with more than 10 reads mapping to them on average 37.83 % of the exons are expressed (expression higher then 0) with a standard deviation of 12.54%, indicating that a large number of exons is not in expressed at all even if the gene is. On average 707 features only had ambiguous reads. And 1109 features had 10 times more ambiguously mapping reads then specific mapping reads.

In case reads are mapped to transcripts rather than gene identifiers (ID), the ambiguous proportion vastly increases with 299% (Standard deviation: 82%) more reads mapping ambiguously than uniquely.

Intronic reads

On average 3512 genes had 10 fold more reads mapping to introns then to exons. This may indicate that the expression of the introns is affecting the expression of the exons, since intronic regions can contain antisense and or miRNAs for proximal genes.

Flanking reads

One of the extra options that ReadCounter has is the option to count reads mapping to regions prior or post a gene. This option was added since we believe expression of this region can be biologically relevant. For 2 features the number of reads mapping on the flanking side were exceptionally high: *Linc00273* and *Mir3687* accounting for 73% of the reads mapping on the right flanking gene region indicating these regions may contain un-annotated features. Excluding these 2 features on average 0.84% of the reads overlapped each of these regions.

Differences due to considering overlap size and ambiguity

To validate the results obtained from our tool we added an option to disregard the overlap size of reads with genes/features. When we enable this option the results are identical to those of HTSeq. However, we were curious to the extent of the difference if we do consider the overlap size, meaning we map reads to the genes that have the biggest overlap with a preference for genes mapping both sides of a read pair. We found that for the dataset we used in the speed test, 1.2% more reads were mapped (approximately 1.7 million reads). These reads would have otherwise been considered mapping ambiguously and now map specifically to one gene. Even though this percentage is seemingly low, for 499 features it meant an at least 2 fold higher number of overlapping read counts including 143 features that would have otherwise had no expression at all (features with less than 10 overlapping reads were excluded in this calculation).

Exon specific counts

It is possible for 2 reads of the same pair to map to the introns flanking both sides of an exon. In this case the fragment is overlapping the exon, even though the reads are not. As a result, the current tools available would not count the fragment to the gene it is overlapping. ReadCounter correctly counts this fragment to the exon it is overlapping. On the particular dataset we tested the difference using this option. The difference in this particular dataset appeared to be negligible, which is possibly due to the limited fragment size (742 reads in this sample (out of a total of approximately 222 million reads)).

Discussion

As stated before, RNA-seq has the unique ability to measure expression levels of specific exons and introns [252]. Shifts in exon expression within one gene can be detected with this

information. This will indicate that a different transcript is being transcribed, which would go unnoticed if one would solely count whole gene expression. Intronic expression is relevant because reported intergenic regions can be transcribed [423] and have been indicated to play an important role in transcription regulation [423-425]. A reasonable number of reads solely overlaps intronic region and the relevance of these transcripts may be larger than their numbers suggest. It is not uncommon for miRNAs to be present in the intronic regions of a gene, which are transcribed and can produce mature miRNAs [426]. We expect gene flanking regions to have a similar relevance to that of intronic regions. Expression of these regions may influence the expression and translation levels of the transcripts obtained from the connected genes.

One common problem with counting reads is that a read can overlap multiple features located on the same region of the genome. This creates the problem that reads occasionally overlap multiple features, making it impossible to determine the origin of the read. Using default options, other tools will count the read toward none of the features. As a result, genes that are in a region that is fully overlapped by another gene, will be considered as never expressed. This problem significantly increases when counting reads toward transcripts rather than genes, as multiple transcripts originating from the same gene often utilize overlapping exons. This occurs at a very high rate and thus creates strong biases resulting in a false representation of the data. Most current tools have an option that allows ambiguously mapped reads to be counted to multiple genes, but in this case it remains unclear whether a gene is mapping to multiple genes or just one. ReadCounter solves this problem by reporting both ambiguous and non-ambiguous reads in separate columns. Under default options these reads are not counted toward any of the transcripts using regular tools, but with ReadCounter it is possible to determine the number of reads that are unique to that transcript as well as the number of

reads that is also part of other transcripts. Combined with the additional exon specific information this will give a clearer indication of the transcript created, as well as allowing for the possibility of identifying un-annotated transcripts. This allows researchers to reduce this bias, with minimal extra effort.

To further reduce the number of ambiguous reads, ReadCounter considers the overlap size of each read with all genes its overlaps. The read is then counted toward the largest overlapping gene, ultimately reducing the number of ambiguously mapping reads. Other tools merely consider whether there is an overlap and ignore the size of the overlap. One feature that is available when using RNA-seq is paired-end sequencing. This leads to two sequences from each end of the fragment being created. This pair can then be employed to more accurately assess the position/gene the read originated from. In case of paired-end reads, FeatureCounts prioritizes features that overlap both reads of the pair, but does not consider the sizes of the overlap. Like featureCounts, ReadCounter gives priority to features overlapping both sides of the pair even if the size of the overlap is smaller with another gene that only overlaps with one of the sides.

To accurately determine the overlapping features with paired-end reads it is necessary for both reads of the same pair to be evaluated simultaneously. However, in unsorted files, reads of the same pair cannot be accessed simultaneously efficiently. For this reason other tools require files to be sorted. ReadCounter employs the fact that each read contains information about its partner to minimize time required for sorting files. This makes ReadCounter analyze unsorted files at an unprecedented speed.

Another feature that is available with ReadCounter is the option to calculate Transcripts Per Million (TPM) values [289] rather than absolute read counts. This unit represents the relative

molar RNA concentration and is a normalized value that can be used for comparison against other samples. We added this option since we feel most users would prefer normalized read counts over absolute read counts. We opted for TPM values rather than FPKM/RPKM values, since the validity of these values has been debated and the use of TPM values is recommended instead [289].

Lastly, when we initially created ReadCounter there were no tools available that counted reads significantly faster than HTSeq. Since we required faster counting we aimed to create a faster tool, which we did by using a more efficient approach as well as employing multi-threaded technology. Meanwhile featureCounts was published which also utilizes this technology and is faster than ReadCounter albeit not supplying the additional intron and exon specific read count information obtained from ReadCounter. FeatureCounts is more efficient than ReadCounter, however if multiple cores are used the difference is negligible. We would like to note that featureCounts uses significantly less memory, using no more than 100 Megabyte at any time, where both HTSeq and ReadCounter exceed 1 GB. We do not expect this to be an issue since we anticipate computers assigned to RNA-seq analyses to have at least 8 GB of RAM memory (although 2 GB should be sufficient).

We believe this tool will help researchers utilize the additional information obtained from RNA-seq experiments. Especially once a particular gene of interest has been identified, the additional exon and intron information, could give crucial hints into the mechanisms underlying their functionality. However, since the extra information adds another layer of complexity to an already complicated procedure, to fully optimize the use of the intron, exon and gene flanking region as well as ambiguous reads information, we believe additional tools are required that supply researchers with a differential expression analysis employing this

luxurious information. Ultimately these tools would highlight interesting changes in these features, so users no longer have to spend time on the tedious procedure of identifying and extracting the relevant information from the vast amount of numbers resulting from this type of analysis.

Conclusion

We created a tool that determines the overlap between genomic co-ordinates and genes, allowing for the determination of expression of genes defined by read counts per gene. ReadCounter can counts reads mapping to each intron and each exon as well as whole genes. Furthermore, it reduces the number of ambiguously overlapping reads compared to other tools. Contrary to other tools, those reads that do map ambiguously are reported in a separate column. This tool does so at an at least 10 fold faster rate than HTSeq. Additionally, this tool counts reads for unsorted files at an unprecedented speed. ReadCounter is available online at www.GeneFriends.org/ReadCounter/ and is free to use. To further facilitate RNA-seq expression analysis we documented a walkthrough that allows researchers that are completely new to RNA-seq analysis to convert RNA-seq data into read counts per gene files (www.GeneFriends.org/RNA-seqForDummies). Lastly, a script is available that runs through these steps automatically.

References

1. Zhao Y, Li H, Fang S et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs, Nucleic Acids Res 2016;44:D203-208.
2. Bruce Alberts DB, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson Studying Gene Expression and Function, Molecular Biology of the Cell. 4th edition. 2002.
3. Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function, Nat Rev Genet 2004;5:11-22.
4. Stuart JM, Segal E, Koller D et al. A gene-coexpression network for global discovery of conserved genetic modules, Science 2003;302:249-255.
5. Fehrmann RS, Karjalainen JM, Krajewska M et al. Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat Genet 2015;47:115-125.
6. Obayashi T, Hayashi S, Shibaoka M et al. COXPRESdb: a database of coexpressed gene networks in mammals, Nucleic Acids Res 2008;36:D77-82.
7. Walker MG, Volkmuth W, Klingler TM. Pharmaceutical target discovery using Guilt-by-Association: schizophrenia and Parkinson's disease genes, Proc Int Conf Intell Syst Mol Biol 1999:282-286.
8. Chen J, Ma M, Shen N et al. Integration of cancer gene co-expression network and metabolic network to uncover potential cancer drug targets, J Proteome Res 2013;12:2354-2364.
9. Fujiwara T, Hiramatsu M, Isagawa T et al. ASCL1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis, Lung Cancer 2012;75:119-125.
10. Balagurunathan Y, Morse DL, Hostetter G et al. Gene expression profiling-based identification of cell-surface targets for developing multimeric ligands in pancreatic cancer, Mol Cancer Ther 2008;7:3071-3080.
11. Segal E, Friedman N, Koller D et al. A module map showing conditional activity of expression modules in cancer, Nat Genet 2004;36:1090-1098.
12. Sweet-Cordero A, Mukherjee S, Subramanian A et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis, Nat Genet 2005;37:48-55.
13. Torkamani A, Dean B, Schork NJ et al. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia, Genome Res 2010;20:403-412.
14. Ge W, Ma X, Li X et al. B7-H1 up-regulation on dendritic-like leukemia cells suppresses T cell immune function through modulation of IL-10/IL-12 production and generation of Treg cells, Leuk Res 2009;33:948-957.
15. Mootha VK, Lindgren CM, Eriksson KF et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, Nat Genet 2003;34:267-273.
16. Singer GAC, Lloyd AT, Huminiecki LB et al. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection, Molecular Biology and Evolution 2005;22:767-775.
17. Franke L, van Bakel H, Fokkens L et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, Am J Hum Genet 2006;78:1011-1025.

18. McCarroll SA, Murphy CT, Zou S et al. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging, Nat Genet 2004;36:197-204.
19. Ala U, Piro RM, Grassi E et al. Prediction of human disease genes by human-mouse conserved coexpression analysis, PLoS Comput Biol 2008;4:e1000043.
20. van Dam S, Cordeiro R, Craig T et al. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases, BMC Genomics 2012;13:535.
21. Allen JD, Xie Y, Chen M et al. Comparing statistical methods for constructing large scale gene networks, PLoS One 2012;7:e29348.
22. D'Haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering, Bioinformatics 2000;16:707-726.
23. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks, BMC Bioinformatics 2005;6:227.
24. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms, PLoS Biol 2004;2:85-93.
25. Chou WC, Cheng AL, Brotto M et al. Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer, BMC Genomics 2014;15:300.
26. Keller MP, Choi Y, Wang P et al. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility, Genome Res 2008;18:706-716.
27. Presson AP, Sobel EM, Papp JC et al. Integrated Weighted Gene Co-expression Network Analysis with an Application to Chronic Fatigue Syndrome, BMC Syst Biol 2009;2:95.
28. Voineagu I, Wang XC, Johnston P et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology, Nature 2011;474:380-384.
29. Khalil AM, Guttman M, Huarte M et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, Proc Natl Acad Sci U S A 2009;106:11667-11672.
30. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy, Nat Biotechnol 2004;22:535-546.
31. Jin G, Sun J, Isaacs SD et al. Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk, Carcinogenesis 2011;32:1655-1659.
32. Gupta RA, Shah N, Wang KC et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis, Nature 2010;464:1071-1076.
33. Navarro P, Page DR, Avner P et al. Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program, Genes Dev 2006;20:2787-2792.
34. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions, Nat Rev Genet 2009;10:155-159.
35. Plath K, Fang J, Mlynarczyk-Evans SK et al. Role of histone H3 lysine 27 methylation in X inactivation, Science 2003;300:131-135.
36. Gregory RI, Chendrimada TP, Cooch N et al. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing, Cell 2005;123:631-640.
37. Pratt AJ, MacRae IJ. The RNA-induced silencing complex: a versatile gene-silencing machine, J Biol Chem 2009;284:17897-17901.
38. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code, RNA 2008;14:802-813.
39. Pessa HK, Will CL, Meng X et al. Minor spliceosome components are predominantly localized in the nucleus, Proc Natl Acad Sci U S A 2008;105:8655-8660.
40. Esteller M. Non-coding RNAs in human disease, Nat Rev Genet 2011;12:861-874.

41. Mockler TC, Chan S, Sundaresan A et al. Applications of DNA tiling arrays for whole-genome analysis, Genomics 2005;85:1-15.
42. Agarwal A, Koppstein D, Rozowsky J et al. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays, BMC Genomics 2010;11:383.
43. Zhao W, He X, Hoadley KA et al. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling, BMC Genomics 2014;15:419.
44. de Magalhaes JP, Finch CE, Janssens G. Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions, Ageing Research Reviews 2010;9:315-323.
45. Sekhon RS, Briskine R, Hirsch CN et al. Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays, PLoS One 2013;8:e61005.
46. Richard H, Schulz MH, Sultan M et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments, Nucleic Acids Res 2010;38.
47. Wood SH, Craig T, Li Y et al. Whole transcriptome sequencing of the aging rat brain reveals dynamic RNA changes in the dark matter of the genome, Age 2013;35:763-776.
48. Yang X, Coulombe-Huntington J, Kang S et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing, Cell 2016;164:805-817.
49. Kelemen O, Convertini P, Zhang Z et al. Function of alternative splicing, Gene 2013;514:1-30.
50. van Dam S, Craig T, de Magalhaes JP. GeneFriends: a human RNA-seq-based gene and transcript co-expression database, Nucleic Acids Res 2015;43:D1124-1132.
51. Heller MJ. DNA microarray technology: devices, systems, and applications, Ann Rev Biomed Eng 2002;4:129-153.
52. Fullwood MJ, Wei CL, Liu ET et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses, Genome Res 2009;19:521-532.
53. Bray NL, Pimentel H, Melsted P et al. Near-optimal probabilistic RNA-seq quantification, Nat Biotechnol 2016;34:525-527.
54. Patro R, Duggal G, Kingsford C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment 2015.
55. Dobin A, Davis CA, Schlesinger F et al. STAR: ultrafast universal RNA-seq aligner, Bioinformatics 2013;29:15-21.
56. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements, Nat Methods 2015;12:357-360.
57. Baruzzo G, Hayer KE, Kim EJ et al. Simulation-based comprehensive benchmarking of RNA-seq aligners, Nat Methods 2016.
58. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 2009;25:1105-1111.
59. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data, Bioinformatics 2015;31:166-169.
60. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 2014;30:923-930.
61. Trapnell C, Williams BA, Pertea G et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat Biotechnol 2010;28:511-515.
62. Mortazavi A, Williams BA, McCue K et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat Methods 2008;5:621-628.

63. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics 2011;12:323.
64. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biol 2010;11:R25.
65. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in metagenomic samples, Bioinformatics 2015;31:2269-2275.
66. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis, BMC Bioinformatics 2008;9:559.
67. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules, BMC Bioinformatics 2010;11:497.
68. Watson M. CoXpress: differential co-expression in gene expression data, BMC Bioinformatics 2006;7:509.
69. Pontes B, Giraldez R, Aguilar-Ruiz JS. Biclustering on expression data: A review, J Biomed Inform 2015.
70. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proc Natl Acad Sci U S A 2003;100:3351-3356.
71. Ponnapalli SP, Saunders MA, Van Loan CF et al. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms, PLoS One 2011;6:e28072.
72. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat Protoc 2009;4:44-57.
73. Mi HY, Muruganujan A, Casagrande JT et al. Large-scale gene function analysis with the PANTHER classification system, Nat Protoc 2013;8:1551-1566.
74. Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 2005;102:15545-15550.
75. Shannon P, Markiel A, Ozier O et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks, Genome Res 2003;13:2498-2504.
76. Theocharidis A, van Dongen S, Enright AJ et al. Network visualization and analysis of gene expression data using BioLayout Express(3D), Nat Protoc 2009;4:1535-1550.
77. Ghosh S, Chan CK. Analysis of RNA-Seq Data Using TopHat and Cufflinks, Methods Mol Biol 2016;1374:339-361.
78. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation, Bioinformatics 2012;28:1721-1728.
79. Yalamanchili HK, Li Z, Wang P et al. SpliceNet: recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples, Nucleic Acids Res 2014;42:e121.
80. Dillies MA, Rau A, Aubert J et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Brief Bioinform 2013;14:671-683.
81. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis, Bioinformatics 2002;18:1585-1592.
82. Irizarry RA, Hobbs B, Collin F et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 2003;4:249-264.
83. Irizarry RA, Bolstad BM, Collin F et al. Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Res 2003;31:e15.

84. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers, Bioinformatics 2015;31:2123-2130.
85. Li B, Tsoi LC, Swindell WR et al. Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms, J Invest Dermatol 2014;134:1828-1838.
86. Okamura Y, Aoki Y, Obayashi T et al. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems, Nucleic Acids Res 2015;43:D82-86.
87. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana, Bioinformatics 2013;29:717-724.
88. Efron B TR. Monographs on Statistics and Applied Probability., New York: Chapman and Hall 1993;57:An Introduction to the Bootstrap.
89. Steuer R, Kurths J, Daub CO et al. The mutual information: Detecting and evaluating dependencies between variables, Bioinformatics 2002;18:S231-S240.
90. Margolin AA, Nemenman I, Basso K et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC Bioinformatics 2006;7 Suppl 1:S7.
91. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, Pac Symp Biocomput 2000:418-429.
92. Guttman M, Donaghey J, Carey BW et al. lincRNAs act in the circuitry controlling pluripotency and differentiation, Nature 2011;477:295-U260.
93. van Someren EP, Vaes BL, Steegenga WT et al. Least absolute regression network analysis of the murine osteoblast differentiation network, Bioinformatics 2006;22:477-484.
94. Friedman N, Linial M, Nachman I et al. Using Bayesian networks to analyze expression data, J Comput Biol 2000;7:601-620.
95. Kumari S, Nie J, Chen HS et al. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery, PLoS One 2012;7:e50411.
96. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices, BMC Bioinformatics 2012;13:328.
97. Albert R, Barabasi AL. Statistical mechanics of complex networks, Reviews of Modern Physics 2002;74:47-97.
98. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis, Stat Appl Genet Mol Biol 2005;4:Article17.
99. D'Haeseleer P. How does gene expression clustering work?, Nat Biotechnol 2005;23:1499-1501.
100. Gupta S, Ellis SE, Ashar FN et al. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism, Nat Commun 2014;5:5748.
101. Chen J, Bardes EE, Aronow BJ et al. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res 2009;37:W305-311.
102. Spellman PT, Sherlock G, Zhang MQ et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell 1998;9:3273-3297.
103. Eisen MB, Spellman PT, Brown PO et al. Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci U S A 1999;96:10943-10943.
104. Wuchty S, Barabasi AL, Ferdig MT. Stable evolutionary signal in a Yeast protein interaction network, BMC Evol Biol 2006;6:8.

105. Benjamin S, Flotho S, Borchers T et al. Conjugated linoleic acid isomers and their precursor fatty acids regulate peroxisome proliferator-activated receptor subtypes and major peroxisome proliferator responsive element-bearing target genes in HepG2 cell model, J Zhejiang Univ Sci B 2013;14:115-123.
106. Zhou XHJ, Kao MCJ, Huang HY et al. Functional annotation and network reconstruction through cross-platform integration of microarray data, Nat Biotechnol 2005;23:238-243.
107. Quackenbush J. Microarrays - Guilt by association, Science 2003;302:240-241.
108. Mason MJ, Fan G, Plath K et al. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells, BMC Genomics 2009;10:327.
109. Luo F, Yang Y, Zhong J et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory, BMC Bioinformatics 2007;8:299.
110. Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data, Proc Natl Acad Sci U S A 2002;99:12783-12788.
111. Bauer-Mehren A, Rautschka M, Sanz F et al. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks, Bioinformatics 2010;26:2924-2926.
112. Obayashi T, Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals, Nucleic Acids Res 2011;39:D1016-1022.
113. Price MN, Rieffel E. Finding coexpressed genes in counts-based data: an improved measure with validation experiments, Bioinformatics 2004;20:945-952.
114. Kong LJ, Fang M, Zhan HS et al. Tuina-focused integrative chinese medical therapies for inpatients with low back pain: a systematic review and meta-analysis, Evid Based Complement Alternat Med 2012;2012:578305.
115. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet 2000;25:25-29.
116. Joshi-Tope G, Gillespie M, Vastrik I et al. Reactome: a knowledgebase of biological pathways, Nucleic Acids Res 2005;33:D428-432.
117. Holmberg CI, Tran SE, Eriksson JE et al. Multisite phosphorylation provides sophisticated regulation of transcription factors, Trends Biochem Sci 2002;27:619-627.
118. Sterner DE, Berger SL. Acetylation of histones and transcription-related factors, Microbiol Mol Biol Rev 2000;64:435-459.
119. Medvedeva YA, Khamis AM, Kulakovskiy IV et al. Effects of cytosine methylation on transcription factor binding sites, BMC Genomics 2014;15:119.
120. Tootle TL, Rebay I. Post-translational modifications influence transcription factor activity: a view from the ETS superfamily, Bioessays 2005;27:285-298.
121. Yu X, Lin J, Zack DJ et al. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues, Nucleic Acids Res 2006;34:4925-4936.
122. Hu R, Qi G, Kong Y et al. Comprehensive analysis of NAC domain transcription factor gene family in Populus trichocarpa, BMC Plant Biol 2010;10:145.
123. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function, BMC Bioinformatics 2004;5:18.
124. Purmann A, Toedling J, Schueler M et al. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality, Genomics 2007;89:580-587.
125. Marco A, Konikoff C, Karr TL et al. Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster, Bioinformatics 2009;25:2473-2477.

126. De Bleser P, Hooghe B, Vlieghe D et al. A distance difference matrix approach to identifying transcription factors that regulate differential gene expression, Genome Biol 2007;8:R83.
127. Perco P, Kainz A, Mayer G et al. Detection of coregulation in differential gene expression profiles, Biosystems 2005;82:235-247.
128. Bonneau R, Reiss DJ, Shannon P et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo, Genome Biol 2006;7:R36.
129. Vandepoele K, Quimbaya M, Casneuf T et al. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks, Plant Physiol 2009;150:535-546.
130. Antoniotti M, Bader GD, Caravagna G et al. GeStoDifferent: a Cytoscape plugin for the generation and the identification of gene regulatory networks describing a stochastic cell differentiation process, Bioinformatics 2013;29:513-514.
131. Gu QA, Nagaraj SH, Hudson NJ et al. Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle, BMC Genomics 2011;12:23.
132. Budovskaya YV, Wu K, Southworth LK et al. An elt-3/elt-5/elt-6 GATA transcription circuit guides aging in C-elegans, Cell 2008;134:291-303.
133. Veerla S, Hoglund M. Analysis of promoter regions of co-expressed genes identified by microarray analysis, BMC Bioinformatics 2006;7:384.
134. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks, Nature 2000;406:378-382.
135. Langfelder P, Mischel PS, Horvath S. When Is Hub Gene Selection Better than Standard Meta-Analysis?, PLoS One 2013;8:e61505.
136. Chen YQ, Zhu J, Lum PY et al. Variations in DNA elucidate molecular networks that cause disease, Nature 2008;452:429-435.
137. Freeman LC. Centrality in Social Networks Conceptual Clarification, Social Networks 1979;1:215-239.
138. Zhao W, Langfelder P, Fuller T et al. Weighted Gene Coexpression Network Analysis: State of the Art, Journal of Biopharmaceutical Statistics 2010;20:281-300.
139. Hu R, Qiu X, Glazko G et al. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection, BMC Bioinformatics 2009;10:20.
140. Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes, Bioinformatics 2004;20 Suppl 1:i194-199.
141. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression, PLoS Comput Biol 2013;9:e1002955.
142. Hudson NJ, Reverter A, Dalrymple BP. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation, PLoS Comput Biol 2009;5:e1000382.
143. Pierson E, Consortium GT, Koller D et al. Sharing and Specificity of Co-expression Networks across 35 Human Tissues, PLoS Comput Biol 2015;11:e1004220.
144. Zeisel A, Munoz-Manchado AB, Codeluppi S et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, Science 2015;347:1138-1142.
145. Gao Q, Ho C, Jia Y et al. Biclustering of linear patterns in gene expression data, J Comput Biol 2012;19:619-631.
146. Xiao X, Moreno-Moral A, Rotival M et al. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules, PLoS Genet 2014;10:e1004006.

147. Anglani R, Creanza TM, Liuzzi VC et al. Loss of connectivity in cancer co-expression networks, PLoS One 2014;9:e87075.
148. Yue F, Cheng Y, Breschi A et al. A comparative encyclopedia of DNA elements in the mouse genome, Nature 2014;515:355-364.
149. Mele M, Ferreira PG, Reverter F et al. Human genomics. The human transcriptome across tissues and individuals, Science 2015;348:660-665.
150. Salgado H, Gama-Castro S, Peralta-Gil M et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions, Nucleic Acids Res 2006;34:D394-397.
151. Bredenkamp N, Nowell CS, Blackburn CC. Regeneration of the aged thymus by a single transcription factor, Development 2014;141:1627-1637.
152. Barrett T, Troup DB, Wilhite SE et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update, Nucleic Acids Res 2007;35:D760-765.
153. Jayaram S, Gupta MK, Shivakumar BM et al. Insights from Chromosome-Centric Mapping of Disease-Associated Genes: Chromosome 12 Perspective, J Proteome Res 2015;14:3432-3440.
154. Muthukumar R, Alexandar V, Thangam B et al. A systems biological approach reveals multiple crosstalk mechanism between gram-positive and negative bacterial infections: an insight into core mechanism and unique molecular signatures, PLoS One 2014;9:e89993.
155. Haitjema A, Mol BM, Kooi IE et al. Coregulation of FANCA and BRCA1 in human cells, Springerplus 2014;3:381.
156. Ashbrook DG, Williams RW, Lu L et al. A cross-species genetic analysis identifies candidate genes for mouse anxiety and human bipolar disorder, Front Behav Neurosci 2015;9:171.
157. Ashbrook DG, Williams RW, Lu L et al. Joint genetic analysis of hippocampal size in mouse and human identifies a novel gene linked to neurodegenerative disease, BMC Genomics 2014;15:850.
158. Iakova P, Awad SS, Timchenko NA. Aging reduces proliferative capacities of liver by switching pathways of C/EBPalpha growth arrest, Cell 2003;113:495-506.
159. Zhang P, Iwasaki-Arai J, Iwasaki H et al. Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha, Immunity 2004;21:853-863.
160. Rubin MA, Zhou M, Dhanasekaran SM et al. alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer, JAMA 2002;287:1662-1670.
161. Tanwar MK, Gilbert MR, Holland EC. Gene expression microarray analysis reveals YKL-40 to be a potential serum marker for malignant character in human glioma, Cancer Res 2002;62:4364-4368.
162. Mok SC, Chao J, Skates S et al. Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology, J Natl Cancer Inst 2001;93:1458-1464.
163. van de Rijn M, Perou CM, Tibshirani R et al. Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome, Am J Pathol 2002;161:1991-1996.
164. Ye QH, Qin LX, Forgues M et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning, Nat Med 2003;9:416-423.
165. Armstrong SA, Kung AL, Mabon ME et al. Inhibition of FLT3 in MLL. Validation of a therapeutic target identified by gene expression based classification, Cancer Cell 2003;3:173-183.

166. Aid-Pavlidis T, Pavlidis P, Timmusk T. Meta-coexpression conservation analysis of microarray data: a "subset" approach provides insight into brain-derived neurotrophic factor regulation, BMC Genomics 2009;10:420.
167. Hughes TR, Marton MJ, Jones AR et al. Functional discovery via a compendium of expression profiles, Cell 2000;102:109-126.
168. Kim SK, Lund J, Kiraly M et al. A gene expression map for Caenorhabditis elegans, Science 2001;293:2087-2092.
169. Wu CJ, Kasif S. GEMS: a web server for biclustering analysis of expression data, Nucleic Acids Res 2005;33:W596-599.
170. de Magalhaes JP, Curado J, Church GM. Meta-analysis of age-related gene expression profiles identifies common signatures of aging, Bioinformatics 2009;25:875-881.
171. Dennis G, Jr., Sherman BT, Hosack DA et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol 2003;4:P3.
172. Toyota M, Ahuja N, Ohe-Toyota M et al. CpG island methylator phenotype in colorectal cancer, Proc Natl Acad Sci U S A 1999;96:8681-8686.
173. Phillips DR, Jennings LK, Prasanna HR. Ca2+-mediated association of glycoprotein G (thrombinsensitive protein, thrombospondin) with human platelets, J Biol Chem 1980;255:11629-11632.
174. Linsalata M, Giannini R, Notarnicola M et al. Peroxisome proliferator-activated receptor gamma and spermidine/spermine N1-acetyltransferase gene expressions are significantly correlated in human colorectal cancer, BMC Cancer 2006;6:191.
175. Starlets D, Gore Y, Binsky I et al. Cell-surface CD74 initiates a signaling cascade leading to cell proliferation and survival, Blood 2006;107:4807-4816.
176. Schaible UE, Collins HL, Priem F et al. Correction of the iron overload defect in beta-2-microglobulin knockout mice by lactoferrin abolishes their increased susceptibility to tuberculosis, J Exp Med 2002;196:1507-1513.
177. Piredda L, Farrace MG, Lo Bello M et al. Identification of 'tissue' transglutaminase binding proteins in neural cells committed to apoptosis, FASEB J 1999;13:355-364.
178. Fesus L, Piacentini M. Transglutaminase 2: an enigmatic enzyme with diverse functions, Trends Biochem Sci 2002;27:534-539.
179. Parente L, Solito E. Annexin 1: more than an anti-phospholipase protein, Inflamm Res 2004;53:125-132.
180. Grewal T, Enrich C. Annexins--modulators of EGF receptor signalling and trafficking, Cell Signal 2009;21:847-858.
181. Hart PS, Pallos D, Zhang Y et al. Identification of a novel cathepsin C mutation (p.W185X) in a Brazilian kindred with Papillon-Lefevre syndrome, Mol Genet Metab 2002;76:145-147.
182. Hewitt C, McCormick D, Linden G et al. The role of cathepsin C in Papillon-Lefevre syndrome, prepubertal periodontitis, and aggressive periodontitis, Hum Mutat 2004;23:222-228.
183. Anderson RA, Byrum RS, Coates PM et al. Mutations at the lysosomal acid cholesteryl ester hydrolase gene locus in Wolman disease, Proc Natl Acad Sci U S A 1994;91:2718-2722.
184. Tinari N, Kuwabara I, Huflejt ME et al. Glycoprotein 90K/MAC-2BP interacts with galectin-1 and mediates galectin-1-induced cell aggregation, Int J Cancer 2001;91:167-172.
185. Beauchamp NJ, Daly ME, Makris M et al. A novel mutation in intron K of the PROS1 gene causes aberrant RNA splicing and is a common cause of protein S deficiency in a UK thrombophilia cohort, Thromb Haemost 1998;79:1086-1091.

186. Comp PC, Esmon CT. Recurrent venous thromboembolism in patients with a partial deficiency of protein S, N Engl J Med 1984;311:1525-1528.
187. Biddinger SB, Miyazaki M, Boucher J et al. Leptin suppresses stearoyl-CoA desaturase 1 by mechanisms independent of insulin and sterol regulatory element-binding protein-1c, Diabetes 2006;55:2032-2041.
188. Ramji DP, Foka P. CCAAT/enhancer-binding proteins: structure, function and regulation, Biochem J 2002;365:561-575.
189. Sterneck E, Paylor R, Jackson-Lewis V et al. Selectively enhanced contextual fear conditioning in mice lacking the transcriptional regulator CCAAT/enhancer binding protein delta, Proc Natl Acad Sci U S A 1998;95:10908-10913.
190. Gigliotti AP, Johnson PF, Sterneck E et al. Nulliparous CCAAT/enhancer binding proteindelta (C/EBPdelta) knockout mice exhibit mammary gland ductal hyperlasia, Exp Biol Med (Maywood) 2003;228:278-285.
191. Huang AM, Montagna C, Sharan S et al. Loss of CCAAT/enhancer binding protein delta promotes chromosomal instability, Oncogene 2004;23:1549-1557.
192. Tacutu R, Craig T, Budovsky A et al. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing, Nucleic Acids Res 2013;41:D1027-1033.
193. Saftig P, Eskelinen EL. Live longer with LAMP-2, Nat Med 2008;14:909-910.
194. Yajima N, Sakamaki K, Yonehara S. Age-related thymic involution is mediated by Fas on thymic epithelial cells, Int Immunol 2004;16:1027-1035.
195. Coschigano KT, Holland AN, Riders ME et al. Deletion, but not antagonism, of the mouse growth hormone receptor results in severely decreased body weights, insulin, and insulin-like growth factor I levels and increased life span, Endocrinology 2003;144:3799-3810.
196. Miwa N, Uebi T, Kawamura S. S100-annexin complexes--biology of conditional association, FEBS J 2008;275:4945-4955.
197. Chiu CH, Lin WD, Huang SY et al. Effect of a C/EBP gene replacement on mitochondrial biogenesis in fat cells, Genes Dev 2004;18:1970-1975.
198. Karagiannides I, Tchkonia T, Dobson DE et al. Altered expression of C/EBP family members results in decreased adipogenesis with aging, Am J Physiol Regul Integr Comp Physiol 2001;280:R1772-1780.
199. Guruceaga E, Segura V, Corrales FJ et al. FactorY, a bioinformatic resource for genome-wide promoter analysis, Comput Biol Med 2009;39:385-387.
200. O'Connell BC, Adamson B, Lydeard JR et al. A Genome-wide Camptothecin Sensitivity Screen Identifies a Mammalian MMS22L-NFKBIL2 Complex Required for Genomic Stability, Mol Cell 2010;40:645-657.
201. Ohta S, Shiomi Y, Sugimoto K et al. A proteomics approach to identify proliferating cell nuclear antigen (PCNA)-binding proteins in human cell lysates. Identification of the human CHL12/RFCs2-5 complex as a novel PCNA-binding protein, J Biol Chem 2002;277:40362-40367.
202. Boutros R, Lobjois V, Ducommun B. CDC25 phosphatases in cancer cells: key players? Good targets?, Nat Rev Cancer 2007;7:495-507.
203. Sawa M, Masai H. Drug design with Cdc7 kinase: a potential novel cancer therapy target, Drug Des Devel Ther 2009;2:255-264.
204. Blachon S, Gopalakrishnan J, Omori Y et al. Drosophila asterless and vertebrate Cep152 Are orthologs essential for centriole duplication, Genetics 2008;180:2081-2094.
205. Okada M, Cheeseman IM, Hori T et al. The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres, Nat Cell Biol 2006;8:446-457.

206. Suzuki C, Daigo Y, Ishikawa N et al. ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway, Cancer Res 2005;65:11314-11325.
207. Nasmyth K. Segregating sister genomes: the molecular biology of chromosome separation, Science 2002;297:559-565.
208. Foltz DR, Jansen LE, Black BE et al. The human CENP-A centromeric nucleosome-associated complex, Nat Cell Biol 2006;8:458-469.
209. Jallepalli PV, Lengauer C. Chromosome segregation and cancer: cutting through the mystery, Nat Rev Cancer 2001;1:109-117.
210. Hayama S, Daigo Y, Yamabuki T et al. Phosphorylation and activation of cell division cycle associated 8 by aurora kinase B plays a significant role in human lung carcinogenesis, Cancer Res 2007;67:4113-4122.
211. Macurek L, Lindqvist A, Medema RH. Aurora-A and hBora join the game of Polo, Cancer Res 2009;69:4555-4558.
212. Gonzalez S, Klatt P, Delgado S et al. Oncogenic activity of Cdc6 through repression of the INK4/ARF locus, Nature 2006;440:702-706.
213. Giaever G, Nislow C. The yeast deletion collection: a decade of functional genomics, Genetics 2014;197:451-465.
214. Osanai T, Nakamura M, Sasaki S et al. Plasma concentration of coupling factor 6 and cardiovascular events in patients with end-stage renal disease, Kidney Int 2003;64:2291-2297.
215. Beauchemin AM, Gottlieb B, Beitel LK et al. Cytochrome c oxidase subunit Vb interacts with human androgen receptor: a potential mechanism for neurotoxicity in spinobulbar muscular atrophy, Brain Res Bull 2001;56:285-297.
216. Kaput J, Swartz D, Paisley E et al. Diet-disease interactions at the molecular level: an experimental paradigm, J Nutr 1994;124:1296S-1305S.
217. Barstead RJ, Waterston RH. Vinculin is essential for muscle function in the nematode, J Cell Biol 1991;114:715-724.
218. Ramasamy A, Mondry A, Holmes CC et al. Key issues in conducting a meta-analysis of gene expression microarray datasets, PLoS Med 2008;5:e184.
219. Berrier A, Siu G, Calame K. Transcription of a minimal promoter from the NF-IL6 gene is regulated by CREB/ATF and SP1 proteins in U937 promonocytic cells, J Immunol 1998;161:2267-2275.
220. Descombes P, Schibler U. A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA, Cell 1991;67:569-579.
221. Luedde T, Duderstadt M, Streetz KL et al. C/EBP beta isoforms LIP and LAP modulate progression of the cell cycle in the regenerating mouse liver, Hepatology 2004;40:356-365.
222. de Magalhaes JP, Cabral JA, Magalhaes D. The influence of genes on the aging process of mice: a statistical assessment of the genetics of aging, Genetics 2005;169:265-274.
223. Harrison DE, Strong R, Sharp ZD et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice, Nature 2009;460:392-395.
224. Calkhoven CF, Muller C, Leutz A. Translational control of C/EBPalpha and C/EBPbeta isoform expression, Genes Dev 2000;14:1920-1932.
225. Lazo JS, Wipf P. Is Cdc25 a druggable target?, Anticancer Agents Med Chem 2008;8:837-842.
226. Brezak MC, Kasprzyk PG, Galcera MO et al. CDC25 inhibitors as anticancer agents are moving forward, Anticancer Agents Med Chem 2008;8:857-862.

227. Lavecchia A, Di Giovanni C, Novellino E. Inhibitors of Cdc25 phosphatases as anticancer agents: a patent review, Expert Opin Ther Pat 2010;20:405-425.
228. Montagnoli A, Moll J, Colotta F. Targeting cell division cycle 7 kinase: a new approach for cancer therapy, Clin Cancer Res 2010;16:4503-4508.
229. Morgan H, Beck T, Blake A et al. EuroPhenome: a repository for high-throughput mouse phenotyping data, Nucleic Acids Res 2010;38:D577-585.
230. Piao L, Nakagawa H, Ueda K et al. C12orf48, termed PARP-1 binding protein, enhances poly(ADP-ribose) polymerase-1 (PARP-1) activity and protects pancreatic cancer cells from DNA damage, Genes Chromosomes Cancer 2011;50:13-24.
231. Lebon S, Chol M, Benit P et al. Recurrent de novo mitochondrial DNA mutations in respiratory chain deficiency, J Med Genet 2003;40:896-899.
232. Kirby DM, Salemi R, Sugiana C et al. NDUFS6 mutations are a novel cause of lethal neonatal mitochondrial complex I deficiency, J Clin Invest 2004;114:837-845.
233. Loeffen JL, Smeitink JA, Trijbels JM et al. Isolated complex I deficiency in children: clinical, biochemical and genetic aspects, Hum Mutat 2000;15:123-134.
234. Pitkanen S, Feigenbaum A, Laframboise R et al. NADH-coenzyme Q reductase (complex I) deficiency: heterogeneity in phenotype and biochemical findings, J Inherit Metab Dis 1996;19:675-686.
235. Robinson BH. Human complex I deficiency: clinical spectrum and involvement of oxygen free radicals in the pathogenicity of the defect, Biochim Biophys Acta 1998;1364:271-286.
236. Wallace DC. Mitochondrial diseases in man and mouse, Science 1999;283:1482-1488.
237. Janssen RJ, Nijtmans LG, van den Heuvel LP et al. Mitochondrial complex I: structure, function and pathology, J Inherit Metab Dis 2006;29:499-515.
238. Pagliarini DJ, Calvo SE, Chang B et al. A mitochondrial protein compendium elucidates complex I disease biology, Cell 2008;134:112-123.
239. Lage K, Hansen NT, Karlberg EO et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes, Proc Natl Acad Sci U S A 2008;105:20870-20875.
240. Zhang X, Joehanes R, Chen BH et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood, Nat Genet 2015;47:345-352.
241. Azuaje F, Zhang L, Jeanty C et al. Analysis of a gene co-expression network establishes robust association between Col5a2 and ischemic heart disease, BMC Med Genomics 2013;6:13.
242. Wolf DM, Lenburg ME, Yau C et al. Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity, PLoS One 2014;9:e88309.
243. Yang Y, Han L, Yuan Y et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types, Nat Commun 2014;5:3231.
244. De Bodt S, Carvajal D, Hollunder J et al. CORNET: a user-friendly tool for data mining and integration, Plant Physiol 2010;152:1167-1179.
245. Hruz T, Laule O, Szabo G et al. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes, Adv Bioinformatics 2008;2008:420747.
246. Jupiter D, Chen H, VanBuren V. STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data, BMC Bioinformatics 2009;10:332.
247. Obayashi T, Okamura Y, Ito S et al. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals, Nucleic Acids Res 2013;41:D1014-1020.

248. Guan Y, Myers CL, Lu R et al. A genomewide functional network for the laboratory mouse, PLoS Comput Biol 2008;4:e1000165.
249. Guan Y, Gorenshteyn D, Burmeister M et al. Tissue-specific functional networks for prioritizing phenotype and disease genes, PLoS Comput Biol 2012;8:e1002694.
250. Wang K, Narayanan M, Zhong H et al. Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases, PLoS Comput Biol 2009;5:e1000616.
251. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes, Genome Res 1999;9:1106-1115.
252. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet 2009;10:57-63.
253. Kodama Y, Shumway M, Leinonen R et al. The sequence read archive: explosive growth of sequencing data, Nucleic Acids Res 2012;40:D54-D56.
254. Wang X, Song X, Glass CK et al. The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs, Cold Spring Harb Perspect Biol 2011;3:a003756.
255. Flicek P, Amode MR, Barrell D et al. Ensembl 2014, Nucleic Acids Res 2014;42:D749-755.
256. Pan Q, Shai O, Lee LJ et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat Genet 2008;40:1413-1415.
257. Mattick JS. The genetic signatures of noncoding RNAs, PLoS Genet 2009;5:e1000459.
258. Taft RJ, Pang KC, Mercer TR et al. Non-coding RNAs: regulators of disease, J Pathol 2010;220:126-139.
259. Manfield IW, Jen CH, Pinney JW et al. Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis, Nucleic Acids Res 2006;34:W504-509.
260. Iancu OD, Kawane S, Bottomly D et al. Utilizing RNA-Seq data for de novo coexpression network inference, Bioinformatics 2012;28:1592-1597.
261. Franceschini A, Szklarczyk D, Frankild S et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Res 2013;41:D808-815.
262. Ooi HS, Schneider G, Chan YL et al. Databases of protein-protein interactions and complexes, Methods Mol Biol 2010;609:145-159.
263. Wong AK, Park CY, Greene CS et al. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks, Nucleic Acids Res 2012;40:W484-490.
264. Li HD, Menon R, Omenn GS et al. The emerging era of genomic data integration for analyzing splice isoform function, Trends Genet 2014;30:340-347.
265. Leinonen R, Sugawara H, Shumway M et al. The sequence read archive, Nucleic Acids Res 2011;39:D19-21.
266. Bhat P, Yang H, Bogre L et al. Computational selection of transcriptomics experiments improves Guilt-by-Association analyses, PLoS One 2012;7:e39681.
267. Flicek P, Ahmed I, Amode MR et al. Ensembl 2013, Nucleic Acids Res 2013;41:D48-55.
268. Kanehisa M, Goto S, Sato Y et al. KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res 2012;40:D109-114.
269. Hamosh A, Scott AF, Amberger JS et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res 2005;33:D514-517.
270. Mandard S, Muller M, Kersten S. Peroxisome proliferator-activated receptor alpha target genes, Cell Mol Life Sci 2004;61:393-416.

271. Michalik L, Desvergne B, Tan NS et al. Impaired skin wound healing in peroxisome proliferator-activated receptor (PPAR)alpha and PPARbeta mutant mice, J Cell Biol 2001;154:799-814.
272. Dobrin R, Zhu J, Molony C et al. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease, Genome Biol 2009;10:R55.
273. Feng J, Bi C, Clark BS et al. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator, Genes Dev 2006;20:1470-1484.
274. Kalantry S, Purushothaman S, Bowen RB et al. Evidence of Xist RNA-independent initiation of mouse imprinted X-chromosome inactivation, Nature 2009;460:647-651.
275. Giorgetti L, Lajoie BR, Carter AC et al. Structural organization of the inactive X chromosome in the mouse, Nature 2016;535:575-579.
276. Rinn JL, Kertesz M, Wang JK et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs, Cell 2007;129:1311-1323.
277. Tsai MC, Manor O, Wan Y et al. Long noncoding RNA as modular scaffold of histone modification complexes, Science 2010;329:689-693.
278. Glusman G, Yanai I, Rubin I et al. The complete human olfactory subgenome, Genome Res 2001;11:685-702.
279. Zozulya S, Echeverri F, Nguyen T. The human olfactory receptor repertoire, Genome Biol 2001;2:RESEARCH0018.
280. Vanin EF. Processed pseudogenes: characteristics and evolution, Annu Rev Genet 1985;19:253-272.
281. Ben-Arie N, Lancet D, Taylor C et al. Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire, Hum Mol Genet 1994;3:229-235.
282. Trask BJ, Massa H, Brand-Arpon V et al. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome, Hum Mol Genet 1998;7:2007-2020.
283. Gilad Y, Man O, Paabo S et al. Human specific loss of olfactory receptor genes, Proc Natl Acad Sci U S A 2003;100:3324-3327.
284. Kodama A, Karakesisoglou I, Wong E et al. ACF7: an essential integrator of microtubule dynamics, Cell 2003;115:343-354.
285. Chen HJ, Lin CM, Lin CS et al. The role of microtubule actin cross-linking factor 1 (MACF1) in the Wnt signaling pathway, Genes Dev 2006;20:1933-1945.
286. Wu X, Kodama A, Fuchs E. ACF7 regulates cytoskeletal-focal adhesion dynamics and migration and has ATPase activity, Cell 2008;135:137-148.
287. Wu X, Shen QT, Oristian DS et al. Skin stem cells orchestrate directional migration by regulating microtubule-ACF7 connections through GSK3beta, Cell 2011;144:341-352.
288. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias, Exp Cell Res 2014;322:12-20.
289. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples, Theory Biosci 2012;131:281-285.
290. Aanes H, Winata C, Moen LF et al. Normalization of RNA-sequencing data from samples with varying mRNA levels, PLoS One 2014;9:e89158.
291. McDaneld TG. MicroRNA: mechanism of gene regulation and application to livestock, J Anim Sci 2009;87:E21-28.
292. Pillai RS. MicroRNA function: multiple mechanisms for a tiny RNA?, RNA 2005;11:1753-1761.

293. Popadin K, Gutierrez-Arcelus M, Dermitzakis ET et al. Genetic and epigenetic regulation of human lincRNA gene expression, Am J Hum Genet 2013;93:1015-1026.
294. Wood SH, van Dam S, Craig T et al. Transcriptome analysis in calorie-restricted rats implicates epigenetic and post-translational mechanisms in neuroprotection and aging, Genome Biol 2015;16:285.
295. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions, Cell 2009;136:215-233.
296. Merry BJ, Kirk A, Goyns MH. Dietary lipoic acid supplementation can mimic or block the effect of dietary restriction on life span, Mechanisms of Ageing and Development 2008;129:341-348.
297. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data, Nucleic Acids Res 2014;42:D68-73.
298. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure, BMC Bioinformatics 2007;8:22.
299. Yu Y, Fuscoe JC, Zhao C et al. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages, Nat Commun 2014;5:3230.
300. St Laurent G, Shtokalo D, Tackett MR et al. On the importance of small changes in RNA expression, Methods 2013;63:18-24.
301. Ravasz E, Somera AL, Mongru DA et al. Hierarchical organization of modularity in metabolic networks, Science 2002;297:1551-1555.
302. Yip AM, Horvath S. The generalized topological overlap matrix for detecting modules in gene networks 2006.
303. Correia-Melo C, Marques FD, Anderson R et al. Mitochondria are required for pro-ageing features of the senescent phenotype, EMBO J 2016;35:724-742.
304. Boveris A, Navarro A. Brain mitochondrial dysfunction in aging, IUBMB Life 2008;60:308-314.
305. Carlson MRJ, Zhang B, Fang ZX et al. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks, BMC Genomics 2006;7:40.
306. Jeong H, Mason SP, Barabasi AL et al. Lethality and centrality in protein networks, Nature 2001;411:41-42.
307. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Molecular Biology and Evolution 2005;22:803-806.
308. Miller LA, Gunstad J, Spitznagel MB et al. CAMTA1 T polymorphism is associated with neuropsychological test performance in older adults with cardiovascular disease, Psychogeriatrics 2011;11:135-140.
309. Baleriola J, Walker CA, Jean YY et al. Axonally Synthesized ATF4 Transmits a Neurodegenerative Signal across Brain Regions, Cell 2014;158:1159-1172.
310. Endo M, Su L, Nielsen TO. Activating transcription factor 2 in mesenchymal tumors, Human Pathology 2014;45:276-284.
311. Rudraraju B, Droog M, Abdel-Fatah TMA et al. Phosphorylation of activating transcription factor-2 (ATF-2) within the activation domain is a key determinant of sensitivity to tamoxifen in breast cancer, Breast Cancer Research and Treatment 2014;147:295-309.
312. Gizard F, Robillard R, Gross B et al. TReP-132 is a novel progesterone receptor coactivator required for the inhibition of breast cancer cell growth and enhancement of differentiation by progesterone, Molecular and Cellular Biology 2006;26:7632-7644.

313. Dorssers LCJ, Veldscholte J. Identification of a novel breast-cancer-anti-estrogen-resistance (BCAR2) locus by cell-fusion-mediated gene transfer in human breast-cancer cells, International Journal of Cancer 1997;72:700-705.
314. Fu LY, Shi K, Wang JS et al. TFAP2B overexpression contributes to tumor growth and a poor prognosis of human lung adenocarcinoma through modulation of ERK and VEGF/PEDF signaling, Molecular Cancer 2014;13:89.
315. Han J, Guo X, Tan WH et al. The expression of p-ATF2 involved in the chondeocytes apoptosis of an endemic osteoarthritis, Kashin-Beck disease, Bmc Musculoskeletal Disorders 2013;14:209.
316. Vaquerizas JM, Kummerfeld SK, Teichmann SA et al. A census of human transcription factors: function, expression and evolution, Nat Rev Genet 2009;10:252-263.
317. Reis MD, Csomos K, Dias LP et al. Decline of FOXN1 gene expression in human thymus correlates with age: possible epigenetic regulation, Immun Ageing 2015;12:18.
318. Chen LZ, Xiao SY, Manley NR. Foxn1 is required to maintain the postnatal thymic microenvironment in a dosage-sensitive manner, Blood 2009;113:567-574.
319. Gui J, Zhu X, Dohkan J et al. The aged thymus shows normal recruitment of lymphohematopoietic progenitors but has defects in thymic epithelial cells, Int Immunol 2007;19:1201-1211.
320. Li PP, Hua X, Zhang Z et al. Characterization of regulatory features of housekeeping and tissue-specific regulators within tissue regulatory networks, BMC Syst Biol 2013;7:112.
321. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals, Mol Cell 2007;26:753-767.
322. Glass K, Girvan M. Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets, Sci Rep 2014;4:4191.
323. Clarke C, Doolan P, Barron N et al. CGCDB: a web-based resource for the investigation of gene coexpression in CHO cell culture, Biotechnol Bioeng 2012;109:1368-1370.
324. Michalopoulos I, Pavlopoulos GA, Malatras A et al. Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes, BMC Res Notes 2012;5:265.
325. Zhu F, Shi L, Li H et al. Modeling dynamic functional relationship networks and application to ex vivo human erythroid differentiation, Bioinformatics 2014;30:3325-3333.
326. Piro RM, Ala U, Molineris I et al. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction, Eur J Hum Genet 2011;19:1173-1180.
327. Wang P, Qi H, Song S et al. ImmuCo: a database of gene co-expression in immune cells, Nucleic Acids Res 2015;43:D1133-1139.
328. Williams G. Database of Gene Co-Regulation (dGCR): A Web Tool for Analysing Patterns of Gene Co-regulation across Publicly Available Expression Data, J Genomics 2015;3:29-35.
329. Kumar V, Westra HJ, Karjalainen J et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression, PLoS Genet 2013;9:e1003201.
330. Vlasblom J, Zuberi K, Rodriguez H et al. Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli, Bioinformatics 2015;31:306-310.
331. Mostafavi S, Ray D, Warde-Farley D et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function, Genome Biol 2008;9 Suppl 1:S4.
332. Zimmermann P, Hirsch-Hoffmann M, Hennig L et al. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox, Plant Physiol 2004;136:2621-2632.

333. Ostlund G, Lindskog M, Sonnhammer EL. Network-based Identification of novel cancer genes, Mol Cell Proteomics 2010;9:648-655.
334. Guala D, Sjolund E, Sonnhammer EL. MaxLink: network-based prioritization of genes tightly linked to a disease seed set, Bioinformatics 2014;30:2689-2690.
335. Jupiter DC, VanBuren V. A Visual Data Mining Tool that Facilitates Reconstruction of Transcription Regulatory Networks, PLoS One 2008;3:e1717.
336. Kolde R, Laur S, Adler P et al. Robust rank aggregation for gene list integration and meta-analysis, Bioinformatics 2012;28:573-580.
337. von Mering C, Jensen LJ, Snel B et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms, Nucleic Acids Res 2005;33:D433-D437.
338. Srinivasasainagendra V, Page GP, Mehta T et al. CressExpress: A tool for large-scale mining of expression data from Arabidopsis, Plant Physiol 2008;147:1004-1016.
339. De Bodt S, Inze D. A guide to CORNET for the construction of coexpression and protein-protein interaction networks, Methods Mol Biol 2013;1011:327-343.
340. Mutwil M, Klie S, Tohge T et al. PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species, Plant Cell 2011;23:895-910.
341. Wong DCJ, Sweetman C, Drew DP et al. VTCdb: a gene co-expression database for the crop species Vitis vinifera (grapevine), BMC Genomics 2013;14:882.
342. Sato Y, Namiki N, Takehisa H et al. RiceFREND: a platform for retrieving coexpressed gene networks in rice, Nucleic Acids Res 2013;41:D1214-D1221.
343. Cao PJ, Jung KH, Choi D et al. The Rice Oligonucleotide Array Database: an atlas of rice gene expression, Rice 2012;5:17.
344. Lee TH, Kim YK, Pham TTM et al. RiceArrayNet: A Database for Correlating Gene Expression from Transcriptome Profiling, and Its Application to the Analysis of Coexpressed Genes in Rice, Plant Physiol 2009;151:16-33.
345. Lee I, Ambaru B, Thakkar P et al. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana, Nat Biotechnol 2010;28:149-U114.
346. Obayashi T, Kinoshita K. Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways, Journal of Plant Research 2010;123:311-319.
347. Obayashi T, Okamura Y, Ito S et al. ATTED-II in 2014: Evaluation of Gene Coexpression in Agriculturally Important Plants, Plant and Cell Physiology 2014;55.
348. Yim WC, Yu Y, Song K et al. PLANEX: the plant co-expression database, BMC Plant Biol 2013;13:83.
349. Ogata Y, Suzuki H, Sakurai N et al. CoP: a database for characterizing co-expressed gene modules with biological information in plants, Bioinformatics 2010;26:1267-1268.
350. Toufighi K, Brady SM, Austin R et al. The Botany Array Resource: e-Northerns, Expression Angling, and Promoter analyses, Plant Journal 2005;43:153-163.
351. Winter EE, Goodstadt L, Ponting CP. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes, Genome Res 2004;14:54-61.
352. Gillis J, Pavlidis P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks, PLoS Comput Biol 2012;8:e1002444.
353. Zhang B, Gaiteri C, Bodea LG et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease, Cell 2013;153:707-720.
354. Rogers MF, Ben-Hur A. The use of gene ontology evidence codes in preventing classifier assessment bias, Bioinformatics 2009;25:1173-1177.
355. Foroushani ABK, Brinkman FSL, Lynn DJ. Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures, Peerj 2013;1:e229.

356. Tsaparas P, Marino-Ramirez L, Bodenreider O et al. Global similarity and local divergence in human and mouse gene co-expression networks, BMC Evol Biol 2006;6:70.
357. Johnsson P, Lipovich L, Grander D et al. Evolutionary conservation of long non-coding RNAs; sequence, structure, function, Biochimica Et Biophysica Acta-General Subjects 2014;1840:1063-1071.
358. Pellegrino M, Provero P, Silengo L et al. CLOE: Identification of putative functional relationships among genes by comparison of expression profiles between two species, BMC Bioinformatics 2004;5:179.
359. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains, Proc Natl Acad Sci U S A 2006;103:17973-17978.
360. Hay M, Thomas DW, Craighead JL et al. Clinical development success rates for investigational drugs, Nat Biotechnol 2014;32:40-51.
361. Monaco G, van Dam S, Casal Novo Ribeiro JL et al. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels, BMC Evol Biol 2015;15:259.
362. Mikkelsen TS, Hillier LW, Eichler EE et al. Initial sequence of the chimpanzee genome and comparison with the human genome, Nature 2005;437:69-87.
363. Bratic A, Larsson NG. The role of mitochondria in aging, J Clin Invest 2013;123:951-957.
364. Seo AY, Joseph AM, Dutta D et al. New insights into the role of mitochondria in aging: mitochondrial dynamics and more, J Cell Sci 2010;123:2533-2542.
365. Peterson CM, Johannsen DL, Ravussin E. Skeletal muscle mitochondria and aging: a review, J Aging Res 2012;2012:194821.
366. Segal E, Shapira M, Regev A et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat Genet 2003;34:166-176.
367. Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with Lemon-Tree, PLoS Comput Biol 2015;11:e1003983.
368. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements, Nat Genet 2001;29:153-159.
369. Lickwar CR, Mueller F, Hanlon SE et al. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function, Nature 2012;484:251-U141.
370. Zidek LM, Ackermann T, Hartleben G et al. Deficiency in mTORC1-controlled C/EBPbeta-mRNA translation improves metabolic health in mice, EMBO Rep 2015;16:1022-1036.
371. Huynh-Thu VA, Irrthum A, Wehenkel L et al. Inferring regulatory networks from expression data using tree-based methods, PLoS One 2010;5:e12776.
372. Hecker M, Lambeck S, Toepfer S et al. Gene regulatory network inference: data integration in dynamic models-a review, Biosystems 2009;96:86-103.
373. Glass K, Huttenhower C, Quackenbush J et al. Passing messages between biological networks to refine predicted interactions, PLoS One 2013;8:e64832.
374. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, Genomics 2004;83:349-360.
375. Ho JW, Bishop E, Karchenko PV et al. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis, BMC Genomics 2011;12:134.
376. Bar-Joseph Z, Gerber GK, Lee TI et al. Computational discovery of gene modules and regulatory networks, Nat Biotechnol 2003;21:1337-1342.

377. Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data, BMC Bioinformatics 2004;5:31.
378. Boulesteix AL, Strimmer K. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach, Theor Biol Med Model 2005;2:23.
379. Wu G, Ji H. ChIPXpress: using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking, BMC Bioinformatics 2013;14:188.
380. Karlebach G, Shamir R. Constructing logical models of gene regulatory networks by integrating transcription factor-DNA interactions with expression data: an entropy-based approach, J Comput Biol 2012;19:30-41.
381. Greene CS, Krishnan A, Wong AK et al. Understanding multicellular function and disease with human tissue-specific networks, Nat Genet 2015;47:569-576.
382. Westra HJ, Peters MJ, Esko T et al. Systematic identification of trans eQTLs as putative drivers of known disease associations, Nat Genet 2013;45:1238-U1195.
383. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, Science 2015;348:648-660.
384. Alipanahi B, Delong A, Weirauch MT et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat Biotechnol 2015;33:831-838.
385. Agarwal V, Bell GW, Nam JW et al. Predicting effective microRNA target sites in mammalian mRNAs, Elife 2015;4.
386. John B, Enright AJ, Aravin A et al. Human MicroRNA targets, PLoS Biol 2004;2:e363.
387. Vlachos IS, Paraskevopoulou MD, Karagkouni D et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions, Nucleic Acids Res 2015;43:D153-D159.
388. Chou CH, Chang NW, Shrestha S et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database, Nucleic Acids Res 2016;44:D239-247.
389. Naukkarinen J, Surakka I, Pietilainen KH et al. Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes, PLoS Genet 2010;6:e1000976.
390. Corradin O, Saiakhova A, Akhtar-Zaidi B et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, Genome Res 2014;24:1-13.
391. Fernandez-Jimenez N, Castellanos-Rubio A, Plaza-Izurieta L et al. Coregulation and modulation of NFkappaB-related genes in celiac disease: uncovered aspects of gut mucosal inflammation, Hum Mol Genet 2014;23:1298-1310.
392. Hofree M, Shen JP, Carter H et al. Network-based stratification of tumor mutations, Nat Methods 2013;10:1108-1115.
393. Nicolle R, Radvanyi F, Elati M. CoRegNet: reconstruction and integrated analysis of co-regulatory networks, Bioinformatics 2015;31:3066-3068.
394. Reiss DJ, Plaisier CL, Wu WJ et al. cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism, Nucleic Acids Res 2015;43:e87.
395. Marbach D, Costello JC, Kuffner R et al. Wisdom of crowds for robust gene network inference, Nat Methods 2012;9:796-804.
396. Ostlund G, Sonnhammer EL. Avoiding pitfalls in gene (co)expression meta-analysis, Genomics 2014;103:21-30.
397. Djordjevic D, Yang A, Zadoorian A et al. How difficult is inference of mammalian causal gene regulatory networks?, PLoS One 2014;9:e111661.

398. Xue Z, Huang K, Cai C et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing, Nature 2013;500:593-597.
399. Sun D, Luo M, Jeong M et al. Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal, Cell Stem Cell 2014;14:673-688.
400. Kowalczyk MS, Tirosh I, Heckl D et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells, Genome Res 2015;25:1860-1872.
401. Deelen P, Zhernakova DV, de Haan M et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels, Genome Medicine 2015;7:30.
402. Navin NE. The first five years of single-cell cancer genomics and beyond, Genome Res 2015;25:1499-1507.
403. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science, Nat Rev Genet 2013;14:618-630.
404. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics, Nat Rev Genet 2015;16:133-145.
405. Teng M, Love MI, Davis CA et al. A benchmark for RNA-seq quantification pipelines, Genome Biol 2016;17:74.
406. Bernstein BE, Stamatoyannopoulos JA, Costello JF et al. The NIH Roadmap Epigenomics Mapping Consortium, Nat Biotechnol 2010;28:1045-1048.
407. Bersanelli M, Mosca E, Remondini D et al. Methods for the integration of multi-omics data: mathematical aspects, BMC Bioinformatics 2016;17 Suppl 2:15.
408. Robinson PN, Kohler S, Bauer S et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, Am J Hum Genet 2008;83:610-615.
409. Metzker ML. Sequencing technologies - the next generation, Nat Rev Genet 2010;11:31-46.
410. Schuster SC. Next-generation sequencing transforms today's biology, Nat Methods 2008;5:16-18.
411. Shendure J, Mitra RD, Varma C et al. Advanced sequencing technologies: methods and goals, Nat Rev Genet 2004;5:335-344.
412. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies, Genomics 2009;93:105-111.
413. Marioni JC, Mason CE, Mane SM et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, Genome Res 2008;18:1509-1517.
414. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 2009;25:1754-1760.
415. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq, PLoS One 2013;8:e76935.
416. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools, Bioinformatics 2009;25:2078-2079.
417. Pages H ea. IRanges: infrastructure for manipulating intervals on sequences. 2013.
418. Gentleman RC, Carey VJ, Bates DM et al. Bioconductor: open software development for computational biology and bioinformatics, Genome Biol 2004;5:R80.
419. Wellcome Trust Sanger Institute. GFF (General Feature Format) specifications document. 2013:http://www.ensembl.org/info/website/upload/gff.html.
420. Pruitt KD, Tatusova T, Brown GR et al. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy, Nucleic Acids Res 2012;40:D130-135.

421. Langmead B, Trapnell C, Pop M et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 2009;10:R25.
422. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts, Genome Biol 2011;12:R72.
423. Louro R, El-Jundi T, Nakaya HI et al. Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci, Genomics 2008;92:18-25.
424. Nott A, Meislin SH, Moore MJ. A quantitative analysis of intron effects on mammalian gene expression, RNA 2003;9:607-617.
425. Reis EM, Louro R, Nakaya HI et al. As antisense RNA gets intronic, OMICS 2005;9:2-12.
426. Kim YK, Kim VN. Processing of intronic microRNAs, EMBO J 2007;26:775-783.