UPPSALA
UNIVERSITET

UPTEC X 14 029
Examensarbete 30 hp
Juni 2014

Differential and co-expression of long non-coding RNAs in abdominal aortic aneurysm

Joakim Karlsson

UPPSALA UNIVERSITET

Degree Project in Bioinformatics

Masters Programme in Molecular Biotechnology Engineering,
Uppsala University School of Engineering

UPTEC X 14 029	Date of issue 2014-11

Author

Joakim Karlsson

Title (English)

Differential and co-expression of long non-coding RNAs in abdominal aortic aneurysm

Title (Swedish)

Abstract

This project concerns an exploration of the presence and interactions of long non-coding RNA transcripts in an experimental atherosclerosis mouse model with relevance for human abdominal aortic aneurysm development. 187 long noncoding RNAs, two of them entirely novel, were found to be differentially expressed between angiotensin II treated (developing abdominal aortic aneurysms) and non-treated apolipoprotein E deficient mice (not developing aneurysms) harvested after the same period of time. These transcripts were also studied with regards to co-expression network connections. Eleven previously annotated and two novel long non-coding RNAs were present in two significantly disease correlated co-expression groups that were further profiled with respect to network properties, Gene Ontology terms and MetaCore© connections.

Keywords

LncRNA, lincRNA, abdominal aortic aneurysm, differential expression, co-expression, RNA-seq, WGCNA

Supervisors

Bengt Sennblad
Karolinska Institutet

Scientific reviewer

Magnus Lundgren
Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **80**

Biology Education Centre Biomedical Center Husargatan 3, Uppsala
Box 592, S-751 24 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Differential and co-expression of long non-coding RNAs in abdominal aortic aneurysm

Joakim Karlsson

Populärvetenskaplig sammanfattning

Aneurysm (uppsvällning) i bukaortan är en allvarlig sjukdom förenad med hög dödsrisk. Kända riskfaktorer är, liksom vid många andra hjärt- och kärlsjukdomar, bland annat rökning, högt blodtryck och ålderdom. Tillståndet är oftast svårt att upptäcka i tid. Undantaget riskabla kirurgiska ingrepp finns få behandlingsmöjligheter. Kartläggningen av sjukdomens genetiska bakgrund har tagit formen av ett eget forskningsfält. Den ökade kunskap som det här fältet bidrar med förväntas underlätta utvecklandet av nya diagnostikmetoder och terapeutiska tekniker.

Som en del i denna strävan har det här arbetet fokuserat på uttrycksmönster hos så kallade långa ickekodande RNA (lncRNA). Detta är en typ av RNA-molekyler som inte kan översättas till proteiner. Det var sedan tidigare känt att ickekodande RNA kan spela en roll i hjärt- och kärlsjukdomar.

Studiens upplägg baserades på sekvenserat RNA hämtat från ett antal möss som behandlats på så vis att en grupp utvecklat aneurysm i bukaortan, medan en kontrollgrupp inte gjort det. Målet var därmed att undersöka förändringar i nivåer av lncRNA mellan dessa två grupper, för att på så vis finna lncRNA-varianter med eventuell betydelse för sjukdomen. Arbetet identifierade därigenom ett antal lncRNA-kandidater, vars beteenden pekade på kopplingar till aneurysmutvecklingen, som rekommenderas för vidare forskning.

Examensarbete 30 hp
Civilingenjörsprogrammet Molekylär bioteknik

Uppsala universitet, september 2014

Contents

Abbreviations

AAA	Abdominal aortic aneurysm
cDNA	Complementary DNA
FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million mapped reads
GO	Gene Ontology
lincRNA	Long intergenic noncoding RNA
lncRNA	Long noncoding RNA
MDS	Multidimensional scaling
miRNA	Micro RNA
ncRNA	Noncoding RNA
ORF	Open reading frame
PCA	Principal component analysis
PCR	Polymerase chain reaction
RABT	Reference annotation based transcriptome [assembly]
rRNA	Ribosomal RNA
ROC	Receiver operating characteristic
WGCNA	Weighted gene co-expression network analysis

1. Introduction

Long non-coding RNAs (lncRNAs) share many similarities with messenger RNAs (mRNAs): They can be differentially spliced and a 5'-cap and poly-A-tail is often added. However, they lack the ability to be translated into protein (since they either have no, or a too short, open reading frame), hence "non-coding". However, noncoding does not equal nonsense. Interference with both the transcriptional and translational machinery in many different ways has been implied [1]. Some lncRNAs have antisense binding properties and some are spliced into several small- and microRNAs that in turn may have various biological functions. However, in the vast majority of cases, their activity is simply unknown.

Studying the expression patterns of lncRNAs in a particular disease may reveal novel functional connections and expand the mechanistic knowledge of the illness. In this project a particular kind of aneurysm development was studied. Abdominal aortic aneurysm (AAA) is a dilation of a region in the aorta, which may lead to rupture. The disease is asymptomatic throughout most of its progression, yet does very often have a fatal outcome. The most important factors influencing AAA appear to be smoking and hypertension [2], genetics is also known to play a role [2, 3]. Previous studies have profiled protein coding genes that are thought influence the condition, and also implied the involvement of microRNAs [3]. lncRNAs, known to be associated with several other diseases (among them cardiovascular ones [4]), and a fascinating biological phenomenon in itself, also offer to be a field worth exploring in the context of AAA.

At our disposal was a set of RNA-sequenced tissue samples from AAA mouse models. The task was to explore this dataset with respect to the expression of both previously known and novel lncRNAs. Any promising candidates may be subject to further functional evaluation in future RNA interference experiments.

2. Background

2.1 Long non-coding RNAs

Long non-coding RNAs (lncRNAs) are defined as RNA molecules longer than 200 bases that do not have the ability to code for any protein. A subclass of these are called "lincRNAs", where the "i" stands for "intergenic". LincRNAs reside in the regions between protein-coding genes, whereas lncRNAs in general may overlap such genes. Some protein-coding genes, for instance, have long transcript isoforms that are non-coding [1].

Intriguingly, lncRNAs have been found to be at least as abundant as the protein-coding genes expressed by the human genome (as well as the mouse genome) [1]. However, in general their expression levels are lower the those of protein-coding genes. Patterns of expression are also highly variable between tissues [5].

LncRNAs are spliced and processed much like regular protein-coding transcripts. Addition of poly-A tails is not uncommon. In several cases, lncRNAs are spliced into small RNAs, which in turn may have their own biological functions. It also appears that lncRNA genes generally are less conserved than protein coding ones [5].

Evidence has shown that lncRNAs can act by epigenetically altering gene expression. For instance, the lncRNAs Hotair, Kcnq1ot1 and RepA all interact with chromatin to cause gene silencing [1, 4]. According to one estimate, about a third of all lincRNAs interact with chromatin-modifying complexes in order to modify the functionality of various cellular processes [1]. It is also not uncommon that lncRNAs are transcribed antisense to protein-coding genes [5], so one may assume that the sequence complementarity could offer a way to interact with these other genes as well.

2.2 Abdominal aortic aneurysm

Abdominal aortic aneurysm (AAA) is a severe disease in which a focal region of the aorta is dilated. The condition is highly fatal upon aneurysm rupture. Generally, no symptoms are shown leading up to the critical point of the illness. Known risk factors are smoking, hypertension, high age, male gender and caucasian ethnicity [2]. The complex genetic aspects of the disease is an active research area. Since surgery is the only currently employed treatment [3], it would be desirable to discover new and less risk-filled venues of therapy.

The processes behind the formation of aneurysms can briefly be described as a complex interplay of vascular inflammation, dysregulation of the behaviour and proliferation of vascular smooth muscle cells, fragmentation of elastin and degradation

of other components of the arterial wall such as the extracellular matrix [3].

With respect to lncRNAs, a long non-coding RNA known as ANRIL ("antisense non coding RNA in the INK4 locus") has been indicated in human atherosclerosis (as well as other serious conditions such as diabetes and cancer) [4]. Similarly to lncRNAs such as Hotair, it binds to polycomb proteins and causes epigenetic modifications. This in turn leads to altered expression of other genes (CDKN2B in this case). No homolog to ANRIL is currently known to exist in mice, however. Nevertheless, it demonstrates the potential of lncRNAs to have key roles in the pathogenesis of diseases related to AAA.

In addition, another class of non-coding RNA, microRNA (which have also been shown to have gene regulatory function in many cases), have been indicated in AAA formation. Two prominent examples are miR-21 and miR-29b [3]. This further supports the idea that the role of the untranslated transcriptome should not be underestimated.

2.3 ApoE knockout mice

The data used for this work was provided by a group at Stanford University. It had already been RNA sequenced by a company named Centrillion, using an Illumina Hiseq 2000, the EpiBio ScriptSeq v2 library preparation kit and the RiboZero rRNA depletion kit. The sequenced tissue samples were derived from a set of mice modified to lack the gene for apolipoprotein E (ApoE). These are common as model animals in the study of atherosclerosis. Treatment with the blood pressure-regulating protein angiotensin II (AngII) leads to development of abdominal aortic aneurysms[6] (the cases), whereas saline treatment does not (the control group). 12 samples were used, all of them 12 weeks old: four mice harvested three and seven weeks after treatment with AngII infusion (the cases), respectively, and four models treated with saline and harvested after seven weeks (the controls). The sample groups will henceforth be referred to as "A3" and "A7" for the case mice (respectively) and "S7" for the control mice.

2.4 Sequencing and transcript assembly

2.4.1 RNA sequencing

RNA sequencing is a method used as a for quantifying gene expression through estimation of transcript abundance (unless stated otherwise, the term "expression" will refer to "transcript abundance" in this work, rather than to protein expression levels). Transcribed material contained in samples is first fragmented, converted to complementary DNA (cDNA) by reverse transcription, amplified using the polymerase chain reaction (PCR) and tagged with appropriate adapter sequences. The common term for this procedure is "library preparation". Subsequently to this preparation, the fragment library is analysed with any of a number of sequencing machines available. In this case, an Illumina Hiseq 2000 had been used. These instruments read a certain number of nucleotides from the ends of each fragment. The output consists of millions of small nucleotide sequences known as "reads" [7]. In this case, the read length was 100 bases.

Reads can be paired on unpaired, depending on the chosen technology. Paired reads, or "paired end reads", result from copies of the same cDNA fragment being sequenced from each end independently. The use of paired reads is helpful when dealing with repetitive transcripts. Since the distance between these reads can be computationally inferred, this information can be used when assigning the reads to positions on a reference genome.

Library preparation can also be done such that directionality of the fragments is preserved for downstream analysis. This is known as stranded library preparation. Preserving strand information makes it possible to tell whether a transcript originated from the sense of antisense strand of DNA. This information is particularly useful when dealing with noncoding fragments, since these may have regions that are partially complimentary to protein coding transcripts. This may, for instance, be one way in which they could regulate said transcript.

2.4.2 Read mapping

In order to find out which regions of the genome were expressed, it is necessary to map the RNA sequencing reads to an appropriate reference genome. Such reference genomes are available from public databases maintained by, for instance, the National Center for Biotechnology Information (NCBI) [8], University of California Santa-Cruz (UCSC) [9], and the European Bioinformatics Institute (EMBL-EBI) [10].

Several algorithms have been developed to perform mapping of RNA-seq reads. One of the most cited to date is TopHat, developed by Trapnell et al. [11]. The latest version of this software, TopHat2 [12], was used in this project. In short, this aligner works by performing a preliminary local alignment of reads (using the Bowtie2 short read aligner), whereafter consensus regions are assembled and possible splice junctions are generated. After this, reads are mapped anew, taking the splice junctions into account. It is also possible to provide a reference database of known transcripts in order to first map reads to the transcriptome (the set of all known transcripts), thereby improving accuracy of alignment.

2.4.3 Transcript assembly

After reads have been mapped to a genome, they will need to be assembled into fragments in order to discover the original structure of the expressed RNA transcripts. In order to do this, the Cufflinks program, by Trapnell et al. [13], was used. Briefly, the algorithm works by first grouping short reads into consecutive coherent units according to the positions to which they were mapped in the genome. This results in the reconstruction of a preliminary set of transcript fragments. Then, overlaps between these preliminary fragments are found in order to determine what the structure of the original transcripts may have been. Often, this results in a number of conflicting or spurious reconstructed transcripts.

Therefore, a second step is undertaken to find out which of these were most likely to be the correct transcripts. Mutually incompatible fragments are isolated and with the help of a graph theoretic approach, the minimal set of transcripts that can explain the particular fragment groups obtained is sought. This yields a final set of candidate transcript variants (isoforms), whose abundance levels are subsequently

estimated with the use of a likelihood-based statistical model.

A reference transcriptome (set of transcripts) can be used to aid in the estimation of expression levels and to improve accuracy of transcript reconstruction. In particular, a useful method when searching for novel isoforms is an algorithm known as Reference Annotation Based Transcriptome assembly (RABT) [14]. This approach mixes the true sequencing reads with so called "faux reads" generated from the transcriptome reference. The faux reads are supposed to compensate for coverage gaps and provide structural information for use in the reconstruction of transcript variants.

In general, expression levels may be expressed simply in terms of the number of reads mapped to each fragment. But another common way is to normalise the read count with the length of the fragment, resulting in a measure known as "fragments per kilobase per million mapped" (FPKM). The argument for doing so is that RNA-seq is prone to a fragment length bias: since longer fragments generate more reads than shorter fragments, more reads will map to them and they are more easily quantified. It has been found that longer transcripts are more often called differentially expressed than short ones, for this reason. Dividing by the length of the fragment is supposed to help dealing with this problem [13].

2.5 Prediction of novel long noncoding RNAs

Finding novel lncRNAs involves computationally inferring the coding potential of putative transcripts. This can be done in many ways, but the focus here was on the following filtering pipeline: First, transcripts from the novel assembly that were not found in the Ensembl gene annotation reference database [10] were extracted. Then, all transcripts shorter than 200 nucleotides were removed. After that, only candidates without very long open reading frames (ORFs) were kept. The next filtering step was based on an evolutionary inference of protein coding potential from phylogenetic codon substitution frequencies. The transcripts remaining after that were compared against the Pfam protein family database, in order to make sure that none of them resemble known protein domains (functional regions). This strategy was adopted from Sun et al. [15], and follows their "lncRscan" pipeline.

2.6 Functional characterisation

2.6.1 Differential expression testing

Detecting differential expression (significant differences in the number of RNA transcripts generated from genes in the comparison of two or more sample groups) requires sophisticated statistical methods. Several problems need to be dealt with: accounting for various types of bias (such as the fragment length bias discussed earlier), modelling and accounting for the biological divergence of the samples (natural differences, not due to the treatment being studied, among samples of the same condition), finding the significance level of any observed differences, etc. Multiple tools exist for this purpose. The most popular ones at the time of writing, specifically made for RNA sequencing, include DESeq [16], EdgeR [17] and Cuffdiff [18].

The Cuffdiff tool was chosen for this project, since it was specifically designed

to provide accurate estimates of differential expression on an isoform (transcript variant, as opposed to entire gene) level. This is important, since lncRNA transcripts are commonly found to be non-coding isoforms of protein-coding genes. Another benefit of using Cuffdiff over competing tools was the tight integration with the other tools of the so called Tuxedo pipeline, TopHat and Cufflinks [19].

2.6.2 Co-expression network inference

In a transcript level co-expression network, the nodes are transcripts and the edges distances between them. The distance measure used is based on the similarity of the expression profile of each transcript to that of every other transcript (the profile is the set of expression levels of a given transcript across all samples). The problem can thus be seen as that of determining the distance between a number of vectors, and finding out if any transcript form certain groups ("clusters" or "modules") of similar expression. There are many ways of measuring the similarity between a pair of vectors, and it may not be immediately obvious which is best with regards to gene expression. Common (though simple) ways of calculating this distance are the Euclidean norm and the absolute Pearson correlation.

There are also many methods of using these distance measures to find clusters among these vectors (transcripts). The most common are probably k-means [20] and hierarchical clustering[21], and modifications thereof. Others include Weighted Gene Co-expression Network Analysis (WGCNA) [22], Non-negative matrix factorisation (NMF) [23, 24] and Markov clustering [25]. All clustering methods are expected to reveal different structures in a given dataset. Some of these structures may be shared between the methods, but just because a given structure does not appear with the use of all methods does not mean it is not real. It could be a true and biological relevant feature of the data, which is only exposed by that particular model. But there is not guarantee that it is not just an artifact of the clustering either.

Therefore, it is often beneficial to compare the results of different clustering methods on the data. Of course though, some clustering measures are simply more suited to some types of data than others, and a lack of consensus between two given methods could just as well be because both methods perform badly, as it could be due to one of the methods performing well and the other not. It can be hard to tell sometimes, so clustering results will need to be additionally validated using biological information gathered from literature.

2.6.3 Reference Databases

A number of reference databases was used in this study:

NONCODE

NONCODE is a comprehensive database of noncoding RNAs, including lncRNAs, collected by literature mining and from other specialised databases. The database is developed and maintained by the Chinese Academy of Sciences. The number or entries in the database at the time of writing (version 4) was over 210 000 (all species included) [26]. The database was available to download on request, for free. NONODE was used to identify previously annotated lncRNAs in the newly assembled transcriptome.

Gene Ontology

The Gene Ontology (GO) [27] defines and hierarchically organises standardised biological concepts. This enables a systematic categorisation of gene and protein functionality. Three separate main ontologies are defined by the Gene Ontology consortium (the body which created and maintains the GO): "Biological process", "Molecular function" and "Cellular component". Molecular function refers to the various ways a gene product can act biochemically. Examples include "enzyme" and "receptor ligand". Such molecular functions may contribute to what is defined as biological processes, general paths towards a biological goal. Biological processes may be "translation" or "signal transduction". Cellular component refers to the location of a gene product in the cell.

One of the goals of creating this structured and well-defined set of terms was to make it easier for researchers to both unambiguously assign functional annotation to any given gene, and make queries on these terms to quickly find the gene participating in any process of interest [27]. Another useful aspect, which was utilised in this project, was the possibility of functionally annotating a given set of genes. That way, it could for instance be discovered if co-expressed clusters of transcripts are enriched for any particular biological functions.

MetaCore©

MetaCore© [28] is a commercial tool by Thomson-Reuters that is based on a curated collection of interactions between "network objects" such as genes, proteins and nucleic acids. The use of this tool was expected to be a valuable complement to Gene Ontology enrichment, when assessing the biological significance of co-expression network modules.

3. Methods

3.1 Read quality control

The quality control software FastQC [29] was used to inspect sequencing reads with regards to overall sequencing quality aspects. The version utilised was 0.10.1, with default options. The purpose of using this tool was to find out if any overrepresented reads or large quality differences among samples was present.

RSeQC [30] (version 2.3.9) was further employed to profile the reads with respect to mapping aspects. In particular, an estimation of "mate inner distance" was made with this program, based on preliminary alignments. The mate inner distance is defined as the distance (in nucleotides) between the two reads in a pair, with respect to where they have mapped on the genome. The value of this distance, and its standard deviation, can be specified as optional parameters during mapping with TopHat. This software also has the ability to infer experiment type, that is, stranded or unstranded sequencing.

3.2 Read mapping

Read mapping was performed with TopHat 2.0.11. The option "--no-mixed" was used to constrain the set of accepted reads to those where both reads in a pair could be mapped. This rather strict option was used in order to minimise the amount of assembled fragment artifacts in downstream analysis steps, of particular importance when aiming to discover novel transcripts. Another potentially helpful restriction on read mapping could be to also use the "--no-discordant" option, which disqualifies reads in which the members of a read pair map to different chromosomes. This was not used, however, since it caused the current version of TopHat to crash.

In addition, a reference annotation file in GTF format was provided (option "-G"), as it can increase the accuracy of alignment to take into account both the genome and the known transcriptome when mapping [11, 12]. Library type "fr-secondstrand" was specified, as the library preparation was stranded and originated from the second strand of cDNA synthesised form the RNA fragments. The values of "mate inner distance" and "mate-std-deviation" were provided for each sample as they had been estimated by RSeQC on preliminary mapping runs (plots can be found in the supplementary material). The version of the short read aligner Bowtie used by TopHat was 2.2.2.

The reference transcript annotation that was provided to TopHat was the Ensembl [10] mm9 reference annotation (the reason to go with mm9 for the main analysis, rather than mm10, was that the current noncoding RNA reference database NONCODE was only available based on mm9) provided by the Illumina iGenomes

project,which had been specifically made to be suitable with the TopHat and Cufflinks software ("NCBIM37", accessed April 11, 2014).

3.3 Transcript assembly

Transcript assembly was performed with Cufflinks 2.2.0 [13]. The "-g" option was used in order to make use of Reference Annotation Based Transcriptome (RABT) assembly, which is beneficial when searching for novel transcripts [14, 19]. The "fr-secondstrand" library type was also specified.

A run of Cufflinks in *ab initio* mode (default parameters, except for "--library-type=fr-secondstrand"), was also made in order to estimate the FPKM distributions of partially assembled fragments and artifacts in comparison to the distributions of assembled fragments that completely matched the reference (more on this under the section on prediction of novel lncRNA).

Both RABT and *ab initio* assembly was performed on the merged set of all alignments, excluding two samples that was deemed outliers (these samples had considerably worse mapping performance, different overall FPKM distributions and formed their own groups during principal component analysis (PCA) and multidimensional scaling (MDS) analysis; the PCA and MDS analyses were performed with the CummeRbund version 2.6.1 R package [19] on preliminary differential expression results (not shown)).

All transcriptome assemblies were annotated with the Cuffcompare [13] software included in the Cufflinks package (utilising the options "-G", "-C" and "-r"). The reference used was the same Illumina iGenomes provided Ensembl [10] mm9 annotation as during the alignments. During the comparisons, Cuffcompare assigns so called "class codes" to the assembled fragments. These indicate the nature of any overlap to reference transcripts that may be found. The meaning of the class codes is summarised in table 3.1.

Table 3.1: Cuffcompare class codes mentioned in this work. The definitions have been cited from the Cuffcompare manual.

Class	Meaning
=	"Complete match of intron chain"
c	"Contained"
j	"Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript"
i	"A transfrag falling entirely within a reference intron"
o	"Generic exonic overlap with a reference transcript"
x	"Exonic overlap with reference on the opposite strand"
u	"Unknown, intergenic transcript"

3.4 Identification of known long noncoding RNAs

Identification of known lncRNAs was accomplished through the use of the Cuffcompare software (version 2.2.1) included in the Cufflinks package [19]. Provided to the

algorithm was a reference annotation file in GTF format downloaded from the NONCODE database (version 4) [26], together with the RABT assembly of transcripts (besides this, default options were used). Comparisons were made twice, once with the NONCODE annotation as reference ("-r" flag) and the merged assembly as primary input; once with the merged assembly as reference (-r) and the NONCODE annotation as primary input. Then the intersection between the output mapping files were taken. This eliminated cases where the matching was incomplete.

After this, annotated lncRNAs were extracted from the Ensembl reference annotation (the categories "3prime_overlapping_ncrna", "ambiguous_orf", "antisense", "lincRNA", "ncrna_host", "non_coding", "processed_transcript", "retained_intron", "sense_intronic", "sense_overlapping", which according to Ensembl's terminology denote lncRNAs) and the union of the Cuffcompare NONCODE intersection set and the Ensembl lncRNA set was taken. The result was the set of all lncRNAs present in the novel transcriptome assembly, which had been previously annotated by either NONCODE or Ensembl.

3.5 Prediction of novel long noncoding RNAs

The general outline of the search for novel long non-coding RNAs was adapted from Sun et al. [15], whose scripts included in the "lncRscan" pipeline were employed in the execution of some of the filtering steps described below.

Filtering of partially assembled transcripts

In order to reduce the risk of making false predictions, fragments that were entirely contained by known annotated transcripts were filtered away together with other fragments suspected to be mere artifacts. The method for doing so was based on the hypothesis that the overall FPKM values estimated for such fragments would be lower than that of true transcripts. This method has been used previously by, for instance, Sun et al. [15]. A custom pipeline was developed for this purpose, using Perl and R.

Inference of transcript coding potential

There are many strategies for judging how likely a transcript is to be protein coding. The software PhyloCSF (Phylogenetic Codon Substitution Frequency) performs multiple alignments between several species and uses the substitution frequency of bases in a given transcript, throughout the phylogeny, to determine whether it is likely to code for a protein [31]. Another, Coding Potential Calculator (CPC) is based on a support vector machine (SVM) classifier that takes into account features such as alignment scores to protein databases and certain aspects of any open reading frames (ORFs) that are found within the transcripts. iSeeRNA [32], included in the Sebnif [33] lincRNA detection pipeline, is also a support vector machine based tool, which has been shown to compare favourably to CPC. PhyloCSF and iSeeRNA has been shown to perform better than several of the prominent tools, including CPC [31, 32]. For this project, PhyloCSF was used.

After coding potential has been inferred using any of the methods mentioned above, one might further validate the transcripts by searching for similarity to known

protein coding domains in Pfam, for instance. This was the strategy employed here.

3.6 Functional characterisation

If any novel lncRNAs are indicated by the data, it is of course also of interest to find out what, if anything, these transcripts may be doing. One approach to this is to see if they are differentially expressed in the studied dataset. Another, complementary, approach is to look for similarities in expression to known transcripts. This can be done by co-expression network inference. Transcript clusters obtained through the use of such network methods can additionally be mapped to Gene Ontology categories. The differential expression and network analyses are described in detail below. GO analysis was performed with the BiNGO plugin of Cytoscape 3.1.1 [34].

These approaches do not, however, give a final conclusion about what the transcripts are actually doing. Rather, the strategy should be seen as a way of generating hypotheses that can later be tested using, for instance, molecular biology techniques.

3.6.1 Differential expression testing

Transcript level differential expression testing was performed with Cuffdiff 2.2.1 [18]. The Cufflinks RABT assembly that had been performed on the merged BAM files of aligned reads from all samples (excluding outliers) was used as reference. Multi-hit correction was performed with the "-u" option, and the library-type "fr-secondstrand" was specified. Isoforms were considered significantly differentially expressed if they had a p-value less than 0.05 and a q-value less than 0.05. The "getSig" function of the cummeRbund R package (version 2.6.1) [19] was used to select those fragments, using 0.05 as the value of the parameter "alpha".

3.6.2 Co-expression network inference

Co-expression network inference was carried out with the WGCNA R package by Langfelder and Horvath [22]. This network construction method uses absolute Pearson correlations between gene profiles (defined as the expression levels of a gene across all samples). These correlations are then raised to a power β in order to emphasise high correlations more than low correlations. β is considered a soft thresholding power, an alternative to a hard cutoff threshold for forming network modules. This has been shown to aid in accomplishing a scale free topology [22], that is, a topology with a few highly connected nodes, rather than many approximately equally connected ones (a so called random topology). Scale free topologies are thought to better reflect biologically relevant information than random topologies [22]. The transformed correlation coefficients are then the basis for a hierarchical clustering of the genes, whereafter the resulting dendrogram is cut using an algorithm called "dynamic tree cut" [35], yielding clusters of genes (modules). Previous studies have applied this method mostly on microarray gene expression data [36, 37, 38], but also on RNA-seq data [37, 39, 40].

In order to run such an analysis within the resource constraints of the project, it was necessary to reduce the dataset to a reasonable level, as opposed to constructing a network based on all transcripts (over 700 000). The method employed to accomplish such a dimensionality reduction was a filtering based on the coefficient

of variation of each gene in the dataset, after log-transformation of all FPKM values. It was reasoned that more variable transcripts would contribute more valuable information for co-expression network inference. A similar strategy has been used previously by Kugler et al. [40]. Genes and transcripts that do not vary at all would per definition be considered co-expressed with each other, but would not contribute any new knowledge about AAA disease in addition to what is already revealed by the differential expression analysis. Therefore, the dataset was reduced to around 16600 genes (using a coefficient of variation threshold of 0.60).

In order to determine the value of the soft thresholding power β, iterative calculations were made, assessing how scale-free the network topology was for each value of the parameter. Values from 2 to 34, with interval 2, were tested. A value of 16 was found to give a sufficiently scale-free structure, and was accordingly applied for the network construction. The scale-free property of the network for the tested values can be seen in figure 4.5 in section 4.4.

For the dynamic tree-cut algorithm, a minimum module size of 30 was used, together with a "mergeCutHeight" of 0.25 and otherwise default parameters.

Furthermore, the correlation between the first principal component (the "eigengene") of the matrix of transcript profiles of each module and a binary vector of sample treatment information (using the values 1 for the A7 group and 0 for the S7 group) was calculated. This ranking was only performed on the basis of the mice harvested after seven weeks (the A3 group was excluded). This should indicate any relationships between the modules and the treatment.

3.6.3 Clustering comparison

It is not trivial to estimate the performance of an unsupervised learning algorithm on a dataset where no class labels are known beforehand. Rather, a common approach is to visually inspect the clusters and judge whether they seem to make sense biologically. Another is to compare clustering results of different algorithms to see if general conclusions can be drawn from any consensus that might be found between them.

One way of objectively doing the later is to calculate an overlap matrix. This matrix would contain cells representing the number of genes that any pair of clusters (one from each method) have in common. The matrix could then be seen as a contingency table, and common statistical approaches may be used to determine whether the two variables (the clustering methods) have any significant association. One association measure that could be used is Cramér's V. This approach yields a score between 0 and 1, where 0 represents no association whatsoever and 1 signifies perfect association (the two methods have given the exact same results --unlikely in practice). A good association between two clusters would give increased support to the common co-expression patterns identified therein.

There does not, however, seem to exist any universal consensus on what can be regarded as a "strong enough" association when dealing with the Cramér's V statistic. Some publications use values around 0.10 as a cutoff, others around 0.25. In order to get a sense of how the method would perform on two methods that yield largely the same results, a test was made where the k-means clustering method, was compared to itself in two runs with $k = 51$ clusters and n = 1 random starting points each. Usually, one uses hundreds of random starting points to compensate

for entrapment in local optima. Only one starting point was used here in order to provide some variability in the clustering results. The algorithm was applied to the coefficient of variation-filtered transcript dataset. The Cramér's V calculated based on this comparison was 0.76, indicating (as expected) a very high association between the two runs, illustrated in figure 3.1. In the figure, one can see that each module of the first clustering run only matches one other module well in the second clustering.

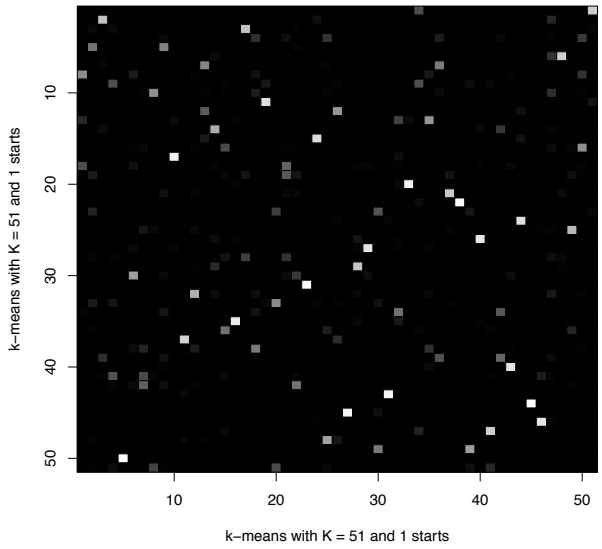


Figure 3.1: The overlap of two runs of k-means clustering on the coefficient of variation filtered dataset. In both cases, k set was equal to 1 and n = 1 starting points of the algorithm was used. The image is a visualisation of the contingency matrix, wherein high values of a given cell is displayed brightly and low or zero overlap between two given clusters is displayed darkly shaded. (Additionally, in this image the number of transcripts in the intersection of each pair of modules have been normalised by the total size of the modules for grater clarity.)

Besides this, a more biology-based cluster validation procedure was undertaken, whereby network modules were uploaded to the commercial knowledge mining tool MetaCore© [28] (version 6.19) in order to inspect the biological aspects of the seemingly co-expressed transcripts. Gene ontology enrichment was also performed for this reason.

4. Results

4.1 Read quality control

The FastQC [29] results of the twelve sequenced mouse samples (one from the A3 group and one from the S7 group) revealed that two of them deviated much in quality from the rest, and principal component analysis (figure not included) confirmed large inconsistencies in gene expression patterns compared to other samples in the same condition groups. These two samples were thus excluded from further analysis.

Warning flags were also raised by the program regarding overrepresented sequences. These were almost exclusively found to be mitochondrial in origin. This was initially not thought to be a problem, however (though, see the discussion in section 5). Patterns of bias at the ends of reads, common to RNA-seq, were also discovered. This bias was thought to be due to lack of sufficient randomness in the "random" hexamer primers used during library preparation [41].

Furthermore, adapter contamination was found to be present. The cause for this was likely the short average fragment size (around 170 nucleotides), which was revealed with the RSeQC [30] quality control software (see figure 4.1).

Too short fragment size means that the sequencing machine will read not just the ends of the fragment, but across the entire fragment and into the adapter ligated to the opposite side.

These issues might lead to fewer reads being mapped by TopHat, although it was initially not expected that they would have any serious effect on transcript assembly and downstream analysis (but see also section 5 for more on this topic).

4.2 Prediction of novel long noncoding RNAs

Filtering of partially assembled fragments

Figure 4.2a shows the (log transformed) FPKM distributions of *ab initio* assembled fragments that partially ("c") and completely ("=") match the reference annotation, respectively. The best threshold for predicting these transcript classes from FPKM values was found to be 1.03, with discriminatory performance according to the ROC curve displayed in figure 4.2b. Both the sensitivity and specificity was found to be quite low, perhaps owing to noisy data. The chosen filtering threshold 1.03 would therefore discard a lot of potentially true lncRNAs, while simultaneously retaining false positives. However, it was considered more important to filter out as many false positives as possible, and thus applying this particular threshold, than to do no filtering at all (and thus retaining more potentially true lncRNA candidates).

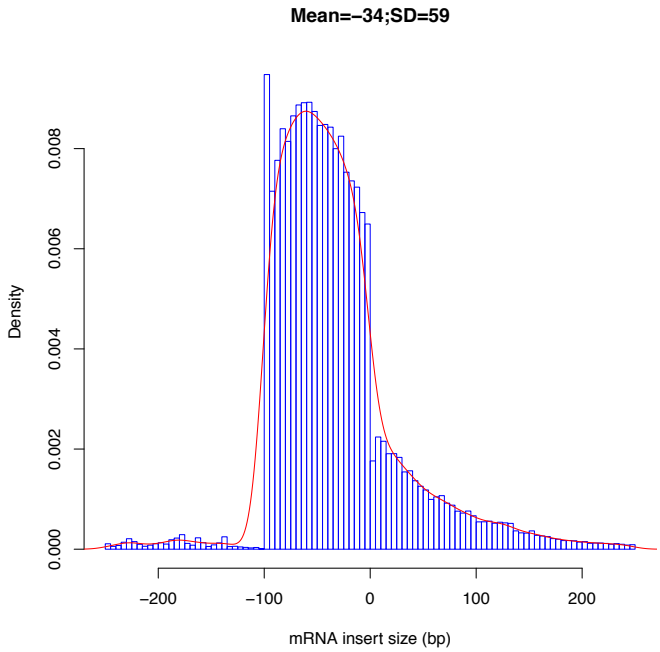
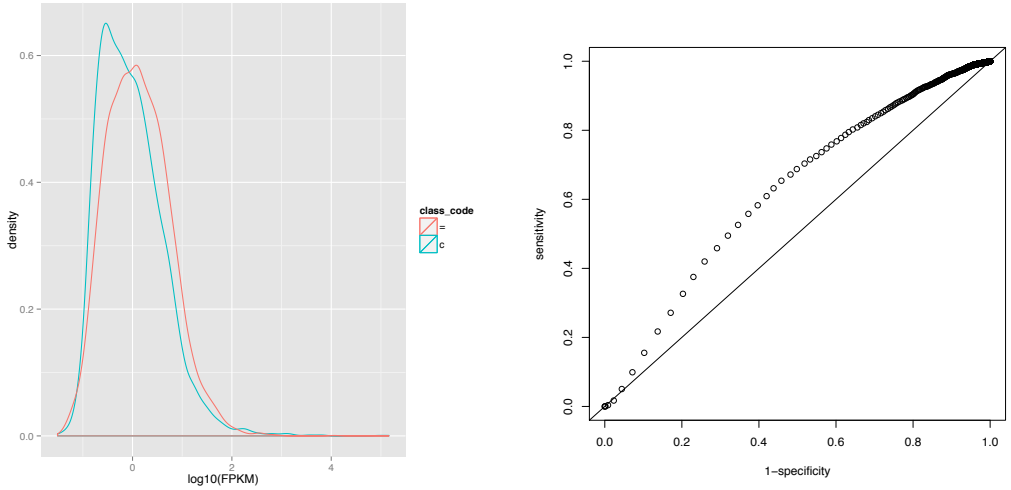


Figure 4.1: Inner distance of paired end reads in one of the samples. The plot was produced on one of the three AngII treated replicates that had been harvested after seven weeks using RSeQC 2.3.6, but is representative for all samples. A negative insert size indicates that reads mapped to the genome overlap with each other.



(a) ***(b)***

Figure 4.2: a) Log 10 transformed FPKM-distributions of fragments that partially ("c") and completely ("=") match the reference annotation. b) ROC curve showing the discriminatory performance of each threshold when classifying transcripts as either partially of completely assembled. The optimal threshold should correspond to the point closest to the upper left corner of the coordinate system (1,0).

Identification of long noncoding RNAs

29 candidate lncRNAs were obtained subsequent to filtering with PhyloCSF and matching against Pfam. Intersecting these candidates with NONCODE v4 revealed an overlap of three transcripts. Furthermore, the genomic coordinates of these 29 candidates were translated (using the LiftOver tool from UCSC [42]) and matched (using Cuffcompare) against the most recent reference genomes (mm10) from Ensembl, RefSeq and UCSC. Only the three lncRNAs that also matched NONCODE were present on any these updated annotations. The remainder were thus 26 potentially novel lncRNAs (table 4.1). Of these, three did not overlap any known reference transcript (class code "u") and could thus be long intergenic non-coding RNA (lincRNA).

Table 4.1: Overview of the 26 novel lncRNA candidates. Nearest reference genes, transcripts and class codes were assigned by Cuffcompare, using the Ensembl mm9 annotation as reference. Isoform IDs are the internal ones assigned to the transcripts by Cufflinks.

Isoform id	Closest gene	Nearest transcript	Class	Length	Locus
TCONS_00026197	Obfc2a	ENSMUST00000027279	j	4944	1:51522798-51535243
TCONS_00037991	Rasal2	ENSMUST00000078308	i	1775	1:159065313-159350366
TCONS_00038774	Mpzl1	ENSMUST00000111435	j	1460	1:167522311-167564672
TCONS_00053753	Pawr	ENSMUST00000095313	i	788	10:107769244-107863779
TCONS_00075759	Ebf1	ENSMUST00000081265	i	996	11:44428427-44821593
TCONS_00076304	Mgat1	ENSMUST00000101293	j	3676	11:49057692-49076532
TCONS_00084766	-	-	u	10332	11:104790337-104801947
TCONS_00095737	Tnk1	ENSMUST00000108631	j	2348	11:69659877-69672232
TCONS_00117924	Ubxn2a	ENSMUST00000020962	j	6341	12:4881930-4914511
TCONS_00121122	Tspan13	ENSMUST00000020896	j	1100	12:36741143-36769087
TCONS_00124446	Sos2	ENSMUST00000035773	i	233	12:70684747-70782839
TCONS_00134561	Gcnt2	ENSMUST00000067778	i	591	13:40955500-41108388
TCONS_00190470	Gpihbp1	ENSMUST00000023243	j	577	15:75426921-75432461
TCONS_00210492	Lpp	ENSMUST00000038053	i	1118	16:24391697-24992662
TCONS_00275488	Gnaq	ENSMUST00000025541	i	5807	19:16207320-16478609
TCONS_00298653	Cstf3	ENSMUST00000028599	j	14027	2:104430679-104552326
TCONS_00307567	-	-	u	615	2:171483568-171489341
TCONS_00355809	Ccnl1	ENSMUST00000154585	j	1102	3:65750072-65762171
TCONS_00385472	-	-	u	335	4:155610115-155610534
TCONS_00445558	Ubn2	ENSMUST00000160583	j	15372	6:38383924-38474825
TCONS_00451604	Slc6a6	ENSMUST00000032185	i	1553	6:91634060-91709057
TCONS_00455818	Clec2d	ENSMUST00000032260	j	817	6:129112594-129136552
TCONS_00479043	Lsm14a	ENSMUST00000085585	i	237	7:35129664-35179798
TCONS_00492246	Lsp1	ENSMUST00000105968	i	2021	7:149646713-149701914
TCONS_00559428	Pts	ENSMUST00000034570	o	1688	9:50329722-50336746
TCONS_00583336	4930468A15Rik	ENSMUST00000114054	i	1108	X:73827948-73848263

By comparison to the NONCODE v4 database and Ensembl mm9, 30490 previously annotated long non-coding RNA were additionally detected from the set of RABT assembled fragments. A relative lack of small non-coding RNAs (and miRNAs) was observed. The reason was thought to be due to greater technological difficulties involved in the detection of short transcripts. Normally, small- and micro RNA studies of RNA-seq requires a specialised type of library preparation in order to be feasible [43].

4.3 Differential expression testing

The total number of significantly differentially expressed transcripts (both coding and non-coding) was 8801 (out of 642012 assembled transcripts in total). In the comparison between between AngII treated mice and control mice harvested after seven weeks (A7 and S7, respectively), the number was 6912.

Of the previously known lncRNAs (annotated in NONCODE v4 and Ensembl mm9), 246 were found to be significantly differentially expressed (p-value less than 0.05 and using a q-value cutoff of 0.05). 185 of those were between A7 and the S7 mice. Among the 26 novel lncRNA candidates, two were significantly differentially expressed between A7 and S7 (table 4.4).

The 35 overall most significantly differentially expressed genes (based on individual isoforms from those genes) between A7 and S7 are presented in table 4.2 (novel genes have been excluded). Similarly, the 35 most differentially expressed long noncoding RNAs are presented in table 4.3.

Some of the genes in table 4.3 have infinite fold change. This only means that they were exclusively expressed in one of the sample groups. One of those was Gm12245, which was a previously predicted pseudogene, whose expression was only observed in the case mice. Given that it is 357 bp long and noncoding, it would be classified as a lncRNA. Not much else is known about it however.

Three of the most differentially expressed genes (infinite fold change) were also Hist1h4j, Hist1h3i and Hist1h2ah, which all code for histone cluster proteins. Histone proteins, important in the context of chromosomal structure, tend are upregulated during DNA replication [44]. Higher expression of histone proteins may therefore be a sign of increased cellular proliferation.

Two other transcript exclusively expressed in case mice was SCARNA17, which stands for Small Cajal body-specific RNA 17, and A930005H10Rik. Very little appears to be known about SCARNA17, other than the fact that Cajal body-specific RNAs seem to play a role in the spliceosomal machinery [45]. The ENSMUST00000067500 isoform of the A930005H10Rik is a 309 bp long processed transcript which does not code for any protein. It is expressed antisense to the gene Dph5.

Among the genes differentially expressed with finite fold change, H19 was the one that deviated most between case and control. H19 encodes a long noncoding RNA and is shown to interact with Igf2 (Insulin-like growth factor 2) via a chromatin remodelling based epigenetic switch [46]. See also section 4.4.

Col11a1 and Col12a1 encode two types of collagen (type XI and XII). Collagens are important constituents of the extracellular matrix. The extracellular matrix (and possible degeneration thereof) is an important part of the arterial wall and plays a large role in the mechanism of AAA disease [3].

Two of the other most differentially expressed genes were Thbs1 and Thbs4, which encode thrombospondin proteins. According to Gonzalez-Quesada et al. (2013) [47], thrombospondin interacts with the extracellular matrix and "is a prototypical matricellular protein that is not part of the normal cardiac matrix network, but is upregulated in cardiac remodeling because of myocardial infarction or pressure overload". Thrombospondin has been found previously to be associated to AAA [48].

Itm2a (integral membrane protein 2A), in turn, encodes a transmembrane protein

Table 4.2: The overall 35 most differentially expressed genes (based on individual isoforms) between AngII treated mice harvested after seven weeks (A7) and control mice harvested after seven weeks (S7). Fold change, p- and q-values (the latter accounting for multiple testing) were calculated with Cuffdiff 2.2.1. A negative fold change indicates down-regulation in the control group (up-regulation in the AngII-treated group). "Class" refers to the class codes detailed in table 3.1.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000117266	Gm12245	=	-Inf	0.00005	0.00265555
ENSMUST00000087714	Hist1h4j	=	-Inf	0.00005	0.00265555
ENSMUST00000099704	Hist1h3i	=	-Inf	0.00005	0.00265555
ENSMUST00000091742	Hist1h2ah	=	-Inf	0.00005	0.00265555
ENSMUST00000158610	SCARNA17	=	-Inf	0.00005	0.00265555
ENSMUST00000067500	A930005H10Rik	=	-Inf	0.00025	0.01074100
ENSMUST00000158415	RNaseP_nuc	=	Inf	0.00065	0.02343650
ENSMUST00000150405	Cpxm2	=	-Inf	0.00110	0.03559290
ENSMUST00000136359	H19	=	-7.36622	0.00005	0.00265555
ENSMUST00000140716	H19	=	-7.06716	0.00005	0.00265555
ENSMUST00000071750	Col12a1	=	-6.56065	0.00015	0.00697443
ENSMUST00000022213	Thbs4	=	-6.25172	0.00005	0.00265555
ENSMUST00000123619	Col11a1	=	-6.08950	0.00030	0.01248340
ENSMUST00000039559	Thbs1	=	-5.74065	0.00030	0.01248340
ENSMUST00000033591	Itm2a	=	-5.67131	0.00005	0.00265555
ENSMUST00000103095	Tnnc2	=	-5.60156	0.00005	0.00265555
ENSMUST00000039178	Tnn	=	-5.53919	0.00005	0.00265555
ENSMUST00000006626	Actn3	=	-5.50895	0.00015	0.00697443
ENSMUST00000001547	Col1a1	=	-5.42860	0.00005	0.00265555
ENSMUST00000055770	Hist1h1a	=	-5.37595	0.00005	0.00265555
ENSMUST00000170872	Thbs2	=	-5.35743	0.00025	0.01074100
ENSMUST00000080511	Hist1h1b	=	-5.32490	0.00005	0.00265555
ENSMUST00000031565	Fscn1	=	-5.21660	0.00005	0.00265555
ENSMUST00000025497	Fbn2	=	-5.17334	0.00005	0.00265555
ENSMUST00000032800	Tyrobp	=	-5.16206	0.00005	0.00265555
ENSMUST00000081035	Mpeg1	=	-5.05190	0.00005	0.00265555
ENSMUST00000119429	Myl1	=	-5.02694	0.00005	0.00265555
ENSMUST00000110455	Hist1h2bk	=	-5.01183	0.00075	0.02627970
ENSMUST00000027885	Angptl1	=	-4.98405	0.00010	0.00490988
ENSMUST00000003643	Ckm	=	-4.96161	0.00005	0.00265555
ENSMUST00000171470	Lox	=	-4.94639	0.00060	0.02200790
ENSMUST00000138511	Col1a2	=	-4.89817	0.00160	0.04753520
ENSMUST00000021231	Abcc3	=	-4.86676	0.00095	0.03168320
ENSMUST00000021506	Serpina3n	=	-4.85405	0.00005	0.00265555
ENSMUST00000005255	Wisp1	=	-4.78107	0.00005	0.00265555

Table 4.3: The overall 35 most differentially expressed lncRNA candidates between the case and control mice that had been harvested seven weeks after treatment (A7 and S7, respectively). Column headers as in table 4.2. Transcript ID:s starting with "ENSMUST" are Ensembl ID:s, those starting with "NONMMUT" refer to transcripts only annotated by NONCODE.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000117266	Gm12245	=	-Inf	0.00005	0.00265555
ENSMUST00000150405	Cpxm2	=	-Inf	0.00110	0.03559290
ENSMUST00000136359	H19	=	-7.36622	0.00005	0.00265555
ENSMUST00000140716	H19	=	-7.06716	0.00005	0.00265555
ENSMUST00000138511	Col1a2	=	-4.89817	0.00160	0.04753520
ENSMUST00000138145	Ccl6	=	-3.72700	0.00030	0.01248340
ENSMUST00000112866	Cxcl12	=	-3.22994	0.00020	0.00891552
ENSMUST00000145974	Col5a3	=	-3.11406	0.00145	0.04401550
ENSMUST00000108631	Tnk1	j	-2.98056	0.00155	0.04638690
ENSMUST00000131787	2410006H16Rik	=	-2.96827	0.00005	0.00265555
ENSMUST00000154058	Gm15639	=	-2.95309	0.00155	0.04638690
ENSMUST00000053085	Nlrc5	=	-2.92918	0.00005	0.00265555
ENSMUST00000164588	Gm17446	=	-2.89507	0.00015	0.00697443
ENSMUST00000162785	Ms4a7	=	-2.79884	0.00005	0.00265555
ENSMUST00000130179	Mtmr1	=	-2.74925	0.00030	0.01248340
ENSMUST00000147681	F630028O10Rik	=	-2.69458	0.00005	0.00265555
ENSMUST00000167285	Usf2	=	-2.66836	0.00020	0.00891552
ENSMUST00000139643	Pnpla7	=	-2.61396	0.00125	0.03928280
ENSMUST00000143673	AI662270	=	-2.59238	0.00005	0.00265555
ENSMUST00000140732	Ptov1	=	-2.57090	0.00090	0.03036040
ENSMUST00000128161	Rgs12	=	-2.49486	0.00020	0.00891552
ENSMUST00000159745	Tmem55b	=	-2.48982	0.00075	0.02627970
ENSMUST00000155083	Ppox	=	-2.47681	0.00005	0.00265555
ENSMUST00000132618	Eif1	=	-2.45017	0.00010	0.00490988
NONMMUT029052	-	=	-2.39013	0.00020	0.00891552
ENSMUST00000161573	Psme2	=	-2.38926	0.00010	0.00490988
ENSMUST00000128015	Plekhg2	=	-2.38253	0.00005	0.00265555
ENSMUST00000153077	Megf8	=	-2.37883	0.00010	0.00490988
ENSMUST00000165292	Gm17282	=	-2.36673	0.00005	0.00265555
ENSMUST00000137837	Lrrc42	=	-2.34961	0.00040	0.01584780
ENSMUST00000152710	Vgll4	=	-2.29871	0.00045	0.01747430
ENSMUST00000034997	Snhg5	=	-2.29561	0.00005	0.00265555
ENSMUST00000132340	Trem2	=	-2.28742	0.00005	0.00265555
ENSMUST00000149391	Gdi1	=	-2.27913	0.00005	0.00265555
ENSMUST00000146254	Cd300lf	=	-2.27148	0.00115	0.03682310

involved in muscle cell differentialion [49] and T-cell activation [50]. It has also been shown to interact with collagens in terms of its involvement in chondrogenic differentiation [51]. A role in AAA via effects on the extracellular matrix of the aortic wall is thus not unthinkable.

The novel differentially expressed transcript had the IDs TCONS_00095737 andTCONS_00559428. TCONS_00095737 was assigned by Cufflinks to the gene Tnk1 with the class code "j", signifying that the transcript shares at least one splice junction with a reference isoform. This means that is could be a novel alternative splice variant. TCONS_00559428 was closest to the Pts gene, sharing a "generic exonic overlap" (class code "o"). This simply means that that the transcript overlaps at least one exon with the reference transcript, but does not share any part of the splicing structure with the reference. It may be a flanking single-exonic transcript that overlaps one of the ends of the Pts gene (though, intuitively, it is more likely to be an artifact as compared to any "j" isoform found).

TCONS_00095737 was the only one of the isoforms of Tnk1 that was differentially expressed, as can be seen in figure 4.3. TCONS_00559428 showed a similar expression pattern as the other isoforms of Pts, only with lower overall expression (figure 4.4). This would perhaps suggest that the expression is a passive consequence of the main protein coding isoform expression, maybe due to inaccuracies of the splicing machinery. In mouse, only one isoform of this gene is described in the Ensembl annotation (the protein coding variant). In human, nine ones are annotated, four of them long noncoding. This raises the likelihood that this novel lncRNA could be a true transcript variant.

Table 4.4: Differentially expressed novel lncRNA candidates between the case and control mice that had been harvested seven weeks after treatment. Column headers as in table 4.2. The transcript IDs correspond to those in table 4.1.

Transcript ID	Closest gene	Class	$\log_2$ fold change	p-value	q-value
TCONS_00095737	Tnk1	j	-2.98	0.002	0.037
TCONS_00559428	Pts	o	-1.39	0.001	0.028

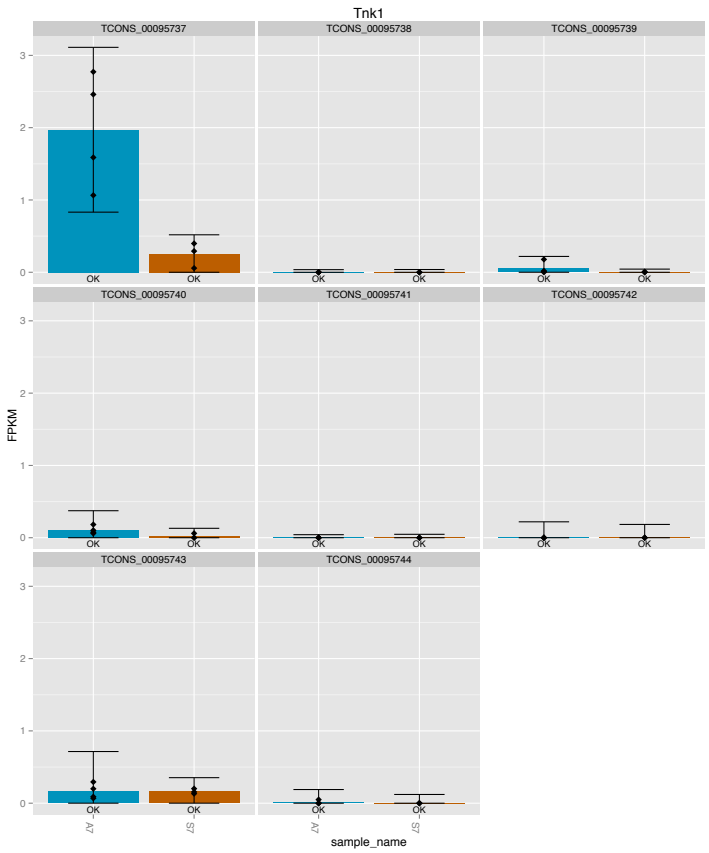


Figure 4.3: Expression of the isoforms of Tnk1, to which the novel transcript TCONS_00095737 was assigned. The blue bar shows expression in the A7 treated group and the orange bar shows expression in the S7 control group. Black dots are the expression levels from individual samples in each group and error bars are confidence intervals based on variance modelling calculated by CuffDiff.

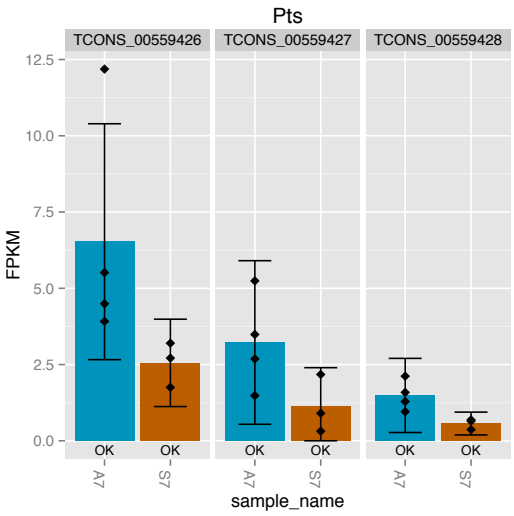


Figure 4.4: Expression of the isoforms of Pts, to which the novel long noncoding transcript TCONS_00559428 was assigned. TCONS_00559427 was also a novel transcript. Axis labels and figure elements the same as in figure 4.3.

4.4 Co-expression network modules

In order to perform *de novo* network inference with the WGCNA method, an optimal soft thresholding power needed to be calculated, as described in section 3.6.2. The choice of this thresholding power was made based on figure 4.5. A value of 16 was found to be appropriate with respect to desired network scale independence and (low) mean connectivity.

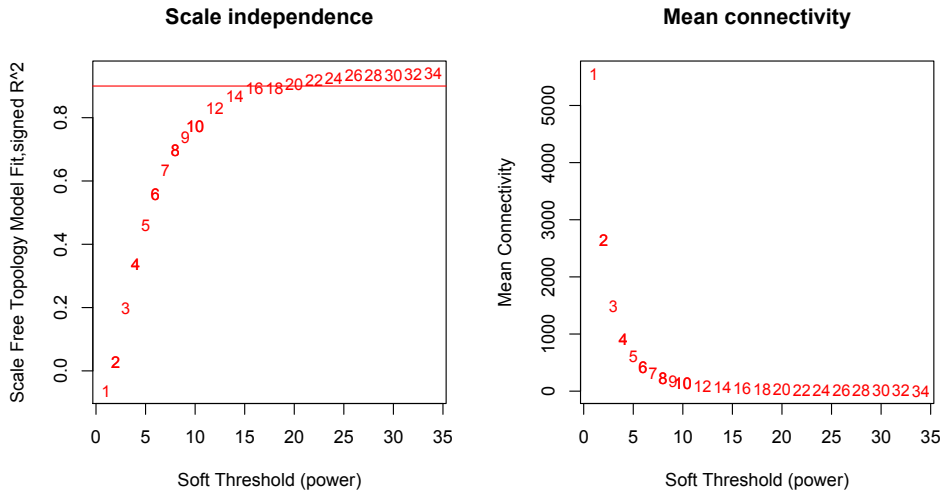


Figure 4.5: Left: Fit of the WGCNA network to a model of scale free topology as a function of the soft thresholding power. Right: Mean connectivity as a function of the soft thresholding power. The red line indicates where the gain in scale-free fit begins to level out. Red numbers indicate the tested soft thresholding powers.

The network construction made with the 16689 transcripts that remained after filtering yielded 51 modules. The dendrogram obtained after hierarchical clustering is shown in figure 4.6, together with the module assignments made with the application of dynamic tree cutting. The modules were assigned colours codes in this process (as shown in figure 4.6), which from now on will be used to refer to them. The colours are merely labels and do not reveal any other information about the modules.

The first principal component of the modules (the "eigengene", in the terminology of Langfelder and Horvath [22]) was correlated with a vector of treatment status as mentioned in section 3.6.2. The modules were subsequently ranked according to the absolute Pearson correlation coefficient obtained, also taking into account the p-value of the correlation.

Intersecting the set of significantly differentially expressed previously known lncRNAs with the transcripts in the most treatment-correlated modules revealed that 20 of those lncRNAs were present in any of those clusters (see table 4.5). Of the 26 novel lncRNA candidates, only two were left after coefficient of variation filtering. Of these, the intergenic TCONS_00385472 was assigned to the "red" cluster. The other one was TCONS_00190470, which was assigned to the "blue" module. Those of the top ranking modules that also contain interesting lncRNAs are shown in table 4.5.

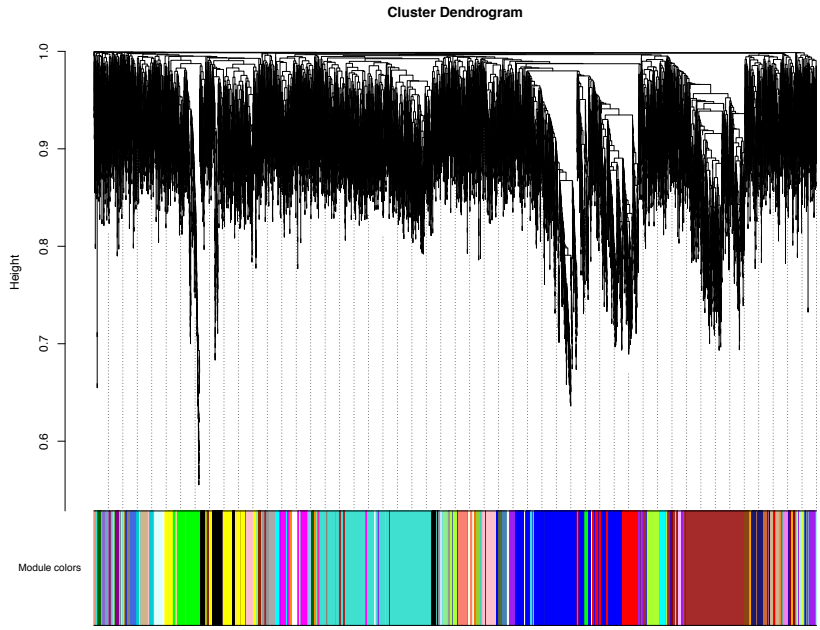


Figure 4.6: An illustration of the WGCNA module assignment process for this dataset. The black dendrogram is obtained through hierarchical clustering of the weighted correlation based network. Modules are represented by individual colours and are the result of applying dynamic tree cutting to the dendrogram.

Given those results, it would seem that two of the modules, "blue" and "red" justify further scrutiny. Partly because they contain both novel and differentially expressed lncRNAs, but also since they were found to be strongly correlated with the treatment vector (see table 4.5).

Table 4.5: Ranking of modules according to Pearson correlation coefficients calculated between the eigengene of the module and a vector of treatment status (see section 3.6.2). Each module is represented with a colour code that was assigned to it during the WGCNA procedure, and which corresponds to the colours in figure 4.13. Only the most significantly correlated (an absolute correlation of at least 0.80 and a p-value less than or equal to 0.05) modules which harbour either differentially expressed lncRNAs or novel lncRNAs are shown here. "Size" refers to the number of transcript isoforms in the module. 'D.e." stands for differentially expressed.

Module	Correlation	p-value	Size	D.e. lncRNAs	Novel lncRNAs
magenta	0.95	0.0004	538	1	0
turquoise	0.95	0.001	2745	2	0
green	0.88	0.009	755	3	0
blue	0.87	0.01	2252	9	1
black	0.82	0.02	628	2	0
red	0.80	0.03	678	3	1

The "blue" module

Nine significantly differentially expressed lncRNAs were detected in the blue module. They are presented in table 4.6. Besides those, a novel lncRNA was also present. It

did not, however, present any significant difference in transcript abundance among any of the sample groups. This lncRNA was denoted TCONS_00190470 in table 4.1 and appeared to be a novel noncoding isoform of the Gpihbp1 (glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1) gene. The main protein otherwise encoded by this gene acts as a lipoprotein lipase transporter in capillary endothelial cells [52].

Table 4.6: Interesting lncRNAs in the "blue' module. "Class" refers to the class code of the transcript (see table 3.1). "Groups" refers to which sample groups the difference was observed in. In all cases, p-values (as obtained from Cuffdiff) were less than 0.05 and the differential expression was considered significant after applying multiple testing correction (an alpha of 0.05 was used as threshold for the "getSig" function in the cummeRbund R package).

Transcript ID	Gene	Class	$\log_2$ fold change	Groups
ENSMUST00000132313	Hmg20b	=	-2.35, -2.21	A3, S7; A7, S7
ENSMUST00000132340	Trem2	=	-2.29	A7, S7
ENSMUST00000162785	Ms4a7	=	1.56, -1.24, -2.80	A3, A7; A3, S7; A7, S7
ENSMUST00000149366	Tmem48	=	-3.00	A3, S7
ENSMUST00000134637	Cdca3	=	2.42	A3, A7
ENSMUST00000140732	Ptov1	=	-2.57	A7, S7
ENSMUST00000136359	H19	=	-7.37	A7, S7
ENSMUST00000145974	Col5a3	=	-2.47, -3.11	A3, S7; A7, S7
ENSMUST00000152109	C430049B03Rik	=	-1.80806	A7, S7

Notable among the transcripts in table 4.6 was an isoform of H19, which had a $\log_2$ fold change of -7.37 (up-regulated in A7 compared to S7). What is known about this noncoding gene is that it can influence the expression of the neighbouring gene Igf2 (insulin-like growth factor 2) [46, 53]. A transcript from Igf2 was present in the same co-expression module as the H19 isoform discussed here. The $\log_2$ fold change of the relevant Igf2 transcript was -3.38 (also up-regulated in A7 compared to S7), although in this case the difference was not statistically significant after multiple testing correction.

The second most differentially expressed lncRNA transcript in this module was a noncoding isoform (with a retained intron) of Col5a3 (collagen, type V, alpha 3). The protein encoded by this gene has been shown to be important for glucose homeostasis in mice [54]. It has also been indicated in a cardiovascular disease affecting connective tissue known as Ehlers-Danlos syndrome (one symptom of which can be arterial rupture) [55]. Col5a3 has also been shown to be targeted by the AAA-related microRNA-29 family [56]. The protein coding transcript (with Ensembl transcript ID ENSMUST00000004201) was also significantly differentially expressed between both A3 and S7, A7 and S7 (the $\log_2$ fold changes were -3.06 and -3.78, respectively for those two group comparisons). See figure 4.7.

The blue module was imported into Cytoscape 3.1.1 [34], which was used to calculate the betweenness centrality and degree of each node, the result of which is shown in figure A.7. In the process of importing the module, connections with lesser weights than 0.29 were removed (due to the time it would take to perform an analysis on all nodes). The calculations were thus made on 274 out of the original 2252 nodes. Highly connected and central (in terms of betweenness centrality) transcripts, in the blue module came from the genes Aspm, Hhipl1, Mirg, Ltbp2, Hist1h3i, Wisp1, Fkbp11, Ndc80, Kif11, Ms4a7, Cenpe, Sdc3, Gpx7, Ptpn13, Sorcs2, Ccdc8, Plk1, Myo7a, Cmtm3, Ednra, Kif15 and Kif4. Four of those transcripts were

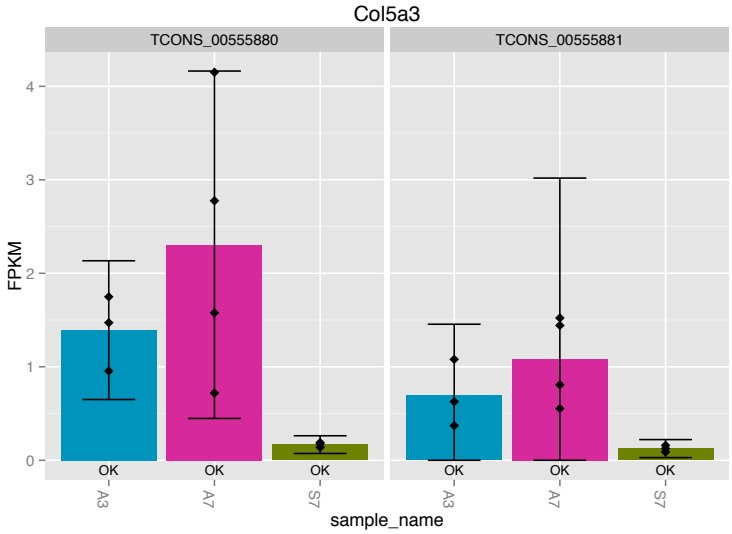


Figure 4.7: Expression levels of the isoforms of Col5a3. TCONS_00555880 is the internal cufflinks ID representing ENSMUST00000004201 (the protein coding isoform) and TCONS_00555881 represents ENSMUST00000145974 (the noncoding isoform). FPKM-values are shown on the vertical axis. Black dots show the expression values of each individual replicate. Blue bars represent the A3 group, pink the seven A7 group and green the S7 control group. Black dots are the expression levels from individual samples in each group and error bars are confidence intervals based on variance modelling calculated by CuffDiff.

novel isoforms of the genes Hhipl1, Mirg and Sdc3. None of them were lncRNAs, though Mirg is a gene harbouring several miRNA.

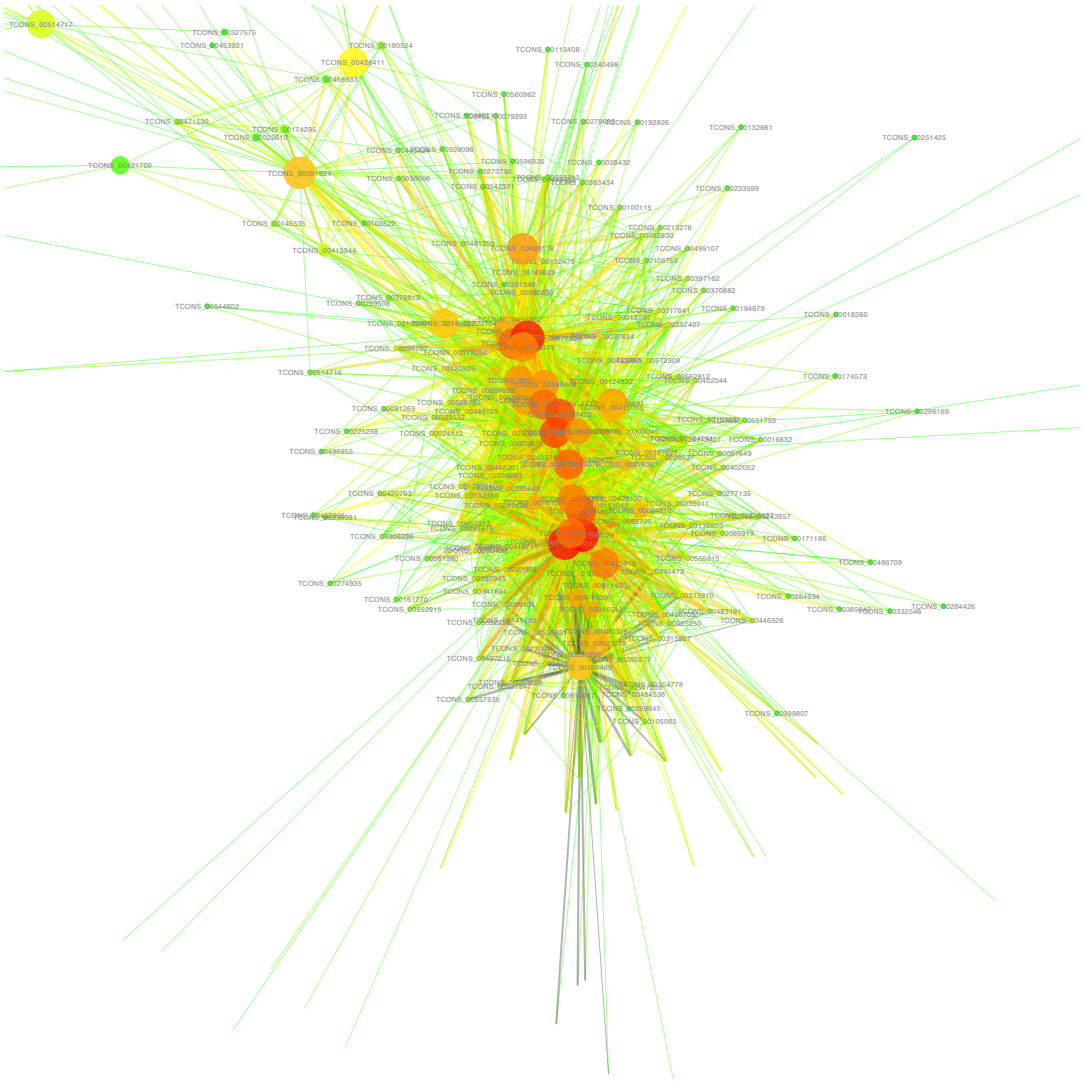


Figure 4.8: The "blue" WGCNA derived module. Larger nodes indicate higher betweenness centrality. Thicker edges have more weight (the edge weights are also displayed on a colour scale from green (low) to red (high)). Degree is indicated with a colour scale: green corresponds to low connectivity and red to high connectivity. The names of the nodes are the internal identifiers assigned to each transcript by the assembly software Cufflinks. A high resolution version of the figure can be obtained from the author upon request.

The "red" module

The three significantly differentially expressed lncRNAs in the red module are summarised in table 4.7. In addition, one novel intergenic lncRNA (lincRNA) was present (TCONS_00385472, see table 4.1), though not with a statistically significant difference in transcript abundance between any of the sample groups (figure 4.12).

Table 4.7: Interesting lncRNAs in the "red" module. "Groups" refers to which sample groups the difference was observed in. Remaining column headers as in table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value	Groups
ENSMUST00000105240	Timeless	=	-3.55	0.00005	0.0027	A3, S7
ENSMUST00000148151	Flcn	=	-2.88	0.00015	0.0070	A3, S7
ENSMUST00000132193	Alkbh6	=	-1.62	0.00055	0.021	A3, S7

One of the three differentially expressed previously annotated lncRNAs in the red module was ENSMUST00000105240, a noncoding isoform of Timeless. The protein coding transcript from this gene was named after its involvement in the regulation of the circadian rhythm in *Drosophila melanogaster*, and its mouse homolog appears to be essential in that context as well [57, 58]. Recently, the gene has also been shown to play a role in coordinated apoptosis of embryonic cells [59]. However, the lncRNA was the only isoform of this gene that was (significantly) differentially expressed.

The second differentially expressed lncRNA in this module was a noncoding variant of Flcn, which otherwise codes for the protein folliculin. The main protein of the gene is believed to be a tumour suppressor [60]. This protein (ENSMUST00000102697) was also significantly differentially expressed (both between the groups A3 and S7, and A7 and S7).

The third one was ENSMUST00000132193. The transcript is a 1112 bp long non-coding isoform of the Alkbh6 (alkylation repair homolog 6) gene, spliced with intron retention. It was the only one of the 12 assembled isoforms of this gene that was found to be significantly differentially expressed. Expression of ENSMUST00000132193 was up-regulated in the AngII-treated mice compared to the controls, though more so in the mice harvested three weeks after treatment than those harvested after seven weeks (see figure 4.11).

Cytoscape 3.1.1 [34] was used to calculate the betweenness centrality and degree of each node. The resulting network can be seen in figure 4.13. Connections with lesser weights than 0.25 were removed (as before, this was mainly done to reduce the computational time of the analysis, but also to focus on the strongest co-expression relations)

The 10 most central nodes of this module are summarised in table 4.8. Notably, one of them (TCONS_00190611) was a predicted novel fragment (although none of the predicted novel lncRNAs). The transcript was present neither in the Ensembl or NONCODE annotation, and attempting to use LiftOver to identify its position on the mm10 genome failed. So either it is a novel fragment or an artifact of the mm9 genome build, though the latter is suspected.

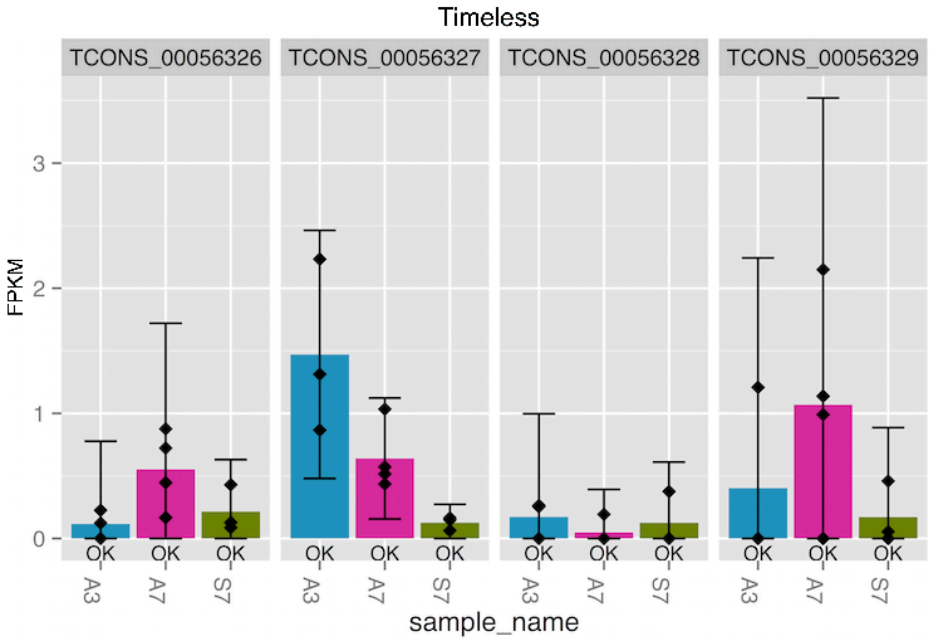


Figure 4.9: Expression levels of four isoforms of Timeless. The lncRNA ENSMUST00000105240 is named TCONS_00056327 here (internal identifier of cufflinks). FPKM-values are shown on the vertical axis. Black dots show the expression values of each individual replicate. Error bars are confidence intervals based on variance modelling calculated by CuffDiff. Blue bars represent the A3 group, pink the seven A7 group and green the S7 control group.

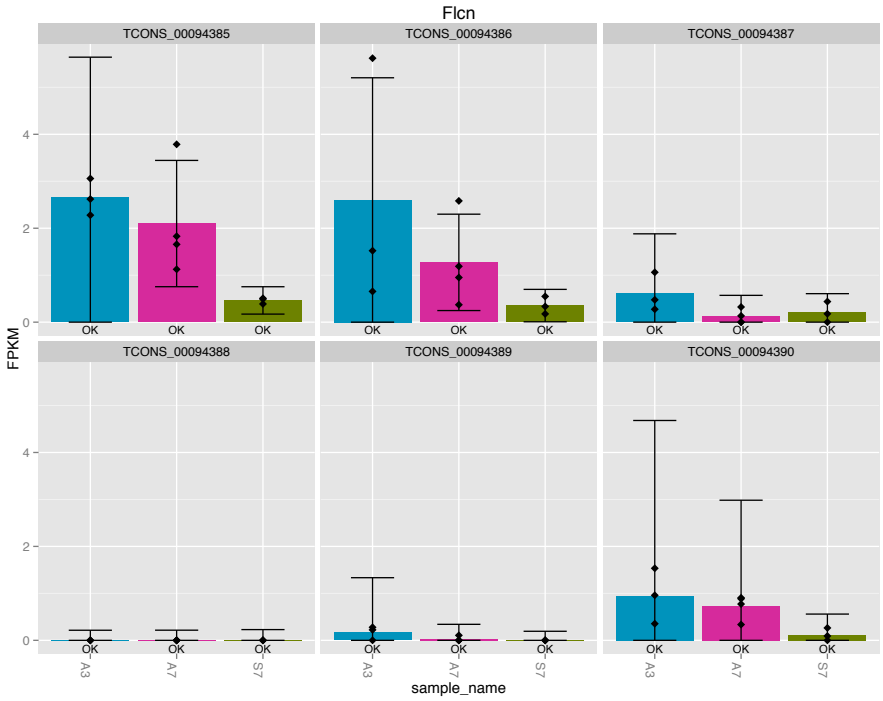


Figure 4.10: Expression levels of the isoforms of Flcn. The lncRNA ENSMUST00000148151 is named TCONS_00094386 here (internal identifier of cufflinks). Axis labels and figure elements the same as in figure 4.9.

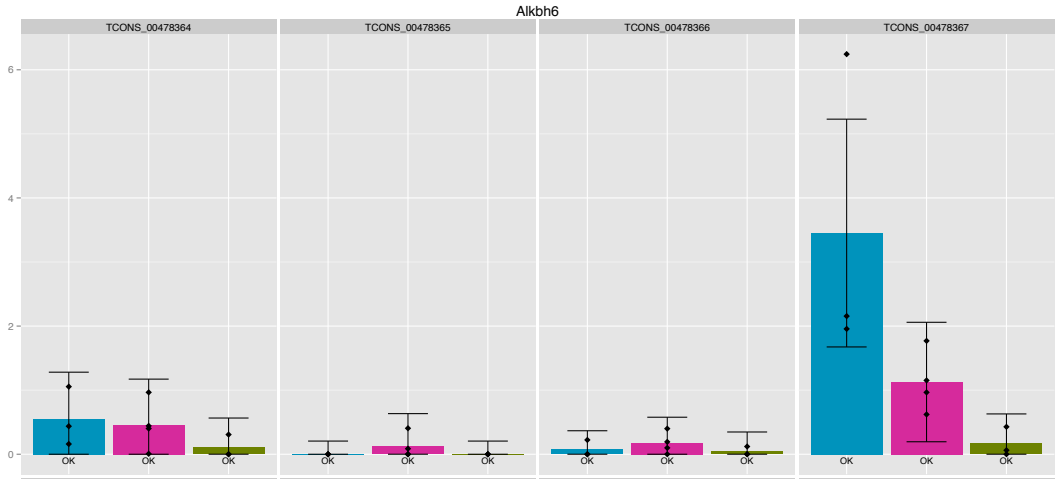


Figure 4.11: Expression levels of four isoforms of Alkbh6. The lncRNA ENSMUST00000132193 is named TCONS_00478367 here (internal identifier of cufflinks). Axis labels and figure elements the same as in figure 4.9.

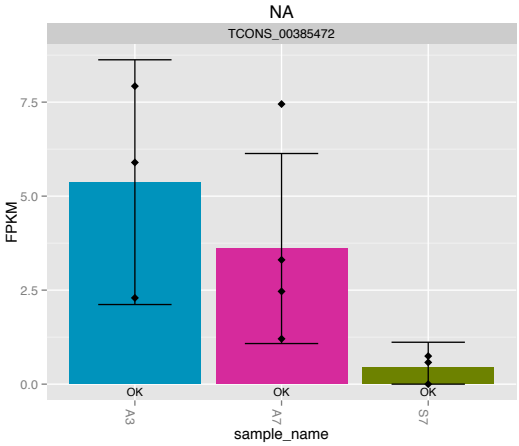


Figure 4.12: Expression levels of the novel lncRNA TCONS_00385472.pdf. Axis labels and figure elements the same as in figure 4.9.

Table 4.8: Summary of the ten of the most central transcripts in the "red" module. The Isoform Ids are internal Cufflinks identifiers. The meaning of the class codes can be found in table 3.1.

Isoform id	Gene	Closest transcript	Class code
TCONS_00079525	Rpain	ENSMUST00000018593	=
TCONS_00161510	Lrp10	ENSMUST00000022782	=
TCONS_00190611	-	-	u
TCONS_00206829	Dhh	ENSMUST00000023737	=
TCONS_00236333	Aars2	ENSMUST00000024733	=
TCONS_00304221	Tm9sf4	ENSMUST00000089027	=
TCONS_00312467	Ccbl1	ENSMUST00000044038	=
TCONS_00426158	Sh3tc1	ENSMUST00000037959	=
TCONS_00524701	Vac14	ENSMUST00000166307	=
TCONS_00534181	Ano8	ENSMUST00000093450	=

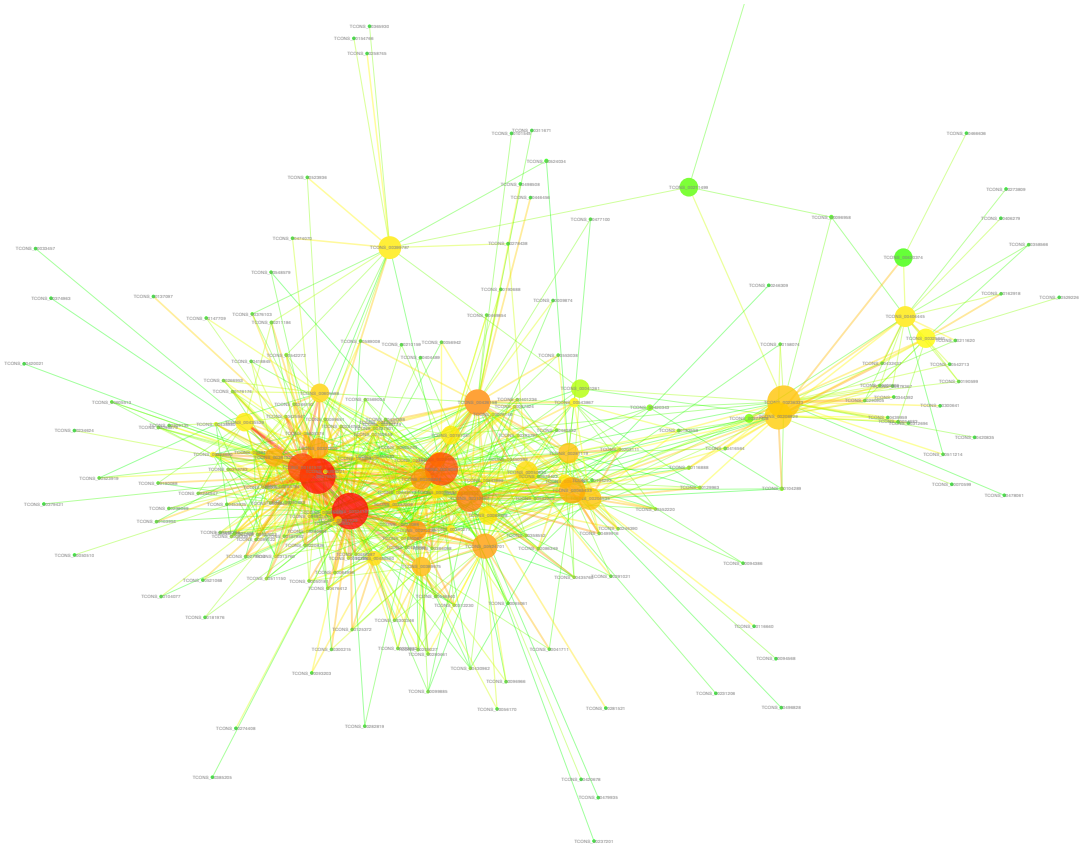


Figure 4.13: The "red" WGCNA derived module. Figure elements are the same as was described in the legend of figure A.7. A high resolution version of the figure can be obtained from the author upon request.

4.5 Gene ontology enrichment

The "blue" module

GO enrichment analysis was performed using the BiNGO plugin of Cytoscape 3.1.1, with default settings used for multiple testing correction. All genes of the transcripts in the blue module were included. In order to interpret those results more easily, the resulting GO terms and false discovery rate (FDR) adjusted p-values were analysed with the GO interpretation tool REVIGO [61] (the software lacks version numbering), using "large" as the size of the output list of terms, "Mus musculus" as reference and otherwise default settings (as they were defined 2014-08-08). This tool uses semantic analysis to group the terms in sensible ways such as to reduce redundancy.

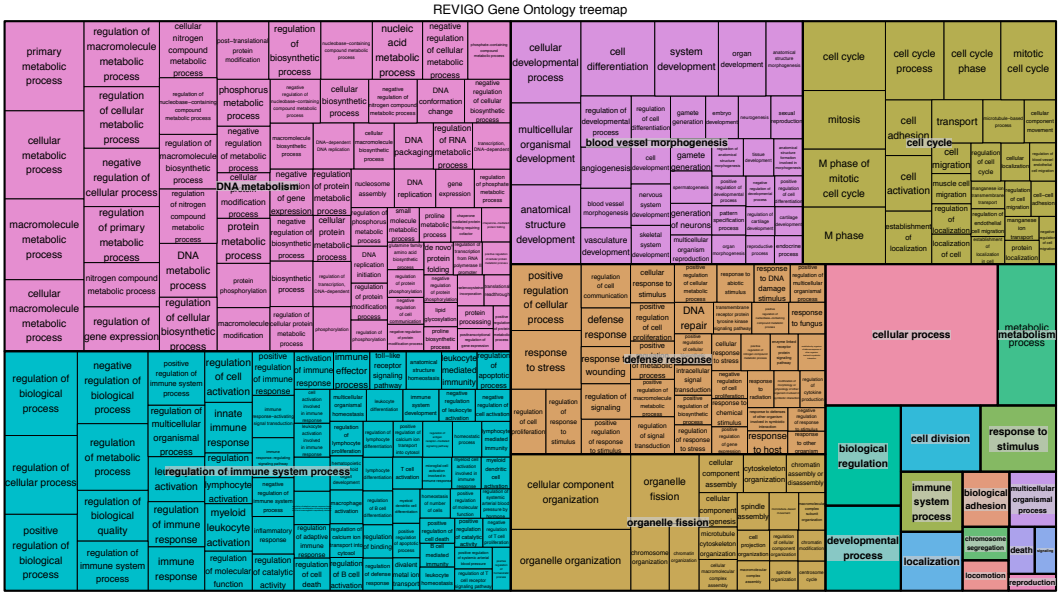
Highly significant categories indicated by REVIGO in the biological process ontology were "DNA metabolism", "regulation of immune system process", "blood vessel morphogenesis", "defence response", "organelle fission", "cell cycle", "cellular process" and "metabolism" (figure 4.14a).

Similarly, in the ontology "molecular function", important categories were: "calcium ion binding", "receptor binding", "protein kinase activity", "binding", "catalytic activity", "GTPase regulator activity", "manganese ion transmembrane transporter activity" and "cytokine receptor activity" (figure 4.14b).

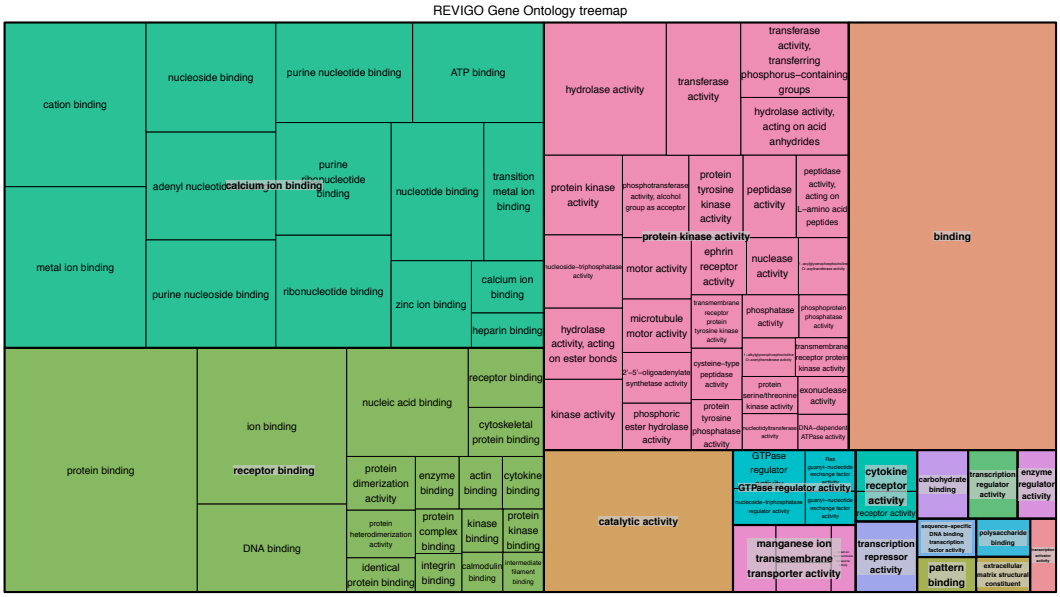
In the "cellular component" category, important terms were "chromosomal part", "cell" (though that term is rather nondescript),"intrinsic to plasma membrane", "organelle", "extracellular region part", "protein-DNA complex", "perinuclear region of cytoplasm", "membrane" and "extracellular matrix' (figure 4.14c).

Figure 4.14: The most prominent terms among the significantly enriched GO terms of the "blue" module; **a)** in the "biological process" ontology, **b)** in the "molecular function" ontology, **c)** in the "cellular component" ontology. The size of the square representing each term is proportional to the absolute log p-value of the enrichment significance of that term. Enrichment was calculated with the BiNGO plugin of Cytoscape 3.1.1 and the semantic grouping in this figure was the result of analysis with the REVIGO software, using the option "large" regarding the size of the resulting list of terms and "Mus musculus" as reference(and otherwise default settings). The $\log_{10}$ p-values visually presented here are the logarithms of the FDR-adjusted p-values from BiNGO. A high resolution version of the figure can be obtained from the author upon request.

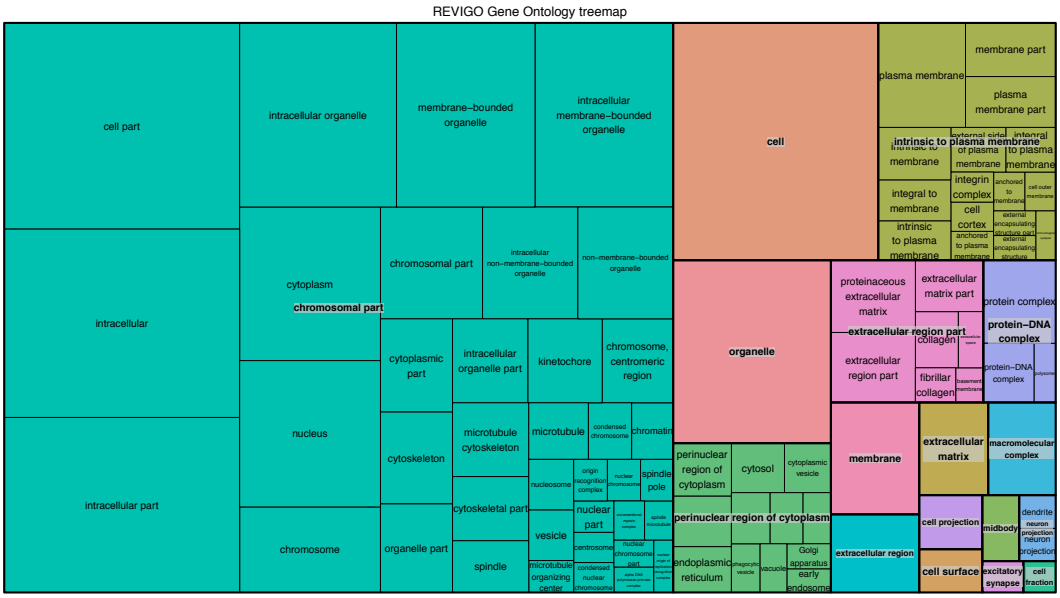
(a)



(b)



(c)



The "red" module

GO enrichment analysis (using the BiNGO plugin of Cytoscape 3.1.1, with default settings used for multiple testing correction) was performed on the "red" module, and the results were parsed and analysed with REVIGO as described in section 4.5.

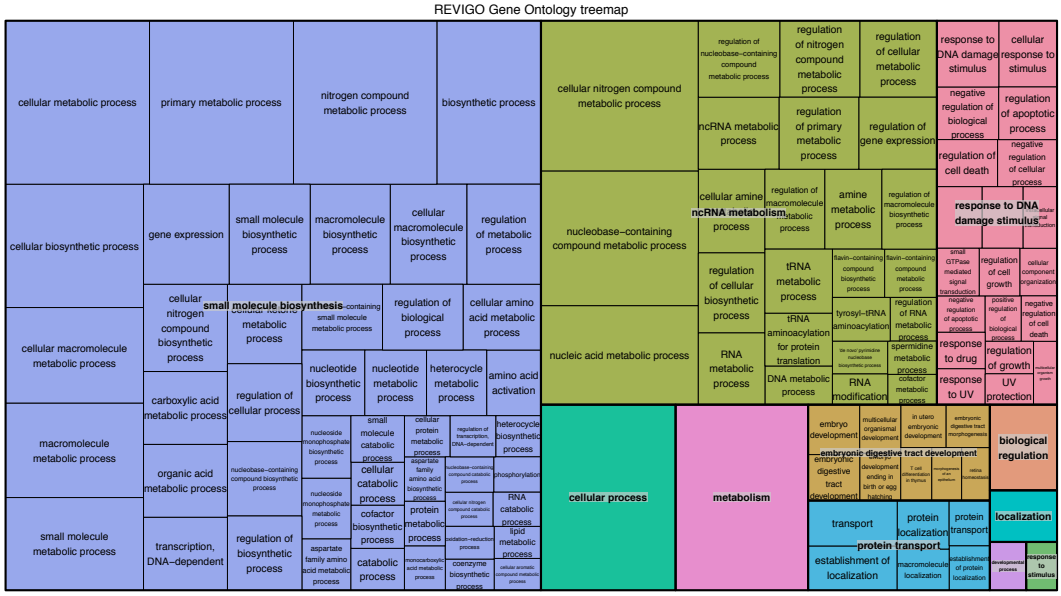
In the "biological process" ontology, the main terms highlighted by analysis with REVIGO were: "small molecule biosynthesis", "ncRNA metabolism", "response to DNA damage stimulus", "cellular process", "metabolism", "embryonic digestive tract development", "protein transport" and "biological regulation" (figure 4.15a).

Similarly, the terms highlighted in the "molecular function" ontology were: "RNA binding", "ligase activity", "binding", "catalytic activity", "identical protein binding", "transcription regulator activity", "cofactor binding", "metal cluster binding" and "RNA polymerase II transcription cofactor activity" (figure 4.15b).

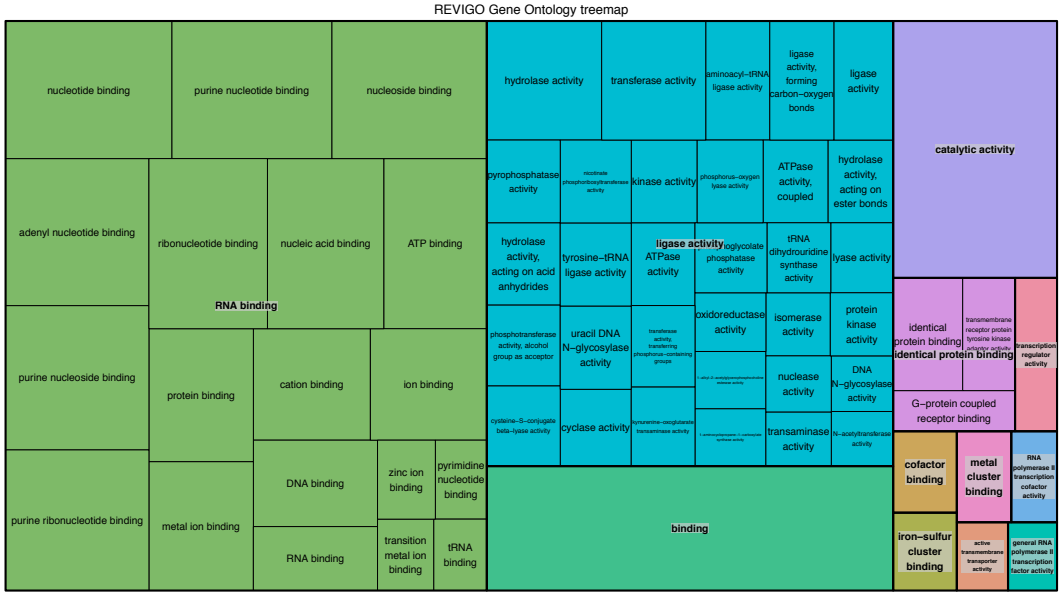
The highlighted "cellular component" categories were: "cytosol", "cell", "organelle", "soluble fraction", "cell fraction", "membrane", "plasma membrane", "endomembrane system", "membrane-enclosed lumen" and "macromolecular complex", with the respective p-value significance ranking visualised in figure 4.15c (the enrichment significance is proportional to the size of the rectangle enclosing each term in the figure).

Figure 4.15: The most prominent terms among the significantly enriched GO terms of the "red" module; **a)** in the "biological process" ontology, **b)** in the "molecular function" ontology, **c)** in the "cellular component" ontology. For a further description of the figure elements, see the legend of figure 4.14. A high resolution version can be obtained from the author upon request.

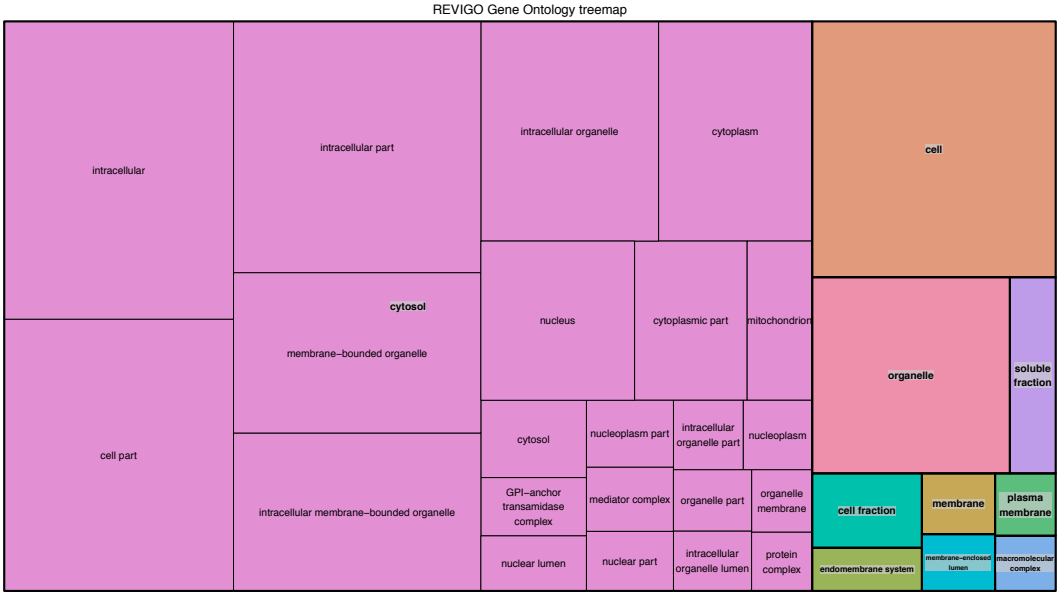
(a)



(b)



(c)


REVIGO Gene Ontology treemap
intracellular
intracellular part
intracellular organelle
cytoplasm
cell
cytosol
membrane−bounded organelle
nucleus
cytoplasmic part
mitochondrion
organelle
soluble fraction
cell part
intracellular membrane−bounded organelle
cytosol
nucleoplasm part
intracellular organelle part
nucleoplasm
GPI−anchor transamidase complex
mediator complex
organelle part
organelle membrane
cell fraction
membrane
plasma membrane
nuclear lumen
nuclear part
intracellular organelle lumen
protein complex
endomembrane system
membrane−enclosed lumen
macromolecular complex

4.6 MetaCore© analysis

The "blue" module

The central nodes of the "blue" module were analysed with MetaCore© [28] (version 6.19). A network was built using the "shortest paths" algorithm (the "expand by one" algorithm would expand the "blue" network with too many connections to handle efficiently), with default settings. The result is shown in figure 4.16.

Among the nodes connected to the central genes of the transcripts of the "blue" module, 7 were connected to either of the following relevant disease terms: "Aneurysm"; "Aortic Aneurysm"; "Aortic Aneurysm, Thoracic"; "Aortic Aneurysm, Abdominal"; "Aneurysm, Dissecting"; "Intracranial Aneurysm"; "Coronary Aneurysm"; "Hypertension". These MetaCore© network objects were ESR1, SMAD3, TGF-beta receptor type I, TGF-beta receptor type II, TGF-beta 1, MMP-9 and a broad group of G protein-coupled receptors called "Galpha(q)-specific peptide GPCRs" (to which one of the original "blue" transcripts was assigned). Additionally, "therapeutic targets" in the network were PLK1 (one of the central genes of the "blue" module), EGFR, ErbB2, SP1 and several others overlapping with the previously mentioned ones.

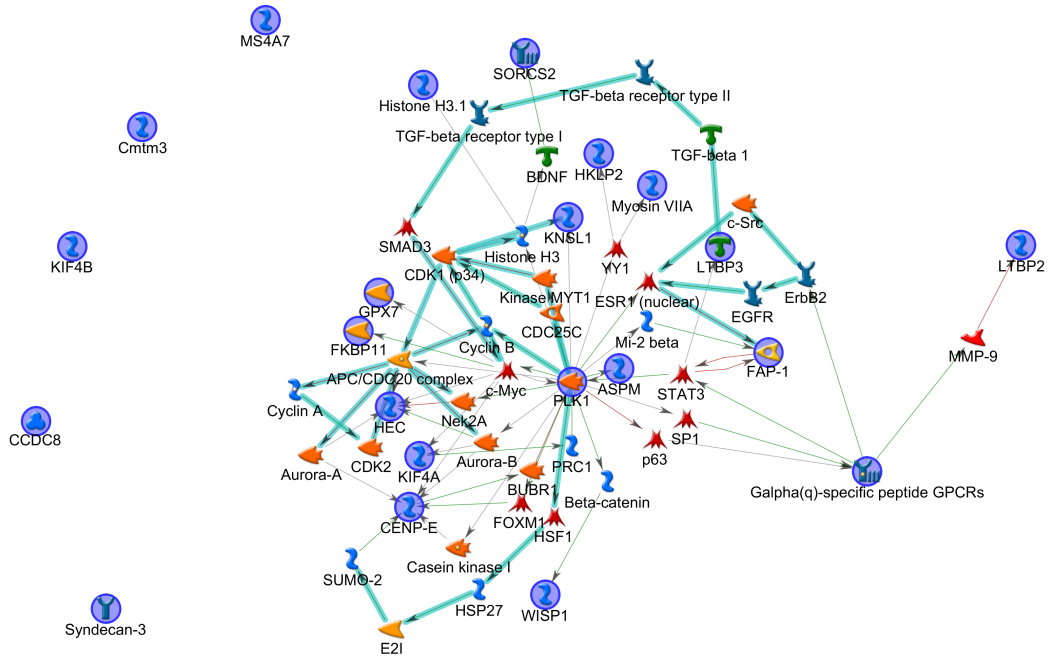


Figure 4.16: MetaCore© Interaction network constructed from central nodes of the "blue" module. The "shortest paths" algorithm was used with default settings. The gene identifiers were in some cases translated to network object carrying the internal naming scheme of MetaCore©, but the original genes are indicated with blue circles. Distances in the image have no meaning.

The "red" module

The nine most central nodes with gene identifiers (table 4.8) of the "red" module were further profiled with MetaCore©. The network shown in 4.17 was built using the "expand by one" algorithm (with default options).

Known therapeutic targets among the nodes in that network were CDK2, GPC3, RELA, VDR, CTNNB1, ESR1 (also present in the MetaCore© network constructed from the central "blue" genes), ESR2, PGR, APH1A and ADAM10. Searching the network for genes related to the terms "Aortic Aneurysm"; "Aneurysm"; "Aortic Aneurysm, Abdominal"; "Aortic Aneurysm, Thoracic"; "Hypertension", yielded the following list of network objects: ESR1, ESR2, CLOCK, SMAD4, PGR, VDR, Caveolin-1 and ATM. A connected compound that might be relevant in the context of cardiovascular disease was cholesterol.

Furthermore, although not indicated by MetaCore© as relevant in the contexts of the conditions mentioned, microRNA-210 was also present in the network. This noncoding RNA has been found to be over-expressed in several cardiovascular diseases [62]. Unfortunately, this RNA could not be detected in the RNA-seq data set, perhaps due to its short length providing challenges during sequencing (in line with the previously mentioned overall relative lack of short noncoding transcripts assembled) (110 bp).

Additionally, the presence of CLOCK may be notable, since the most differentially expressed lncRNA in the "red" module was a noncoding isoform of Timeless (table 4.7), another gene involved in circadian rhythm regulation with which it is known to interact [63].

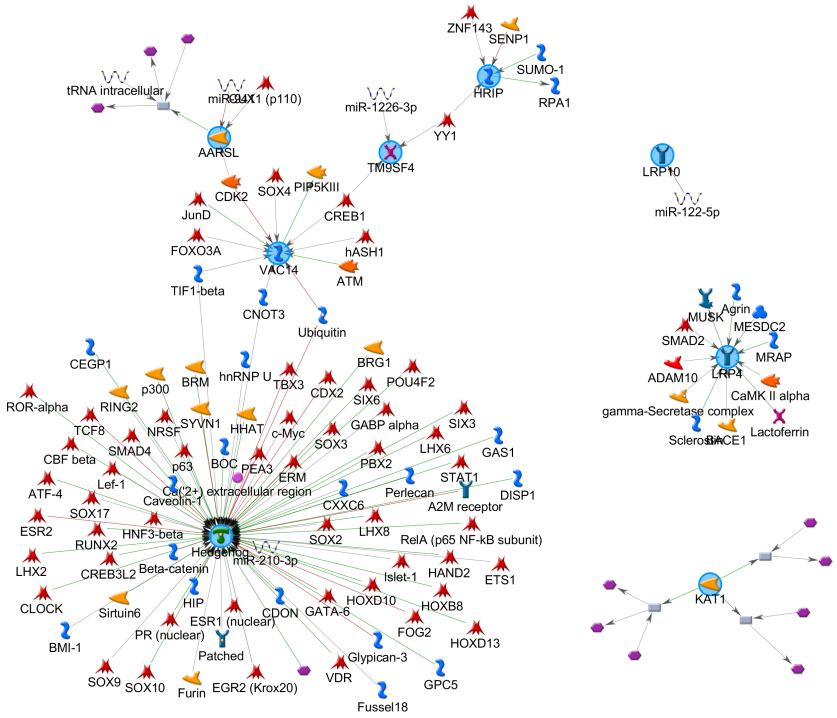


Figure 4.17: MetaCore© Interaction network constructed from nine central nodes of the "red" module. The "expand by one" algorithm was used. The gene identifiers were in some cases translated to network object carrying the internal naming scheme of MetaCore©, but the original genes are indicated with blue circles. Distances in the image have no meaning.

4.7 Clustering comparison

In an attempt to validate the WGCNA clustering, it was compared it with a separate clustering using the k-means method. Since the WGCNA method yielded 51 modules, k was set to 51 for the k-means method as comparison. For k-means, 200 starting points were used, which appeared to be sufficient for convergence.

As can be seen in figure 4.18, where the resulting overlap matrix is visualised. It is apparent that very few pairs of modules share any specific overlap between the WGCNA and k-means clustering. In accordance with this, the association score was a Cramér's V of 0.06, which indicates no significant association between the two methods. Any overlap found is almost entirely due chance (proportional to the sizes of the modules). For instance, the module 4 in the WGCNA method shares a rather large overlap with almost all of the modules obtained from k-means. This is merely the due to the large size of this module. Contrast this with the ideal result illustrated in figure 3.1, where each module only matches well with one module from the other clustering run. This is an intriguing result and exemplifies how different two given clustering methods can perform on the same gene expression dataset.

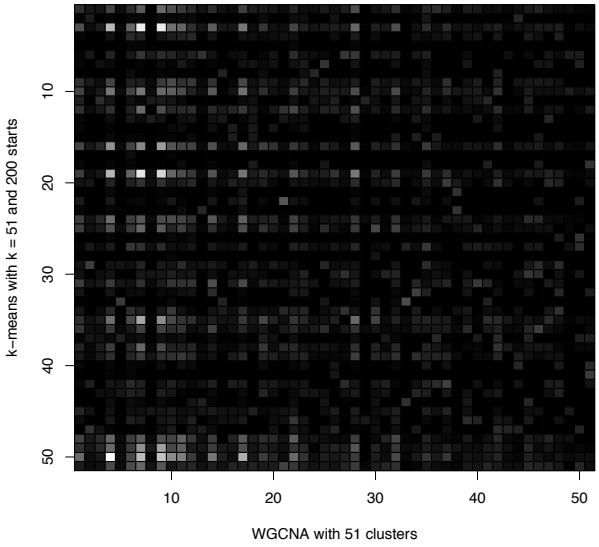


Figure 4.18: The overlap of WGCNA module assignment and k-means clustering on the coefficient of variation filtered dataset. For k-means k was set equal to 51 and n = 200 starting points of the algorithm was used. The image is a visualisation of the contingency matrix, wherein high values of a given cell is displayed brightly and low or zero overlap between two given clusters is displayed darkly shaded. (Additionally, in this image the number of transcripts in the intersection of each pair of modules have been normalised by the total size of the modules for greater clarity.)

This does not, however, invalidate the clustering results. There may be multiple explanations for this observation: Either both methods highlight biologically relevant, yet completely different, co-expression patterns in the data; or one or both of them perform particularly poorly on the data.

One reason for the differences may be that a step in the WGCNA method parses the initial cluster assignment for sub-clusters displaying certain characteristics that

it believes justifies reassigning such sub-clusters to other modules. This step has no equivalent in the k-means method.

Also, since the k-means method is sometimes considered somewhat of a quick and dirty method that may not give the most accurate results on complex datasets (and whose performance highly depends on the user-specified value of the k parameter [64]), one might suspect that the use of this clustering method is simply unsuitable to this complex data, and therefore fails to highlight the same biological aspects as the WGCNA method. But further study is required in order to draw any such conclusions.

This lack of a conclusive objective verification places further weight on the subjective interpretation of biological patterns within the modules as a form of clustering validation.

5. Discussion

Challenges

Due to the expenses of RNA-seq, it has been common to sequence few samples deeply rather than many samples shallowly. This can cause problems when dealing with the, often large, biological variability. In particular, this variability needs to be adequately statistically modelled during differential expression testing in order to make sure that any differences in expression levels are actually caused by the treatment and not just there as a result of stochastic variation. Ideally, one should also test more than one differential expression estimation tool, in order to study the impact of differences in variance modelling on the results. This was not possible here due to time limitations, however.

Another challenge was the fact that not all the assembled fragments were true RNA transcripts expressed by the sampled organism. Some were mere artifacts caused by, for instance, non-specific mapping of reads or failures of the assembly algorithm. Likely, the quality of the library preparation and sequencing runs played a large role in producing these artifacts as well.

Many parameters could be tuned in the Cufflinks assembler, and in hindsight it may have been desirable to mask mitochondrial fragments due to the overrepresentation of these. As it stands, it is suspected that this may have been a reason for the undesirably large overlap between the FPKM distributions of the partially and completely assembled fragments used for setting the expression threshold during the novel transcript filtering. Another potential culprit may have been adapter contamination. Ideally, the potential impact of this should have been studied more in detail. Additional analyses are being performed. (Note added in proof: The results of these analyses are presented in appendix A.)

It is, in general, quite challenging to objectively validate the performance of an unsupervised learning method. At best these methods can be seen as tools for hypothesis construction. The results here showed that the WGCNA module assignment deviated very much from that of k-means clustering. Though one may argue that this should be expected, since the WGCNA method was (after all) developed to provide an alternative clustering algorithm that was better suited to expression data. The poor association can have multiple explanations. However, ultimately, it does not tell us as much about the biological data itself as it informs us of the differences between the algorithms. Had there, however, been found to be a significant association it would have made it much easier to justify conclusions of transcript co-expression. Perhaps k-means was not be best choice for comparison. Another clustering method such as NMF, which has been shown to perform well on biological data [65], would most likely have been a better choice.

Results

Among the 29 novel long non-coding RNA that remained after the filtering procedure undertaken, three were already present in the noncoding RNA database NONCODE (26 thus remaining as novel). This would appear to offer some reassurance that these may indeed be actual lncRNAs, rather than mere artifacts. Due to the inability of finding a well performing FPKM based artifact filtering threshold, however, it is hard to completely rule out that possibility. Regardless, actual verification with biotechnology techniques is required in order to confirm them.

In total, 248 lncRNAs were found to be differentially expressed. 187 of those were between the AngII treated mice harvested after seven weeks and the controls. Two of those were novel. In the comparison between the treated mice harvested after three weeks and the controls, 131 lncRNA (one of those novel) were differentially expressed.

In several of the cases where lncRNAs were differentially expressed, they seemed to follow the expression patterns of the protein coding isoforms of the same genes. One may therefore suspect that the lncRNA expression in those cases was mainly a passive byproduct of the overall gene expression, rather than a consequence of independent functional significance. Though such independent roles can not be precluded at this stage. The exclusively lncRNA producing gene H19, on the other hand, would then appear more likely to have some role in the disease (perhaps via its connection to Igf2).

Furthermore, transcript were clustered into co-expression groups (modules), which were subsequently correlated to treatment. The results highlighted a few of those, which appeared to display expression patterns related to the disease. Of those, a couple contained both novel and known lncRNAs. Those modules were subject to further analysis, which involved Gene Ontology enrichment and MetaCore© profiling of genes belonging to central (in terms of network properties) transcript isoforms.

The results of GO enrichment of the first (here called "blue") module showed an enrichment of terms such "DNA metabolism", "regulation of immune system process", "blood vessel morphogenesis" and "defence response". Since immune reactions, such as vascular inflammation, are known to be involved in the disease [3] these results should not be surprising if the genes in that module were indeed related to the disease. Another interesting term that was enriched was "extracellular region part" (encapsulating the term "collagen") and "extracellular matrix", which would perhaps be related to processes affecting the arterial wall in the progression of the disease (collagen protein producing genes were also found to be among the most significantly differentially expressed ones).

The second module analysed (called "red"), also displayed enrichment of a number of GO terms. "small molecule biosynthesis", "ncRNA metabolism" (processes involving tRNA seemed to be common under this term), "response to DNA damage stimulus", "RNA binding", and "ligase activity" were among the highlighted ones. These terms may be broad, but connections to the disease could be the terms "regulation of cell death" and "regulation of cell growth" encapsulated by "response to DNA damage stimulus". Enrichment of GO terms in a module would appear to strengthen claims of some degree of actual co-expression.

Extracting the genes belonging to central transcripts in these modules and performing analysis of curated interactions with MetaCore© revealed that fairly close

ties between those genes could be found, in networks that also included therapeutic targets and a number of genes connected to aortic aneurysm diseases and hypertension. Network objects such as ESR1, SMAD3, TGF-beta receptor type I, TGF-beta receptor type II, TGF-beta 1, MMP-9 (in the MetaCore© network build from central genes of the "blue" module) and ESR1, ESR2, CLOCK, SMAD4, PGR, VDR, Caveolin-1 and ATM (based on the "red" module) may offer some clues as to functionality represented in those groups of putatively co-expressed transcripts. Long noncoding RNAs included in those modules could, by guilt of association, be interesting to study further.

Of the differentially expressed previously annotated lncRNAs, 11 ones were present in either of these two highly treatment-correlated co-expression modules. These modules also contained one novel lncRNA each (which, however, were not significantly differentially expressed).

6. Conclusions

This work has revealed 26 novel long noncoding RNA candidates in mouse, two of which were up-regulated in the AngII treated group (which develops aneurysms) in comparison with controls. In addition, 185 previously annotated lncRNAs were also indicated as differentially expressed between those sample groups. One lncRNA in particular displayed a high fold change between the groups: H19.

Protein coding isoforms that displayed large differences in expression between case and control involved those from histone proteins, collagens (important parts of the extracellular matrix), thrombospondin proteins (known to interact with the extracellular matrix) and Itm2a (involved in T-cell activation and muscle cell differentiation). These observations are interesting since degradation of extracellular matrix in the arterial wall is an important part in the progression of abdominal aortic aneurysm disease, and so are immune reations.

Furthermore, a network of transcriptional co-expression was built using weighted gene co-expression network analysis. Several lncRNAs were found to be present in a couple of co-expression modules obtained from this procedure, whose expression patterns appeared to be related to treatment. Investigating these modules further revealed ties to genes and processes that could be connected to abdominal aortic aneurysm disease. Those results would suggest that the lncRNAs included in those modules could warrant further research.

Understanding the role of noncoding RNAs in abdominal aortic aneurysm may provide knowledge that can open paths to new diagnostic possibilities of a disease that rarely present any symptoms until it is too late. In a best case scenario, new therapies that can supplement or provide alternatives to current invasive last-minute surgical procedures could be found along the way to towards the goal of unraveling the complex genetic mechanisms underlying this life-threatening disease.

7. Acknowledgements

Karolinska Institutet
Bengt Sennblad
Lars Maegdefessel
Jesper Gådin
Mattias Frånberg
Stanford University
Joshua Spin
Uppsala University
Magnus Lundgren
SciLifeLab
Mikael Huss

And last, but not least: all those I forgot to include here.

8. References

[1] John L Rinn and Howard Y Chang. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81:145–66, January 2012.

[2] Sourabh Aggarwal, Arman Qamar, Vishal Sharma, and Alka Sharma. Abdominal aortic aneurysm: A comprehensive review. *Experimental and clinical cardiology*, 16:11–15, 2011.

[3] Lars Maegdefessel, Ronald L Dalman, and Philip S Tsao. Pathogenesis of abdominal aortic aneurysms: microRNAs, proteases, genetic associations. *Annual review of medicine*, 65:49–62, January 2014.

[4] Ada Congrains, Kei Kamide, Mitsuru Ohishi, and Hiromi Rakugi. ANRIL: Molecular Mechanisms and Implications in Human Health. *International journal of molecular sciences*, 14(1):1278–92, January 2013.

[5] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–89, September 2012.

[6] Daiana Weiss, John J Kools, and Robert W Taylor. Angiotensin II-Induced Hypertension Accelerates the Development of Atherosclerosis in ApoE-Deficient Mice. *Circulation*, 103(3):448–454, January 2001.

[7] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[8] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, et al. RefSeq: an update on mammalian reference sequences. *Nucleic acids research*, 42(Database issue):D756–63, January 2014.

[9] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, et al. The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, 42(Database issue):D764–70, January 2014.

[10] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, et al. Ensembl 2014. *Nucleic acids research*, 42(Database issue):D749–55, January 2014.

[11] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, May 2009.

[12] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, April 2013.

[13] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5, May 2010.

[14] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)*, 27(17):2325–9, September 2011.

[15] Lei Sun, Zhihua Zhang, Timothy L Bailey, Andrew C Perkins, Michael R Tallack, et al. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *BMC bioinformatics*, 13:331, January 2012.

[16] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, January 2010.

[17] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40, January 2010.

[18] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, January 2013.

[19] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78, March 2012.

[20] James Macqueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. *Elements*, 1:337–387, 2009.

[22] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, January 2008.

[23] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, June 1994.

[24] Daniel D Lee and Sebastian H Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[25] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–84, April 2002.

[26] Chaoyong Xie, Jiao Yuan, Hui Li, Ming Li, Guoguang Zhao, et al. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research*, 42(Database issue):D98–103, January 2014.

[27] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000.

[28] Yuri Nikolsky, Eugene Kirillov, Roman Zuev, Eugene Rakhmatulin, and Tatiana Nikolskaya. Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform. *Methods in molecular biology (Clifton, N.J.)*, 563:177–196, 2009.

[29] Simon Andrews. FastQC: A quality control tool for high throughput sequence data. *[Online] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*, 2010.

[30] Liguo Wang, Shengqin Wang, and Wei Li. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, 28:2184–2185, 2012.

[31] Michael F Lin, Irwin Jungreis, and Manolis Kellis. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics (Oxford, England)*, 27(13):i275–82, July 2011.

[32] Kun Sun, Xiaona Chen, Peiyong Jiang, Xiaofeng Song, Huating Wang, et al. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics*, 14 Suppl 2(Suppl 2):S7, January 2013.

[33] Kun Sun, Yu Zhao, Huating Wang, and Hao Sun. Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)–application in human skeletal muscle cells. *PloS one*, 9(1):e84500, January 2014.

[34] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13:2498–2504, 2003.

[35] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)*, 24(5):719–20, March 2008.

[36] Liang Wang, Hui Tang, Venugopal Thayanithy, Subbaya Subramanian, Ann L Oberg, et al. Gene networks and microRNAs implicated in aggressive prostate cancer. *Cancer research*, 69(24):9490–7, December 2009.

[37] Ovidiu D Iancu, Sunita Kawane, Daniel Bottomly, Robert Searles, Robert Hitzemann, et al. Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics (Oxford, England)*, 28(12):1592–7, June 2012.

[38] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, January 2008.

[39] Qi Liao, Jia Shen, Jianfa Liu, Xi Sun, Guoguang Zhao, et al. Genome-wide identification and functional annotation of Plasmodium falciparum long non-coding RNAs from RNA-seq data. *Parasitology research*, 113(4):1269–81, April 2014.

[40] Karl G Kugler, Gerald Siegwart, Thomas Nussbaumer, Christian Ametz, Manuel Spannagl, et al. Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (Triticum aestivum L.). *BMC genomics*, 14(1):728, January 2013.

[41] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131, July 2010.

[42] Angie S Hinrichs. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34:D590–D598, 2006.

[43] Susanne Motameny, Stefanie Wolters, Peter Nürnberg, and Björn Schumacher. Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. *Genes*, 1(1):70–84, January 2010.

[44] William F Marzluff, Eric J Wagner, and Robert J Duronio. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature reviews. Genetics*, 9:843–854, 2008.

[45] Kazimierz T Tycowski, Alar Aab, and Joan A Steitz. Guide RNAs with 5' Caps and Novel Box C/D snoRNA-like Domains for Modification of snRNAs in Metazoa. *Current Biology*, 14:1985–1995, 2004.

[46] Adele Murrell, Sarah Heeson, and Wolf Reik. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nature genetics*, 36:889–893, 2004.

[47] Carlos Gonzalez-Quesada, Michele Cavalera, Anna Biernacka, Ping Kong, Dong Wook Lee, et al. Thrombospondin-1 induction in the diabetic myocardium stabilizes the cardiac matrix in addition to promoting vascular rarefaction through angiopoietin-2 upregulation. *Circulation Research*, 113:1331–1344, 2013.

[48] Joseph V Moxon, Matthew P Padula, Paula Clancy, Theophilus I Emeto, Ben R Herbert, et al. Proteomic analysis of intra-arterial thrombus secretions reveals a negative association of clusterin and thrombospondin-1 with abdominal aortic aneurysm. *Atherosclerosis*, 219:432–439, 2011.

[49] Dave Van Den Plas and Joseph Merregaert. Constitutive overexpression of the integral membrane protein Itm2A enhances myogenic differentiation of C2C12 cells. *Cell Biology International*, 28:199–207, 2004.

[50] Jacqueline Kirchner and Michael J Bevan. ITM2A is induced during thymocyte selection and T cell activation and causes downregulation of CD8 when overexpressed in CD4(+)CD8(+) double positive thymocytes. *The Journal of experimental medicine*, 190:217–228, 1999.

[51] Karen Pittois, Jan Wauters, Paul Bossuyt, Willy Deleersnijder, and Joseph Merregaert. Genomic organization and chromosomal localization of the Itm2a gene. *Mammalian Genome*, 10:54–56, 1999.

[52] Stephen G Young, Brandon S J Davies, Constance V Voss, Peter Gin, Michael M Weinstein, et al. GPIHBP1, an endothelial cell transporter for lipoprotein lipase. *Journal of lipid research*, 52(11):1869–84, December 2011.

[53] Anne Gabory, Marie-Anne Ripoche, Tomomi Yoshimizu, and Luisa Dandolo. The H19 gene: regulation and function of a non-coding RNA. *Cytogenetic and genome research*, 113(1-4):188–93, January 2006.

[54] Guorui Huang, Gaoxiang Ge, Dingyan Wang, Bagavathi Gopalakrishnan, Delana H Butz, et al. α3(V) collagen is critical for glucose homeostasis in mice due to effects in pancreatic islets and peripheral tissues. *The Journal of clinical investigation*, 121:769–783, 2011.

[55] Guntram Borck, Peter Beighton, Christian Wilhelm, Jürgen Kohlhase, and Christian Kubisch. Arterial rupture in classic Ehlers-Danlos syndrome with COL5A1 mutation. *American Journal of Medical Genetics, Part A*, 152:2090–2093, 2010.

[56] Lars Maegdefessel, Junya Azuma, and Ryuji Toh. Inhibition of microRNA-29b reduces murine abdominal aortic aneurysm development. *The Journal of clinical ...*, 122(2):497–506, 2012.

[57] Amita Sehgal, Jeffrey L Price, Bernice Man, and Michael W Young. Loss of circadian behavioral rhythms and per RNA oscillations in the Drosophila mutant timeless. *Science (New York, N.Y.)*, 263(5153):1603–6, March 1994.

[58] Keziban Unsal-Kaçmaz, Thomas E Mullen, William K Kaufmann, and Aziz Sancar. Coupling of human circadian and cell cycles by the timeless protein. *Molecular and cellular biology*, 25:3109–3116, 2005.

[59] Linda P O'Reilly, Simon C Watkins, and Thomas E Smithgall. An unexpected role for the clock protein timeless in developmental apoptosis. *PloS one*, 6(2):e17157, January 2011.

[60] Nancy F da Silva, Dean Gentle, L uke B Hesson, Dion G Morton, Farida Latif, et al. Analysis of the Birt-Hogg-Dubé (BHD) tumour suppressor gene in sporadic renal cell carcinoma and colorectal cancer. *Journal of medical genetics*, 40(11):820–4, November 2003.

[61] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, January 2011.

[62] Cecilia Devlin, Simona Greco, Fabio Martelli, and Mircea Ivan. miR-210: More than a silent player in hypoxia. *IUBMB life*, 63(2):94–100, February 2011.

[63] Lauren P Shearman, Sathyanarayanan Sriram, David R Weaver, Elizabeth S Maywood, Ines Chaves, et al. Interacting molecular loops in the mammalian circadian clock. *Science (New York, N.Y.)*, 288(5468):1013–9, May 2000.

[64] Jianhua Ruan, Angela K Dean, and Weixiong Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC systems biology*, 4:8, January 2010.

[65] Renaud Gaujoux and Cathal Seoighe. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics*, 11:367, January 2010.

[66] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, May 2011.

[67] Thomas H Barker, Gretchen Baneyx, Marina Cardó-Vila, Gail A Workman, Matt Weaver, et al. SPARC regulates extracellular matrix organization through its modulation of integrin-linked kinase activity. *The Journal of biological chemistry*, 280(43):36483–93, October 2005.

[68] Elena Tomasello and Eric Vivier. KARAP/DAP12/TYROBP: three names and a multiplicity of biological functions. *European journal of immunology*, 35(6):1670–7, June 2005.

[69] Harini Krishnan, Jhon A Ochoa-Alvarez, Yongquan Shen, Evan Nevel, Meenakshi Lakshminarayanan, et al. Serines in the intracellular tail of podoplanin (PDPN) regulate cell motility. *The Journal of biological chemistry*, 288(17):12215–21, April 2013.

[70] Jillian L Astarita, Sophie E Acton, and Shannon J Turley. Podoplanin: emerging functions in development, the immune system, and cancer. *Frontiers in immunology*, 3:283, January 2012.

[71] Gaetano Santulli. Angiopoietin-like proteins: a comprehensive look. *Frontiers in endocrinology*, 5:4, January 2014.

A. Alternative analysis

A reiteration of the methods presented in the main body of this work was performed in order to examine the impact of trimming reads with respect to sequencing adapter oligonucleotides, as well as the effect of mitochondrial read overrepresentation. Additionally, some parameters of mapping and assembly were changed to alternatives that were deemed more likely to be appropriate for the purpose of this work.

A.1 Methods

A.1.1 Read trimming

Sequences representing the oligonucleotide adapters of the EpiBio Scriptseq v2 stranded sequencing protocol were removed from the ends of reads using the read trimming tool Cutadapt [66] version 1.4.1, using the "-b" option (which instruct the algorithm to match the oligonucleotide sequences against both the 3' and 5' ends, as well as the interior of the reads). Poly-A and poly-T tails of reads were also trimmed using the same method. Low quality bases (as determined using FastQC [29] base quality plots) at the ends were removed prior to adapter trimming using fastx-trimmer version 0.0.13 (16 nucleotides from the beginning and 15 nucleotides from the end of each read). Since each read needs to have a corresponding paired partner, "singleton" reads (without partner) were removed.

A.1.2 Read mapping

Mapping was performed largely identically as described in the methods section of the main report, with the exception of the added option "–no-mixed", which instructs the program to only report alignments where both reads in a pair can be mapped. It was reasoned that this option would decrease the risk of assembling transcript artifacts downstream, due to otherwise potentially unreliable singleton read alignment. In addition, the "-r" and "–mate-std-dev" option was specified according to the results of RSeQC [30] read mapping inspection (the procedure of which was described in the methods section of the main body of the report).

A.1.3 Transcript assembly

Transcriptome assembly was performed as described in the main methods section, with the addition of the "-M" option, which purpose was to mask reads mapping to mitochondrial transcripts. The mask file was derived from the Ensembl [10] transcriptome annotation file that was used for mapping and assembly. Assemblies were merged with Cuffcompare [13] (using the -M option and "-g", which instructs

the program to run RABT assembly, as described regarding the cufflinks runs in the main methods section). The merged assembly was annotated with the Ensembl annotation using cuffcompare (options "-G", "-C" and "-r").

A.1.4 Differential expression testing

Differential expression testing was performed as as described in section 3.6.1, with the exception that the "-u" option was not used (due to problems pertaining to glitches of the Cuffdiff [18] software and since it did not seem to affect the results in any significant way). Version 2.2.1 of Cuffdiff encountered serious issues of unknown cause that were not seen in the previous runs ("Segmentation fault"), so 2.1.1 was used instead. The "-b" option was also used, which inspects the ends of the reads for overrepresentation of certain patterns, and tries to account for this overrepresentation by adjusting FPKM values if such bias is found. The question was raised that this was perhaps not appropriate when read ends have been trimmed (since any such biased sequence motifs would have been stripped away), but since the algorithm only performs correction if bias is actually detected (and would otherwise do nothing at all), it was argued that it would not have any negative consequences to use the option.

A.1.5 Identification of known long noncoding RNAs

Identification of previously annotated lncRNAs in the set of assembled fragments was performed identically as previously described in section 3.4.

A.1.6 Prediction of novel long noncoding RNAs

The approach used previously to filter away potential artifacts (using the FPKM distributions of completely and partially assembled transcripts) was deemed unsatisfactory for two reasons. For one, the time and system resources required for running cufflinks on merged alignment files was very significant. Secondly, the performance of the previously constructed FPKM threshold classifier was very poor.

Therefore a new approach was used, whereby FPKM values, confidence intervals and transcript abundance estimations status ("OK" or not) as given by Cuffdiff (using merged assemblies from Cuffmerge [19]) was used instead to filter out unreliable transcripts. Only transcripts whose estimated FPKM confidence interval was above zero, and whose status were "OK" were kept. Additionally, only those of the transcripts having FPKM greater than zero in at least two mice in each condition were retained. After thereby discarding low quality assemblies, the same filtering pipeline as described in section 3.5 was employed to discover novel lncRNA candidates.

A.1.7 Co-expression network analysis

Co-expression network analysis was performed as described in section 3.6.3, with the exception of using 12 as the soft thresholding power, setting minimum module size to 100 and merge cut height to 0.15. In addition, the dataset was pre-filtered in order to reduce its size to around 16000. This filtering was composed of two steps: First, only keeping transcripts from genes that were either in the Ensembl

annotation, novel lncRNA or NONCODE lncRNA. Then, filtering the dataset in order to only keep the top ca 16000 transcripts with highest coefficient of variation.

A.2 Results

A.2.1 Identification of long noncoding RNAs

69 novel long noncoding lncRNA candidates were discovered. Three of these overlapped with the NONCODE annotation (with IDs TCONS_00141617, TCONS_00722068 and TCONS_00722067 in the trimmed read based analysis), 66 thus remaining as truly novel. One transcript overlapped with the previously identified novel lncRNA candidates, (having the ID TCONS_00129545 in the new analysis and TCONS_00084766 in the previous one). 30 of the novel lncRNA did not have any kind of overlap with annotated genes and were thus considered intergenic (class code "u", table A.1). The remaining 36 appeared to be novel isoforms of previously annotated transcripts (table A.2).

Regarding already known transcript isoforms, 31624 long noncoding RNA were identified from previous annotations by NONCODE (v4) and Ensembl (mm9).

Table A.1: Overview of the 30 novel intergenic lncRNA candidates. Nearest reference genes, transcripts and class codes were assigned by Cuffcompare, using the Ensembl mm9 annotation as reference. Column headers as in main text table 4.1.

Isoform id	Gene	Nearest transcript	Class	Length	Locus
TCONS_00078697	-	-	u	1174	10:62259045-62260412
TCONS_00122305	-	-	u	1070	11:58409413-58410669
TCONS_00126277	-	-	u	694	11:84684559-84685378
TCONS_00129545	-	-	u	3434	11:104796669-104801381
TCONS_00156551	-	-	u	403	12:32068232-32068765
TCONS_00170486	-	-	u	1074	12:113950607-113951761
TCONS_00189355	-	-	u	1250	12:107120334-107121995
TCONS_00224111	-	-	u	4478	13:74597747-74603043
TCONS_00305463	-	-	u	669	15:86592532-86593366
TCONS_00328957	-	-	u	1642	16:19892287-19893993
TCONS_00359545	-	-	u	2538	17:95166243-95174862
TCONS_00373910	-	-	u	2500	17:88351484-88354043
TCONS_00388566	-	-	u	2163	18:76813155-76815764
TCONS_00452647	-	-	u	844	2:171771282-171772189
TCONS_00496710	-	-	u	1543	3:69572143-69573834
TCONS_00502760	-	-	u	998	3:98061089-98062695
TCONS_00545054	-	-	u	1839	4:3016213-3018590
TCONS_00622031	-	-	u	511	5:28198977-28199610
TCONS_00637091	-	-	u	904	5:114545674-114547464
TCONS_00663966	-	-	u	517	6:135306139-135306768
TCONS_00681106	-	-	u	1238	6:90843187-90845069
TCONS_00713893	-	-	u	613	7:6928150-6928993
TCONS_00721729	-	-	u	2453	7:65876002-65879125
TCONS_00722564	-	-	u	974	7:69183232-69184816
TCONS_00738833	-	-	u	1493	8:19706436-19708036
TCONS_00742571	-	-	u	3569	8:43100252-43103910
TCONS_00743635	-	-	u	943	8:48001763-48002902
TCONS_00769513	-	-	u	2114	8:76172586-76174856
TCONS_00770093	-	-	u	536	8:79075977-79076569
TCONS_00829084	-	-	u	1035	X:41701449-4170340

Table A.2: Overview of the 36 novel lncRNA candidates that appeared to be new isoforms of previously known transcripts. Nearest reference genes, transcripts and class codes were assigned by Cuffcompare, using the Ensembl mm9 annotation as reference. Column headers as in main text table 4.1.

Isoform id	Gene	Nearest transcript	Class	Length	Locus
TCONS_00002246	Gm16720	ENSMUST00000130710	j	3030	1:16647632-16652359
TCONS_00044520	Als2cr12	ENSMUST00000055313	j	1832	1:58713908-58752833
TCONS_00058021	9230116N13Rik	ENSMUST00000140195	j	2077	1:138308120-138336735
TCONS_00080011	RP23-60H16.1.1	ENSMUST00000177104	j	1963	10:70773492-70777007
TCONS_00122763	Gm12273	ENSMUST00000137418	j	1592	11:61764276-61766238
TCONS_00126764	Coil	ENSMUST00000135325	j	579	11:88831107-88852927
TCONS_00191604	Gm17177	ENSMUST00000169144	j	1305	13:3120377-3122425
TCONS_00199845	Gm904	ENSMUST00000099519	j	1151	13:50738596-50741207
TCONS_00315218	Cd200r2	ENSMUST00000102805	j	1176	16:44867209-44915953
TCONS_00335531	2310005G13Rik	ENSMUST00000067173	j	3363	16:57036923-57071459
TCONS_00386270	Gm4951	ENSMUST00000031549	j	5278	18:60371224-60408364
TCONS_00416304	Slc22a28	ENSMUST00000065651	j	933	19:8136698-8206472
TCONS_00425961	Gm13293	ENSMUST00000131188	j	1719	2:11261115-11264989
TCONS_00479186	2310005A03Rik	ENSMUST00000142569	j	1815	2:154923287-154925817
TCONS_00495995	RP23-445D13.2.1	ENSMUST00000176925	j	1393	3:64251244-64264017
TCONS_00545120	Gm11784	ENSMUST00000118568	j	1210	4:3313992-3315945
TCONS_00555933	Orm3	ENSMUST00000006687	j	888	4:63017086-63020545
TCONS_00562897	Gm12789	ENSMUST00000106914	j	1489	4:101659230-101662836
TCONS_00579840	Gm12505	ENSMUST00000146041	j	743	4:55423234-55431714
TCONS_00589000	Skint10	ENSMUST00000068851	j	980	4:112383751-112447495
TCONS_00590275	Gm12886	ENSMUST00000106266	j	2330	4:121086420-121095704
TCONS_00621502	Mir3096	ENSMUST00000116685	o	2228	5:23214891-23218029
TCONS_00621825	Speer4a	ENSMUST00000079447	j	3079	5:26359122-26366046
TCONS_00640073	Gm454	ENSMUST00000160126	j	2029	5:138643516-138648896
TCONS_00650640	Vmn1r4	ENSMUST00000176838	j	845	6:56874015-56908094
TCONS_00662554	Clec2h	ENSMUST00000032518	j	2218	6:128612403-128627547
TCONS_00663622	5530400C23Rik	ENSMUST00000048459	j	1782	6:133242189-133246057
TCONS_00675723	A530053G22Rik	ENSMUST00000060147	j	2808	6:60345557-60353737
TCONS_00686793	5930416I19Rik	ENSMUST00000112152	j	1183	6:128306021-128312929
TCONS_00686826	Klrb1a	ENSMUST00000032512	j	2088	6:128559085-128573419
TCONS_00686877	Klrb1b	ENSMUST00000032472	j	3660	6:128763467-128777652
TCONS_00690369	Gm3104	ENSMUST00000171004	j	1454	7:3085852-3088038
TCONS_00714087	Nlrp4d	ENSMUST00000086269	j	637	7:10944249-10974565
TCONS_00757200	Cd209c	ENSMUST00000127592	j	1447	8:3940193-3954746
TCONS_00803647	Gm3867	ENSMUST00000086108	j	1222	9:36064625-36065962
TCONS_00823024	Enox	ENSMUST00000140767	j	3866	X:100688894-100701683

A.2.2 Differential expression testing

Among the previously known lncRNAs, 74 were significantly differentially expressed (48 of those between the A7 and S7 groups). Of the novel lncRNAs, two were differentially expressed between the case and control mice sacrificed after seven weeks. In total, 3508 transcripts were differentially expressed between A7 and S7 (8277 if all pairwise comparisons between A3, A7 and S7 are included).

The overall most significantly differentially expressed transcript isoforms (between A7 and S7) are summarised in table A.3. Similarly, the most differentially expressed lncRNA (both novel and previously known) are presented in table A.4.

Concordant with the results presented in the main analysis in section 4.3, histone proteins are among the ones displaying the largest differences in transcript abundance estimates in the comparison of the A7 and S7 groups of mouse samples (table A.3).

Sparc, however, was a gene not highlighted in the original results. The protein produced by this gene takes part in the regulation of extracellular matrix formation [67]. As previously mentioned in the main report, processes affecting the composition and eventual degradation of the extracellular matrix are involved in AAA disease.

Tyrobp has roles in the immune system and takes part in inflammation "via its coupling to myeloid receptors, such as the triggering receptors expressed by myeloid cells (TREM) displayed by neutrophils, monocytes/macrophages and dendritic cells" [68]. This is interesting in context of the inflammatory response observed during abdominalaortic aneurysm formation, in which macrophages (among other immune cells) are involved [3]. Another highly differentially expressed gene with connections to the immune system is Pdpn, which besides regulating cell motility [69] also "plays crucial roles in the biology of immune cells, including T cells and dendritic cells" [70]. An additional connection to the immune system among the genes in table A.3 is Mpeg1, which stands for "Macrophage Expressed Gene 1".

Angptl1 stands for Angiopoietin-Like Protein 1 and is known to display anti-angiogenic properties by "inhibiting the proiferation, migration, tube formation, and adhesion of endothelial cells" [71] and has additionally "been shown to exhibit antiapoptotic activity in human endothelial cells" [71] (endothelial cells line the inside of blood vessels). Given these descriptions and its high differential epression between case and control, Angptl1 could be suspected to have a role in the disease studied here.

It should also be mentioned, although its implications are rather unclear, that other than the above discussed genes, a significant number of predicted genes and pseudogenes were also present in table A.3.

Regarding the lncRNAs, ENSMUST00000130642 was the most differentially expressed transcript (table A.4). It stems from the gene "Sparc", which was discussed above. However, as such, one would probably suspect that its expression levels are a passive result of the overall expression of the gene (and its main protein), rather than indicative of any independent functionality for the noncoding isoform. However, that hypothesis remains to be tested.

Several NONCODE lncRNAs were differentially expressed. Though these remain to be characterized, and not much can be said about them. The exclusively lncRNA-producing genes 2410006H16Rik, F630028O10Rik, BC029722 and C330006A16Rik are also found among the transcripts in table A.4, though they remain unclassified. Snhg5 also remains to characterize further (as not much information could be found

about it). Similarly to table A.3, a multitude of predicted genes are also represented in table A.4.

Table A.3: The overall 35 most differentially expressed genes (based on individual isoforms) between AngII treated mice harvested after seven weeks (A7) and control mice harvested after seven weeks (S7). Column headers as in main text table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000109141	Gm15427	=	-Inf	0.00005	0.00636546
ENSMUST00000130642	Sparc	=	-Inf	0.00005	0.00636546
ENSMUST00000087714	Hist1h4j	=	-Inf	0.00005	0.00636546
ENSMUST00000078369	Hist1h2ab	=	-Inf	0.00005	0.00636546
ENSMUST00000099704	Hist1h3i	=	-Inf	0.00005	0.00636546
ENSMUST00000091751	Hist1h2an	=	-Inf	0.00005	0.00636546
ENSMUST00000091701	Hist1h3a	=	-Inf	0.00005	0.00636546
ENSMUST00000074245	Gm10119	=	-Inf	0.00005	0.00636546
ENSMUST00000089396	Gm6723	=	-Inf	0.00005	0.00636546
ENSMUST00000088648	Gm8759	=	-Inf	0.00005	0.00636546
ENSMUST00000097682	Rpl27-ps3	=	-Inf	0.00005	0.00636546
ENSMUST00000051412	Gm5512	=	-Inf	0.00005	0.00636546
ENSMUST00000119459	H3f3c	=	-Inf	0.00005	0.00636546
ENSMUST00000169025	Gm17510	=	-Inf	0.00005	0.00636546
ENSMUST00000122336	Gm12751	=	-Inf	0.00005	0.00636546
ENSMUST00000118323	Gm11936	=	-Inf	0.00005	0.00636546
ENSMUST00000117129	Pgam1-ps2	=	-Inf	0.00005	0.00636546
ENSMUST00000121308	Gm5931	=	-Inf	0.00005	0.00636546
ENSMUST00000083987	U6	=	-Inf	0.00010	0.01118960
ENSMUST00000117045	Dynlt1-ps1	=	-Inf	0.00010	0.01118960
ENSMUST00000117533	Gm12062	=	-Inf	0.00035	0.03008360
ENSMUST00000093651	U6	=	Inf	0.00040	0.03375630
ENSMUST00000050586	5430419D17Rik	x	-Inf	0.00055	0.04222160
ENSMUST00000006626	Actn3	=	-5.58226	0.00010	0.01118960
ENSMUST00000032800	Tyrobp	=	-5.35215	0.00005	0.00636546
ENSMUST00000055770	Hist1h1a	=	-5.31370	0.00005	0.00636546
ENSMUST00000077389	Gm7536	=	-5.28799	0.00065	0.04835630
ENSMUST00000030317	Pdpn	=	-5.25626	0.00005	0.00636546
ENSMUST00000027885	Angptl1	=	-5.18182	0.00015	0.01562130
ENSMUST00000080511	Hist1h1b	=	-5.15822	0.00010	0.01118960
ENSMUST00000040359	Arsi	=	-5.11103	0.00005	0.00636546
ENSMUST00000021506	Serpina3n	=	-5.09322	0.00050	0.03911860
ENSMUST00000003643	Ckm	=	-5.08662	0.00005	0.00636546
ENSMUST00000081035	Mpeg1	=	-5.06781	0.00005	0.00636546
ENSMUST00000163360	D17H6S56E-5	=	-5.05824	0.00065	0.04835630

Table A.4: The overall 35 most differentially expressed lncRNA (based on individual isoforms) between AngII treated mice harvested after seven weeks (A7) and control mice harvested after seven weeks (S7). Column headers as in main text table 4.2. Matches to protein coding genes may be observed, but those refer to noncoding isoforms of those genes.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000130642	Sparc	=	-Inf	0.00005	0.00636546
ENSMUST00000117533	Gm12062	=	-Inf	0.00035	0.03008360
ENSMUST00000131787	2410006H16Rik	=	-3.26653	0.00005	0.00636546
ENSMUST00000153077	Megf8	=	-2.46350	0.00025	0.02314610
TCONS_00545054	-	u	2.45603	0.00020	0.01973000
ENSMUST00000023292	Arsa	=	-2.45424	0.00005	0.00636546
ENSMUST00000034997	Snhg5	=	-2.44712	0.00005	0.00636546
ENSMUST00000137359	Gm15512	=	-2.40868	0.00060	0.04528880
ENSMUST00000147681	F630028O10Rik	=	-2.34503	0.00010	0.01118960
ENSMUST00000110601	Gpr137b-ps	=	-2.29320	0.00005	0.00636546
ENSMUST00000124586	BC029722	=	-2.25654	0.00005	0.00636546
ENSMUST00000133808	C330006A16Rik	=	-2.23649	0.00005	0.00636546
NONMMUT050162	NONMMUG031060	=	-2.21508	0.00035	0.03008360
NONMMUT022945	NONMMUG014195	=	-2.16479	0.00005	0.00636546
ENSMUST00000155507	Pgs1	=	-2.14105	0.00025	0.02314610
ENSMUST00000172805	Map3k8	=	-2.10502	0.00025	0.02314610
NONMMUT069993	NONMMUG043325	=	-2.03938	0.00005	0.00636546
ENSMUST00000155949	6530402F18Rik	=	-2.03331	0.00020	0.01973000
ENSMUST00000151488	Yipf3	=	-2.02016	0.00050	0.03911860
ENSMUST00000173269	Neu1	=	-2.00370	0.00010	0.01118960
ENSMUST00000164646	E530011L22Rik	=	-2.00007	0.00010	0.01118960
ENSMUST00000148314	Gm13889	=	-1.96165	0.00005	0.00636546
ENSMUST00000116685	Mir3096	o	-1.93738	0.00045	0.03700300
ENSMUST00000175820	A230050P20Rik	=	-1.92682	0.00005	0.00636546
NONMMUT045116	NONMMUG027834	=	-1.92268	0.00050	0.03911860
ENSMUST00000138323	Pfdn2	=	-1.76269	0.00010	0.01118960
NONMMUT032612	NONMMUG020085	=	-1.72129	0.00020	0.01973000
ENSMUST00000097503	Gm10524	=	-1.71945	0.00025	0.02314610
ENSMUST00000163793	Gm17255	=	-1.71467	0.00005	0.00636546
ENSMUST00000171743	Tmem134	=	-1.70072	0.00050	0.03911860
ENSMUST00000165292	Gm17282	=	-1.68657	0.00005	0.00636546
ENSMUST00000156380	2900053A13Rik	=	-1.65174	0.00065	0.04835630
ENSMUST00000134226	Gm12590	=	-1.62306	0.00015	0.01562130
NONMMUT072544	NONMMUG044953	=	-1.56763	0.00005	0.00636546
ENSMUST00000099459	Gm10780	=	-1.49290	0.00005	0.00636546

A.2.3 Co-expression network modules

Using WGCNA [22], 11 co-expression modules were obtained. A visualisation of the WGCNA dynamic tree cut module assignment can be seen in figure A.1. As previously, the modules were ranked based on correlation between the module eigengene and the treatment status vector. The best ranking modules are found in table A.5. As before, the modules were assigned labels in the form of arbitrary colours by the WGCNA software. From now on, the modules with be referred to by these colour names.

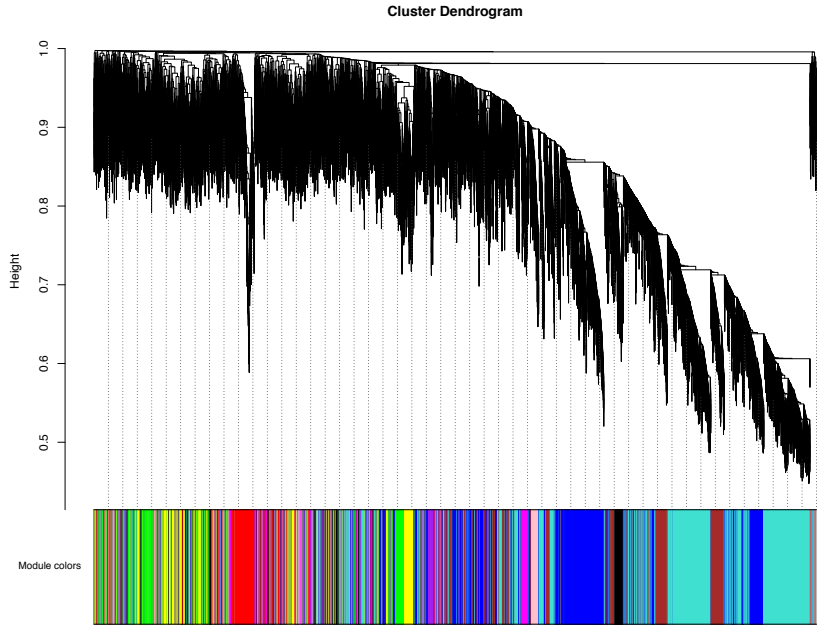


Figure A.1: The WGCNA module assignment of this dataset. The black dendrogram is obtained through hierarchical clustering of the weighted correlation based network. Modules are represented by individual colours and are the result of applying dynamic tree cutting to the dendrogram.

Table A.5: Ranking of modules according to absolute Pearson correlation between the eigengene of the module and a vector of treatment status. Each module is represented with a colour code corresponding to figure A.1. Only the most significantly correlated modules are shown here (absolute correlation greater than or equal to 0.80 and p-value less or equal to 0.05). "Size" refers to the number of transcript isoforms in the module.

Module	Correlation	p-value	Size	D.e lncRNAs	Novel lncRNAs
yellow	0.94	0.002	1313	2	14
pink	0.89	0.007	547	1	0
blue	0.88	0.008	3421	19	6
turquoise	0.84	0.02	4245	21	7

The "yellow" module

As can be seen in table A.5, the transcript isoforms in the "yellow" module appears to be expressed in patterns that correlate highly with the treatment. The genes of the

most central transcripts (in terms of betweenness centrality and degree, calculated with Cytoscape and visualised in figure A.2) of the yellow module were Gm16576, Ttk, Cd109, Fmod, BC034090, Trip13, Rasgrf1, 7SK, Tlr9, Ddah1, and C1qtnf3. Gm16576 is a predicted lincRNA gene with unknown function. BC034090 appears to an unclassified protein coding gene.

The expression of Gm16576, which can be seen in figure A.3, was higher in the "A7" mice than in the "S7" control, and the difference was statistically significant according to the "getSig" function of the CummeRbund R package, using an alpha cutoff of 0.05. Other significantly differentially expressed among the central nodes were Tlr9 (figure A.4), Fmod (figure A.5) (both in the comparison A7, S7 and A3, A7) and Trip13 (figure A.6) (in the comparison A7, S7). Tlr9 stands for "Toll-like receptor 9", Fmod stand for "fibromodulin" and Trip13 for "Thyroid Receptor Interacting Protein 13".

In total, 34 transcripts (6 lncRNAs) were significantly differentially expressed between any two samples in the "yellow" module, 21 (2 lncRNAs) between A7 and S7. Gm16576 was the only one of these that was previously annotated in Ensembl. The others were present in NONCODE. The remaining five differentially expressed lncRNAs were novel transcripts. The overall most differentially expressed transcripts in the "yellow" module can be seen in table A.7.

Table A.6: Differentially expressed lncRNAs in the "yellow" module. Column headers as in main text table 4.7.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value	Groups
ENSMUST00000128342	Gm16576	=	-1.2	5e-05	0.00636546	A7, S7
TCONS_00722067	-	u	-2.47005	5e-05	0.00636546	A3, A7

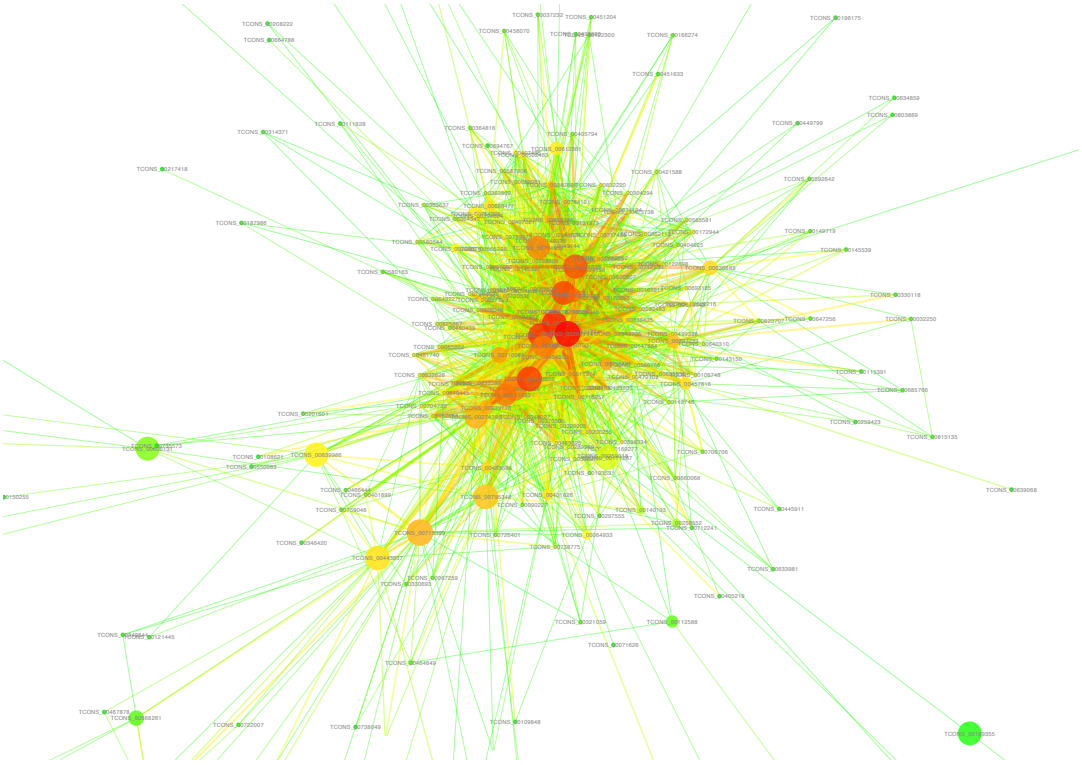


Figure A.2: The "yellow" WGCNA derived module. Figure elements are the same as was described in the legend of main text figure A.7. A high resolution version of the figure can be obtained from the author upon request.

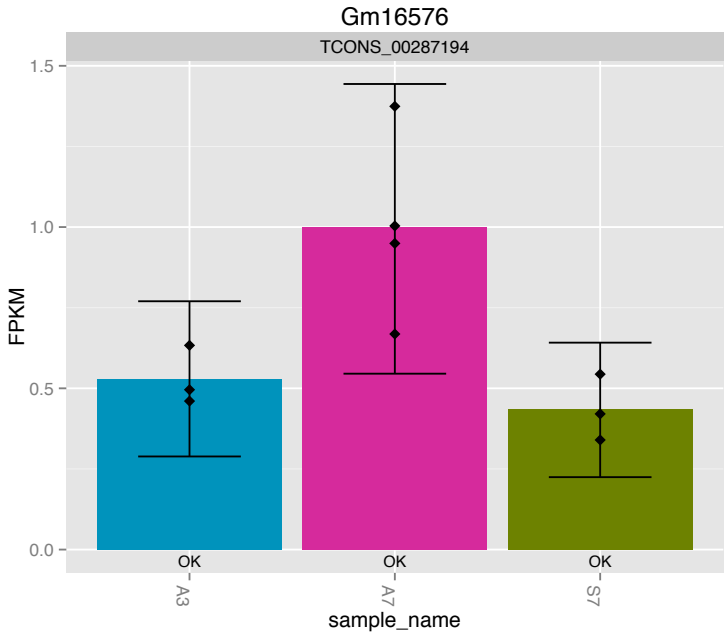


Figure A.3: Expression of the lincRNA Gm16576, the most central transcript of the "yellow" WGCNA module. The blue bar corresponds to the "A3" sample, pink to "A7" and green to "S7".

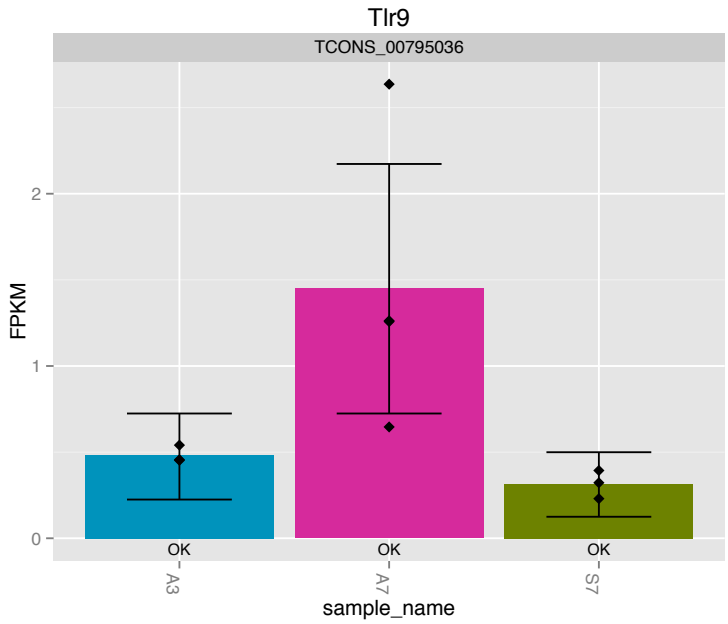


Figure A.4: Expression of Tlr9, one of the most central transcripts of the "yellow" WGCNA module. The blue bar corresponds to the "A3" sample, pink to "A7" and green to "S7".

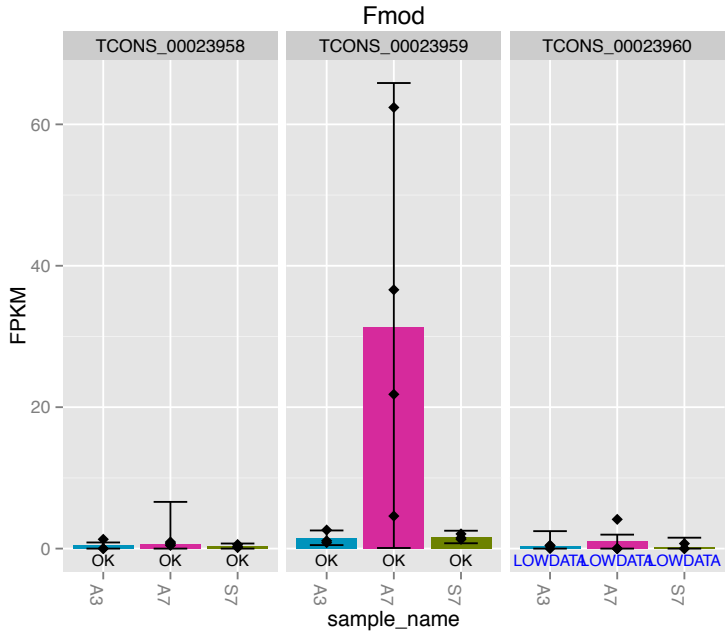


Figure A.5: Expression of Fmod. TCONS_00023959 was one of the most central transcripts of the "yellow" WGCNA module. TCONS_00023958 was a novel transcript overlapping TCONS_00023959. The blue bar corresponds to the "A3" sample, pink to "A7" and green to "S7".

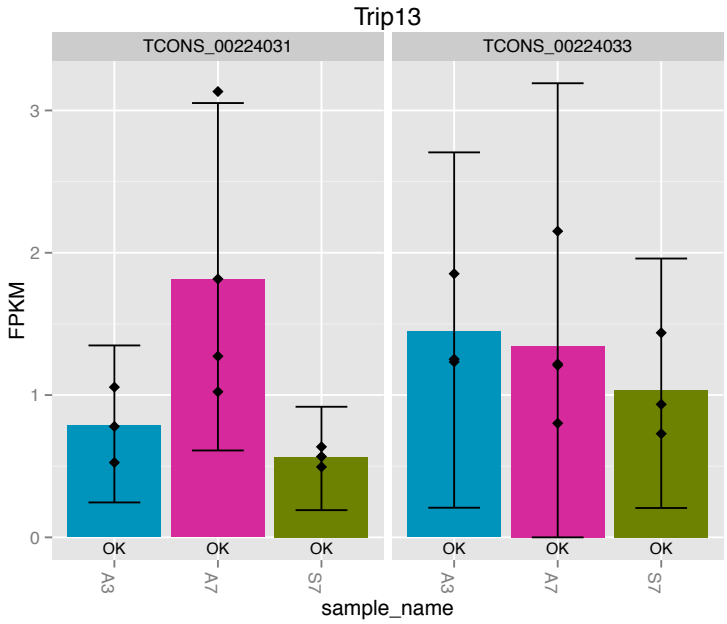


Figure A.6: Expression of Trip13. TCONS_00224031 was one of the most central transcripts of the "yellow" WGCNA module. TCONS_00224033 was a novel transcript isoform overlapping TCONS_0022403. The blue bar corresponds to the "A3" sample, pink to "A7" and green to "S7".

Table A.7: Differentially expressed transcripts between A7 and S7 in the "yellow" module. Column headers as in main text table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000087316	Capn6	=	-4.92985	0.00005	0.00636546
ENSMUST00000048183	Fmod	=	-4.25981	0.00005	0.00636546
ENSMUST00000060710	Cdc25c	=	-3.10131	0.00065	0.04835630
ENSMUST00000072119	Ccnb1	=	-2.74186	0.00020	0.01973000
ENSMUST00000029848	Col24a1	j	-2.61988	0.00020	0.01973000
ENSMUST00000076505	Pyroxd2	=	-2.57141	0.00005	0.00636546
ENSMUST00000020899	Matn3	=	-2.38223	0.00005	0.00636546
ENSMUST00000128727	Actr3b	=	-2.24899	0.00035	0.03008360
ENSMUST00000083546	Mir133b	=	-2.20249	0.00005	0.00636546
ENSMUST00000062241	Tlr9	=	-2.20233	0.00005	0.00636546
ENSMUST00000048246	Fgb	=	2.15004	0.00015	0.01562130
ENSMUST00000098382	Adamts17	j	-2.04165	0.00060	0.04528880
ENSMUST00000058825	Ccdc121	=	-1.96315	0.00005	0.00636546
ENSMUST00000032341	Art4	=	-1.82901	0.00020	0.01973000
ENSMUST00000098953	Mex3a	=	-1.74837	0.00005	0.00636546
ENSMUST00000090473	Gpr88	=	-1.72988	0.00005	0.00636546
ENSMUST00000022053	Trip13	=	-1.67750	0.00005	0.00636546
ENSMUST00000109212	Gm5431	=	-1.48573	0.00005	0.00636546
ENSMUST00000043183	Ces2g	=	-1.41782	0.00050	0.03911860
ENSMUST00000128342	Gm16576	=	-1.20000	0.00005	0.00636546
TCONS_00314402	-	u	1.06805	0.00030	0.02649240

The "blue" module

The central nodes of the "blue" module (visualised in figure A.7) were transcript isoforms from the genes Gns, Pctp, Heatr5a, Wdfy2, Eny2, Ext1, Ifnar1, Dscr3, B3gnt1, Ppp3ca, Hey1, Pole, Rpn1, Dera and C230081A13Rik. Several lncRNA were present in this module. Those can be found in table A.8. The 25 most significantly differentially expressed transcript isoforms in the "blue" module are shown in table A.9.

Seven of the central nodes corresponded to differentially expressed isoforms: Pctp, Wdfy2, Ifnar1, Dscr3, B3gnt1, Hey1 and Dera.

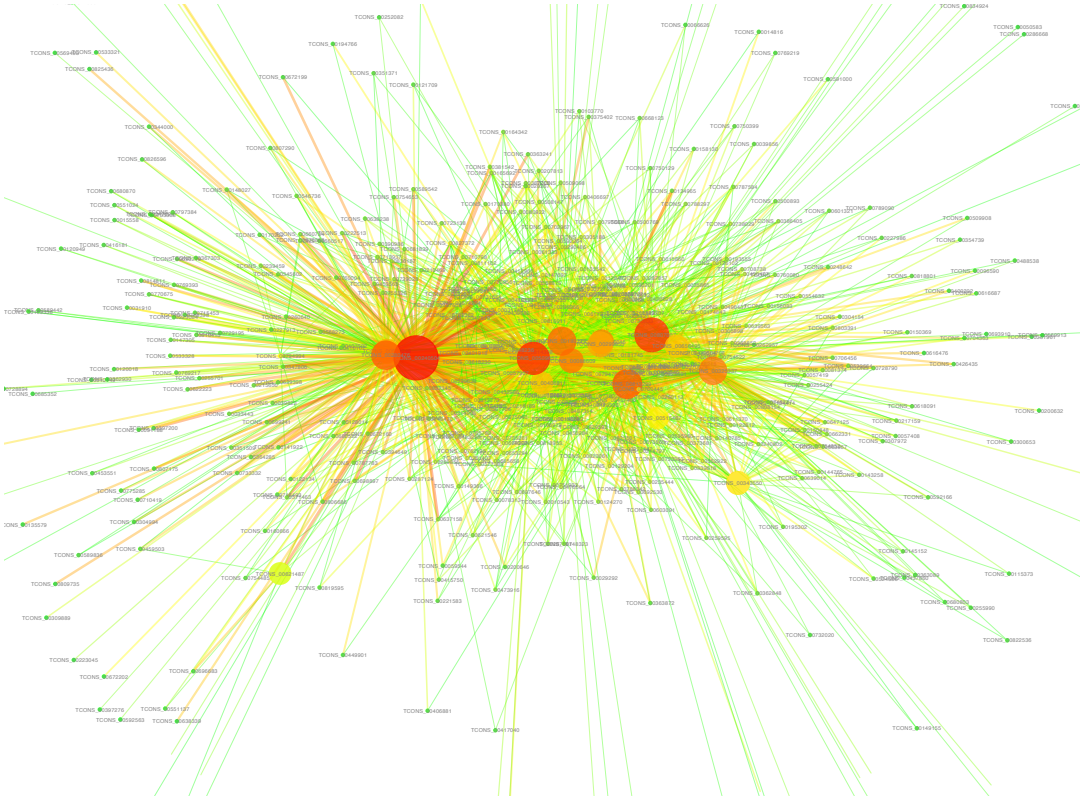


Figure A.7: The "blue" WGCNA derived module. Larger nodes indicate higher betweenness centrality. Figure elements are the same as was described in the legend of main text figure A.7. A high resolution version of the figure can be obtained from the author upon request.

Table A.8: Differentially expressed lncRNAs in the "blue' module. Column headers as in main text figure 4.7

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value	Groups
ENSMUST00000131787	2410006H16Rik	=	-3.27	0.00005	0.0064	A7, S7
ENSMUST00000153077	Megf8	=	-2.46	0.00025	0.023	A7, S7
ENSMUST00000023292	Arsa	=	-1.86; -2.45	0.00055; 0.00005	0.042; 0.0064	A3, S7; A7, S7
ENSMUST00000137359	Gm15512	=	-2.41	0.00060	0.045	A7, S7
ENSMUST00000147681	F630028O10Rik	=	-2.35	0.00010	0.011	A7, S7
ENSMUST00000110601	Gpr137b-ps	=	-2.29	0.00005	0.0064	A7, S7
TCONS_00592563	-	u	-2.22	0.00035	0.030	A7, S7
NONMMUT022945	NONMMUG014195	=	-2.16	0.00005	0.0064	A7, S7
ENSMUST00000155949	6530402F18Rik	=	-2.03	0.00020	0.020	A7, S7
ENSMUST00000151488	Yipf3	=	-2.02	0.00050	0.039	A7, S7
ENSMUST00000164646	E530011L22Rik	=	-2.00	0.00010	0.011	A7, S7
ENSMUST00000175820	A230050P20Rik	=	-1.93	0.00005	0.0064	A7, S7
ENSMUST00000171743	Tmem134	=	-1.70	0.00050	0.039	A7, S7
ENSMUST00000165292	Gm17282	=	-1.50; -1.69	0.00040; 0.00005	0.034; 0.0064	A3, S7; A7, S7
ENSMUST00000134226	Gm12590	=	-1.62	0.00015	0.016	A7, S7
ENSMUST00000146416	Rbm12b	=	-1.48	0.00005	0.0064	A7, S7
ENSMUST00000168014	Gm17401	=	-1.33	0.00005	0.0064	A7, S7
ENSMUST00000105140	AW011738	=	-1.21; -1.32	0.00050; 0.00015	0.039; 0.016	A3, S7; A7, S7
ENSMUST00000131147	Gm13387	=	-1.14	0.00045	0.037	A7, S7

Table A.9: The 25 most differentially expressed transcript isoforms between A7 and S7 in the "blue" module. Column headers as in main text table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000055770	Hist1h1a	=	-5.31370	0.00005	0.00636546
ENSMUST00000030317	Pdpn	=	-5.25626	0.00005	0.00636546
ENSMUST00000040359	Arsi	=	-5.11103	0.00005	0.00636546
ENSMUST00000081035	Mpeg1	=	-5.06781	0.00005	0.00636546
ENSMUST00000043069	Cyth4	=	-4.97811	0.00005	0.00636546
ENSMUST00000025497	Fbn2	=	-4.93516	0.00005	0.00636546
ENSMUST00000099703	Hist1h2bb	=	-4.77073	0.00005	0.00636546
ENSMUST00000024118	Clec4n	=	-4.76126	0.00005	0.00636546
ENSMUST00000021685	Hhipl1	=	-4.71655	0.00005	0.00636546
ENSMUST00000002883	Sfrp4	=	-4.68946	0.00005	0.00636546
ENSMUST00000006956	Saa3	=	-4.62616	0.00005	0.00636546
ENSMUST00000015664	Ctsk	=	-4.55248	0.00005	0.00636546
ENSMUST00000034774	Itga11	=	-4.54105	0.00005	0.00636546
ENSMUST00000024981	Hn1l	=	-4.48850	0.00005	0.00636546
ENSMUST00000018918	Cd68	=	-4.43995	0.00005	0.00636546
ENSMUST00000028897	Cpxm1	=	-4.42749	0.00005	0.00636546
ENSMUST00000060484	Clec4a1	=	-4.40459	0.00025	0.02314610
ENSMUST00000030202	Glipr2	=	-4.37610	0.00060	0.04528880
ENSMUST00000025419	Ppic	=	-4.31323	0.00005	0.00636546
ENSMUST00000087557	Tspan6	=	-4.29439	0.00005	0.00636546
ENSMUST00000003445	Fkbp11	=	-4.28323	0.00005	0.00636546
ENSMUST00000034742	Ccnb2	=	-4.27510	0.00010	0.01118960
ENSMUST00000004587	Clec11a	=	-4.21670	0.00005	0.00636546
ENSMUST00000004201	Col5a3	=	-4.20165	0.00005	0.00636546
ENSMUST00000086763	Emr1	=	-4.16421	0.00005	0.00636546

The "turquoise" module

The central nodes of the "turquoise" module (visualised in figure A.8) corresponded to the genes Tmem131, Atf6, Fig4, Tspan31, Itsn2, Arf6, Cpsf2, Ubr7, Hist1h4d, Fam120a, 2610301G19Rik, Ep300, Scaf8, Sos1, Trim56, Gm5578, Atp6v0d1. Significantly differentially expressed isoforms among these were from the genes Tmem131, Fig4, Arf6, Hist1h4d, 2610301G19Rik, Gm5578 and Atp6v0d1.

Differentially expressed lncRNA in this module are summarised in table A.10. The overall 25 most differentially expressed transcript isoforms, both coding and noncoding, are shown in table A.11.

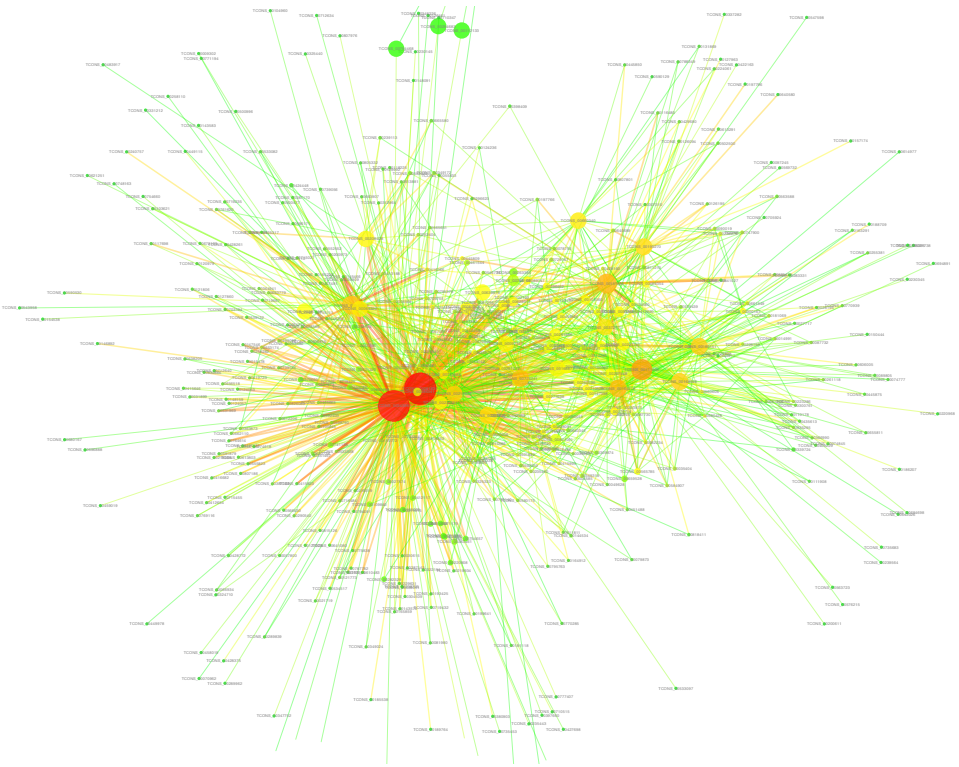


Figure A.8: The "turquoise" WGCNA derived module. Figure elements are the same as was described in the legend of main text figure A.7. A high resolution version of the figure can be obtained from the author upon request.

Table A.10: Differentially expressed lncRNAs in the "turquoise" module. Column headers as in main text table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value	Groups
ENSMUST00000148314	Gm13889	=	-3.97; -1.96; -2.00	0.00005; 0.00005; 0,00050	0.0064; 0.0064; 0.0064	A3, S7; A7, S7; A3, A7
ENSMUST00000023207	Apod	=	-3.59; -2.95	0.00010; 0,00050	0.011; 0.039	A3, S7; A3, A7
ENSMUST00000168817	Fcgr3	=	-3.49	0.00055	0.042	A3, S7
ENSMUST00000133808	C330006A16Rik	=	-2.91; -2.24	0.00005; 0.00005	0.0064; 0.0064	A3, S7; A7, S7
ENSMUST00000034997	Snhg5	=	-2.63; -2.45	0.00010; 0.00005	0.011; 0.0064	A3, S7; A7, S7
ENSMUST00000155507	Pgs1	=	-2.58; -2.14	0.00005; 0.00025	0.0064; 0.023	A3, S7; A7, S7
ENSMUST00000172805	Map3k8	=	-2.52; -2.11	0.00005; 0.00025	0.00637; 0.023	A3, S7; A7, S7
TCONS_00453772	-	=	-2.39	0.00045	0.037	A3, S7
TCONS_00789319	-	=	-2.38; -2.04	0.00005; 0.00005	0.0064; 0.0064	A3, S7; A7, S7
ENSMUST00000173269	Neu1	=	-2.00	0.00010	0.011	A7, S7
NONMMUT045116	NONMMUG027834	=	-1.99; -1.92	0.00045; 0.00050	0.037; 0.039	A3, S7; A7, S7
ENSMUST00000132305	Clcf1	=	-1.90	0.00045	0.037	A3, S7
TCONS_00386479	-	=	-1.82; -1.72	0.00015; 0.00020	0.016; 0.020	A3, S7; A7, S7
ENSMUST00000123544	Gm16516	=	-1.78	0.00020	0.020	A3, S7
ENSMUST00000135037	2810001G20Rik	=	-1.74	0.00060	0.045	A3, S7
ENSMUST00000070085	AI504432	=	-1.66	0.00005	0.0065	A3, S7
TCONS_00274088	-	=	-1.61; -1.28	0.00005; 0.00005	0.0064; 0.0064	A3, S7; A7, S7
ENSMUST00000164804	Gm17056	=	-1.60	0.00050	0.039	A3, S7
TCONS_00710288	-	=	-1.41	0.00045	0.037	A3, S7
ENSMUST00000070085	AI504432	=	-1.05	0.00015	0.016	A7, S7

Table A.11: The 25 most differentially expressed transcripts between A7 and S7 in the "turquoise" module. Column headers as in main text table 4.2.

Transcript ID	Gene	Class	$\log_2$ fold change	p-value	q-value
ENSMUST00000032800	Tyrobp	=	-5.35215	0.00005	0.00636546
ENSMUST00000021506	Serpina3n	=	-5.09322	0.00050	0.03911860
ENSMUST00000119853	Gm12174	=	-4.68687	0.00005	0.00636546
ENSMUST00000038144	Esm1	=	-4.63327	0.00005	0.00636546
ENSMUST00000005548	Hmox1	=	-4.56687	0.00005	0.00636546
ENSMUST00000116487	Lgals3	=	-4.41037	0.00005	0.00636546
ENSMUST00000118928	Gm13456	=	-4.33618	0.00005	0.00636546
ENSMUST00000033004	Il4ra	=	-4.22683	0.00005	0.00636546
ENSMUST00000025486	Lmnb1	=	-4.16269	0.00005	0.00636546
ENSMUST00000040772	Fermt3	=	-4.12244	0.00005	0.00636546
ENSMUST00000034214	Mt2	=	-4.07249	0.00005	0.00636546
ENSMUST00000071130	Alox5ap	=	-4.02396	0.00005	0.00636546
ENSMUST00000113440	Ccdc88b	=	-3.99798	0.00020	0.01973000
ENSMUST00000002678	Tgfb1	=	-3.98090	0.00065	0.04835630
ENSMUST00000100198	Bin2	=	-3.93078	0.00005	0.00636546
ENSMUST00000030651	Sh3bgrl3	=	-3.90295	0.00005	0.00636546
ENSMUST00000079957	Fcer1g	=	-3.88466	0.00005	0.00636546
ENSMUST00000021011	Ccl7	=	-3.88466	0.00020	0.01973000
ENSMUST00000069988	Xpnpep1	=	-3.86411	0.00005	0.00636546
ENSMUST00000058914	Tuba1c	=	-3.84341	0.00005	0.00636546
ENSMUST00000034215	Mt1	=	-3.69782	0.00005	0.00636546
ENSMUST00000102881	Plek	=	-3.69213	0.00005	0.00636546
ENSMUST00000034339	Cdh5	=	-3.66873	0.00005	0.00636546
ENSMUST00000111315	Adamts4	=	-3.64842	0.00050	0.03911860
ENSMUST00000021676	0610007P14Rik	=	-3.64035	0.00045	0.03700300

A.2.4 MetaCore© analysis

The "yellow" module

Constructing a network of curated relations among the most central nodes of the "yellow" module using the "expand by one" algorithm in MetaCore© gave the results shown in figure A.9. Among the genes connected to the central yellow nodes in the MetaCore© network, 16 were annotated as being relevant in the context of at least one of the following diseases: "Aneurysm, Ruptured"; "Aneurysm"; "Aortic Aneurysm"; "Aortic Aneurysm, Thoracic"; "Aortic Aneurysm", "Abdominal"; "Hypertension"; "Atherosclerosis"; "Intracranial Aneurysm".

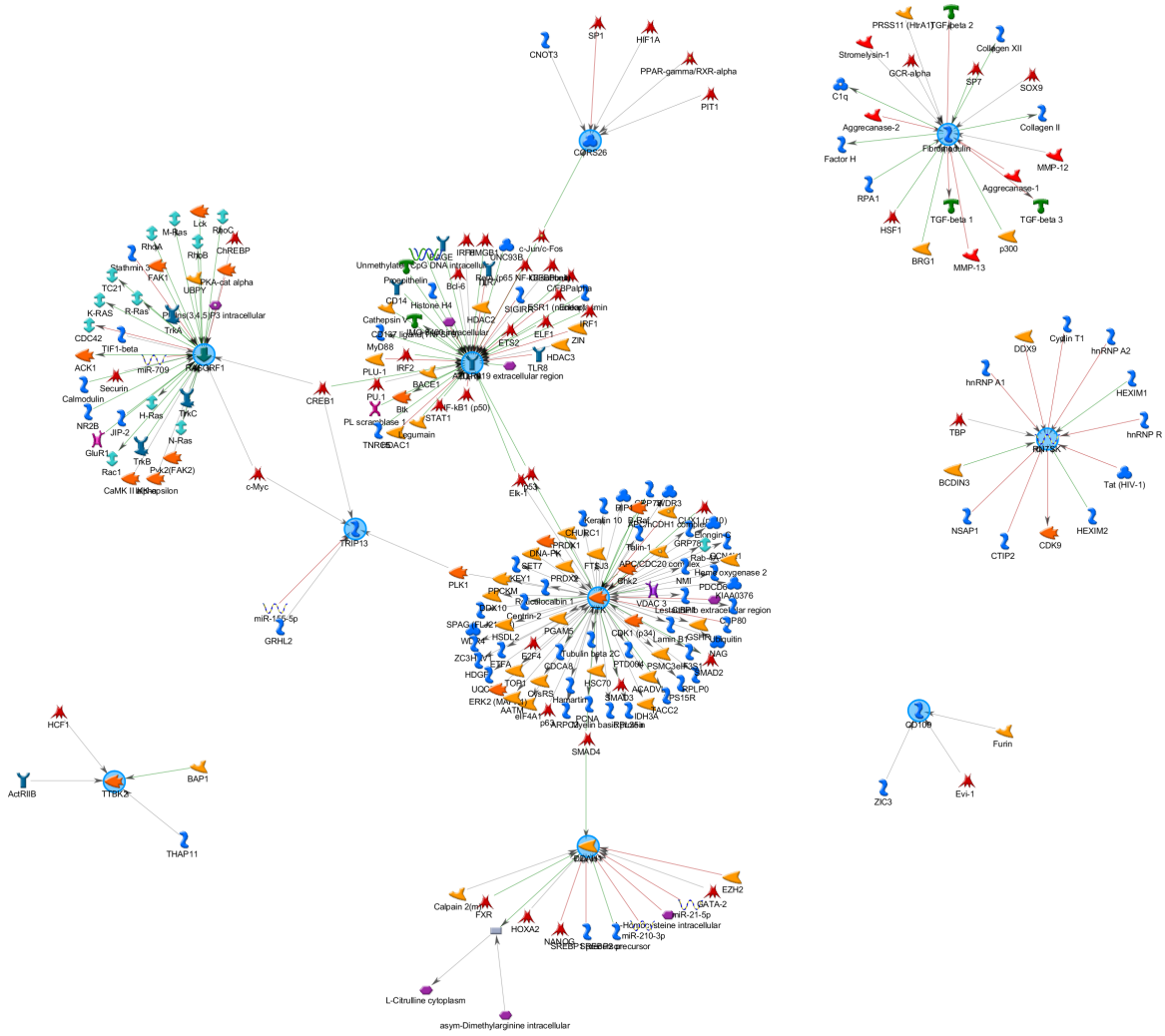


Figure A.9: MetaCore© interaction network constructed from central nodes of the "yellow" module. The "expand by one" algorithm was used (default settings). The gene identifiers were in some cases translated to the internal naming scheme of MetaCore©, but the original genes are indicated with blue circles. A high resolution version of the figure can be obtained from the author upon request.

The "blue" module

Constructing a network of curated relations among the most central nodes of the "blue" module using the "shortest paths" algorithm (without the option "use canonical pathways") gave the results shown in figure A.10. Among the nodes in the MetaCore© interaction network constructed from these genes, Tyk2, Calcineurin A and the androgen receptor were annotated as therapeutic targets. HEY1, 26S proteasome and the androgen receptor were related to the term "hypertension". Tyk2, IFN-alpha/beta receptor, DNA polymerase epsilon, c-Myc, c-Jun, the androgen receptor, HEY1, Calcineurin A, Huntingtin, 26S proteasome and CREB1 were related to "cardiovascular disease" according to the Metacore© annotation.

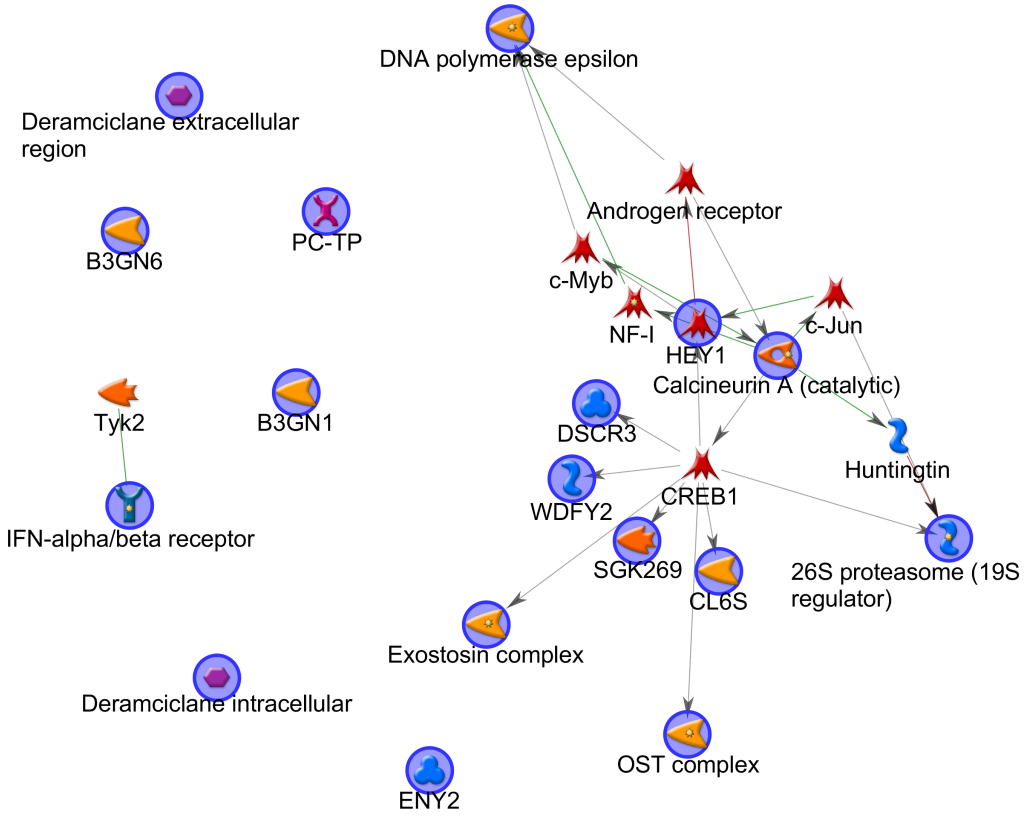


Figure A.10: MetaCore© interaction network constructed from central nodes of the "blue" module. The "shortest paths" algorithm was used (with "use canonical pathways" turned off). The gene identifiers were in some cases translated to the internal naming scheme of MetaCore©, but the original genes are indicated with blue circles.

The "turquoise" module

The MetaCore© network constructed from the genes belonging to the central transcripts from the "turquoise" module was built using the "shortest paths" algorithm (without the option "use canonical pathways"). Searching the network for therapeutic targets highlighted 13 nodes: H-Ras, ESR1, CDK1, TORC2, c-Src, p38 MAPK, HDAC3, HIF1A, the androgen receptor, c-Abl, EGFR, FGFR1 and VEGFR-3. Network objects related to either "Aortic Aneurysm; Abdominal", "Aortic Aneurysm", 'Aneurysm" or "Hypertension" were ATM, GCR-beta, AHR, Sirtuin 1, ESR1, HIF1A, PGC1-alpha, p85-alpha, BMAL1 and the androgen receptor.

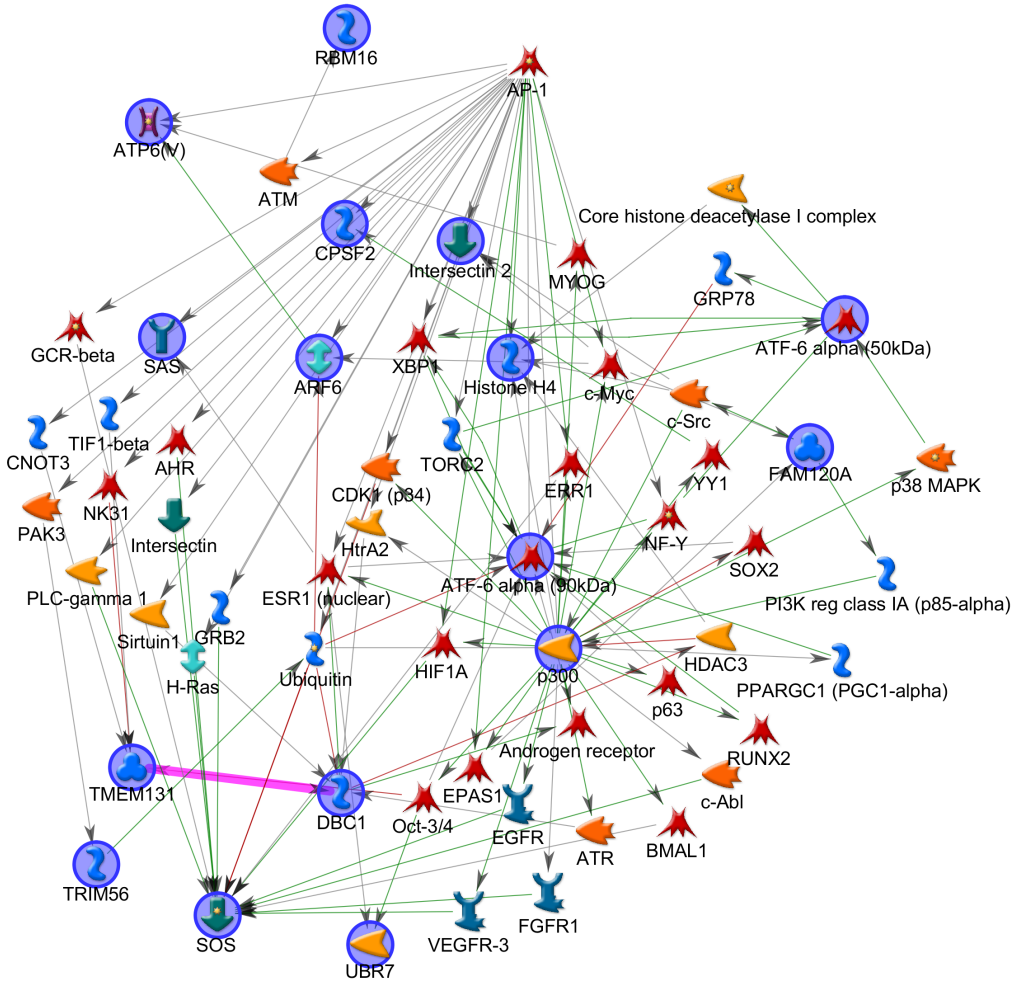


Figure A.11: MetaCore© interaction network constructed from central nodes of the "turquoise" module. The "shortest paths" algorithm was used (with "use canonical pathways" turned off). The gene identifiers were in some cases translated to the internal naming scheme of MetaCore©, but the original genes are indicated with blue circles.

A.2.5 Clustering comparison

Using the method mentioned in section 3.6.3, a Cramér's V of 0.025 was obtained when testing for association between WGCNA and k-means clustering. For a visualisation of the overlap between modules (clusters), see figure A.12. These results indicate that no significant association between k-means and WGCNA clustering was present (see the discussion in section 5).

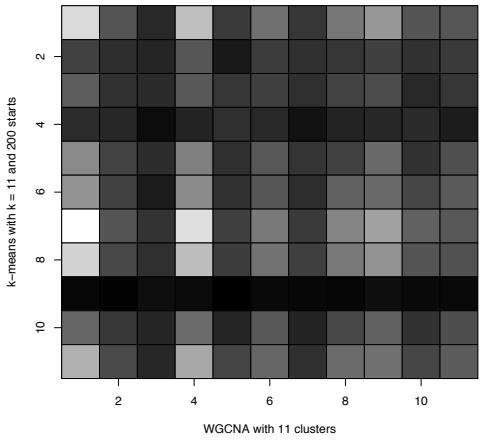


Figure A.12: The overlap of WGCNA module assignment and k-means clustering on the coefficient of variation filtered dataset. For k-means k was set equal to 11 and n = 200 starts of the algorithm was used. The image is a visualisation of the contingency matrix, wherein high values of a given cell is displayed brightly and low or zero overlap between two given clusters is displayed darkly shaded. (Additionally, in this image the number of transcripts in the intersection of each pair of modules have been normalised by the total size of the modules for greater clarity.)

A.3 Discussion

The results of the trimmed read analysis revealed a rather different picture. Other genes and long noncoding RNAs were highlighted as most likely to be important in context of the disease. Other novel lncRNAs were also discovered, (with the exception of one, which was identified in the original analysis as well).

It remains hard to determine which of the two analyses is the more correct one, especially given the multitude of parameters changed between the two:

The trimming procedure can be done in many ways, and there is no guarantee that the one performed here is optimal, or even better than performing no trimming at all.

In addition, the effectiveness of bias correction in Cuffdiff is unclear and may need to be examined further. One concern when using trimmed reads and bias correction in combination is that, since bias correction examines the ends of reads for overrepresented patterns, necessary information for effectively performing this correction may be lost during trimming. Though this would mostly be a concern if bias correction is thought to indeed be valuable to perform, something that was not entirely obvious throughout the analysis performed here. Further testing of the impact of such parameters is thus recommended.

Also, two different artifact filtering schemes were employed in the novel lncRNA discovery step of the two analyses, so those results need not be different entirely due to the trimming performed. And since multiple parameters of mapping, assembly and differential expression testing were changed (besides the trimming of the reads) between the original analysis and the supplementary one, it is not certain which factors contributed the most to the differences observed.

There is, however, one indication which makes the original analysis appear more trustworthy: The trimmed read based transcriptome assembly yielded many more "novel" RNA transcript isoforms (around 200 000 more than in the original analysis). There are two likely explanations for this: The first is that shorter reads map less specifically to the reference genome. The other, and perhaps the most important one, is that stricter read filtering leads to more coverage gaps in the structure of individual transcripts, so that spurious isoforms are discovered. Inspection of several of these "extra" novel fragments indeed revealed generally poor coverage (in comparison to annotated variants from the same genes). The vast majority of these are thus likely artifacts. It is not unlikely that poor isoform reconstruction may have negatively affected transcript abundance estimation (and therefore also differential and co-expression testing).

As such, it is recommended that these alternative results rather be seen as a complementary set of hypotheses to the old ones, while noting that the original is somewhat more likely to be correct. Finally, appendix B presents a table (B.1) comprising the top 50 long noncoding RNAs deemed most promising for further research in the context of AAA disease, weighing in results from both the original and the alternative analysis.

B. Candidate lncRNAs

Table B.1 presents 50 candidate lncRNAs recommended for further research. The list is based mainly on the results from the original (main article) analysis, with notes about differences to the alternative one (appendix A) included. The reason for choosing to focus on the original analysis was motivated by the principle of parsimony, based on the fact that the trimmed read analysis yielded far more (about 200 000 extra) novel assembled transcripts, many of which are most likely artifacts. The presence of these suspected artifacts additionally appeared to "dilute" the read coverage of many established (previously annotated) transcript isoforms, contributing to worse statistical significance in differential expression testing (and likely also affecting the co-expression analysis).

In selecting candidates for this list, intergenic lncRNAs were given higher priority than lncRNAs overlapping with protein coding genes. When such isoforms of protein coding transcripts appeared to closely follow the expression levels of the corresponding coding variants, these lncRNAs were excluded (if the relative abundance of the coding and non-coding variant does not change, one could suspect that the expression of the non-coding isoform is merely a passive side-effect of the expression of the protein coding variant, rather than indicative of independent lncRNA function). Additionally, differentially expressed lncRNAs were given higher priority than lncRNAs that were merely co-expressed in treatment correlated modules. Transcript isoforms with inconsistent annotation in online databases were excluded.

Table B.1: 50 candidate lncRNAs recommended for further research, ranked by absolute $\log_2$ fold change between the A7 and S7 groups (case and control at seven weeks) in the original analysis. "Module" refers to co-expression module assignment in the original anaysis (only modules in table 4.5 taken into account). LncRNAs that were significantly differentially expressed (on transcript level) between A7 and S7 in both the original and the alternative analysis were placed on top of the list. Differences between the two analyses are noted.

Transcript ID	Gene	Rank in orig.	Rank in alt.	Note	Module
ENSMUST00000131787	2410006H16Rik	10	3		
ENSMUST00000147681	F630028O10Rik	16	9		
ENSMUST00000034997	Snhg5	32	7		
ENSMUST00000133808	C330006A16Rik	43	12		
ENSMUST00000097503	Gm10524	69	28		
ENSMUST00000175820	A230050P20Rik	84	24		
ENSMUST00000164646	E530011L22Rik	114	21		
ENSMUST00000099459	Gm10780	118	35		
ENSMUST00000131248	5430416O09Rik	148	166		
ENSMUST00000156243	4732414G09Rik	169	46		
ENSMUST00000108969	Gm14401	173	49		
ENSMUST00000117266	Gm12245	1	-	*	
ENSMUST00000136359, ENSMUST00000140716	H19	3, 4	-	**	blue
ENSMUST00000143673	AI662270	19	-	*	
ENSMUST00000155083	Ppox	23	-	**	
ENSMUST00000132618	Eif1	24	-	*	
TCONS_00234018/TCONS_00348192 (novel)	17:30713409-30714566	25	-	***	
ENSMUST00000132340	Trem2	33	-	*	blue
ENSMUST00000145164, ENSMUST00000152828	Pisd-ps1	36, 75	-	*	
ENSMUST00000127901	Nsun7	38	-	*	turquoise
ENSMUST00000127652	Tmem59	39	-	**	
ENSMUST00000145089	Cd37	42	-	*	
ENSMUST00000129649, ENSMUST00000127981	A530020G20Rik	47, 70	-	*	
ENSMUST00000118528	Gm14293	54	-	*	
ENSMUST00000171670, ENSMUST00000166221	Snhg1	55, 82	-	*	
ENSMUST00000164690	Gm17586	56	-	**	
ENSMUST00000169511	Gm9917	63	-	*	
ENSMUST00000170675	Xpo6	66	-	**	
ENSMUST00000162030	Gm10075 (AC098736.2)	67	-	*	
ENSMUST00000131275	Hoxb3os	68	-	**	
ENSMUST00000170093	Snhg7	73	-	**	
ENSMUST00000148900	D4Wsu53e (Rsrp1)	74	-	**	
ENSMUST00000151043	1300002E11Rik	76	-	*	
ENSMUST00000152109	C430049B03Rik	79	-	**	blue
ENSMUST00000147076	Rrp1b	83	-	*	magenta
ENSMUST00000146721	Ndufaf4	87	-	*	
ENSMUST00000123292	Dynll1	90	-	**	
ENSMUST00000177539	RP23-164N15.3.1 (Gm20645)	94	-	*	
ENSMUST00000139026	Vsig10	97	-	*	
ENSMUST00000132305	Clcf1	99	-	*	
ENSMUST00000149667	9430008C03Rik	101	-	**	
ENSMUST00000063040	Minos1	113	-	**	
ENSMUST00000170099	D930016D06Rik	135	-	*	
ENSMUST00000152147	1810058I24Rik	136	-	*	
ENSMUST00000154405	Qrsl1	140	-	*	
ENSMUST00000145894	Gm14703	-	-	**	green
ENSMUST00000156068	6330403K07Rik	-	-	**	green
ENSMUST00000105240	Timeless	-	-	*	red
ENSMUST00000132193	Alkbh6	-	-	*	red

*Significantly differentially expressed on gene level in both analyses. Only significantly differentially expressed on transcript level in the original analysis.

**Only significantly differentially expressed in the original analysis. Overall expression pattern looks very similar.

***Novel lncRNA. Significantly differentially expressed in both the original and the alternative analysis ($\log_2$ fold changes -2.39/-2.64.). Not present in table A.4 since it was removed in the pre-filtering step of the novel lncRNA detection pipeline in the alternative analysis (likely due to too low read coverage).