Dissecting the genetic basis of transcriptional and translational regulation in heart and liver

Franziska Witte

Dissertation zur Erlangung des Grades

eines Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik

Berlin, 28.11.2018

Gutachter:

Prof. Dr. Martin Vingron

Prof. Dr. Norbert Hübner

Tag der Disputation: 14.05.2019

Acknowledgements

I owe my deep gratitude especially to Prof. Dr. Norbert Hübner for giving me the opportunity to work on this topic, for the continuing scientific support, and for the critical and inspiring comments throughout the years. I thank Prof. Dr. Martin Vingron for agreeing to supervise this PhD thesis and giving me the opportunity to join the lecture series of Max Planck Institute for Molecular Genetics. This allowed me to discuss work progress and problems of the project from a bioinformatics point of view.

Many thanks go to the entire Hübner and Lee group for the friendly working atmosphere, their reliability and their constant readiness to help. Special thanks go to Sebastiaan and Eleonora for their direct help with this project. I would like to also thank Valentin, Allison and Catherine for the end-less brainstorming sessions to solve computational problems, and Kathrin for always being there with an open ear. I would like to thank Matthias for the great introduction into the work of quantitative statistics and his patience.

And I want to thank Christin and Sarah for their friendship and support throughout the whole time. It was always nice to meet you for a chat in the kitchen or a walk around the campus.

Thanks go to Mona and Jana for joining the very early morning swim in Buch and the breakfast afterwards. I enjoyed it a lot and would have loved to continue it a bit longer.

Outside the lab, I would like to thank my best friend Steffi, who greatly supported me with being such a good friend during the time of this thesis that came with all ups and downs.

I thank my husband Roger for all the support during the past years. It would not have been possible to finish this thesis without his patience and his good advices. Thank you for your unconditional love and everlasting confidence in me.

My sincerest gratitude goes to my family, especially to my parents and my sister, who have supported me at all time.

Statement of contribution by others

Tissue was provided by Michal Pravenec (BxH/HxB rat RI panel, Congenic rat strains) and Daniel Sanchis (C57BL/6 EndoG -/- mice). Experimental work in order to generate RNA-Seq and Ribo-Seq data was carried out by Sebastiaan van Heesch, Anita Müller, Susanne Blachut and Eleonora Adami.

Oliver Hummel helped in setting up the genotype map used for QTL mapping. Giannino Patone performed demultiplexing and assisted in aligning all datasets.

The validation experiments mentioned in chapter 5.5 have been carried out by Samreen Falack and Maria Bikou (qPCR and Western Blots), Marieluise Kirchner (Mass spectrometry) and Camilla Ciolli Mattioli (Polysome profiling).

Summary

Gene expression regulation is a multi-layered process and genetic variation can modulate expression levels at various stages, for example by changing the activity of regulatory DNA elements or via nonsense mutations. One layer of regulation for which the genome-wide genetic effects have not been fully assessed yet is translational regulation. The main objective of this thesis is to elucidate the role of genetic variants on transcriptional and translational levels of gene expression. We therefore perform a combinatorial RNA-Seq, Ribo-Seq and genotyping approach to liver and heart tissue of a rat genetic disease model system for spontaneous hypertension and metabolic disease.

We use the well-established HxB/BxH rat recombinant inbred (RI) panel that consists of 30 lines that have been derived from a reciprocal cross of the spontaneously hypertensive rat (SHR.Ola) and the normotensive Brown Norway rat (BN.Lx/Cub). The RI lines, which have been previously comprehensively described on multiple levels, enable us to perform quantitative trait loci (QTL) mapping in order to link causal genetic variants to quantitative differences in both transcription and translation.

We identify local and distant associations on the transcriptional (eQTL) and the translational (riboQTL) level in each tissue. The majority of detected associations occurred between SNPs and genes that are in close proximity to each other and are termed local QTL. However, we also detect genetic variants that affect gene expression levels over longer distances, termed distant associations. In this thesis, we assessed both types of QTL in detail in order to explain the mechanistic and genetic basis of these QTL. Interestingly, we observe 17 genes regulated by a single distant riboQTL on chromosome 3. To follow-up this finding, we performed an independent RNA-Seq and Ribo-Seq experiment in 2 congenic rat strains that differ specifically in this riboQTL region. The differential expression analysis on both layers of regulation reproduces the results and suggests that the chromosome 3 locus is in control of translation of many additional genes.

In hypertensive hearts, we observe global translatome changes that appear protein-length dependent. Translation rates of long proteins are downregulated and shorter proteins, including many ribosomal subunits, tend to be upregulated. As this phenotype is absent from the liver, we hypothesize that the chromosome 3 locus hosts a genetic variant responsible for a cardiac-specific state of translational stress in strained hearts.

These data provide insight into the genetic effects on multiple layers of gene expression regulation and show that translational regulation is an important mediator of molecular phenotypes in complex diseases.

Zusammenfassung

Die Regulation der Genexpression ist ein vielschichtiger Prozess, welcher auf verschiedenen Ebenen von genetischen Varianten modifiziert werden kann. Beispiele sind Mutationen, die die Aktivität von regulatorischen DNA-Elementen verändern oder die zu einem Aminosäureaustausch führen. Die genetische Regulation der Translation wurde bisher noch nicht genomweit untersucht. Das Ziel dieser Doktorarbeit ist es, die Rolle genetischer Varianten in der Genexpression auf Transkriptions- und Translationsebene zu verstehen. Dafür haben wir eine Kombination aus RNA-Seq, Ribo-Seq und Genotypisierung in der Leber und im linken Ventrikel durchgeführt.

Für diese Studie wird das Rattenmodell der rekombinanten Inzuchtlinie HxB/BxH benutzt, bestehend aus 30 Rattenstämmen, welche durch wechselseitige Kreuzung eines Rattenmodells für spontan entwickelten Bluthochdruck (SHR.Ola) und einem Modell mit normalem Blutdruck (BN.Lx/Cub), generiert wurde. Diese Inzuchtlinie erlaubt es uns, Regionen eines quantitativen Merkmals (*quantitative trait loci* (QTL)) zu bestimmen, um kausale genetische Varianten mit quantitativen Unterschieden, wie zum Beispiel Transkriptions- und Translationslevel, zu assoziieren.

Auf transkriptionaler (eQTL) und translationaler (riboQTL) Ebene haben wir in beiden Geweben hunderte von Assoziationen identifiziert. Die meisten Assoziationen wurden in Genen gefunden, die in der Nähe der assoziierten genetischen Variante liegen und als lokale QTL bezeichnet werden. Allerdings haben wir ebenfalls genetische Varianten gefunden, die Gene über weite Distanzen beeinflussen, diese werden entfernte QTL genannt. In dieser Arbeit werden beide QTL-Arten im Detail untersucht, um deren Mechanismen und genetische Grundlagen zu verstehen.

Interessanterweise, haben wir 17 Gene gefunden, die von einem einzigen, entfernt liegenden riboQTL auf Chromosom 3 reguliert werden. Um diese Beobachtung näher zu untersuchen, haben wir ein unabhängiges RNA-Seq und Ribo-Seq Experiment in zwei kongenen Rattenstämmen durchgeführt, die sich spezifisch in der Region des entfernten riboQTL unterscheiden. Die differentielle Genexpressionsanalyse dieser kongenen Ratten auf beiden Ebenen konnte die Ergebnisse reproduzieren und validieren. Dies deutet darauf hin, dass die Region auf Chromosom 3 in der Lage ist, die Translation vieler anderer Gene zu kontrollieren. Wir beobachten, dass globale Veränderungen des Translatoms in hypertrophen Herzen mit der Länge des translatierten Proteins korrelieren. Das Translationslevel von langen Proteinen scheint herunterreguliert zu sein und das von kurzen Protein hochreguliert.

Dieser Datensatz liefert Einblicke in die Rolle genetischer Varianten auf zwei Ebenen der Genexpressionsregulation und zeigt, dass die Regulation der Translation ein wichtiger Mediator vom molekularen Phänotypen zu einer komplexen Krankheit ist.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Genetic research

The aim of genetic researchers is to elucidate the role of heritable factors shaping a phenotype, which is a trait or characteristic of an organism that can be observed or measured. Those phenotypes can be physiological traits, such as size, age and hair colour; or molecular traits like transcription or translation levels of a gene. Phenotypic variation can be caused by variable expression of genes, the influence of environmental factors or the combination of both. The intent is to understand the genetic impact on the phenotype by linking these factors to molecular mechanism.

As many human diseases are known to have a genetic component, the understanding of molecular mechanism shaping a disease phenotype is very important, because these insights can guide the development of therapies and disease marker screens. On the other hand it allows to study the basis of molecular disease processes.

In molecular genetics the term heritable traits corresponds to genes in the genome. Diploid organism, such as animal, plants and human have two variants of each gene, that are also called alleles and they are building the basis of genetic diversity. An allele can be homozygous, if both alleles are the same or heterozygous with respect to the gene, if the alleles are different. Sometimes these different alleles change the phenotypic traits, but most genetic variations result in little or no observable difference in the expression of a gene.

Eukaryotic gene expression regulation ranges from accessibility or structure of the chromatin to transcription factors that can bind to specific DNA sequences close to the gene in order to promote or repress the transcription into pre-mRNA. The pre-mRNA is processed into the mature mRNA after splicing, capping and polyadenylation, which enables the mRNA to be exported from the nucleus to the cytoplasm, where each mRNA has a certain lifetime that determines how many proteins can be made from this mRNA. Next, the mRNA is translated into a polypeptide chain and subsequently further processed in order to create a functional protein (see Figure 1.1.1).

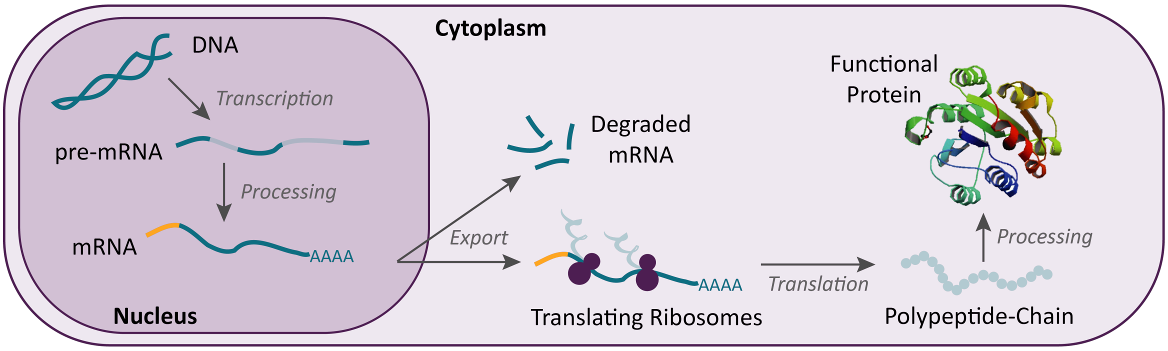


Figure 1.1.1: Schematic of gene expression

This shows that gene expression is a multi-layered process and it can be regulated at almost every layer by genetic variants. In this work, we therefore study the role of genetic variants at two layers of gene expression regulation, transcription and translation, in order to understand how genes can be regulated at different or even at multiple stages. We make use of molecular genetic tools to perform genetic mapping of quantitative trait loci (QTL), which are genetic loci that alter the expression of a gene.

1.2 Linking genetics to quantitative traits

Complex traits are shaped by genetic loci that cause a quantitative alteration on a physiological or molecular phenotype; and are called quantitative trait loci (QTL). Each genomic locus carries one or multiple genetic variants that can be linked to the abundance of a certain expression level. QTL that describe transcriptional changes are called eQTL, those that describe translational changes are riboQTL and protein changes are often abbreviated as pQTL.
A significantly associated QTL is a genetic variant that segregates a population into two groups of animals or humans carrying a specific genotype pattern matching the abundance of a gene. This is depicted schematically in Figure 1.2.1 on the example of a SNP changing the translation levels of a gene (riboQTL).

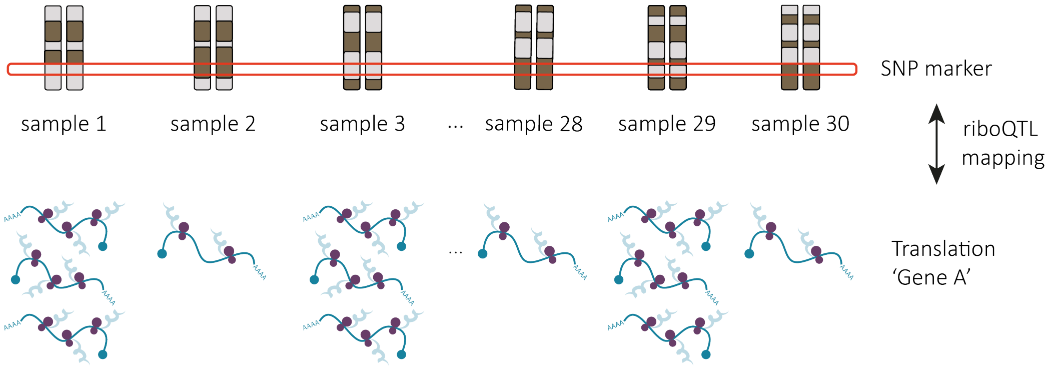


Figure 1.2.1: Schematic of a gene regulated by a riboQTL

QTL can be further defined by their location with respect to the gene that is altered. The genetic variant can either be located in close proximity to the gene (local QTL) or the variant is more distantly located with respect to the gene (distant QTL). Local QTL can act in *cis*, which means that the gene expression is modulated in an allele-specific manner, or in *trans*, via additional factors. In order to distinguish between local *cis* and *trans* QTL, an allele-specific expression analysis of heterozygous variants can be performed, where an observed imbalance suggests a *cis*-regulatory expression change.

Local QTL can be either coding variations, which are variants located in the coding sequence of a gene, that lead to amino acid changes, premature stop codons or splice defects that ultimately result in a non-functional protein. On the other hand regulatory variants can cause an expression change by affecting regulatory sequences. Distant QTL affect the expression of a gene through intermediate factors over longer distances. If the expression of multiple genes is altered by a single variant located elsewhere in the genome this variant forms a distant QTL cluster (Figure 1.2.2). The distance between the genetic locus and the affected gene is ranging from 10 kilobases (kb) in yeast to several megabases (Mb) in human in order to be defined as a distant association.

The first genome-wide eQTL mapping was carried out by Brem et al. in 2002 comparing the mRNA abundance of all expressed genes among a recombinant panel derived from two yeast strains [1]. eQTL mapping enables us to understand how the information encoded on the DNA level is translated to a certain phenotype or even a disease condition via molecular pathways and gene expression modulation. Since then, many eQTL studies have been published including eQTL mapping in mammalian hearts. In this thesis, a combined eQTL-riboQTL mapping approach in mammalian heart and liver tissue is presented. This enables us for the first time to understand the differ-

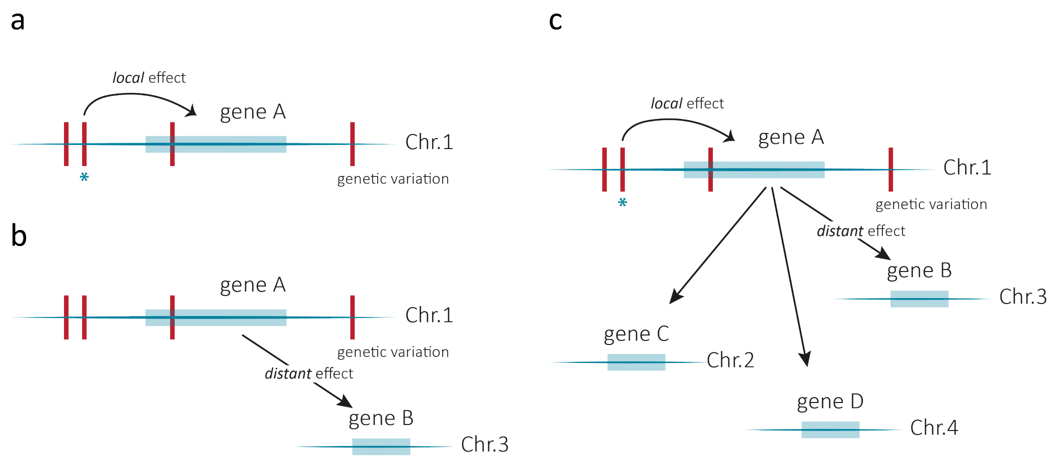


Figure 1.2.2: Different types of QTL: a) local QTL, b) distant QTL and c) distant QTL cluster

ences in regulation between transcription and translation caused by genetic variants. Adding translation as a second layer of gene expression regulation allows us to study an intermediate phenotype between the transcriptome and the proteome that ultimately lead to the observed phenotype. To date only few eQTL studies and no riboQTL involve mammalian hearts and only two were performed on human heart tissue [2, 3].

1.2.1 Studying the role of genetic variants in gene expression regulation

Translation can be regulated by genetic variants via various different mechanisms. Variants can either affect the transcription and/or translation of a single mRNA, a subset of mRNAs or change the transcriptional/translational landscape of all mRNAs. Transcriptional regulation can be modulated directly by variants changing a cis-regulatory element of an affected gene, as for example the promotor, enhancer, transcription start-site (TSS) or splice-site of the gene. Furthermore, eQTL can change the transcription levels that affect the expression of a gene in trans and over longer distances via intermediary factors, such as transcription or splice factors.

Translational gene expression changes can be forwarded from the transcriptional level with or without additional translational regulation. Local riboQTL can be explained by changes in regulatory sequences, such as internal ribosomal entry sites (IRES) or binding sites for translational regulators like RNA binding proteins (RBPs) and microRNAs (miRNAs). Coding variants can change translation levels by newly created or removed upstream open reading frames (uORFs) that affect the translation levels of the main open reading frame (mORF) or by truncating variants that lead to the translation of a shorter version of the protein. Global changes are often caused by activation or repression of components of the translational machinery through the action of *trans*-acting proteins that bind to the *cis*-elements of the regulated genes (Figure: 1.2.3).

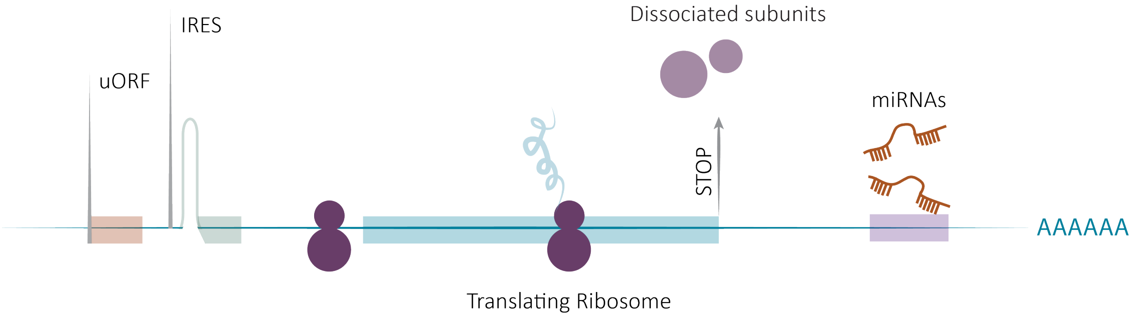


Figure 1.2.3: Schematic of different translational regulators that might explain local riboQTL. uORF - upstream open reading frame, IRES - internal ribosomal entry site, miRNA - micro RNA.

Trans-acting molecules, such as miRNAs and RBPs, can lead to changes in translation rates by binding to the mRNA and blocking the translation initiation site for ribosomes or stabilise the mRNA. miRNAs are known translational regulators that can stimulate the degradation of their target mRNA. A direct inhibition of translation by blocking ribosomal entry sites or prevention of the ribosome moving along the mRNA was previously described as a mechanism how miRNAs affect translation of their target mRNAs [4, 5]. Additionally, miRNAs have been shown to affect nascent peptide turnover [4], by e.g. promoting the translation of uORFs.

The translation of uORFs can lead to a decrease of translation of the mainORF (mORF) by competing ribosomes [6–8]. There are two classes of uORFs that have been described in the literature [9]. Those uORFs that are independent of the mORF and located in the 5' UTR of the gene (Figure 1.2.4 a), and overlapping uORFs, which initiate in the 5' UTR and overlap the coding region of the gene (Figure 1.2.4 b).

Even though overlapping uORFs are in a different frame than the mORF, it is more difficult to qualitatively and quantitatively assess them. The role of uORFs in disease is not clear to date. On the one hand, it was described that mutations in uORFs play a role in various diseases including cardiac disease [7], on the other hand Chew et al. claim that not all uORFs are able to efficiently repress the translation of the mORF

[8].

Another class of translational regulator are RBPs, which are involved in many regulatory processes during gene expression ranging from RNA processing to translation. There are thousands of known RBPs in humans [10]. RBPs that are part of the translation machinery and result in global changes are for example initiation or elongation factors. Another class of RBPs are ribosomal proteins that have the potential to change the composition of the ribosome. A third class of RBPs bind and regulate their target genes specifically and cause expression changes of their targets [11].

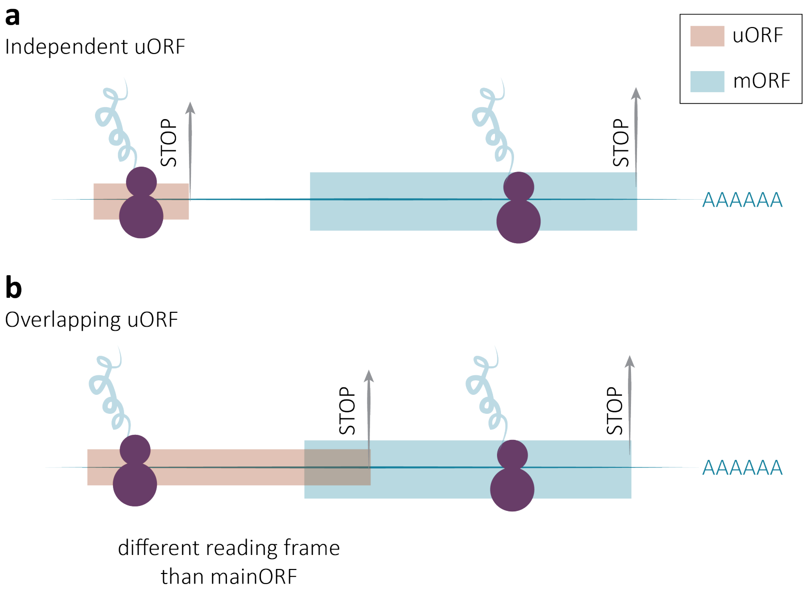


Figure 1.2.4: Schematic of independent and overlapping uORFs. a) Independent uORFs are located upstream of the mORF and do not overlap with it. b) Overlapping uORFs also start upstream of the mORF but overlap with it, with translation occurring in a different reading frame.

Even though the structure and function of the ribosome has been studied in great detail, we still lack knowledge about the regulation of translation [12]. The combination of transcription and translation levels, serve as a good proxy for the synthesis rate of a protein [13–15]. Especially in biosynthetically active cells (e.g. cardiomyocytes, monocytes, fibroblasts), protein synthesis accounts for a large proportion of the energy budget of the cell [16], which makes it particularly important to understand the molecular mechanism that shape the phenotype. In order to study the two layers of gene expression regulation, we measure the transcriptome levels by RNA-Sequencing (RNA-Seq) and the translatome levels by Ribosome-Sequencing (Ribo-Seq).

1.2.2 Quantifying genetic information in mammalian tissue samples

RNA-Seq is a next-generation sequencing approach to generate high quality transcriptome data on a genome-wide scale. In order to achieve this massive parallel sequencing of cDNA libraries, cells are lysed, total RNA is extracted and selected for polyadenylation. The polyadenylated (polyA) transcripts are enriched, fragmented and synthesized into cDNA libraries, followed by an amplification of the final library by a polymerase chain reaction (PCR). Ultimately, the libraries are sequenced to reveal a quantitative measure of RNAs in a biological sample at a given time point (Figure: 1.2.5) [17].

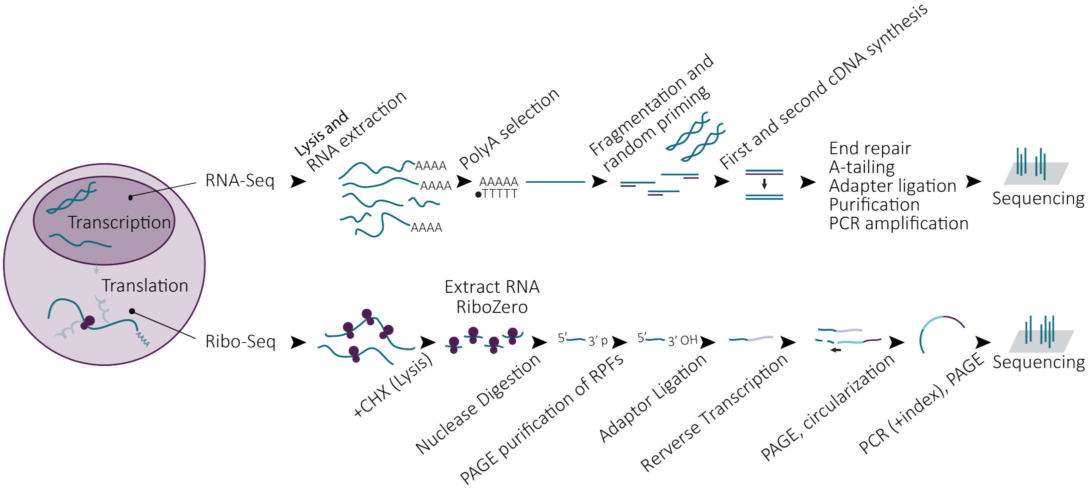


Figure 1.2.5: Schematic of RNA-Seq and Ribo-Seq protocol

RNA-Seq allows identification of various changes in the transcriptome, ranging from gene splicing and genetic variants to changes of gene expression over time, under different conditions/treatments or different genetic backgrounds [18–20]. Additionally, RNA-Seq data can be used to generate a new transcriptome assembly for species without a reference or to improve incomplete reference transcriptomes [21]. This approach can be applied to different types of RNAs, such as polyadenylated (polyA) mRNA, total RNA, microRNA or ribosome protected mRNA fragments with only minor changes to the protocol for library preparation.

Sequencing ribosome protected mRNA fragments on a genome-wide scale is called Ribo-Seq or ribosome profiling. Ribo-Seq allows to quantify the number of ribosomes that are actively translating mRNA into proteins and experimentally annotate the

coding regions of a gene. Ribo-Seq was the first global, *in vivo* approach to measure translation rates in a comparable way to RNA-Seq described by Ingolia et al. in 2009 in yeast [22]. The protocol of Ribo-Seq is similar to the RNA-Seq protocol, but the first crucial step is to fix the ribosomes on the RNA. This step is followed by a nuclease digestion step that removes the RNA which is not protected by the ribosome. Subsequently, ribosomes are removed by adding proteases. The remaining mRNA fragments, also called ribosome footprints, are reverse transcribed and sequenced similarly to RNA-Seq (Figure: 1.2.5).
The ribosome harbours three binding sites for tRNAs: i) the A-site, which is the first binding site of the ribosome where the incoming aminoacyl-tRNA bind; ii) the P-site that links the tRNA to the growing peptide chain and iii) the E-site which is the site where the tRNA is released as depicted in Figure 1.2.6.

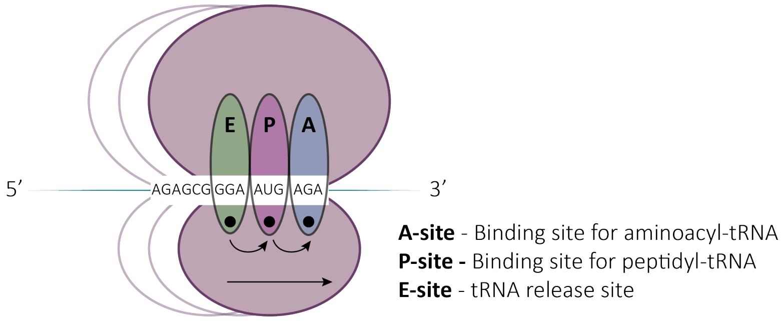

E
P
A
5'
AGAGCG
GGA
AUG
AGA
3'
A-site - Binding site for aminoacyl-tRNA
P-site - Binding site for peptidyl-tRNA
E-site - tRNA release site

Figure 1.2.6: Three ribosomal binding sites: A-site - Binding site for aminoacyl-tRNA, P-site - Binding site for peptidyl-tRNA and E-site - tRNA release site. During active translation the ribosome moves in a 3nt-periodic manner (codon-wise), which allows to infer the reading frame of the ribosome.

Ribo-Seq captures on average 30 nucleotides (nt) footprint reads that can be mapped to the original mRNA to define the exact location of the ribosome. This allows the position of the P-sites of the reads to be estimated, enabling the periodic movement of the translating ribosome to be inferred.
In addition to quantifying the translation rates of known protein-coding genes, Ribo-Seq enables the identification of novel translated short open reading frames (sORFs), such as genes that have been previously annotated to be non-coding and ORFs that are located in the 5' and 3' untranslated region (UTR) of the gene, upstream and downstream open reading frames (uORFs and dORFs).

1.3 QTL mapping in a model system for hypertension and metabolic syndrome

In this work we study the impact of genetic variants on transcriptional and translational regulation in the BxH/HxB recombinant inbred panel (RI panel), which was derived by crossing the parental strains, a disease model for hypertension and metabolic syndrome, the Spontaneously Hypertensive rat (SHR.Ola) and a normotensive control strain Brown Norway rat (BN.Lx/Cub). We make use of this model system for a complex human disease, to gain insights into the molecular mechanism that shape the disease phenotype. The use of animal models to study human diseases has several advantages, such as the ability to control for environmental factors. This allows to study gene expression differences that are caused to a large degree by genetic variation between two strains. For studying cardiovascular diseases the rat is the most commonly used model organism [23].

1.3.1 Phenotype of the parental strains SHR.Ola and BN.Lx/Cub

SHR.Ola combines genetic components of two diseases leading to cardiac hypertrophy - hypertension and the metabolic syndrome. The SHR strain was created by Okamoto et al. in 1963, by selective outbreeding of Wistar-Kyoto (WKY) rats with high blood pressure [24]. The SHR inbred line served as a model to study the genetic components of multiple diseases.

SHR rats develop hypertension at an age of 5-6 weeks with a systolic blood pressure (SBP) of 180-200mmHg. First characteristics of cardiovascular diseases, such as cardiac hypertrophy, can be observed at an age of 10-12 months. In addition to hypertension the SHR animals show typical features of metabolic syndrome with increased visceral adiposity, insulin resistance and abnormal lipid and glucose metabolism. [25].

In order to study these two disease components, we assess the gene expression in two disease relevant tissues, namely heart and liver. Analysing gene expression changes in the heart can help to understand molecular changes due to high blood pressure and beginning cardiac hypertrophy. As described before the metabolic syndrome is mainly straining the liver and we want to observe first changes on the molecular level that contribute to the metabolic changes in liver.

The BN.Lx/Cub, here abbreviated by BN.Lx, represents a normotensive control strain and is a congenic rat strain derived by introducing the mutant *Lx* gene of the polydactylous rat onto the BN background. Polydactylous syndrome leads to malformations of the hindlimb. As this locus does not play a role in heart disease, we can use this strain as a control. The BN reference strain is known to be susceptible to respiratory inflammation and to the development of mercury-induced autoimmunity [26]. The BN genome is divergent from most other rat strains and so it was used to encode the rat reference genome. The BN and BN.Lx/Cub rat differ at about 102,000 SNPs, 627,000 indels and 13,000 structural variants whereas SHR.Ola differs at 3.6 million SNPs and 343,243 indels from the reference genome [27].

1.3.2 Role of transcriptional and translational regulation in SHR

In the past, gene expression differences between SHR.Ola and BN.Lx/Cub have been studied comprehensively [28–32], including the Hübner lab, that show extensive translational regulation by performing differential expression analysis on RNA-Seq and Ribo-Seq data of the two strains BN.Lx/Cub and SHR.Ola [32].
The findings from Schäfer et al. illustrate that gene expression regulation cannot be explained by transcriptional regulation alone (Figure 1.3.1). Up to this point, transcription levels have been the most studied layer of gene expression to understand regulation. Schäfer at al. state that additional 40% of differentially expressed genes were picked up by adding translational information.
Figure 1.3.2 shows the enrichment of differentially transcribed and translated genes for known pathways that play a role in hypertension and metabolic syndrome. They identified a number of differential genes involved in fatty acid metabolism as well as in hypertrophic cardiomyopathy. Adding Ribo-Seq information enables to observe the enrichment in two cardiac pathways that would have been missed otherwise, which suggests that besides transcriptional regulation also translational regulation is shaping the SHR phenotype.
In order to understand the genetic components of these translational changes observed in the two parental strains, we use the rat RI panel to perform QTL mapping in the same two tissues on the transcriptional and translational level.

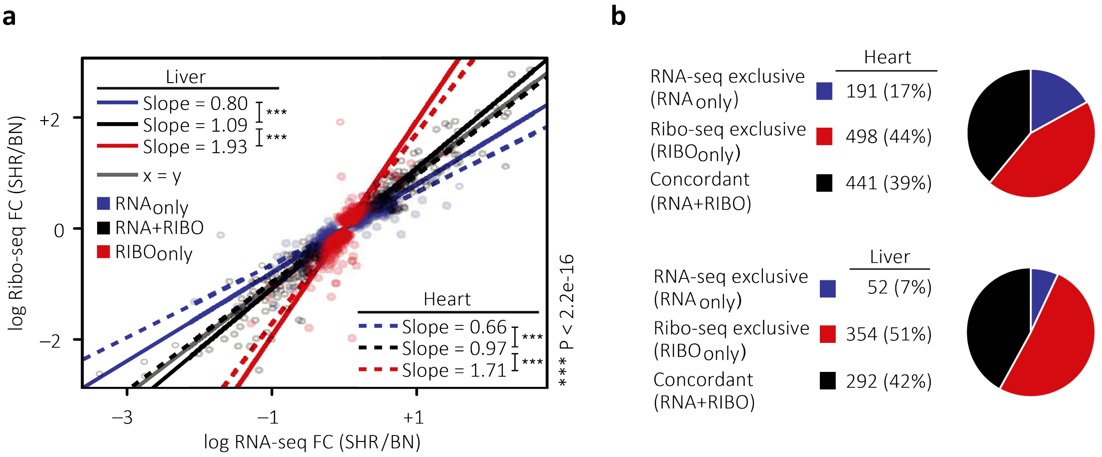
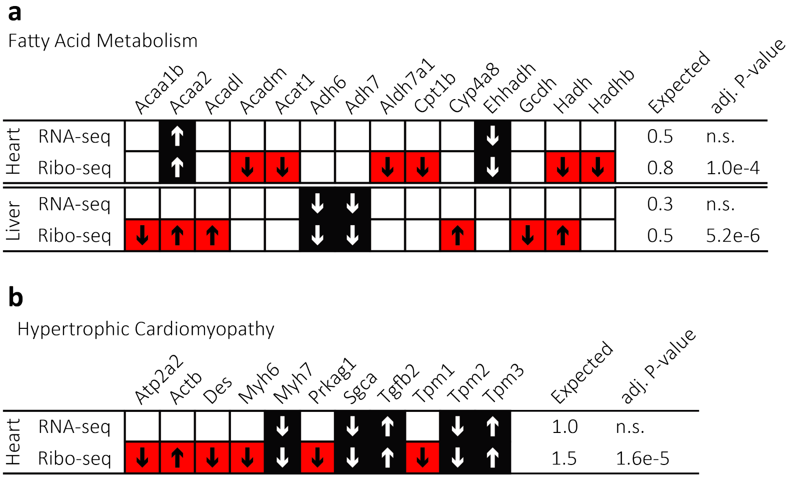


Figure 1.3.1: Transcriptional and translational regulation in the SHR heart and liver: a) RNA-Seq fold change (FC) FC(SHR/BN) vs. Ribo-Seq FC (SHR/BN) shows the extent of translational regulation by grouping genes based on whether they show a significant difference on the RNA and Ribo-Seq level (black), only on the RNA level (blue) or only on the Ribo-Seq level (red). The slopes that have been assessed by standardised major axis estimations show that the blue and the red line deviate from the black line that follows the diagonal. b) Number and percentages of genes that have been detected for the three categories RNA_{only}, $Ribo_{only}$ and RNA+Ribo in heart and liver. (Taken from [32])


a

Fatty Acid Metabolism

		Acaa1b	Acaa2	Acadl	Acadm	Acat1	Adh6	Adh7	Aldh7a1	Cpt1b	Cyp4a8	Ehhadh	Gcdh	Hadh	Hadhb	Expected	adj. P-value
Heart	RNA-seq		↑									↓				0.5	n.s.
	Ribo-seq		↑		↓	↓			↓	↓		↓		↓	↓	0.8	1.0e-4
Liver	RNA-seq						↓	↓								0.3	n.s.
	Ribo-seq	↓	↑	↑			↓	↓			↑		↓	↑		0.5	5.2e-6

b

Hypertrophic Cardiomyopathy

		Atp2a2	Actb	Des	Myh6	Myh7	Prkag1	Sgca	Tgfb2	Tpm1	Tpm2	Tpm3	Expected	adj. P-value
Heart	RNA-seq					↓		↓	↑		↓	↑	1.0	n.s.
	Ribo-seq	↓	↑	↓	↓	↓	↓	↓	↑	↓	↓	↑	1.5	1.6e-5

Figure 1.3.2: Translational and transcriptional regulation in SHR of genes in a) fatty acid metabolism and b) Hypertrophic Cardiomyopathy. Arrows in the boxes indicate whether a gene is up- or down-regulated in the SHR compared to BN. (Taken from [32])

1.3.3 Recombinant inbred panel

The rat recombinant inbred (RI) panel, which consists of 30 fully inbred animals, was used to elucidate the role of genetic variants on gene expression regulation in a heart disease context.

In detail, the RI strains have been generated by crossing BN.Lx and SHR.Ola resulting in recombined F2 animals (Figure 1.3.3). Independent segregation and recombination of homologous chromosomes lead to unique combinations of maternal and paternal genes in each F2 strain. Next, randomly selected F2 animals were inbred by brother sister mating over 80 generations. Gender reciprocal crossing was performed to generate a second set that only differs in the source of mitochondrial DNA and the Y chromosomes. HxB abbreviated strains are derived by crossing female SHR.Ola and male BN.Lx rats and BxH by crossing male SHR.Ola and female BN.Lx [33] (Figure 1.3.3).

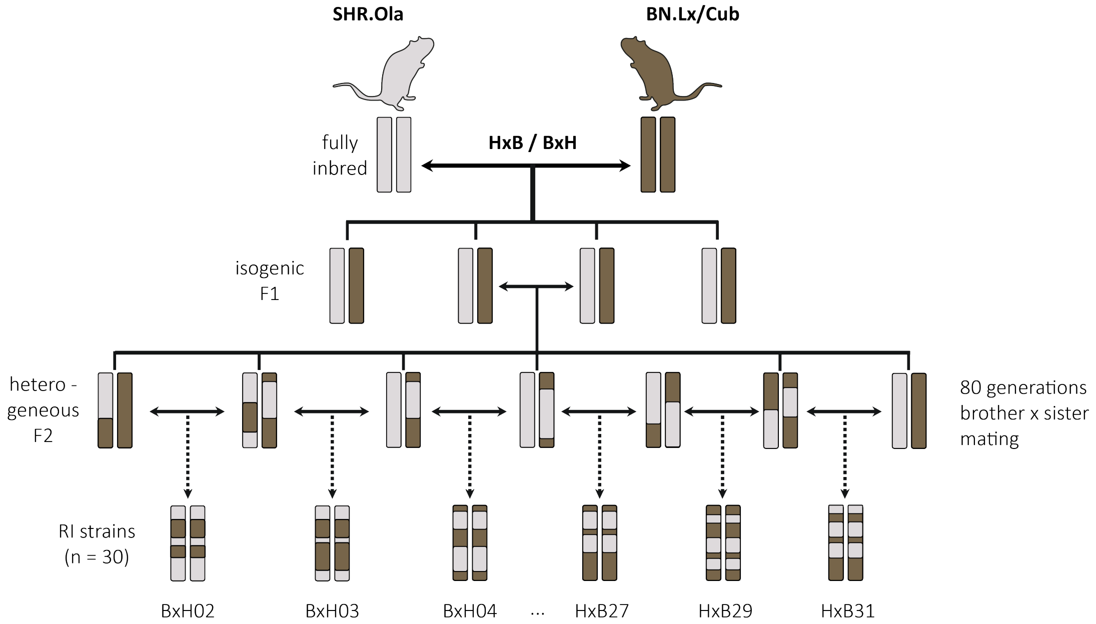


Figure 1.3.3: Breeding scheme of rat recombinant inbred panel that was derived from the two parental strains Bn.Lx/Cub and SHR.Ola by 80 generations of brother and sister mating.

All animals are homozygous for the BN.Lx/Cub or SHR.Ola allele at every position in the genome. Animal models of segregating populations derived by crossing two phenotypically contrasting inbred lines, are well suited to carry out association studies in order to link genetic variants and molecular traits such as gene abundance. Their lower complexity compared to human genomes and their reduced dimensionality of the different data collected, enables to derive associations that act locally, but also distant associations with smaller sample sizes than those that would be necessary to perform a genome-wide QTL mapping study in a human cohort.

In the past, various labs, including the Hübner lab used the RI panel to study the genetic component of numerous cellular-level and whole-animal traits. In 2005, one of the first multi-tissue eQTL analyses was carried out using the RI panel [34]. The authors first assessed all differentially expressed genes in the parental strains BN.Lx and SHR.Ola in fat and kidney tissue and compared those results with eQTL findings of the RI panel in the same tissues. Thousands of commonly and specifically regulated genes in fat and kidney have been identified. In general, the authors observed that distant associations tend to be more tissue specific than local ones.

In 2014, Rintisch et al. published an integrative approach combining eQTL mapping in heart and liver with genetic variants linking to histone modifications (histoneQTLs) in order to describe the impact of histone modifications on gene expression regulation [35]. Rintisch et al. were able to build models that can be used to predict gene expression based on the genetic variant and the histone level. Additionally, the RI panel has been used to study the genetic impact on whole genome phenotypes, such as blood pressure, cardiac mass and lipid levels and to identify a variety of disease genes [32, 34, 36–42].

In summary this shows, that the RI panel provides the perfect framework to study the genetic basis of translationally regulated genes observed by Schäfer et al. [32].

1.4 Motivation to perform eQTL and riboQTL mapping in the rat RI panel

The aim of this project is to identify genetic variants that drive transcriptional and/or translational changes in rats that suffer from the metabolic syndrome and spontaneous hypertension (Figure 1.4.1).

To reach that goal we performed a combinatorial Ribo-Seq, RNA-Seq and genotyping approach in liver and left-ventricular tissue of 30 animals belonging to the HxB/BxH recombinant inbred (RI) panel. Because each of these 30 RI lines is a fully inbred genomic mosaic of the two parental strains BN.Lx/Cub and SHR.Ola, we are able to perform quantitative trait loci (QTL) mapping to link causal genetic variants to quantitative differences in transcription and translation.

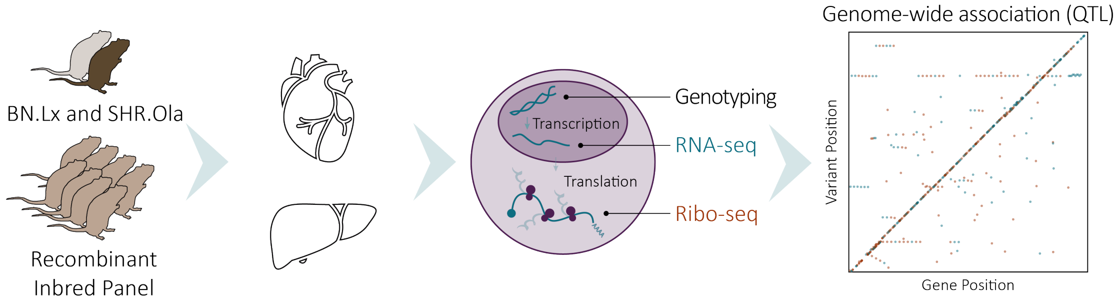


Figure 1.4.1: Project overview - we performed a combinatorial RNA-Seq, Ribo-Seq and genotyping approach to liver and left-ventricular tissue of the rat RI panel and the two founder strains BN.Lx/Cub and SHR.Ola in order to perform genome-wide eQTL and riboQTL mapping.

Currently little is known about translational control by genetic variants. Since the first protein QTL analysis was performed in 1994 on a limited number of 74 genes in maize [43], just a small number of pQTL studies on incomplete sets of proteins have been performed. Genome-wide protein QTL studies are at the moment not feasible, because the assessment of the protein levels in a genome-wide scale is very laborious and due to the wide dynamic range of proteins, particularly in the heart, very challenging. Even the largest mass spectrometry dataset in the heart from Doll et al. 2017 comprising 16 human heart regions and 4 cardiac cell types can only assess protein levels of roughly 6000 genes [44].

RNA-Seq and Ribo-Seq data are read count based quantitative measures of gene expression and can therefore be studied in a comparable way. Ribo-Seq experiments enable to study an additional layer of gene expression regulation that is closer to protein levels and ultimately to the physiological phenotype. As we get genome-wide measures of translation rates, we can understand post-transcriptional changes that are potentially forwarded onto the protein level. Ribo-Seq data can serve as a good proxy for protein levels allowing to obtain a better understanding of genetics in disease.

There are many advantages of studying the rat RI panel as a model for hypertension and metabolic disease compared to studying human patients that suffer from essential hypertension and the metabolic syndrome. Disease relevant mutations are difficult to study in human as they are usually rare and only detectable in very large cohorts, whereas in the rat RI panel every variant in present in approximately 15 lines and in a homozygous form. Additionally, patients with cardiovascular diseases

often suffer from other diseases at the same time, take different medications or other environmental factors are present, that contribute to the molecular phenotype. In the rat RI panel, the rats are all fed, raised and treated the same way which reduces confounding factors to a minimum and allows us to map the molecular traits that contribute to disease. The rat in comparison to other animal models resembles the genetic and phenotypic features of humans very well and is described as one of the best models for studying cardiometabolic diseases. Due to their size, measuring of cardiac phenotypes and deriving tissue is more feasible compared to mice [45].
We therefore, generated RNA-Seq and Ribo-Seq data in all 30 rat RI lines of two tissues - left ventricle and liver - in order to study the genetic impact on the metabolic syndrome and hypertension.

2 Datasets

2.1 RNA-Seq and Ribo-Seq data of the rat recombinant inbred panel in left ventricle and liver

The main dataset consists of RNA-Seq and Ribo-Seq data generated from liver and left-ventricular tissue in the 30 recombinant inbred lines that form the RI panel. Thirty male animals, each representing one RI strain, have been sacrificed at the age of 6 weeks and left ventricles (LV) and livers have been collected and immediately frozen. The animals were fed with a normal diet and housed under normal conditions in the Institute of Physiology, Academy of Sciences of the Czech Republic, Prague and provided by Michal Pravenec. RNA-Seq (according to TruSeq Stranded Total RNA and mRNA protocol) and Ribo-Seq (according to the TruSeq Ribo Profile protocol) was performed as described in [32]. Genotype information was derived using the Affymetrix RATDIV SNP Array of the HxB/BxH rat RI panel with >500K high quality markers as described in [46]. The RATDIV array includes in total 805,399 SNPs selected based on available sequence data for 13 rat strains (SHR, SD, SHRSP, SS, WKY, GK, F344, SR, BB, FHH, LEW, PVG, DA). This allows the determination of all genetic differences between the two founder rats BN.Lx and SHR.Ola, and therefore all combinations of the RI panel are covered in the SNP Array.

Ribo-Seq and RNA-Seq cDNA libraries have been sequenced on an Illumina HiSeq 2000 platform. For RNA-Seq we derive 2x101bp reads using a stranded, paired-end approach and for Ribo-Seq we observe shorter single-end (1x50bp) reads according to the protocol. The average RNA integrity number (RIN) of all samples is above 9.1, which indicates high quality of the RNA that was used for sequencing (Figure: 2.1.1).

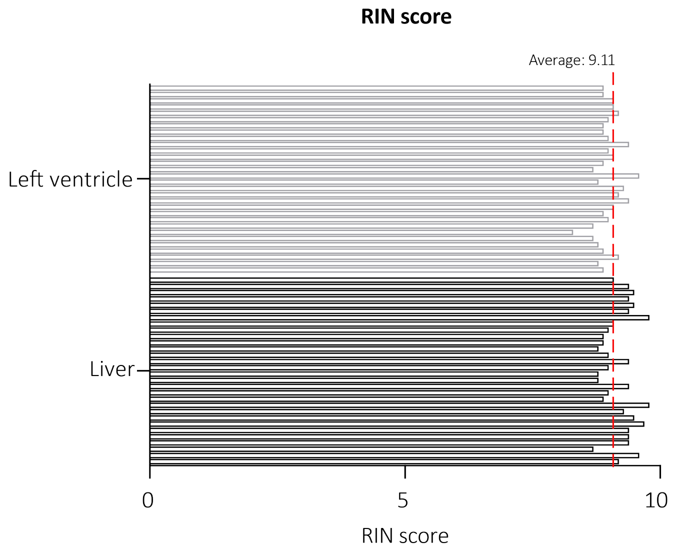


Figure 2.1.1: RNA integrity number (RIN) of all 30 RI lines in left ventricle and liver. The average RIN is at 9.11 and the lowest RIN at 8.3.

Additional quality measures, as for example the sequencing depth, sequencing quality and library complexity are discussed in detail in the chapter: 4.1.
The RI panel contains 2,957 different local combinations of the BN.Lx and SHR.Ola genomes. Each individual strain is homozygous at every locus across the whole genome for either the BN.Lx or the SHR.Ola genotype. The individual strains represent 15 replicates at each locus, which gives sufficient power for QTL mapping. Nevertheless, QTLs that can be identified in these studies span large chromosomal regions of sizes up to several Mb (average size of an SDP is 0.75Mb), leaving the underlying genes and mechanisms unknown. Additional fine-mapping of the QTL region is usually necessary to pinpoint the causal gene and variant as many genes can be located in the associated loci that are equally likely to be implicated in disease.

2.2 RNA-Seq and Ribo-Seq data of congenic rats and EndoG knock-out mice in left ventricle

A recent publication discovered endonuclease G (EndoG) as a driver for cardiac hypertrophy and mitochondrial function in the heart by using systems approaches combining QTL mapping and differential expression analysis in two congenic rat strains of this region [47]. This QTL region is located on chromosome 3 and overlaps the gene *EndoG*. The SHR.Ola rat carries a frameshift insertion of 37nt in the first exon of this gene that leads to the absence of the EndoG protein. McDermott et al. showed that EndoG is a determinant of cardiac hypertrophy and mitochondrial function. Additionally, this region is implicated a changed heart weight without changed the blood pressure [47] .

The two congenic rat strains (Figure 2.2.1) have been derived by a selective breeding regimen, which was employed to integrate segments of BN.Crl (Charles River Laboratories) chromosome 3 onto an SHR.Ola background.

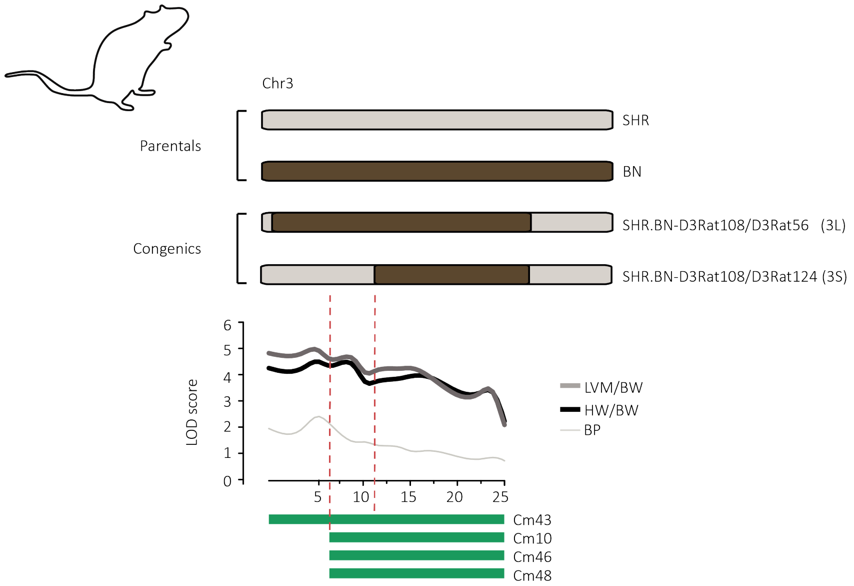


Figure 2.2.1: Illustration of the two rat congenic lines with association result of mapping the traits HW (heart weight) and LVM (left-ventricular mass) corrected for body mass of the animal to the previously described cardiac mass QTL on chromosome 3 in the BN.Lx and SHR.Ola F2 offspring. The red dashed lines indicate the limits of the refined QTL (modified from [47]).

The two congenic strains are called SHR.BN-(3L) and SHR.BN-(3S). The SHR.BN-(3L) has a 60Mb segment of BN.Crl introgressed onto SHR.Ola chromosome 3, while the SHR.BN-(3S) contains a 49Mb segment of BN.Crl. All other genomic loci are homozygous for SHR.Ola alleles.

In Figure 2.2.1 the cardiac mass locus on chromosome 3 including the association results of two additional relevant cardiac phenotype traits are shown, namely heart weight (HW) and left-ventricular mass (LVM) corrected for body weight (BW) and blood pressure. The dashed red lines indicate the boundaries of the blood-pressure independent heart weight QTL. To confirm these findings, McDermott et al. performed a comparative analysis using the two congenic strains, which narrows down the QTL to 11Mbp by exclusion and confirmed the blood pressure independent effect on the mass difference regulated by this locus.

Additionally, the authors generated a knock-out mouse that has been derived by genetic disruption of the *EndoG* gene, which leads to the absence of the EndoG protein in these mice [47]. In a comparative analysis of C57BL/6 wild-type and C57BL/6 EndoG-/- mice they showed that EndoG drives the cardiac mass differences without changing the blood pressure.

The second dataset consists of these two congenic strains and the knock-out mouse model. For the study described in this thesis, we use both models as an independent validation dataset, for a distant riboQTL, that we identify in the RI panel and that is largely overlapping the cardiac mass QTL (described in chapter 4.5.2). The two congenic strains and the knock-out mouse model allow us to fine-map this locus and help to understand whether *EndoG* is driving the translational changes we observe. RNA-Seq and Ribo-Seq was generated from five male replicates of each strain in the same way as for the RI panel. In accordance with the RI panel samples, the animals were sacrificed at an age of 6 weeks and left ventricles were collected and frozen immediately.

3 Computational approaches

3.1 General pre-processing and quality control steps for RNA-Seq and Ribo-Seq data

In the following paragraph, our pre-processing pipeline for RNA-Seq and Ribo-Seq data is described. The pre-processing starts by demultiplexing the raw sequencing reads using bcl2fastq to convert bcl files derived from the Illumina HiSeq 2000 platform into fastq files. This step is followed by removing the sequencing adapters from fastq files of the RNA-Seq and Ribo-Seq data. At the same time, we remove RNA species, which are different from mRNA and trim the RNA-Seq data to the same length as the Ribo-Seq reads. Both datasets are aligned against the reference genome using Tophat2 [48] and bam files for each sample are produced. Bam files are binary files that contain sequence alignment data. These bam files are used to quantify the data and to perform quality control and the identification of all translated genes. After pre-processing the data, we perform QTL mapping and differential expression analysis. This pipeline is summarised in Figure 3.1.1 and the rationale for the selected approach is described in the following paragraph.

Both sequencing approaches share many characteristics, such as their purpose to derive read-based quantitative measures of gene expression. Nevertheless, there are differences between the derived data. RNA-Seq reads normally span 75-100bp, while Ribo-Seq reads are limited to the footprint size of the ribosome, which are mainly 29bp. To circumvent an imbalance in shorter read length resulting in better alignment rates [49], we trim the RNA-Seq reads to 29bp allowing for a comparable alignment (as previously described in [32]). Thus, only similar challenges for both datasets in regard to repetitive sequences or overlapping genes remain. The following code combines all fastq files sequenced on different lanes and in a paired-end manner to one merged file in order to next remove the adapter using fastx_clipper and trim the reads to 29bp using fastx_trimmer.

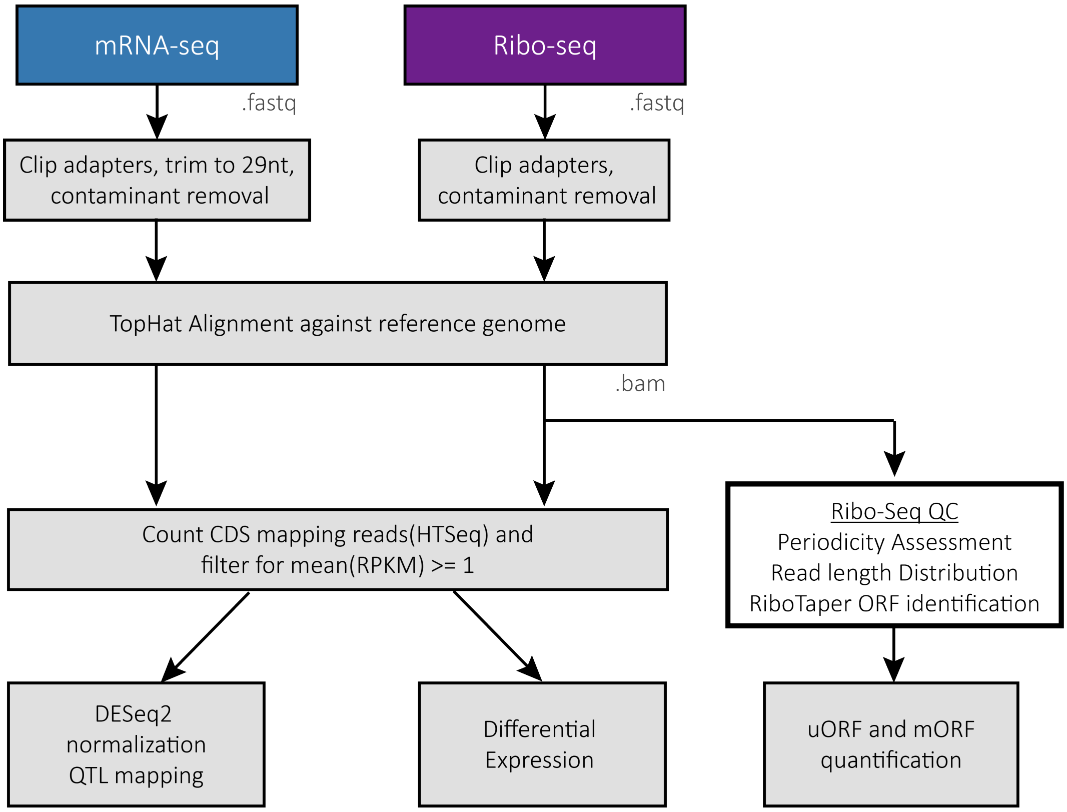

mRNA-seq
Ribo-seq
.fastq
.fastq
Clip adapters, trim to 29nt, contaminant removal
Clip adapters, contaminant removal
TopHat Alignment against reference genome
.bam
Count CDS mapping reads(HTSeq) and filter for mean(RPKM) >= 1
Ribo-Seq QC
Periodicity Assessment
Read length Distribution
RiboTaper ORF identification
DESeq2 normalization QTL mapping
Differential Expression
uORF and mORF quantification

Figure 3.1.1: Pipeline for pre-processing and quality control of the RNA-Seq and Ribo-Seq data. Ranging from clipping the adapter sequences over the TopHat Alignment to the reference genome to the identification of all translated genes and ultimately to perform QTL mapping.

Unix Code:

```
cat <sample>.fastq.gz > <sample>_merged.fastq.gz

zcat <sample>_ merged.fastq.gz | \
fastx_clipper -a AGATCGGAAGAGCACACGTCT -l 20 -c -n | \
fastx_trimmer -f 1 -z -o <sample>_merged_trimmed.fastq.gz
```

Parameters:

```
fastx_clipper:
-a - adapter sequence
-l - minimal length
-c - discard non-clipped sequences
-n - keep sequences with unknown nucleotides

fastx_trimmer:
-f - first base to keep
-z - output compressed as gzip file
-o - output file
```

Ribo-Seq data tend to include RNA species in addition to mRNA. Sequencing of ribosomal RNA (rRNA) fragments resulting from the nuclease-treatment step of Ribo-Seq reduces the sequencing space, especially in conditions where global translation levels are low. To remove reads arising from non-mRNA sources, we align the Ribo-Seq data, as well as the RNA-Seq data, to known abundant sequences from these RNA species: transfer RNA (tRNA), mitochondrial RNA (mtRNA) and ribosomal RNA (rRNA), using Bowtie2 [50]. The abundant sequences were downloaded from the following resources:

- ribosomal RNA from IGenomes (Illumina) - providing fasta sequences of all rRNAs
- mitochondrial RNA from Ensembl - providing fasta sequences of all mitochondrial DNAs [51]
- transfer RNA from the GtRNAdb from UCSC - providing whole genome scans for tRNAs [52, 53]

The RNA-Seq and the Ribo-Seq data are filtered by aligning the trimmed and cleaned fastq files against a merged file of all abundant sequences using the following script.

Unix Code:

```
cat tRNAs.fa rRNA.fa MT.fa > rn6_contaminants.fa

bowtie2-build -f rn6_contaminants.fa rn6_cont_index

bowtie2 -p 2 -x rn6_cont_index -L 20 \
              --un-gz <sample>_trimmed_clean.fastq.gz \
              -U <sample>_merged_trimmed.fastq.gz \
              -S <sample>_removed_sequences.sam
```

Parameters:

```
bowtie2:
-f - input fasta file
-p - number of threads
-x - genome of abundant RNA sequences
-L - length of seed substrings
--un_gz - output gezipped
-U - file with unpaired reads (input)
-S - aligned sequences in sam format (reads aligning against abundant
   genome)
```

For sequencing alignment, we use Tophat2 [48], which is a widely used alignment tool that allows for spliced reads, which is important to map RNA-Seq and Ribo-Seq data. In order to obtain all existing splice junctions, we first align the full-length 2x101nt RNA-seq data against the Ensembl rat reference genome (Rattus Norvegicus Ens82). In a second alignment step, the 29mer reads of both datasets are mapped using the splice junction information gathered from the alignment of the full-length RNA-Seq reads. Establishing all splice junctions allows for the accurate alignment of short reads that are split across exons. This is done by the following command.

Unix Code:

```
tophat2 --num-threads 6 --prefilter-multihits
             --read-mismatches 2 --read-edit-dist 2
             --read-realign-edit-dist 0 --tmp-dir <tmp-dir>
             -o <outdir> -G --raw-juncs <juncs_files>
             <genome annotation> <bowtie_index_genome>
             <forward_reads> <reverse_reads>
```

Parameters:

```
--num-threads - number of threads
--prefilter-multihits - reads that map to multiple positions are removed
--read-mismatches - number of allowed mismatches per read
--read-(realign)-edit-dist - maximum read (realign) edit distance
--tmp-dir - temporary directory
-o - output directory
-G - GTF-file
-raw-juncs - junction file from full-length mapping
```

The quantification of aligned reads was carried out using HTSeq [54] with default parameters. As Ribo-Seq reads only map to the coding part of the gene, we decided to only quantify reads mapping to the coding DNA sequence (CDS) of the genes to make RNA-Seq and Ribo-Seq data more comparable.

Next, the RNA-Seq and Ribo-Seq count data was normalised using the median-of-ratios normalization method in a joint manner for both datasets. This method is implemented in the function *estimateSizeFactors* in the R package DESeq2 [55]. The basic idea for this normalization procedure is to fit the RNA-Seq and Ribo-Seq data to a negative binomial (NB) distribution with dispersions α and means μ. The read counts K_{gid} for gene g in sample i and dataset d can be described as [56]:

$$K_{gid} \sim NB(\mu_{gid}, \alpha_{gid},) \tag{3.1}$$

After fitting the count data to the NB distribution, the scaling factor s_i can be determined by:

$$s_i = median \frac{K_{gid}}{(\Pi_{v=1}^{m} K_{gid})^{1/m}}, \quad \text{where m describes the number of samples} \tag{3.2}$$

This joint normalization was previously described by the developers from the XTAIL, which is a R package to perform genome-wide differential translation analysis [56]. This tool cannot be used directly for our purpose of performing riboQTL mapping, because it does not offer sample-wise normalised counts. But, we can make use of the general idea to performing joint normalization to make RNA-Seq and Ribo-Seq data more comparable. This way both count matrices follow the same distribution, which is important for evaluating gene expression differences that are present on both layers or only on one, but also to be able to calculate the translational efficiency - a measure of how efficiently an mRNA molecule is translated into a protein.

R Code:

```
Normalize_rna_ribo_together = function(mrna, rpf)
{
        pool_sizeFactor <- estimateSizeFactorsForMatrix(cbind(mrna,rpf))
        mrna_sizeFactor <- pool_sizeFactor[1:ncol(mrna)]
        rpf_sizeFactor <- pool_sizeFactor[(ncol(mrna)+1:ncol(mrna)+ncol(
            rpf))]
}
```

There are various approaches described in the literature, how to assess the translational efficiency (TE) from RNA-Seq and Ribo-Seq levels [22, 56, 57]. Ingolia et al. suggest taking the ratio of Ribo-Seq reads over RNA-Seq reads [22]:

$$\text{TE} = \frac{\text{Ribo-Seq}}{\text{RNA-Seq}}. \tag{3.3}$$

For differential expression analysis comparing two strains with multiple replicates this approach works well, but in the case of QTL mapping it leads to higher variance levels, because noise is introduced twice - on the RNA-Seq and on the Ribo-Seq level. Meanwhile, various tools have been published to derive the mean TE and to perform differential TE analysis by comparing two different conditions/strains. Two widely used approaches are RiboDiff [57] and XTAIL [56], both based on DESeq2, that are taking the variability across replicates into account to properly define the translational efficiency and perform differential testing. For QTL mapping these tools are not applicable as values for each sample are required rather than average values per group.

In this project, the translational efficiency is derived by regressing the RNA-Seq effect from the Ribo-Seq levels using a linear model. This allows to derive residuals for each sample - gene pair. This measure for translational regulation can be used for QTL mapping as a separate trait. Subsequently, TE refers to the residuals of the linear model.

R Code:

```
Translational Efficiency =
                    resid(lm(normalized_ribo_read_counts ~
                        normalized_rna_read_counts))
```

3.2 Identification of actively translated genes by inferring periodicity using RiboTaper

The scope of this thesis is to study translational regulation. Therefore, we need to define all actively translated genes. To identify all translated genes from Ribo-Seq data, we need to assess the 3-nt periodicity, which can be done by assessing the P-site tracks for each gene in all the three possible reading frames and evaluating their periodicity. There are different approaches, how to transform the quantitative P-site tracks into their periodic components. Schuster et al. described an transformation that is called a periodogram [58], which gives an estimate of the spectral density through Fourier transformation. A second approach is based on the multitaper method [59], where prior to Fourier transformation, different smoothing windows are used to derive a less noisy signal. Both approaches are based on transforming the quantitative data into their spectrum of fixed periodic components. We decided to use the tool RiboTaper [60], which it is based on the multitaper approach that gives a less noisy representation of the periodicity. For RiboTaper, F-values get extracted from the coefficient vector representing the contribution of each frequency component, that are closest to 3-nt periodicity. Next, p-values are calculated from the F-statistic using 2 and 2k-2 degrees of freedom, where k is the number of tapers used [60]. The RiboTaper algorithm is summarized in Figure 3.2.1.

We perform a two-step RiboTaper approach. First, RiboTaper is run on each sample separately using only the optimum read lengths and frames. Those parameters were determined for each sample by counting the number of reads showing a specific read length and inferring the periodicity by assessing the P-site distribution in a meta gene plot. This allows us to identify all open reading frames for each sample.

In a second step, the aligned Ribo-Seq reads are pooled for each tissue using samtools [61, 62] and RiboTaper is run on the merged dataset. Merging samples on the one hand allows to increase the power to detect translated genes, on the other hand it enables to filter the results based on shared detection of ORFs across samples.

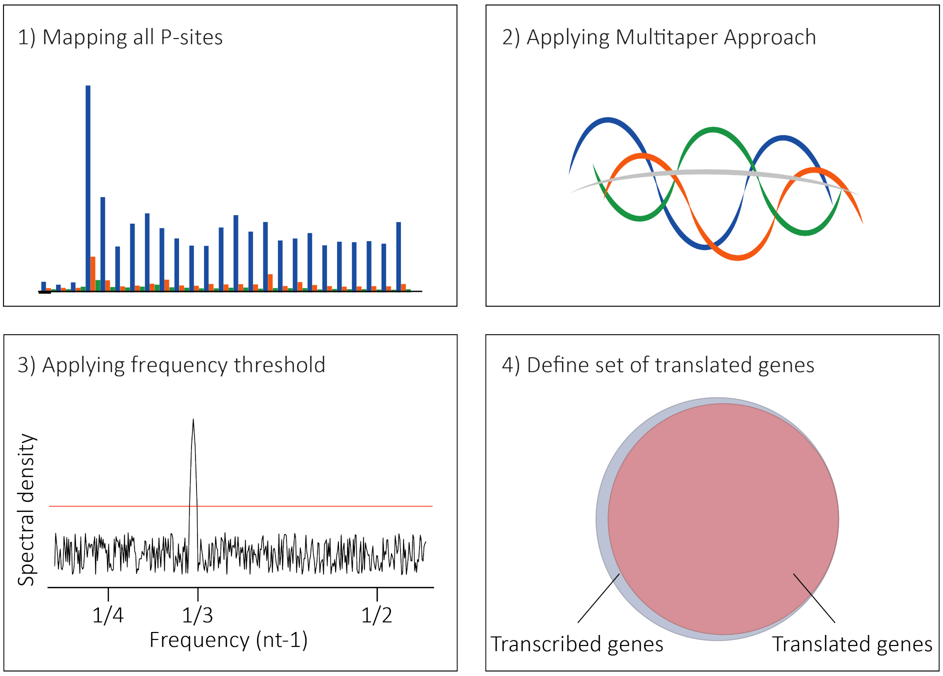


Figure 3.2.1: RiboTaper workflow: 1) The P-sites of each read are determined in the aligned data files, 2) the P-site tracks are tested for 3-nt periodicity using the multitaper method for all three possible frames. 3) via Fourier transformation the original signal is transformed into the fixed periodic component and evaluated as such, 4) after filtering and quality control steps the output from RiboTaper defines the set of translated genes that are used for further analysis.

All detected ORFs are filtered by the following criteria:

- Detected as translated in at least 10 samples
- RNA-Seq expression cutoff of mean (RPKM) ≥ 1
- Categorized as:
  - ORFs_ccds (ORFs that overlap the canonical CDS) or
  - nonccds_coding_ORFs (ORFs that match a protein-coding gene but not the canonical CDS) or
  - ncORFS (non-coding ORFs)
- For lncRNAs the ORF should not overlap an annotated CDS on the same strand

In addition to define genes as translated, RiboTaper provides information on independent and overlapping uORFs. Independent uORFs are defined in the RiboTaper output, whereas overlapping uORFs need to be derived from the output by filtering as following:

- the ORF starts in the 5'UTR and overlaps the annotated CDS of the same transcript
- the uORF must not be in the same frame as the mORF and so the resulting peptide should be non-overlapping
- mORF is transcribed with a mean RPKM ≥ 1
- ORF was detected in at least 10 samples

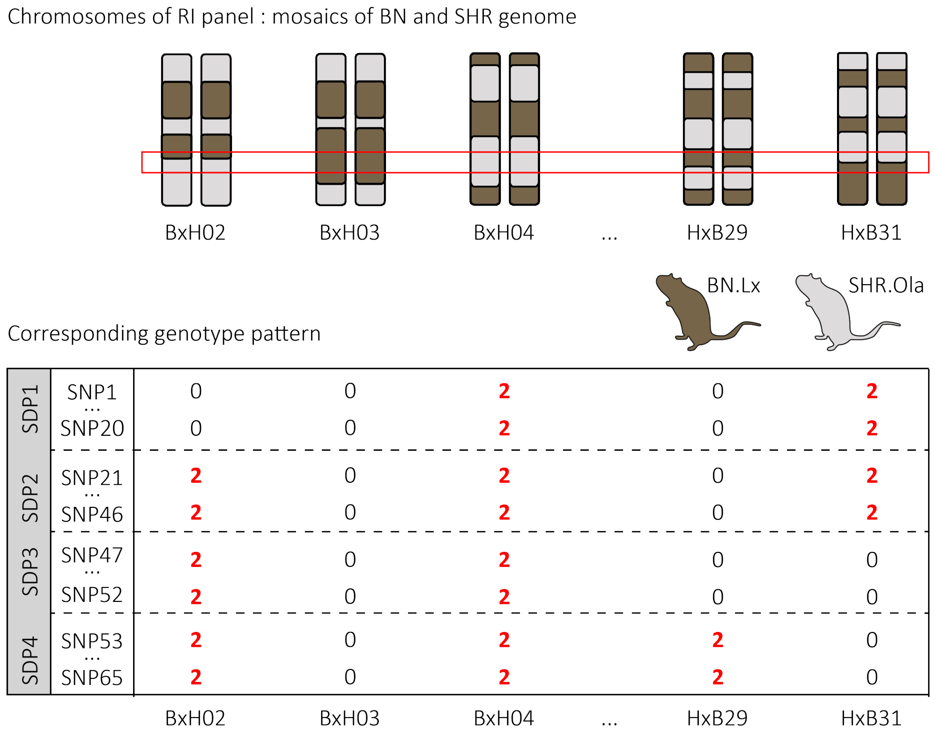
3.3 Generating a genotype map of the rat RI panel for QTL mapping

The genotypes used for this study are based on the Affymetrix RATDIV SNP Array with >500k high quality markers [46]. The original 25-mer Affymetrix probes have been remapped to the latest Ensembl rat genome (Rnor6.0) as described in [46] using BLAST [63]. The wild type or variant probe was required to map uniquely within the entire genome.

Single nucleotide polymorphisms (SNPs) in the HxB/BxH RI panel form blocks of identical segregation patterns, where each SNP within such a block has the exact same association strength with the trait. We collapse these blocks into 2,957 strain distribution patterns (SDPs) to improve the power of the QTL mapping. An SDP is defined as a genomic region, where the distribution of BN.Lx and SHR.Ola alleles is not changing. If the genotype pattern across the 30 strains changes, a new SDP is defined (Figure 3.3.1).

Some SDPs occur more than once in the genome. Those SDPs have been simplified to globally uniquely occurring SDPs (n = 1,835). We preserve the information to be able to link all local SDPs to global SDPs.

Prior, to QTL mapping, we filter for putative misannotation under the hypothesis that an annotation error occurred when a single SDP is surrounded by two SDPs showing exactly the same genotype pattern. Applying this filter leads to a total number of 1,685 SDPs used for QTL mapping.


Chromosomes of RI panel : mosaics of BN and SHR genome
BxH02
BxH03
BxH04
...
HxB29
HxB31
BN.Lx
SHR.Ola
Corresponding genotype pattern

		BxH02	BxH03	BxH04	...	HxB29	HxB31
SDP1	SNP1	0	0	2		0	2
	...						
	SNP20	0	0	2		0	2
SDP2	SNP21	2	0	2		0	2
	...						
	SNP46	2	0	2		0	2
SDP3	SNP47	2	0	2		0	0
	...						
	SNP52	2	0	2		0	0
SDP4	SNP53	2	0	2		2	0
	...						
	SNP65	2	0	2		2	0

Figure 3.3.1: Schematic explaining the definition of strain distribution patterns (SDPs) - SNPs that show an identical segregation pattern are grouped together as a SDP.

3.4 Assessing genetic effects on gene expression by QTL mapping

We perform quantitative trait loci (QTL) mapping using two different approaches to understand the impact of genetic variants on gene expression regulation. The most frequently used model to perform QTL mapping is a simple linear regression model. In the following linear model, Y correspond to the observed gene expression level and X to the genotype pattern of a particular SNP, which describes the predictor variable.

$$Y = \alpha + \beta X + \epsilon, \quad \text{where} \quad \epsilon \sim i.i.d.N(0, \sigma^2). \tag{3.4}$$

The association for each gene-SNP pair is determined assuming that the relation is linear. The genotypes of a SNP are encoded by 0 for homozygous reference,1 for heterozygous and 2 for homozygous alternative allele. In order to solve this linear regression model a number of variables need to be determined first to determine the test statistic, which could be of different nature according to the quantitative trait that is tested and ultimately the p-value is calculated. The linear model needs to be solved for each gene-SNP pair, which is computationally demanding. In the past many tools have been developed to decrease the computational burden to perform genome-wide QTL mapping.

We use two different approaches to derive QTL, which have been developed to allow for fast processing. The first approach is based on the identification of associations in a pairwise manner by testing each gene-genotype pair separately. Whereas, the second approach uses hierarchical clustering of genes to identify shared regulation of multiple genes or of different experiments, tissues or conditions.

3.4.1 Pairwise association testing using MatrixEQTL

The first QTL mapping strategy is based on the R package MatrixEQTL that was developed by Shabalin et al. in 2012 [64]. This method is a state of the art approach for performing QTL mapping on large datasets, as it offers flexible input options and fast processing.

In each MatrixEQTL run a correlation matrix of genotype (s) and gene expression patterns (g) across all samples is calculated. Shabalin et al. [64] states that each test statistics can be described by the sample correlation equation:

$$r = cor(g, s). \tag{3.5}$$

The absolute value of the sample correlation is chosen as a test statistic to determine significant gene-SNP associations. In order to simplify the calculations the genotype and gene expression variables are standardised to have zero mean and unit sum of squares, which results in

$$r_{gs} = cor(s, g) = \frac{\sum(s_i - \bar{s})(g_i - \bar{g})}{\sqrt{\sum(s_i - \bar{s})^2 \sum(g_i - \bar{g})^2}} = \sum s_i g_i = \langle s, g \rangle. \tag{3.6}$$

Here, $\langle s, g \rangle$ describes the inner product between the vectors s and g (taken from [64]). Each row in the genotype matrix S contain the measurements of each SNP across samples and each row of the gene expression matrix G contains the values for a single gene across samples. The columns, here the different samples, match in both matrices and therefore a matrix of all gene-SNP pair correlations can be determined a in one large matrix multiplication $G * S^T$. To allow for fast processing even with large genotype and/or gene expression matrices, the linear regression model is solved in blocks of 10,000 x 10,000 entries (Figure 3.4.1) [64].

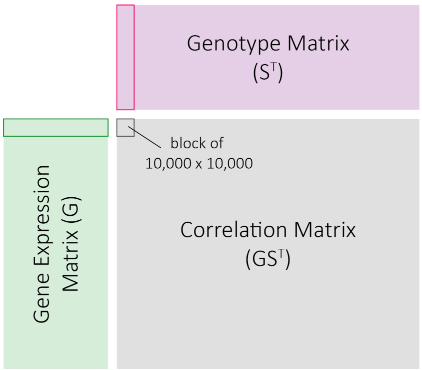


Figure 3.4.1: MatrixEQTL is based on generating a correlation matrix (G*ST) by using multiplication of the gene expression matrix G and the transposed genotype matrix S. Association tests are performed in blocks of 10,000 x 10,000. Figure adapted from Shabalin et al. [64]

MatrixEQTL offers a variety of parameters that can be used according to the aim of the study and the nature of the dataset. Though MatrixEQTL allows the use of covariates and abbreviated processing e.g. only detecting local associations or only significant associations, we did not use any covariates and performed all tests to derive p-values for each association. The grouping of non-unique SDPs, makes it necessary to infer non-significant associations of neighbouring SDPs in order to identify the real associated SDP in the genome. After the association testing, the position of those SDPs that occur several times in the genome can be assigned to determine whether a local or distant effect was observed. Our hypothesis is that real associations are visible in neighbouring SDPs to a lower extent, as the changes of the genotype pattern from one SDP to the next are normally rather small. Therefore, we examine all associations of the surrounding SDPs to identify the origin of each QTL.
After assigning all SDPs to their genomic position, we group local and distant effects separately and perform a Benjamini Hochberg correction [65] on each group of associations. This enables us to use different FDR cutoffs for local and distant QTLs based on the p-value distribution within each group. In this study, we require distant QTL to appear on a different chromosome than the affected gene, because some genotype blocks are very large, so that it is difficult to define a general distance threshold between a gene and an SDPs in order to identify an association as distant.
We perform permutation testing to determine a significance threshold for each dataset by deriving the distribution of test statistics under the null hypothesis that there is no

association. We therefore randomly shuffle the samples in the gene expression matrix and perform 10,000 runs of MatrixEQTL on the original genotype matrix. To speed up the permutation testing, we remove genes from further permutation runs, if they show 15 times more extreme associations than the most significant association in the original MatrixEQTL run. A significant local association was defined as having an empirical p-value ≤ 0.0015 (less or 15 times more extreme p-values in 10,000 permutations) and FDR ≤ 0.1, and for distant associations the empirical p-value ≤ 0.0015 and FDR ≤ 0.2. We use a less stringent FDR cutoff for distant associations, because the effect sizes of distant associations are smaller.

3.4.2 Hierarchical association testing using MT-Hess

The second approach, that was used to perform QTL mapping in this thesis is called MT-Hess [66] and detects associations simultaneously in multiple datasets, such as different tissues, cell-types or even different traits, e.g. RNA-Seq and Ribo-Seq.
The basic idea is that gene-SNP associations are derived for multiple conditions at the same time by modelling a linear regression for each gene expression measurement that is linked by a hierarchical model [66]:

$$G_k = A_k - SB_k, \quad N(I_n, E_k) \tag{3.7}$$

where, G denotes a matrix containing all observed gene expression levels for gene k and the matrix S describes the genotype matrix. B_k is a regression coefficient matrix, where each entry (b_{kjl}) relates a genotype of a SNP j, the expression level of a gene k in condition l. The intercept A_k and the covariance matrix E_k for between-condition are specific to each gene k [66]. The Hess pipeline can also be applied to a single condition or tissue, with the aim to increase the power for detecting distant gene cluster, e.g. shared regulation of one genetic variant on the gene expression of multiple genes. The basis of the modelling approach is a Bayesian variable selection method that acts on three levels at the same time. In the first step, the gene expression matrices of each trait are transformed into sparse regression matrices on all genotype patterns. Therefore, an additional binary matrix is introduced that summarises for all conditions, whether an association between gene expression and genotype pattern was observed. For a given gene k, we denote by y_k the row binary vector for the regression coefficient matrix B and the genotype matrix S. Conditionally, on the binary matrix the sparse regression for all genes is given by:

$$G_k = A_k - S_{yk}B_{yk}, \quad N(I_n, E_k) \tag{3.8}$$

In a next step, a multi-variate regression is performed jointly on all sparse regression matrices for each gene to detect shared regulation across all traits. This provides increased power for the detection of SDPs that regulate multiple traits of a gene. In the last step, a multi-variate regression analysis is performed across all genes to identify similarly regulated genes that share prior parameters. The multi-variate regression analysis across all genes, enables the detection of gene cluster regulated by the same genetic locus (Figure 3.4.2).

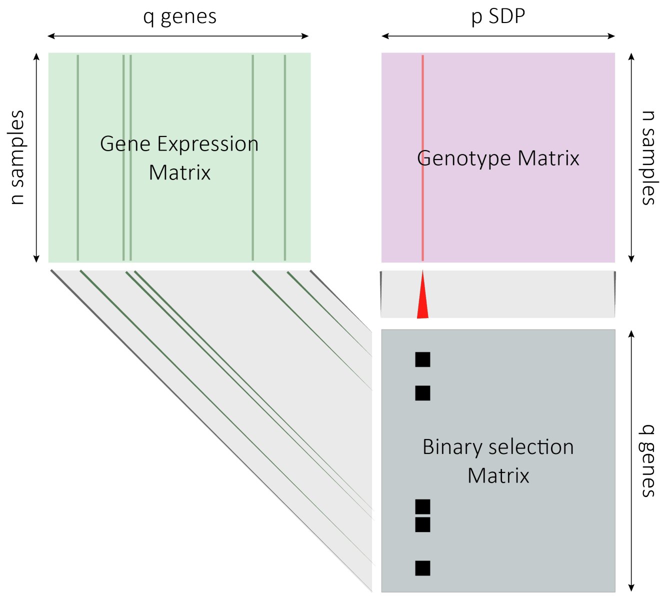
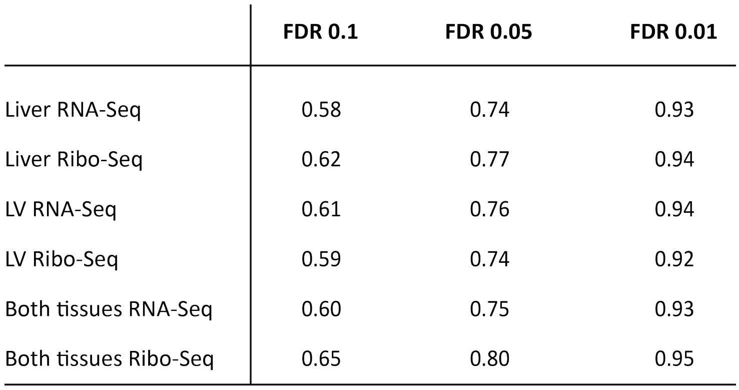


Figure 3.4.2: Illustration of the Bayesian variable selection of the Hess approach based on a multi-variate regression based on sparse regression matrices

For MT-Hess all three steps are performed and for the Hess approach performed on one condition only step 1 and 3 are performed. We use this approach first to identify shared regulation between left-ventricular and liver tissue, and second to validate and enhance the number of distantly regulated genes that have been identified by the global QTL mapping using MatrixEQTL.

The output from both Hess approaches are marginal posterior probability of inclusion (MPPI) matrices, which give an evidence measure for the association of a genotype and a trait. To account for multiple testing, we calculated MPPI cutoffs that correspond to the FDR cutoffs 1%, 5% and 10% (Table 3.1). The selected algorithm to calculate the FDR cutoffs is commonly used to control for hierarchical errors in QTL studies [66].



	FDR 0.1	FDR 0.05	FDR 0.01
Liver RNA-Seq	0.58	0.74	0.93
Liver Ribo-Seq	0.62	0.77	0.94
LV RNA-Seq	0.61	0.76	0.94
LV Ribo-Seq	0.59	0.74	0.92
Both tissues RNA-Seq	0.60	0.75	0.93
Both tissues Ribo-Seq	0.65	0.80	0.95

Table 3.1: MPPI cutoffs for the different datasets and three different FDR cutoffs 1%, 5% and 10%.

3.5 Performing RBP motif enrichment in translationally regulated genes

In order to assess the regulatory potential of RNA binding proteins (RBP) we predict putative targets among regulated genes using RBP binding motifs. We therefore, downloaded all known rat RBP binding motif matrices from the CISBP-RNA database [67], an online library of RBPs and their motifs. The motif matrix $M = M_{w,\alpha}$ is a position frequency matrix (PFM) and records how frequently a nucleotide α has been observed at position w within the alignment.

In total, the database contains 189 distinct RBP motifs for the rat. We evaluate the potential binding affinity for each RBP by using the Transcription Factor Affinity Prediction (TRAP) algorithm that was published by Manke et al. in 2008 [68, 69]. This allows us to test for enrichment of all provided binding motifs in a set of sequences.

Manke et al [68] describes basic idea for determining the affinity of binding motif as such: $M = M_{w,\alpha}$ is a Wx4 motif matrix of a known binding site, where $w = (1...W)$ are the positions in the motif and α describes the four possible nucleotides A, C, G, T. The entries in the PFM describe the nucleotide counts, which can be used to match each motif against any other DNA sequence of width W. The position-specific contributions to the mismatch energy can be described as:

$$\varepsilon_w = \frac{1}{\lambda} log \left(\frac{M_{w,max}}{M_{w,\alpha}} \right). \tag{3.9}$$

The different affinities of a given binding motif are ranked for each sequence region and tested for an enrichment in a set of sequences.

We adapt this approach in order to predict binding affinities for RBP motifs instead of transcription factor motifs in all translated genes. First, we run TRAP on the DNA sequences including all isoforms of every translated gene in random order to define a background set. Next, we generate our search datasets based on the QTL mapping results, ordered by the most significant association per gene. We test the genes showing a translational regulation for an enrichment of RBP binding sites. For distant associations, we additionally test whether the detected RBP is located in the distant locus of association.

4 Results

4.1 Assessing the quality of gene expression sequencing data

RNA-Seq as well as Ribo-Seq data is very dependent on the quality of the RNA, the sequencing library preparation and sequencing depth. Therefore, the quality of sequencing data can vary across datasets and this can impact the results of all downstream analyses.

In order to avoid differences in datasets due to varying data quality, we establish a quality control pipeline (Chapter 3.1), that contains 7 analysis steps ranging from assessing the read length distribution and periodicity of each Ribo-Seq sample, to evaluating the complexity of each sequenced library and similarity across samples. These different steps are described in more detail in the following paragraphs.

4.1.1 Determining Ribo-Seq specific parameters and removing non-mRNA sequences

As ribosome footprints span roughly 28-29nt, the expected read length after removal of sequencing adaptors peaks at this length. We assess the read length distribution of each sample to estimate the quality of the Ribo-Seq experiments. Figure 4.1.1 depicts the pooled read length distribution of all 30 animals of the rat RI panel and the majority of reads match the expected read length.

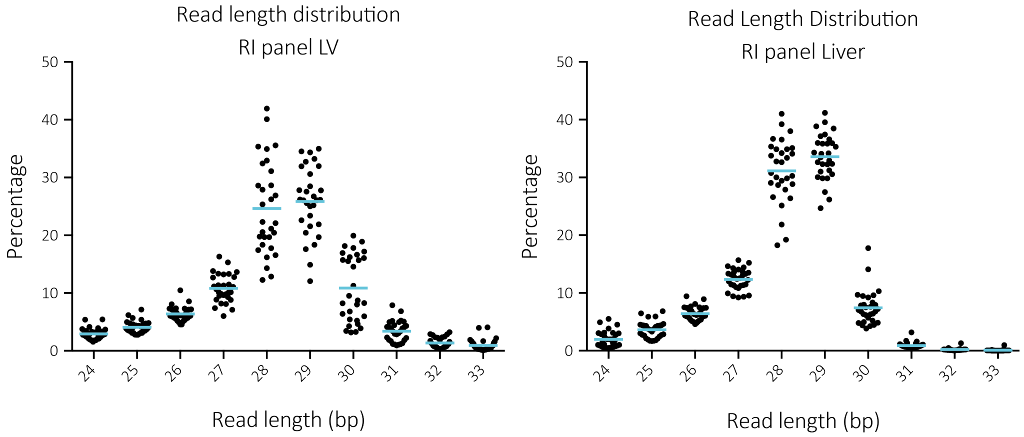


Figure 4.1.1: Read length distribution of the Ribo-Seq data of 30 rat RI samples in left ventricle and liver tissue. X-Axis shows the read length in base pairs (bp), the y-axis shows the percentage of reads that have a specific length.

During library preparation of both sequencing protocols, a cleaning step that depletes rRNA is performed, but this step is not removing all non-mRNA sequences. So, both datasets are aligned against the abundant sequences to remove and quantify tRNA, rRNA and mtRNA (as described in chapter 3.1). This allows us to only map and quantify reads mapping to mRNA. The average percentages of rRNA, tRNA and mtRNA in each dataset are summarised in Table 4.1 and the individual percentages per sample are depicted in Figure 4.1.2.

	LV RNA-Seq	Liver RNA-Seq	LV Ribo-Seq	Liver Ribo-Seq
rRNA	1.1%	0.6%	23.7%	36.0%
tRNA	0.03%	0.3%	6.1%	0.5%
mtRNA	21.4%	11.1%	19.9%	1.2%

Table 4.1: Average percentages of rRNA, tRNA and mtRNA in RNA-Seq and Ribo-Seq data of both tissues

In left ventricle higher mtRNA levels are observed in both RNA-Seq and Ribo-Seq data compared to liver. This was previously described by Mercer et al [70], who show that mitochondrial transcript abundance is increased in tissues with high-energy demands, such as heart and muscle. Furthermore, we observe high rRNA levels in the liver Ribo-Seq data, as previously described by [32]. This reflects the actively translating state of hepatocytes, which are metabolically very active and

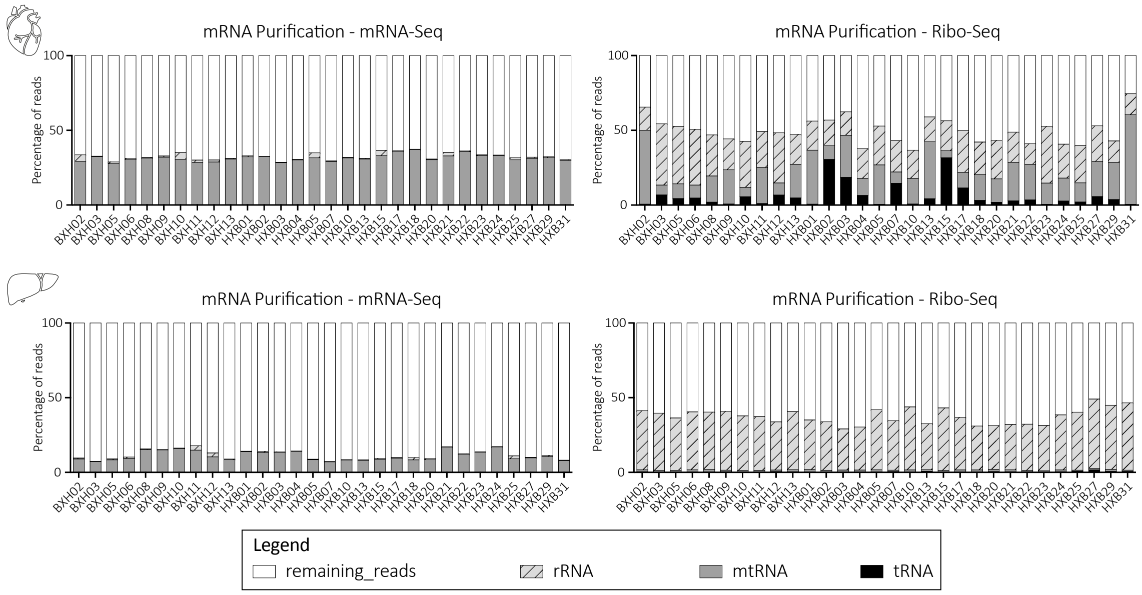


Figure 4.1.2: Percentage of reads mapping to specific RNA species in left-ventricular tissue (top) and liver (bottom). The two left panels show the distribution of reads in polyA RNA-Seq data and the right panels show this accordingly in the Ribo-Seq data.

produce many proteins. In contrast, the left ventricular ribosomes are less actively translating and as a result make less protein compared to liver.

4.1.2 Determination of 3-nt periodicity

One important quality measure of Ribo-Seq data is the number of reads that map to the primary open reading frame. We determine the P-site of every read, which corresponds to the 12th nucleotide, and map this position to the first 100nt of each gene. The P-sites map to the first nucleotide of a codon triplet. This strategy allows to observe the periodic nature of the ribosomes. The count data is not filtered for periodic reads, but the percentage of periodic reads serves a quality measure for each sample. For the identification of actively translated genes using RiboTaper, we require 70% of the reads of a specific length to be in-frame, resulting in the inclusion of the two most prominent read lengths (28nt and 29nt). Figure 4.1.3 shows a meta periodicity plot of all RI samples per tissue.

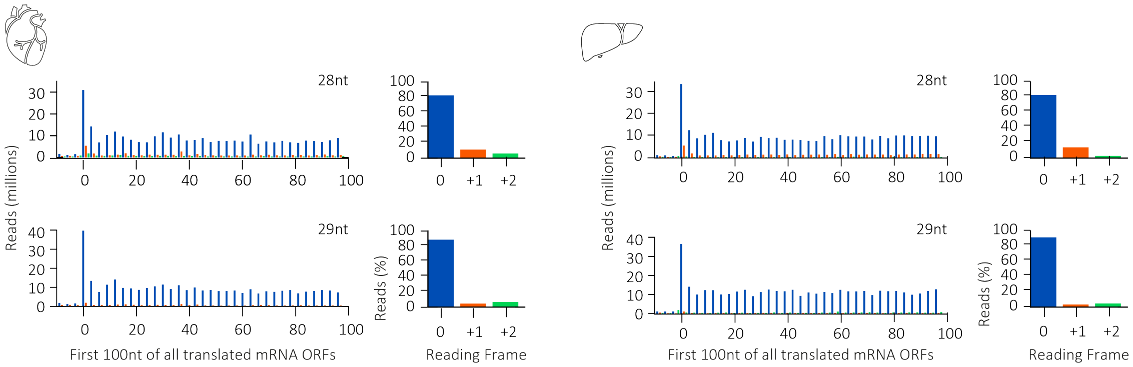


Figure 4.1.3: Ribo-Seq periodicity of left ventricle and liver samples matching the read length 28nt and 29nt. On the x-axis the first 100nt of all translated mRNAs are shown and on the y-axis the corresponding number of reads in millions. The bar plot on the right of each periodicity plot illustrates the percentage of reads falling into the 1st, 2nd or 3rd frame (0, +1, +2, respectively). The high quality of the data can be observed by the large percentage of reads that map to the primary open reading frame (frame "0").

4.1.3 Alignment of sequencing libraries

Selecting the correct reference genome for the sequence alignment is a crucial step for all following analyses. As each rat of the RI Panel is a mosaic of the two founder strains: BN.Lx/Cub and SHR.Ola, the alignment rates might differ due to genetic variability. The genome of BN.Lx is very similar to the BN reference genome, with only about 102,000 SNPs [27]. On the other hand, the SHR differs at 3.6 million SNPs from the reference genome [29]. The imbalance in genetic similarity towards the reference genome might lead to mapping biases towards the BN.Lx samples.

To address this, we aligned RNA-Seq and Ribo-Seq data of the SHR parental animals 1) to the reference genome Rattus Norvegicus 6.0 (Rnor6.0) from Ensembl version 82 and 2) to a SNP-infused genome, where we used the reference genome as a basis and inserted all known strain-specific SNPs.

We performed a differential expression analysis using DESeq2, in order to assess the impact of genetic variants on the alignment. In left ventricle, we identified 68 differential genes in the RNA-Seq and 62 in the Ribo-Seq data. Similarly, we found 48 genes on the RNA-Seq and 37 on the Ribo-Seq level in liver using an FDR $\leq$ 0.05. Applying a fold change cutoff of log2(±1.5), leads to the loss of all differentially expressed genes in both tissues and datasets. As the number of differential genes is very low, we decided to align the RI panel samples to the reference genome. We allow two mismatches per 29mer read in order to account for SHR SNPs.

As described in the introduction, Ribo-Seq reads only map to the coding region of the gene whereas mRNA-Seq reads map to the entire mRNA transcript, thus including the 5’ and 3’ untranslated regions. To be able to assess the extent of translational regulation, we only quantify those reads that map to the annotated coding DNA sequence (CDS) of the gene. If there is no CDS annotated, we count all reads that map to the exons of each gene and merge these counts into a matrix.
In Figure 4.1.4 the number of sequenced reads is shown for each RI animal in the left ventricle (upper panel) and liver (lower panel).

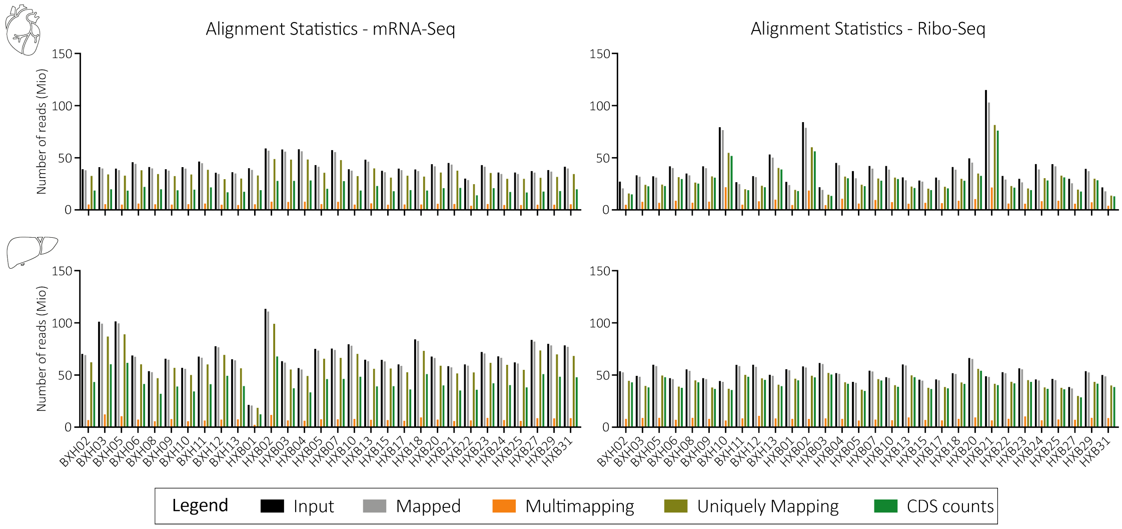


Figure 4.1.4: Alignment statistics of left ventricle and liver RNA-Seq and Ribo-Seq data. The upper panels show the left ventricle statistics; left of RNA-Seq and right Ribo-Seq. The lower panels show these numbers for liver accordingly. Black bars represent the number of input reads, grey all mapped reads, which are broken down into uniquely mapping and multi-mapping reads in brown and orange. The green bars show the CDS counts that enter the analysis.

Despite the total number of sequenced reads varying between the two sequencing approaches, we get a similar number of uniquely aligned counts mapping to the coding DNA sequence. On average, we observe in RNA-Seq LV 20.3 Mio. reads and Ribo-Seq LV 27.8 Mio reads, RNA-Seq Liver 42.7 Mio. reads and Ribo-Seq Liver 41.5 Mio. reads. The library specific differences in depth can have biological or technical reasons, but we account for them by applying a joint normalization strategy that is described in chapter 3.1.
In the liver dataset, HxB01 occurs as an outlier in terms of library depth in the RNA-Seq data. Experimental parameters such as the RIN value, which describes

the RNA integrity number of sample HxB01 was comparable to other samples with higher coverage (RIN: 8.9) and therefore the sample is considered for the analysis and further investigated at later quality control steps.

4.1.4 Assessing the library complexity

The library complexity of RNA-Seq as well as Ribo-Seq data provides another measure of quality. Low-complexity libraries contain a large number of sequenced reads that correspond to the same gene. Experimentally, low complexity of sequencing libraries is caused by over-amplification during polymerase chain reaction (PCR). Increasing the sequencing depth would mainly gain redundant information. In order to evaluate the complexity, a library complexity prediction using PRESEQ [71] is performed to all 4 datasets. PRESEQ defines complexity as the expected number of distinct molecules that can be observed in each set of sequenced reads. In Figure 4.1.5, the library complexity of all four datasets is illustrated. In general, the RNA-Seq and Ribo-Seq library complexity plots look slightly different, which can be explained by different sequencing depth of RNA-Seq compared to Ribo-Seq resulting in higher numbers of uniquely mapping reads. The red line indicates the average depth at which the data was sequenced and it confirms that none of the samples were sequenced to saturation. Mapping biases caused by variable read length can be excluded, as the RNA-Seq data was trimmed to 29mer reads and aligned similarly. All four datasets are of high quality with a 71-87% ratio of uniquely mapping reads to the total number of sequenced reads.

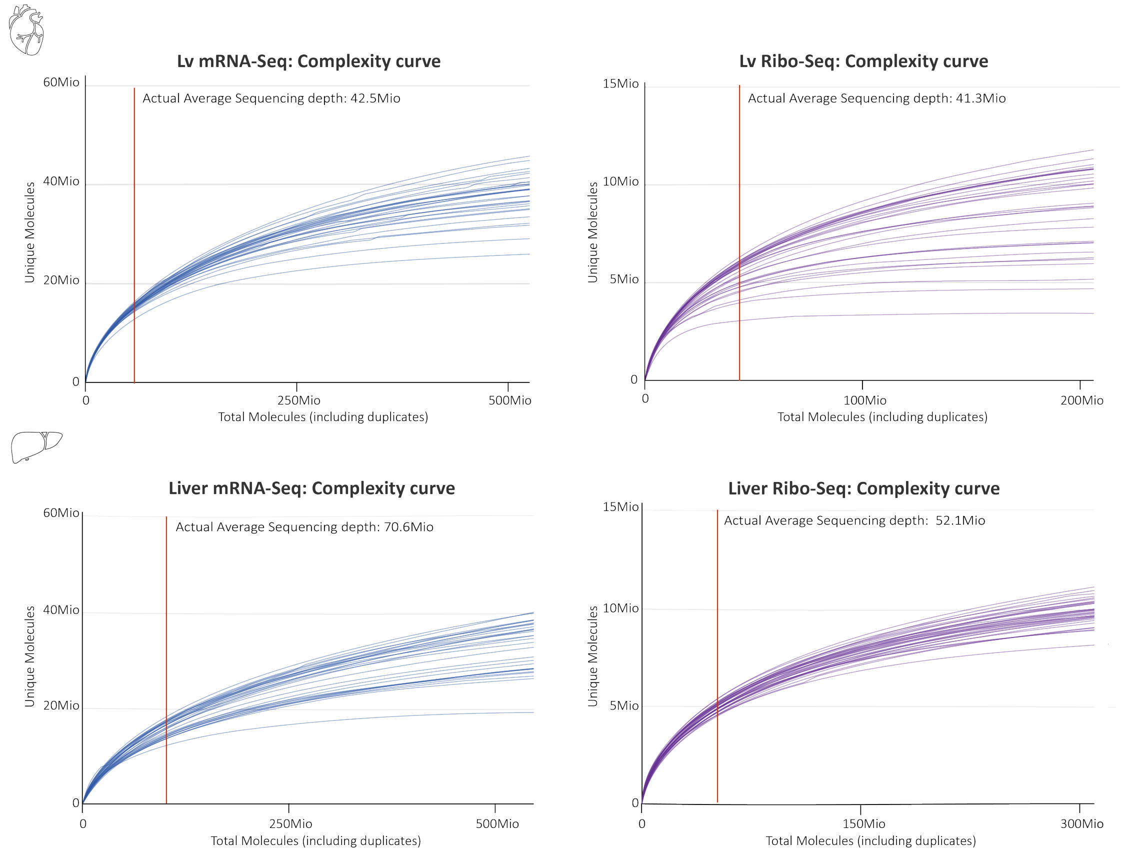


Figure 4.1.5: Library complexity results of all 30 libraries in each experiment based on extrapolation of the sequencing results. left ventricle datasets in the upper panel; left RNA-Seq and right Ribo-Seq. Liver datasets in the lower panel accordingly. The red vertical line indicates the average sequencing depth of our samples.

4.1.5 Sample-to-sample correlation

In order to test the datasets for outliers, a pairwise Pearson correlation was calculated. The mean sample correlation is required to show a Pearson's $r^2 \geq 0.85$. All samples meet the correlation requirements. Figure 4.1.6 summarises the cross sample Pearson correlation of all datasets. In general, we observe a slightly better correlation across the RNA-Seq samples compared to the Ribo-Seq data.

Interestingly, the RNA-Seq liver sample HxB01, which was observed as a putative outlier in the alignment and count statistics, correlates very well with all other samples of the same experiment. This suggests that this sample is of high and comparable quality to the other samples despite that sample HxB01 has a lower sequencing depth. The RI panel on average consists of 15 animals sharing the same genetic background at every genomic locus, resulting in 15 replicates per SDP. The generation of RNA-

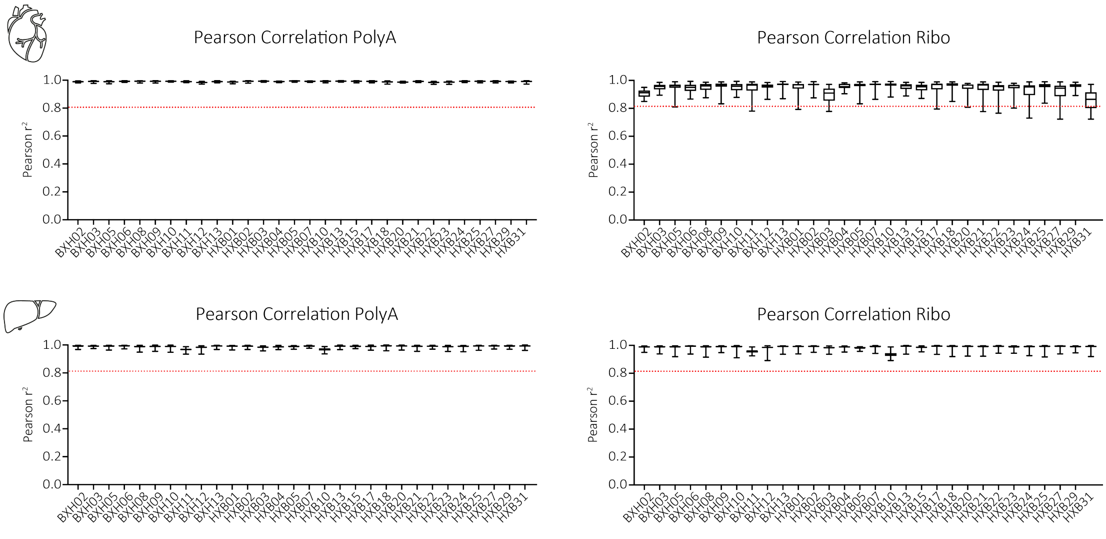


Figure 4.1.6: Sample correlation using Pearson r^2 of mRNA-Seq data. Red line indicates the threshold $r^2 \geq 0.85$ to define outliers.

Seq and Ribo-Seq data of true biological replicates would easily upscale the setup to hundreds of experiments. In order to estimate the technical variability across the 30 RI strains in a controlled setting, Ribo-Seq data of 3 additional biological replicates of two RI strains was generated and compared. The Pearson correlation showed an $r^2 > 0.98$. This data of the biological replicates indicate that the data generation is reproducible and we only expect little variability across samples due to technical biases (Figure 4.1.7).

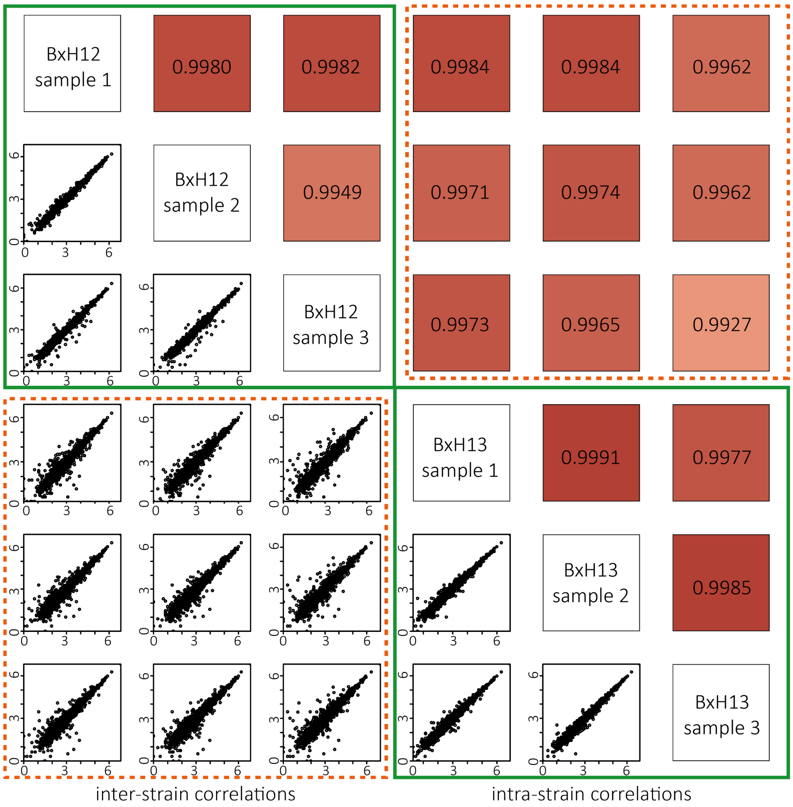


Figure 4.1.7: Pearson Correlation analysis of 3 biological replicates in two RI strains using Ribo-Seq of liver tissue. The numbers indicate the Pearson r^2 and the scatterplots show the comparison of corresponding genes in two different samples.

4.1.6 Determination of putative confounding factors

In comparative analyses, such as a differential expression analysis and QTL mapping, accounting for putative confounding factors, such as date of library preparation, sequencing date, RIN score (RNA integrity number) and library concentration, but also different laboratory technicians, can have large impact on the results. All datasets have been tested for confounding factors as described in chapter 3.1. Besides the small effect of the sequencing date in the liver RNA-Seq data, all tested technical covariates revealed no significant impact on the data (Table 4.2).
The animals, that are described in this thesis, were treated and selected under very controlled settings, including sex and age matching, identical housing and nutrition. This taken together with the three biological replicates showing little technical variations across experiments, suggest that there are no major confounding factors that need to be considered for the QTL mapping.

	LV RNA-Seq	Liver RNA-Seq	LV Ribo-Seq	Liver Ribo-Seq
Date of library preparation	0.78	0.64	0.35	0.50
Date of Sequencing	0.61	0.01	NA	NA
RIN score	0.72	0.85	NA	NA
Library concentration	NA	NA	0.28	0.97
Date of PCR	NA	NA	0.11	0.30

Table 4.2: Resulting ANOVA p-values testing whether a technical factor has impact on the data. We used the first principle component (PC1) that describes 50-80% of the variance in the data.

4.2 Characterising and comparing translated genes in rat left ventricle and liver

Quantification tools such as HTSeq [54] provide read counts for every gene listed in a provided annotation file. Further filtering of the gene set by a certain expression level is important for the following analyses. An expression cutoff for RNA-Seq data is applied to define the set of genes that is expressed in a specific sample to avoid dealing with lowly expressed genes. For this dataset, we first apply an expression threshold of mean(RPKM) ≥ 1 for every gene. RPKM is the abbreviation for reads per kilobase of transcript per million mapped reads and this is a standard way to normalise RNA-Seq data taking the library-size and gene-length into account, in order to make values comparable across genes and samples.
This filtering step provides us with a set of 11,858 transcribed left ventricle genes and 10,362 genes in liver. Figure 4.2.1 shows the distribution of all RPKM values across the 30 animals in left ventricle and liver, respectively.

The light grey bars show all measured genes and the overlapping dark grey bars show the RPKM distribution of the genes that passed the filtering criterion. Some individual values are below 1, as the mean RPKM was used.
The aim of this thesis is to study genes that are translated in the two tissues of interest. To define all actively translated genes, the 3nt periodicity in our data inferred by the tool RiboTaper to define actively translated genes. Additionally, RiboTaper enables the identification of novel translation events, such as genes that have been previously defined as non-coding (lincRNAs, antisense RNAs and processed transcripts)

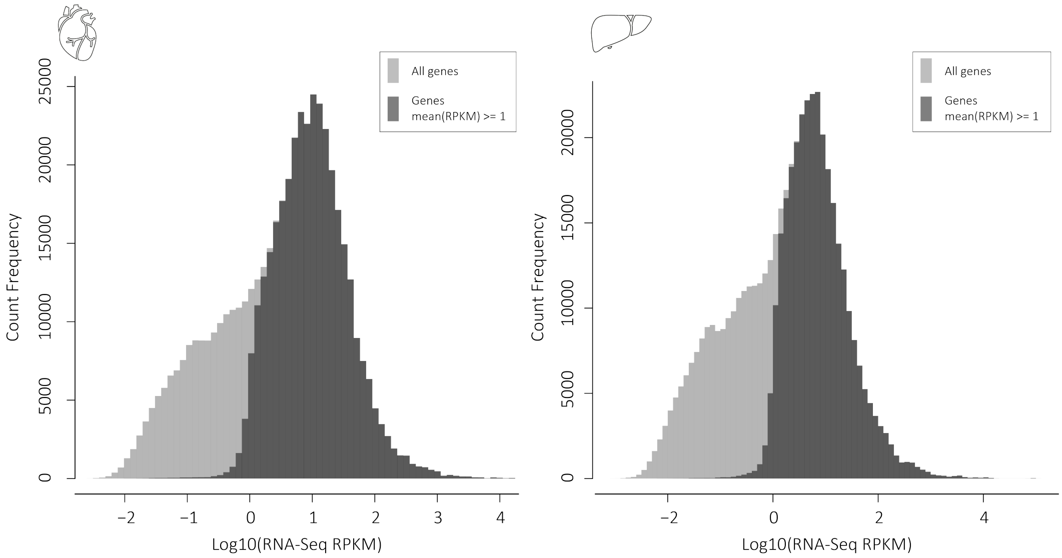


Figure 4.2.1: The histogram illustrates the count distribution of all measured genes in light grey and the overlaying dark grey bars highlight the count distribution of the expressed genes that have mean(RPKM) ≥ 1.

or small open reading frames in untranslated regions (uORFs and dORFs).

The combination of filtering for transcribed and translated genes defines the subsequently used gene set for left ventricle and liver, resulting in 9,336 liver genes and 10,531 genes in the left ventricle. The vast majority (8,622 genes) are found to be expressed in both datasets (Figure 4.2.2 a). In total, we identify 98 translated lncRNAs in the rat liver and left ventricle, 42 of them are left ventricle-specific and 24 liver-specific (Figure 4.2.2 b).

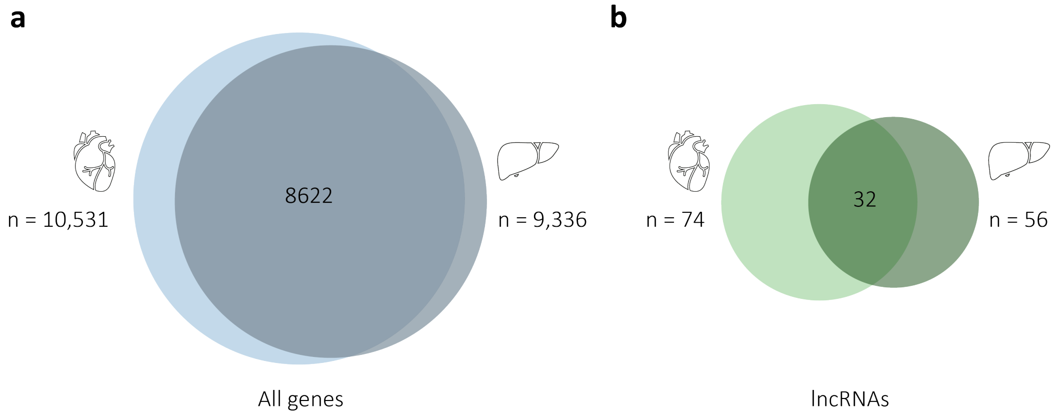


Figure 4.2.2: Number of shared genes in left ventricle and liver. a) shows all genes and b) zooms into the potentially translated lncRNAs.

The lncRNAs include 46 lincRNAs, 1 antisense gene and 27 processed transcripts in the left ventricle and 38 lincRNAs, no antisense RNAs and 19 processed transcripts in liver.

Among the 10,531 genes expressed in the left ventricle, we observe 10,457 protein-coding, 74 lncRNAs, 24 pseudogenes and one snoRNA to be translated in our data. Of the 9,336 genes expressed in the liver, 9,279 are protein-coding, 57 are translated lncRNA and 16 are categorised as pseudogenes (Figure 4.2.3).

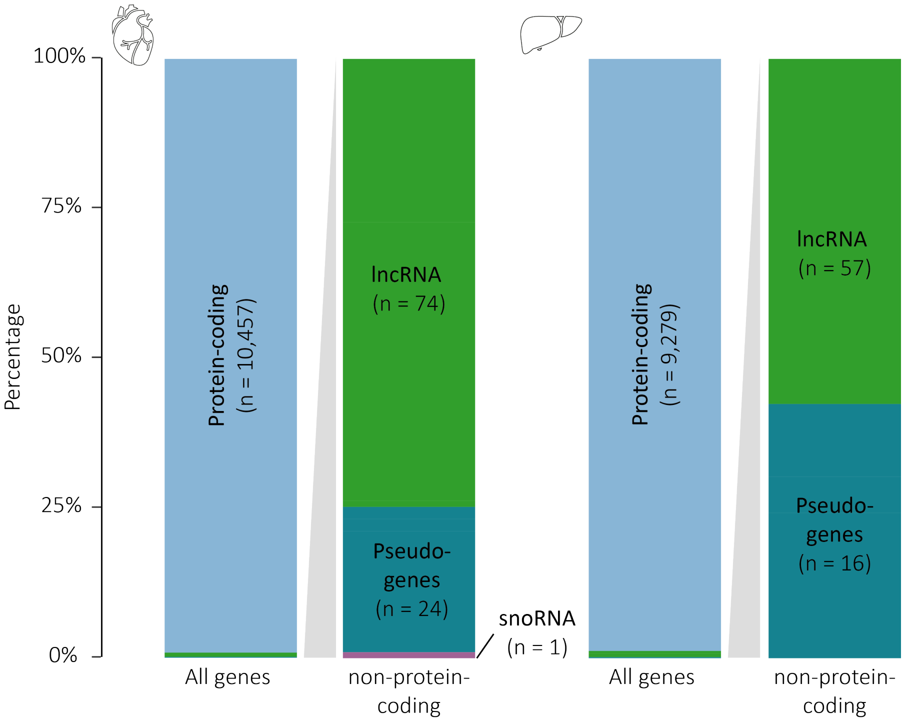


Figure 4.2.3: Genes with RiboTaper evidence are categorised by gene biotype, which are used for further analyses.

For further analysis, we used these two sets of genes that are found to be translated. Furthermore, the similarity of the jointly normalised RNA-Seq and Ribo-Seq data was evaluated by Pearson correlation as described in chapter 3.1. In both tissues, we observe Pearson r of nearly 0.9, which indicates that both datasets are well normalised and can be treated as comparable for the following analyses (Figure 4.2.4).

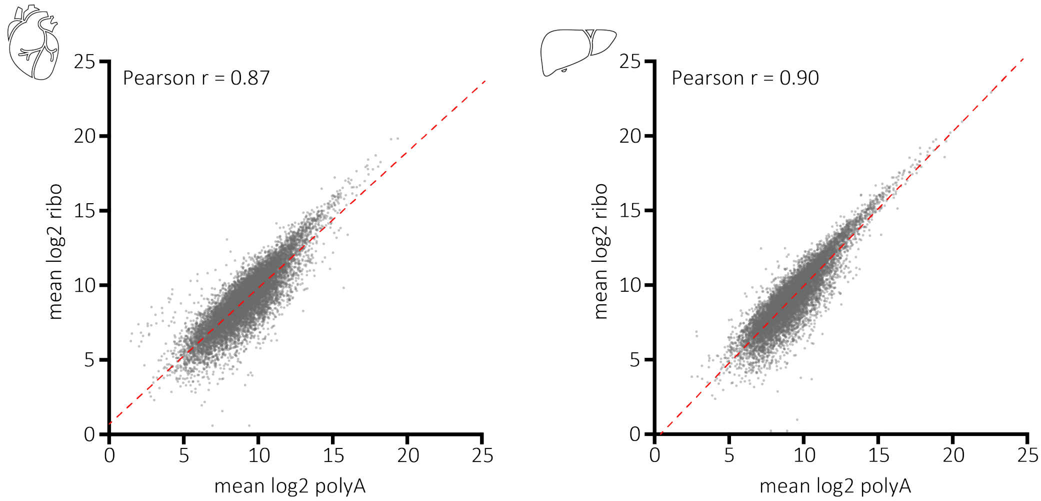


Figure 4.2.4: Pearson correlation of polyA RNA-Seq vs. Ribo-Seq on filtered and jointly normalised and log transformed counts as described in 3.1.

4.3 Identification of gene expression quantitative trait loci (QTL)

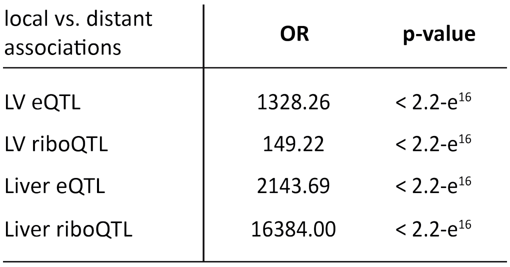
In order to identify genes that are regulated on the transcriptional and translational level with a genetic basis, quantitative trait loci (QTL) mapping is performed. In this thesis two QTL mapping approaches have been used: a) MatrixEQTL to perform pairwise QTL mapping of gene-SDP pairs and b) MT-Hess to identify common regulation of different datasets of a group of genes that goes beyond pairwise comparisons.

The resulting number of associated genes derived by pairwise comparisons is summarised in table 4.3. In left-ventricular tissue, we identify 2,385 genes showing a significant association on the transcriptional level (eQTL) and 1,146 genes on the translational level (riboQTL), in liver 1,278 genes show an eQTL and 1,270 a riboQTL, using an FDR of 0.05 for local and 0.1 for distant associations.

	local	distant		local	distant
eQTL	2355	30	eQTL	1268	10
riboQTL	1029	117	riboQTL	1269	1

Table 4.3: Number of genes that have been detected to be significantly associated to at least one SDP in the individual MatrixEQTL runs per tissue. The groups are not exclusive to each other.

Local associations are strongly enriched (Table 4.4) compared to distant associations, which can be explained by lower power to detect distant associations, but also by the stringent criterion to define associations as being distant.



local vs. distant associations	**OR**	**p-value**
LV eQTL	1328.26	$< 2.2\text{-e}^{16}$
LV riboQTL	149.22	$< 2.2\text{-e}^{16}$
Liver eQTL	2143.69	$< 2.2\text{-e}^{16}$
Liver riboQTL	16384.00	$< 2.2\text{-e}^{16}$

Table 4.4: Results of Fisher's Exact test evaluating the enrichment of significant local associations compared to distant associations.

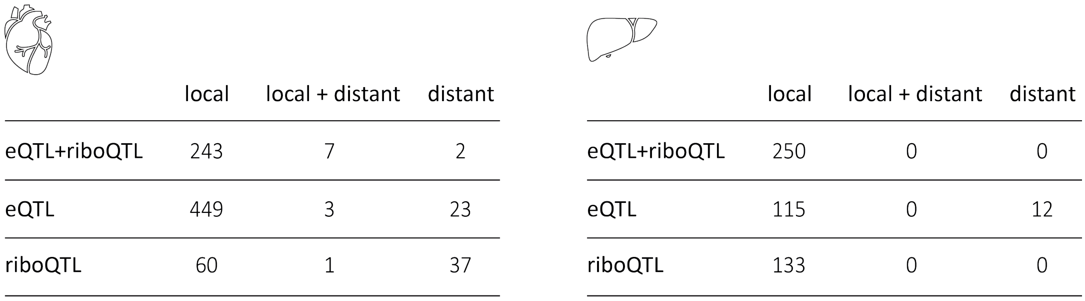
Permutation testing on each dataset for local and distant associations separately is performed in order to identify true positive hits and to filter for false positives. QTL mapping on 10,000 permutations for all traits vs. all genotype patterns is carried out by shuffling the labels of the samples in the genotype matrix and keeping the order in the gene expression matrix. This destroys the association of gene expression and genotype pattern. Based on the number of permutation tests that are performed and the observation of more extreme p-values for a certain gene, we can calculate an empirical p-value for each gene-SDP pair.

$$\text{Empirical p-value} = \frac{\text{number of more extreme p-values}}{\text{number of permutations}} \tag{4.1}$$

The permutation testing allows us to generate a reliable set of genes that are regulated by genetic variants. These genes can be grouped by the location of the gene in relation to the genetic variant. The significance threshold used for defining regulated genes are $\text{FDR}_{local} \leq 0.1$, $\text{FDR}_{distant} \leq 0.2$ and an empirical p-value ≤ 0.0015, which results from less than 15 more extreme associations per gene in 10,000 permutations. Furthermore, the setup of the experiment allows the classification of regulated genes by their level of gene expression that is affected by the genetic variant. By overlapping the different groups of eQTL and riboQTL, genes which show a change on the transcriptional and translational level can be identified, subsequently defined as eQTL+riboQTL. Genes only showing a significant change on the transcriptional level that is absent on the translational level are called eQTL, and genes where the gene expression change only occurs on the translational level are called riboQTL (Table

4.5).

This categorisation results in hundreds of local eQTL and riboQTL in both tissues. In left ventricle and liver we observe that genes, that show both a transcriptionally and translationally changed gene expression pattern, overlap to a large extent. This suggests that gene expression regulation which initiates on the transcriptional level is often forwarded to the translational level. In the left ventricle, multiple distant associations especially enriched on the translational level are found, whereas in liver no distant riboQTL are observed.



	local	local + distant	distant
eQTL+riboQTL	243	7	2
eQTL	449	3	23
riboQTL	60	1	37

	local	local + distant	distant
eQTL+riboQTL	250	0	0
eQTL	115	0	12
riboQTL	133	0	0

Table 4.5: Grouped QTL results after permutations

In conclusion, the study design allows to classify the associated genes based on three different criteria: 1) positional relation between associated variant and gene, 2) the layer of regulation that shows a gene expression change, and 3) the tissue in which the association was detected. The following sections will describe these classifications in more detail.

4.3.1 Local and distant associations

First we classify the QTL results based on the positional relation between associated variant and gene. Figure 4.3.1 illustrates this positional relation between gene loci and SDP loci in a genome-wide plot of all gene-SDP pairs showing the significant associations in left ventricle and liver.

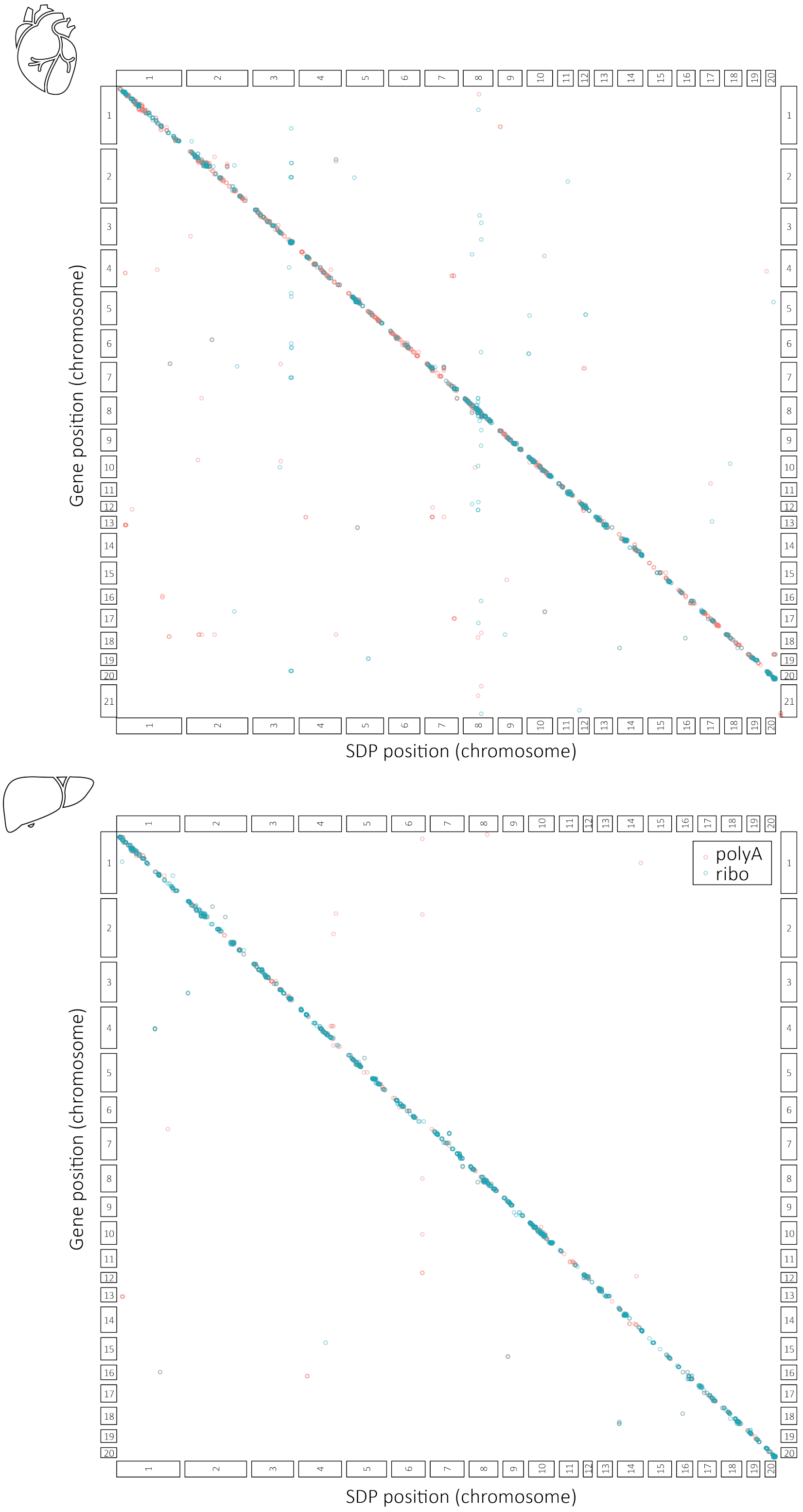


Figure 4.3.1: Genome-wide association plot of left ventricle and liver QTL mapping. Red circles represent eQTL and blue riboQTL. Local associations are visualised by dots along the diagonal and distant associations by off-diagonal dots. Genes that are regulated by the same SDP on both layers of regulation are shown by two overlaying circles.

As expected for QTL analysis, we observe a strong enrichment of associations along the diagonal which represent local associations, whereas the off-diagonal dots show the distant associations. In this figure the eQTL and riboQTL are depicted in one graph to show where associations occur on both layers of regulation. Additionally to the spurious off-diagonal dots, we also observe vertical stretches of associations, that represent SDPs that regulate multiple genes. The two vertical blue lines in the upper figure (left ventricle) represent distant associations that are purely driven by translational changes. Those distant riboQTL cluster will be further discussed in chapter 4.4.4.

Overall, hundreds of local associations have been identified in left ventricle and liver. The vast majority of all local QTL are located in a 10Mb window around the transcription start site (TSS) of the regulated genes. The relation between distance and strength of association is similar across tissues and layer of regulation (Figure 4.3.2).

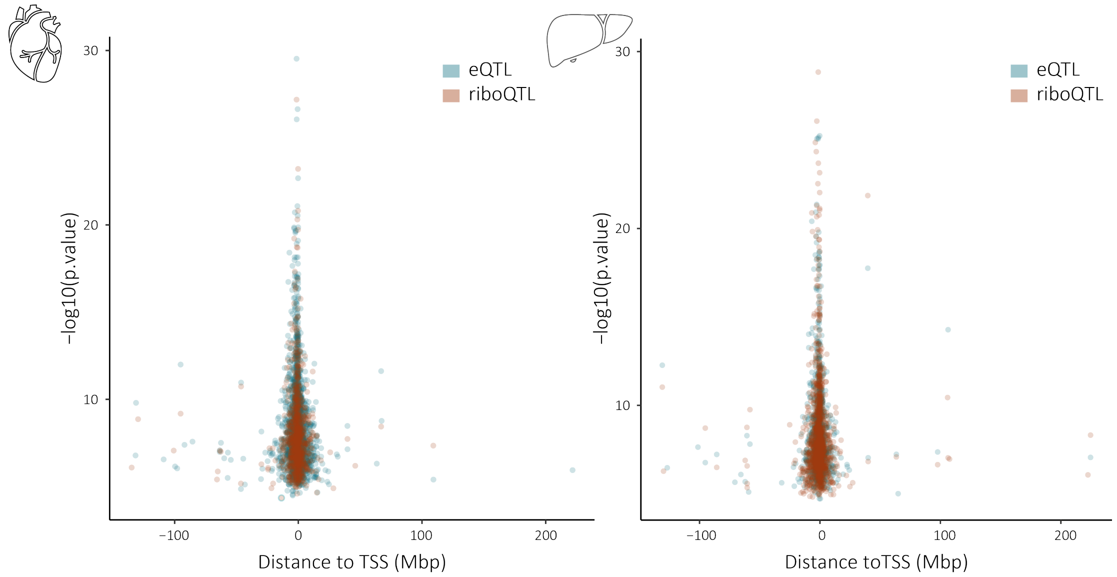


Figure 4.3.2: Distance of significantly associated SDP to the TSS of the associated gene. This plot is limited to associations on the same chromosome.

4.3.2 Disentangling transcriptional and translational regulation - Performing translational efficiency QTL (teQTL) mapping

The full repertoire of translational regulation includes more than only changes that purely occur on the translational level. Also changes that initiate on the transcriptional level and are absent on the translational level, as well as transcriptional differences that are enhanced on the translational level, describe translational regulation. By only defining translational regulation in genes that only show a riboQTL, we would miss many genes that are regulated translationally. This means that simply classifying genes based on their presence/absence of eQTLs and/or riboQTLs is not sufficient to understand the regulatory potential of genetic variants on translation.
Therefore, the translational efficiency (TE) is calculated and used as an additional quantitative trait using the same parameters as before. The translational efficiency allows to correct Ribo-Seq counts for the RNA-Seq effects. Using this approach and applying permutation testing as described before results in 71 local and 68 distant associations in the left ventricle and 88 local and 7 distant teQTLs in liver. The same significance cutoffs $\text{FDR}_{local} \leq 0.1$, $\text{FDR}_{distant} \leq 0.2$ and empirical p-value ≤ 0.0015 have been used in order to define significant teQTLs (Figure 4.3.3).

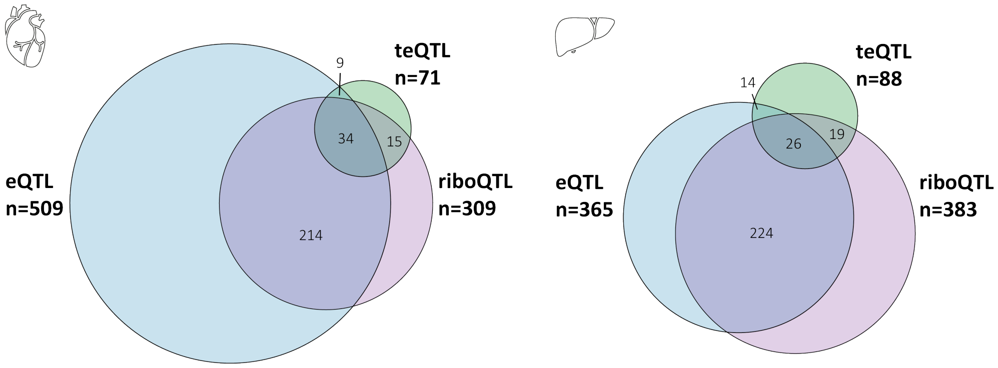


Figure 4.3.3: Overlaps of local QTL in left ventricle and liver

By overlapping the teQTL and riboQTL results, we are able to identify genes that are regulated on the translational layer only. In the left ventricle data, 9 genes show the same QTL for translation levels and translational efficiency. Among these 9 genes (*Lama4, Mlec, Lrp1, Lamb1, Ost4, Agrn, Hspg2, Scarf1* and *B4galt1*), 5 genes are regulated by the same SDP on chromosome 3, which will be further discussed in chapter 4.5. *Mlec, Ost4* and *Scarf1* are regulated by different SDPs in close proximity on chromosome 8. This putative trans cluster will be further discussed in chapter 4.4.4. In liver we do not observe any genes that are regulated purely on the translational level, indicated by having a teQTL and a riboQTL pointing to the same locus and genes.

4.3.3 Tissue comparison of genetically regulated genes

The total number of genes that are expressed in left ventricle and liver according to our transcription and translation thresholds are 8,266 genes, which means that the vast majority of the genes is expressed in both tissues. We detect 1,909 genes that are only expressed in left ventricle but not in liver, and 714 genes in liver specifically. In order to understand the role of translational regulation shaping a disease phenotype in more detail, it is important to study the role of genetic variants in different tissues as shown by the Genotype-Tissue Expression (GTEx) Project [72]. The experimental setup enables us to define genes that are only expressed in one of the two tissues and to identify tissue-specific regulation in these genes. Shared regulation in the two tissues can be studied by overlapping the results from pairwise association mapping or by using a joint QTL analysis performed with MT-Hess.

Among the 8,622 genes we identified common regulation in 126 genes on RNA-Seq level that are locally associated to the same SDP in both tissues. 81 genes carry a common local riboQTL and 20 genes a local teQTL using the pairwise association testing approach (Figure 4.3.4).

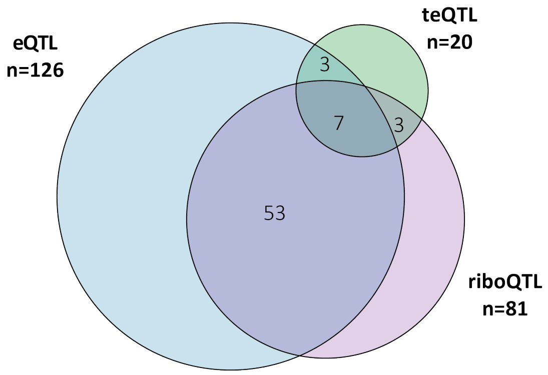


Figure 4.3.4: Overlap of genes that are expressed in both tissues and have a shared local QTL

The results suggest that common regulation mostly initiates on the transcriptional level and is often forwarded to the translational level. The low number of genes that also show a teQTL in both tissues can be explained on the one hand by lower effect sizes of TE that are caused by a smaller dynamic range. But on the other hand it suggests that regulation initiating on the translational level is more tissue-specific than transcriptional regulation.

Only one gene is distantly regulated in both tissues, the transmembrane and coiled-coil domain family 2 (*Tmcc2*) gene, which is located on chromosome 13 and regulated by an SDP on chromosome 1. Figure 4.3.5 shows two overlaying Manhattan plots, one for left ventricle in red and one in green for liver. The y-axis shows the strength of the associations and the x-axis the genomic location of each SDP. The black box shows the genomic location of *Tmcc2*. The overlay illustrates that the gene does not show any additional associations locally or distantly in either of the two tissues. So, the gene expression regulation is shared between both tissue and specifically regulated by a distant locus on chromosome 1.

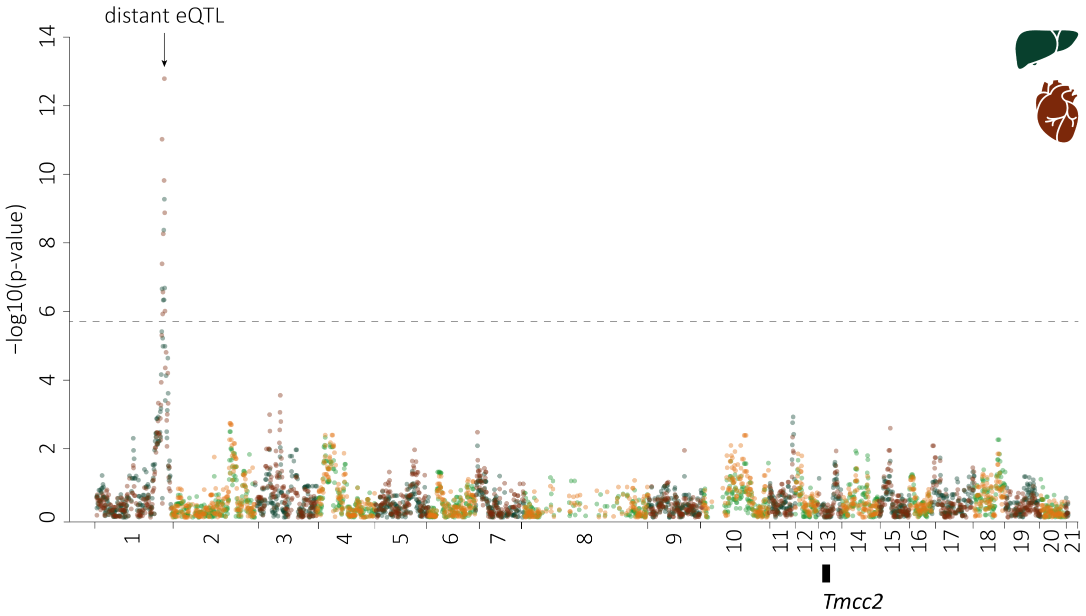


Figure 4.3.5: Manhattan plot of *Tmcc2* illustrating the distant eQTL on chromosome 1. The association results in the left ventricle are depicted in red and orange and the results in liver are light and dark green.

Tmcc2 shows an up-regulation on the transcriptional level in animals carrying the SHR genotype at the distant QTL. The translation levels of *Tmcc2* are showing no significant difference between the two genotype alleles, which suggests translational buffering (Figure 4.3.6).

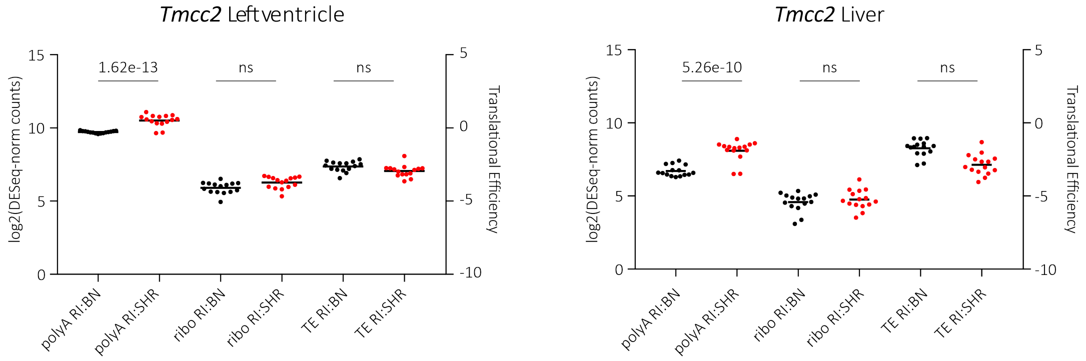


Figure 4.3.6: Gene expression of *Tmcc2* separated by shared distant eQTL in left ventricle and liver.

In order to enhance these findings of shared regulation across tissues, we perform a joint QTL mapping analysis of left ventricle and liver using the MT-Hess approach. Interestingly, no shared eQTL are observed in the two tissues. *Timp1*, which is located on the X chromosome, shows a comparable expression pattern in both tissues resulting in a riboQTL.

The fact that we cannot find shared gene expression regulation across the two tissues can have various explainations. In order to reduce the amount of false positive associations, very stringent cutoffs are chosen, which could result in missing those associations with smaller effect sizes across tissues. Langley et al published 158 *trans*-eQTL cluster across seven tissues in the rat RI panel [39] using a similar QTL mapping approach. Only 13 out of these *trans* eQTL cluster have been identified in at least two tissues. This suggests on the one hand that adding information of additional tissues might increase the power to detect shared regulation, but on the other hand that even having seven tissues only results in a small number of genes regulated across tissues over longer distances.

These results require careful interpretation, as the absence of a significant association cannot prove tissue-specific regulation. In order to derive tissue-specific regulation that only occurs in one of the two tissue, we assess the effect sizes of each gene-SDP pair and calculate the ratio of the effect sizes. A gene-SDP pair is defined as tissue-specific, if it only shows a significant association in one of the tissues and if the absolute ratio of effect sizes is 10 times more extreme in one or the other tissue (±3.322 on a log2 scale; Figure 4.3.7 a).

The absolute ratio of effect sizes is defined as abs(effect size left ventricle$_{S,G}$ / effect size liver$_{S,G}$) for each SDP (S) - gene (G) pair.

We identify 88 genes showing left ventricle-specific local eQTL and 54 liver-specific genes with a local eQTL (Figure 4.3.7 b). On the translational level, we observe 38 locally regulated genes in left ventricle and 66 in liver. The majority of associations that act over longer distances have been observed in the left ventricle including 5 genes on the transcriptional and 16 on the translational level. 3 liver-specific genes have been identified to show a distant eQTL.

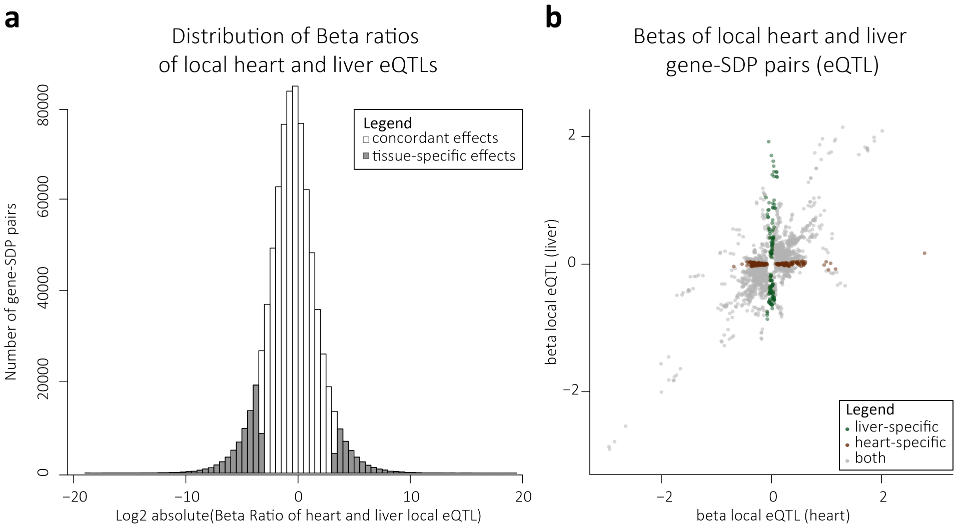


Figure 4.3.7: a) Example distribution of absolute beta ratios on a log2 scale for every SDP-gene pair that is showing local eQTL in liver and/or left ventricle. The grey bars represent the selected gene-SDP pairs that show an 1/10 ≤ absolute beta ratio ≥ 10. b) Corresponding classification of gene-SDP pairs as being liver- (green) or left ventricle-specific (red) associations exemplary on the local eQTL results.

In summary, we identify 16-25% tissue-specific regulation among local eQTL and riboQTL and 20-45% of genes that are regulated in a tissue-specific manner over longer distances. Even though this approach is very stringent, it allows the separation of genes just missing the significance threshold from those that indeed show tissue specific regulation.

4.4 Studying different mechanisms that regulate translation

As previously pointed out, genetic variants can be coding or non-coding in order to affect translation levels of a gene. The following paragraph describes known mechanisms of translational regulation. Genetic variants that change the coding sequence can for example remove or create a start or stop codon or change the reading frame of a protein. Non-coding variants occur outside the coding DNA sequence of a gene and change translation rates e.g. by disrupting or newly creating binding sites of miRNA or RBPs. Furthermore, variants can change an IRES or an uORF resulting in a translational regulation.

4.4.1 Identifying causal variants in local QTL

Causal variants define genetic alterations that have a consequence on the protein sequence of a gene, e.g. they result in a stop gain or stop loss, change the reading frame (frameshift) or the amino acid sequence (missense) of a gene. In order to identify these variants, we predict their effects with the Variant Effect Predictor (VEP) tool from Ensembl [73]. This can guide the identification of causal variants in local QTL. Therefore, first all SNPs located in the range of the associated local SDPs are identified and the consequences of each variant are predicted. This allows us to link putative consequences to the detected local QTL.
VEP predicts a damaging effect of two genetic variants on the corresponding associated proteins, which both show an eQTL and a riboQTL in liver. Additionally, we find 39 variants that have a predicted amino acid change with moderate impact on the amino acid sequence of the gene. In order to evaluate the impact of the missense variants, we annotate them with a protein function prediction tool that is applicable to rat - the scale-invariant feature transform (SIFT) [74] score (Table 4.6). Among those 39 only 6 variants are predicted to be deleterious missense mutations. The two examples with a predicted damaging effect will be further discussed in the following paragraph. They include a stop gain mutation in *Ces2e* that leads to a premature stop codon in exon 10, and a variant in an essential splice site of *Mst1* at the donor site of exon 6.

SNP	Consequence	Gene	QTL	SIFT
19:97061 T/G	Stop gain	*Ces2e*	Liver eQTL&riboQTL	-
8:116859358 G/A	Splice donor variant	*Mst1*	Liver eQTL&riboQTL	-
18:73263215 G/C	Missense variant	*Hdhd2*	LV eQTL	deleterious
19:42068673 A/G	Missense variant, Splice region variant	*Dhodh*	LV eQTL&riboQTL	deleterious
2:157987019 G/A	Missense variant	*Veph1*	LV eQTL	deleterious
2:261317097 T/C	Missense variant	*Tnni3k*	LV eQTL	deleterious
20:12866554 C/T	Missense variant	*Lss*	LV eQTL,riboQTL&teQTL Liver eQTL&riboQTL	deleterious
20:40543729 T/C	Missense variant	*Hs3st5*	LV eQTL	deleterious

Table 4.6: Variant effect prediction results of all local associations that either have high impact or moderate impact on the gene that is regulated.

By definition, a stop gain mutation leads to a premature stop codon that often results in a nonsense mediated decay (NMD) unless it occurs in the same exon as the canonical stop. NMD is a process that is present in all eukaryotic cells, which eliminates RNA transcripts that contain premature stop codons. During mRNA processing the exon-junction complex binds to each exon-exon junction and this complex normally gets removed by the ribosome in the first round of translation. If there is a premature stop codon and the ribosome does not reach the exon-junction complex it will retain at the mRNA. This activates the NMD pathway via up-frameshift proteins (UPF), which initiate the degradation of the mRNA [75, 76].

The degradation of the mRNA carrying the mutated allele often reduces the mRNA as well as the translation levels. Sometimes the wild type allele can compensate for the degradation of the mutant allele, which leads to unchanged transcription and translation levels of the gene. By assessing the allele specific expression of a gene, a shift in allele ratios from heterozygous to homozygous transcription can be observed. The observed stop gain mutation is occurring in the 10th of 13 exons in the gene carboxylesterase 2E (*Ces2e*). We observe that the stop mutation in animals carrying the SHR genotype correspond to a down-regulation on the transcriptional and translational level (Figure 4.4.1), which can be explained by partial NMD. In line with this hypothesis, we observe decreased levels on both layers of gene expression in the samples that carry the premature stop.

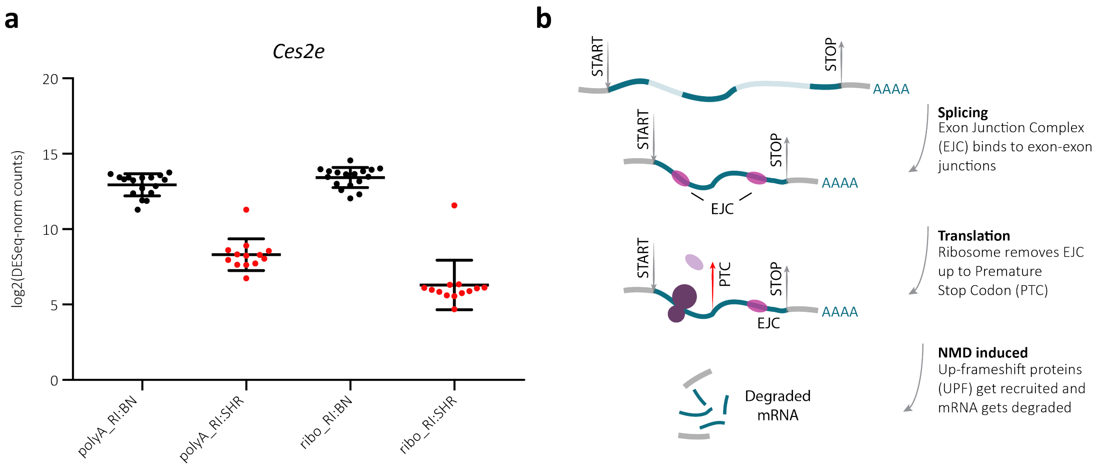


Figure 4.4.1: a) Expression levels of *Ces2e* on the RNA-Seq and Ribo-Seq separated by local QTL. b) Schematic of induced non-sense mediated decay by presence of pre-mature stop codon.

The second example is the macrophage stimulating gene 1 (*Mst1*), which carries a variant in an essential splice-site and shows reduced gene expression in animals that have a loss of the splice donor site (Figure 4.4.2. a). Similar to *Cse2e* this gene is only expressed in the liver data. According to the human gene expression data of the Genotype-Tissue Expression (GTEx) Project [72] *Mst1* is liver-specific across 53 human tissues (Figure 4.4.2 b).

We performed a targeted splicing analysis on this gene based on the SDP that carries the splice donor variant. Figure 4.4.2 c shows the percentage of spliced-in (PSI) reads per exon, which is the ratio of reads that are matching the exon and those that span the exon [77]. The PSI analysis is not showing significant differences between animals that carry the BN or the SHR genotype at this locus.

The splice site mutation occurs at the donor site of exon 6, which leads to lower levels of *Mst1* in both RNA-Seq and Ribo-Seq. The additional down-regulation on the Ribo-Seq level indicates an additional translational regulation. The disrupted essential splice-site likely leads to intron retention of intron 6 that activates the non-sense mediated decay machinery ultimately degrading the mRNA that was not properly spliced.

Both examples illustrate that understanding the exact mechanism of how genetic variants regulate gene expression is very challenging, even when potentially damaging variants in the QTL region are known. It is likely that the described two variants in *Ces2e* and *Mst1* cause the reduced expression levels, but there can be additional genetic effects which have not been considered here.

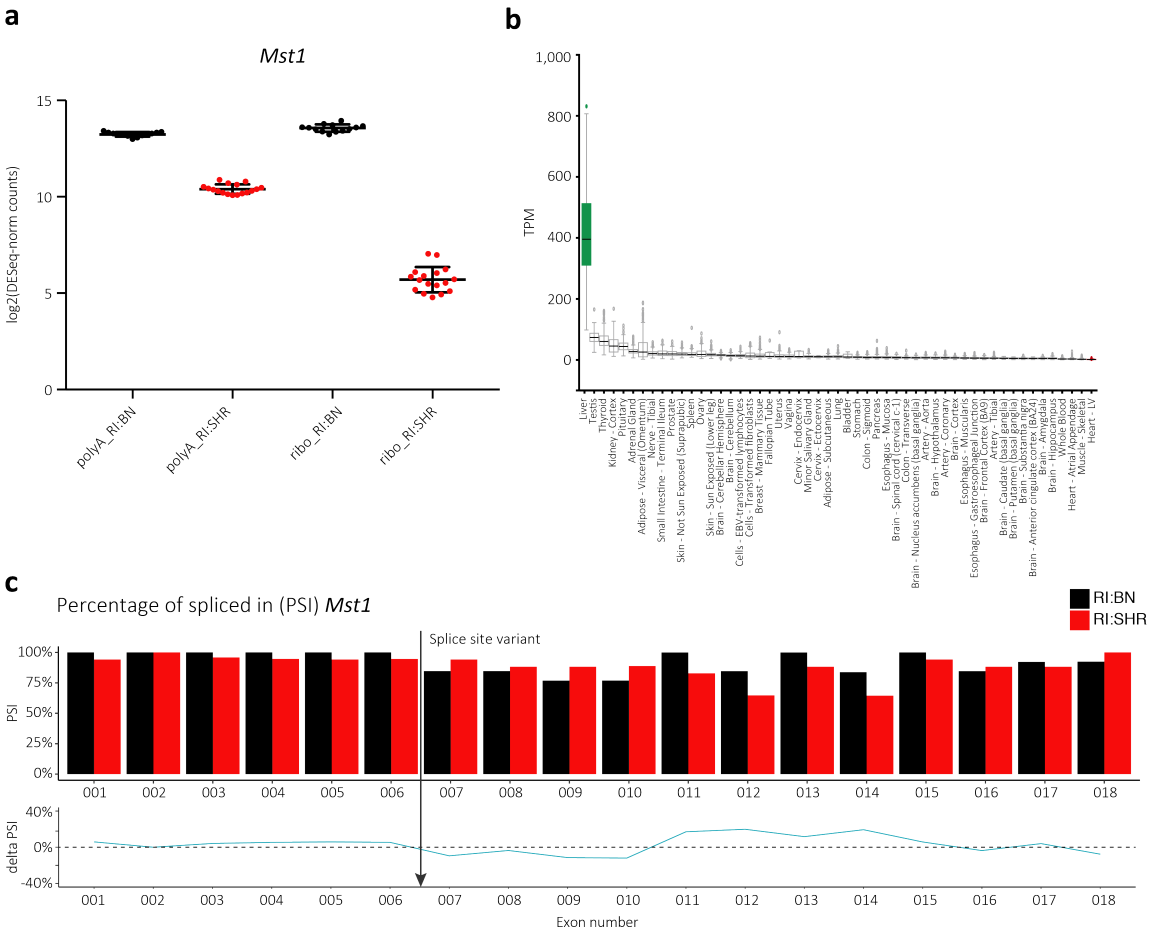


Figure 4.4.2: a) shows the gene expression levels in the animals that carry the BN allele and those that carry the SHR allele in the SDP that was significantly associated to the gene expression levels of *Mst1*. b) RNA-Seq gene expression of *Mst1* across 53 human tissues according to the Genotype-Tissue Expression (GTEx) Project. c) This PSI plot shows the splicing results. The lower panel represents a delta PSI plot $mean(PSI_{SHR}) - mean(PSI_{BN})$.

4.4.2 Understanding the contribution of uORFs to translational regulation

Upstream ORFs (uORFs), which are short ORFs that are located in the 5' untranslated region (5'UTR) of a gene, are known translational regulators. There are two different types of uORFs, one that starts and ends upstream of the CDS of a gene and one that starts upstream of the gene and overlaps with the mainORF (mORF). Both types of uORFs have the capability to decrease the translation rates of the mORF by competing for translating ribosomes (Figure 1.2.4).

The general idea is that uORF translation levels and mORF translation levels are anti-correlated, so that high uORF translation rates lead to low mORF translation rates and vice versa, as illustrated in Figure 4.4.3 [6].

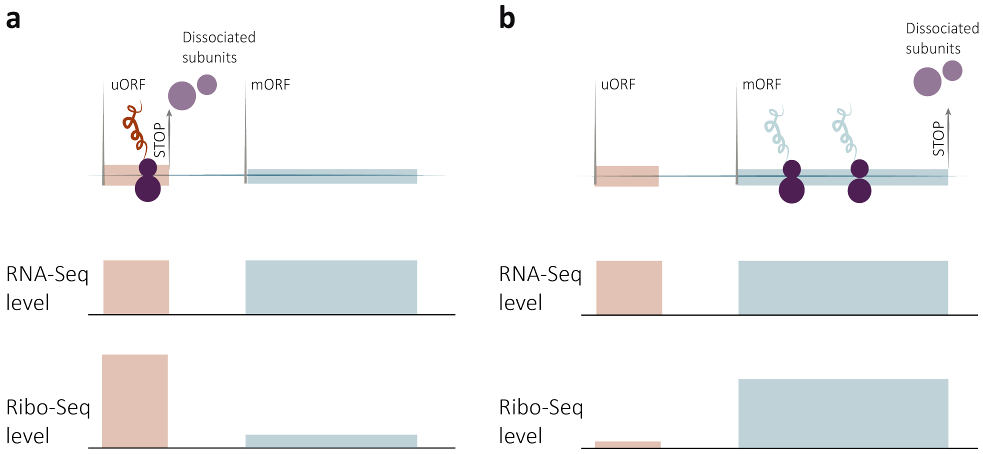


Figure 4.4.3: Translational regulation by uORFs a) The uORF is translated to a high extent and leads to lower translation levels of the mORF because the ribosomes dissociate at the stop codon of the uORF. The transcription levels of uORF and mORF are similar. b) The uORF is not translated, which results in increased translation levels of the mORF compared to the uORF while the transcription level of the gene is unchanged

To see whether this assumption can be validated in our data, we first use the tool RiboTaper to identify all uORFs in left ventricle and liver according to our data. We classify two types of ORFs as uORFs: independent uORFs, which are positioned upstream and fully separated from the mORF, and overlapping uORFs, which initiate translation upstream of the mORF in a different reading frame and only terminate downstream of the mORF translation initiation site (TIS). The independent uORFs can be easily filtered from the RiboTaper output, as they are pre-classified as uORF. The overlapping uORFs can be derived by filtering the output as following: the ORF has to start in the 5'UTR and overlap the annotated CDS of the same transcript. The uORF should not be in the same frame as the mORF and so the predicted resulting peptide should show no homology with the canonical ORF.

We identify a total of 795 uORFs (in 690 genes) that are independent and 330 overlapping uORFs (in 326 genes) in the left ventricle. In liver we obtain 929 independent (in 849 genes) and 330 overlapping uORFs (in 327 genes). Approximately, one third of the genes that have any type of uORF are shared between left ventricle and liver as depicted in Figure 4.4.4.

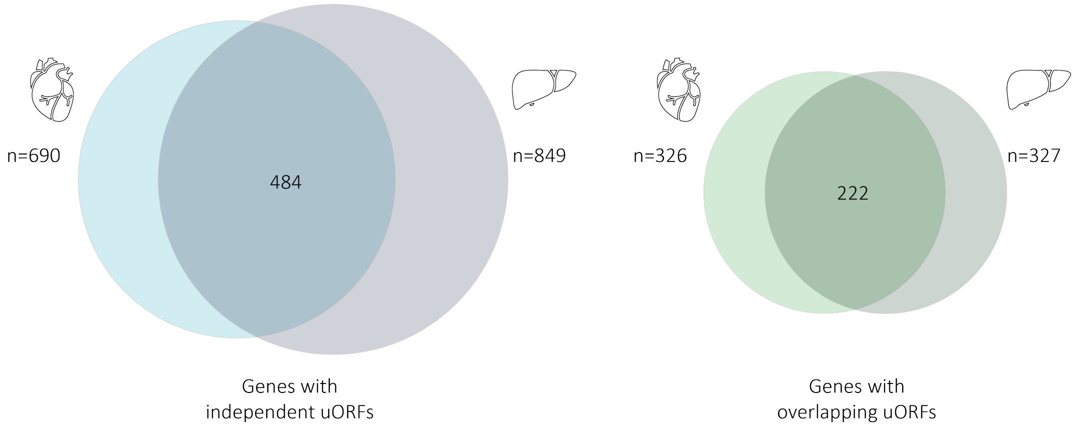


Figure 4.4.4: Venn diagram showing the number of detected uORFs in left ventricle and liver

We observe a higher number of detected uORFs in the liver data, which indicates that translational regulation in hepatic tissue is mostly mediated locally. This would explain the lower number of distant riboQTL and teQTL in the liver.

In order to assess possible translational regulation caused by uORFs two different approaches have been used. First, we calculate Spearman's correlation of every uORF-mORF pair using the uORF Ribo-Seq levels and mORF translational efficiency. The Ribo-Seq coverage of each uORF was normalised by its length and library depth to make it comparable to the TE of the mORF.

The hypothesis is that increased translation rates of the uORF results in a decreased translational efficiency of the mORF, shown by an anti-correlation of the uORF-mORF pair. Surprisingly, the majority of uORF-mORF pairs do not show any correlation rather than a positive or negative correlation, see Figure 4.4.5.

In left ventricle, we observe 8 uORF-mORF pairs that show a negative Spearman's rho $\leq$ -0.5 (*Wipi2, Dexi, Papola, Rb1cc1, Brd3, Tox4, Ext1* and *Rfwd2*) and 4 in liver (*Gprc5c, Yeats4, Tlr3* and *Icam2*). *Icam2* shows the strongest anti-correlation with a Spearman's rho of -0.64.

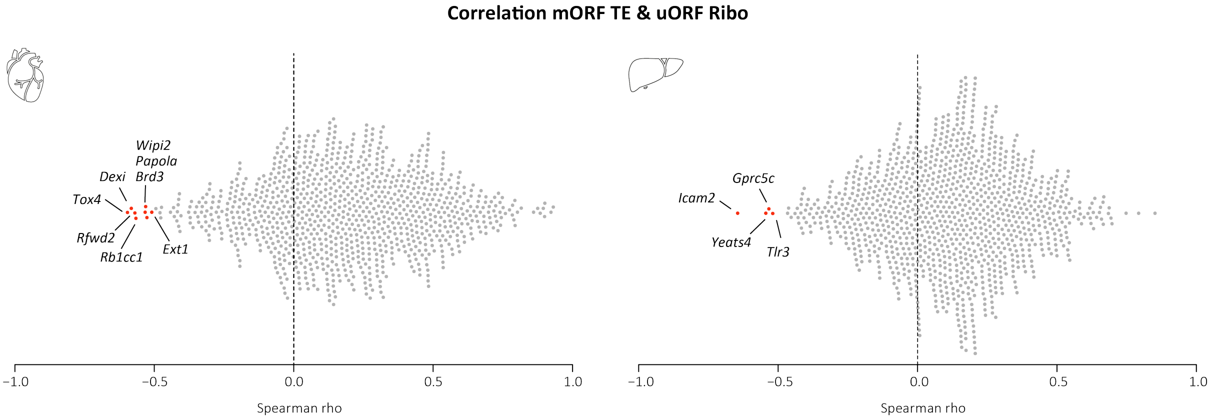


Figure 4.4.5: Beeswarm plot showing the Spearman correlation of all uORF-mORF pairs in both tissues on the x-axis. In red are uORF-mORF pairs with Spearman's rho ≤ -0.5 highlighted.

In Figure 4.4.6 two examples that show a significant anti-correlation of the uORF and their mORF translation are depicted.

The result of this analysis suggests that uORFs have a moderate but measurable impact on the efficiency of mORF translation even though for the majority of cases we do not observe an anti-correlation.

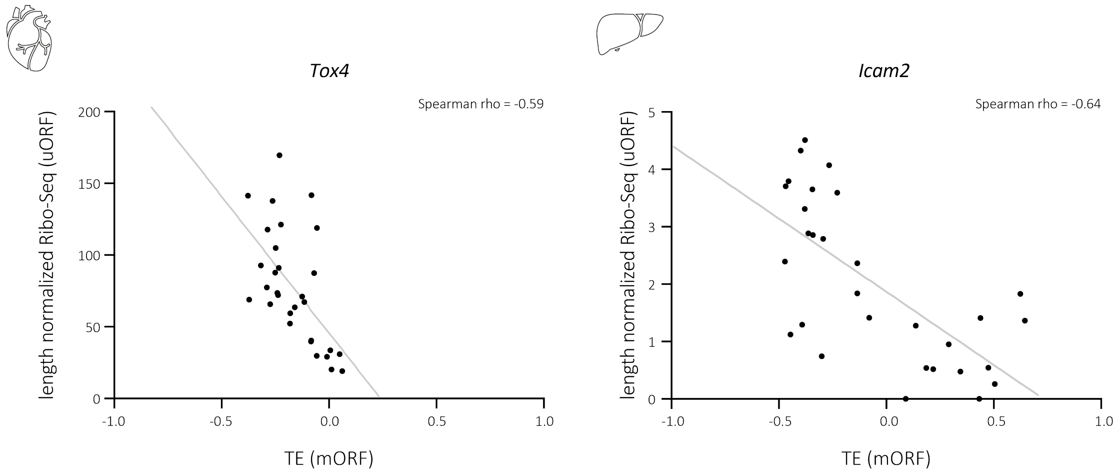


Figure 4.4.6: Two examples of anti-correlated uORF and mORF translation.

To investigate this further, we applied a second approach. QTL for the translation rates of uORFs have been assessed and the resulting uORF riboQTL have been overlapped with those observed in the mORF. Only local associations have been computed, as the assessment of distant QTL is not necessary, because distant genetic variants influencing the translation rates of uORFs are not expected.

In total, 46 uORF riboQTL in 26 genes in the left ventricle and 18 uORF riboQTL in 13 genes in liver have been detected. Overlapping these uORF riboQTL on a gene level with the mORF riboQTL or teQTL results in five genes carrying a uORF riboQTL in the left ventricle (*Rtel1, Vwa7* and *RGD1306001* in overlapping uORFs and *Adamts12* and *Gpr22* in independent uORFs). *Rtel1* and *Adamts12* additionally show a local teQTL.

In liver, five genes with a local riboQTL for the mORF and the uORF are observed. *Pvrl1* shows the associations with an independent uORF and additionally four overlapping uORFs have been linked to genetics (*Rtel1, Evc, Mocos* and *Lrrc45*). Exemplary, the expression levels of *Rtel1* in left ventricle are shown in Figure 4.4.7 separated by the local genotype that has been associated to both the uORF and the mORF.

These data suggest that not all uORFs regulate the translation rates of the mORFs via its abundance as previously observed by Chew et al. [8]. The regulation of mORF translation seems to correlate with the presence of a uORF, which might be explained by the resulting peptide that interferes with the translation of the mORF.

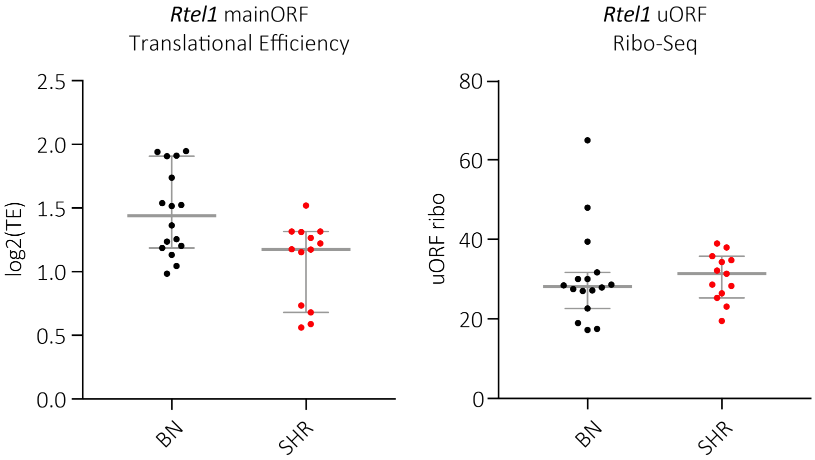


Figure 4.4.7: mORF and uORF expression of *Rtel1* separated by local genotype of the gene. The SHR genotype leads to a decrease translation of the mORF and an increased translation of the uORF.

4.4.3 Studying RBP binding affinities in order to understand their ability to alter translation levels

Another group of potential translational regulators are RNA binding proteins (RBP), which are proteins that bind RNA and carry structural motifs to bind to their target genes in order to regulate gene expression in various ways. RBPs are known to play a role in mRNA stabilisation, transport and localisation, but RBPs also regulate splicing and polyadenylation [78].

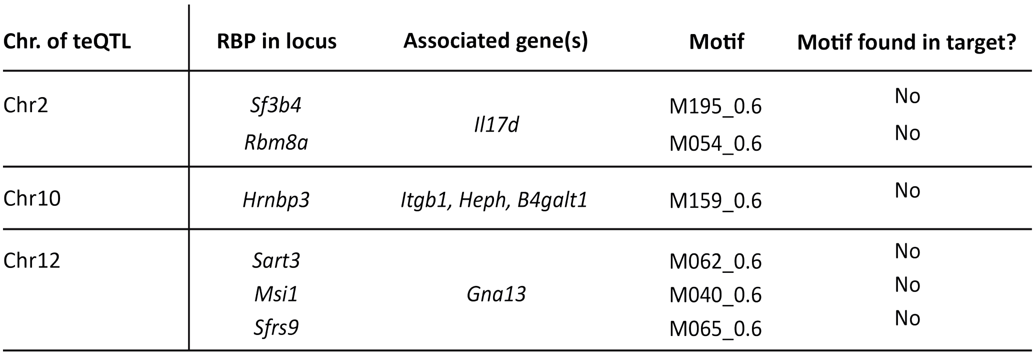
RBPs can recognise their targets via two mechanisms: 1) by the geometry and shape of the target gene [10] and 2) by a specific RNA motif that is located within the gene body of the target gene [11]. We predict the affinity of RBPs to all translationally regulated genes showing a local QTL. Hereby, we specifically focus on the role of genetic variants. A genetic variant can alter the gene expression levels of the RBP directly, which can lead to higher or lower levels of binding the target gene. Moreover, a variant can be located at the target site changing the binding motif of the RBP and act locally in close proximity to the target gene.

In the different QTL sets, we observed a slight enrichment of genes harbouring a putative RBP binding site in genes showing a local teQTL (Table 4.7). In left ventricle, we find 23 genes with a local teQTL and a predicted RBP binding site, which results in a significant enrichment (Fisher's exact test p-value 1.03e-46) suggesting that purely translationally regulated genes might be modulated to a large proportion by local variants that change the mRNA (e.g. motif or structure) of their target gene so that the RBP can no longer bind.

	Left ventricle	Liver
Local eQTL	37/699 (5.2%)	15/366 (4.1%)
Local teQTL	23/71 (32%)	10/88 (11%)

Table 4.7: Amount of transcriptionally and translationally regulated genes harbouring predicted RBP binding sites compared to the total number of transcriptionally and translationally regulated genes.

Besides the *cis*-regulatory role of RBPs, they have can change the translation levels of their targets over longer distances. To study these effects, we overlap distant QTL regions with a list of known RBPs in the rat based on the CISBP-RNA database. We find 8 RBPs in distant QTL regions that are contained in the database. In order to understand, whether these RBPs are in charge of regulating the translation levels of the associated genes, we performed a target prediction for RBPs with a known binding motif in the CISBP-RNA database. Binding affinities have been calculated for all translated genes in order to identify their targets. None of the 6 RBPs show a local QTL on either level of regulation. Unfortunately, the predicted binding targets do not overlap the genes with a distant riboQTL and/or teQTL (Table 4.8).



Chr. of teQTL	RBP in locus	Associated gene(s)	Motif	Motif found in target?
Chr2	*Sf3b4* *Rbm8a*	*Il17d*	M195_0.6 M054_0.6	No No
Chr10	*Hrnbp3*	*Itgb1, Heph, B4galt1*	M159_0.6	No
Chr12	*Sart3* *Msi1* *Sfrs9*	*Gna13*	M062_0.6 M040_0.6 M065_0.6	No No No

Table 4.8: Distant teQTL that harbour RBPs listed in the CISBP-RNA database and have a binding motif including the associated genes with those teQTL. Results of target prediction did not overlap the associated genes.

Even though, we were not able to draw conclusions about a specific relation between translational QTLs and RBPs due to the limited number of RBP binding motifs that are known for the rat, the data suggest that RBPs play a role in translational regulation, which was observed by the enrichment of RBP targets among the translationally regulated genes.

4.4.4 Assessing distantly regulated cluster of genes

Multiple genes linking to a single QTL form a distantly regulated QTL cluster. In this study, we call QTL with at least 3 associated genes over longer distances a distant QTL cluster. In total, 11 cluster have been observed in liver and left ventricle. The majority of cluster are detected in left ventricle and mostly on the translational level, i.e. a riboQTL and/or a teQTL.

In Table 4.9 all distant cluster are listed with the corresponding SDP ID indicating the location of the QTL (e.g. "SDPG_03000054231" describes a locus on chromosome 3 starting at 54,231bp), all associated genes and the type of association found to be significant for the group of genes.

QTL	Associated Genes	Type of Association
SDPG_03000054231	*Stt3a, Sema3c, LOC679811, Hmcn1*	LV distant teQTL
SDPG_03003960268	*Lama4, Lrp1, Lamb1, Agrn*	LV distant riboQTL&teQTL
	Fat4, Notch2	LV distant riboQTL
	Dchs1, Kdr, Hmcn1, Igf2r, Fat1, Sema3c, Uggt1	LV distant teQTL
SDPG_03006314843	*Lama4, Lrp1, Lamb1, Agrn, Hspg2*	LV distant riboQTL&teQTL
	Notch2, Strn3, Fat4, Ssc5d	LV distant riboQTL
	Dchs1, Fat1, Igf2r, Sema3c, Stt3a	LV distant teQTL
SDPG_03009805651	*Lrp1, Lama4, Agrn*	LV distant riboQTL&teQTL
	Fat4	LV distant riboQTL
	Kdr, Hmcn1, Dchs1, Igf2r	LV distant teQTL
SDPG_06006563392	*Gas7, Ccdc64, Cyp2c13, Car12*	Liver distant eQTL
SDPG_08051402844	*Rasgrp1, Atp6ap2, Strbp*	LV distant riboQTL
	Dlgap1	LV local eQTL, distant riboQTL
	Enpp5, Atp2b2	LV local eQTL, distant teQTL
	Acot7, Tspan7, Fam171b, Trim37, Ppp1r9a, Nlgn2	LV distant riboQTL&teQTL
SDPG_08052253560	*Nr2c2ap, Phf3*	LV distant riboQTL
	Ost4	LV distant riboQTL&teQTL
	Znrf1	LV local riboQTL, distant teQTL
	Cmss1	LV distant teQTL
SDPG_08059217527	*Rnpepl1, Mfhas1, Nold1, Eif4g3*	LV distant riboQTL
SDPG_10106181521	*Itgb1, Heph*	LV distant teQTL
	B4galt1	LV distant riboQTL&teQTL

Table 4.9: List of all distant cluster with at least 3 genes being associated to the SDP.

We further grouped distant QTL cluster, that fulfil the following criteria in order to increase the resolution:

- Neighbouring SDP
- large overlap in number of regulated genes

- same layer of regulation
- similar strain distribution pattern ($\leq$ 2 differences)

We observe two groups of QTL that show an overlap in regulated genes, one on chromosome 3 and one on chromosome 8. Three out of four cluster identified on chromosome 3 were grouped based on the above mentioned criteria. These three cluster largely overlap in their group of regulated genes and mode of regulation. Additionally, the SDPs differ only marginally from each other (Figure 4.4.8). The three QTL are treated as one cluster subsequently.

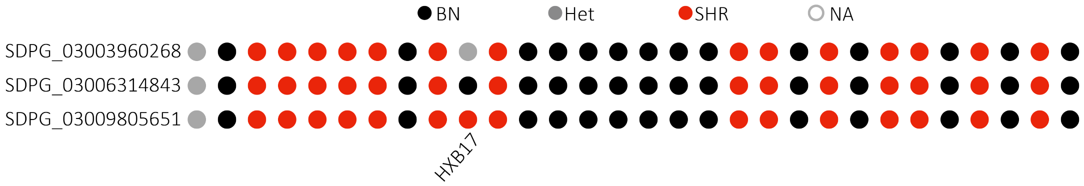


Figure 4.4.8: Strain distribution pattern of the SDPs SDPG_03003960268, SDPG_03006314843 and SDPG_03009805651 that form the distant cluster on chromosome 3. Each dot represents the genotype - BN in black and SHR in red in the three SDPs across all RI lines.

The QTL cluster on chromosome 8 only show a very marginal overlap in regulated genes and different strain distribution patterns (Figure 4.4.9), so they are not merged for further analyses.

For all seven cluster, a Gene Ontology (GO) enrichment analysis using g:profiler [79] was applied in order to understand their common function or involvement in similar pathways. All expressed genes in left ventricle and liver respectively have been used as background set in order to test for enrichment of GO terms, Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways and binding sites for miRNAs and transcription factors (TF) . This results in a significant enrichment for two translationally regulated cluster, located on chromosome 3 and 8.

In total, 17 genes are regulated by the chromosome 3 locus and have been tested for enrichment. The most enriched term is "basement membrane" accompanied by other terms related to the extracellular matrix. Three regulated genes are linked to extracellular matrix receptor interaction (*Laminin alpha 4, Laminin beta 1* and *Agrin*). Additionally, 7 genes are associated to calcium ion binding. Both terms can be linked to heart disease and therefore might be of high relevance for shaping the

SHR phenotype (Table 4.10).

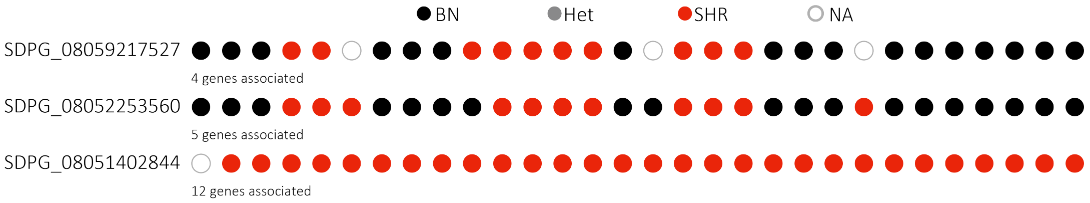
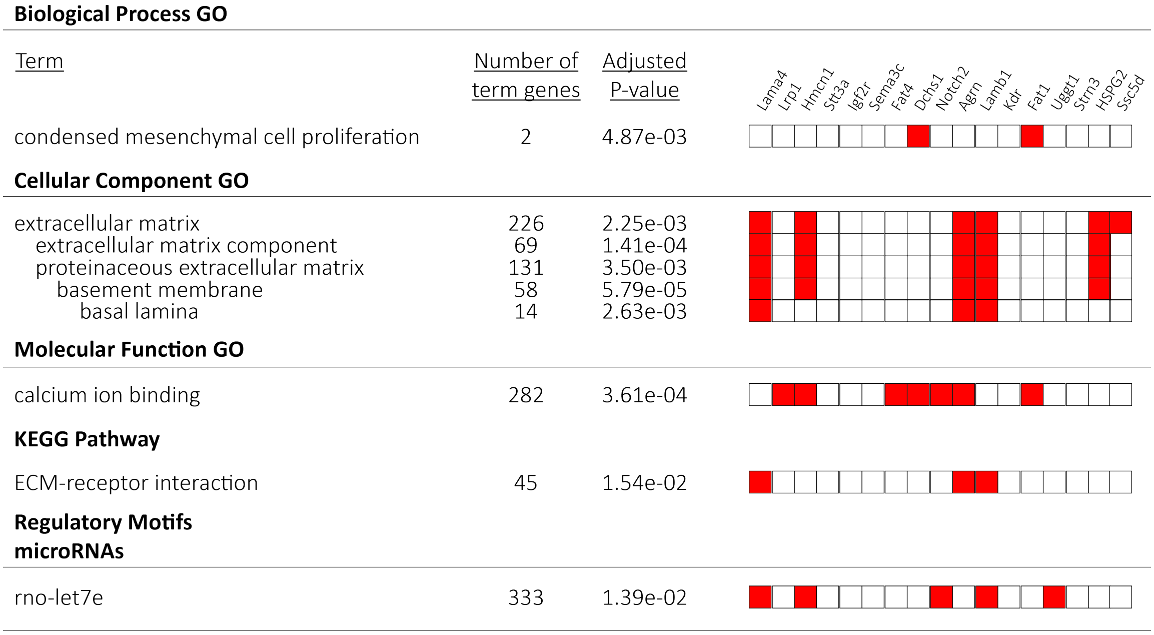


Figure 4.4.9: Strain distribution pattern of the SDPs SDPG_08059217527, SDPG_08052253560 and SDPG_08051402844 that are located on chromosome 8 and regulate a number of genes on the translational level. Each dot represents the genotype - BN in black and SHR in red in the three SDPs across all RI lines.



Term	Number of term genes	Adjusted P-value	Lama4	Lrp1	Hmcn1	Stt3a	Igf2r	Sema3c	Fat4	Dchs1	Notch2	Agrn	Lamb1	Kdr	Fat1	Uggt1	Strn3	HSPG2	Ssc5d
Biological Process GO																			
condensed mesenchymal cell proliferation	2	4.87e-03								🟥					🟥				
Cellular Component GO																			
extracellular matrix	226	2.25e-03	🟥		🟥							🟥	🟥					🟥	🟥
extracellular matrix component	69	1.41e-04	🟥		🟥							🟥	🟥					🟥	
proteinaceous extracellular matrix	131	3.50e-03	🟥		🟥							🟥	🟥					🟥	
basement membrane	58	5.79e-05	🟥		🟥							🟥	🟥					🟥	
basal lamina	14	2.63e-03	🟥									🟥	🟥						
Molecular Function GO																			
calcium ion binding	282	3.61e-04		🟥	🟥				🟥	🟥	🟥	🟥			🟥				
KEGG Pathway																			
ECM-receptor interaction	45	1.54e-02	🟥									🟥	🟥						
Regulatory Motifs microRNAs																			
rno-let7e	333	1.39e-02	🟥		🟥						🟥		🟥			🟥			

Table 4.10: Results from g:profiler - testing for enriched terms for all 17 genes that are regulated by distant cluster that is located on chromosome 3.

Of the three cluster on chromosome 8, only "SDPG_08051402844", associated with 12 genes, show a significant GO enrichment. The GO enrichment suggests a role in aggresome assembly based on the 2/12 genes (*Dlgap1* and *Trim37*).
All distant associations are visualised in a circos plot (Figure 4.4.10), which shows the genomic positions of SDPs and genes on the outer circle and the associations as connecting lines in the middle.

The distant QTL that have been identified on at least one of the three layers, RNA-Seq, Ribo-Seq and translational efficiency are displayed. By definition, distant associations occur between SDPs and genes located on different chromosomes. We can see an enrichment of connections initiating on chromosome 3 and 8 that represent the distant cluster described above. The following chapter focusses on the distant riboQTL, which is located on chromosome 3 (depicted in green), to further assess biological mechanisms and validate initial findings.

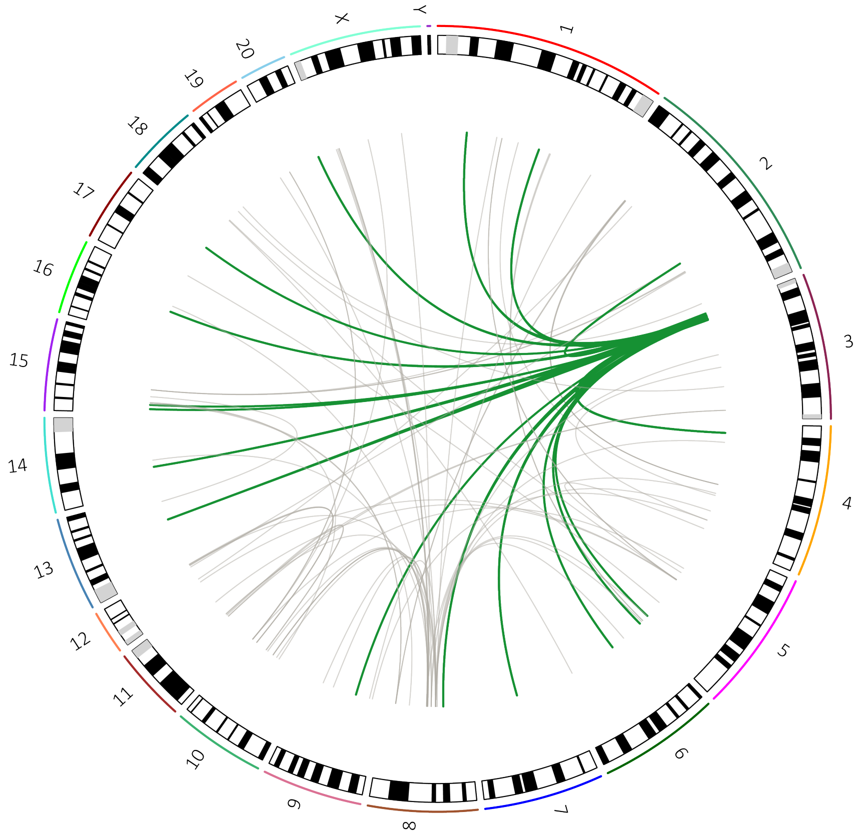


Figure 4.4.10: Circos plot of all distant cluster in the left ventricle. All 17 genes that show either a distant riboQTL or a distant teQTL for the cluster on chromosome 3 are highlighted in green.

4.5 Investigating the translational regulation of extracellular matrix production by a QTL on chromosome 3

The global eQTL and riboQTL mapping in the rat RI panel identified an approximately 7Mb large locus on chromosome 3 that seems to affect the translation levels of a number of genes located elsewhere in the genome. In total 17 genes have been linked to this locus purely on the translational level.

Figure 4.5.1 shows exemplarily six genes that illustrate the comparable association patterns across the detected genes.

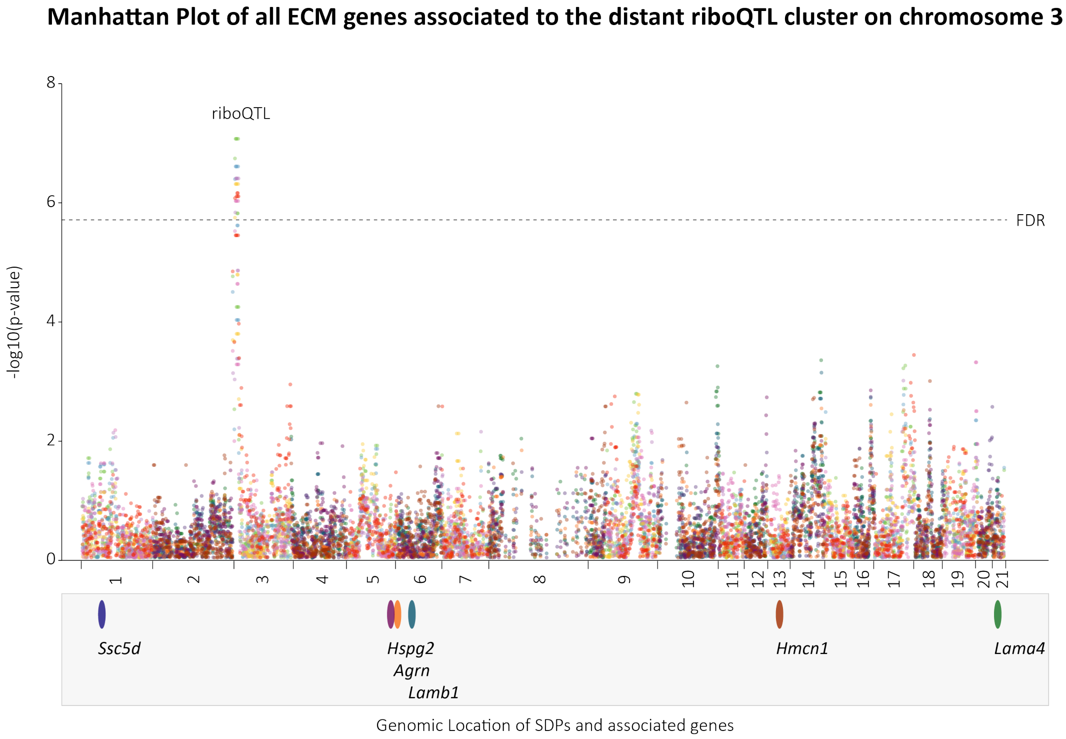
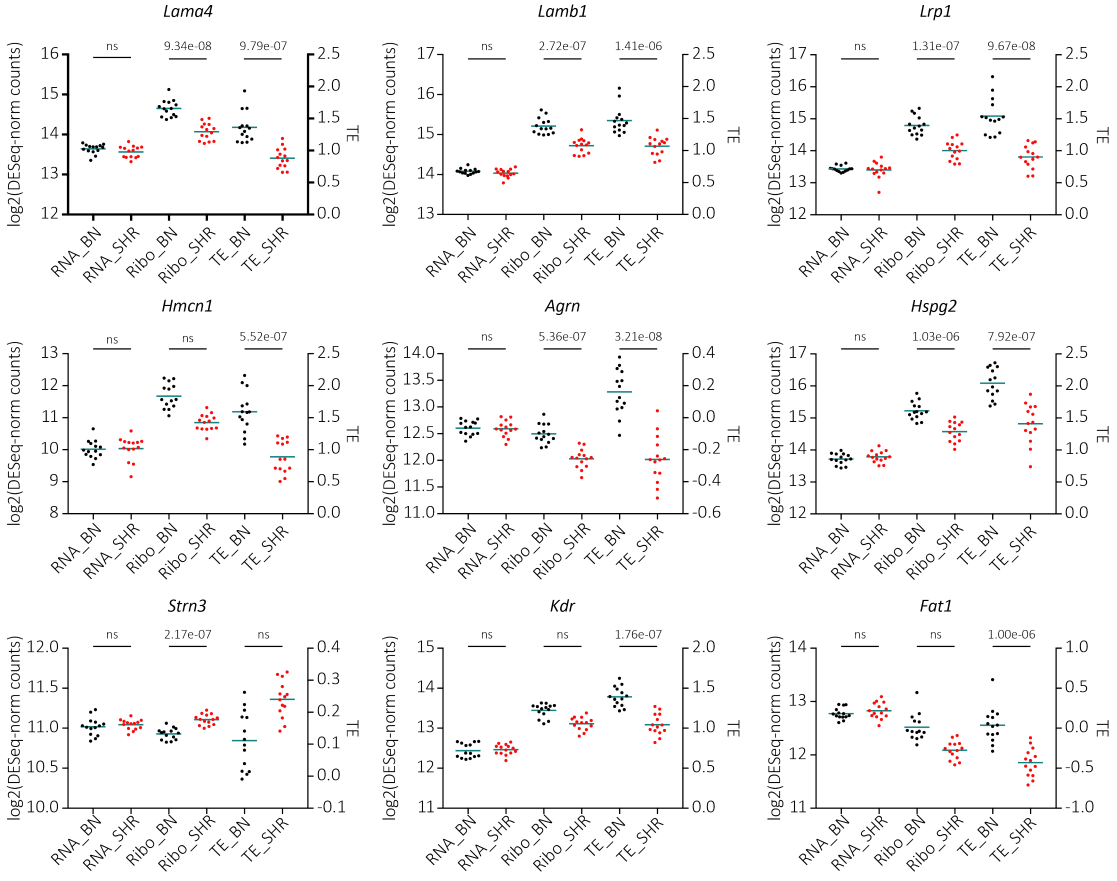


Figure 4.5.1: Overlay of six individual manhattan plots illustrate the strength of association of six example genes that have been linked to the chromosome 3 locus. The manhatten plots are color-coded based on the genes that show the associations. The genomic locus of the SDP is plotted on the x-axis and the strength of association as negative log10 p-value on the y-axis. Each dot represents a single gene-SDP pair. The FDR for this experiment is at a p-value cutoff of 1.9e-06 and represents an FDR ≤ 0.1.

A strong association to the distant riboQTL on chromosome 3 and no additional effects from other SDPs are consistently observed in these genes. On the RNA-Seq level, these genes show no association, which indicates that the effect only initiates on the translational level and therefore is this effect is mainly translationally driven. To validate this finding the riboQTL and teQTL results have been overlapped resulting in the detection of 5 out of 9 genes associated to Ribo-Seq levels and the translational efficiency. Additionally, the TE levels of 8 genes have been found to link to this locus. The expression levels of all 17 associated genes are depicted in Figure 4.5.2.

Here, the 30 RI animals are separated by genotype at the QTL and abbreviated as RNA_/Ribo_/TE_BN corresponding to the expression levels of animals carrying the BN allele and RNA_/Ribo_/TE_SHR for the SHR allele. The transcription levels do not show a significant difference between animals of different genotype at SDPG_03003960268, whereas a clear difference on the translation levels is observed. All associated genes except of *Striatin 3* (*Strn3*) show reduced translation levels in animals carrying the SHR genotype.



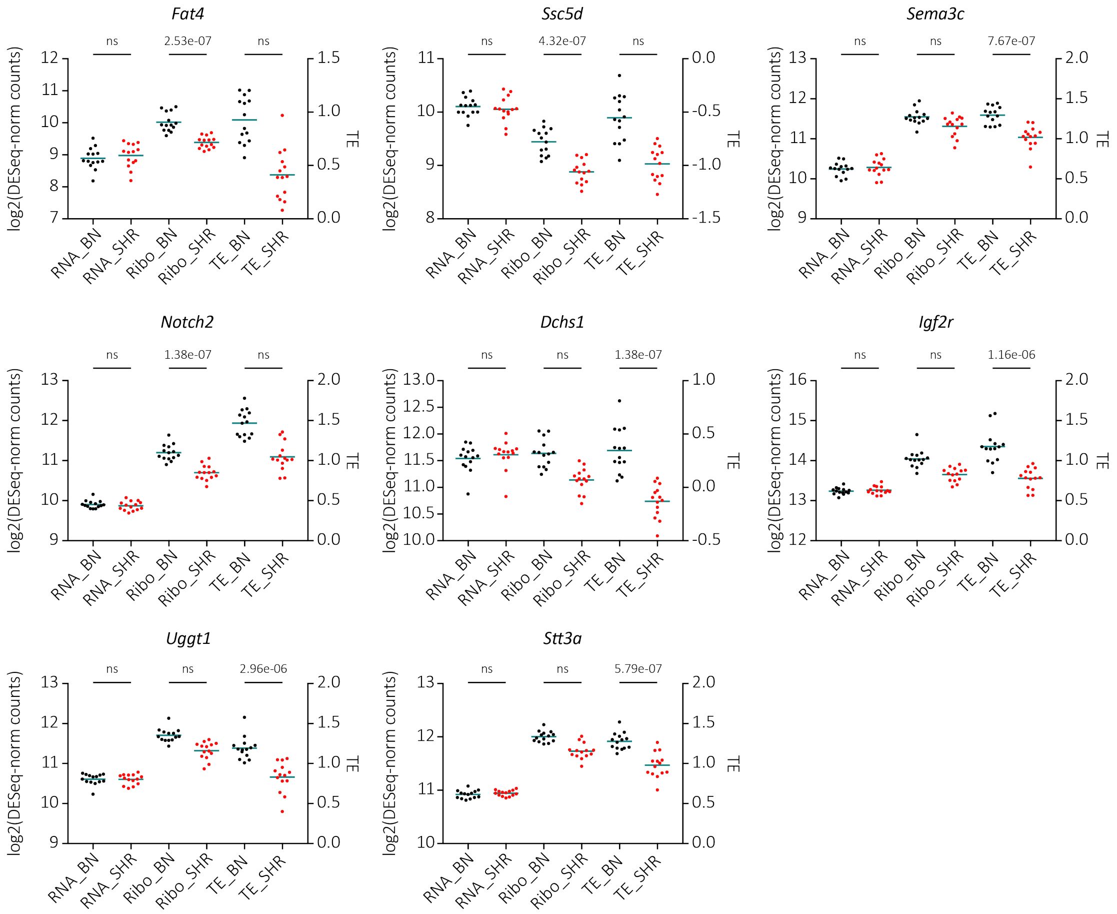


Figure 4.5.2: Expression levels of all 17 genes that are translationally regulated by the chromosome 3 cluster separated by genotype of SDPG_03003960268. Nominal p-values shown in figure. If the association was not significant according to the thresholds FDR ≤ 0.2 and empirical p-value ≤ 0.0015 it is indicated accordingly.

This common regulation and the strong enrichment for genes linking to the GO Terms "extracellular matrix" and "calcium ion binding" indicate a common regulatory mechanism. There are no associations of these genes observed in liver, which suggests that the observed mechanism is specific for the left ventricle and therefore might be relevant for heart disease.

4.5.1 Using MT-Hess to gain a greater understanding of the distantly regulated genes

The pairwise QTL mapping approach is quite stringent and only picks up gene-SDP pairs with sufficient effect sizes. In order to get a more complete list of genes that are translationally regulated by the chromosome 3 locus, a grouped QTL mapping approach is used to increase power specifically for shared regulation. We use the MT-Hess approach in a single tissue mode to identify groups of genes that show common regulation patterns. Figure 4.5.3 compares the results from MatrixEQTL and MT-Hess in left ventricle.

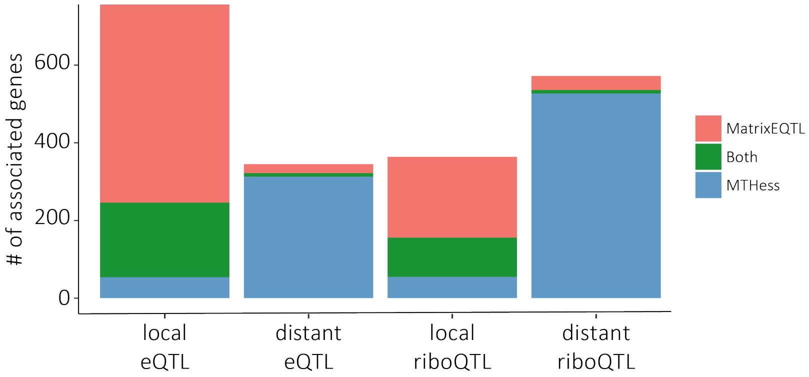


Figure 4.5.3: MatrixEQTL and MT-Hess results in left ventricle

To assess the significant associations determined by MT-Hess, we pick comparable FDR cutoffs as for the MatrixEQTL results. For local associations require a FDR $\leq$ 0.05, which corresponds to a MPPI (marginal posterior probability of inclusion) ≤ 0.74 in Ribo-Seq and 0.76 in RNA-Seq, distant associations have a cutoff that correspond to an FDR $\leq$ 0.1, which is a MPPI $\leq$ 0.59 and 0.61 in Ribo-Seq and RNA-Seq, respectively. Figure 4.5.3 illustrates the strength of MT-Hess in the identification of distant associations. As local associations are mostly specific to the gene in close proximity, MT-Hess cannot make use of shared patterns and therefore only a small number of local associations are found by MT-Hess.

The strength of MT-Hess is the detection of large gene cluster regulated by the same genetic locus. Additionally, to the 17 genes identified by MatrixEQTL, 45 additional genes, purely regulated on the translational level by the SDP on chromosome 3, have been found. Performing a GO enrichment analysis on this set of genes resulted in a strong enrichment for extracellular matrix genes with a total of 16 ECM genes (p-value: 7.61e-11, see Figure 4.5.4).

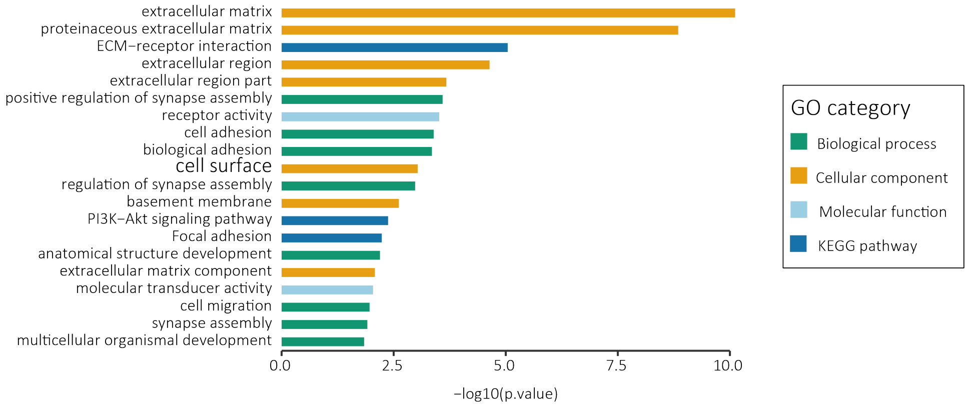


Figure 4.5.4: Top 20 GO-Terms and KEGG pathways enriched in the 45 genes associated with SDPG_03003960268. The figure shows a strong enrichment for extracellular matrix protein-related terms

Figure 4.5.5 shows the components of the extracellular matrix that have been found to be translationally regulated either by MatrixEQTL alone (n=2), by MT-Hess alone (n=12) or by both approaches (n=4) in a simplified schematic.

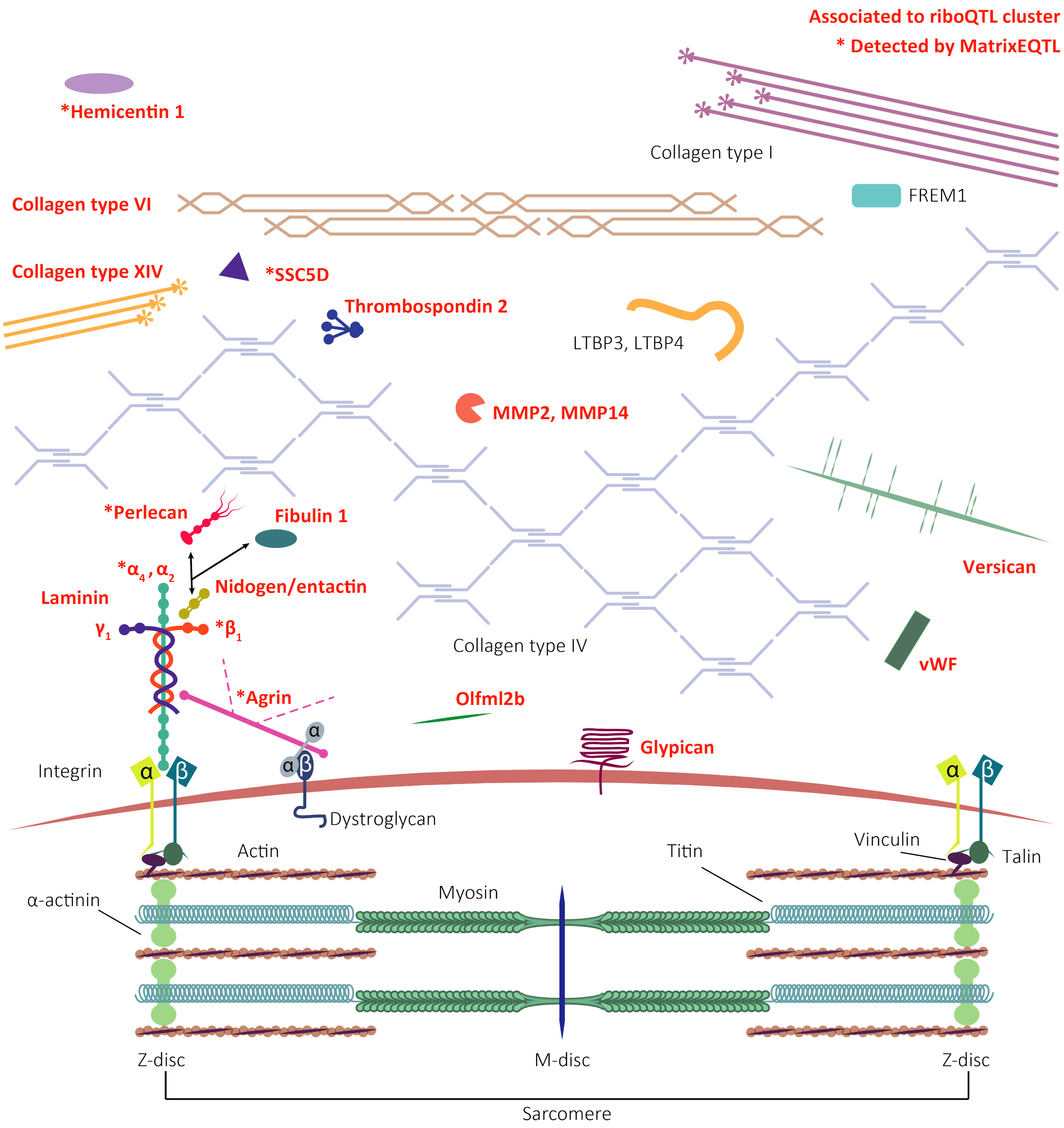


Figure 4.5.5: Schematic of extracellular matrix proteins highlighting the genes that have been identified to be translationally regulated by the distant ribo-QTL cluster on chromosome 3. All genes that have been associated in the MatrixEQTL or MT-Hess run are highlighted in red. Those that have been found in the initial run using MatrixEQTL are indicated with an asterisk.

4.5.2 Zooming into the regulatory locus for translational regulation on chromosome 3

The regulatory locus for translational regulation of a number of ECM genes spans a locus on chromosome 3 from 3.96Mb to 10.94Mb including 65 genes that are translated in the rat left ventricle. We performed a variant effect prediction using VEP from Ensembl on all SNPs located in this locus resulting in missense mutations in 20 genes. The majority of missense mutations are predicted to be tolerated, 7 genes show a deleterious missense mutation that might have an impact on the transcription or translation levels of these genes (Table 4.11).

SNP	Consequence	Gene	SIFT	Granthan Score
3:5608532 G/A	Missense variant	*Mymk*	-	74
3:5126449 G/A	Missense variant	*Abo3*	deleterious	21
3:7486953 G/A	Missense variant	*Ddx31*	deleterious	21
3 4042119 G/A	Missense variant	*Egfl7*	deleterious	98
3:8758093 T/C	Missense variant	*Kyat1*	deleterious	0
3:9814822 T/C	Missense variant	*RGD1305178*	deleterious	0
3:7657786 C/T	Missense variant	*Ttf1*	deleterious	81

Table 4.11: Variant effect prediction for genes that are located in the QTL region on chromosome 3 and carry a deleterious missense mutation.

Furthermore, the Granthan score [80] is added for each missense variant, in order to estimate how evolutionarily distant the reference and variant amino acids are. All 7 genes show a Granthan score below 100, which means that the effect is unlikely having an effect on the gene.

Besides assessing the genetic differences of genes in the regulators locus, we also assessed regulatory candidates based their known function or other features. For example, the gene EGF-like domain-containing protein 7 (*Egfl7*) carries a G-to-A mutation corresponding to a substitution of a glycine to glutamic acid, which could affect the stability or three-dimensional conformation of the gene. Perhaps this is altering its ability to bind other genes. *Egfl7* carries an EMI domain, which is a small cysteine-rich protein domain of around 75 amino acids that is often present in ECM proteins. Known interaction partners of Egfl7 also include Notch-class proteins, which have been found to be regulated by the chromosome 3 locus. Therefore, Egfl7 might have the ability to regulate the translation levels of other ECM genes and is a good candidate gene.

Additionally, the locus contains other candidate regulators, such as the RNA binding proteins Fubp3, Rexo4 and Ddx3. To date their target genes are unknown and therefore experimental follow-up would be required to understand their potential role in translational regulation of these genes.

To narrow down the list of potential regulator genes, literature research pointed to a previously described rat phenotype QTL for cardiac hypertrophy that overlaps part of the 7Mb large riboQTL [34, 47]. Figure 4.5.6 displays the overlapping region of all three QTL and the genes that are located in this region. As remodelling of the ECM is one characteristic often described in hypertension and cardiac hypertrophy, this smaller QTL region could also contain the causal gene regulating ECM protein translation.

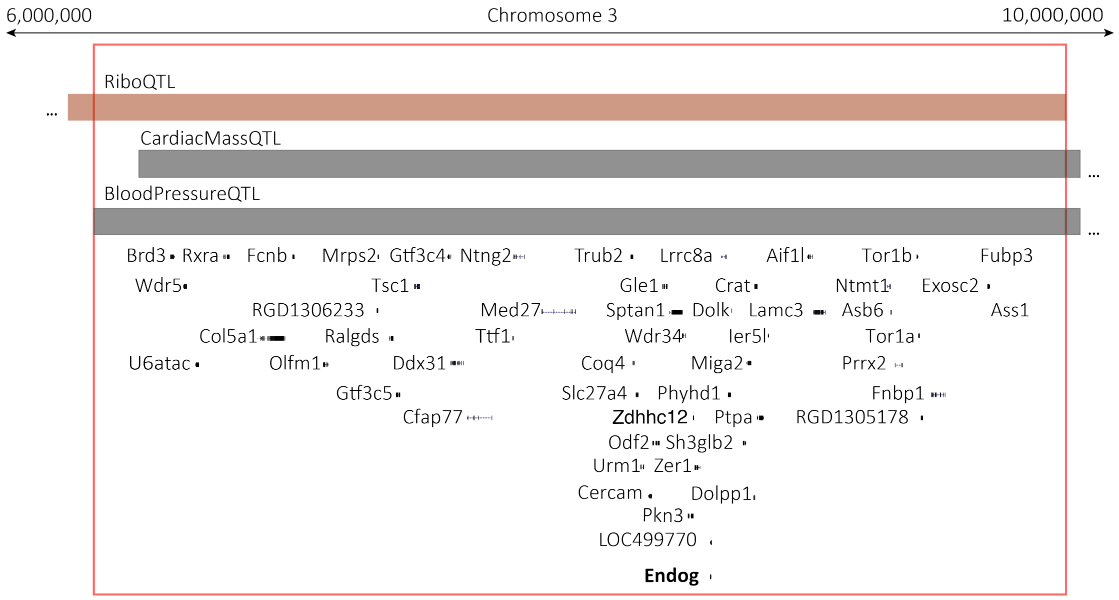


Figure 4.5.6: The distant riboQTL/teQTL on chromosome 3 overlaps with two cardiac phenotype QTL

In order to validate the cardiac phenotype QTL two congenic rat strains have been generated including the riboQTL locus. RNA-Seq and Ribo-Seq data was generated and used as an independent validation dataset for the distant riboQTL cluster.

4.5.3 Independent validation of distant trans cluster using rat congenic strains

The two congenic rat strains have been derived by a selective breeding regimen that was employed to integrate segments of the BN genome onto an SHR background [47]. They have have differently sized BN insertions on the SHR background. The strain carrying the longer fragment of BN is called SHR.BN-(3L) and the shorter SHR.BN-(3S). The two strains differ genomically in 11Mb. This locus almost completely overlaps the distant riboQTL that we detected in the RI panel.

In order to ascertain whether the two congenic strains show the same expression pattern that was observed in the RI panel, a differential expression analysis on RNA-Seq and Ribo-Seq using DESeq2 was performed [55] resulting in 924 differentially transcribed and 912 differentially translated genes. 716 of the 912 genes are located on a different chromosome than the QTL and do not show differential expression on the RNA-Seq level. Among those 716 genes, 44 genes are associated to the ECM resulting in a GO enrichment p-value of 2.04e^{-08} (Figure 4.5.7).

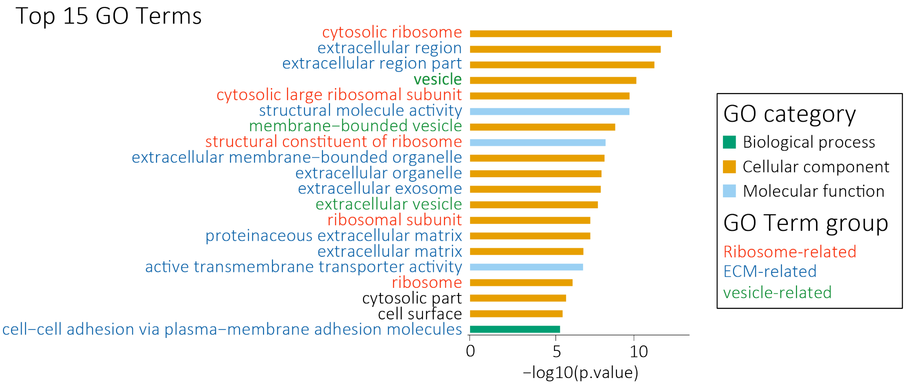


Figure 4.5.7: GO enrichment results from testing 716 genes that are purely regulated on the translational level and not located on chromosome 3. The enrichment is depicted as bars representing the p-value on the x-axis with the bars coloured according to GO category; biological process (green), cellular component (orange) and molecular function (blue). Additionally, the GO terms are further classified into "Ribosome-", "ECM-" and "Vesicle"-associated GO Terms, which represent the three main GO groups we observed for these genes.

Of the ECM genes that have been linked by QTL mapping using MatrixEQTL and MT-Hess to the distant cluster, 12 genes replicate in the congenic rat strains. In Figure 4.5.8, 6 example ECM genes are shown to illustrate the concordant observations in the RI panel (separated by genotype at the riboQTL) and the two congenic strains that show the BN.Lx genotype at this locus or the SHR genotype.

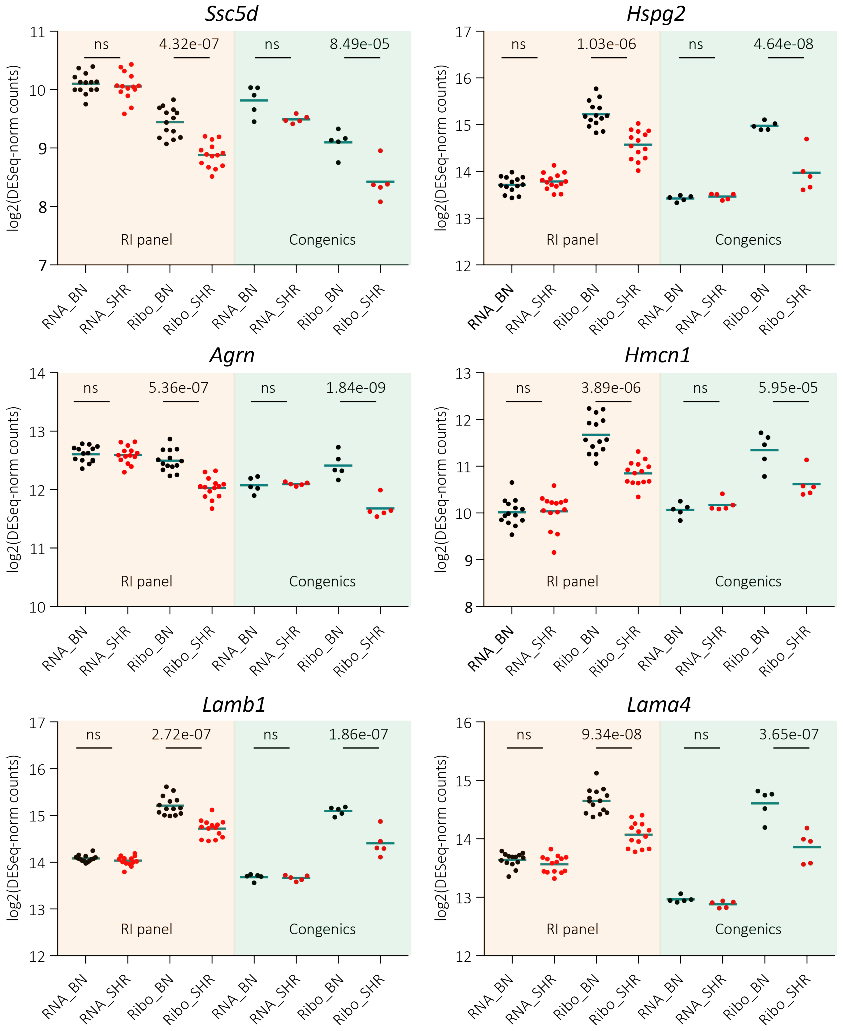


Figure 4.5.8: Expression plots of 6 ECM genes in RI panel (orange) and congenic rats (green). We see concordant down-regulation in animals carrying the SHR genotype in the RI panel and the congenic rat strains.

In both experiments, no differences at the RNA-Seq level are observed, but a significant decrease in translation. This data enables us to validate the distant riboQTL cluster in the congenic rats.

The GO Enrichment analysis not only shows a strong enrichment for ECM genes

but also shows an enrichment for various ribosome and vesicle associated GO terms (Figure 4.5.7).
Both datasets, the RI panel and the congenic strains show mainly a down-regulation in translation of the ECM genes in animals carrying the SHR genotype at the riboQTL locus. Therefore, we next assessed the direction of translational regulation in all three GO Term groups in order to understand whether these genes show similar expression patterns. We plot the $\log_2$ fold changes of SHR.BN-(3S)/SHR.BN-(3L), which corresponds to SHR/BN.Lx for this locus. All RNA-Seq and Ribo-Seq fold changes are calculated and ordered by the Ribo-Seq fold changes from small to large, in order analyse the differentially translated genes for patterns according to the different GO terms groups (Figure 4.5.9).

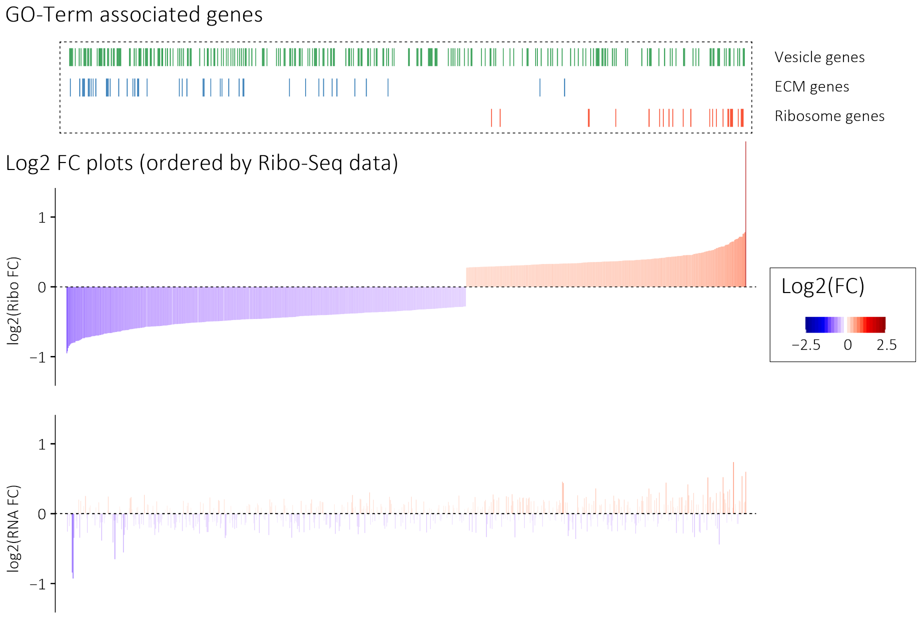


Figure 4.5.9: Log2 fold change plots of 716 genes that are purely regulated on the translational level and not located on chromosome 3 on RNA-Seq and Ribo-Seq data. The upper panel shows which genes are associated with the three GO term groups described above.

On the top of the figure, genes that have been linked to the three GO term groups "Vesicle", "Ribosome" and "ECM" are highlighted. Ribosome-associated genes seem to be mainly up-regulated in SHR.BN-(3S) congenic strains and ECM-associated genes are mainly down-regulated, whereas vesicle genes do not show a specific pattern of translational regulation. The two up-regulated ECM genes are Superoxide dismutase (*Sod1*) and the Fibroblast Growth Factor 1 (*Fgf1*) are both known to have cardio-protective effects upon over-expression [81, 82]. The 42 down-regulated ECM

genes include well-known heart disease examples such as *Lama2* and *Mmp2*.
All genes that have been associated with the ribosome show an up-regulation of translation, without a significant change on the RNA-Seq level.
ECM genes tend to be rather large, conversely, ribosomal genes, which include mainly 60S ribosomal subunit proteins (RPL) and 40S ribosomal subunit proteins (RPS), are small. In order to understand the putative relation between up-regulation of small proteins and down-regulation of large proteins, we assessed correlation of CDS-length and fold changes by applying the standardised major axis approach [83] (Figure 4.5.10).

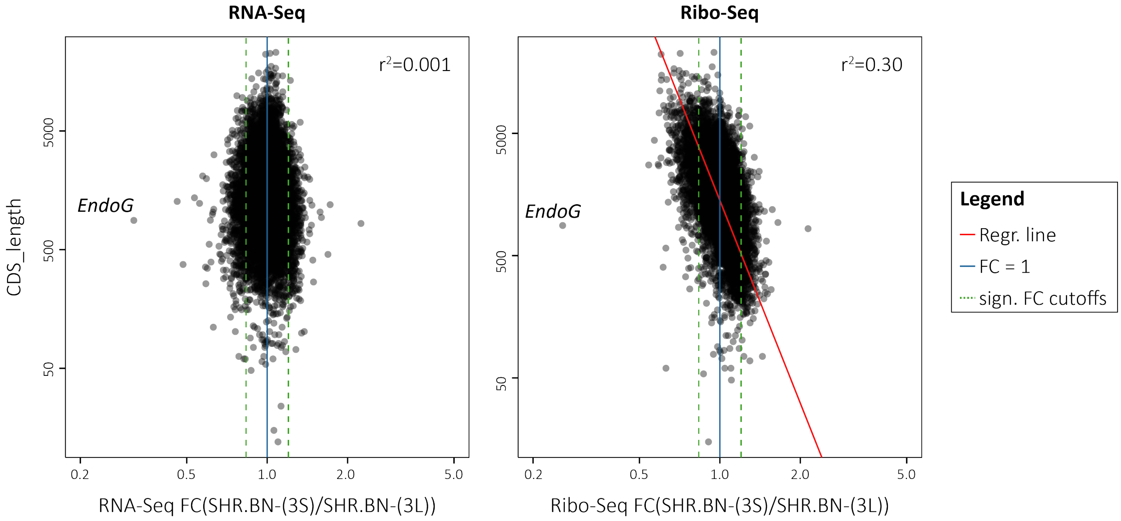


Figure 4.5.10: RNA-Seq and Ribo-Seq fold changes of congenic strains carrying the SHR genotype over those that carry the BN genotype plotted against the CDS-length of each gene. The r^2 was determined by standardised major axis estimations. The red line indicates the regression line on the Ribo-Seq plot. The blue line indicates the FC = 1 and the dashed green lines indicated the fold change cutoffs for significance.

No correlation has been observed between CDS-length and RNA-Seq fold changes, but on the Ribo-Seq level a slight correlation between fold change and the length of the coding sequence of a gene is seen. This suggests that the identified riboQTL is not only a regulator for the translation of extracellular matrix genes, but rather a global regulator for translational regulation resulting in increased translation of small proteins and reduced translation of large proteins upon disease.
In Figure 4.5.10 we see one dot that shows a strong down-regulation on the RNA-Seq and Ribo-Seq level, which corresponds to the gene *endonuclease G* (*EndoG*), a gene

that is located directly in the distant riboQTL locus. This gene was previously shown to be a regulator for blood pressure independent hypertrophy [47] and mitochondrial function. The SHR.BN-(3S) carries a frame shift mutation in *EndoG* that leads to low expression rates of *EndoG* in this congenic rat strain. The lack of *EndoG* in the animals carrying the SHR genotype, which affects mitochondrial function and therefore might lead to lower rates of GTP and ATP, could explain the lower translation rates of long proteins and up-regulation of small proteins that need less energy for translation.

In order to make sure that this effect is not just a general pattern that corresponds to the way how Ribo-seq compares to CDS-length, we additionally investigated different loci in the left ventricle as well as in liver in the rat RI panel. We were not able to see a length specific regulation of translation except of this locus on chromosome 3 (Figure 4.5.11).

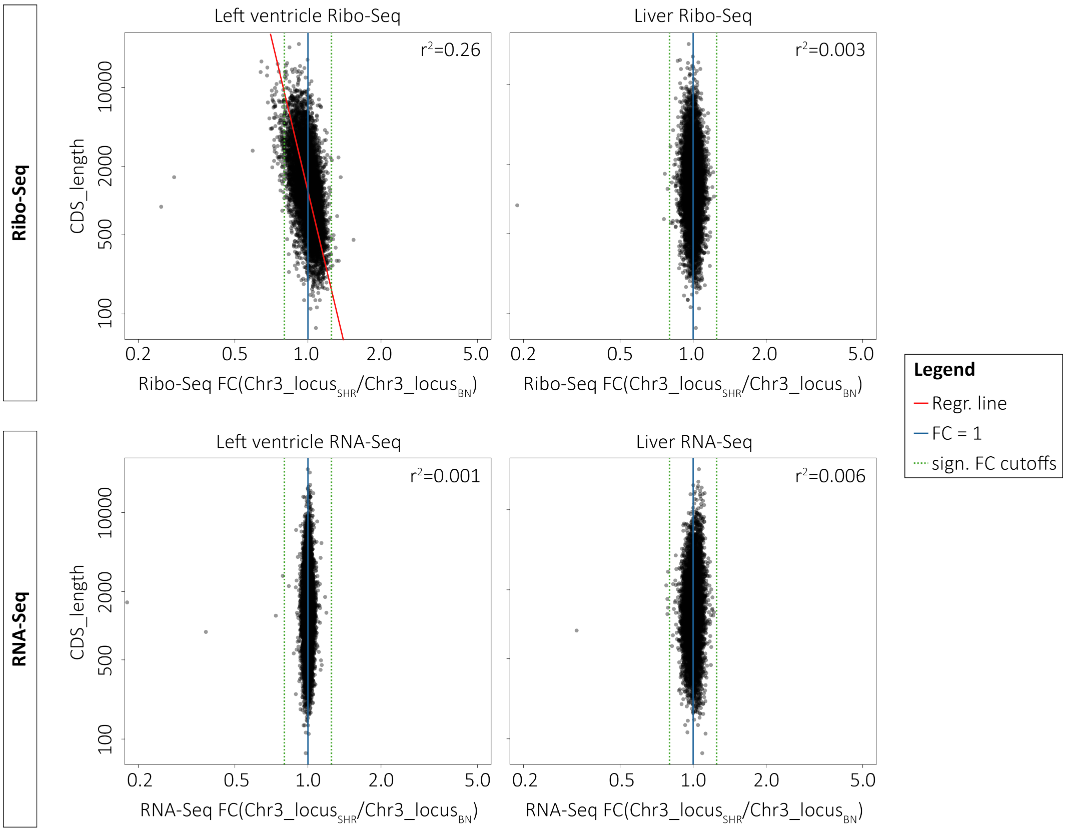


Figure 4.5.11: RNA-Seq and Ribo-Seq fold changes of the RI lines carrying the SHR genotype over the ones carrying the BN genotype at the distant cluster on chromosome 3 plotted against the CDS-length of each gene for left ventricle (left) and liver (right). The R2 was determined by standardised major axis estimations. The blue line indicates the FC = 1 and the dashed green lines indicated the fold change cutoffs for significance.

The fold changes of the RI lines carrying the SHR genotype over the ones carrying the BN genotype at the distant cluster on chromosome 3 are plotted against the CDS-length of the corresponding genes. In left-ventricular Ribo-Seq data we observe a slight correlation between CDS-length and FC, whereas in the left ventricle RNA-Seq as well as in the RNA-Seq and Ribo-Seq of the hepatic tissue, we do not see any correlation. This suggests a very specific response on the translation levels of left-ventricular tissue.

4.5.4 Studying the impact of a EndoG knock-out in a mouse model

In order to test the hypothesis that *EndoG* is responsible for the translational phenotype observed in the heart, RNA-Seq and Ribo-Seq data have been generated from the *EndoG* knock-out mice and all differentially expressed genes in C57BL/6 wild type versus C57BL/6 *EndoG-/-* mice have been derived.
In total, 38 genes are differentially translated and 1,133 genes differentially transcribed. None of the 17 previously identified genes, described in chapter 4.5, that have been linked to the chromosome 3 locus in the RI panel and the congenics, were found to be differentially expressed.
To check whether the effect sizes in mice are smaller and whether a translational down-regulation of example ECM genes is observed, the gene expression of those genes was assessed. In general, no decrease in C57BL/6 *EndoG-/-* mice on the translational level was detected (Figure 4.5.12).

In the congenics, as well as in the RI panel (discussed in chapter 5.3) we observe a correlation between CDS-length and Ribo-Seq fold change, which suggests general decrease of translation in long proteins and an increase in shorter proteins. Even though the ECM genes do not show significant changes on the translational level, the relationship between CDS-length and the RNA-Seq and Ribo-Seq fold change was tested for correlation. No correlation between Ribo-Seq fold change and the length of the transcript was observed (Figure 4.5.13).
The results based on the RNA-Seq and Ribo-Seq data of the *EndoG* knock-out mouse seem to contradict our hypothesis of *EndoG* playing a role in regulating the translation of ECM genes. But the *EndoG* knock-out mouse might simply not be suitable for picking up global translation changes as a disease response, because the mice were neither stressed nor sick and *EndoG* might affect translational regulation only as a

response to stress. This hypothesis will be further discussed in chapter 5.4

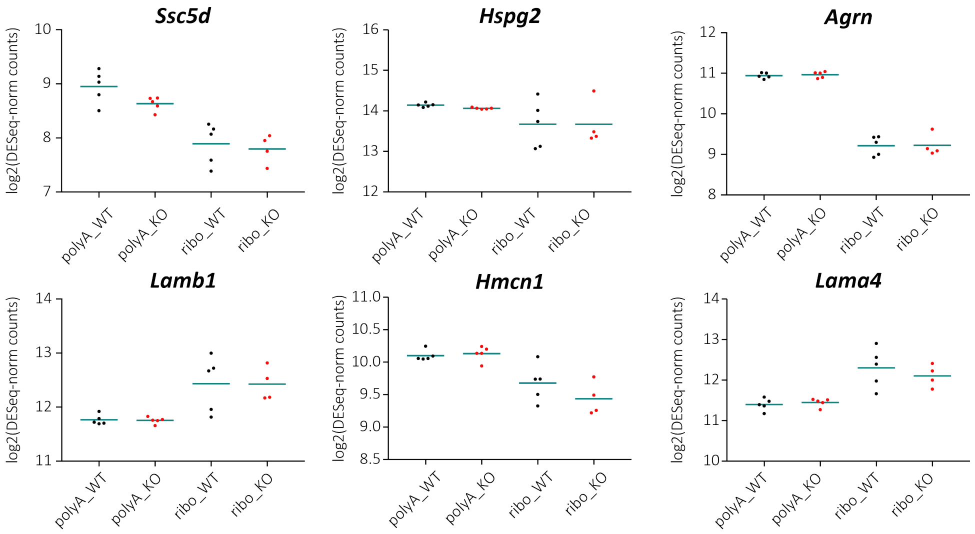


Figure 4.5.12: Expression plots of 6 ECM genes in the C57BL/6 wild type versus C57BL/6 *EndoG*-/- mice. The RNA-Seq data is shown on the left and Ribo-Seq data on the right in log2 transformed normalised counts. None show a significant shift in wild type vs. knock-out.

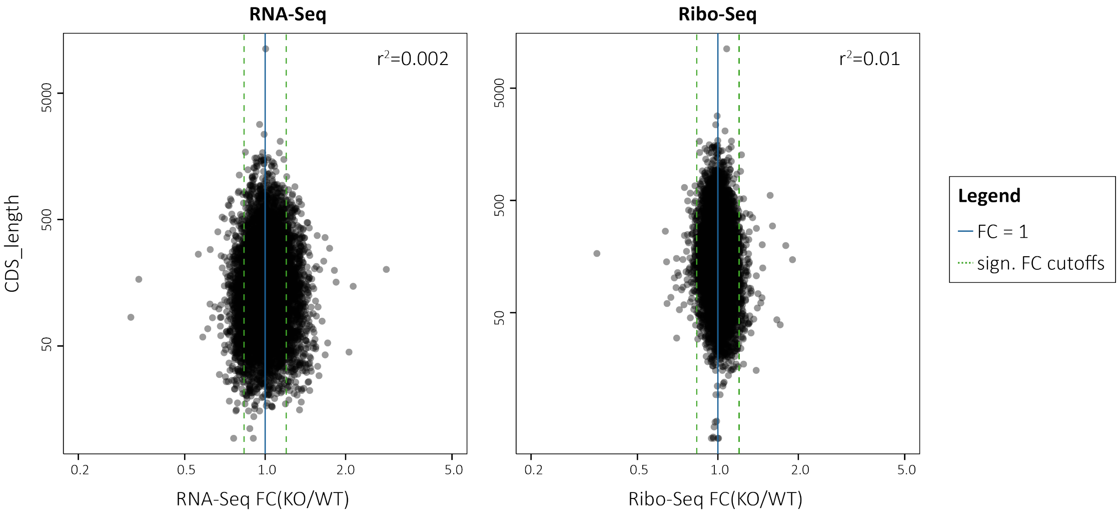


Figure 4.5.13: RNA-Seq and Ribo-Seq fold changes of C57BL/6 wild type versus C57BL/6 *EndoG*-/- mice plotted against the CDS-length of each gene. The r^2 was determined by standardised major axis estimations. The blue line indicates the FC = 1 and the dashed green lines indicated the fold change cutoffs for significance.

5 Discussion

5.1 Identification of prominent cardiac and hepatic genes regulated by genetic variants

In this thesis I presented a QTL mapping approach on two layers of gene expression, namely transcription and translation, in the rat RI panel that resulted in hundreds of local associations in left ventricle and liver on both levels. Additionally, we also identified a number of genes of which the gene expression is altered by genetic variants that act over longer distances. In 2014, Rintisch et al. [35] carried out a similar eQTL mapping study in left-ventricular and liver tissue of the rat RI panel. Despite differences in experimental design, such as the use of single-end, unstranded RNA-Seq data and a lower resolution (20k SNP) genotype map in the Rintisch et al study, eQTL mapping results correspond for approximately 50% in left ventricle (148 out of 301 genes) and 40% in liver (83 genes out of 209), with the most strongly regulated genes identified independently by both approaches. Examples are the previously reported genetically regulated cardiac and hepatic genes *Cyp17a1* [84] and *Endog* [47], which, in addition to a local eQTL, also show a forwarded expression change on the translational level in our data. These results reassure us that our data and QTL mapping strategy are solid and capable of producing reproducible findings, which indicates we can confidently apply the same strategy for studying genetic effects on translational regulation.

5.2 Tissue-specific regulation in left ventricle and liver

All tissues in an organism share common regulation and processes, but they also differ in their gene expression patterns, which suggests that tissue specificity can be derived by different regulatory programs across tissues ([85]). Sonawane et al. observed that regulatory genes, such as transcription factors, are more likely to show similar expression across tissues whereas their target genes tend to be more tissue specific.

If we apply this hypothesis to translational regulation, it would mean that the gene expression of translational regulators like RBPs is more often shared across tissues, compared to their target genes. We do not observe a significant enrichment of transcription factors among shared eQTL or RBPs in shared riboQTLs in our data. Nevertheless, we observe 13 RBPs among the 85 genes that share translational regulation across tissues that can be attributed to local genetic variants (Fisher's exact p-value = 0.51). This suggests that translational regulation of RBPs might not be genetically regulated to a high extent or that one would need to include more tissues in order to identify common expression of regulatory nodes, such as transcription factors and RBPs.

We make two additional observations about tissue-specificity according to our data that might be transferable to other species and tissues. We see that shared regulation across tissues observed on the translation basis is mostly already initiated on the transcriptional level, which means that expression changes that are commonly regulated evolve on the transcriptional layer rather than on the translational layer. On the other hand we see that especially distantly regulated genes are often only detected in one tissue, which could be explained by a smaller effect size in one of the tissue, that make it more difficult to detect the distant QTL in both tissues.

Even by using MT:Hess, which is meant to gain power for the detection of shared regulation across tissues, we do not identify common regulation over longer distances. In a previous study on eQTL mapping in the rat RI panel across seven tissues, 13 distant eQTL clusters that are shared across at least two out of seven tissue have been reported [39]. This low number indicates that even by using more data, it seems as if there is only limited common regulation across longer distances.

5.3 Length-dependent translational regulation

Ribosome profiling is still a relatively new technique that has not been fully assessed in terms of technical biases caused by the protocol or during processing of the data. We were able to show the reproducibility of the findings by replicating the initial identification of 17 ECM genes regulated by the chromosome 3 locus in an independent Ribo-Seq experiment in two rat congenic strains that only differ genomically by this locus. Also, we do not observe the down-regulation of long and up-regulation of short proteins in liver, which indicates a very specific response in the left ventricle. We checked different loci in the left ventricle as well as in liver and we were not able to see a length specific regulation of translation except of this locus on chromosome 3 (Figure 4.5.11).

In left-ventricular Ribo-Seq data we observed a slight correlation between CDS-length and FC, whereas in the left ventricle RNA-Seq as well as in the RNA-Seq and Ribo-Seq of the hepatic tissue, we do not see any correlation. This suggests a very specific response on the translation levels of left-ventricular tissue.

It is possible that the up-regulation of small proteins is caused by a normalization bias in the data, which results from the down-regulation of large proteins, i.e. fewer reads that map to large genes in the total sequencing library. As a consequence, these reads are equally distributed elsewhere in the genome. In the end this small increase of reads might have an effect on all other genes that is most prominent on short genes, which is observed as an up-regulation of these genes specifically.

5.4 Identification of a potential regulator of the translational landscape in the heart

We identified a genomic locus on chromosome 3 of 7 Mb size, which harbours 65 genes expressed in the rat heart, that might regulate a number of other genes purely on the translational level. Both the riboQTL and teQTL mapping on the rat RI panel and the differential translation analysis in the rat congenic strains suggested a more general effect on translational regulation. The findings from the congenic strains indicate that instead of the identification of a master regulator for the translation of extracellular matrix proteins only, it seems as if we identified a genomic locus that has an effect on the translation levels of genes in general. So it could be that we are rather searching for a regulator of another pathway that has the ability to change the translational landscape as a secondary effect.

The protein Endonuclease G (EndoG) is one among the 65 expressed genes in the riboQTL locus and has a high potential for being involved in the translational regulation in the heart. EndoG is an enzyme localised in the mitochondrial intermembrane space bound to Hsp70 (Heat shock protein 70) and CHIP (E3 ligase C terminus of Hsc-70 interacting protein). Under oxidative stress EndoG gets released and translocated to the nucleus and initiates DNA and proteasomal degradation. EndoG was previously described to play a role in ageing, cancer and neuronal diseases [86]. Altered expression levels of this gene might increase or decrease DNA degradation [86]. EndoG is an important enzyme in the maintenance of normal tissue homeostasis, which was shown by its role in the mitochondrial death pathway and the cardiac myocyte apoptosis pathway [87]. In the rat RI panel, the locus on chromosome 3 was previously linked to hypertension-independent hypertrophy and the authors were able to identify EndoG as the causal gene by testing the same congenic rat strains and knock-out mice as described in this thesis [47]. The question is now whether the effect that we observe on the translational landscape is also driven by EndoG itself, by the phenotype that is caused by the absence of EndoG in the animals carrying the SHR genotype, or whether it is completely unrelated to EndoG.

The first hypothesis is that the absence of EndoG leads to a global change of translational regulation. EndoG is mostly located in mitochondria, which are known for their function in producing and providing energy of the cell. If EndoG is not present it can no longer fulfil its role in maintaining the normal tissue homeostasis. This means that the energy metabolism of the cell might get disturbed and this might lead to ER-stress as a consequence of lower energy capacity than demands. A global change in the translational program of a cell that prefers to produce small proteins that need less energy for proper folding is possible. Additionally, the chances that small proteins do not properly fold are lower and this helps to avoid an accumulation of unfolded or misfolded proteins. An additional accumulation of unfolded proteins would again result in ER-stress. Considering EndoG as the causal gene for the translational change would suggest that the absence of EndoG leads to an accumulation of misfolded proteins, as they can no longer be cleaved and degraded by EndoG. The unfolded protein response (UPR) would get activated, which leads to a transcriptional and translational reprogramming of the whole cell in order to ensure gene expression of specific proteins that help to maintain the homeostasis. There are two main pathways that regulate the unfolded protein response, which are PERK/eIF2α/ATF4 signalling and mTOR signalling. The mammalian target of rapamycin, (mTOR) is a kinase that acts as a main activator for cell growth and proliferation. But it also plays a big role in translational regulation as a response to ER-stress. MTOR signalling gets activated by growth factor signalling and by the sufficient supply of nutrients (e.g. amino acids and glucose) and energy, which leads to an up-regulation in translation proliferation and ribosome production. Factors such as insufficient energy supplies, ER-stress and DNA damage lead to an inhibition of mTOR signalling that down-regulate processes such as translation and ribosome production. In the SHR rat, where the lack of EndoG that is important to maintain the homeostasis of protein production and supply of energy and nutrients by degrading unfolded proteins, mTOR signalling is very likely in the inhibited state. This would lead to a general down-regulation of translation and ribosome biogenesis [88], whereas in our data we observe a specific down-regulation of large proteins and an up-regulation of small proteins.

If the observed effect is caused by UPR, it might be caused by the transcriptional and translational control network of PERK/eIF2α/ATF4 signalling rather than mTOR signalling. This pathway is known to repress global translation and at the same time translational up-regulation of specific mRNAs that are involved in UPR gene expression regulation. Upon UPR, PERK phosphorylates the translation initiation factor eIF2α, which leads to reduced amount of GTP bound eIF2. Overall, this decreases the levels of translation initiation, which can be determined, for example, by polysome profiling as a reduction of large polysome chains coinciding with increased levels of translating monosomes [89]. This response to ER-stress might explain why we observe an up-regulation of smaller proteins compared to a down-regulation of larger ones. The effect on small proteins changing the translation program from polysomal translation to monosomal translation might simply not have the same impact on translational efficiency as on large proteins, where the number of ribosomes that could translate an mRNA in parallel is simply much higher due to space capacity (Figure 5.4.1).

Another possibility is that the translational response that we are observing is a secondary effect from the blood pressure independent increase of the heart weight that was described for this locus by McDermott et al., in both the RI panel as well as the congenic rat strains. Elevated heart weight that is observed without hypertension is often observed accompanying diseases such as obesity, type 2 diabetes and was shown to be associated with mitochondrial dysfunction and lipotoxicity [47].

The main changes in the heart during cardiac hypertrophy are an abnormal thickening of the heart muscle that is caused by increased size of cardiomyocytes, but also by the remodelling of other cardiac muscle components such as the ECM (increased fibrosis). It could very well be that a very early response to the hypertrophy is accompanied by ER-stress. The initiation of the remodelling of the ECM leads to a switch in protein production of specific ECM components that cause as an intermediate state an imbalance between the demand and capacity of protein load. One has to keep in mind that these animals are only 6 weeks old and do not show extreme phenotypes yet, so an early response to the cardiac growth would make sense. As a result of the

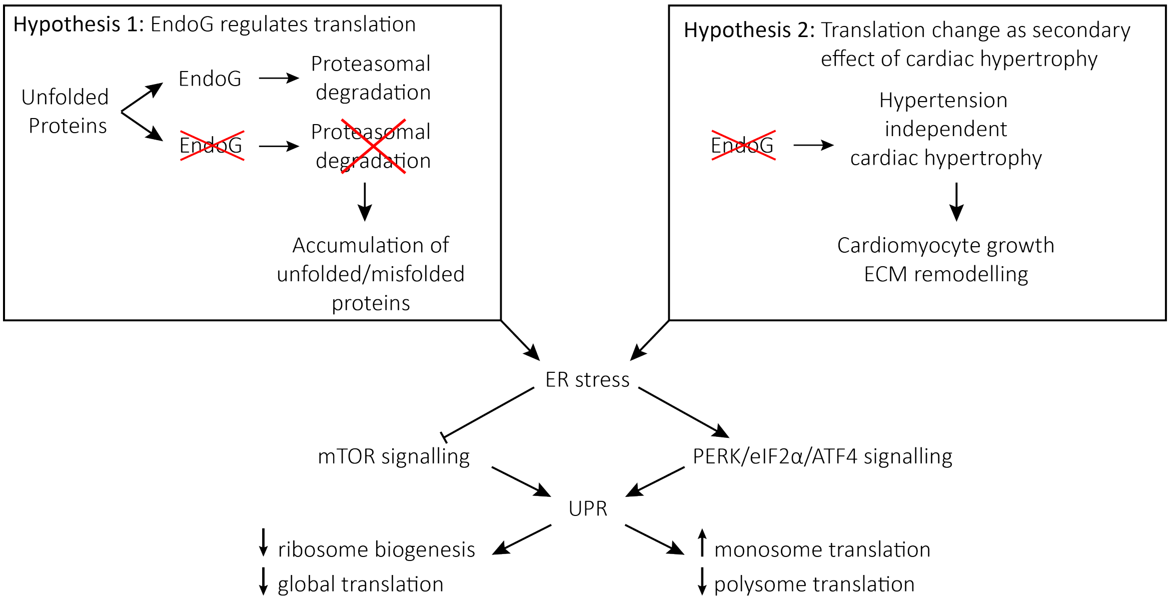

Hypothesis 1: EndoG regulates translation
Unfolded Proteins
EndoG
Proteasomal degradation
~~EndoG~~
~~Proteasomal degradation~~
Accumulation of unfolded/misfolded proteins
Hypothesis 2: Translation change as secondary effect of cardiac hypertrophy
~~EndoG~~
Hypertension independent cardiac hypertrophy
Cardiomyocyte growth
ECM remodelling
ER stress
mTOR signalling
PERK/eIF2α/ATF4 signalling
UPR
↓ ribosome biogenesis
↓ global translation
↑ monosome translation
↓ polysome translation

Figure 5.4.1: Schematic illustrating the two hypotheses for how EndoG might play a role in the global change in translation observed in the SHR genotype rats in the chromosome 3 QTL region. The first hypothesis describes how EndoG might be regulating the translational program in the SHR genotype due to a lack of the protein. The second hypothesis is based on the assumption that the translational change observed in the animals carrying the SHR genotype is a secondary effect due to the cardiac hypertrophy that is induced by the lack of EndoG.

imbalance caused by the ECM remodelling, the protein synthesis is reduced and the expression of specific proteins is up-regulated in order to increase the capacity for the protein income. This would fit quite well with the increased translation levels of ribosomal proteins and down-regulation of larger proteins. This would suggest that the hypertrophy that is caused by EndoG induces ER-stress by the high protein demand during remodelling of the ECM. And the change on the translational landscape that we observe is an early response on the ER-stress in order to treat the imbalance in protein production.

The fact that we do not see a translational response in the EndoG knock-out mice could either be explained by a rat specific translational regulation that is not conserved across rodents or it could also mean that the translational change can only be observed if there is an additional stimuli such as cardiac hypertrophy or cardiac stress at the same time. This would strengthen the hypothesis that the translational program we are seeing is a secondary response to the disease. On the other hand it could also mean that animals carrying the EndoG mutation cannot cope with the extra stress caused by the disease. In order to be able to prove this we would need to induce either ER-stress by known agents or we would need to induce cardiac hypertrophy for example by a transverse aortic constriction (TAC) in these mice. A TAC mouse is an experimental model for blood-pressure overload-induced cardiac hypertrophy. Combined with the EndoG knock-out this should show a similar effect as observed in the rats if this hypothesis is true (Figure 5.4.1).

A third possibility could be that the translational response is an effect that is completely independent from EndoG, but caused by another gene located in the same locus. After a thorough literature search on all expressed genes that are located in the chromosome 3 locus, we did not find an obvious candidate gene that could act as global translational regulator, but there are a number of interesting genes in there such as RNA-binding proteins, miRNAs and other ECM components. The difficulty is to identify the causal gene without any prior knowledge on possible relations. In general, the QTL regions tend to be rather large, not only in the rat RI panel but also in other studies; it is challenging to identify the causal gene that affects gene expression or whole body phenotypes. QTL regions often include many genes and SNPs, in the case of the chromosome 3 locus, there are 65 expressed genes and over 12.000 SNPs. If there were genes in the same locus with a local QTL of the same SDP, they would make it a good candidate for further assessment. However, many genes can be affected in a non-quantitative way, which means that they do not have to be differentially expressed to have a quantitative impact on target genes. For example, they may carry a missense mutation that reduces affinity of a protein's RNA binding site for target genes, or a mutation in any position that only affects gene function and not gene quantity.

5.5 Summarising remarks and future steps

This thesis summarises the bioinformatic analysis of a project that aims to understand the role of genetic variants in gene expression regulation in heart and liver in a heart disease context. The QTL mapping analysis resulted in a number of good candidates for functional follow-up in order to translate the finding to human diseases.

The most promising finding is the riboQTL locus on chromosome 3 that seems to affect the translational regulation of all other genes, specifically in the heart. In the discussion, I discussed possible mechanisms of how the global translational changes could arise. The next steps for this project will include multiple experimental validations in order to identify the causal gene and the mechanism behind the global modulation of gene expression.

We already started some experiments to validate the translational effect that was observed with Ribo-Seq. We performed q-PCR to show that there is no effect on the transcriptional level and western blot experiments to show the effect on the translational level in the two rat congenic strains. Instead of conventional western blotting, we also performed western blots using the automated western blotting machine WES, which is supposed to make the results more reproducible and allow the quantitative assessment of the protein concentration. Thus far, we have not been able to validate the difference in gene translation on the protein level using either of these two techniques. This could be explained by insufficient sensitivity to detect the small changes of translation rates (Ribo-Seq fold changes $<$ 2-fold).

To follow-up whether we see an effect of the global translational regulation on the protein level using a different technique than western blotting, we performed a mass spectrometry experiment on the two congenic strains (on a QExactivePlus machine with a long gradient (6 hours) on monolithic column). In general, it is rather difficult to perform mass spectrometry on heart tissue, because of the high complexity and wide dynamic range of protein abundance. Heart tissue is known to have high-abundance proteins that make it harder to detect and measure proteins with lower abundance. This should not play a big role for looking at ECM proteins as those have a decent abundance, but smaller less abundant proteins such as ribosomal subunits might be missed. We were not able to validate a global change on protein levels using mass spectrometry. This suggests that the changes on the translational level are either not forwarded to the protein level, because the imbalance is compensated post-translationally or mass spectrometry does not have the resolution to make small changes like this visible.

In a next step, we are planning to perform polysome profiling on the two congenic rat strains and generate RNA-Seq data of the different fractions. This would allow us to prove our hypothesis that translational regulation is modulated due to lower rates of translation initiation that results in higher levels of translation by monosomes rather than by polysomes. The pilot experiment for the polysome profiling was already performed and we could observe a clear difference of the polysome profile between the two congenic strains. We will next perform RNA-Seq on the fractions of the polysome profiling in order to understand whether longer proteins and shorter proteins end up in different fractions or if polysomal translation is very specific to proteins that are involved in the untranslated protein response.

After validating that the global change on translation can be seen with different approaches than Ribo-Seq, we want to identify the causal gene and genetic variant that drives the gene expression change. In order to figure out whether EndoG might be the causal gene, we could for example perform Ribo-Seq on animals carrying an EndoG rescue in the SHR.Ola background. Animals like this have been previously generated, but the EndoG rescue only works with low efficiency, which means that the levels of EndoG are still not comparable to wild type BN.Lx expression of the gene. As I mentioned before, we also tested an EndoG knock-out mouse model, that did not show any difference on the translational level. If the hypothesis that UPR only occurs as a secondary effect to the cardiac hypertrophy, it could be that we need to challenge the mice with agents that lead to ER-stress or perform a TAC experiment in order to induce cardiac hypertrophy.

If the experiments suggest that not EndoG, but one of the other 64 genes that are located in the chromosome 3 locus are causing the translational regulation it will be more difficult to pinpoint the causal gene. Fine-mapping of the QTL region by generating sub-strains could enable us to identify the global translational regulator.

Besides following-up on the translational trans cluster, this thesis opens up many opportunities for new projects, e.g. the elucidation of the exact mechanisms of local riboQTL regulation by for instance uORFs and the effects of different splice isoforms on translational efficiency.

Bibliography

[1] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, 296(5568):752–5, apr 2002.

[2] Tamara T. Koopmann, Michiel E. Adriaens, Perry D. Moerland, Roos F. Marsman, Margriet L. Westerveld, Sean Lal, Taifang Zhang, Christine Q. Simmons, Istvan Baczko, Cristobal dos Remedios, et al. Genome-Wide Identification of Expression Quantitative Trait Loci (eQTLs) in Human Heart. *PLoS ONE*, 9(5):e97380, may 2014.

[3] Matthias Heinig, Michiel E. Adriaens, Sebastian Schafer, Hanneke W. M. van Deutekom, Elisabeth M. Lodder, James S. Ware, Valentin Schneider, Leanne E. Felkin, Esther E. Creemers, Benjamin Meder, et al. Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy. *Genome Biology*, 18(1):170, dec 2017.

[4] Marco Antonio Valencia-Sanchez, Jidong Liu, Gregory J Hannon, and Roy Parker. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development*, 20(5):515–24, mar 2006.

[5] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry*, 79(1):351–379, jun 2010.

[6] D R Morris and A P Geballe. Upstream open reading frames as regulators of mRNA translation. *Molecular and cellular biology*, 20(23):8635–42, dec 2000.

[7] Cristina Barbosa, Isabel Peixeiro, and Luísa Romão. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genetics*, 9(8):e1003529, aug 2013.

[8] Guo-Liang Chew, Andrea Pauli, and Alexander F. Schier. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications*, 7:11663, may 2016.

[9] Sara K Young and Ronald C Wek. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *The Journal of biological chemistry*, 291(33):16927–35, aug 2016.

[10] Ze-Lin Wang, Bin Li, Yu-Xia Luo, Qiao Lin, Shu-Rong Liu, Xiao-Qin Zhang, Hui Zhou, Jian-Hua Yang, and Liang-Hu Qu. Comprehensive Genomic Characterization of RNA-Binding Proteins across Human Cancers. *Cell Reports*, 22(1):286–298, jan 2018.

[11] Jack D. Keene. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8(7):533–543, jul 2007.

[12] Cédric Gobet and Felix Naef. Ribosome profiling and dynamic regulation of translation in mammals. *Current Opinion in Genetics & Development*, 43:120–127, apr 2017.

[13] Nicholas T. Ingolia, Liana F. Lareau, and Jonathan S. Weissman. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*, 147(4):789–802, nov 2011.

[14] Gerben Menschaert, Wim Van Criekinge, Tineke Notelaers, Alexander Koch, Jeroen Crappé, Kris Gevaert, and Petra Van Damme. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics : MCP*, 12(7):1780–90, jul 2013.

[15] Jeroen Crappé, Elvis Ndah, Alexander Koch, Sandra Steyaert, Daria Gawron, Sarah De Keulenaer, Ellen De Meester, Tim De Meyer, Wim Van Criekinge, Petra Van Damme, et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research*, 43(5):e29–e29, mar 2015.

[16] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–35, apr 2014.

[17] James Dominic Mills, Yoshihiro Kawahara, and Michael Janitz. Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling. *Current genomics*, 14(3):173–81, may 2013.

[18] Elaine R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, sep 2008.

[19] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, jan 2009.

[20] Michael L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, jan 2010.

[21] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, jul 2011.

[22] Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924):218–23, apr 2009.

[23] Timothy J Aitman, John K Critser, Edwin Cuppen, Anna Dominiczak, Xose M Fernandez-Suarez, Jonathan Flint, Dominique Gauguier, Aron M Geurts, Michael Gould, Peter C Harris, et al. Progress and prospects in rat genetics: a community view. *Nature Genetics*, 40(5):516–522, may 2008.

[24] K Okamoto and K Aoki. Development of a strain of spontaneously hypertensive rats. *Japanese circulation journal*, 27:282–93, mar 1963.

[25] S Doggrell and Lindsay Brown. Rat models of hypertension, cardiac hypertrophy and failure. *Cardiovascular Research*, 39(1):89–105, jul 1998.

[26] Jörn Lange, Thomas Barz, Axel Ekkernkamp, Barbara Wilke, Ingrid Klöting, and Niels Follak. Phenotypic and Gene Expression Differences between DA, BN and WOKW Rats. *PLoS ONE*, 7(6):e38981, jun 2012.

[27] Roel Hermsen, Joep de Ligt, Wim Spee, Francis Blokzijl, Sebastian Schäfer, Eleonora Adami, Sander Boymans, Stephen Flink, Ruben van Boxtel, Robin H van der Weide, et al. Genomic landscape of rat strain and substrain variation. *BMC Genomics*, 16(1):357, dec 2015.

[28] Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, apr 2004.

[29] Santosh S Atanur, Inanç Birol, Victor Guryev, Martin Hirst, Oliver Hummel, Catherine Morrissey, Jacques Behmoaras, Xose M Fernandez-Suarez, Michelle D Johnson, William M McLaren, et al. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome research*, 20(6):791–803, jun 2010.

[30] Michal Pravenec and Theodore W Kurtz. Recent advances in genetics of the spontaneously hypertensive rat. *Current hypertension reports*, 12(1):5–9, feb 2010.

[31] Santosh S. Atanur, Ana Garcia Diaz, Klio Maratou, Allison Sarkis, Maxime Rotival, Laurence Game, Michael R. Tschannen, Pamela J. Kaisaki, Georg W. Otto, Man Chun John Ma, et al. Genome Sequencing Reveals Loci under Artificial Selection that Underlie Disease Phenotypes in the Laboratory Rat. *Cell*, 154(3):691–703, aug 2013.

[32] Sebastian Schafer, Eleonora Adami, Matthias Heinig, Katharina E. Costa Rodrigues, Franziska Kreuchwig, Jan Silhavy, Sebastiaan van Heesch, Deimante Simaite, Nikolaus Rajewsky, Edwin Cuppen, et al. Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nature Communications*, 6(1):7200, dec 2015.

[33] Morton P. Printz, Martin Jirout, Rebecca Jaworski, Adamu Alemayehu, and Vladimir Kren. Invited Review: HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. *Journal of Applied Physiology*, 94(6), 2003.

[34] Norbert Hubner, Caroline A Wallace, Heike Zimdahl, Enrico Petretto, Herbert Schulz, Fiona Maciver, Michael Mueller, Oliver Hummel, Jan Monti, Vaclav Zidek, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37(3):243–253, mar 2005.

[35] Carola Rintisch, Matthias Heinig, Anja Bauerfeind, Sebastian Schafer, Christin Mieth, Giannino Patone, Oliver Hummel, Wei Chen, Stuart Cook, Edwin Cuppen, et al. Natural variation of histone modification and its impact on gene expression in the rat genome. *Genome research*, 24(6):942–53, jun 2014.

[36] Victor Guryev, Kathrin Saar, Tatjana Adamovic, Mark Verheul, Sebastiaan A A C van Heesch, Stuart Cook, Michal Pravenec, Timothy Aitman, Howard Jacob, James D Shull, et al. Distribution and functional impact of DNA copy number variation in the rat. *Nature Genetics*, 40(5):538–545, may 2008.

[37] Boris Tabakoff, Laura Saba, Morton Printz, Pam Flodman, Colin Hodgkinson, David Goldman, George Koob, Heather N Richardson, Katerina Kechris, Richard L Bell, et al. Genetical genomic determinants of alcohol consumption in rats and humans. *BMC Biology*, 7(1):70, oct 2009.

[38] M.L. Jirout, R.S. Friese, N.R. Mahapatra, M. Mahata, L. Taupenot, S.K. Mahata, V. Kren, V. Zídek, J. Fischer, H. Maatz, et al. Genetic regulation of catecholamine synthesis, storage and secretion in the spontaneously hypertensive rat. *Human Molecular Genetics*, 19(13):2567–2580, jul 2010.

[39] Sarah R. Langley, Leonardo Bottolo, Jaroslav Kunes, Josef Zicha, Vaclav Zidek, Norbert Hubner, Stuart A. Cook, Michal Pravenec, Timothy J. Aitman, and Enrico Petretto. Systems-level approaches reveal conservation of trans-regulated genes in the rat and genetic determinants of blood pressure in humans. *Cardiovascular Research*, 97(4):653–665, mar 2013.

[40] Michelle D. Johnson, Michael Mueller, Martyna Adamowicz-Brice, Melissa J. Collins, Pascal Gellert, Klio Maratou, Prashant K. Srivastava, Maxime Rotival, Shahena Butt, Laurence Game, et al. Genetic Analysis of the Cardiac Methylome at Single Nucleotide Resolution in a Model of Human Cardiovascular Disease. *PLoS Genetics*, 10(12):e1004813, dec 2014.

[41] Laura M. Saba, Stephen C. Flink, Lauren A. Vanderlinden, Yedy Israel, Lutske Tampier, Giancarlo Colombo, Kalervo Kiianmaa, Richard L. Bell, Morton P. Printz, Pamela Flodman, et al. The sequenced rat brain transcriptome - its use in identifying networks predisposing alcohol consumption. *FEBS Journal*, 282(18):3556–3578, sep 2015.

[42] O. Kuda, M. Brezinova, J. Silhavy, V. Landa, V. Zidek, C. Dodia, F. Kreuchwig, M. Vrbacky, L. Balas, T. Durand, et al. Nrf2-Mediated antioxidant defense and peroxiredoxin 6 are linked to biosynthesis of palmitic acid ester of 9-Hydroxystearic acid. *Diabetes*, 67(6), 2018.

[43] C Damerval, A Maurice, J M Josse, and D de Vienne. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics*, 137(1):289–301, may 1994.

[44] Sophia Doll, Martina Dreßen, Philipp E. Geyer, Daniel N. Itzhak, Christian Braun, Stefanie A. Doppler, Florian Meier, Marcus-Andre Deutsch, Harald Lahm, Rüdiger Lange, et al. Region and cell-type resolved quantitative proteomic map of the human heart. *Nature Communications*, 8(1):1469, dec 2017.

[45] Dominique Gauguier. Application of quantitative metabolomics in systems genetics in rodent models of complex phenotypes. *Archives of Biochemistry and Biophysics*, 589:158–167, jan 2016.

[46] Amelie Baud, Victor Guryev, Oliver Hummel, Martina Johannesson, Amelie Baud, Victor Guryev, Oliver Hummel, Martina Johannesson, Roel Hermsen, Pernilla Stridh, et al. Genomes and phenomes of a population of outbred rats and its progenitors. *Scientific Data*, 1, jul 2014.

[47] Chris McDermott-Roe, Junmei Ye, Rizwan Ahmed, Xi-Ming Sun, Anna Serafín, James Ware, Leonardo Bottolo, Phil Muckett, Xavier Cañas, Jisheng Zhang, et al. Endonuclease G is a novel determinant of cardiac hypertrophy and mitochondrial function. *Nature*, 478(7367):114–8, oct 2011.

[48] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, apr 2013.

[49] Sagar Chhangawala, Gabe Rudy, Christopher E Mason, and Jeffrey A Rosenfeld. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome biology*, 16(1):131, jun 2015.

[50] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, mar 2009.

[51] Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, et al. The Ensembl gene annotation system. *Database*, 2016:baw093, jun 2016.

[52] P. P. Chan and T. M. Lowe. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, 37(Database):D93–D97, jan 2009.

[53] Patricia P. Chan and Todd M. Lowe. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, 44(D1):D184–D189, jan 2016.

[54] S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, jan 2015.

[55] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014.

[56] Zhengtao Xiao, Qin Zou, Yu Liu, and Xuerui Yang. Genome-wide assessment of differential translations with ribosome profiling data. *Nature Communications*, 7:11194, apr 2016.

[57] Yi Zhong, Theofanis Karaletsos, Philipp Drewe, Vipin T. Sreedharan, David Kuo, Kamini Singh, Hans-Guido Wendel, and Gunnar Rätsch. RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics*, 33(1):139–141, jan 2017.

[58] Arthur Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Journal of Geophysical Research*, 3(1):13, mar 1898.

[59] D.J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.

[60] Lorenzo Calviello, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler. Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, 13(2):165–170, feb 2016.

[61] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.

[62] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, nov 2011.

[63] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, dec 2009.

[64] Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, 28(10):1353–8, may 2012.

[65] Y Hochberg and Y Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7):811–8, jul 1990.

[66] Alex Lewin, Habib Saadi, James E. Peters, Aida Moreno-Moral, James C. Lee, Kenneth G. C. Smith, Enrico Petretto, Leonardo Bottolo, and Sylvia Richardson. MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics*, 32(4):523–532, feb 2016.

[67] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, sep 2014.

[68] Thomas Manke, Helge G. Roider, and Martin Vingron. Statistical Modeling of Transcription Factor Binding Affinities Predicts Regulatory Interactions. *PLoS Computational Biology*, 4(3):e1000039, mar 2008.

[69] Morgane Thomas-Chollier, Andrew Hufton, Matthias Heinig, Sean O'Keeffe, Nassim El Masri, Helge G Roider, Thomas Manke, and Martin Vingron. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*, 6(12):1860–1869, nov 2011.

[70] Tim R Mercer, Shane Neph, Marcel E Dinger, Joanna Crawford, Martin A Smith, Anne-Marie J Shearwood, Eric Haugen, Cameron P Bracken, Oliver Rackham, John A Stamatoyannopoulos, et al. The human mitochondrial transcriptome. *Cell*, 146(4):645–58, aug 2011.

[71] Timothy Daley and Andrew D Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–7, apr 2013.

[72] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683, oct 2017.

[73] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, dec 2016.

[74] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, jul 2009.

[75] Kristian E Baker and Roy Parker. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Current Opinion in Cell Biology*, 16(3):293–299, jun 2004.

[76] Yao-Fu Chang, J Saadi Imam, and Miles F Wilkinson. The nonsense-mediated decay RNA surveillance pathway. *Annual review of biochemistry*, 76(1):51–74, jun 2007.

[77] Sebastian Schafer, Kui Miao, Craig C. Benson, Matthias Heinig, Stuart A. Cook, Norbert Hubner, Sebastian Schafer, Kui Miao, Craig C. Benson, Matthias Heinig, et al. Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). In *Current Protocols in Human Genetics*, pages 11.16.1–11.16.14. John Wiley & Sons, Inc., Hoboken, NJ, USA, oct 2015.

[78] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–86, jun 2008.

[79] Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. g:Profiler?a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, jul 2016.

[80] T. Yamazaki and T. Maruyama. Evidence for the Neutral Hypothesis of Protein Polymorphism. *Science*, 178(4056):56–58, oct 1972.

[81] P Wang, H Chen, H Qin, S Sankarapandi, M W Becher, P C Wong, and J L Zweier. Overexpression of human copper, zinc-superoxide dismutase (SOD1) prevents postischemic injury. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):4556–60, apr 1998.

[82] Meindert Palmen, Mat J.A.P. Daemen, Leon J. De Windt, Jodil Willems, Willem R.M. Dassen, Sylvia Heeneman, Rene Zimmermann, Marc Van Bilsen, and Pieter A. Doevendans. Fibroblast growth factor-1 improves cardiac functional recovery and enhances cell survival after ischemia and reperfusion: A fibroblast growth factor receptor, protein kinase c, and tyrosine kinase-dependent mechanism. *Journal of the American College of Cardiology*, 44(5):1113–1123, sep 2004.

[83] Warton D.I. Falster, D.S. and I.J. Wright. Smatr. 2006.

[84] Teck Yew Low, Sebastiaan van Heesch, Henk van den Toorn, Piero Giansanti, Alba Cristobal, Pim Toonen, Sebastian Schafer, Norbert Hübner, Bas van Breukelen, Shabaz Mohammed, et al. Quantitative and Qualitative Proteome Characteristics Extracted from In-Depth Integrated Genomics and Proteomics Analysis. *Cell Reports*, 5(5):1469–1478, dec 2013.

[85] Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding Tissue-Specific Gene Regulation. *Cell reports*, 21(4):1077–1088, oct 2017.

[86] Sabrina Büttner, Lukas Habernig, Filomena Broeskamp, Doris Ruli, F Nora Vögtle, Manolis Vlachos, Francesca Macchi, Victoria Küttner, Didac Carmona-Gutierrez, Tobias Eisenberg, et al. Endonuclease G mediates α-synuclein cytotoxicity during Parkinson's disease. *The EMBO journal*, 32(23):3041–54, nov 2013.

[87] M Chiong, Z V Wang, Z Pedrozo, D J Cao, R Troncoso, M Ibacache, A Criollo, A Nemchenko, J A Hill, and S Lavandero. Cardiomyocyte death: mechanisms and translational implications. *Cell Death & Disease*, 2(12):e244–e244, dec 2011.

[88] Jose Aramburu, M Carmen Ortells, Sonia Tejedor, Maria Buxadé, and Cristina López-Rodríguez. Transcriptional regulation of the stress response by mTOR. *Science signaling*, 7(332):re2, jul 2014.

[89] Bo-Jhih Guan, Dawid Krokowski, Mithu Majumder, Christine L Schmotzer, Scot R Kimball, William C Merrick, Antonis E Koromilas, and Maria Hatzoglou. Translational control during endoplasmic reticulum stress beyond phosphorylation of the translation initiation factor eIF2α. *The Journal of biological chemistry*, 289(18):12593–611, may 2014.

Abbreviations

ATP Adenosine-5'-triphosphate.

BN Brown Norway with polydactyly-luxate.

bp base pair.

CDS coding DNA sequence.

ECM extracellular matrix.

eQTL expression QTL (trait - RNA-Seq coverage).

FC fold change.

FDR false discovery rate.

GTP Guanosine-5'-triphosphate.

IRES internal ribosomal entry site.

kb kilo base pairs (1,000 bp).

lncRNA long non-coding RNA (including antisense genes, long intergenic non-coding genes and processed transcripts).

LV left ventricle.

Mb mega base pairs (1,000,000 bp).

miRNA micro RNA.

mORF main ORF.

mRNA messenger RNA.

mtRNA mitochondrial RNA.

NMD nonsense mediated decay.

nt nucleotide.

ORF open reading frame.

PFM position frequency matrix format.

pQTL protein QTL.

q-PCR quantitative polymerase chain reaction.

QTL quantitative trait loci.

RBP RNA binding protein.

RI recombinant inbred.

riboQTL ribosome profiling QTL (trait - Ribo-Seq coverage).

RIN RNA integrity number.

Rnor rattus norvegicus.

RPKM Reads per Kilobase of transcript, per Million mapped reads.

RPL large ribosomal subunit protein.

RPS small ribosomal subunit protein.

rRNA ribosomal RNA.

SBP systolic blood pressure.

SDP strain distribution pattern.

SHR Spontaneously hypertensive rat.

SNP single nucleotide polymorphism.

sORF short/small open reading frame.

TAC Transverse aortic constriction.

TE translational efficiency.

teQTL translational efficiency QTL (trait - residuals of the lm(Ribo RNA)).

TRAP Transcription factor affinity prediction.

tRNA transfer RNA.

TSS transcription start site.

uORF upstream open reading frame.

UPR Unfolded protein response.

UTR untranslated region.

For reasons of data protection, the curriculum vitae is not included in the online version.

For reasons of data protection, the curriculum vitae is not included in the online version.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Doktorarbeit selbstständig verfasst und die verwendeten Hilfsmittel vollständig angegeben habe. Alle Stellen, die aus anderen Werken in Wortlaut oder dem Sinn entsprechend übernommen wurden, habe ich mit Quellenangaben kenntlich gemacht.

Ich erkläre weiterhin, dass die vorliegende Arbeit noch nicht im Rahmen eines anderen Promotionsverfahrens eingereicht wurde.

Berlin, 28.11.2018 ____________________

Franziska Witte