Linking gene expression and orthology in mammals

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Julia Franziska Elisabeth Söllner
aus Augsburg

Tübingen
2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.06.2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatterin:	Apl. Prof. Dr. Kay Nieselt
2. Berichterstatter:	Prof. Dr. Stephan Ossowski

"I hope you'll make mistakes. If you're making mistakes, it means you're out there doing something."
- Neil Gaiman, Making Good Art

Abstract

The overall aim of biomedical research is to understand disease mechanisms and to provide a drug to eventually cure the disease. This challenging endeavour requires an early research phase that deals with identifying target genes or proteins playing an important role in the disease. At this stage one uses animal models mimicking human disease to determine differences between healthy and diseased animals. Once potential drug targets have been found, compounds are screened and promising compounds go into the preclinical phase where their efficacy and, most importantly, safety are assessed. Those having proven to be efficacious and safe proceed to toxicology where the maximum tolerable dosage is assessed in, mainly, non-rodent species.

According to the Bundesministerium für Ernährung und Landwirtschaft, more than 2 million animals were used for animal testing in German laboratories in 2017. The majority of these animals were mice and rats but also dogs, cats and monkeys are model organisms used for testing. While it is commonly accepted that other mammalian species resemble human biology to a great extent, one has to bear in mind that there are species-specific differences.

One of the aims of this thesis was to investigate how similar widely used model species are to human and to each other on a molecular level. For this purpose we assessed the relationship between protein sequence identity and gene expression correlation with an emphasis on mouse and rat. We found that the majority of genes are highly similar, both on sequence and gene expression level. There were, however, cases with low sequence identity but high expression correlation. These cases were investigated in greater detail and the hypothesis that sequences annotated in widely used databases like Ensembl, UniProt, or RefSeq, may contain errors or are incomplete, was confirmed.

Therefore, we investigated whether sequence information from related species can be used to derive a target's sequence in a species with poor annotation. The a&o-tool was developed to exploit sequence similarity between related species and short-read RNA-Seq data to refine or validate target sequences. Since long-read RNA-Seq data would greatly improve the results as entire transcripts are sequenced as a whole, we conducted a pilot study for comparing short- and long-read sequencing data. Even though PacBio's SMRT sequencing technology still shows some issues with respect to data quality, it is a very promising approach that is going to prove valuable for sequence refinement.

Another important goal of this thesis was to develop a score to assess a human target's conservation across several model species. Publicly available data on the homology relationships between genes and RNA-Seq data build the basis for this score. Using a set of presumably highly conserved genes in human and mouse, we found that the proposed score yields reasonable results. An enrichment of Gene Ontology terms further strengthened our confidence in the conservation score.

Zusammenfassung

Das übergeordnete Ziel der biomedizinischen Forschung ist es, die einer Krankheit zugrunde liegenden Mechanismen zu verstehen und Medikamente zu finden, mit deren Hilfe die Krankheit letztendlich geheilt werden kann. Hierfür müssen zunächst geeignete Gene oder Proteine, welche eine wichtige Rolle bei der Entstehung und dem Verlauf der jeweiligen Krankheit spielen, identifiziert werden. Um Unterschiede zwischen gesunden und erkrankten Individuen zu ermitteln, werden häufig Tiermodelle eingesetzt, welche die Krankheit bzw. bestimmte Aspekte der Krankheit nachbilden. Sobald geeignete Zielmoleküle identifiziert wurden, werden mögliche Wirkstoff evaluiert, welche die Aktivität des Zielmoleküls beeinflussen können. Vielversprechende Wirkstoffe gehen dann in die präklinische Phase, in der die Wirksamkeit sowie die Sicherheit des Wirkstoffes nachgewiesen werden müssen, bevor in der Toxikologie die maximal tolerierte Dosis ermittelt wird. Während alle bisherigen Versuche zumeist mit Nagern durchgeführt werden, kommen hier höhere Säugetiere zum Einsatz.

Laut des Bundesministeriums für Ernährung und Landwirtschaft wurden im Jahr 2017 in deutschen Laboratorien über 2 Millionen Versuchstiere verwendet. Die Mehrheit dieser Tiere waren Mäuse und Ratten, aber auch Hunde, Katzen und Affen stellen wichtige Spezies dar. Im Allgemeinen geht man davon aus, dass Säugetiere die Vorgänge im menschlichen Körper gut abbilden. Allerdings sollte man sich immer darüber im Klaren sein, dass es Spezies-spezifische Unterschiede gibt, die es zu berücksichtigen gilt.

Eines der Ziele dieser Dissertation war es deshalb zu untersuchen, wie ähnlich weit verbreitete Tiermodelle zueinander und zum Menschen auf molekularer Ebene sind. Hierfür wurde die Sequenzidentität mit der Korrelation der Genexpression verglichen, was zu dem Ergebnis führte, dass der Großteil der orthologen Gene sowohl auf Sequenz- als auch auf Expressionsebene sehr ähnlich sind. Allerdings gab es auch Fälle, in denen wir eine hohe Expressionskorrelation bei niedriger Sequenzidentität beobachteten. Diese Fälle wurden genauer untersucht und unsere Hypothese, dass die in Datenbanken wie z.B. Ensemble, UniProt oder RefSeq annotierten Sequenzen möglicherweise fehlerhaft oder unvollständig sind, wurde bestätigt.

Daher eruierten wir die Möglichkeit bekannte Sequenzen von verwandten Spezies zu verwenden, um mangelhaft annotierte Sequenzen zu verbessern. Als Resultat stellen wir das a&o-tool vor, welches sich Orthologiebeziehungen und short-read RNA-Seq Daten zu Nutze macht, um die Sequenz eines Zielproteins zu vervollständigen bzw. zu validieren. Dieser Ansatz würde von sogenannten long-read RNA-Seq Technologien profitieren, da Transkripte hiermit über ihre volle Länge sequenziert werden können. Deshalb führten wir eine Pilotstudie zum Vergleich von short- und long-read Technologien durch. Obwohl die SMRT Sequenziertechnologie von PacBio noch einige Schwächen aufweist, implizieren unsere Ergebnisse, dass es sich um eine vielversprechende Plattform handelt.

Ein weiteres Ziel dieser Arbeit war es, eine Metrik zu entwickeln, anhand derer man den Grad der Konserviertheit eines humanen Zielgens in verschiedenen Tiermodellen bestimmen kann. Als Datengrundlage dienen frei verfügbare Informationen bezüglich der Homologiebeziehungen zwischen Spezies sowie RNA-Seq Daten. Um unseren Ansatz zu validieren, wurden humane Gene verwendet, welche in Mäusen vermutlich hoch konserviert sind. Basierend auf diesem Datensatz und einer Gene Ontology Überrepräsentationsanalyse in Genen mit niedrigen und hohen Werten bezüglich unserer Metrik, können wir schlussfolgern, dass die Metrik zu plausiblen Ergebnissen führt.

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisors Prof. Dr. Kay Nieselt at University of Tübingen, and Dr. Eric Simon at Boehringer Ingelheim. Throughout the last couple of years they provided me with invaluable support and guidance. I am very grateful for the fruitful discussions with both of them. While Kay has helped me to find my way in the academic world, Eric has helped me to get along in the very different environment of a pharmaceutical company. My special thanks also to Kay for motivating me to get through the difficult times that are probably inevitable during a Ph.D. thesis. I would also like to thank Prof. Dr. Stephan Ossowski for agreeing to review my thesis.

Most of my time was spent at Boehringer Ingelheim, therefore, I would like to thank Dr. Udo Maier, Dr. Elia Stupka, Dr. Christian Haslinger, and Eric for making this thesis possible. I am very grateful to all members of the former Target Discovery Research, now Comp Bio, department who shaped my work by discussing ideas or issues I faced. I am especially thankful to those with whom I had the pleasure to write publications with: Dr. Germán Leparc, Dr. Matthias Zwick, Dr. Holger Klein, Dr. Tobias Hildebrandt, and Dr. Tanja Schönberger. A special thanks goes to Germán for his previous work on the approach implemented in the a&o-tool and his general support. I am also grateful for the advice regarding the more technical aspects of bioinformatics that I received from Holger, Matthias, Dr. Katrin Fundel-Clemens, Dr. Francesc Fernandez-Albert, Dr. Fidel Ramírez and Dr. Shen Yang. It was a pleasure to work with Werner Rust who never got tired of my questions regarding the wet-lab part of RNA sequencing, thank you very much. Many thanks to Dr. Nathan Lawless for being such an encouraging and supportive colleague. I would also like to mention the CompBio IT team who have been of great help when it came to finding my way in the sometimes very complicated IT environment of the company.

I am especially grateful to Kay, Eric, Dr. Michaela Willi, and Dr. Thorsten Schweikardt for proofreading my thesis.

I also want to thank my colleagues at University of Tübingen, André Hennig, Dr. Alexander Peltzer, Dr. Alexander Seitz, Judith Neukamm, Andreas Friedrich, and Sven Fillinger for always welcoming me with open arms and providing new perspectives.

Additionally, I would like to thank those who have made life in Biberach enjoyable: Dagmar Knebel, Dr. Bettina Knapp and her family, Christian Wohnhaas, Dr. Miao Sun, Dr. Benjamin Wahl, Dr. Jonas Weinmann, Kai Zuckschwerdt, Karin Fiesel, Dr. Coralie Viollet, Dr. Kolja Becker and Michaela.

Last but not least I would like to thank my family and friends for their support throughout the past years, without them I would not have been able to handle the ups and downs of being a Ph.D. student. Thank you for all your encouragement!

Contents

List of Figures **viii**

List of Tables **x**

List of Abbreviations **xi**

1 Introduction **1**

2 Background **5**

2.1 From target identification to a marketed drug 5

2.2 Homology 7

2.3 RNA sequencing 8

2.3.1 Short-read sequencing 9

2.3.2 Analysis of short-read data 9

2.3.3 Long-read sequencing 11

2.4 A modern approach to data processing 13

3 Assessing the relationship between orthology and gene expression correlation **15**

3.1 Background 15

3.2 Methods 16

3.2.1 RNA-Seq data 16

3.2.2 RNA-Seq data mapping, counting and normalisation . . . 16

3.2.3 Inter-species expression patterns 17

3.2.4 Intra-species expression patterns 17

3.2.5 Comparing gene expression and sequence identity 17

3.3 Results 18

3.3.1 Inter-species expression patterns 18

3.3.2 Intra-species expression patterns 21
3.3.3 Comparing gene expression and sequence identity 25
3.4 Discussion 25

4 Exploiting orthology and *de novo* transcriptome assembly to refine target sequence information 27
4.1 Background 27
4.2 Methods 30
4.2.1 Data description 30
4.2.2 Proportion of genes to be improved 30
4.2.3 *De novo* transcriptome assemblies 31
4.2.4 Evaluation of detection rates in human and related species from RNA-Seq assemblies 31
4.2.5 Generalised refinement pipeline 33
4.3 Results 35
4.3.1 Proportion of genes to be improved 35
4.3.2 *De novo* transcriptome assemblies 35
4.3.3 Evaluation of detection rates in human and related species from RNA-Seq assemblies 36
4.3.4 Application of the generalised refinement pipeline 38
4.4 Discussion 39

5 A pilot study for the comparison of long- and short-read RNA-Seq data 42
5.1 Introduction 42
5.2 Methods 43
5.2.1 PacBio and Illumina RNA-Seq data 43
5.2.2 Comparing PacBio isoforms to the rat genome 44
5.2.3 Comparing PacBio isoforms to known sequences from rat and human 45
5.2.4 Comparing PacBio and Illumina data 46
5.3 Results 47
5.3.1 PacBio and Illumina RNA-Seq data 47
5.3.2 Comparing PacBio isoforms to the rat genome 48
5.3.3 Comparing PacBio isoforms to known sequences from rat and human 48

5.3.4 Comparing PacBio and Illumina data 54
5.4 Discussion and outlook 58

6 Assessing a target's conservation across model organisms 62
6.1 Introduction 62
6.2 Methods 64
6.2.1 Target conservation score 64
6.2.2 Application to all protein-coding human genes 67
6.2.3 targetcon 68
6.3 Results 70
6.3.1 Distribution of subscores 70
6.3.2 Subscores based on co-expression networks 72
6.3.3 Subscore impact 72
6.3.4 GO enrichment for genes with high and low scores 74
6.3.5 Validation using MGI human - mouse disease connection data 76
6.4 Discussion 76

7 Conclusion & outlook 79

A Appendix 81
A.1 a&o-tool 81
A.2 PacBio pilot study 84
A.3 Conservation score 92
A.4 List of publications 95

Bibliography 96

List of Figures

2.1 Drug development pipeline. 6
2.2 Schematic illustration of homology. 7
2.3 PacBio's Iso-Seq protocol. 12

3.1 Schematic illustration of the terms query and target sequence identity. 17
3.2 The percentage of explained variance. 18
3.3 Principal component analysis of mouse and rat gene expression values. 19
3.4 Hierarchical clustering. 20
3.5 Tissue-specific correlation of gene expression in mouse and rat. . . 22
3.6 Principal component analysis of merged mouse and rat gene expression. 23
3.7 Hierarchical clustering of combined expression data. 24
3.8 Comparison of sequence identity and expression correlation. . . . 25

4.1 Comparison of the number of annotated transcripts and curated proteins. 29
4.2 Workflow diagram of the reciprocal best hit BLAST approach. . . 32
4.3 Schematic overview of the a&o-tool. 34
4.4 Proportion of known human proteins with an RBH. 36
4.5 Expression levels of proteins with an RBH and those without one. 37
4.6 Multiple sequence alignment resulting from the application of the a&o-tool to DNAJC11. 38
4.7 a&o-tool results for pig sequences. 39

5.1 SQANTI transcript classification. 45
5.2 Distribution of structural categories identified by SQANTI. 49
5.3 Number of transcripts per gene from PacBio and Illumina data. . 50
5.4 Alignment of PacBio isoforms to the rat genome at the C3 gene locus as an example for PacBio artefacts. 50

5.5 Coverage and sequence identity distributions from the RBH search of PacBio isoforms and human proteins. 53
5.6 Expression levels of genes and transcripts associated with PacBio isoforms with and without an RBH in human UniProtKB/Swiss-Prot. 54
5.7 Comparison of gene expression based on short-read data between highly, lowly and not covered PacBio isoforms. 56
5.8 Comparison of gene expression between genes with an associated PacBio isoform and those without. 57
5.9 Overlap of transcripts detected with PacBio and Illumina. 58

6.1 Phylogenetic tree. 63
6.2 Database schema of targetcon. 69
6.3 Raw scores. 72
6.4 Subscore impact. 73
6.5 Grouping human protein-coding genes according to their conservation score. 74
6.6 GO Biological Process terms overrepresented in the group of genes with a low and high conservation score. 75
6.7 Distribution of the mouse species score for human genes in the HMDC validation data. 76

A.1 Distribution of the difference in sequence identities. 82
A.2 Intersections between genes identified as potentially poorly annotated across species. 82
A.3 Principal component analysis of expression from Illumina short-read data. 85
A.4 Distribution of logarithmised TPM expression values of Illumina samples. 86
A.5 Clustered sample correlation matrix. 87
A.6 Expression of 2125 liver-specific genes from Illumina short-read data. 89
A.7 Comparison of the query length and the percentage of the query covered by the HSP. 90
A.8 Distribution of mapped reads normalised by the number reads in the sample. 91
A.9 Impact of varying correlation thresholds on *s_netNumCorr* and *s_netNumOrtho*. 93
A.10 Impact of varying correlation thresholds on *s_species* and *s_total*. 94

List of Tables

4.1 RNA-Seq data sets used for the analyses. 30

5.1 Productivity metrics of the Sequel run. 47
5.2 Results of the reciprocal best hit BLAST search using PacBio isoforms. 51
5.3 Quantiles of coverage and sequence identity from the RBH search of PacBio isoforms and human proteins. 52
5.4 Sensitivity and precision at several levels achieved with Illumina and PacBio. 55

A.1 Quantitative metrics for the tissue-specific assemblies computed by TransRate. 83
A.2 Details on intermediate results of the PacBio analysis. 84
A.3 Basic information on the 12 rat samples that were sequenced with Illumina for comparison to PacBio. 88
A.4 Contingency table contrasting the reciprocal best BLAST hit category of PacBio isoforms with their associated gene and transcript category assigned by SQANTI. 88
A.5 Input parameters for the targetcon Nextflow pipeline. 92

List of Abbreviations

BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA
CHO	Chinese hamster ovary
contig	Contiguous sequence
EMA	European Medicines Agency
FDA	U.S. Food and Drug Administration
FL	Full-length
FLNC	Full-length non-chimeric
FPKM	Fragments per kilobase of exon per million reads mapped
FSM	Full splice matches
GI	Gastrointestinal
GO	Gene Ontology
GOC	Gene order conservation
GTEx	Genotype-Tissue Expression Project
GTF	General Transfer Format
HMDC	Human - Mouse: Disease Connection
HPO	Human Phenotype Ontology
HQ	High-quality
HSP	Highest Scoring Pair
IGV	Integrative Genomics Viewer
INS1	Insulin 1 gene
ISM	Incomplete splice match
LQ	Low-quality
MGI	Mouse Genome Informatics
MSA	Multiple sequence alignment
MYA	Million years ago
NCBI	National Center for Biotechnology Information
nFL	Non-full-length
NGS	Next generation sequencing
NIC	Novel in catalogue
NNC	Novel not in catalogue
OLC	Overlap-layout-consensus
OMIM	Online Mendelian Inheritance in Man

ORF	Open reading frame
PacBio	Pacific BioSciences
PC	Principal component
PCA	Principal component analysis
RBH	Reciprocal best hit
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
rPCR	Real-time polymerase chain reaction
RPKM	Reads per kilobase of exon per million reads mapped
SAM	Sequence alignment map
SMRT	Single-molecule real-time
TMM	Trimmed mean of M-values
TPM	Transcripts per million
WGA	Whole-genome alignment
ZMW	Zero-mode waveguide

Chapter 1

Introduction

One of the main pillars of biomedical research is the application of mammalian animal models to better understand human diseases. In essence, animal models allow researchers to mimic human disease phenotypes in order to investigate the underlying molecular mechanisms. This is of particular importance during the early research phase of drug development, which comprises the identification and validation of disease-related genes or proteins to derive possible drug targets. Subsequently, known compounds are screened to identify candidate drugs altering the target's activity. Finally, promising compounds are tested for their safety and efficacy. Although advanced *in vitro* assays, such as organs-on-a-chip, and *in silico* methods have recently been developed and are gaining increasing attention, animal experiments are still the method of choice for many steps in biomedical research [45].

Common model organisms include mouse, rat, dog, pig, and cynomolgus monkey, which all greatly differ from human in their obvious physical appearance, their susceptibility to disease, and their response to environmental and experimental influences. For example, substances that are carcinogenic in mice are often not carcinogenic in humans, and vice versa [8].

On a genetic level, these model species are, however, fairly similar to human. Comparative analyses of the human and mouse genomes have revealed that only a few human genes do not have an equivalent one in mouse [23, 76, 77, 116, 118]. For example, the UniProt Consortium [118] states that 75 % of protein-coding genes in C57BL/6J mice have a one-to-one orthologue in human.

The term "orthology" is a more specific form of the general evolutionary principle, called "homology". Two genes are homologous if they originate from a common ancestor via gene duplication or a speciation event during evolution [35, 36]. In biomedical research one is particularly interested in orthologous genes, i.e., genes in two distinct species, which were derived from a common ancestor by a speciation event.

As it is generally assumed that homologous genes share similar function, homology information is often used to assess whether experimental findings obtained in one species can be translated into another mammalian species.

Homology relationships are, however, mainly derived from sequence similarity and may not suffice to derive functional similarity. Although orthologues share the same sequence, they may be involved in different pathways and thus show diverging functions.

Therefore, it is crucial to keep the possibility of differences, with severe effects, in mind when planning animal experiments. A tragic example is the case of TGN1412, a CD28 superagonist antibody, for which efficacy and safety were shown in cynomolgus monkey. However, all six healthy volunteers of the phase I clinical study suffered from a life-threatening cytokine storm and had to be treated in an intensive care unit [10, 43, 112]. Hansen and Leslie [43] stress the fact that minor differences between the human and cynomolgus monkey sequence of the drug's target CD28 probably alter the binding affinity and contributed to the severe side effects observed in human.

By integrating additional information like gene expression data, we can investigate functional similarity of orthologous genes. Here, RNA sequencing (RNA-Seq) is the technology of choice to derive genome-wide gene expression by sequencing the RNA transcribed from all genes, the so called transcriptome.

In this thesis, the link between sequence similarity and gene expression was investigated in greater detail by studying various model species that are of interest to biomedical research. The ultimate goal was to assess whether combining sequence similarity and gene expression can improve early biomedical research and reduce the number of experiments conducted in animals.

One issue researchers often encounter in drug development, is the incompleteness and incorrectness of sequence information available in public databases. As Steven L. Salzberg [103] has put it, despite the rapid improvement in genome sequencing and assembly technologies, "errors in annotation are just as prevalent as they were in the past, if not more so." While there are tools focusing on improving the overall gene structure [56, 131], biomedical researchers would often benefit from an approach to reconstruct the sequence of a specific target protein, even in species with poor genome annotation, as many of these, such as rat and dog, are highly relevant for biomedical research. Therefore, we aim to develop a tool that makes use of a well annotated protein sequence from a closely related species to derive the target's protein sequence in a *de novo* transcriptome assembly of the species of interest. Thereby, we can provide a small contribution to improving the overall situation in genome annotation stated by Salzberg.

As the approach described above would greatly benefit from directly sequencing entire RNA molecules without the need of assembly, one project dealt with the possibility to use long read sequencing technologies. Here, the focus was to compare short- and long-read sequencing and the possibility of obtaining full-length transcripts.

Another important aspect of this thesis is target prioritisation. To our knowledge, this is mainly done on information about the target gene in human. The scoring approach implemented in OpenTargets [18] incorporates mouse models into their target-disease-association score. However, we wanted to take it a step further and provide a method that takes sequence and expression based information of a variety of model species into account. The resulting target conservation score does not only facilitate target prioritisation, but also species selection as we are able to determine the model species in which the human target gene is conserved best.

Overall, this thesis aims to provide insight into the link between sequence information and gene expression across a broad spectrum of model species—in particular those relevant for biomedical research—and human. If we were able to ensure that we are using the correct target sequence and the most suitable model species *in silico*, we could reduce the number of projects failing at late stages of drug development due to species related differences and thus spare thousands of animal lives.

Outline

This thesis is structured as follows: Chapter 2 covers background information required for the following chapters. The background chapter starts with a description of the drug development process emphasising the role of animal models and the high number of failing drug discovery projects. Then the evolutionary concept of homology will be explained in greater detail, before RNA sequencing is introduced. Here, we cover short- and long-read sequencing technologies and explain the different bioinformatics analysis strategies. A brief introduction of Nextflow, a framework for building scalable and reproducible analysis pipelines, concludes the background chapter.

In Chapter 3, we investigate gene expression patterns within and between mouse and rat, before we examine the relationship between orthology and gene expression correlation in the two rodent species.

Chapter 4 focuses on available genome annotations and the fact that sequence information provided via public databases may contain errors or may be incomplete. We assess the proportion of affected genes and investigate whether exploiting orthology relationships and *de novo* transcriptome assemblies from RNA-Seq data can improve these sequences. Finally, we introduce and apply the a&o-tool that was developed to perform sequence refinement for poorly annotated genes by using an orthologous bait protein and a *de novo* transcriptome assembly of short-read RNA-Seq data. The approach behind the a&o-tool would greatly benefit from long-read sequencing data as the assembly step could be skipped. Therefore, Chapter 5 presents the results of a pilot study for the comparison of long- and short-read RNA-Seq data.

In Chapter 6 a score is proposed which provides a measure for the conservation of a human target gene across several model species. The scores resulting from the application to all human protein-coding genes using mouse and rat as model species, are used to investigate the individual subscores as well as their contribution to the overall conservation score. Results are validated via a Gene Ontology enrichment of genes with low and high conservation scores. Furthermore, we use a data set containing human and mouse disease connections, provided by the Jackson Laboratory, for validation of the proposed conservation score.

Chapter 7 concludes this thesis and provides ideas for further research.

Chapter 2

Background

This chapter paves the way for the following ones by providing background on concepts that are important for the understanding of this thesis. As the thesis was conducted at the University of Tübingen and Boehringer Ingelheim, the pharmaceutical context is established before we define homology. Then, short- and long-read RNA sequencing are introduced and important bioinformatics analysis steps that were used in this thesis are explained. A brief introduction of modern approaches to data processing concludes this chapter.

2.1 From target identification to a marketed drug

A disease and an unmet clinical need lie at the root of the drug development pipeline which starts with basic research to elucidate the molecular mechanisms of the disease (see Figure 2.1). During target identification, researchers look into molecular mechanisms which are causally linked to the disease and that are suitable for pharmacological intervention, i.e., gene or protein inhibition or activation [49]. These genes or proteins are referred to as drug targets and have to meet certain requirements regarding their efficacy, safety and druggability. The latter describes the fact that a therapeutic compound can bind to the target and trigger a measurable response. Following its selection, the target is validated using *in vitro* assays and animal models.

The aim of the discovery phase is to determine candidate compounds affecting the target by, e.g., performing a high-throughput screening of large compound sets against the target. Once such hit molecules are found, they are further investigated and optimised with respect to their potency and selectivity (mainly *in vitro*) as well as their efficacy and safety (mainly *in vivo*). Successful compounds then move on to *in vivo* studies where their pharmacology, i.e., pharmacokinetics and pharmacodynamics, as well as their toxicology are investigated in different model species. Pharmacokinetics examines how long it takes for the drug to be absorbed

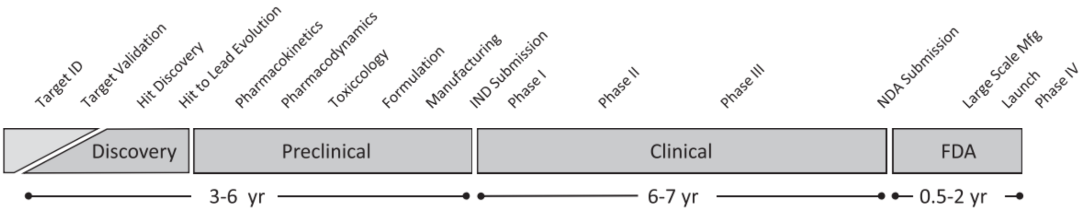


Figure 2.1: Required steps to proceed from basic research to a drug on the market. The European equivalent to the American FDA (U.S. Food and Drug Administration) is the EMA (European Medicines Agency). Target ID: target identification; IND: investigational new drug; NDA: new drug application; Mfg: manufacturing. Figure reprinted with permission from [102].

and metabolised, while pharmacodynamics looks into the concentration of the drug and how that is related to the drug's effects, both beneficial and adverse ones [70]. Toxicological experiments aim at assessing the toxicity of the drug and its metabolites to ensure safety before the drug is tested in clinical trials.

Once a drug has passed the preclinical studies, where it is only tested in *in vitro* assays and model species, it is transferred into the clinics. In the clinical trials the drug is administered to humans, first to healthy volunteers an then to patients, to investigate its effects and its therapeutic value in human.

Finally, after an average of 12 years [102] of development, a drug successfully passing the clinical trials has to be approved by regulatory agencies like the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA).

Clearly, the attrition rate of the drug development process is high, i.e., the initial compound screening leads to hundreds of thousands of hits of which many do not show suitable drug-like properties [92]. The term drug-like refers to a wide range of structural, physio- and biochemical, as well as pharmacokinetic and toxicity related properties, such as solubility, molecular weight or bioavailability [57]. The majority of the selected hits then fail due to insufficient efficacy or safety. Cook et al. [21] have analysed the small-molecule drug projects at AstraZeneca between 2005 and 2010 with respect to project success and the reasons for project failure. They found that 66 % of the projects in the preclinical phase and 59 % of those in the clinical trial phase I were successful. However, of these only 15 % succeeded in phase II clinical trials. During the investigated period of time only five projects reached phase III clinical trials of which 60 % were successful. In preclinical and phase I clinical trials safety was the major cause of project failure while in phase II clinical trials, a lack of efficacy lead to project closure in most of the cases.

One of the main reasons for the high number of compounds failing due to insufficient efficacy in human is probably the diverging molecular function of drug targets in human and the model species used for identifying the target and assessing the drug's therapeutic effects.

2.2 Homology

Translating experimental findings between species is a critical step occurring at different stages of the drug development process like target identification or moving from pre-clinical testing to clinical trials. To investigate a target gene in different model organisms, one has to make sure that the gene exists and is highly similar in each of the species. The term “homology”, coined by Fitch [35, 36], describes the fact that two genes originated from a common ancestor. Depending on the type of evolutionary event separating the two genes, one further distinguishes “orthologues” and “paralogues”. A speciation event leads to orthologous genes while a gene duplication results in paralogous genes (see Figure 2.2).

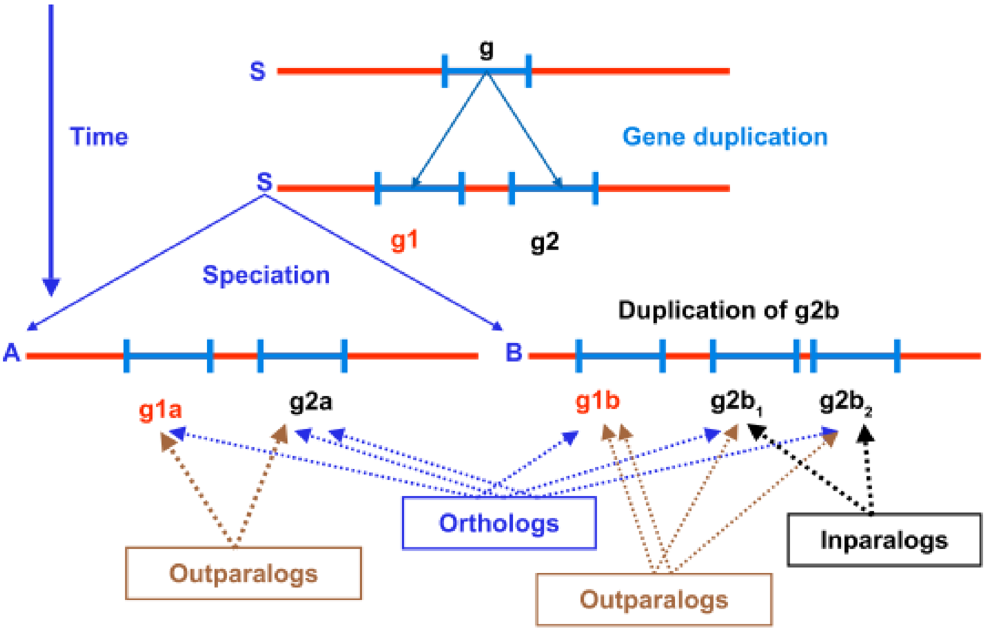


Figure 2.2: This example shows three species (S, A and B) and speciation as well as gene duplication evens to explain the terms “homologue”, “paralogue” and “orthologue”. The genes $g1$ and $g2$ are inparalogues as they resulted from a gene duplication in S. A speciation event lead to species A and B containing the orthologous genes $g1a$ and $g1b$. Because $g2b$ was duplicated in species B, the resulting genes $g2b_1$ and $g2b_2$ are both orthologous to $g2a$ in A. Due to the duplication of gene g in S, $g1a$ and $g2a$ are referred to as outparalogues. The figure is reprinted with permission from [115].

As we do not know how genes really evolved across different species, two approaches for orthology inference have been established: Sequence similarity based ones, such as InParanoid/MultiParanoid [94, 111] or eggNOG [48], and phylogeny-based methods, like EnsemblCompara GeneTrees [132] or UPhO [12]. A more comprehensive list of homology prediction methods can be found in [115].

Comparative genomics is often based on the general assumption that orthologous genes are functionally similar and can therefore be used for functional annotation of closely related species. It is also assumed that orthologues are functionally more similar than paralogous genes. This is referred to as the “orthology conjecture”. However, up to date there are numerous contradicting studies on whether the ortholog conjecture is true. Nehrt et al. [79] have used Gene Ontology (GO) annotation and microarray analysis in human and mouse to test the ortholog conjecture and found that paralogues are more functionally similar to each other than orthologues. Inspired by the conclusion drawn by Nehrt et al., Chen and Zhang [19] critically evaluated the suitability of GO terms for the assessment of functional similarity. Furthermore, they used RNA-Seq instead of microarrays to investigate the expression similarity of orthologues and paralogues. In contradiction to Nehrt et al., they found that expression similarity is greater between orthologues than between within-species paralogues and that GO terms suffer from biases that render them unsuitable for the assessment of the ortholog conjecture. Another study compared tissue-specificity of gene expression between orthologues and paralogues and found that tissue-specificity is conserved in orthologous genes while it differs for paralogous genes [60].

2.3 RNA sequencing

Examining the gene expression of potential drug targets is a crucial part of target identification and validation as it provides insight into the targets’ disease link [49] and their expression profile across different tissues. Nowadays RNA sequencing is the technology of choice to determine gene expression on a genome-wide level. The preparation of a complementary DNA (cDNA) library from the isolated RNA precedes the actual sequencing as most technologies do not sequence RNA directly but its cDNA. Briefly, RNA is isolated from a tissue sample, filtered for, e.g., mRNA and reverse transcribed to cDNA. Sequencing adapters are then added to both ends of the cDNA fragments. During library preparation one also assesses the RNA quality, in particular with respect to RNA degradation by computing the sample’s RNA integrity number (RIN). Together with the amount of RNA the RIN is used to filter for high quality samples which then are sequenced.

The sequencing itself results in nucleotide sequences referred to as “reads”. There are several RNA-Seq methods which can be classified into short- and long-read sequencing methods according to the length of these reads.

2.3.1 Short-read sequencing

The most widely used sequencing method is sequencing by synthesis [39,52] offered by Illumina. Here read lengths up to 300 bp can be reached [51].

Sequencing by synthesis

Sequencing by synthesis is performed on glass slides, so-called flow cells, that are made up of lanes coated with two types of oligos required for the hybridisation of cDNA fragments to the the flow cell. By generating a sequence complementary to the immobilised fragment, a double stranded DNA molecule is formed. This double-stranded DNA is then denaturated and the original sequence template is washed off. Using a method called bridge amplification the resulting sequence molecule is clonally amplified. This process is repeated until millions of clusters, each containing copies of a single cDNA fragment, are produced.

The actual sequencing then starts by extending the sequencing primer and adding fluorescently labelled nucleotides that bind to the fragments on the flow cell. By exciting the clusters with light and detecting the emitted flourescent signals, the base is determined. This process of adding nucleotides and determining the added base via imaging is repeated until the desired read length is achieved.

If the cDNA fragments are sequenced in only one direction, so-called single-end reads are generated. In an additional step, the reverse strand is synthesised and the sequencing steps are repeated to obtain a second read from the other side of the cDNA fragment. Since this method yields two reads per fragment, it is referred to as paired-end sequencing. Even though paired-end sequencing is more expensive, it is beneficial for the identification of novel isoforms, the analysis of isoform expression, or the investigation of poorly annotated species [20, 38, 55].

Raw data analysis

An initial quality control of raw reads precedes all further analyses and includes general sequence quality, GC content, sequence length distribution, duplication levels, adapter content, and the presence of overrepresented sequences [15, 20]. Samples with sufficient quality are then subjected to further processing which depends on the desired analysis.

2.3.2 Analysis of short-read data

Mapping-based analysis

Here, we are going to present the steps required for expression analysis in species with an annotated genome. By aligning raw reads to the reference genome, expressed genes are detected and their expression level can be quantified based on the number of mapped reads. Usually, the mapping process results in an alignment stored in SAM (sequence alignment map) or BAM (binary alignment map) format. Together with an annotation file (usually in gene transfer format (GTF)) the alignment is then used for counting reads mapping to each annotated gene region. Since factors like the total number of sequenced reads in a sample (sequencing depth) or the gene length affect the number of reads mapping to the gene, these counts have to be normalised before they can be compared within or between samples. RPKM (reads per kilobase of exon per million reads mapped) [75], for single-end sequencing, and FPKM (fragments per kilobase of exon per million reads mapped) [122], for paired-end reads, are widely used normalisation methods that aim at eliminating the effect of sequencing depth and gene length. Another normalisation method that also takes sequencing depth and gene length into account is the TPM (transcripts per million) [64]. When used for differential gene expression, all three units, do however still suffer from the issue that longer genes are more likely to be called as differentially expressed [17]. Quantile normalisation, the trimmed mean of M values (TMM) [97], or DESeq [66] aim at making expression levels truly comparable between samples.

De novo transcriptome assembly from short reads

The analysis described above relies on the availability of an annotated reference genome. Depending on the quality of the available reference, transcripts might be missed and some species entirely lack a reference genome. In both cases one can resort to *de novo* transcriptome assembly from the short-read RNA-Seq data.

Most assemblers apply one of two basic approaches: Overlap-layout-consensus (OLC) graphs or de Brujin graphs. The OLC approach involves the computation of all pairwise overlaps between reads and searching for a path through the graph visiting every node exactly once (Hamiltonian path problem). Because the Hamiltonian path problem is NP-complete, the OLC approach is computationally intense and not solvable in an efficient manner for millions of short reads [50, 91] as in a typical RNA-Seq experiment. In the de Brujin graph approach, reads are split into all possible substrings of length k, so-called k-mers. The de Brujin graph is constructed by using unique (k-1)-mers as nodes and connecting two nodes v_i and v_j, if there is a k-mer whose first k-1 bases match v_i and whose last k-1 bases match v_j. Contiguous sequences (contigs) are then generated by solving an Eulerian path problem, i.e., finding paths that visit every edge once, on this graph. In contrast to a Hamiltonian path, an Eulerian path can be determined efficiently [91]. Hence, de Brujin graphs are the method of choice for short-read data.

Both, the OLC and the de Brujin graph approach, are applicable to genomic as well as transcriptomic data, however, transcriptome assemblers have to take some challenging properties of RNA-Seq reads into account [68]. First of all, highly abundant transcripts lead to a greater number of reads. Many genome assemblers would remove these transcripts as they would be misunderstood as repetitive regions. Secondly, RNA-Seq protocols may be strand-specific. To detect overlaps between forward and reverse strand reads, one often includes both versions of the reads in the de Brujin graph. Thirdly, transcriptome assemblers have to be splice-aware. Several transcripts might be alternative isoforms of one gene and may, therefore, contain identical subsequences because they share exons.

There are several published transcriptome assemblers such as Trinity [40], Oases [105], SPAdes-rna [13], or BinPacker [65], to just name a few. Recently, Hölzer and Marz [46] compared several *de novo* transcriptome assemblers across different species and found that, although some assemblers perform well on most of the data sets, it is still advisable to test different assemblers and assembly parameters as there is not a single tool that performs best on all data sets.

Once the transcriptome has been assembled, one can, for example, investigate the assembled sequences or map the reads back to the assembled transcriptome to derive gene expression levels.

2.3.3 Long-read sequencing

While short-read sequencing enables the analysis of gene expression at relatively low cost and high throughput, it is less suitable for *de novo* isoform identification because the short reads have to be assembled into transcripts. Long-read sequencing technologies like the Single Molecule, Real-Time (SMRT) sequencing offered by Pacific Biosciences (PacBio), recently acquired by Illumina, generate reads with a median read length of 50 kb and up to more than 100 kb (PacBio's Sequel system). Therefore, most transcripts can be sequenced as a whole, eliminating the need for transcript assembly as required for short-read data.

In this thesis long-read data from PacBio's Sequel system has been analysed, thus PacBio's SMRT sequencing approach will be introduced in greater detail below.

SMRT sequencing

In contrast to Illumina's sequencing by synthesis, SMRT sequencing does not involve clonal amplification of the synthesised cDNA molecules. Instead, a single cDNA molecule is transformed into a circular DNA molecule by adding hairpin structures of single-stranded DNA to both sides of the double-stranded DNA. Together with a polymerase, this molecule is then loaded into a zero-mode waveguide (ZMW) [30] where it is sequenced [85]. The DNA-polymerase complex is immobilised at the bottom of the ZMW and a solution containing all four fluorescently labelled nucleotides is added. Once the polymerase incorporates one of the nucleotides into the generated DNA strand, the fluorophore is detached from the nucleotide and a camera records the emitted light signal. This results in so-called polymerase reads (see Figure 2.3A) which are split into subreads by removing the adapter DNA sequence. In an error-correction step these subreads are summarised in a circular consensus sequence (CCS). These CCSs are then processed with PacBio's bioinformatics pipeline.

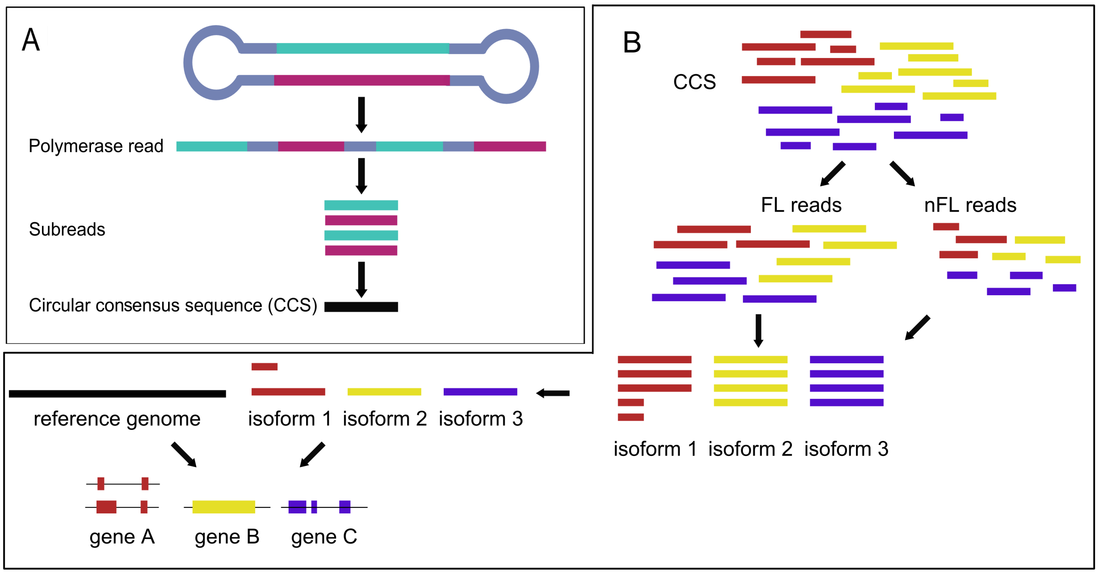


Figure 2.3: PacBio's Iso-Seq protocol. A) The cDNA molecule is extended by single-stranded DNA and this SMRTbell template is then sequenced in a ZMW. The resulting polymerase read is split into subreads which are summarised by a CCS. B) All CCSs generated in a sequencing run are classified into full-length (FL) and non full-length (nFL) sequences. The FL reads are then clustered by similarity and high-quality isoforms are aligned to the reference genome to obtain unique, collapsed isoforms. This figure is a combination of adapted figures from [86, 125].

PacBio's raw read data analysis

The generated CCSs are classified into full-length (FL) and non-full-length (nFL) reads (see Figure 2.3B), i.e., if a CCS contains the 5′ and 3′ adapter sequences and a poly-A sequence, it is a full-length read. These FL and nFL reads are clustered by sequence similarity to obtain error-corrected candidate sequences. In the following polishing step, these candidate transcripts are split into high- and low-quality isoforms based on their accuracy and the high-quality ones are then mapped to a reference genome, if available. The alignment to a reference allows the set of isoforms to be further cleaned as redundant isoforms are collapsed. The resulting set of unique isoforms is referred to as “PacBio isoforms” in this thesis.

2.4 A modern approach to data processing

The analysis of RNA-Seq data involves several independent tools (see section 2.3.2) being applied to many (potentially thousands of) samples. In recent years, there have been several attempts to ease the process of developing such bioinformatics pipelines and to provide a framework for reproducible data analysis. Snakemake [59], Bpipe [101], and Nextflow [25] are examples for workflow languages with Nextflow gaining increasing attention by the (bioinformatics) community.

A Nextflow pipeline consists of processes that are connected by channels through which data is passed from one process to the next. Each process contains a script which is executed as a bash script in the host system. Therefore, it can be written in any programming language supported by the host system.

Let us look at an example process which generates the genome index required by STAR, a popular aligner used to align RNA-Seq reads to a reference genome (see Listing 2.1). In lines seven and eight we declare the input of the process. The path to the reference genome which is to be indexed, is provided via the Channel genome_file that has been created earlier in the script. Lines ten and 11 specify that the directory genome_dir is made available to the following process performing the actual STAR alignment via the Channel genome_dir_ch. The process script performs the actual task of creating the output directory genome_dir_ch and generating the genome index which is then stored in genome_dir_ch.

Incorporating Docker [72] and Singularity [63] further increases reproducibility and also portability because the required software is packed into so-called containers in which the Nextflow processes are then run. That way the software does not have to be installed on the host system and one can even define separate environments with, for example, differing software versions for individual processes. Whenever possible, processes are run in parallel, e.g., the quality analysis of raw reads can be done for all samples in parallel as these processes do not depend on one another. Furthermore, Nextflow facilitates highly scalable data analysis as it is easily configurable to run the processes on a high performance cluster or a cloud platform.

Although we did not use any of their pipelines in this thesis, we want to point out that the nf-core [33] makes Nextflow pipelines for common tasks such as RNA-Seq analysis easily accessible and usable, even for people new to Nextflow. The initiative evolved from the highly active community of Nextflow users and aims at establishing a collection of pipelines fulfilling certain standards.

Listing 2.1: An example Nextflow process creating the genome index for the STAR aligner.

```
/*
* create the STAR genome index
*/

process make_star_genome_index {

input:
file(genome) from genome_file

output:
file(genome_dir) into genome_dir_ch

script:
"""
mkdir genome_dir

STAR --runMode genomeGenerate \
--genomeDir genome_dir \
--genomeFastaFiles ${genome} \
--runThreadN 16
"""
}
```

Chapter 3

Assessing the relationship between orthology and gene expression correlation

The content of this chapter was partially published in:

> J. F. Söllner, G. Leparc, T. Hildebrandt, H. Klein, L. Thomas, E. Stupka and E. Simon. *An RNA-Seq atlas of gene expression in mouse and rat normal tissues.* Scientific Data 2017: 4

I analysed the data, conducted the downstream analyses, and generated the figures presented in this thesis. G. Leparc implemented and described the data analysis pipeline and carried out the primary analysis of the data. T. Hildebrandt directed the RNA preparation and sequencing of the samples and was involved in the design of the study. H. Klein contributed code and supervised the design of the primary data analysis pipeline. L. Thomas performed the in-vivo analyses. E. Stupka reviewed the draft of the paper. E. Simon analysed data (results not included in this thesis), wrote the first draft of the manuscript and supervised the complete study.

3.1 Background

All living organisms consist of a variety of different tissues and organs which play very diverse roles in maintaining the body's functionality. The existence of a wide range of highly specialised cells expressing distinct sets of proteins facilitates this process.

In diseased individuals the cells' function is affected and biomedical research aims at determining the underlying mechanisms to differentiate healthy from diseased individuals. Ideally, one would investigate the proteome directly. Unfortunately, there are still limitations like the detection of lowly expressed proteins or technical issues when separating large protein complexes [41]. Therefore, researchers often resort to RNA sequencing as a surrogate for proteomics analyses, which is also reflected by the fact that repositories of proteomics data are used less commonly [130] than those containing RNA-Seq data.

To determine differences between healthy and diseased tissue, one has to acquire samples from both tissue types. In human studies, it is often not feasible to obtain healthy tissue, imagine brain biopsies, and in animal models it would also be desirable to reduce the number of sacrificed animals. Therefore, data repositories such as the Genotype-Tissue Expression Project (GTEx) and the rat BodyMap [141], providing access to RNA-Seq data from healthy individuals, are a valuable resource. The aim of the project upon which this chapter is based, was to add gene expression data sets from mouse and rat acquired in a well controlled experiment and capturing a wide range of tissues, to the public domain.

In this chapter we will investigate the descriptive analysis of the generated RNA-Seq data sets. Furthermore, the combined data from mouse and rat are examined to assess whether gene expression patterns are tissue- or species-specific and whether conserved genes share gene expression patterns.

3.2 Methods

3.2.1 RNA-Seq data

Samples from 13 tissues were taken from three male BL/6J mice (C57BL/6J) and three male Wistar Han rats (Crl:WI(Han)). The investigated tissues were: Brain, kidney, heart, thymus, pancreas, esophagus, stomach, duodenum, jejunum, ileum, colon, liver and muscle (quadriceps). Total RNA was extracted and libraries were prepared according to standard protocols (see our publication [109] for details). Sequencing was then performed as 50 bp, single-end reads and seven bases index reads on an Illumina HiSeq2000.

3.2.2 RNA-Seq data mapping, counting and normalisation

We had to exclude one sample (mouse_11_heart) from the analysis due to technical issues. FastQC [15] (v0.11.2) was used to evaluate the read quality of all remaining samples. RNA-Seq reads from mouse and rat were aligned to the corresponding Ensembl 84 reference genome with STAR [26] (v2.5.2a11) and alignment quality metrics were obtained with RNASeQC [24] (v1.1812). Duplication rates were identified using bamUtil [54] (v1.0.11) and assessed with the dupRadar [104] (v1.4) Bioconductor [47] R [93] package. We used Cufflinks [121] (version 2.2.114) to get the reads per kilobase of transcript per million mapped reads (RPKM) and the featureCounts [108] software package to obtain read counts. The subsequent analysis was conducted in R: Count values were normalised to trimmed mean of M-values (TMM) [98] using the calcNormFactors() function from the edgeR package [69] and these were voom normalised using the corresponding function from the limma [95] package.

3.2.3 Inter-species expression patterns

As an initial quality control we performed the following on the voom-transformed gene expression values of mouse and rat separately: Distribution analysis, principal component analysis (PCA), and hierarchical clustering.

3.2.4 Intra-species expression patterns

To investigate whether mouse and rat share gene expression patterns, we queried the one-to-one rat orthologue for all protein-coding mouse genes using the biomaRt package [27, 28] (version 2.32.1) with Ensembl version 84. Based on these orthologous pairs we compiled merged expression matrices, one with median tissue RPKMs for correlation analysis and one with voom-normalised counts for PCA. To assess the tissue-specific correlation of the two species, we computed the Pearson's correlation coefficient and generated scatter plots for each tissue.

3.2.5 Comparing gene expression and sequence identity

When querying the rat orthologues from Ensembl, we also retrieved the sequence identity which is calculated from the pairwise protein sequence alignment. For each pair of orthologous genes, Ensembl reports a query and a target sequence identity (see Figure 3.1). We summarised these two values using their mean value which was then compared to the Pearson's correlation coefficient of median tissue expression (RPKM).

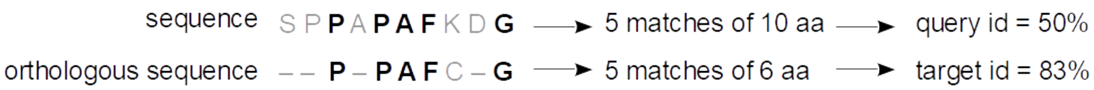


Figure 3.1: Schematic illustration of the terms query and target sequence identity as used by Ensembl. Black letters indicate matching amino acids while grey ones represent mismatches. The proportion of the sequence of interest (top) covered by the pairwise protein sequence alignment with an orthologous sequence (bottom) is referred to as "query identity". The percentage of the orthologous sequence covered by the alignment is called "target identity".

3.3 Results

3.3.1 Inter-species expression patterns

Based on the percentage of variance explained by each principal component (see Figure 3.2) we concluded that in mouse the majority of variance was explained by the first two components (PC1 and PC2), while in rat PC1 explained the greatest proportion of variance in the data. Due to the high dimensionality of the data, the percentage of variance explained by the other principal components only declined slowly.

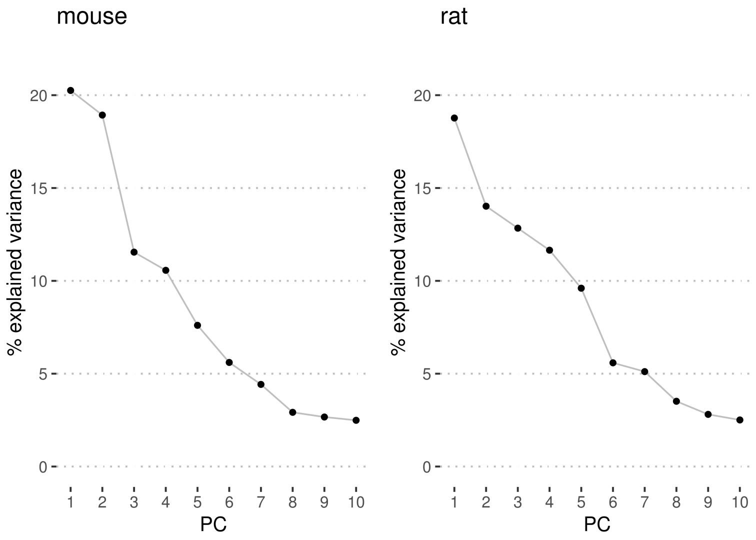
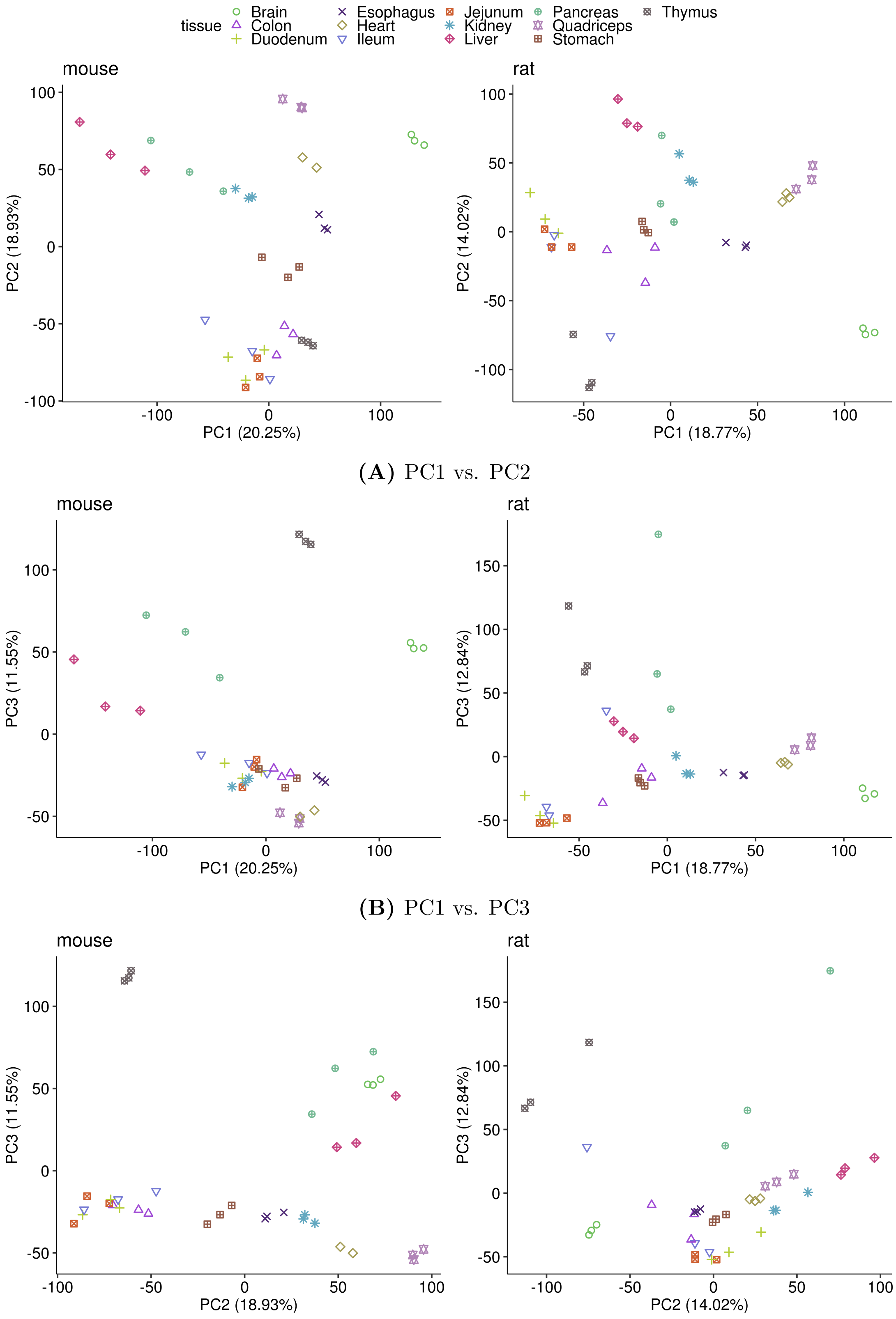


Figure 3.2: The percentage of total variance explained by the first ten principal components for mouse (left) and rat (right).

Projecting the samples into the space spanned by PC1 and PC2, revealed that the effect tissue origin had on the gene expression of mouse and rat was greater than the animal effect (see Figure 3.3A). Samples from muscular tissue, i.e., heart and quadriceps, were close to each other. This was particularly pronounced in rat. Brain samples formed a cluster which was separated from all other clusters. A larger, more mixed cluster consisted of samples from the gastrointestinal (GI) tract, i.e., duodenum, jejunum, ileum and colon, and thymus. Apart from the GI samples, liver and pancreas were the tissues with the most variability. In mouse, PC3 separated thymus samples from the other tissue clusters (see Figures 3.3B and 3.3C). A hierarchical clustering analysis of the two data sets confirmed the observation that the tissue effect is greater than the animal effect (see Figure 3.4). In mouse, the samples from the GI tract again formed a mixed cluster. In rat, we noted that the colon samples were grouped together and that their cluster was more similar to stomach and esophagus than to the other GI tissues. Furthermore, one pancreas sample in rat was separated from the other two pancreas samples.



(A) PC1 vs. PC2

(B) PC1 vs. PC3

(C) PC2 vs. PC3

Figure 3.3: Principal component analysis of mouse and rat gene expression values. Samples are coloured by tissue and the numbers in brackets correspond to the proportion of variance explained by the respective principal component.

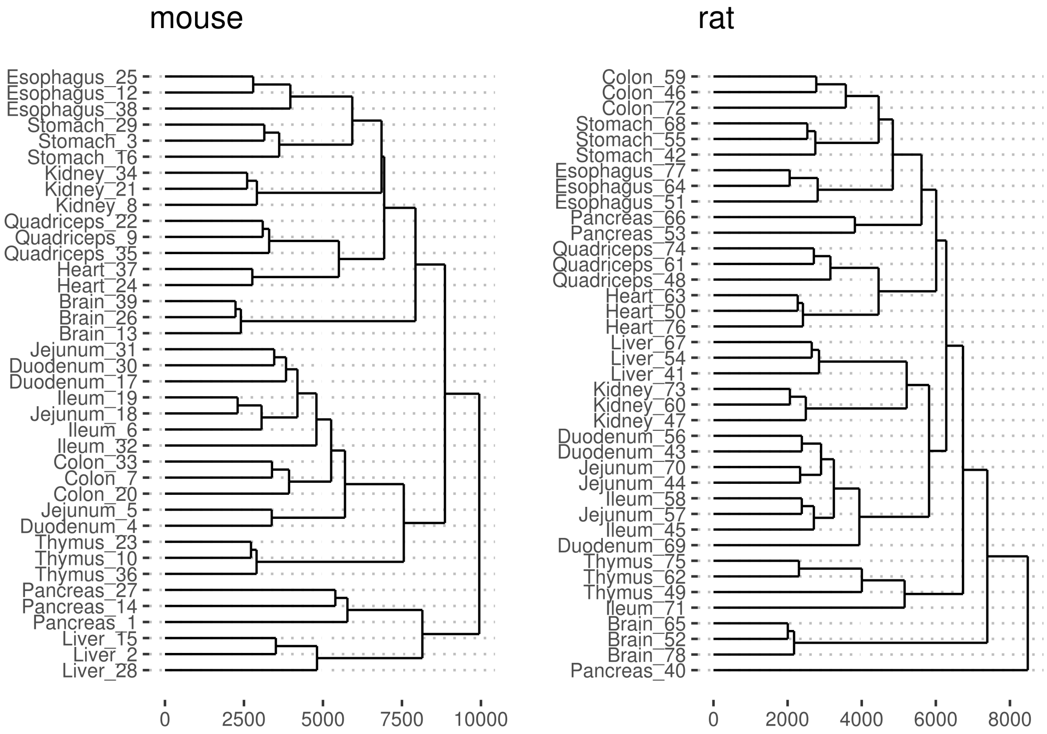
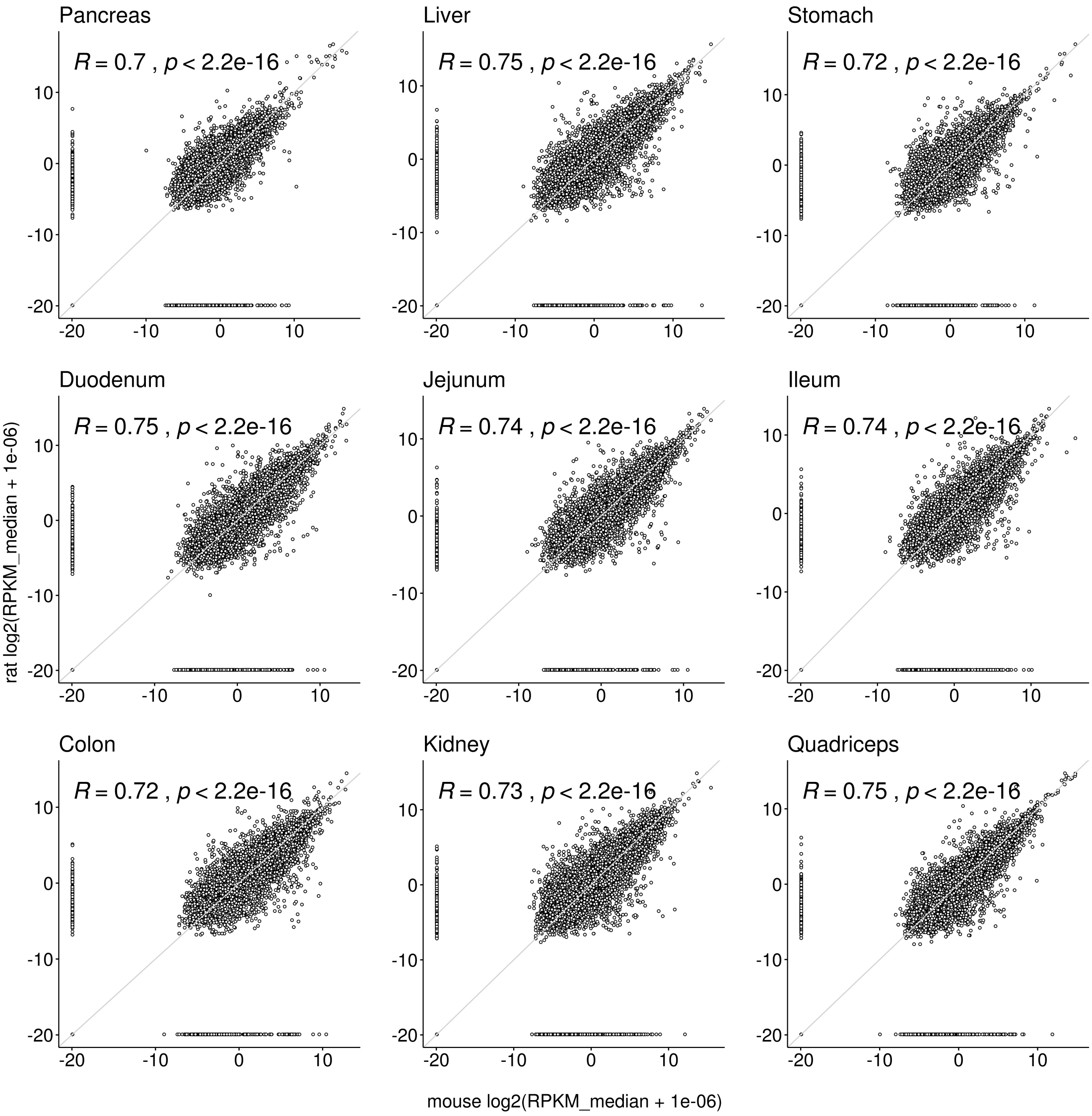


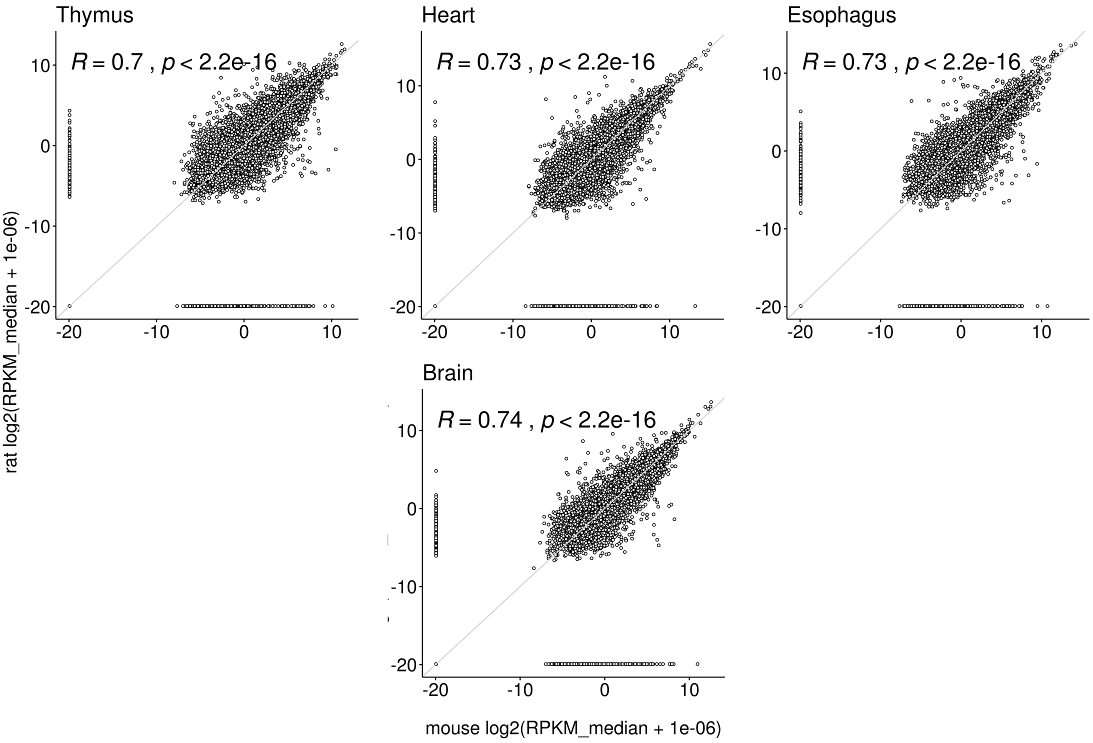
Figure 3.4: Hierarchical clustering of mouse (left) and rat samples (right) based on voom-transformed log(counts per million). The Canberra distance between samples and the complete linkage method were used for clustering.

3.3.2 Intra-species expression patterns

A comparison of the gene expression between mouse and rat showed that, in general, the expression of one-to-one orthologues is well correlated (see Figure 3.5). There are, however, a number of genes which are not expressed in only one of the two species. Therefore, the overall Perason's correlation coefficient was around 0.7 in all tissues even though the majority of the points cluster along the diagonal.



(A) Pancreas, Liver, Stomach, Duodenum, Jejunum, Ileum, Colon, Kidney, and Quadriceps.



(B) Thymus, Heat, Esophagus, and Brain.

Figure 3.5: Tissue-specific correlation of gene expression in mouse and rat. R: Pearson's correlation coefficient; grey line: diagonal.

Dimensionality reduction revealed that the first three PCs are those explaining the greatest portion of the total variance in the data set (see Figure 3.6A). Projecting the data into the space spanned by PC1 and PC2 (see Figure 3.6B), a clustering by tissue instead of species was observed. The GI tract again formed a rather mixed up supercluster which was separated from the other tissues. Quadriceps and heart samples from both species clustered together, with the esophagus cluster being the closest neighbouring cluster. The greater spread of pancreas samples from both species that we observed in the variability analysis in section 3.3.1 was also visible in Figure 3.6b. Brain samples from both mouse and rat formed the most distinct clusters. Figures 3.6C and 3.6D clearly showed that PC3 separated the two species.

In the hierarchical clustering of the combined mouse and rat expression data (see Figure 3.7), we observed that, for the majority of tissues, samples from both species cluster together. For example, kidney samples from mouse and rat form two clusters that can be merged into a joint kidney cluster. The two muscular tissues are an exception because mouse heart and quadriceps as well as rat heart and quadriceps form a cluster. These two clusters are, however, again similar to each other.

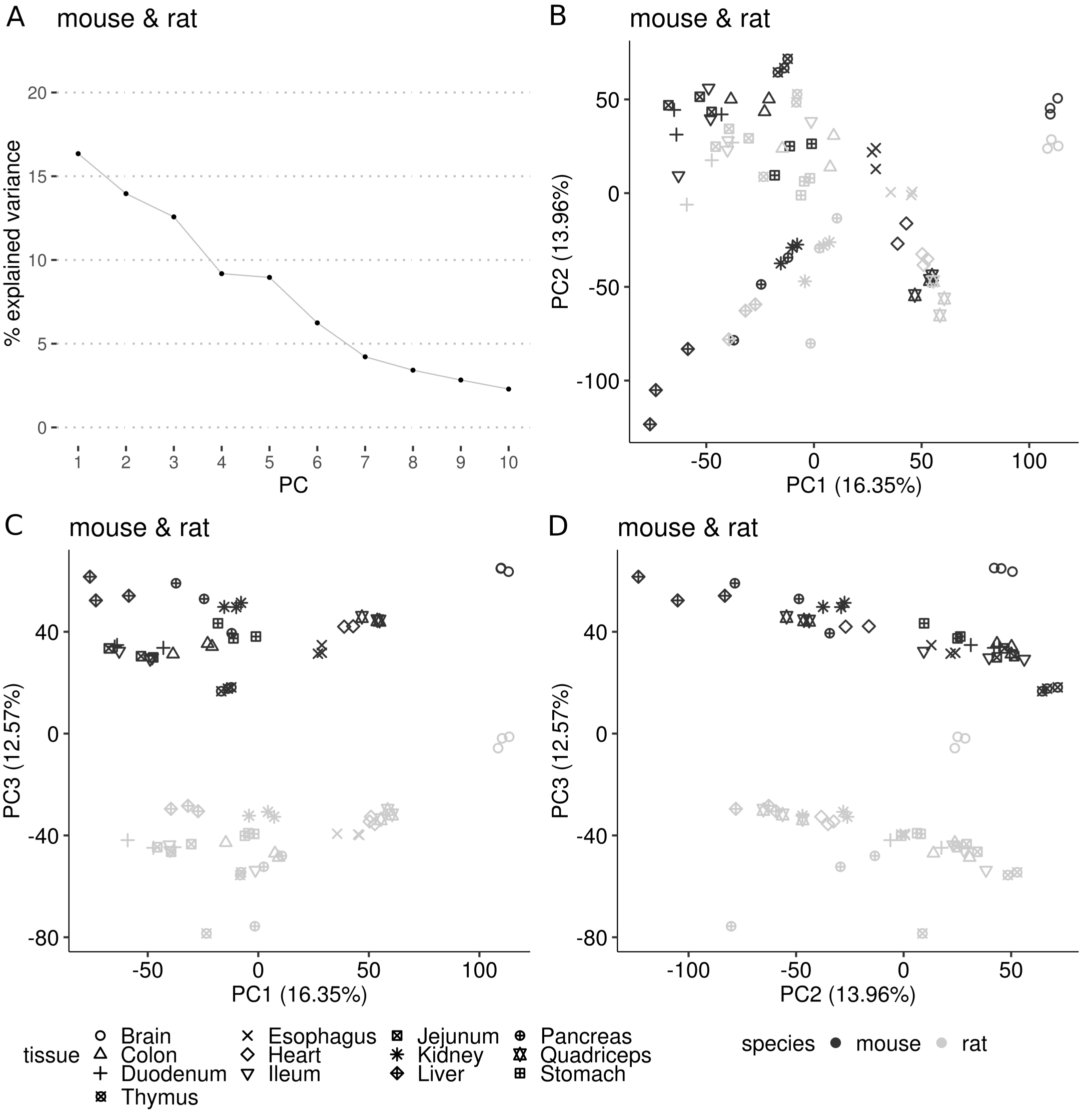


Figure 3.6: Principal component analysis of merged mouse and rat gene expression. A) Percentage of total variance explained by the first ten principal components (PCs). B)-D) Gene expression transformed to the space spanned by different combinations of PC1 to PC3. Point shape and colour represent the tissue and species of origin, respectively. Numbers in brackets correspond to the explained variance.

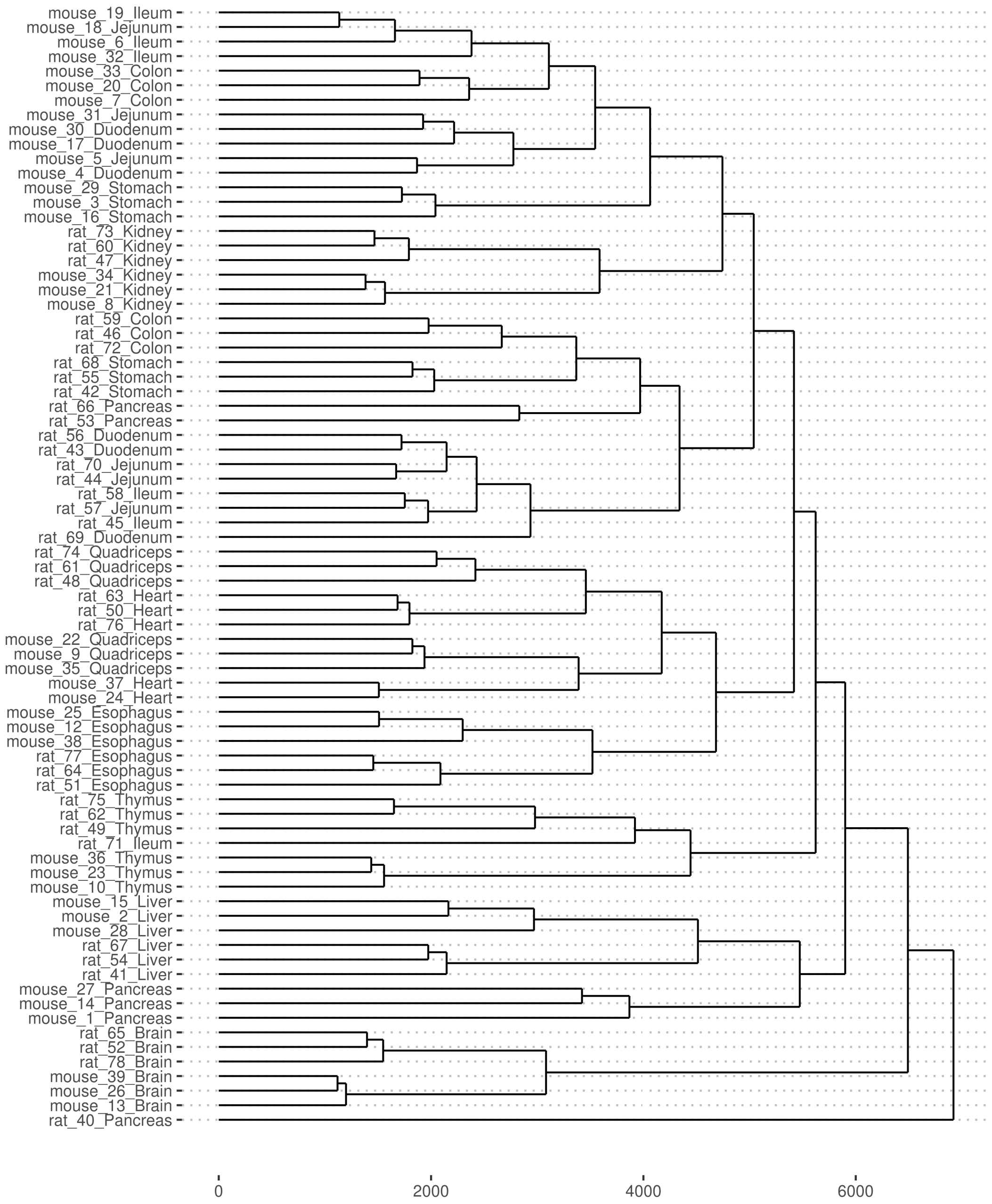


Figure 3.7: Hierarchical clustering of the combined voom-transformed log(counts per million) for one-to-one orthologues in mouse and rat. The Canberra distance between samples and the complete linkage method were used for clustering.

Interestingly, the similarity of gastrointestinal tissues is higher within each of the two species and the distance between the two GI clusters is relatively high. The rat pancreas samples also do not cluster with the corresponding samples in mouse.

3.3.3 Comparing gene expression and sequence identity

To investigate our hypothesis that highly conserved genes share gene expression patterns across tissues, we compared the expression correlation and the annotated sequence identity (see Figure 3.8). As expected, the majority of the orthologous gene pairs clustered in the upper right corner, i.e., both their protein sequences and their expression patterns across tissues were very similar. However, 11 % of the orthologues had a sequence identity greater than 80 % but a low positive (< 0.5) or even a negative expression correlation. A low expression correlation across tissues indicates that the gene was expressed in different tissues in the two species. In 4 % of the cases we observed lower protein sequence identity (< 80 %) while the expression correlation was high (> 80 %).

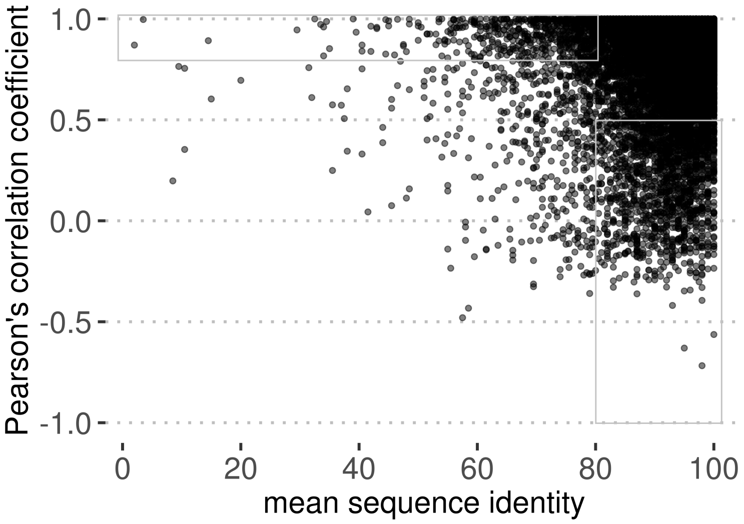


Figure 3.8: Comparison of sequence identity and expression correlation between mouse and rat one-to-one orthologues. The grey rectangles mark orthologous genes with either high sequence similarity and low expression correlation or low sequence identity and high correlation coefficients.

3.4 Discussion

In this chapter we have presented the results from the descriptive analyses of two RNA-Seq data sets from mouse and rat, each comprising 13 tissues. When analysing the two data sets separately, we observed that the difference in gene expression was greater between tissues than between animals. The biological similarity between tissues was also visible in our analyses as samples from muscular tissues (heart and quadriceps) were close to each other and samples from the gastroinestinal tract formed a bigger cluster which could be distinguished from other tissues. The tissue clusters were preserved when performing the same analysis on the combined data set, however, PC3 separated the two species from each other. Therefore, it would be interesting to investigate PC3 of the combined data in greater detail to determine the genes driving this differentiation. The hierarchical clustering confirmed that, in general, the tissue of origin has a greater effect on gene expression than the animal and its species. The analysis of the human transcriptome by Melé et al. [71] also found that expression varies more between tissues than between individuals. Furthermore, it would be surprising to see large differences between animals because we investigated gene expression in inbred strains.

The comparison of gene expression correlation and reported amino acid sequence identity of the integrated data showed that for the majority of one-to-one orthologous gene pairs our assumption that homologous genes share expression patterns across tissues holds true. Since homology refers to genes sharing a common ancestor [87], it is not surprising to find orthologous genes whose sequence is highly similar but their gene expression correlation is low. We hypothesise that these genes have developed diverging functions in mouse and rat and a more detailed investigation of the corresponding genes with respect to their biological function would be a valuable next step. As noted by Pearson [87], low sequence similarity is often observed for homologous genes. On the one hand, our observation that there are genes with low sequence identity but high expression correlation confirms this statement. On the other hand, some of these cases might also be caused by erroneous sequence information. In the next chapter we will investigate whether poor sequence annotation is an issue and how to deal with it.

Chapter 4

Exploiting orthology and *de novo* transcriptome assembly to refine target sequence information

The content of this chapter was partially published in:

> J. F. Söllner, G. Leparc, M. Zwick, T. Schönberger, T. Hildebrandt, K. Nieselt and E. Simon. *Exploiting orthology and de novo transcriptome assembly to refine target sequence information.* BMC Medical Genomics 2019: 12, 69

I implemented the Nextflow pipeline, performed the analyses and wrote the paper. G. Leparc designed the prototype of the underlying method. G. Leparc and M. Zwick provided advice regarding the implementation. M. Zwick reviewed the paper. T. Schönberger prepared the cynomolgus monkey samples. T. Hildebrandt directed the RNA preparation and sequencing of the samples and was involved in the design of the cynomolgus monkey study. K. Nieselt and E. Simon supervised the study and edited and reviewed the paper.

4.1 Background

Before a compound can proceed into the clinics where it is tested in humans, it has to prove its efficacy and safety in non-human species throughout the drug development process. Once a drug progresses towards the preclinical phase of the drug development pipeline, toxicological studies are performed to assess its safety and for dose selection. Although *in vitro* and *in silico* methods have gained increasing interest in recent years [5, 14, 34, 78, 82, 99], it is still indispensable to perform *in vivo* testing. To prevent unnecessary animal testing and failure of drug discovery projects in late stages of the pipeline, it is crucial to correctly assess the compound's activity on the target early on. Here established cellular and biochemical *in vitro* assays play an important role. By introducing a DNA template of the known target protein into bacteria or cell lines, one achieves the expression of the target protein in the host system. Errors in the introduced sequence may, however, negatively impact the interpretability of results as the compound's activity is over- or underestimated, the selected dose is too high or too low, and finally, *in vivo* experiments are misinterpreted.

Genes, transcript, or protein sequences are usually retrieved from public databases like Ensembl [142], UniProt [118], and RefSeq [83] which contain information for most model species that are of interest to biomedical research. One should, however, note that the knowledge regarding genome annotation and the availability of reliable sequences varies greatly between different species. This is emphasised by the comparison of the number of manually reviewed protein sequences in UniProtKB/Swiss-Prot (accessed: 26/04/2018) to the number of annotated transcripts in Ensembl (version 91). The high number of annotated transcripts and reviewed proteins indicates that human and mouse are the two species whose genomes have been investigated in greater detail than those of other species (see Figure 4.1). This is easily explained by the fact that mouse is a commonly used model species in biomedical research aiming to cure human diseases. Although non-human primates such as cynomolgus monkey (Macaca fascicularis) and rhesus macaque (Macaca mulatta) are important model organisms, their genomes are not as well characterised as those of human or mouse. Interestingly, there is also very little information available for the Chinese hamster genome even though Chinese hamster ovary (CHO) cells are important production systems for biopharmaceuticals such as monoclonal antibodies. The sequence information available through the public databases originates from a combination of sophisticated computational pipelines and expert curation. One example is the NCBI's automated genome annotation pipeline which builds its predicted gene models based on the alignment of known protein sequences, transcripts, and RNA-Seq data to a reference genome [119]. Existing sequences from related species are also considered for in the alignment process. This obviously results in an annotation bias as there is more reliable information available for well characterised species, like human or mouse, which leads to a more extensive and more trustworthy genome annotation than in less well characterised species. Therefore, the sequence information stored in the databases may be incomplete or erroneous [11].

There are also tools that make use of transcriptome assemblies to derive gene models. Scipio [56], for example, aligns query proteins to a reference genome to determine the exon-intron structure as well as splice sites in the corresponding gene. MIKADO [131] follows a different approach, it aims to improve transcript models by combining multiple transcriptome assemblies. Both tools focus on the gene's structure.

Instead of determining a gene's overall structure, we want to reconstruct the protein sequence of an individual target by exploiting orthology relationships and RNA-Seq reads. Therefore, we implemented the a&o-tool. It uses RNA-Seq data

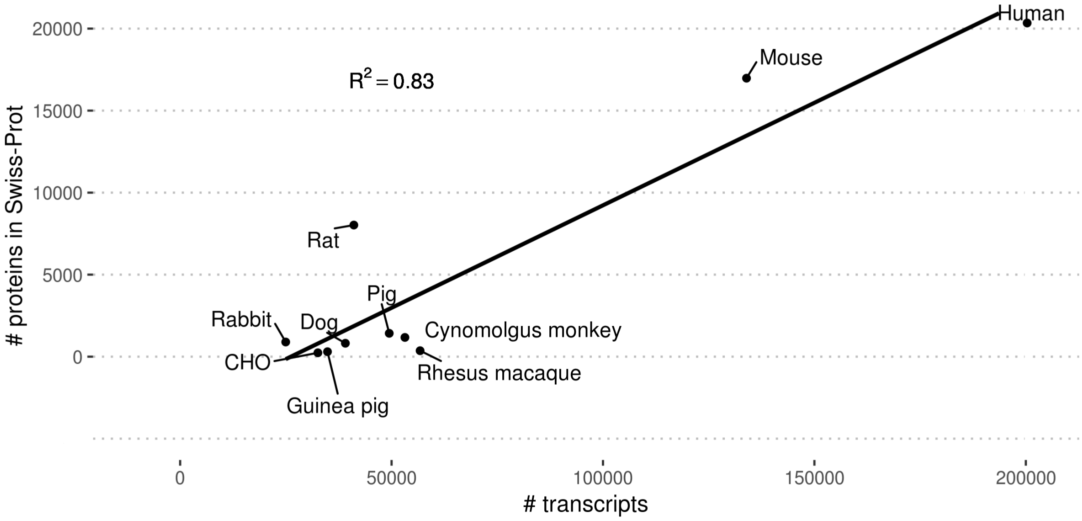


Figure 4.1: The x-axis corresponds to the number of annotated transcripts in Ensembl while the y-axis shows the number of manually reviewed protein sequences in UniProtKB/Swiss-Prot that are available for some of the commonly used model species in biomedical research. R^2 and the regression line emphasise the correlation between the two numbers. CHO: Chinese Hamster Ovary cells.

from the species of interest and a bait protein sequence from a related species to first compute a *de novo* transcriptome assembly and to then search the assembled contiguous sequences that best match the bait. The target's protein sequence is derived by searching open reading frames (ORFs) in the best matching contig and translating these into an amino acid sequence. A multiple sequence alignment as well as descriptive metrics based on pairwise alignments offer the possibility to asses the quality of the refined sequence.

We begin this chapter by investigating how many sequences are presumably poorly annotated in five model species commonly used in biomedical research. The general idea to use RNA-Seq data and an orthologous protein to refine such poorly annotated sequences, is then evaluated via a reciprocal best hit BLAST approach. RNA-Seq reads from all five species and human were assembled into tissue-specific transcriptomes (brain, liver, and kidney) and the 20,350 manually reviewed human protein sequences in UniProtKB/Swiss-Prot (from now on referred to as "known human protein sequences") were used as bait sequences. Finally, we introduce the a&o-tool, an automated sequence refinement pipeline, and apply it to a set of presumably poorly annotated sequences to assess its performance.

4.2 Methods

4.2.1 Data description

The analyses in this chapter are based on paired-end RNA-Seq raw read data from two publicly available and one internally generated data sets (see Table 4.1). For mouse, rat, dog and pig we used data published by Fushan et al. [37] which includes samples from brain, kidney and liver. The Human Protein Atlas [128, 129] provides human RNA-Seq data from various tissues including those available in the Fushan et al. data. The internal data set consists of data from two cynomolgus monkeys (*Macaca fascicularis*) and contains samples from brain, kidney and liver, amongst other tissues. Raw read data were processed the same way as in Chapter 3 to obtain normalised expression values. For all six species—human, mouse, rat, dog, pig, and cynomolgus monkey—we retrieved manually reviewed protein sequences from UniProtKB/Swiss-Prot (accessed on June 9th, 2018).

Table 4.1: RNA-Seq data sets used for the analyses.

species	data source	RNA-Seq details
human	Uhlén et al.	17×10^6 reads per sample
mouse, rat, dog, pig	Fushan et al.	15×10^6 reads per sample
cynomolgus monkey	internal	55×10^6 reads per sample, RIN median 8.7, 2x85 bp on HiSeq3000

4.2.2 Proportion of genes to be improved

We queried all orthologous gene pairs between human and the other five species using the biomaRt [27, 28] R [93] package (version 2.32.1) with Ensembl version 92. The target and query protein sequence identities (see Figure 3.1) of these gene pairs were used to calculate the difference in sequence identity as:

$$\Delta seq\_id = target\_identity - query\_identity \tag{4.1}$$

The difference Δseq_id was used to determine genes in the non-human species that showed a diverging protein sequence compared to their human one-to-one orthologue. We hypothesised that if, for example, 99 % of the orthologous protein sequence matches the human protein but the human sequence only matches the orthologous protein sequence in 78 % of its amino acids, the orthologous sequence might be incomplete or contain errors. A query identity that is higher than the target identity could, for example, indicate that the orthologous protein is too long which may be the result of a missing stop codon.

To estimate the number of potentially incomplete or incorrect sequences, we computed Δseq_id for all one-to-one orthologous gene pairs between human and the other species (mouse, rat, dog, pig, and cynomolgus monkey). A gene in the non-human species was called “affected” if its absolute Δseq_id was greater than the species’ mean + 2 times standard deviation. We are aware that this threshold is rather conservative. Our aim was to detect genes with the most significant deviation from their human orthologue and based the choice of the threshold on the distribution of Δseq_id (see Figure A.1).

Of course, differing protein sequences are also the result of divergence during evolution and these cases should not be considered for further analysis. To rule out that the difference in sequence identity reflects evolutionary events instead of poor annotation, we filtered for genes with a high absolute Δseq_id in only one of the five non-human species, i.e., in the other four species the protein sequence was conserved.

4.2.3 *De novo* transcriptome assemblies

BinPacker [65] (version v1.0) and rnaSPAdes, which is part of the SPAdes package [13] (version 3.11.1), were used for *de novo* assembly of tissue-specific transcriptomes. Both assemblers were run with their default parameters.

We compared the two assemblers based on TransRate [107] metrics and found that the rnaSPAdes assemblies contained a high number of contigs with less than 200 bp. These short contigs may result in the a&o-tool mainly detecting fragments instead of entire transcripts. Together with the observation that the coverage of known human sequences was higher with BinPacker, this led to the decision to use BinPacker for further analyses and in the a&o-tool.

4.2.4 Evaluation of detection rates in human and related species from RNA-Seq assemblies

As we aimed to use protein sequences of closely related species to derive the corresponding protein sequence from the assembled transcriptome of a poorly annotated species, we first validated this approach by using curated human UniProtKB/Swiss-Prot and assemblies from human RNA-Seq data.

A reciprocal best hit (RBH) BLAST search (see Figure 4.2) was used to determine the best matching contig in the assembled transcriptome for a given human bait protein. In the forward step, we determined the best matching contig by aligning each human protein sequence to all assembled contigs (tblastn, NCBI BLAST+ [6] version 2.7.1). In the backward search, the best hits were then aligned back to all known human sequences (blastx, NCBI BLAST+ version 2.7.1). All BLAST parameters were left at their default values, except for the e-value threshold which we set to 1×10^{-4} to remove highly insignificant highest scoring pairs (HSPs) from the results.

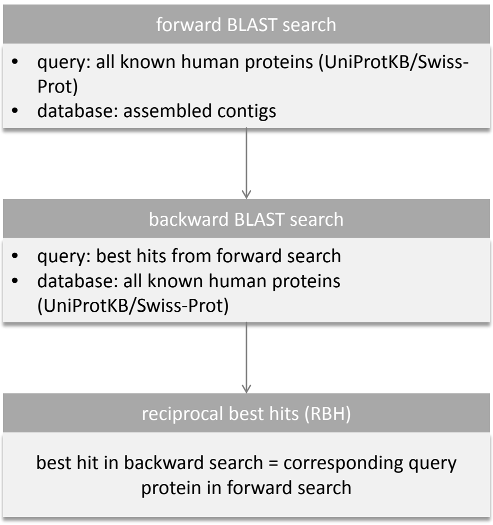

forward BLAST search
• query: all known human proteins (UniProtKB/Swiss-Prot)
• database: assembled contigs
backward BLAST search
• query: best hits from forward search
• database: all known human proteins (UniProtKB/Swiss-Prot)
reciprocal best hits (RBH)
best hit in backward search = corresponding query protein in forward search

Figure 4.2: Workflow diagram of the reciprocal best hit BLAST approach. In the forward search all known human proteins were searched in the assembled contigs. For each query proteins the best hit was determined and used as query in the backward BLAST search against all known human proteins. Based on the BLAST results the proteins with a reciprocal best hit were determined.

The detection rate, i.e., the rate of RBHs, was calculated as the proportion of known human proteins that led to the correct human protein as the best hit in the backward search.

We hypothesised that proteins which were not found as RBH (non-RBH) showed a lower gene expression than those found as an RBH. To investigate this, we compared gene expression levels of these two groups. The biomaRt R package and Ensembl (version 92) were used to map human UniProt accession numbers to Ensembl gene identifiers and to query orthologous Ensembl gene identifiers.

We also examined the sequence identity from the BLAST results as well as the coverage of the human protein sequence. The coverage was calculated as the proportion of the number of amino acids in the human protein covered by the alignment (alignment end - start position) in relation to its length.

The detection rates were also determined for mouse, rat, dog, pig, and cynomolgus monkey. Tissue-specific assemblies from each species were used to construct separate databases for the forward (tblastn) search to which the known human protein sequences were aligned. In the backward (blastx) search the best hits were aligned to the database of all known human proteins.

4.2.5 Generalised refinement pipeline

The reciprocal best hit BLAST approach described above was used to evaluate whether a protein sequence from a closely related species can reliably identify the corresponding protein in the transcriptome assembly of a species of interest. As BLAST HSPs do not necessarily represent the entire contig and the RBH approach requires a comprehensive set of curated protein sequences, which is not available for most model organisms, we have implemented a generalised and automated refinement pipeline, the a&o-tool (see Figure 4.3). For the successful application of the a&o-tool it is crucial that the input meets two criteria: 1) The RNA-Seq data is of high quality and originates from paired-end sequencing of samples in which the target protein's transcript is expected to be expressed at reasonable levels. Alternatively, a high-quality, pre-computed transcriptome assembly is provided. 2) A reliable orthologous protein sequence (referred to as "bait") is available.

If paired-end RNA-Seq reads are provided, a *de novo* transcriptome assembly is generated using BinPacker (version v1.0; with the default parameters). Alternatively, a pre-computed assembly can serve as the entry point for the pipeline. We recommend to pre-compute the transcriptome assembly in case the same RNA-Seq data are to be used for repeatedly running the a&o-tool with different bait sequences. Firstly, the assembly process is computationally intense and secondly, the quality of the assembly might be improved by trying different assemblers and/or parameter settings.

The transcriptome assembly is then used as the database for a tblastn (NCBI BLAST+ version 2.7.1) search in which the orthologous bait sequence is aligned to the assembly. We sort the BLAST result by bitscore and e-value to determine the best matching assembly contigs. The n best matching contigs are extracted and their cDNA sequence is fetched from the transcriptome assembly. n is a user-defined parameter. By setting it to n=1, only the best matching contig is chosen, values greater than one lead to multiple hits that are processed in the following steps of the pipeline. For assemblies derived from short-read RNA-Seq reads, we recommend $n>$ 1 (default: 5) as transcripts are often represented by multiple contigs that were not combined properly during the assembly process.

We use TransDecoder [42] (version 5.2.0) to search the n resulting cDNA sequence(s) for open reading frames and, for each of them, we chose the translation of the longest ORF into an amino acid sequence as the output protein.

To validate the resulting protein sequence(s), we use visualisation of a multiple sequence alignment (MSA) as well as quantitative metrics. In the MSA we align the resulting protein sequence(s) with the bait sequence and optional orthologous proteins from other related species. If an annotated version of the target sequence is available, it can also be included in the set of optional orthologues to allow for the comparison of the annotated and the refined sequence(s). The a&o-tool uses MUSCLE [29] (version 3.8.31) to compute and visualise the MSA.

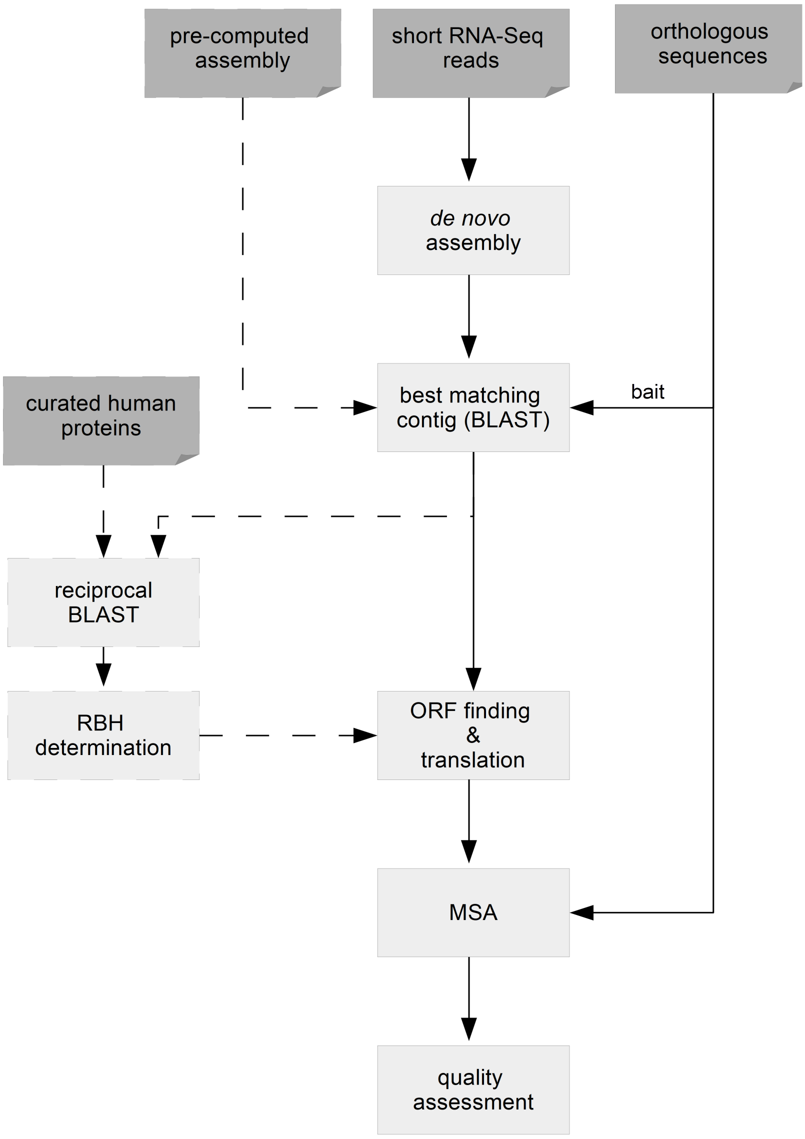


Figure 4.3: Schematic overview of the a&o-tool, an automated sequence refinement pipeline. Dark grey boxes mark input files while light grey boxes represent processing steps. Dashed lines indicate alternative processes. Entry points for the pipeline are either short-read RNA-Seq data which are assembled into a transcriptome or a pre-computed transcriptome assembly. The FASTA file of orthologous sequences must contain a bait sequence which is searched in the (pre-computed) assembly to retrieve best matching contigs. If the bait protein is from human, the curated human UniProtKB/Swiss-Prot sequences may provided as an additional input that is used for a reciprocal best hit BLAST search. In the resulting contigs, ORFs are determined and translated into protein sequences. These, as well as the orthologous sequences including the bait, are visualised in a MSA and quality control metrics are computed. All required and optional input is provided via a JSON file. MSA: Multiple sequence alignment; ORF: Open reading frame; RBH: Reciprocal best hit; BLAST: Basic local alignment search tool.

If an annotated protein sequence of the target is available, we perform two alignments: 1) The annotated sequence is aligned to the bait protein and 2) the refined sequence(s) are aligned to the bait protein. In both cases we compute Δseq_id. If an annotated sequence is provided, we compare the resulting Δseq_id values to determine whether the refinement has led to a decreased difference in sequence identity. If no existing sequence is available, we only report Δseq_id for the alignment of the refined sequence(s) to the orthologous bait protein.

Nextflow [25] was used to automate the procedure described above. A configuration file in JSON format provides all required information regarding the pipeline input and its parameters. One can either pass the paths to the paired-end RNA-Seq reads or the path to a pre-computed assembly. A FASTA file containing the orthologous bait protein sequence as well as other optional sequences, which should be included in the MSA, has to be provided.

The software required by a&o-tool (BinPacker, BLAST+, MUSCLE, and Python) are provided via a Docker [72] container. To make the pipeline easily accessible, we have set up a GitHub repository (https://github.com/Julia-F-S/a-o-tool) that also includes some example data.

4.3 Results

4.3.1 Proportion of genes to be improved

The number of potentially poorly annotated genes in the five non-human species ranged from 474 in dog (3 % of all dog genes which have one-to-one orthologues in human) to 259 in mouse, i.e., 1.5 % of all mouse genes which have one-to-one orthologues in human (see Figure A.2).

4.3.2 *De novo* transcriptome assemblies

The assembly of RNA-Seq data from six species and the three tissues brain, liver, and kidney, resulted in 18 tissue-specific transcriptomes. The mean contig length in the human assemblies was 1,369 bases (mean across assemblies from all tissues) and 29.2 % of the contigs (again mean across all three assemblies) contained an open reading frame (ORF). The mean percentage of the contig covered by the ORF was 43.1 %. Details on the quantitative metrics for all 18 assemblies can be found in Table A.1.

4.3.3 Evaluation of detection rates in human and related species from RNA-Seq assemblies

A reciprocal best BLAST search was applied to determine whether we can use RNA-Seq data to improve or validate existing protein sequences.

In the tissue-specific transcriptomes of human, 64 % of all known human proteins were detected as RBH (mean across tissues; see Figure 4.4). The average number of genes with an FPKM ≥ 1, a lower boundary for genes to be considered as expressed [128], was 14,265 (70 % of the 20,350 human proteins in UniProtKB/Swiss-Prot). Therefore, we concluded that the majority of expressed protein-coding genes was detected by the RBH approach.

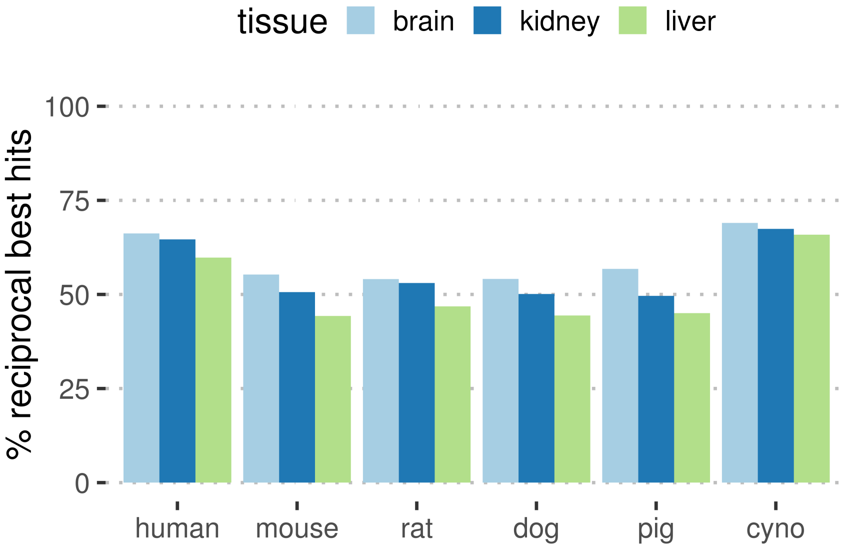
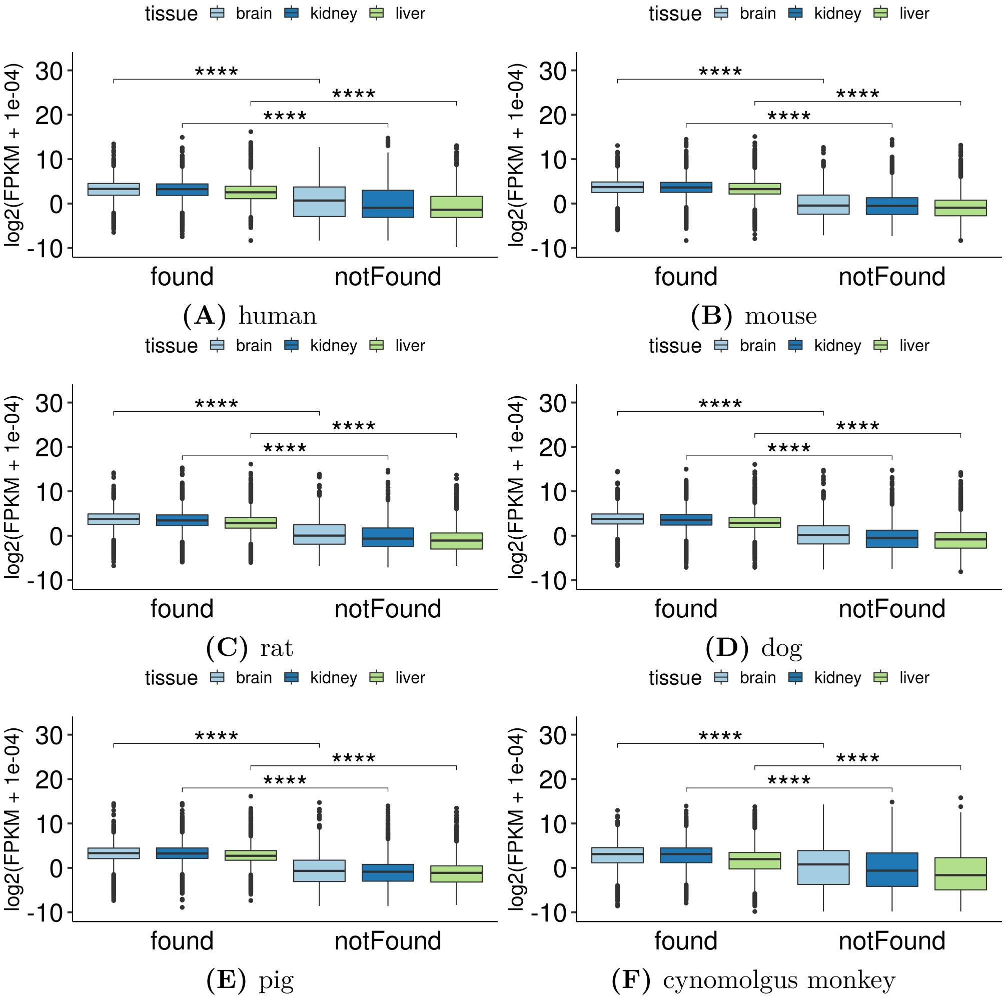


Figure 4.4: Percentage of all 20,350 known human proteins that had an RBH in the respective assembly. Bars indicate the detection rate resulting from the search with the assembled transcriptome from the individual species and tissue.

The reciprocal BLAST search with all known human sequences and the tissue-specific transcriptome assemblies from mouse, rat, dog and pig resulted in a lower detection rate of around 50 % (see Figure 4.4). In comparison to the search of the known human proteins in the human assemblies, the reciprocal BLAST search with the known human proteins and the tissue-specific cynomolgus monkey assemblies led to an increase in detection rate of about 3.9 % (mean increase across tissues).

To investigate our hypothesis that proteins without an RBH are not represented in the transcriptome because they are lowly or not expressed, we compared the expression levels of genes associated with proteins leading to an RBH and those not having an RBH. This comparison confirmed our hypothesis (Wilcoxon rank sum test: p-value $<1 \times 10^{-4}$ for all tissues in all species; see Figure 4.5). Interestingly, we observed that there are also genes associated with proteins without an RBH that are highly expressed. A more detailed investigation of these genes revealed that they code for multiple proteins which are represented by different UniProt accession numbers. One example is the human GNAS complex locus (Ensembl identifier ENSG00000087460): Its median expression in kidney is FPKM > 330 and it is associated with the UniProt accession numbers Q5JWF2, P84996, O95467, and P63092. We do, however, only find an RBH in the kidney transcriptome for Q5JWF2, P84996, and O95467 but not for P63092. In summary, the reciprocal best hit BLAST analysis confirmed that an orthologous protein sequence can be used to reliably detect the corresponding orthologous sequence in the transcriptome assembly of a species of interest.



(A) human

(B) mouse

(C) rat

(D) dog

(E) pig

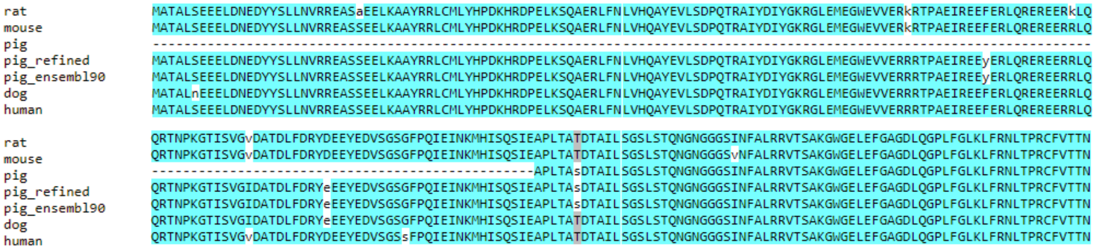
(F) cynomolgus monkey

Figure 4.5: Expression levels of proteins with a reciprocal best BLAST hit (found) and those without (notFound). A Wilcoxon rank sum test was used to determine pairwise significance (****: p-value ≤ 0.0001).

4.3.4 Application of the generalised refinement pipeline

With the a&o-tool we present an automated sequence refinement pipeline that can be used to refine or validate a target protein's sequence (see Figure 4.3).

An example for the successful application of the a&o-tool is the pig orthologue of the human protein DnaJ homolog subfamily C member 11 (DJC11_HUMAN, UniProt Accession Q9NVH1). In comparison to its human orthologue, the pig sequence (Ensembl version 88, ID ENSSSCP00000003669.2) was shorter (169 amino acids missing at N-terminus) while the human sequence was well conserved in mouse, rat, and dog (>94 % protein sequence identity). By applying the a&o-tool, we were able to retrieve a full-length protein sequence that matched the human protein almost perfectly. The resulting sequence was further supported by an update of the pig sequences with Ensembl (version 90) as it matched the sequence of ENSSSCP00000003669.3 (see Figure 4.6). To determine whether the a&o-tool is able to improve the sequence information of a variety of proteins, we applied it to the set of pig proteins that are presumably poorly annotated (see Figure A.2) with human proteins as the orthologous bait sequences. For 73 of the 293 proteins we did not get any result because they either did not have a hit in the initial BLAST search or the detected contig did not contain an ORF. The remaining 220 proteins were filtered for those having an RBH. Furthermore, we excluded proteins from further analysis for which the query identity or the target identity differed by more than 3 % between the values provided by Ensembl and those derived from the alignment with the Swiss-Prot sequence. This led to 131 proteins for which we examined the results of the a&o-tool and found that a decrease in Δseq_id was achieved for 98 proteins (mean decrease: 19.5 %; see Figure 4.7).



```
rat             MATALSEEELDNEDYYSLLNVRREASaEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERkLQ
mouse           MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERkRTPAEIREEFERLQREREERRLQ
pig             ------------------------------------------------------------------------------------------------------------------------
pig_refined     MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEyERLQREREERRLQ
pig_ensembl90   MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEyERLQREREERRLQ
dog             MATALnEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ
human           MATALSEEELDNEDYYSLLNVRREASSEELKAAYRRLCMLYHPDKHRDPELKSQAERLFNLVHQAYEVLSDPQTRAIYDIYGKRGLEMEGWEVVERRRTPAEIREEFERLQREREERRLQ

rat             QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
mouse           QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAILSGSLSTQNGNGGGSvNFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig             -------------------------------------------------APLTAsDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_refined     QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTAsDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
pig_ensembl90   QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTAsDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
dog             QRTNPKGTISVGIDATDLFDRYeEEYEDVSGSGFPQIEINKMHISQSIEAPLTATDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
human           QRTNPKGTISVGvDATDLFDRYDEEYEDVSGSsFPQIEINKMHISQSIEAPLTATDTAILSGSLSTQNGNGGGSINFALRRVTSAKGWGELEFGAGDLQGPLFGLKLFRNLTPRCFVTTN
```

Figure 4.6: Multiple sequence alignment of the human protein DnaJ homolog subfamily C member 11 and its orthologous protein sequences in rat, mouse, pig, and dog. pig_refined corresponds to the refined sequence and pig_ensembl90 is the updated pig sequence released with Ensembl 90. The figure has been published in [110] under the Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`).

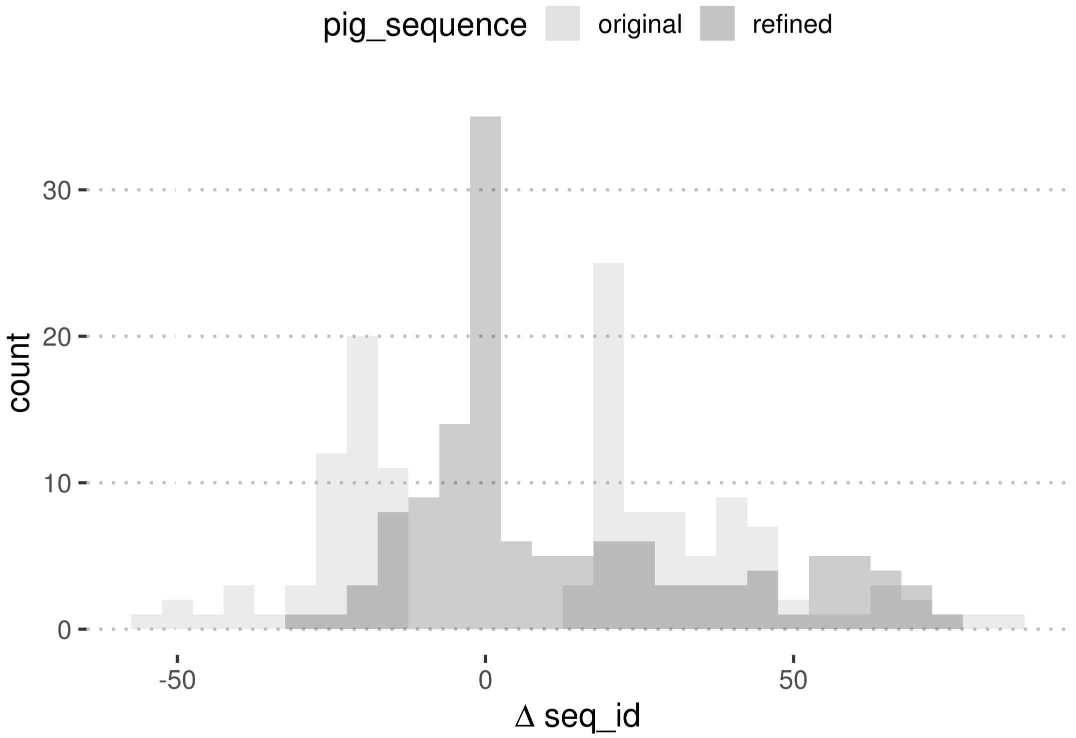


Figure 4.7: The x-axis shows the difference between target and query sequence identity, Δseq_id, and the y-axis corresponds to the number of genes with a certain Δseq_id. Light grey bars show the distribution of Δseq_id for all presumably poorly annotated pig proteins, i.e., those with $|\Delta seq\_id|$ greater than the specie's mean + 2 times standard deviation. Dark grey bars show the distribution after the application of the a&o-tool and illustrate the overall reduction of Δseq_id.

4.4 Discussion

In this chapter we investigated whether *de novo* assembled transcriptomes from paired-end RNA-Seq data and an orthologous protein sequence can be used to reliably identify the corresponding sequence in a species of interest. Furthermore, we introduced the a&o-tool, an automated pipeline to refine or validate incomplete or erroneous protein sequences by exploiting orthology and *de novo* transcriptome assembly.

The evaluation of detection rates in a reciprocal best hit BLAST search using human protein sequences and *de novo* assembled tissue-specific transcriptomes from a variety of species—human, mouse, rat, dog, pig, and cynomolgus monkey—showed that this approach leads to high detection rates which are in line with the proportion of expressed proteins. The Human Protein Atlas webpage, for example, states that 59 % of all examined proteins are expressed in liver, 68 % in kidney and 74 % in brain [2].

When interpreting the hit rates in the RBH analysis, one should keep in mind that UniProtKB/Swiss-Prot typically does not contain all isoforms of a gene, i.e., it is a simplified collection that does not appropriately account for the fact that most genes have several isoforms that are often expressed tissue-specifically. Furthermore, most sequences that were not detected corresponded to lowly expressed gene. These two reasons may explain the slightly lower hit rates in comparison to the expected number of expressed genes.

It should also be noted that, although UniProtKB/Swiss-Prot contains a comprehensive set of well curated human proteins, it does not represent an exhaustive set of all human proteins. Pertea et al. [90] have analysed the GTEx [117] data and, as a result, released a new gene and transcript catalogue that adds roughly 1,000 protein-coding genes to those present in UniProtKB/Swiss-Prot.

We observed differences in the achieved hit rates depending on the source of the RNA-Seq data used for the assembly. The highest hit rates were obtained with the cynomolgus monkey assemblies followed by human. In the four species for which sequencing data were published by Fushan et al. (mouse, rat, dog, and pig), we saw the lowest proportion of RBHs. These differences are probably due to varying quality of the input RNA used for sequencing and the resulting read quality. The RNA from cynomolgus monkey had an RNA integrity number (RIN; median across samples) of 8.7 and 4.96×10^7 sequencing reads were uniquely mapped to the reference genome (median across samples). With a median RIN of 7.5 the quality of the Fushan data for mouse, rat, pig, and dog was significantly lower and as a result only 1.36×10^7 (median of per species medians) reads were uniquely mappable. We did not have RIN values for the Human Protein Atlas data but with 1.15×10^7 the number of uniquely mapped reads (again median across samples) was even lower.

The RBH analysis also led to the conclusion that using human proteins as bait sequences is a valid approach if one does not have access to a reliable sequence from a more closely related species. We do, however, recommend to choose the bait protein from a species with minimum evolutionary distance to the species of interest to rule out that the sequence is missed due to evolutionary divergence.

Here we used short-read RNA-Seq data for *de novo* transcriptome assembly. One should bear in mind that, due to the inherent properties of short-read sequencing data, the contigs in the assembled transcriptome do not necessarily correspond to full-length transcripts. This may lead to the identification of a contig which belongs to the correct transcript but which is too short. As a result, the contig cannot be translated into a complete protein sequence because, for example, the initiation codon is missing. Another scenario one might be faced with is that the resulting protein is too long due to 5' UTRs containing initiation codons that belong to upstream open reading frames [136].

To solve these two issues, we suggest to remove short contigs or ones hardly covered by reads from the assembly by adding a filtering step to the a&o-tool. One could also replace or combine the short-read RNA-Seq data with long-read data from, for example, PacBio or Oxford Nanopore sequencing. These platforms generate supposedly full-length transcripts and could therefore eliminate the error-prone assembly step.

Regarding the application of the a&o-tool to targets from families of highly similar proteins, we advise special caution when interpreting the results. For similar bait proteins our tool might identify the same contig as best match and thus return the same translated sequence for different members of a protein family.

After the thorough evaluation of the general approach behind the a&o-tool, we applied the tool to all pig genes that were found to be potentially poorly annotated. We observed a shift of the median difference in sequence identity towards 0. Therefore, we concluded that, for the majority of proteins, the sequence information could be improved and that the target-centric a&o-tool facilitates target identification and validation.

Finally, it may be concluded that RNA-Seq data and orthologous protein sequences from closely related species can be used to improve or validate protein sequences in species with poor annotation. With the a&o-tool we automated this approach and provide a tool that can be easily incorporated at different stages of the drug development process.

Chapter 5

A pilot study for the comparison of long- and short-read RNA-Seq data

5.1 Introduction

In recent years long-read sequencing technologies, such as those developed by Pacific BioSciences and Oxford Nanopore, have gained increasing attention for genome and transcriptome sequencing. While Illumina's short-read sequencing has become very cost-efficient and easily accessible, it currently only generates a maximum read length of 300 bp [51]. In case of RNA-Seq this is sufficient for quantitative gene expression analysis but in order to investigate transcript sequences and alternative splicing events, the reads first have to be assembled into contiguous sequences that ideally correspond to the whole transcripts. The quality of the resulting genomic or transcript sequences depends on the chosen assembler and may suffer from assembly errors. With PacBio's Sequel systems performing single-molecule real-time sequencing, an average read length of 46 kb is achieved [84]. Especially in case of RNA sequencing, this drastic increase in read length has revolutionised the field, as it is now possible to sequence entire transcripts and the assembly step is rendered obsolete for basically all transcripts. For example, protein-coding rat transcripts annotated in Ensembl version 92 had a mean length of 2,254 bp and a maximum length of 25,658 bp.

There are already numerous studies that investigate alternative splicing on single-gene [16,127] or whole-transcriptome level using long-read sequencing. For example, Anvar et al. [9] have used the human breast cancer cell line MCF-7 and three human tissues to examine the interplay of transcription initiation, splicing, and polyadenylation. While the majority of long-read sequencing technologies are primarily applied within the research community, they are slowly advancing to clinical applications [7].

In this chapter we present the results of a pilot study comprising one rat liver sample which has been sequenced on PacBio's Sequel system. To assess the quality and compare results between long- and short-read sequencing, we also sequenced this sample, as well as additional samples, on an Illumina NextSeq500. The main questions we wanted to answer with this pilot study were, 1) how many of the annotated transcripts are detected, 2) how do the detected sequences compare to well-curated data, 3) are the PacBio sequences supported by short-read data and if yes, how well, 4) are lowly expressed transcripts detected with long-read sequencing?

5.2 Methods

5.2.1 PacBio and Illumina RNA-Seq data

For this study, samples were collected from three tissues (brain pre-frontal cortex, kidney cortex and liver) from four male Brown Norway rats. All 12 samples were split into aliquots for short-read sequencing with Illumina. These aliquots were sequenced on a NextSeq500 to obtain paired-end reads with a read length of 76 bases. For the long-read sequencing on a PacBio Sequel, a pilot study with just one liver sample (sample id 677_3) was conducted.

PacBio reads were processed with SMRTLink (v6.0.1) and the Ensembl *Rattus norvegicus* reference genome (Rnor_6.0) according to the PacBio Iso-Seq protocol (see Figure 2.3B) by the sequencing laboratory c.ATG. The report also contained some information generated during the sequencing process. For example, the SMRT cells were loaded using the diffusion method with 4 nM and a movie time of 600 min was used. Furthermore, the following statistics on the productivity of the zero-mode waveguides were provided: P0 refers to the percentage of ZMWs which did not produce a polymerase read, P1 is the percentage of ZMWs that produced exactly one polymerase read and P2 is the percentage of ZMWs which produced inconsistent signals.

A Boehringer in-house RNA-Seq pipeline was used to process the Illumina reads. Quality control of FASTQ files was performed with FastQC [15] (version v0.11.5). Samples were aligned to the Ensembl *Rattus norvegicus* reference genome (Rnor_6.0) using STAR [26] (version 2.5.2b). The resulting BAM files were subjected to further quality control with picardmetrics [106] (version 0.2.4). dupRadar [104] (version 1.2.2) was run to assess duplication rates. Duplicated reads were, however, not removed.

The BAM files were sorted with SAMtools [4] (version 1.7) and transcripts were quantified with RSEM [64] (version 1.3.0) to obtain their expression as transcripts per million (TPM) and fragments per kilobase million (FPKM). In addition to RSEM, featureCounts from the subread package [108] (version subread-1.5.1) was used to summarise read mappings. Based on the variance-stabilised counts we performed a principal component analysis (PCA) using DESeq2 [66] to detect potential outliers and investigated the distribution of expression values.

As the liver samples showed a high number of lowly expressed genes, we used the independent rat data published by Fushan et al. [37] to determine liver-specific genes according to the following specificity score (unpublished work by Eric Simon): Given a set of tissues $\mathbf{T}$ of length t and a vector of gene expression values $\mathbf{e} = [e_1, e_2, \ldots, e_t]$ we computed the arithmetic mean expression across all tissues as:

$$\mu = \frac{\sum_{i=1}^{t} e_i}{t} \tag{5.1}$$

The specificity vector, **spec** is then computed as:

$$spec_i = \frac{e_i - \mu}{e_i + \mu + 10^{-9}} \tag{5.2}$$

Where each element is the tissue-specificity of the gene for tissue$_i$ $\in \mathbf{T}$. To determine whether a gene is specifically expressed in a tissue one could either define an arbitrary threshold or use the tissue with the highest specificity score among the examined ones. Later in this chapter, we are going to use the latter approach as we did not want to rely on an arbitrarily chosen threshold.

5.2.2 Comparing PacBio isoforms to the rat genome

SQANTI [114] was applied to compare PacBio isoforms to the Ensembl reference genome (Rnor_6.0) and the corresponding annotation. The tool first maps the PacBio isoforms to the provided reference genome using GMAP [138] to determine splice junctions. These detected splice junctions are then compared to annotated splice junctions and transcripts are classified into structural categories accordingly (see Figure 5.1). Full splice matches (FSM) are transcripts that perfectly match a reference transcript with respect to its splice junctions. An incomplete splice match (ISM) is a fragment of an annotated transcript, i.e., it matches parts of the reference perfectly but does not contain all splice junctions. If PacBio isoforms can be mapped to a reference gene but use known splice junctions to form an unknown transcript, they are classified as novel in catalogue (NIC). In case an unknown transcript of an annotated gene arises from splice junctions which are not annotated, the PacBio isoform is referred to as novel not in catalogue (NNC). Then there transcripts which are referred to as “novel genes”. These are further classified into: Genic Intron (lying in a gene’s intronic region), Genic Genomic (overlapping intronic and exonic region), Intergenic (located between to genes), Fusion Transcripts, and Antisense (matching the complementary sequence of an annotated gene).

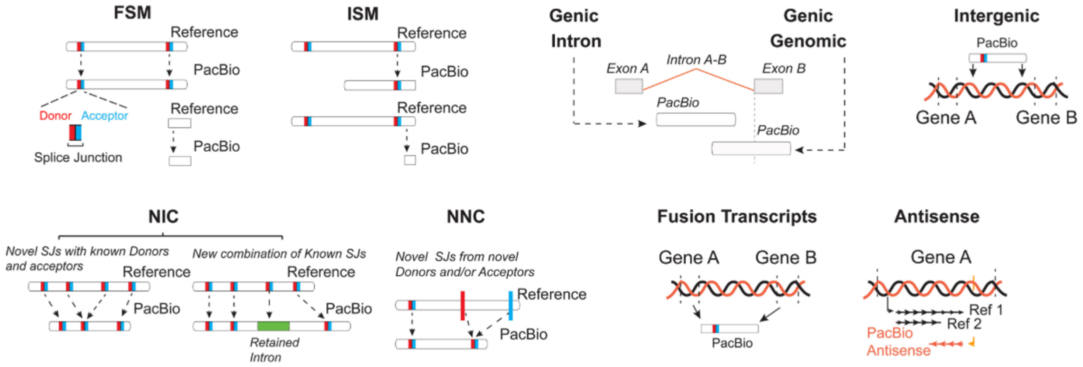


Figure 5.1: SQANTI transcript classification. FSM: Full splice match, ISM: Incomplete splice match, NIC: Novel in catalogue, NNC: Novel not in catalogue, SJ: Splice junction. This figure is part of Figure 1 in Tardaguila et al. [114] which has been published under the creative commons license 4.0 `https://creativecommons.org/licenses/by/4.0/legalcode`.

We also compared the number of transcripts per gene derived from SQANTI results to that from Illumina RSEM results and investigated alignments using the Integrative Genomics Viewer (IGV) [96, 120] (version 2.3.98).

5.2.3 Comparing PacBio isoforms to known sequences from rat and human

We aligned the PacBio isoforms to all rat and human protein sequences in UniProtKB/Swiss-Prot and the larger set of rat and human cDNA sequences available in Ensembl. In both cases a Nextflow pipeline performing a reciprocal best hit BLAST approach was used. PacBio isoforms were aligned to the human amino acid sequences with blastx and to nucleotide sequences with blastn. The best hits were then aligned back to the set of PacBio isoforms using tblastn (proteins) or blastn (transcripts). All BLAST parameters, except for the e-value threshold which was set to 1×10^{-4}, were used with their default values implemented in NCBI BLAST+ version 2.7.1. The results were investigated with respect to whether a reciprocal best hit was found as well as the sequence identity of BLAST HSPs and the coverage of the query sequence. Furthermore, we incorporated the results from SQANTI to compare the expression between isoforms which had a reciprocal best hit and those which did not. We also contrasted the classification of isoforms into known and novel genes/transcripts with their RBH property.

5.2.4 Comparing PacBio and Illumina data

Mapping rates

Illumina reads obtained from sample 677_3 were assembled with StringTie [89] using the BAM file generated by STAR in the in-house RNA-Seq pipeline and the Ensembl 92 reference annotation. As a result, StringTie generates a GTF file containing annotated and novel features. For PacBio isoforms such an annotation file was produced by SQANTI. Using gffcompare [88] (version 0.11.2) we compared both GTF files to the Ensembl reference annotation and retrieved sensitivity and precision values at several levels, e.g., base, exon, and intron level. We did not set the -R option of gffcompare, therefore, the sensitivity values were not adjusted, i.e., reference genes not supported by any data were also included in the calculations.

Coverage of PacBio isoforms by Illumina reads

To compare the data we obtained from the Illumina sequencing runs to those from PacBio sequencing, we merged all FASTQ files from the four liver samples and mapped them to the PacBio isoforms using STAR. Each PacBio isoform was considered to be a reference sequence, i.e., a “chromosome”. We applied SAMtools to index the resulting BAM file and computed the number of mapped reads per reference sequence.

We applied the same procedure to the Illumina reads from the single liver sample 677_3 to get a direct comparison between Illumina and PacBio data obtained from the same sample. Furthermore, we also aligned the short reads from kidney and brain to the PacBio isoforms to assess whether their coverage of the PacBio liver isoforms differs from that with short-read data from liver.

Detection limit

To get an idea for the expression threshold sufficient for detecting a transcript with PacBio sequencing, we compared transcript expression, based on RSEM results, between transcripts that were associated with a PacBio isoform and those that were not.

Overlap of detected transcripts

We compared the set of transcripts associated with PacBio isoforms by SQANTI to the set of transcripts expressed based on RSEM results of short-read data. A transcript was considered to be expressed if its FPKM was greater than 5.

5.3 Results

5.3.1 PacBio and Illumina RNA-Seq data

The productivity statistics of the ZMWs (see Table 5.1) showed that the library was overloaded, i.e. a high number of ZMWs contained more than one molecule. Overloaded ZMWs produce inconsistent signals because multiple polymerase-template complexes are in the same ZMW and/or there is high background signal. In summary, 40% of ZMWs produced a polymerase read.

In total we obtained 4.5 million filtered subreads that were processed with SMRT-Link and eventually led to 8,106 error-corrected PacBio isoforms. Details on the results of intermediate processing steps can be found in Table A.2.

Table 5.1: Productivity metrics of the Sequel run as provided by the sequencing facility. P0, P1 and P2 refer to the productivity category which is described in the column "description". "value" corresponds to the percentage of all ZMWs that fall in the respective category. The column "ideal" provides reference values for each category.

productivity	value [%]	ideal	description
P0	3.965	P2>P0	ZMWs which did not produce a polymerase read
P1	39.893	30-45	ZMWs which produced a polymerase read
P2	56.142	< 10	ZMWs which produced inconsistent signals (multiple polymerase-template-complexes, high background signal)

Illumina samples had an average of 30 million reads of which more than 85 % were uniquely mapped to the reference genome (see Table A.3). Only one kidney sample (677_8) had fewer (74 %) uniquely mapped reads. In the PCA, samples clustered by tissue (see Figure A.3). Interestingly, the liver samples formed a cluster which was very distinct from the other tissue clusters in PC1. Their distribution of TPM expression values also differed from those of kidney and brain (see Figure A.4) and revealed that liver samples had a high number of lowly expressed transcripts. To determine whether this reflected a batch effect or a true biological difference, we compared all Illumina samples to the rat data published by Fushan et al. [37] (see Chapter 4) and found that the pattern of liver samples forming a separate cluster which is very distinct from the cluster consisting of kidney and brain samples, was confirmed by a correlation analysis (see Figure A.5). By computing the tissue specificity score defined in Equation 5.2 for each gene with an expression above the detection limit of 1 TPM [2], we determined 2,146 genes with the highest specificity score in liver. Comparing the distribution of expression levels of these genes across all 12 Illumina samples (see

Figure A.6), showed that the median expression of liver-specific genes was higher in liver samples than in brain or kidney samples (t-test p-value: 3.54×10^{-28}). Hence, we concluded that the higher number of lowly expressed genes was not due to technical issues and the data were suitable for further analyses.

5.3.2 Comparing PacBio isoforms to the rat genome

The reference genome and available annotation are valuable resources when characterising isoforms obtained with PacBio long-read sequencing. By aligning the candidate isoforms to an annotated reference one can, for example, assess how many isoforms map to known genes and how many are not present in the annotation. The analysis of the 8,106 PacBio isoforms with SQANTI showed that 7,475 PacBio isoforms were mapped to a set of 4,363 annotated genes. 631 PacBio isoforms could not be mapped to an annotated gene, however, 99 of them mapped to the complementary strand of a known gene. SQANTI uses the term “novel gene” to refer to a genomic locus without an annotated gene but mapping PacBio isoforms, i.e., several PacBio isoforms can be assigned to the same novel gene. Thus, more than 50 % of all 8,106 PacBio isoforms matched an annotated transcript, either in all splice junctions or a subset of splice junctions. The majority of those in the FSM and ISM category contained an open reading frame (ORF) and were thus predicted to be protein-coding (see Figure 5.2). We also investigated the number of transcripts per gene, both for long- and short-read data (see Figure 5.3), and found that PacBio sequencing led to more genes with multiple transcripts. In particular, genes with five and more transcripts were almost exclusively detected in the PacBio data. To examine these genes in greater detail, we looked at the alignment of the 12 genes with more than 10 transcripts and observed an interesting pattern. In eight out of 12 loci, the length of the PacBio isoforms declined successively at the $5'$ end while they were identical in the retained parts of the sequence (see Figure 5.4 for an example). We think that these are artefacts.

5.3.3 Comparing PacBio isoforms to known sequences from rat and human

Since the annotation of the rat genome is incomplete and contains errors (see Chapter 4), we aligned the PacBio isoforms to all rat/human protein sequences in UniProtKB/Swiss-Prot and the larger set of rat/human cDNA sequences available in Ensembl.

The reciprocal best hit BLAST search of the PacBio isoforms with all rat proteins in UniProtKB/Swiss-Prot led to 79.2 % of the isoforms being part of an HSP and 35.2 % of all PacBio isoforms having an RBH (see Table 5.2). Using human instead of rat protein sequences raised the number of PacBio isoforms resulting in an HSP by more than 10 % to 89.8 %. With 51.2 % RBH isoforms, there were also

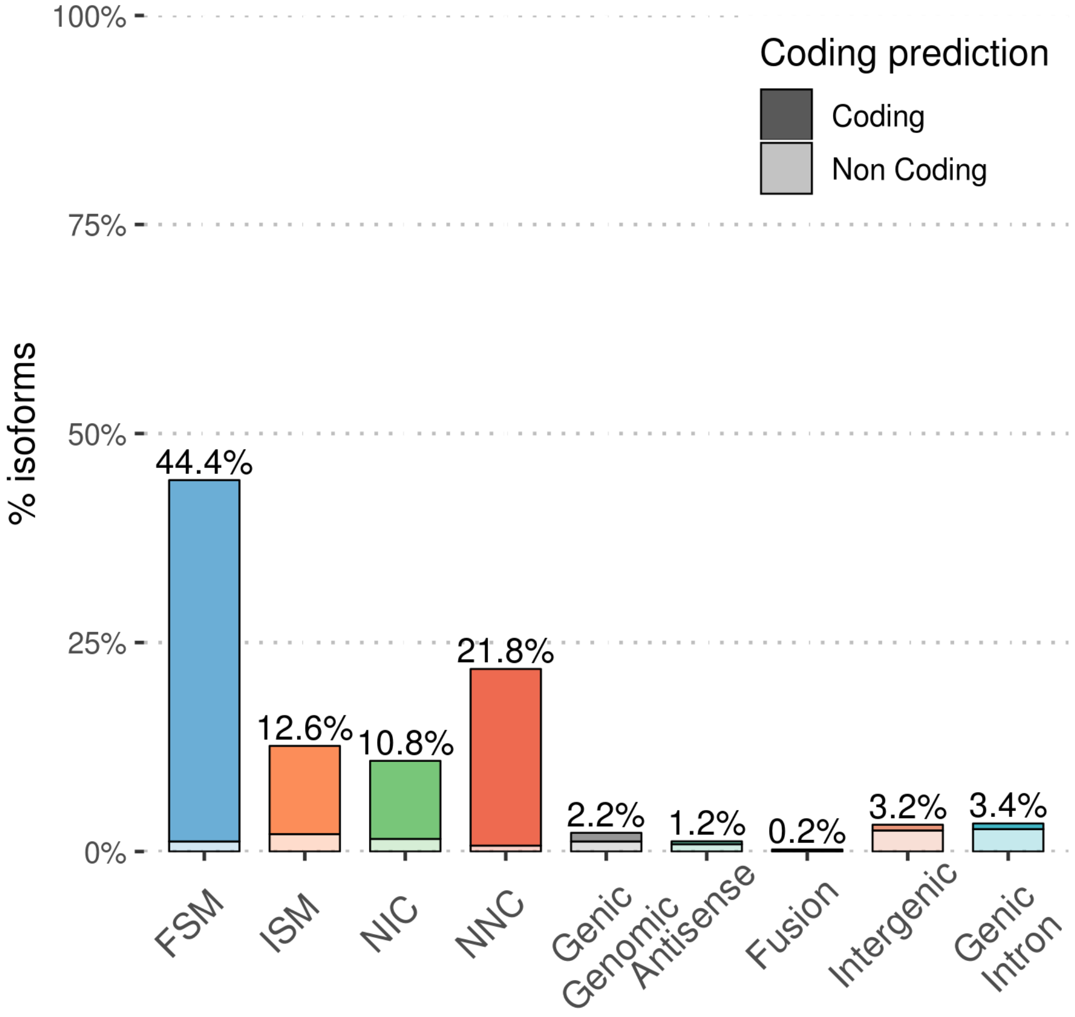


Figure 5.2: Distribution of structural categories identified by SQANTI. FSM: Full Splice Match, ISM: Incomplete Splice Match, NIC: Novel in Catalogue, NNC: Novel Not in Catalogue. The opacity reflects the proportion of transcripts predicted to be protein-coding in each category. The figure was created with minor adaptations using the original SQANTI [114] code.

almost 16 % more isoforms having an RBH in human UniProtKB/Swiss-Prot than in rat UniProtKB/Swiss-Prot. When we looked at the number of RBHs in relation to the number of proteins with a hit in the forward BLAST search, i.e., those we can actually find in the backward search, 44.5 % (rat) and around 80 % (human) were found. Further relaxing the constraints such that the PacBio isoform does not have to be the best hit in the backward search but only a significant one, yielded a hit rate of 57.3 % of all isoforms and 72.3 % of initially hit isoforms for rat. Using human sequences again led to an increase in both hit rates, 74 % of all isoforms and 82.4 % of initial hits had an RBH.

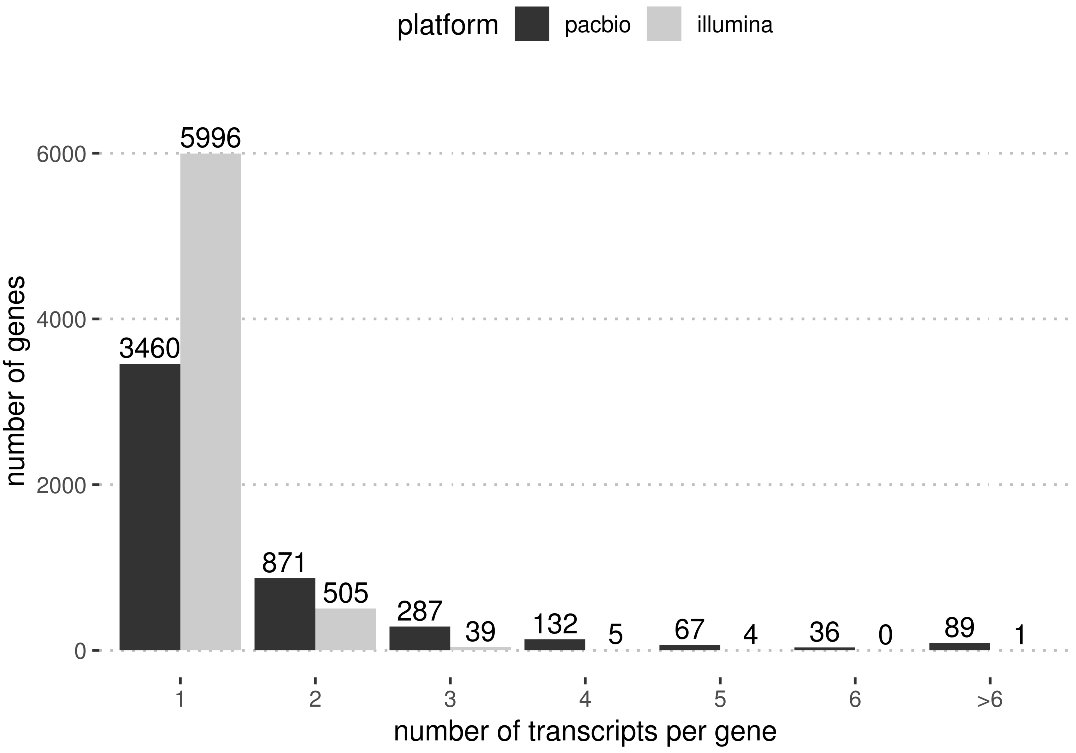


Figure 5.3: Comparison of the number of transcripts per gene between PacBio and Illumina. For the Illumina counts we only considered transcripts with an FPKM>5.

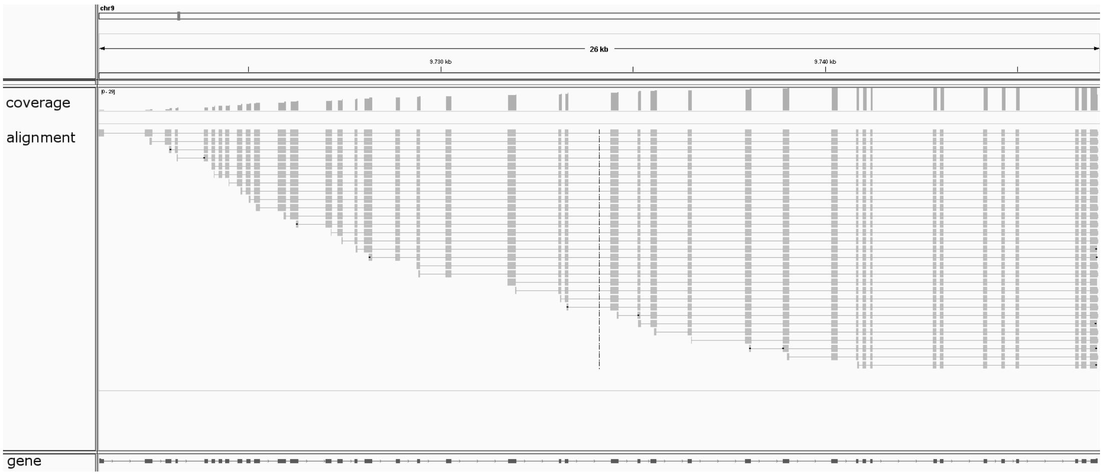


Figure 5.4: Alignment of PacBio isoforms to the rat genome at the C3 gene locus (chr9:9721105-9747167) as an example to illustrate PacBio artefacts. The gene's annotated exon-intron-structure is shown at the bottom. The aligned PacBio isoforms are drawn in the panel in the middle and the barplot on top of the middle panel summarises the coverage.

Table 5.2: Results of the reciprocal best hit BLAST search. The columns "query" and "DB" correspond to query and database in the forward BLAST search, respectively. "pb_iso" stands for PacBio isoforms, "sp" indicates that UniProtKB/Swiss-Prot sequences were used, "ens95" means that sequences came from Ensembl, and the abbreviations "rn" and "hs" denote the species. "iso" is the percentage of rat PacBio isoforms being part of an HSP in the forward BLAST step. "RBH_iso" is the number of reciprocal best BLAST hits in relation to the total number of query sequences while "RBH_init" corresponds to the number of reciprocal best BLAST hits in relation to the number of initial hits in the forward search. "bf_ab_iso" and "bf_ab_init" are the number of isoforms whose best hit in the forward search leads to the isoform as a hit (not necessarily the best one) in the backward search, divided by the number of isoforms or the number of initial hits.

query	DB	iso [%]	RBH_iso [%]	RBH_init [%]	bf_ab_iso [%]	bf_ab_init [%]
pb_iso	sp_rn	79.20	35.25	44.50	57.28	72.32
pb_iso	sp_hs	89.80	51.16	56.97	73.96	82.36
pb_iso	ens95_rn	97.27	55.37	56.92	80.58	82.84
pb_iso	ens95_hs	78.32	49.19	62.80	68.52	87.48
sp_rn	sp_hs	NA	NA	95.43	NA	98.38
ens95_rn	ens95_hs	NA	NA	66.05	NA	90.26

Since UniProtKB/Swiss-Prot contains protein sequences, it does not reflect the entire transcriptome. To allow for the detection of non-coding isoforms, we also performed the reciprocal best hit BLAST approach with the set of cDNA sequences in Ensembl. Using rat Ensembl sequences led to higher numbers in all measured categories in comparison to the search against rat UniProtKB/Swiss-Prot sequences. 97.3 % of PacBio isoforms had an initial hit in the forward BLAST step and more than 82 % of them had an RBH. Interestingly, the number of initial hits decreased by more than 10 % when aligning PacBio isoforms to human cDNA instead of protein sequences. This might be due to a higher number of similar transcripts per gene in the Ensembl data set compared to the number of proteins in the UniProtKB/Swiss-Prot data set. As a result, all other numbers corresponding to a percentage of all PacBio isoforms declined. Those relating to the number of initial hits increased by about 5 %. We also compared the rat cDNA sequences from Ensembl to the protein sequences in UniProtKB/Swiss-Prot to get a theoretical upper limit for the number of hits to expect. Practically, this number cannot be reached because not all genes are expressed at the same time in a given tissue. Furthermore, the sets of sequences retrieved from the databases are not complete, so the numbers of hits we obtained are only an estimate for the theoretical upper limit. In case of the UniProtKB/Swiss-Prot sequences a reciprocal best hit rate of 95.4 % was observed and 98.4 % of the rat proteins had a significant reciprocal hit. When comparing the rat Ensembl sequences to those in human one gets 66 % reciprocal best, and 90.3 % reciprocal hits.

Since the reciprocal best hit BLAST search of PacBio isoforms and the set of well-curated human protein sequences in UniProtKB/Swiss-Prot led to almost 90 % of isoforms having an initial hit of which 57 % had a reciprocal best hit, we decided to proceed with the analysis based on this comparison.

A more detailed investigation of the BLAST results of the reciprocal best hits showed that the coverage of query isoforms by their corresponding HSP varied heavily, both in the forward and the backward search (see Figure 5.5A). The majority of PacBio isoforms were, however, covered well (see Table 5.3 for quantiles of the distribution). The pairwise alignments of the HSPs had an overall high number of matches (see Figure 5.5B and Table 5.3).

Table 5.3: Quantiles of coverage and sequence identity from the RBH search of PacBio isoforms and human proteins. Coverage refers to the percentage of the PacBio isoform covered by the HSP and sequence identity is the percentage of matches in the alignment. The forward search is that aligning PacBio isoforms to human proteins and in the backward step best hits from the forward search are aligned back to PacBio isoforms.

	min	1^{st} quartile	median	3^{rd} quartile	max
coverage (forward)	2.34	40.74	58.87	74.73	98.74
coverage (backward)	2.34	40.78	58.88	74.81	98.38
identity (forward)	32.34	82.44	90.29	96.11	100.00
identity (backward)	32.34	82.47	90.32	96.13	100.00

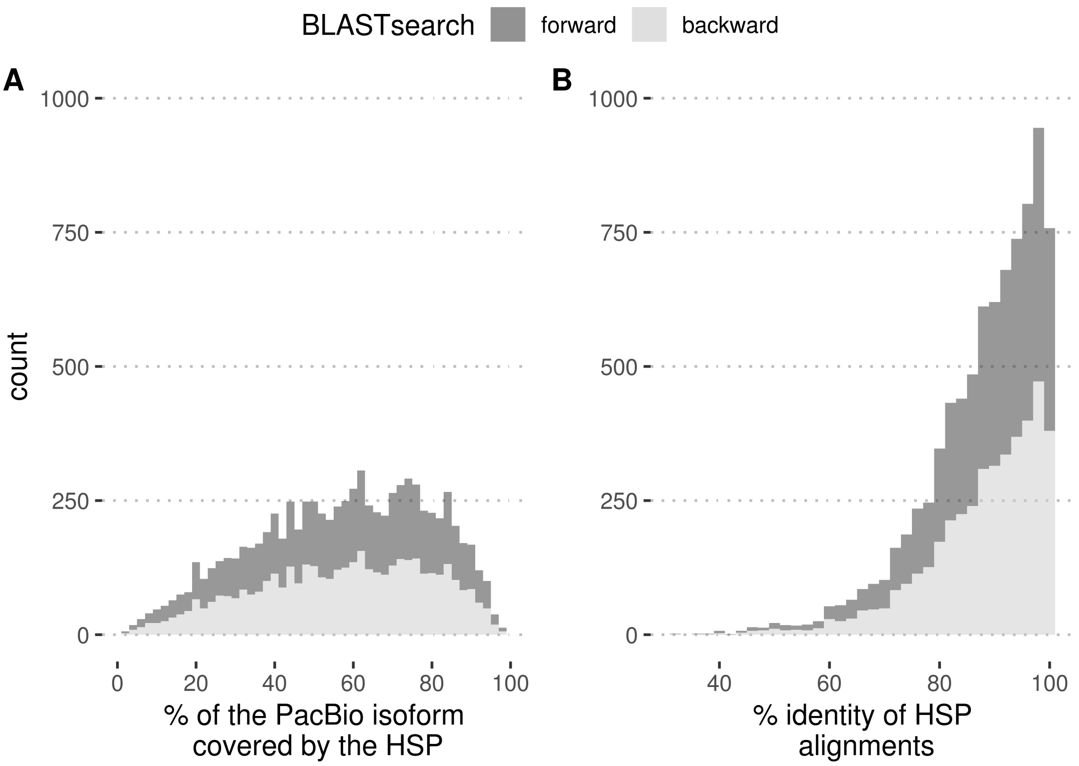


Figure 5.5: Distribution of the percentage of the PacBio transcript covered by the HSP (A) and the percentage of matches in the HSP alignment (B). Only data of reciprocal best hits is shown. The dark grey histogram results from the search of PacBio isoforms in human UniProtKB/Swiss-Prot (forward). The light grey histogram reflects the distribution based on the search of best hits from the forward step in the set of all PacBio isoforms (backward).

SQANTI provided us with an Ensembl gene as well as transcript identifier for all PacBio isoforms which were successfully aligned to the rat genome. Using these identifiers we compared the expression of isoforms having a reciprocal best hit in our search of PacBio isoforms in human UniProtKB/Swiss-Prot sequences and those which did not lead to a reciprocal best hit (see Figure 5.6). Even though the visual difference between the tissues in the two groups is not very pronounced, pairwise Wilcoxon rank sum tests revealed that all tissues, except for kidney on the isoform expression level, showed a significantly different mean expression between the RBH and the notRBH isoforms (significance level: 0.05).

We hypothesised that the set of isoforms not leading to an RBH contains more novel genes and transcripts than those having a reciprocal best hit. Investigating both sets of PacBio isoforms with respect to the gene and transcript categories into which SQANTI classified them (see Table A.4), showed that there are more novel genes and transcripts among the notRBH isoforms than in the RBH ones. A two-sided Fisher's exact test confirmed that the RBH category and the gene/transcript categories are not independent (gene category p-value: 2.8×10^{-147}; transcript

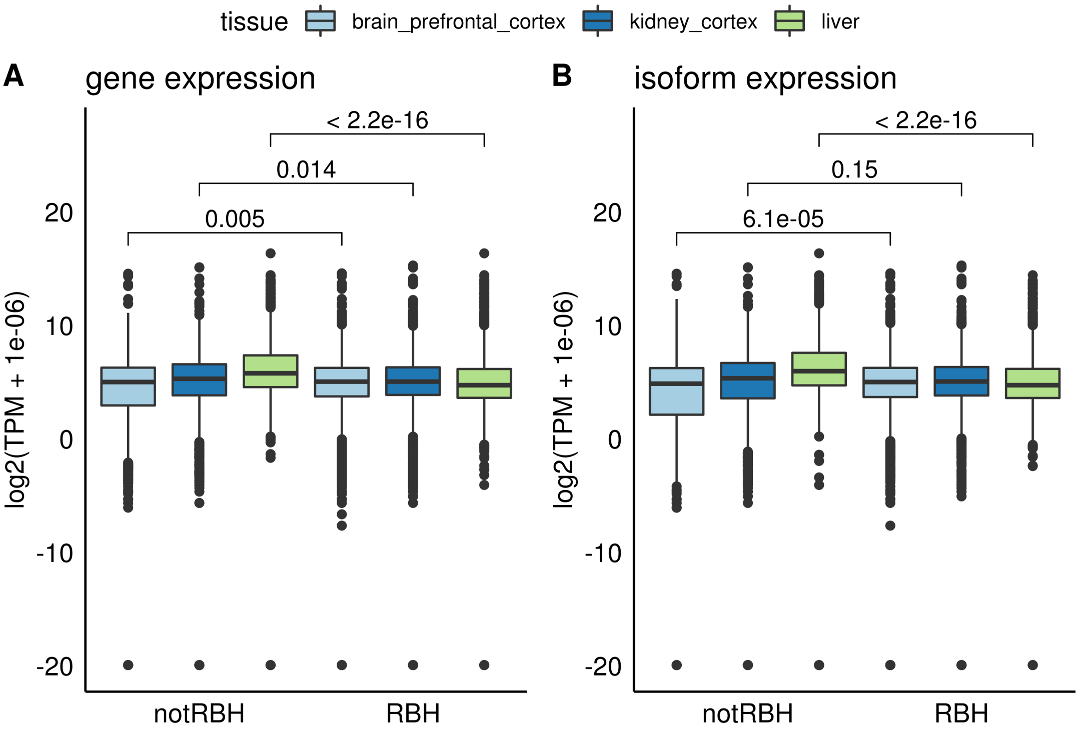


Figure 5.6: Expression levels of genes (A) and transcripts (B) associated with PacBio isoforms having a reciprocal best hit (RBH) and those without one (notRBH) in the BLAST search of PacBio isoforms in human UniProtKB/Swiss-Prot.

category p-value: 2.0×10^{-115}). The 22 PacBio isoforms classified as novel by SQANTI but having a reciprocal best BLAST hit when searched in human UniProtKB/Swiss-Prot proteins were of particular interest. We therefore investigated these further by looking at their alignments to the rat reference genome. In eight cases there is an annotated gene in RefSeq to which the respective PacBio isoform matched well. Nine isoforms are supported by Illumina reads and we therefore suspect they are genes which are missing in the Ensembl annotation. Five isoforms seem to be artefacts because they were of poor quality, i.e., they contained many insertions, deletions and mismatches, and were mostly not supported by Illumina reads.

5.3.4 Comparing PacBio and Illumina data

Mapping rates

We have examined the alignment to the reference genome, both for PacBio and Illumina data, with respect to sensitivity and precision at several levels (see Table 5.4). On average, the sensitivity obtained with the short-read data was 34 % higher than that achieved with long-read data. This means that more annotated features (exons, transcripts, loci etc.) were detected with Illumina than with PacBio. The precision, i.e., the proportion of annotated features in the set of input features, achieved with both sequencing technologies was comparable (mean difference of 8 %).

Table 5.4: Sensitivity and precision at several levels achieved with Illumina and PacBio when comparing the data to the reference genome. The intron chain level refers to the intron-exon structure excluding the terminal exons. The transcript level is a more stringent version of the intron chain level as it also requires matching terminal exons (max. 100 differing bases at the outer boundaries).

	Illumina		PacBio	
	sensitivity	precision	sensitivity	precision
Base level	54.9	74.1	15.4	71.5
Exon level	52.4	88.6	16.3	82.6
Intron level	55.3	95.8	18.3	93.3
Intron chain level	40.3	64.4	10.0	42.7
Transcript level	37.1	53.7	7.5	37.9
Locus level	38.1	59.6	9.2	61.3

Coverage of PacBio isoforms by Illumina reads

To assess whether the long-read data is supported by short-read data, we aligned Illumina reads to PacBio isoforms. A sanity check for the alignments was performed by examining the distribution of the number of mapped reads when aligning Illumina reads from different tissues to our long-read liver data. As expected, we found that the alignment of the Illumina reads from liver samples yielded the highest fraction of mapped reads (Figure A.8). The results from aligning reads from sample 677_3 and the merged liver samples to PacBio isoforms were investigated in greater detail.

In the single sample case, 53 PacBio isoforms were not covered by any Illumina read and 36 of them were classified as novel transcripts by SQANTI. Using the merged liver samples decreased these numbers to 19 not covered isoforms (10 novel transcripts). In both cases none of the isoforms had a reciprocal best BLAST hit when compared to human UniProtKB/Swiss-Prot sequences. A comparison of the number of isoforms per locus (based on SMRTLink processing) between isoforms to which no Illumina read was mapped and those covered by short reads, showed that isoforms which were not covered, belonged to loci with a higher number of isoforms. The 19 PacBio isoforms not covered in the alignment of merged liver reads to PacBio isoforms are from loci with an average of 14.20 isoforms while the rest of the isoforms come from loci with an average of 1.64 isoforms (t-test p-value: 1.7×10^{-2}). As we saw earlier in this chapter (see section 5.3.2), most of these isoforms are probably artefacts.

For the nine isoforms not covered by Illumina reads but successfully mapped to the reference, we also compared reference transcript length and the number of exons and found that they were associated with longer transcripts (t-test p-value: 1.5×10^{-2}) which contain more exons (t-test p-value: 8.6×10^{-5}). This was also confirmed by the data from the single sample alignment, though less pronounced (data not shown).

When we investigated the expression of genes associated with highly, lowly (excluding not covered) and not covered isoforms we saw that the highly covered isoforms corresponded to highly expressed genes while those with few mapped Illumina reads corresponded to genes with lower expression (see Figure 5.7). Interestingly, those not covered mapped to genes with even higher median expression than those highly covered with short reads. In case of lowly and highly covered isoforms we did not observe a difference between the single sample 677_3 and the merged liver samples. Merging the liver samples did, however, reduce the number of not covered isofoms and it emphasised the fact that isoforms to which not a single Illumina read mapped, were associated with highly expressed genes.

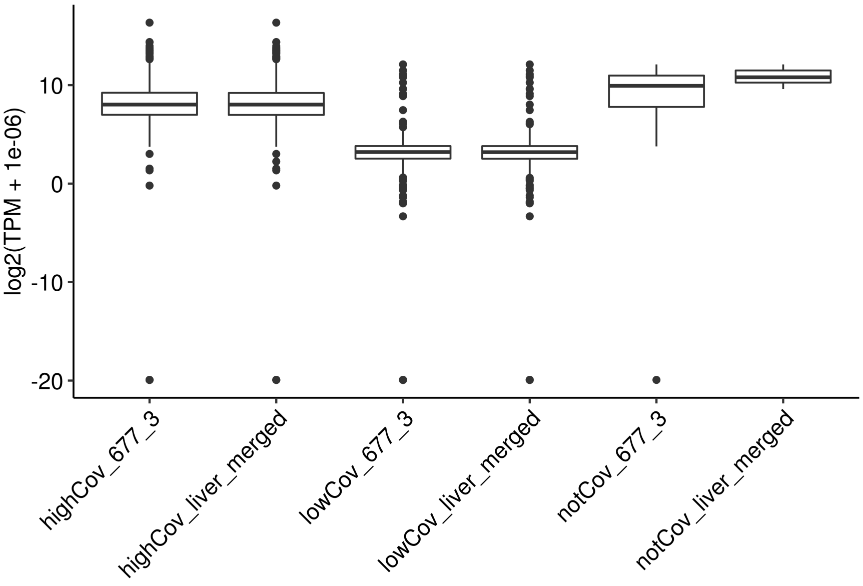


Figure 5.7: Expression of genes associated with PacBio isoforms based on short-read data for PacBio isoforms highly, lowly and not covered by Illumina reads (either from sample 677_3 only or from the pool of all liver samples). High and low coverage was defined using the 3^{rd} and the 1^{st} quartile. Lowly covered isoforms do not include those with zero mapped reads.

Detection limit

One of the main questions we wanted to answer with this pilot study, was which expression level is required for a transcript to be detected by long-read sequencing. A comparison of the expression based on the short-read data from sample 677_3, showed that genes to which a PacBio isoform was mapped had a median TPM of around 30 while those without a matching PacBio isoform had significantly lower expression levels (see Figure 5.8). There were 27 transcripts which had an associated PacBio isoform but an expression of 0.

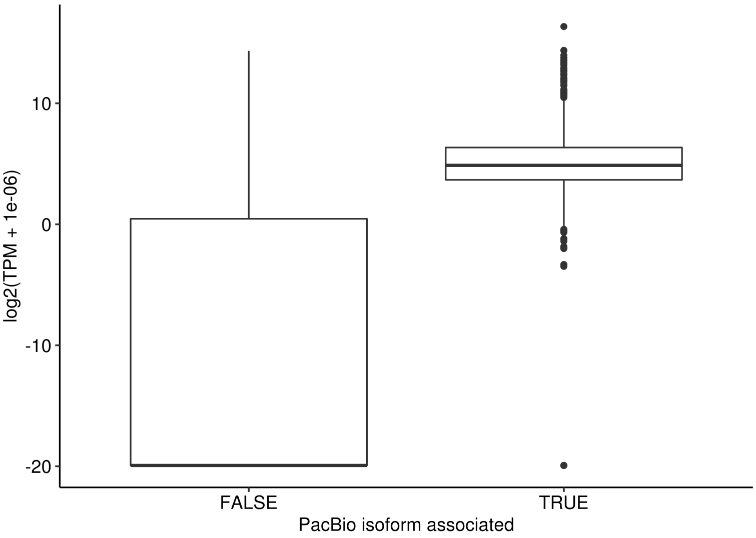
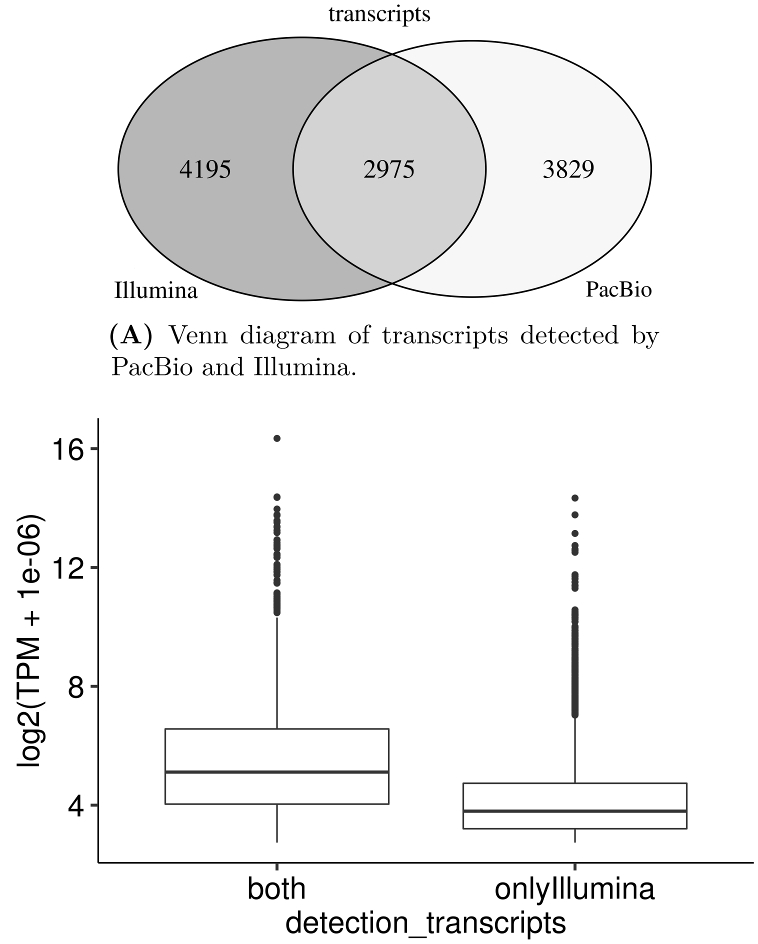


Figure 5.8: Comparison of gene expression between genes to which a PacBio isoform was mapped and those without an associated isoform.

Overlap of detected transcripts

Based on the analysis of PacBio isoforms with SQANTI, we compared the set of associated transcripts to that of expressed genes detected with Illumina data (see Figure 5.9A). Interestingly, there are more transcripts detected by only one of the two platforms than ones detected by both. Especially the set of transcripts only detected with Illumina appears to be quite large. A look at the expression of these transcripts compared to those found by both technologies, showed that they mainly correspond to lowly expressed transcripts (see Figure 5.9B).



(A) Venn diagram of transcripts detected by PacBio and Illumina.

(B) Comparison of transcript expression based on short-read data between transcripts detected with PacBio and Illumina and those only detected with Illumina.

Figure 5.9: Overlap of transcripts detected with PacBio and Illumina.

5.4 Discussion and outlook

In this chapter the results from a pilot study were presented in which rat liver, brain and kidney samples were examined with the aim to compare short- and long-read transcriptomic data. All samples were sequenced with Illumina to obtain short-read data and one liver sample was sequenced using PacBio's Sequel system to generate long reads.

During the initial quality control of the short-read data we noticed that the liver samples had a high number of lowly expressed genes. Since Yu et al. [141] also found that liver and muscle had the lowest numbers of expressed genes (FPKM ≥ 1) and we observed this in other data sets, we were confident that this was not a technical batch effect. Also, our primary aim was to use these data to confirm isoforms which were detected with PacBio's long-read sequencing and not to compare expression between tissues, therefore, we did not expect the high number of lowly expressed genes to negatively impact our results.

From the productivity values of the Sequel run we concluded, that the SMRT cells were overloaded (high P2). Although the 40 % of ZMWs that produced one polymerase read (P1) was within the ideal range provided by the sequencing laboratory (30-45 %), it appeared quite low in comparison to recommendations by the developer of the Iso-Seq pipeline, Elizabeth Tseng [84, 123] (50-75 %). Therefore, the productivity of our run was not optimal but it should not impair the results of our study (personal communication with David Stucki, Senior Scientist at PacBio). In her presentation [84], Elizabeth Tseng provided additional reference numbers obtained from processing human data from a "best case" scenario with the IsoSeq3 pipeline. The number of CCS (264,467 vs. 572,406) as well as the number of full-length reads (234,476 vs. 430,257) also seemed rather low. The percentage of full-length reads was, however, comparable to the number provided in Tseng's presentation (88 % and 75 %, respectively). Of course, one has to keep in mind that the values used as a reference stem from an ideal sequencing run using samples from another species.

After clustering full-length reads, the IsoSeq3 pipeline polishes the resulting isoforms to generate consensus sequences that are split into high and low quality isoforms. Both have $\geq$ two full-length read support but high quality isoforms have an accuracy of $\geq$99 % while the accuracy of low quality isoforms is less than 99 % [126]. Since we observed more low quality isoforms than expected for the IsoSeq3 protocol (personal communication David Stucki), we had a high number of polished isoforms with low predicted accuracy. Here, one should, however, keep in mind that a hard cut-off of 99 % is used and isoforms with an accuracy of 98.9 % are already classified as having low quality.

When we investigated the alignment of PacBio isoforms to the reference genome, we observed a number of loci with enrichment in isoforms that successively declined in length. Initially, we expected them to be merged during the clustering step in the IsoSeq pipeline. Their difference at the $5'$ end was, however, too big for them to be identified as the same isoform since the IsoSeq3 pipeline clusters full-length reads by considering two reads to be the same isoform if their: 1) difference at $5'$ end is $<$100 bp, 2) difference at $3'$ end is $<$30 bp and 3) gaps in exons are $<$10 bp with no limit on the number of gaps [126]. These successively shorter isoforms could be either due to alternative transcription start sites, incomplete reverse transcription [134], or RNA degradation. For the latter to be ruled out, a $5'$ cap selection would have to be done to correctly determine transcription start sites and hence ensure that transcripts are really full-length. This would probably reduce the already low number of detected isoforms but it may also increase the chances of finding transcripts with lower expression since less reads are used for these truncated transcripts.

Another approach would be to simply remove the redundant isoforms by either using the longest representative or collapsing them [62, 124, 133, 134]. A validation of PacBio transcripts with PCR, performed by the authors of SQANTI, showed that it is generally advisable to apply further filtering steps to the IsoSeq output [114]. SQANTI also includes a classifier which uses FSM transcripts (positive class) and NNC transcripts with noncanonical splice junctions (negative class) derived from the input to remove artefacts. Unfortunately, the thorough evaluation of this method and its application to the pilot study's data were outside the scope of this thesis.

In general, the points discussed above hint at a low sequencing depth and we were advised to look into rarefaction curves (personal communication with David Stucki) to determine whether more data would increase our isoform yield. Unfortunately, this is beyond the scope of this thesis.

Mapping the PacBio isoforms to the reference genome of rat showed that the majority could be associated with an annotated transcript and were predicted to be protein-coding. The average length of isoforms classified as FSM by SQANTI (2130 bp) compared well to that of the matching reference transcripts (2146 bp).

Analysing the overlap of transcripts detected with PacBio and Illumina showed that there is a substantial number of transcripts detected with both platforms (27 %). There are, however, also many transcripts detected by only one of the technologies. We found that those only detected with Illumina were generally lower expressed than those detected by both platforms. Considering the potential artefacts we identified, the 35 % only found with PacBio should be investigated in greater detail. Weirather et al. [135] have evaluated the performance of Illumina, PacBio, and Oxford Nanopore on a gold standard data set comprising 68 transcripts with various alternative splicing events. They found that the two long-read sequencing technologies outperform Illumina, i.e., we can be confident that some of these transcripts only detected with PacBio are true novel isoforms which should be validated.

The fact that we found more reciprocal (best) hits in the reciprocal BLAST of PacBio isoforms vs. human protein sequences than in the search against rat proteins, confirmed the observation we made in Chapter 4 that the rat annotation is incomplete and rat is sufficiently closely related to human to exploit human annotation to improve that of rat.

The high number of reciprocal (best) hits in the comparison to rat cDNA sequences annotated in Ensembl (version 95) showed that the long-read data covered the rat transcriptome well. Interestingly, we observed a decrease in initial hits when aligning PacBio isoforms to human cDNA sequences. Those having a reciprocal best hit showed a lower sequence identity across HSPs (median: 85.78 %; data not shown) than in the search against human proteins (median: 90.29 %; data not shown). Therefore, we concluded that the difference between rat PacBio isoforms and human is relatively high on a nucleotide sequence level but, due to the degenerated genetic code, the resulting proteins are similar. This finding may hint at a still high error-rate in PacBio isoforms which contradicts the findings of another study that has shown that the error correction in the IsoSeq method reduces the error rate from 14.20 % in subreads to 1.72 % in CCS [135].

The alignment of Illumina reads to PacBio isoforms showed that only a negligible number of isoforms was not at all covered by short reads and the majority had a high number of mapped reads. Some of the isoforms not covered by any reads were also not successfully mapped to the rat genome and may thus be true novel isoforms. One should, however, not only look at the number of mapped reads but also at their distribution along the isoform to determine whether it is uniformly covered or whether, for example, the $5'$ end is less well covered. To make this analysis feasible for thousands of isoforms, an appropriate metric to summarise the density of mapped reads along the PacBio isoform would have to be found.

We were also able to show that, in our data set, a transcript was likely to be discovered with long-read sequencing if its expression was above 30 TPM. There were, however, 27 isoforms with a PacBio isoform but 0 expression for the associated transcript. Looking into these cases revealed that these transcripts belonged to genes with either several very similar annotated transcripts or to ones that have a high number of different transcripts. For example, the PacBio isoform PB.2596.1 was mapped to the transcript Trim39-204 (Ensembl identifier: ENSRNOT00000084559) of the gene Trim39 (Ensembl identifier: ENSRNOG00000000785) which had a TPM of 0 in sample 677_3. Visualisation of the alignment showed that PB.2596.1 matched Trim39-204 well, i.e., it was correctly aligned to Trim39-204 instead of one of the other three protein-coding, annotated transcripts. Therefore, we assumed that in case of these 27 transcripts, either the RSEM isoform expression quantification might be incorrect or the transcripts were simply not captured with short-read sequencing.

Despite the fact that the long-read data generated in this study suffered from low accuracy for many isoforms and a $5'$ cap selection is highly recommended to increase the number of true full-length isoforms, we obtained encouraging results. The long-read isoforms were mostly confirmed by the comparison to 1) the rat genome, 2) human, well curated protein sequences and 3) short-read data. Therefore, we recommend sequencing all tissue samples with PacBio performing a $5'$ cap selection to then look into tissue specific isoforms or investigate the novel isoforms in greater detail. We are confident that the genome annotation of such a common model organism as rat can greatly benefit from a follow-up study.

A recently published method, LoReAn [22], already combined short- and long-read RNA-Seq data, as well as protein information, to improve genome annotation. As it has, so far, only been applied with two fungal and two plant data sets, it would be very interesting to investigate the LoReAn approach in greater detail and to apply it to mammalian genomes, if feasible.

Chapter 6

Assessing a target's conservation across model organisms

6.1 Introduction

The analysis of reasons for pharmaceutical project closure performed by Cook et al. [21] showed that the majority of the projects were terminated due to efficacy and safety issues (see chapter 2.1 for more details) with a shift from safety to efficacy related reasons the farther the project has advanced along the drug development pipeline. This lack of efficacy and safety may be due to differences between the model species—like mouse, rat, dog, pig, or cynomolgus monkey—and human. Despite their general genetic, anatomical and physiological similarity to human, one has to critically assess the suitability of certain model organisms for drug development.

The most commonly used model species, mouse and rat, shared their last common ancestor with human around 90 million years ago (MYA) (see Figure 6.1). The two species frequently used for non-rodent safety studies, dog and pig, are even more distantly related to human with their last common ancestor with human being about 96 MYA. Cynomolgus monkey is closest to human with their last common ancestor having occurred roughly 30 MYA. Consequently, we must expect targets to act differently in different species. By determining these discrepancies in advance *in silico*, we could spare millions of animal lives. This would be in line with the aim to reduce animal experiments as defined in the principles of the 3Rs [100] (**R**eplacement, **R**eduction and **R**efinement). Furthermore, not performing experiments whose result cannot be translated into human due to species differences, reduces cost and time, both valuable assets for pharmaceutical companies. To prevent unnecessary animal testing, it is important to carefully investigate and compare different types of information about the target in human and the potential model species. Initiatives like "Illuminating the Druggable Genome" aim to assemble a comprehensive collection of data sets such as literature based target information, gene expression or disease associations. The gathered

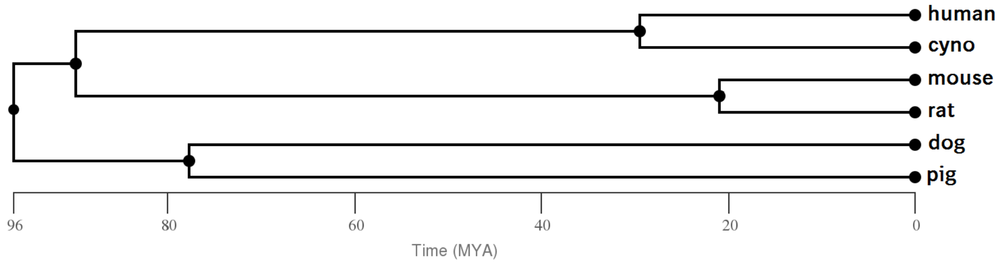


Figure 6.1: Phylogenetic tree of human, cynomolgus monkey, mouse, rat, dog and pig. The branch length represents the divergence time in million years ago (MYA). This figure is an adapted version of a figure created with TimeTree [61].

information is accessible via Pharos [81] and also includes a list of orthologous genes. OpenTargets [18] takes it a step further and considers mouse models when computing a score measuring target-disease-association. These approaches are valuable for target prioritisation and to decrease the number of projects failing due to target related issues in efficacy and safety. To our knowledge there is, however, no method that assesses a target's properties in rodent and non-rodent model species. In addition to ranking potential targets, such a score would facilitate the *a priori* determination of suitable model species and allow for the determination of the conservation level of specific gene sets.

Therefore, we present the target conservation score which takes sequence, gene expression and function related information into account to provide a measure for the conservation of a target gene across different model organisms. In the following, we first introduce the total conservation score, which is the weighted mean of altogether eight subscores, that are explained in detail below. We will apply it to all protein-coding human genes and analyse the results. A brief description of the database implemented for storage and retrieval of the computed scores is going to be presented. Attempts to validate the proposed score and a discussion of the results conclude this chapter.

6.2 Methods

6.2.1 Target conservation score

The target conservation score is the weighted mean of eight subscores, that model three different sequence homology aspects—expression, functional, and network conservation—for orthologous gene pairs between human and a chosen model species.

The conservation score for a human gene g^{Hs} in species s is computed as the weighted sum of subscores that are calculated for all orthologous pairs (g^{Hs}, g^s_o):

$$\begin{aligned} s_{species}(g^{Hs}, s) =& w_1 * s_{hom} \\ &+ w_2 * \frac{\sum_{o=1}^{n_s} s_{id}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_3 * \frac{\sum_{o=1}^{n_s} s_{goc}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_4 * \frac{\sum_{o=1}^{n_s} s_{wga}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_5 * \frac{\sum_{o=1}^{n_s} s_{exp}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_6 * \frac{\sum_{o=1}^{n_s} s_{spec}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_7 * \frac{\sum_{o=1}^{n_s} s_{net}(g^{Hs}, g^s_o)}{n_s} \\ &+ w_8 * \frac{\sum_{o=1}^{n_s} s_{interpro}(g^{Hs}, g^s_o)}{n_s} \end{aligned} \tag{6.1}$$

The total conservation score for g^{Hs} in then th weighted sum of the species scores in all considered species:

$$s_{total}(g^{Hs}) = \sum_{s=1}^{S} w_s * \left(s_{species}(g^{Hs}, s) \right) \tag{6.2}$$

With:

- S: number of species
- n_s: number of orthologues for g^{Hs} in species s
- w_s: species weight, e.g., $\frac{1}{S}$
- $w_1, \dots, w_8$: score weights, e.g., $\frac{1}{8}$

- s_{hom}: subscore based on the number of orthologues ($\frac{1}{n_s}$)
- s_{id}: subscore based on sequence identity
- s_{goc}: subscore based on synteny
- s_{wga}: subscore based on genome alignment
- s_{exp}: subscore based on gene expression correlation across tissues
- s_{spec}: subscore based on tissue specificity of gene expression
- s_{net}: subscore based on an expression correlation network
- $s_{interpro}$: subscore based on Interpro annotation

All subscores are normalised to the range of $[0, 1]$ to make them comparable.

Homology subscores

To model homology conservation, we make use of Ensembl's cross-species resource, EnsemblCompara [44], which provides different scores for many organisms both on the sequence and the gene level. We use three subscores, s_{id}, s_{goc}, and s_{wga}.

Ensembl generates pairwise protein alignments for each orthologous gene pair and computes the target and query percent identity corresponding to the percentage of the sequence covered by the alignment (see Figure 3.1). s_{id} is the average of these two identities.

Ensembl's gene order conservation (GOC) score (s_{goc}) captures conserved synteny, i.e., co-localisation, of the orthologous genes by comparing order, orientation and homology of two genes up- and downstream of the orthologues. If all four genes match, the orthologue's GOC is 100%, for each mismatch 25% are deducted. While the GOC score accounts for large-scale rearrangements, the whole genome alignment (WGA) score focuses on nucleotide level differences. It is based on the assumption that the genomic sequence of orthologues should align well and is calculated from the pairwise genome alignments available in Ensembl. For each gene in an orthologous gene pair, the coverage over exons and introns is computed separately and a weighted score is calculated such that the emphasis is on the exon coverage [31]:

$$s_{wga} = \frac{cov_{exons} + cov_{introns} * perc_{intron} * (1 - cov_{exons})}{100}$$

where:

- cov_{exons}: alignment coverage on exons
- $cov_{introns}$: alignment coverage on introns
- $perc_{introns}$: percentage of the gene structure consisting of intronic sequence

By applying different thresholds (depending on the last common ancestor of the involved species) to s_{id}, s_{goc}, and s_{wga}, Ensembl tags homology relationships as "high confidence" [31]. For the computation of the conservation score we did, however, only use the raw scores s_{id}, s_{goc}, and s_{wga} because the thresholds are chosen arbitrarily.

Expression subscores

To compare the human target gene to an orthologue with respect to their expression pattern across different tissues, we compute Pearson's correlation coefficient ρ of normalised gene expression (tissue median RPKM). This coefficient is then transformed to be in the range of $[0, 1]$:

$$s_{exp} = \left(\frac{\rho + 1}{2}\right)^2 \tag{6.3}$$

We chose to transform ρ according to Equation 6.3 to penalise negative correlation because we do not consider a gene to be conserved if it is, for example, lowly expressed in human liver but highly expressed in mouse liver.

Another expression based subscore aims at capturing similar tissue-specificity patterns using the tissue-specificity score defined in Equation 5.2. Computing this score for two orthologous genes results in two vectors with a specificity value for each tissue. The subscore s_{spec} is then computed as Kendall's tau [80] of these two vectors. A value of 0 indicates that there is no relationship while 1 means there is a perfect relationship between the two tissue-specificity vectors.

Expression correlation can be exploited further to not only assess direct correlation between orthologous genes but also their co-expression pattern within each species separately. This helps us to determine the level of interaction with other genes in the transcriptome.

Based on the between-species correlation matrices we determine all genes in the investigated species s which are highly correlated to the human target g^{Hs} ($\rho >= 0.8$, see section 6.3.2 for details on the threshold selection). This set of genes is referred to as $set^{s}_{between}$. Given the gene expression matrices for human and species s ($genes \times tissues$), we compute the pairwise Pearson's correlation coefficient across tissues for all genes in each species separately. From these within-species correlation matrices we determine all highly correlated genes for g^{Hs} and each gene $g^s \in set^s_{between}$ $(\rho >= 0.8)$. Let us call the resulting sets $highCorr_{g^{Hs}}$ and $highCorr_{g^s}$. Based on these sets we determine the difference in the number of the proportional counts of highly correlated genes in the two species:

$$s\_netNumCorr = 1 - \left| \frac{|highCorr_{g^{Hs}}|}{num\_genes_{hs}} - \frac{|highCorr_{g^s}|}{num\_genes_{species}} \right| \tag{6.4}$$

As a potential alternative to the purely count based network score, we compute the Jaccard index [53] of the two sets to assesses which fraction of the highly correlated genes in each species are orthologous:

$$s\_netNumOrtho = \frac{|ortho(highCorr_{g^{Hs}}, s) \cap highCorr_{g^s}|}{|ortho(highCorr_{g^{Hs}}, s) \cup highCorr_{g^s}|} \tag{6.5}$$

with $ortho(highCorr_{g^{Hs}}, s)$ being the orthologues of the genes in $highCorr_{g^{Hs}}$ in species s.

Functional subscore

To assess functional similarity between orthologues, we compare associated protein families and domains. Here we exploit the annotation with InterPro [74] accession numbers provided by Ensembl and compute $s_{interpro}$ as the Jaccard index:

$$s_{interpro} = \frac{|set_{hs} \cap set_s|}{|set_{hs} \cup set_s|} \tag{6.6}$$

where set_{hs} contains all InterPro accession numbers of the human gene and set_s holds all InterPro accession numbers associated with the orthologous gene in a chosen species s.

6.2.2 Application to all protein-coding human genes

We computed the target conservation score defined in Equation 6.2 for all human protein-coding genes annotated in Ensembl version 92 based on their relation to orthologues in mouse and rat. The gene expression related subscores are based on the Fushan et al. data [37] comprising three tissues (brain, liver, and kidney), because they contain a wide range of species that can be incorporated after successful evaluation of the conservation score.

Based on the resulting scores we investigated the distribution of individual subscores and evaluated the two proposed co-expression network scores. Both network scores $s\_netNumCorr$ (Equation 6.4) and $s\_netNumOrtho$ (Equation 6.5) are calculated by comparing the set of highly correlated genes in two species and do therefore depend on the correlation threshold used to call a pair of genes highly correlated. We tested different thresholds (0.7, 0.8, and 0.9) and compared the distribution of $s\_netNumCorr$, $s\_netNumOrtho$ as well as the aggregated species scores and the conservation score s_{total}. Due to the little difference between the distributions for 0.7 and 0.8, we show kernel density estimates (geom_density() from ggplot2 [137]) instead of gene counts to facilitate visual comparison.

The impact each subscore has on the total conservation score was evaluated by following a leave-one-out approach and comparing the resulting score distribution to that obtained when using all subscores.

Furthermore, we determined groups of genes with low and high scores via k-means clustering using the kmeans function in R. The number of clusters k was set to five as one can visually identify roughly five peaks in the distribution of s_{total}. We performed a Gene Ontology enrichment for the two groups of genes with lowest and highest scores to validate the conservation score. The GO enrichment was done using the Bioconductor package clusterProfiler [140] which performs a hypergeometric test. The set of conservation scores for all human protein-coding genes were used as the gene universe to which the clusters were compared.

The Jackson Laboratory provides the Human - Mouse: Disease Connection (HMDC) data [3], which combines information from the Mouse Genome Informatics database (MGI) with human disease and phenotype annotation obtained from NCBI, the Online Mendelian Inheritance in Man (OMIM) and the Human Phenotype Ontology (HPO) [1]. Due to the lack of a list of genes which were proven to be well conserved, we used the HMDC data to validate the proposed conservation score. This approach is based on the assumption that genes being associated with a certain disease in different species, are most likely well conserved with respect to sequence and gene expression information. To obtain a list of presumably well conserved genes, we extracted orthologous human-mouse gene pairs associated with the same disease ontology [58] identifier. The resulting set of Entrez [67] identifiers was mapped to Ensembl identifiers using the biomaRt package (version 2.32.1) with Ensembl version 92 and the distribution of $s_{species}$ for mouse was investigated.

6.2.3 targetcon - A relational database storing conservation scores

To provide easy access to the computed scores for all protein-coding human genes we have set up a PostgreSQL database, called targetcon, which stores the underlying Ensembl and expression data as well as the derived conservation scores. An overview of the database tables and their relations is depicted in Figure 6.2.

Targetcon contains 13 tables which can be grouped into three groups: 1) Ones that contain data from Ensembl on genes, homology relationships and Interpro annotation, 2) those which hold the expression data and group them into different studies, and 3) tables storing the aggregated conservation scores $s_{species}$ and s_{total} as well as all subscores and the weights used for the score computation. Currently, all species and all subscores are weighted equally (see equations 6.1 and 6.2).

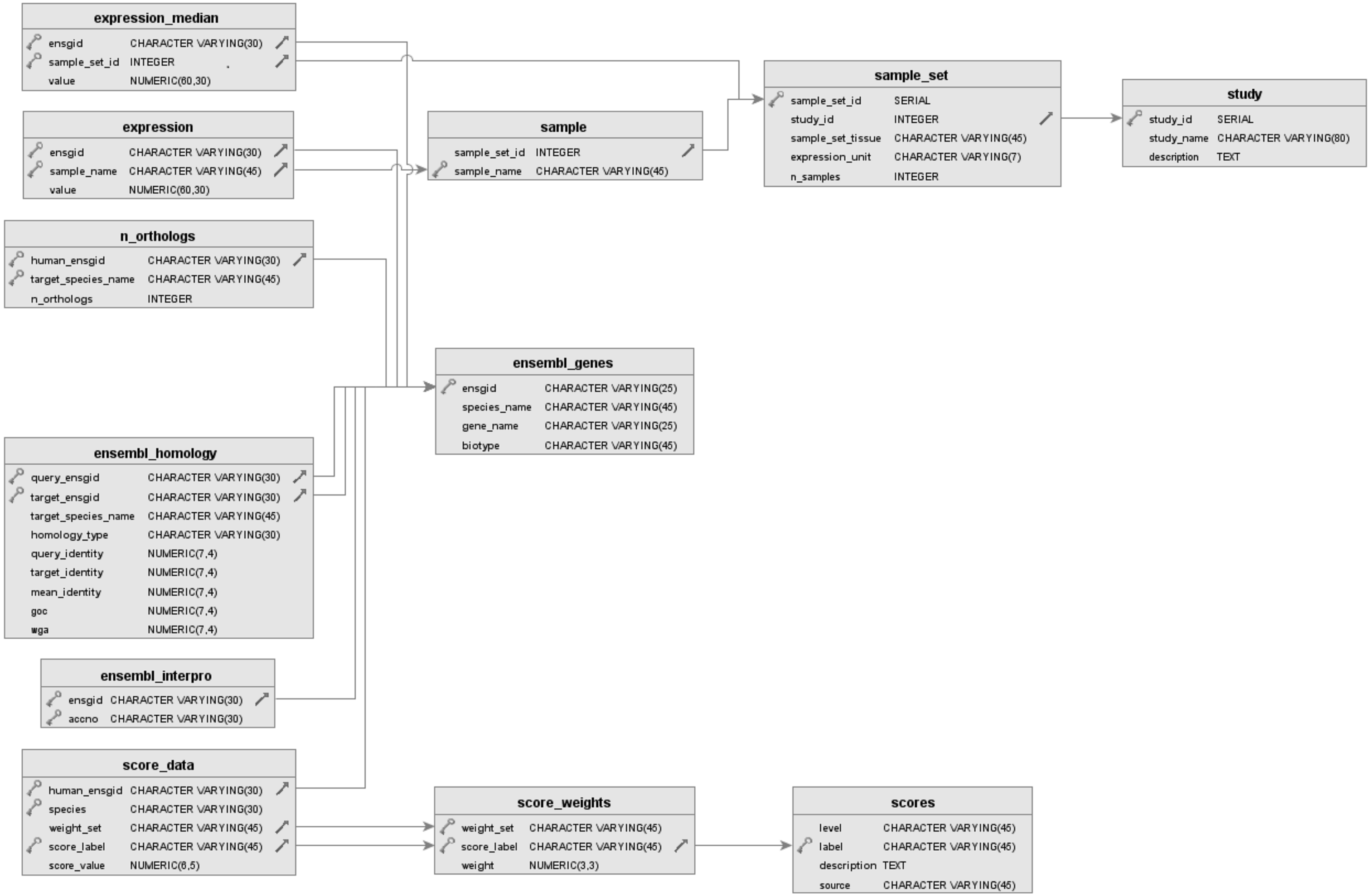

expression_median
ensgid CHARACTER VARYING(30)
sample_set_id INTEGER
value NUMERIC(60,30)
expression
ensgid CHARACTER VARYING(30)
sample_name CHARACTER VARYING(45)
value NUMERIC(60,30)
n_orthologs
human_ensgid CHARACTER VARYING(30)
target_species_name CHARACTER VARYING(45)
n_orthologs INTEGER
ensembl_homology
query_ensgid CHARACTER VARYING(30)
target_ensgid CHARACTER VARYING(30)
target_species_name CHARACTER VARYING(45)
homology_type CHARACTER VARYING(30)
query_identity NUMERIC(7,4)
target_identity NUMERIC(7,4)
mean_identity NUMERIC(7,4)
goc NUMERIC(7,4)
wga NUMERIC(7,4)
ensembl_interpro
ensgid CHARACTER VARYING(30)
accno CHARACTER VARYING(30)
score_data
human_ensgid CHARACTER VARYING(30)
species CHARACTER VARYING(30)
weight_set CHARACTER VARYING(45)
score_label CHARACTER VARYING(45)
score_value NUMERIC(8,5)
sample
sample_set_id INTEGER
sample_name CHARACTER VARYING(45)
ensembl_genes
ensgid CHARACTER VARYING(25)
species_name CHARACTER VARYING(45)
gene_name CHARACTER VARYING(25)
biotype CHARACTER VARYING(45)
score_weights
weight_set CHARACTER VARYING(45)
score_label CHARACTER VARYING(45)
weight NUMERIC(3,3)
sample_set
sample_set_id SERIAL
study_id INTEGER
sample_set_tissue CHARACTER VARYING(45)
expression_unit CHARACTER VARYING(7)
n_samples INTEGER
scores
level CHARACTER VARYING(45)
label CHARACTER VARYING(45)
description TEXT
source CHARACTER VARYING(45)
study
study_id SERIAL
study_name CHARACTER VARYING(80)
description TEXT

Figure 6.2: Database schema of targetcon. Boxes correspond to tables and show the table's columns with their type. Keys to the left of column names indicate primary keys while arrows to the right of column names mark foreign keys. Tables containing a foreign key are connected to the referenced table with an arrow.

As the information in targetcon changes with new Ensembl releases but we wanted to keep old versions available to ensure compatibility with legacy data sets but also for comparison, a new version of targetcon is created when one wishes to update to a new Ensembl version.

A fully automated Nextflow pipeline was developed to create the database, to fetch the expression matrices from the file system and the required Ensembl information, to prepare the data for the database upload and the subsequent computation of the conservation scores, and finally to compute and upload the conservation scores for all human protein-coding genes. The pipeline receives all its input parameters via a configuration file with the fields described in Table A.5.

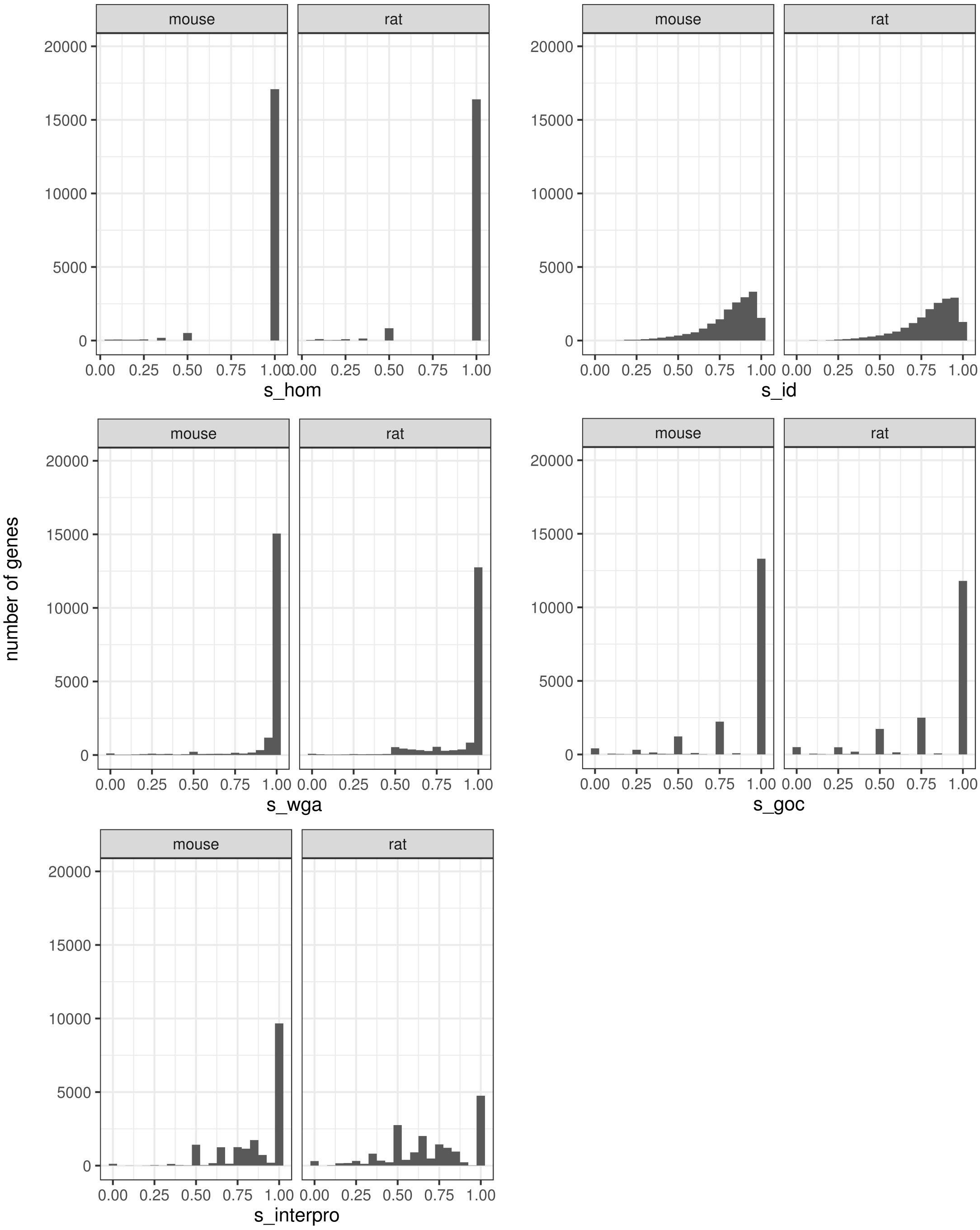
6.3 Results

6.3.1 Distribution of subscores

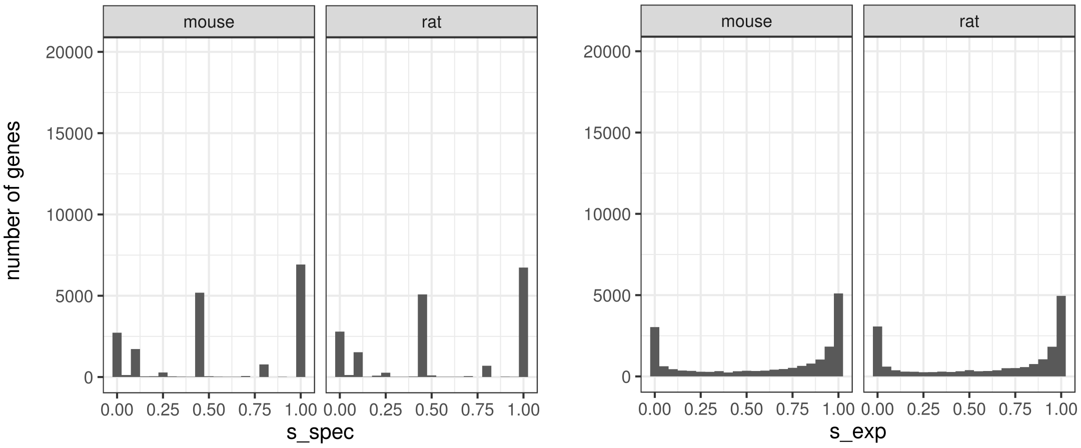
Investigating the distribution of the individual subscores (see Figure 6.3), revealed that the vast majority of orthology relationships between human protein-coding genes and genes in mouse or rat were one-to-one homologies (s_{hom}=1). In most cases, the orthologous gene pairs showed conserved synteny (s_{goc}) and a high coverage on nucleotide level (s_{wga}). Interestingly, the protein sequence identity (s_{id}) showed a broader distribution.

The distribution of the expression correlation across tissues (s_{exp}) indicates that the majority of orthologous gene pairs were either strongly correlated or almost perfectly anti-correlated. Regarding the tissue-specificity subscore s_{spec}, many orthologous genes were well conserved, there was, however, also a substantial proportion of cases with weak or no relationship between the tissue-specificity vectors.

The subscore distributions described so far, only showed minor differences between human-mouse and human-rat cases. For $s_{interpro}$ we observed that the scores for rat were lower than those for mouse, indicating that there is a greater overlap between Interpro annotations of human with those in mouse than those in rat. One should, however, bear in mind that mouse has been investigated to a greater extent and therefore more annotation is available which is also more reliable.



(A) Distribution of s_{hom}, s_{id}, s_{wga}, s_{goc}, and $s_{interpro}$.



(B) Distribution of s_{spec} and s_{exp}.

Figure 6.3: Distribution of (A) the Ensembl-based subscores s_{hom}, s_{id}, s_{goc}, s_{wga}, and $s_{interpro}$ as well as (B) the expression based subscores s_{exp} and s_{spec} between human and each of the two rodent species mouse and rat.

6.3.2 Subscores based on co-expression networks

Due to the low dynamic range covered by the network score based on the orthology relationships between the two species-specific networks ($s\_netNumOrtho$; see Figure A.9A), we decided to only use the one comparing the sizes of the species-specific networks ($s\_netNumCorr$).

From the distribution of $s\_netNumCorr$ (see Figure A.9B), we deduced that a correlation threshold of 0.9 is too stringent as it led to a high number of genes having a network score of zero. We did not observe a great difference between 0.7 and 0.8, but since the distribution based on 0.7 was slightly shifted towards lower scores around 0.5, we decided to use 0.8 as the correlation threshold in the computation of $s\_netNumCorr$. In the aggregated scores ($s_{species}$ and s_{total}) the two lower thresholds led to a smoother distribution for higher scores between 0.75 and 1 (see Figure A.10).

6.3.3 Subscore impact

The leave-one-out approach to assess the impact of each subscore on s_{total}, showed that all subscores contribute to the conservation score (see Figure 6.4). For each comparison a Kolmogorov-Smirnov test was performed and all FDR-corrected p-values were significant at a level of 0.01.

Visually, leaving out the expression based subscores s_{exp} and s_{spec} led to the most drastic change in the distribution (see Figure 6.4) emphasising the importance of expression when assessing the conservation across species.

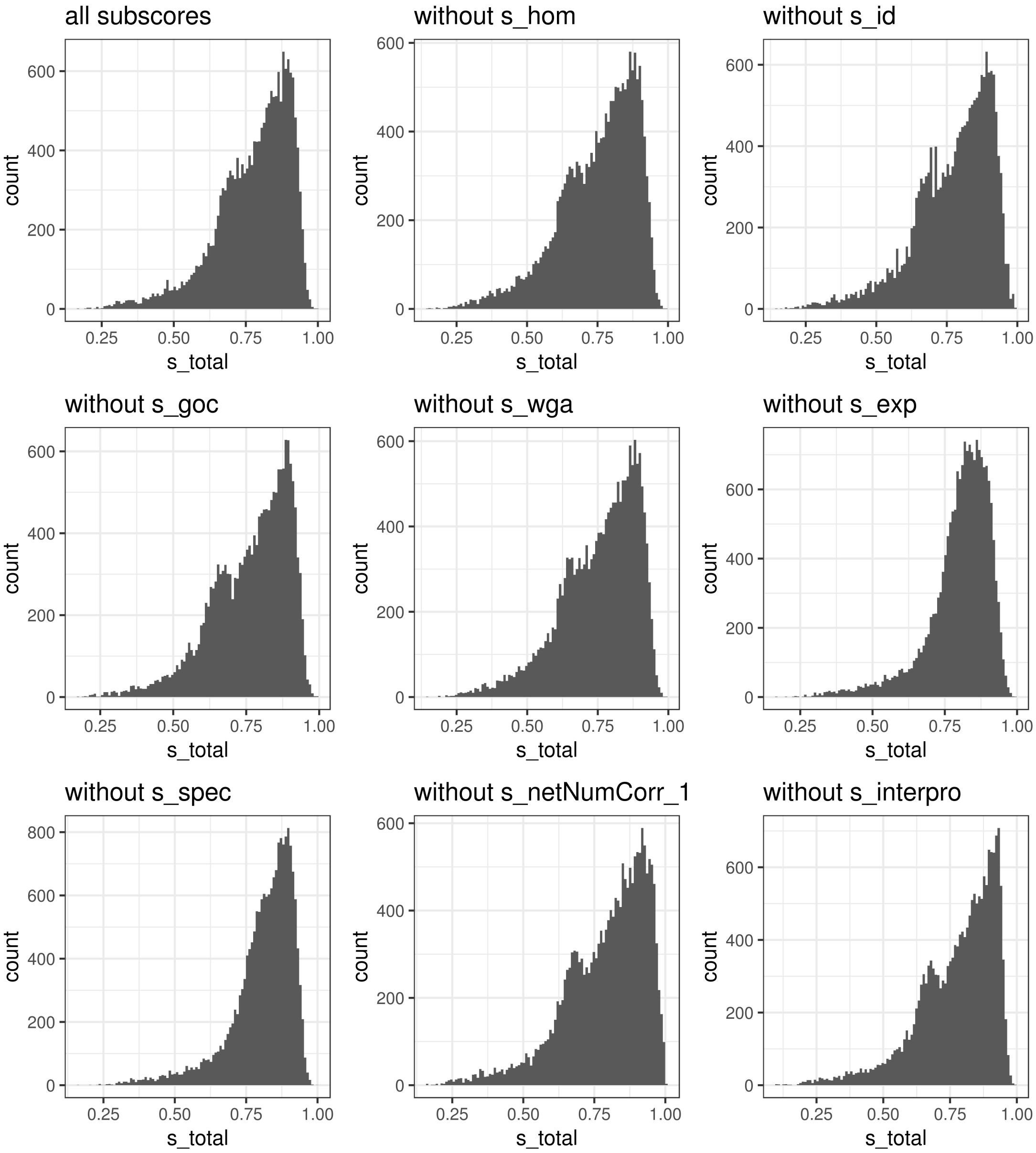


Figure 6.4: Distribution of the aggregated conservation score s_{final} using all subscores (top left) and when leaving out one subscore at a time.

6.3.4 GO enrichment for genes with high and low scores

For the group of low scoring genes (cluster 5 in Figure 6.5; cluster centre: 0.4), GO terms related to sensory detection of chemical stimuli and smell were the most overrepresented terms, followed by processes linked to the immune system (see Figure 6.6). Since mice and rats are a lot more dependent on their olfactory sense for survival than humans, it is reasonable that these genes have diverged between human and the two rodent species [32, 139]. Even though mice are commonly used as model species in immunological studies, several differences between the immune system of human and mouse have been reported [73].

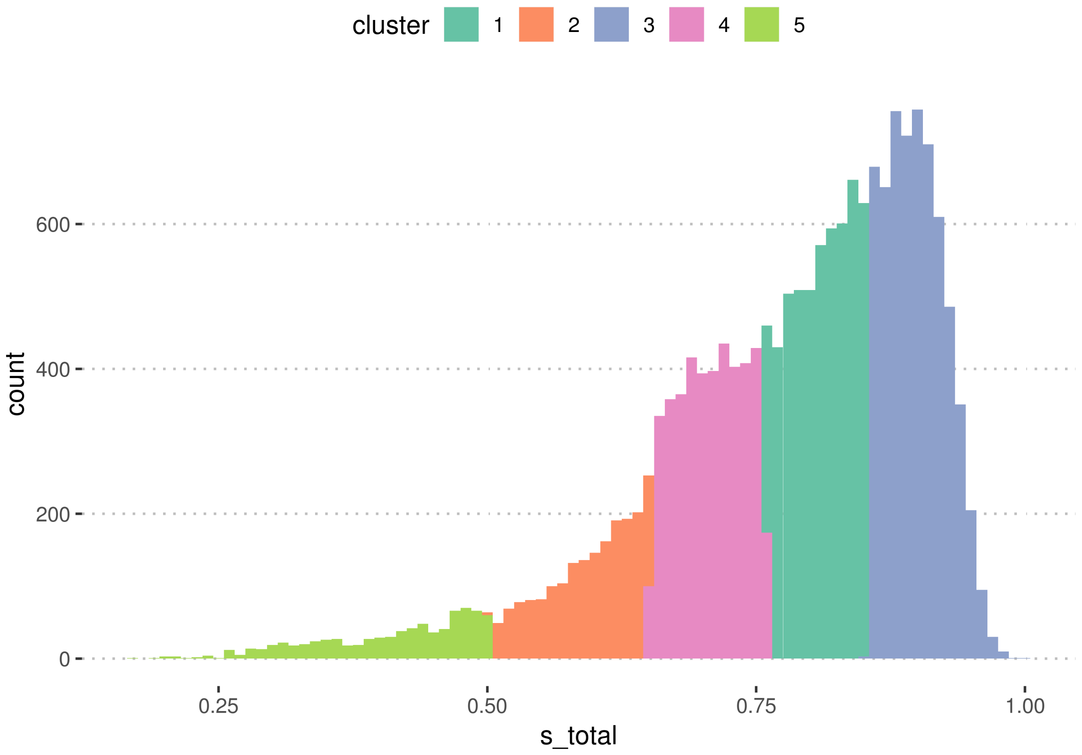


Figure 6.5: Grouping human protein-coding genes according to their conservation score using k-means clustering.

The group with high scoring genes (cluster 3 in Figure 6.5; cluster centre: 0.9) was enriched for terms related to synaptic signal transduction (see Figure 6.6) which makes sense as synapses are a common and conserved component of the mammalian nervous system. Furthermore, there is research confirming the conservation of gene expression patterns across brain regions between human and mouse [113].

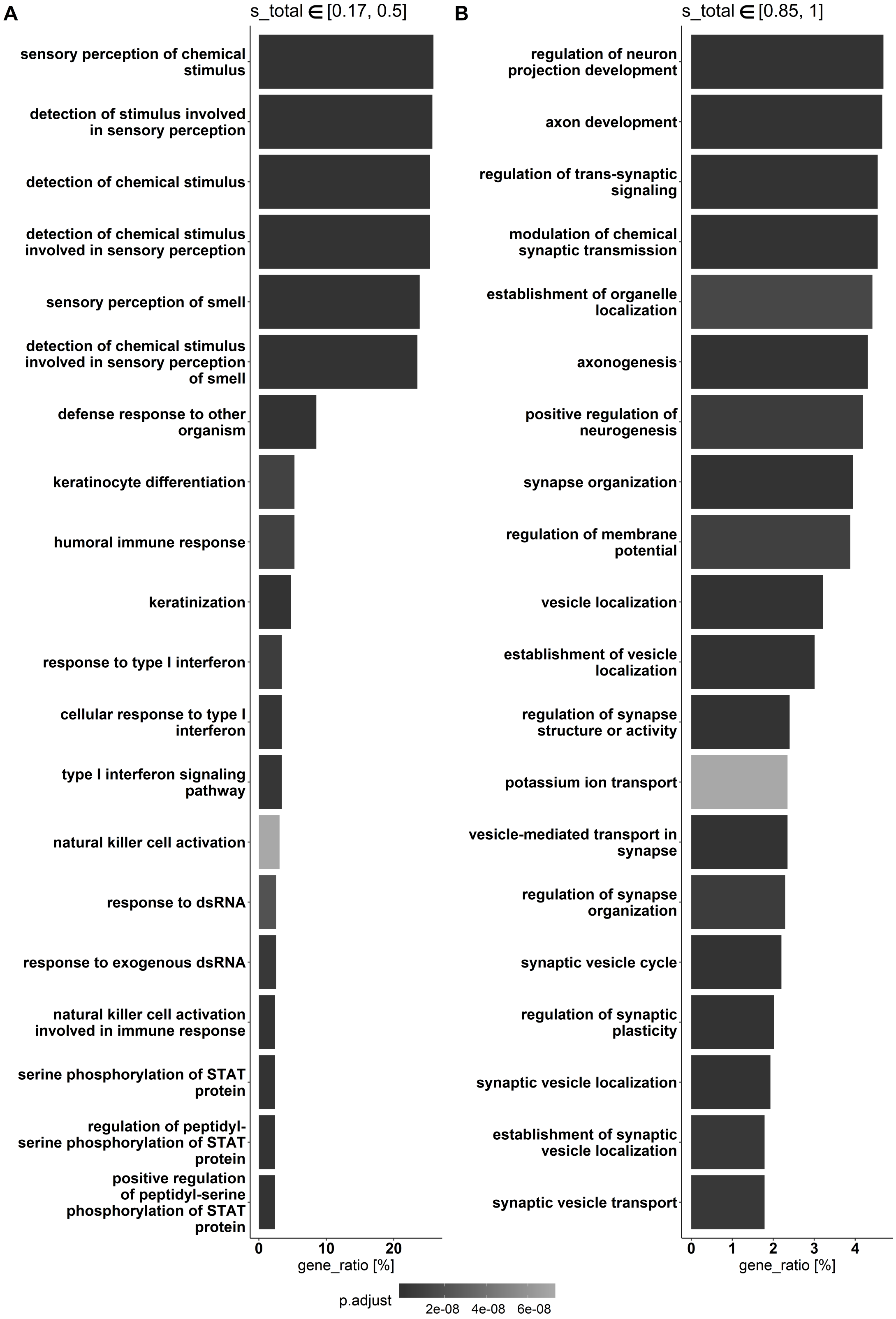


Figure 6.6: GO Biological Process terms overrepresented in the group of genes with low (A) and high (B) conservation scores. p.adjust corresponds to the FDR adjusted p-value and numbers on the x-axis indicate the percentage of genes in the cluster associated with each GO term.

6.3.5 Validation using MGI human - mouse disease connection data

The distribution of the 984 presumably well conserved genes retrieved from the HMDC data set showed a median species score of 0.86 in mouse (see Figure 6.7). None of the human genes reached a score of exactly 1 (maximum: 0.99) and the lowest score was 0.42. Comparing the mean $s_{species}$ for mouse of human genes in the HMDC set and those that were not in the HMDC set, we found that the mean species score is higher for HMDC genes (p-value: 9×10^{-27}).

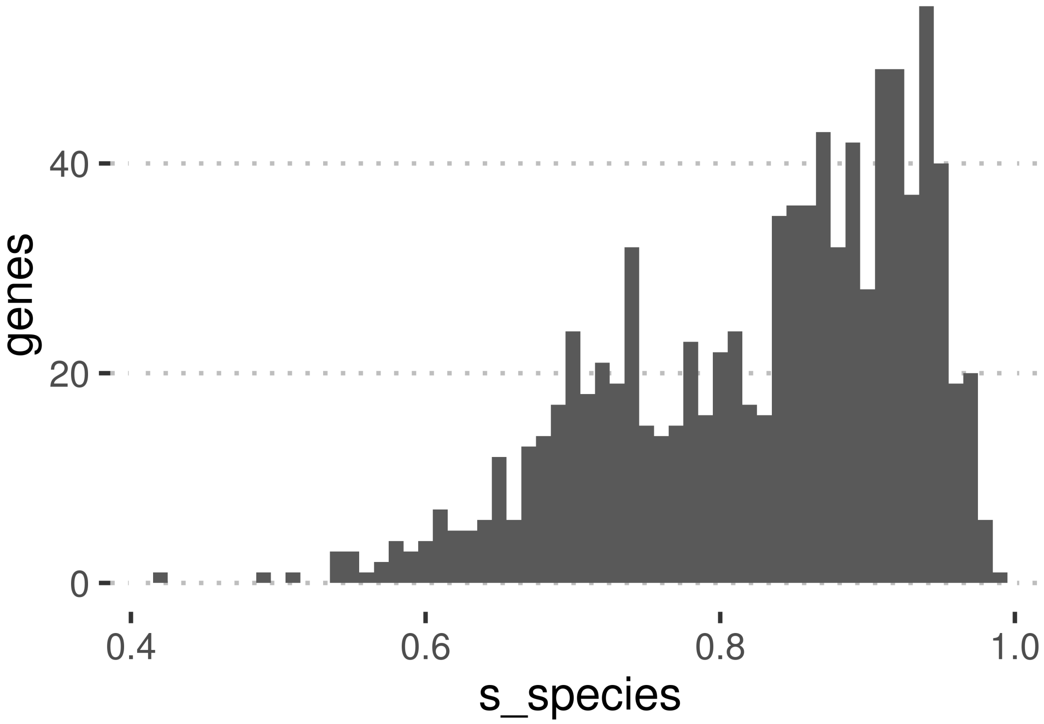


Figure 6.7: Distribution of the mouse species score (see Equation 6.1) for human genes in the HMDC validation data.

6.4 Discussion

Resources like Pharos or OpenTargets provide extensive information on human target genes and thereby support researchers during the often tedious process of prioritising drug targets. Although a target gene may have a strong link to the investigated disease in human, it might lead to problems farther down the drug development pipeline if, for example, one fails to find a suitable model species or effects observed in the animal model cannot be confirmed in human clinical trials.

In both cases the reason could be that the drug itself or its targeted protein or pathway are species specific and this aspect has not been addressed appropriately.

In this chapter we proposed a score to assess the conservation of a human target gene across model species. To our knowledge, we are the first to integrate sequence and gene expression based information to determine the degree of similarity between the human genes and its orthologues in several species.

By applying the proposed score to all human protein-coding genes using mouse and rat as the related species, we were able to investigate the different components of the score. The observed broad distribution of s_{id} could be due to diverging exon usage in the different species as it has previously been found that between 14 and 53 % of human alternative splice junctions are not conserved in mouse [116].

A leave-one-out analysis showed that all subscores contribute to the overall conservation score with the expression based ones playing a particularly important role. This emphasises that the different sequence and expression based subscores provide non-redundant information which is beneficial for the assessment of conservation across multiple species.

For the network subscore we compared two approaches, one only considering the number of highly correlated genes and one taking the orthology relationships between the two species into account. Even though the latter would provide valuable insight into conserved pathways in both species, we had to refrain from using it for our total score as it only covered a very narrow and low dynamic range and therefore shifted the conservation score to low values. We hypothesise, that this shift is caused by 1-to-many orthology relationships. If, for example, a human gene, which is highly correlated to the target gene, has 20 orthologues in mouse but only one of them is in the co-expression network of the target's orthologous gene, we get a low overlap between the two networks. On the one hand this effect could be attenuated by only using orthologous genes marked as "high confidence" based on s_{id}, s_{wga}, and s_{goc}. On the other hand we might introduce a bias into our analyses as we filter the input by the sequence related features we use to assess conservation.

In order to deduce some functional information from the co-expression networks, we tried to integrate Interpro annotation and compare them between the two species. This did, however, lead to very low dynamic ranges of the resulting network score, therefore, results were not shown in this thesis. Alternative annotations one could try to incorporate are Gene Ontology terms or KEGG pathways.

We also tried to extend the network score by not only taking direct neighbours into account but also those at distance two. However, we did not observe a difference between the two approaches and therefore did not present details here.

One critical point regarding the network based score is the choice of the correlation threshold. Considering the lack of a validation set of well conserved genes, we had to rely on our gut feeling when comparing the score distributions obtained with different thresholds. Therefore, we think that the the chosen threshold of 0.8 is suitable for the evaluation of the conservation score, but we do not claim that it is appropriate for all use cases. In mouse, for example, we observed that a more stringent threshold of 0.9 revealed three pronounced peaks in the score distribution which might contain scientifically interesting genes. A thorough evaluation of thresholds in different scenarios with a comprehensive validation data set is recommended.

The GO enrichment in low and high scoring human protein-coding genes revealed that these two groups contained genes which could be assumed to be poorly or well conserved based on their biological function. Since we only account for the expression in three tissues, with one of them being brain, the GO enrichment might be biased towards terms related to brain function. Therefore, we think it would be exciting to also incorporate a wider range of tissues.

Our attempt to validate the proposed conservation score using human-mouse disease connection data containing potentially highly conserved genes, showed that the majority of these genes are detected as highly conserved by the conservation score. There were, however, also genes which had a rather low score. We think that this is probably due to the data set also containing gene-disease links without reliable evidence.

Regarding the data used to compute the score, one has to keep in mind that maybe not all required information is available for all orthologous gene pairs. In the current implementation, these cases are discarded. Ideally, one would implement a reliability score reflecting the completeness of the data. A naive approach would deduce a certain value from one for each part of information that is missing.

Furthermore, we have to emphasise, that in the current implementation, only mouse and rat were used to compute the conservation score. The database and the Nextflow pipeline are, however, designed such that one can easily incorporate additional species.

In summary, we have presented a score which has great potential to decrease the amount of animal testing, time, and cost during the drug development process. Its primary application is target gene ranking based on the conservation level across several species of interest. By comparing the conservation score of a certain gene set to the scores of all human protein-coding genes, one could draw conclusions about how well, for example, a specific pathways is conserved across different species. Another important use case is determining the best model species for a target gene by comparing the aggregated species scores of several species. In retrospect, one could also use the conservation score to check whether an observed species specific drug response can be explained by differences on the sequence or expression level.

Chapter 7

Conclusion & outlook

In this thesis we wanted to systematically evaluate how gene expression across normal tissues is linked to orthology in several commonly used mammalian model species. Increased understanding of this link can then help to reduce animal testing in biomedical research as it can, for example, be used to assess a drug target's potential to successfully translate from animal models to human patients.

The analysis of RNA-Seq data sets from mouse and rat containing samples from a wide range of tissues, has shown that the variability in gene expression between tissues was greater than the variability between animals. This was preserved when combining the data from both species, however, a species effect was observed, too. Comparing sequence identity and gene expression correlation of one-to-one orthologues, we found that the majority are highly similar on sequence and expression level. There were, however, cases with high expression correlation despite a low sequence identity. We hypothesised that some of these cases are due to incorrectly or incompletely annotated sequences in the public databases like Ensembl.

We systematically estimated the number of genes which might be poorly annotated in mouse, rat, dog, pig, and cyno. The results indicated that there is a substantial number of sequences in Ensembl that apparently contain errors or are incomplete. Using a curated bait sequence from a closely related species and RNA-Seq data of the species of interest, we were able to improve a large proportion of these sequences. With the a&o-tool we have developed a tool performing this otherwise tedious process in an easy-to-use and automated manner. The a&o-tool has already been successfully used in a number of drug discovery projects at Boehringer Ingelheim and has thereby proven its value for biomedical research although details cannot be shown here.

One issue we see with our approach is, that our results heavily depend on the quality of the *de novo* transcriptome assembly. We tried two assemblers on our data and did not observe great differences. A recent review [46] has, however, nicely point out, that the choice of the assembler can greatly impact the quality of the resulting assembly and they concluded that it is best to try different assemblers as there is not a single one performing best on all different data sets they used for their study. Long-read sequencing technologies have the potential to solve this problem as they are able to sequence most transcripts as a whole, rendering the assembly step obsolete.

To investigate the potential of long-read sequencing technologies, we have conducted a pilot study using PacBio's SMRT sequencing. Basic quality control of the PacBio results showed that the pilot run was successful, but not optimal. Most PacBio transcripts were confirmed by aligning them to the rat genome and to human well curated protein sequences as well as by the comparison to short-read data. We did, however, observe potential quality issues at the 5' ends of some transcripts which might be eliminated by using a different sample preparation protocol. On the other hand these observations might reflect, at least partially, real transcripts with alternative transcription start sites. Therefore, we recommend to conduct another PacBio sequencing project as well as one using Oxford Nanopore's long-read sequencing platform to gain greater insight into the maturity of these technologies and their applicability in pharmaceutical research.

With the target conservation score, we were the first to integrate sequence and gene expression based metrics to assess the conservation of a human gene across several species. First validation attempts implied that the score is capable of differentiating genes with presumably low conservation from those that, judging by the biological processes they are involved in, should be highly conserved.

Further research regarding the subscore derived from the comparison of co-expression networks would greatly improve the conservation score. We think it would be very interesting to incorporate pathway information to better capture common biological activity.

A graphical user interface to facilitate easy data retrieval from targetcon would be desirable. First of all, precomputed scores and the underlying information could be queried for a single human target gene. These results could then be used for target prioritisation or the choice of the most suitable model species.

Another feature could include the upload of a gene list or pathway for which we could determine how well they are conserved in comparison to all human protein-coding genes.

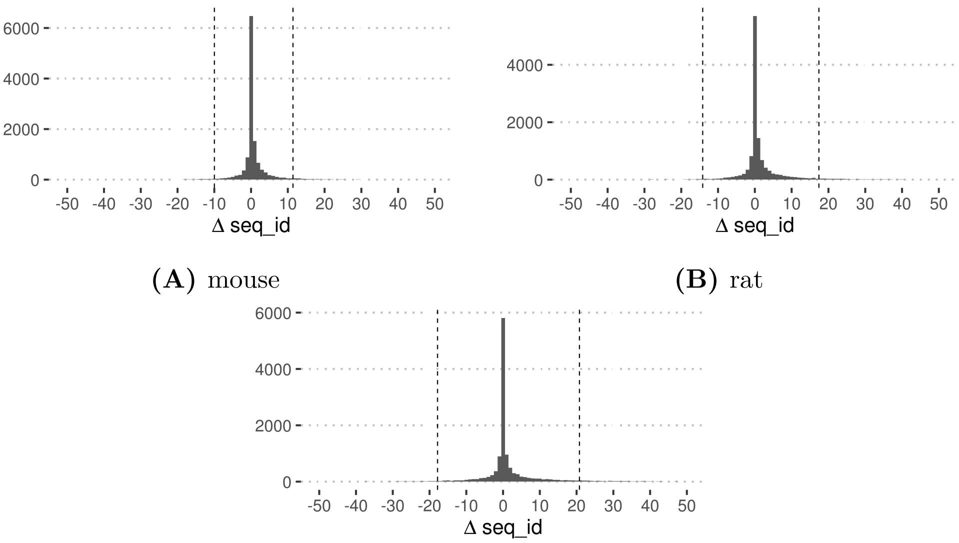
In contrast to only using precomputed scores, we could also provide a more dynamic approach where the user can compute the score for a specific target or a list of targets based on custom expression data to tailor the results to their scientific question.

In summary, we have shown that by exploiting sequence similarity of orthologues and their gene expression correlation, the tools and methods developed in this thesis contribute to the improvement of the drug discovery process. Our work helps to prevent mistakes being made due to wrongly chosen model species or wrong dose selection based on an incorrect sequence in toxicological experiments.

Appendix A

Appendix

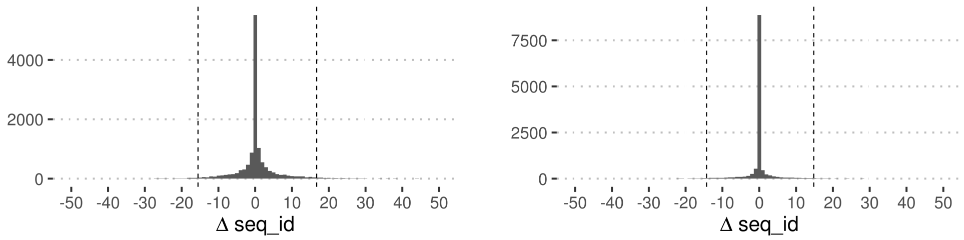
A.1 a&o-tool



(A) mouse

(B) rat

(C) dog



(D) pig

(E) cynomolgus monkey

Figure A.1: Distribution of the difference in sequence identities for all human genes having a one-to-one orthologue in the respective species. The target sequence identity corresponds to the percentage of the orthologous sequence matching the human sequence in the amino acid sequence alignment, and query identity is the percentage of human sequence matching the orthologous sequence. Dashed lines mark the threshold (mean +/- 2 times standard deviation) for considering a gene for refinement.

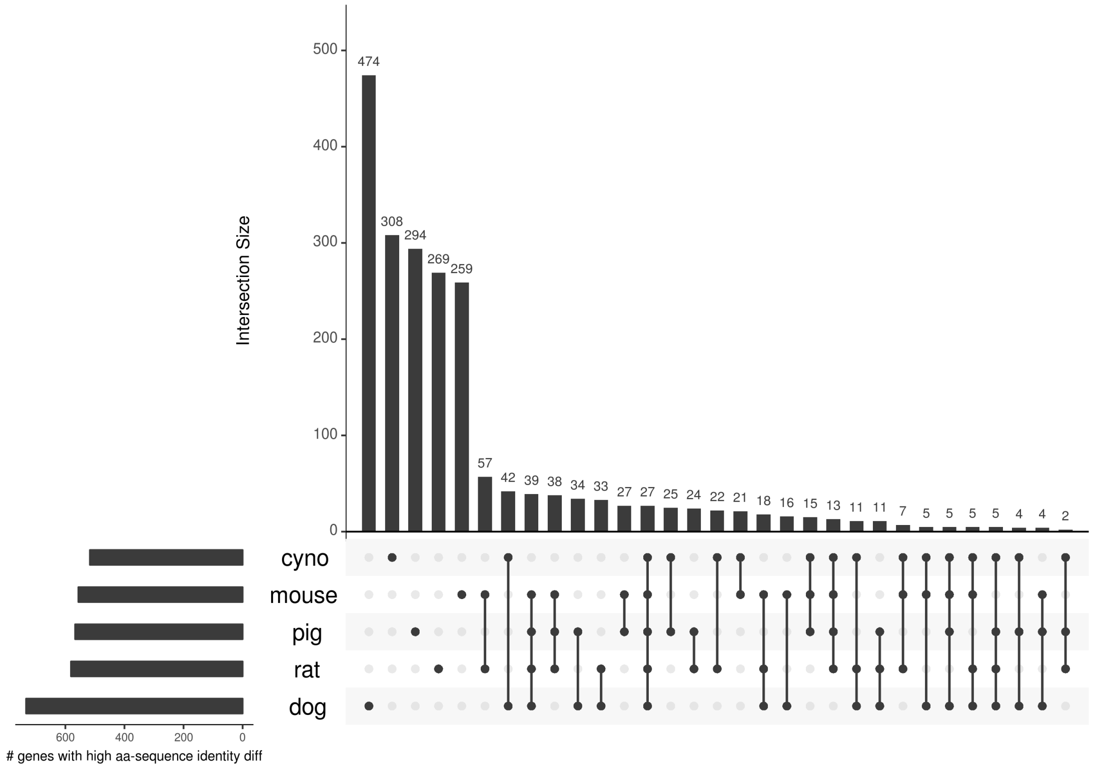


Figure A.2: An UpSet plot depicting the intersections between genes identified as potentially poorly annotated across species.

Table A.1: Quantitative metrics for the tissue-specific assemblies computed by TransRate. The number of contigs, the number of base pairs in the longest contig and the mean contig length provide information on the basic characteristics of the assembly. The number of contigs containing an open reading frame (# contigs with ORF) and the mean percentage of the contig being covered by the ORF (coverage of contigs with ORF [%]) help to assess the protein-coding potential of the assembled contigs.

species	tissue	# contigs	max. contig length [*bp*]	mean contig length [*bp*]	# contigs with ORF	coverage of contigs with ORF [%]
cyno	brain	346391	32640	1230.6	62637	34.7
cyno	liver	317055	32053	1153.4	53337	34.8
cyno	kidney	325290	27390	1180.6	56922	35.0
dog	brain	44703	27224	1614.5	20315	51.2
dog	liver	35143	17944	1321.3	15247	53.9
dog	kidney	38950	22476	1415.3	18019	53.5
human	brain	165810	30531	1367.9	42581	39.8
human	liver	88083	27458	1383.3	29226	45.9
human	kidney	128530	23822	1355.6	37009	43.6
mouse	brain	68734	23400	1548.1	24630	46.2
mouse	liver	35944	17404	1396.1	14697	52.8
mouse	kidney	42561	21694	1555.8	18036	51.4
pig	brain	61933	24820	1518.6	24822	49.9
pig	liver	37620	18047	1279.3	16114	55.3
pig	kidney	39922	17558	1343.6	18032	54.4
rat	brain	59378	21744	1605.8	22829	46.4
rat	liver	52636	17162	1308.2	17576	48.3
rat	kidney	42176	28641	1616.8	18856	51.2

A.2 PacBio pilot study

Table A.2: Details on intermediate results of the PacBio analysis.

metric	value	description
polymerase read bases	9,493,170,789	total number of sequenced bases
polymerase reads	402,505	total number of polymerase reads
polymerase read length (mean)	24,135	—
subreads	5,313,295	total number of subreads
subread length (mean)	1,787	—
CCS	264,467 (65.71%)	total number of generated circular consensus sequences (CCS), percentage in relation to all available polymerase reads
full-length reads	234,476	total number of CCS containing 5' and 3' adapter sequences and poly-A sequence
full-length reads, non-chimeric (FLNC)	231,025	—
FLNC mean read length	2,102	—
non-full-length reads	29,968	—
HQ isoforms	20,879	candidate error-corrected high-quality transcripts
LQ isoforms	108,664	candidate error-corrected low-quality transcripts
HQ isoform length (min)	381	—
HQ isoform length (max)	20,879	—
HQ isoform length (mean)	2,086	—
unique isoforms	8,106	HQ isoforms collapsed into unique set of transcript isoforms by aligning them to the reference genome

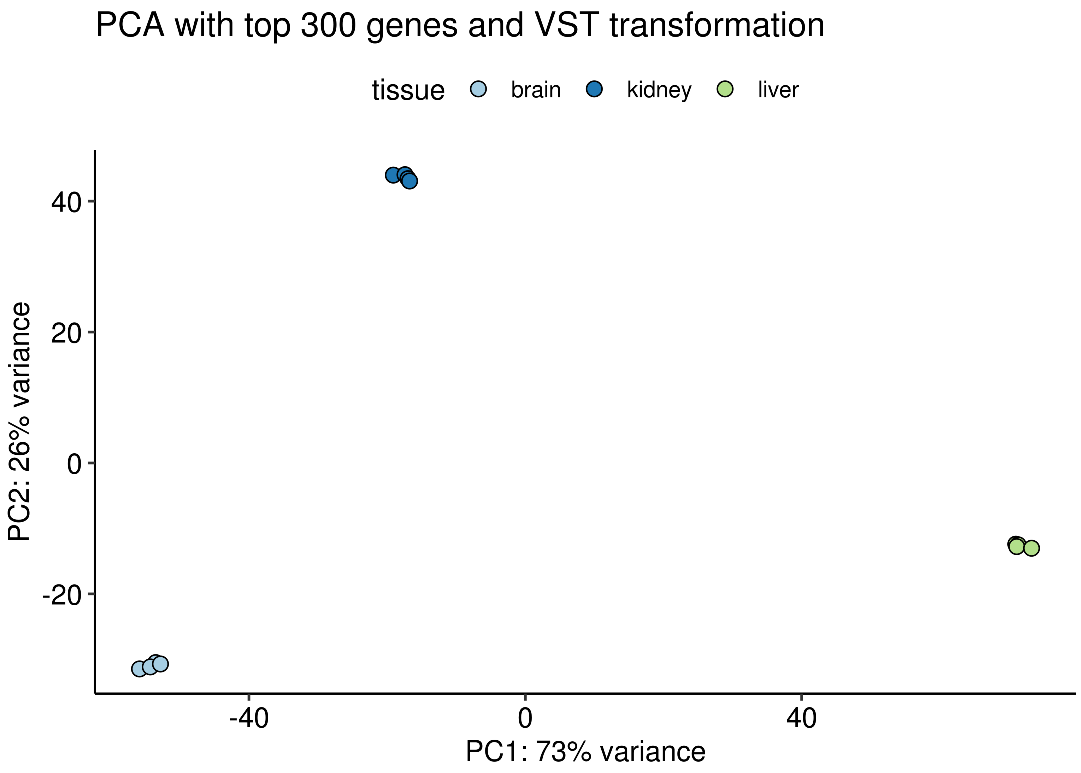


Figure A.3: Principal component analysis of expression from Illumina short-read data. Colours represent tissues. The numbers in the axis labels correspond to the variance explained by the principal component.

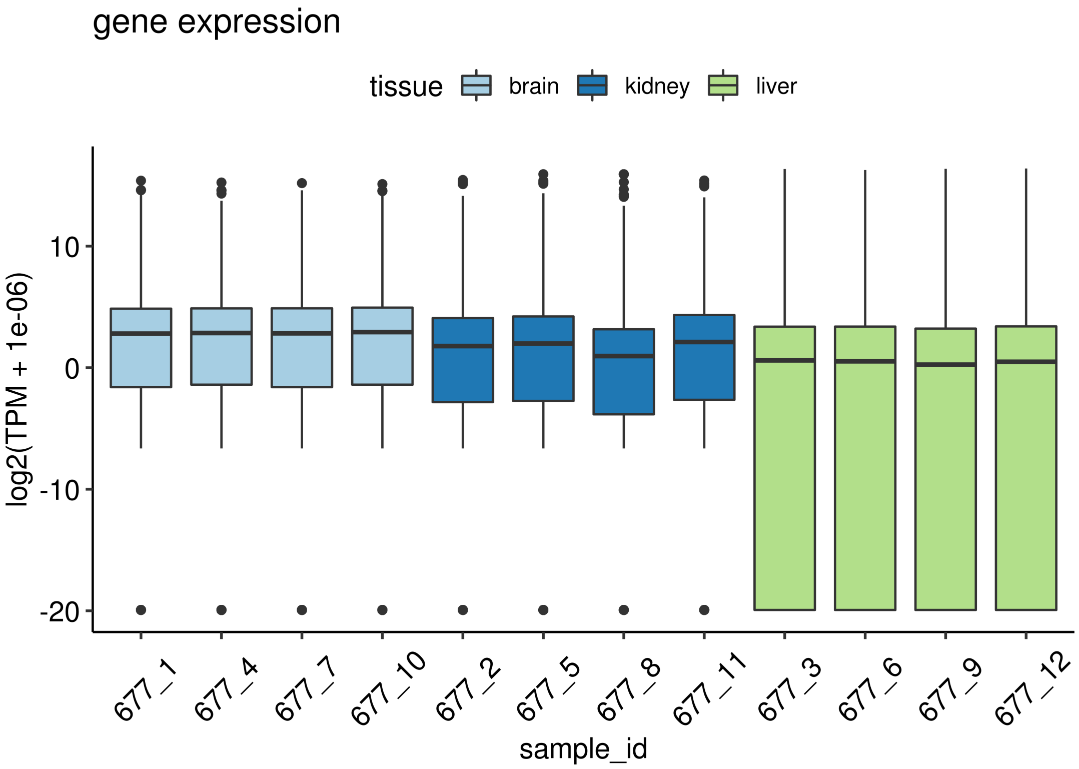


Figure A.4: Distribution of logarithmised TPM expression values of Illumina samples. Colours correspond to tissues.

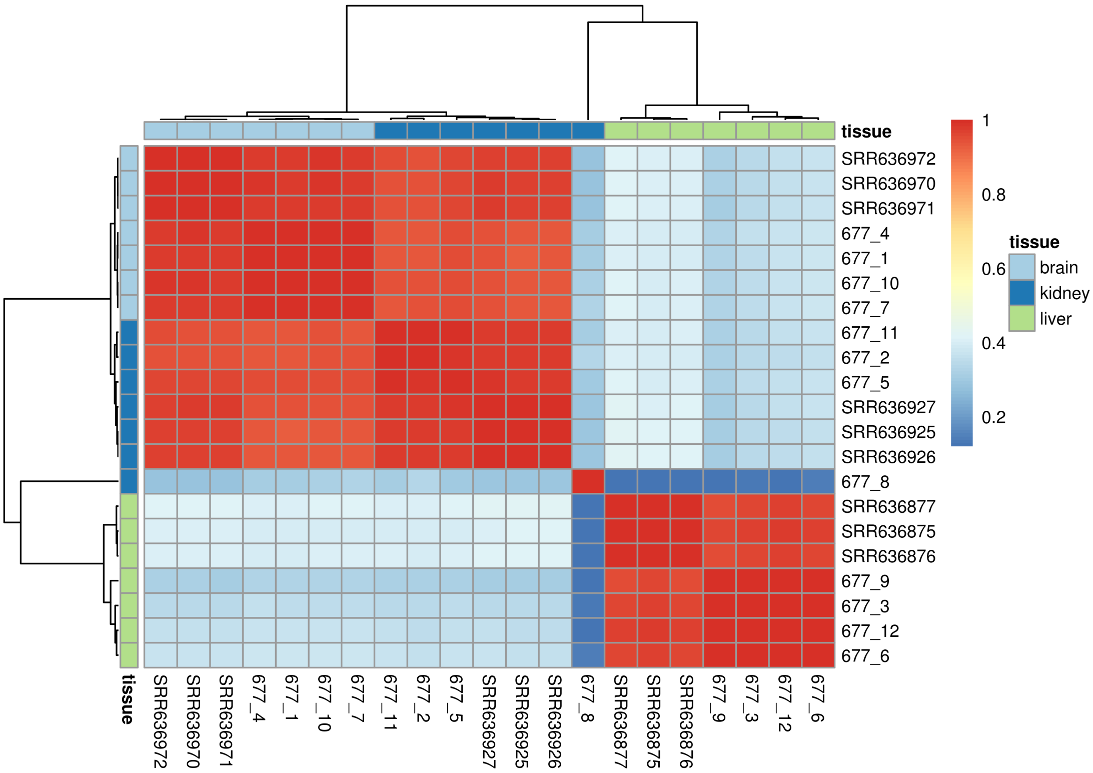


Figure A.5: Clustered sample correlation (Pearson's correlation coefficient) matrix. Sample names starting with "SRR" are the Fushan et al. data and those with the prefix 677 are the internal Illumina samples.

Table A.3: Basic information on the 12 rat samples regarding the animal and tissue they came from. “total_sequences” refers to the number of Illumina reads and “uniquely_mapped” is the percentage of these reads that were uniquely mapped to the Ensembl *Rattus norvegicus* (Rnor_6.0) reference genome.

sample_id	tissue	animal	total_sequences	uniquely_mapped [%]
677_1	brain	101	26,247,834	88.65
677_2	kidney	101	26,894,820	85.93
677_3	liver	101	29,145,488	86.23
677_4	brain	102	31,218,437	88.32
677_5	kidney	102	31,027,119	86.25
677_6	liver	102	30,773,234	86.50
677_7	brain	103	30,357,943	88.26
677_8	kidney	103	30,513,775	74.14
677_9	liver	103	29,950,055	86.38
677_10	brain	104	28,638,346	89.28
677_11	kidney	104	33,267,517	86.34
677_12	liver	104	30,099,932	86.76

Table A.4: Contingency table contrasting the reciprocal best BLAST hit category of PacBio isoforms with their associated gene and transcript category assigned by SQANTI. Isoforms classified as “known” mapped to an annotated gene/transcript, “novel” means the isoforms were not mapped to an annotated gene/transcript and “novel_known_AS” refers to PacBio isoforms that “overlap the complementary strand of an annotated transcript”[114].

	SQANTI_gene_category			SQANTI_transcript_category	
	known	novel	novel_known_AS	known	novel
notRBH	3366	510	83	1753	2206
RBH	4109	22	16	2872	1275

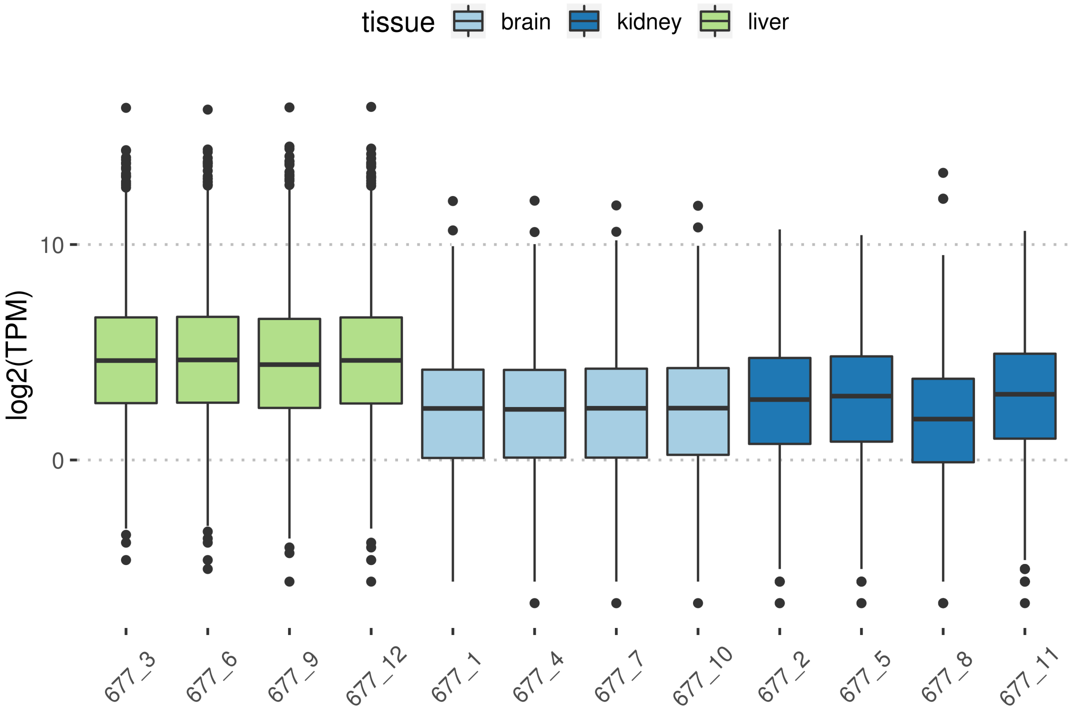


Figure A.6: Expression of 2125 liver-specific genes from Illumina short-read data.

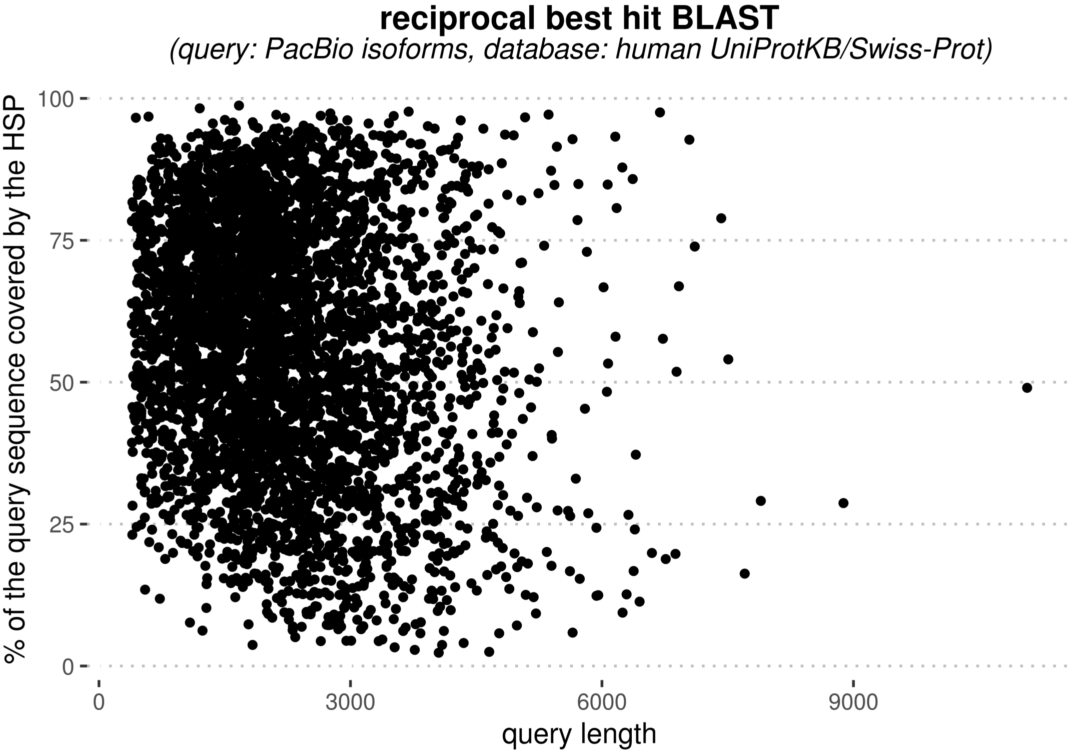


Figure A.7: Comparison of the query length and the percentage of the query covered by the HSP. Each dot corresponds to the numbers in the forward search of an RBH.

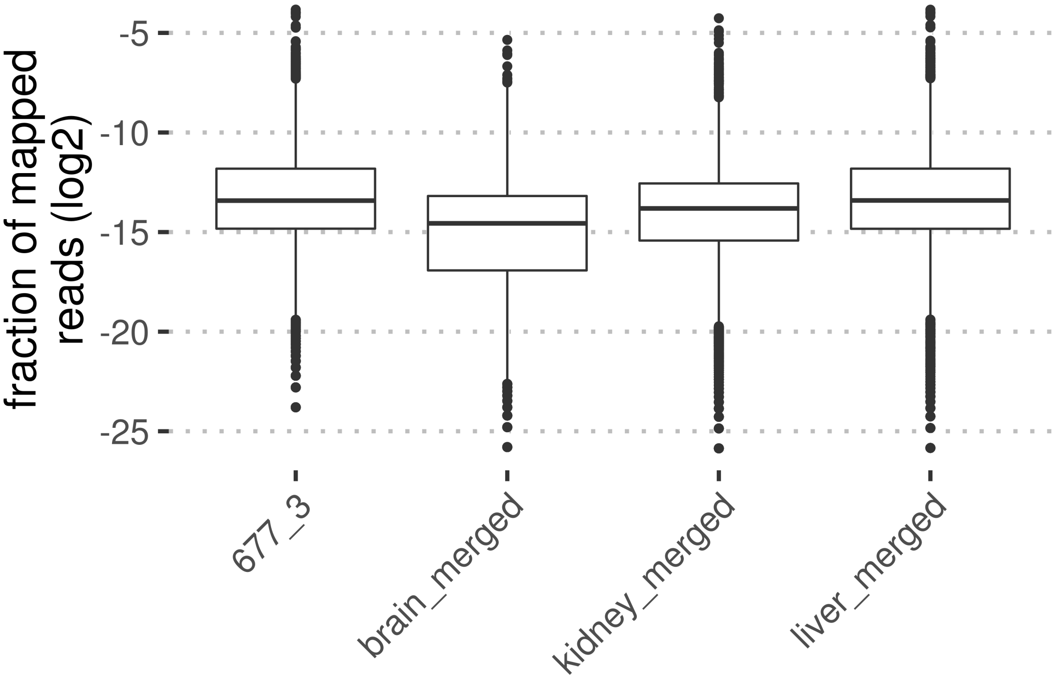


Figure A.8: Distribution of mapped reads normalised by the number reads in the sample.

A.3 Conservation score

Table A.5: Input parameters for the targetcon Nextflow pipeline provided via a configuration file in JSON format.

field name	description
tgtcon_dbVersion	version of the PostgreSQL database to be created
tgtcon_ensemblVersion	Ensembl version to be used
tgtcon_host	host where the database is to be created
tgtcon_user	user with all rights required to create a database
tgtcon_pwd	password for tgtcon_user
outdir	directory where pipeline output should be stored
target_species	list of target species to be used during score calculation
medExp_input	path to a tab-separated file with the columns “species”, “rpkm_path”, and “design_path”; “species” must match target_species
study_input	path to a tab-separated file with the columns “species”, “study_name”, i.e., the name to be stored in the database to differentiate expression data sets; “species” must match target_species

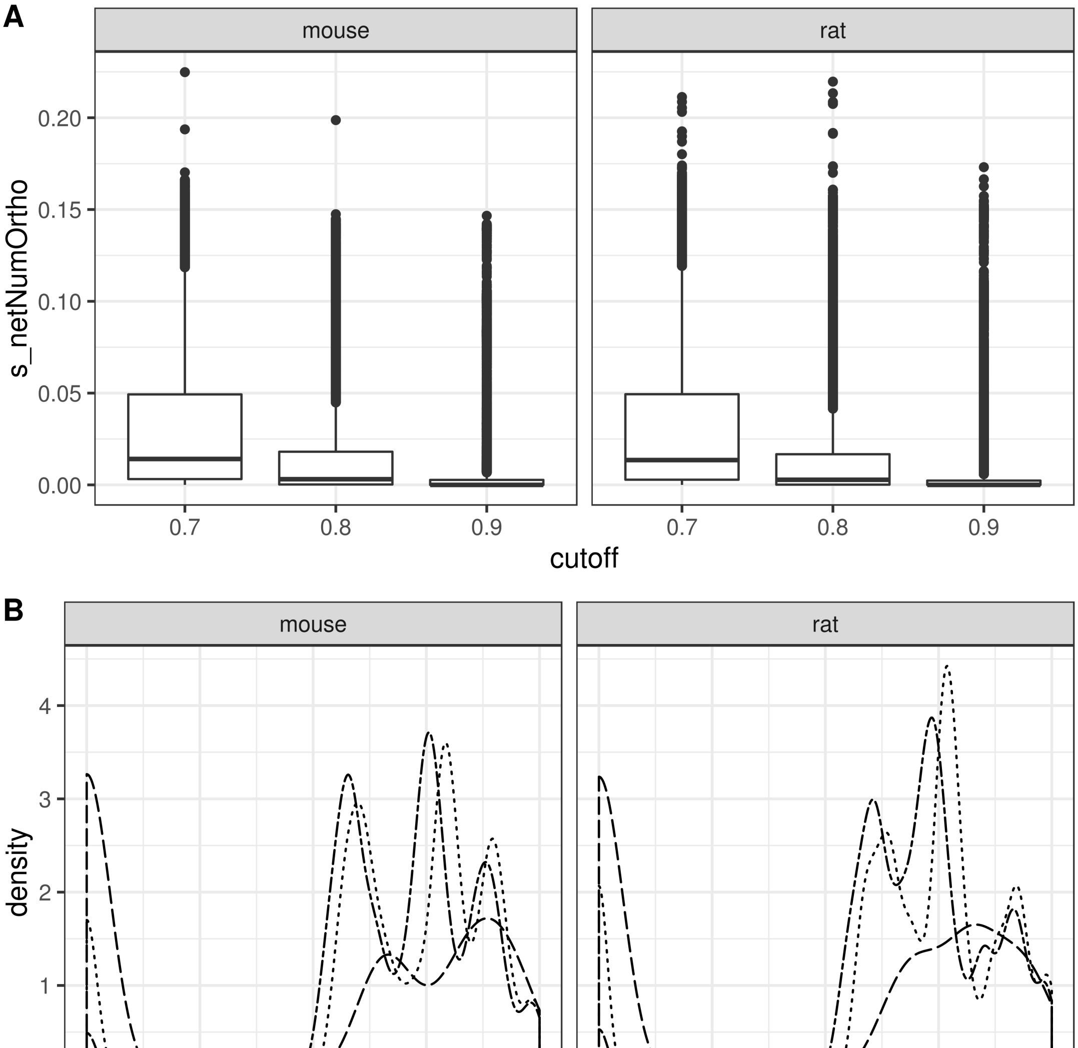


Figure A.9: Distribution of the network score based on (A) the overlap of highly correlated genes and (B) the number of highly correlated genes between human and the two rodent species with varying correlation thresholds.

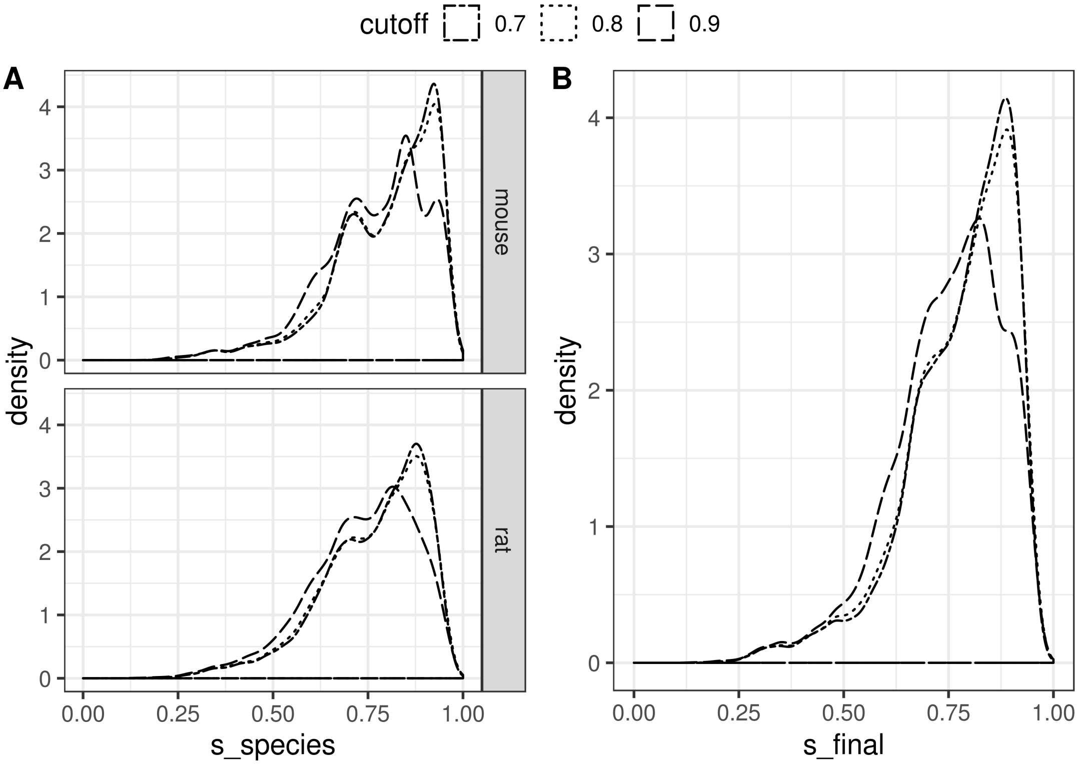


Figure A.10: Impact of varying correlation thresholds on the aggregated species (A) and total conservation score (B).

A.4 List of publications

- **J. F. Söllner**, G. Leparc, M. Zwick, T. Schönberger, T. Hildebrandt, K. Nieselt and E. Simon. *Exploiting orthology and de novo transcriptome assembly to refine target sequence information.* BMC Medical Genomics 2019: 12, 69

- **J. F. Söllner**, G. Leparc, T. Hildebrandt, H. Klein, L. Thomas, E. Stupka and E. Simon. *An RNA-Seq atlas of gene expression in mouse and rat normal tissues.* Scientific Data 2017: 4

- The following publications originated from my post-graduate research at the Institute of Computational Biology at the Helmholtz Zentrum München. As part of the statistics team, I participated in the Prostate Cancer DREAM Challenge and we published our results in an F1000Research article. The overall results of the DREAM Challenge were published in Lancet Oncology and JCO Clinical Cancer Informatics.

    - I. Kondofersky, M. Laimighofer, C. Kurz, N. Krautenbacher, **J. F. Söllner**, P. Dargatz, H. Scherb, D. P. Ankerst, C. Fuchs. *Three general concepts to improve risk prediction: good data, wisdom of the crowd, recalibration [version 1; peer review: 2 approved with reservations].* F1000Research 2016: 5:2671 (Together with the other authors, I preprocessed the data, established first analyses, and contributed to the manuscript.)
    - J. Guinney, T. Wang, T. D. Laajala, et al. and **the Prostate Cancer Challenge DREAM Community**. *Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data.* Lancet Oncology 2017: 18, 132-42
    - F. Seyednasrollah, D. C. Koestler, T. Wang, et al. and **Prostate Cancer DREAM Challenge Community**. *A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer.* JCO Clinical Cancer Informatics 2017: 1, 1-15

Bibliography

[1] About the Human - Mouse: Disease Connection at the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine. `http://www.informatics.jax.org/mgihome/projects/aboutHMDC.shtml`. Accessed: 2019-07-18.

[2] The human protein atlas. `https://www.proteinatlas.org/about/assays+annotation`. Accessed: 2019-03-10.

[3] The Human - Mouse: Disease Connection (HMDC) at the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine. `http://www.informatics.jax.org/downloads/reports/MGI_DO.rpt`. Accessed: 2019-07-18.

[4] 1000 Genome Project Data Processing Subgroup, A. Wysoker, B. Handsaker, G. Marth, G. Abecasis, H. Li, J. Ruan, N. Homer, R. Durbin, and T. Fennell. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[5] Y. Adeleye, M. Andersen, R. Clewell, M. Davies, M. Dent, S. Edwards, P. Fowler, S. Malcomber, B. Nicol, A. Scott, S. Scott, B. Sun, C. Westmoreland, A. White, Q. Zhang, and P. L. Carmichael. Implementing Toxicity Testing in the 21st Century (TT21C): Making safety decisions using toxicity pathways, and progress in a prototype risk assessment. *Toxicology*, 332:102–111, 2015.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[7] A. Ameur, W. P. Kloosterman, and M. S. Hestand. Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnology*, 37(1):72–85, 2019.

[8] V. N. Anisimov, S. V. Ukraintseva, and A. I. Yashin. Cancer in rodents: does it tell us about cancer in humans? *Nature Reviews Cancer*, 5(10):807–819, 2005.

[9] S. Y. Anvar, G. Allard, E. Tseng, G. M. Sheynkman, E. de Klerk, M. Vermaat, R. H. Yin, H. E. Johansson, Y. Ariyurek, J. T. den Dunnen, S. W. Turner, and P. A. C. ‘t Hoen. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biology*, 19(1):46, 2018.

[10] H. Attarwala. TGN1412: From Discovery to Disaster. *Journal of Young Pharmacists*, 2(3):332 – 336, 2010.

[11] K. F. Au, V. Sebastiano, P. T. Afshar, J. D. Durruthy, L. Lee, B. A. Williams, H. van Bakel, E. E. Schadt, R. A. Reijo-Pera, J. G. Underwood, and W. H. Wong. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):E4821–E4830, 2013.

[12] J. A. Ballesteros and G. Hormiga. A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Molecular Biology and Evolution*, 33(8):2117–2134, 2016.

[13] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. O. N. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. X. A. Alekseyev, and P. A. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[14] M. Beilmann, H. Boonen, A. Czich, G. Dear, P. Hewitt, T. Mow, P. Newham, T. Oinonen, F. Pognan, A. Roth, J.-P. Valentin, F. Van Goethem, R. Weaver, B. Birk, S. Boyer, F. Caloni, A. Chen, R. Corvi, M. Cronin, M. Daneshian, L. Ewart, R. Fitzgerald, G. Hamilton, T. Hartung, J. Kangas, N. Kramer, M. Leist, U. Marx, S. Polak, C. Rovida, E. Testai, B. van der Water, P. Vulto, and T. Steger-Hartmann. Optimizing drug discovery by investigative toxicology: Current and future trends. *ALTEX - Alternatives to animal experimentation*, 36(2):289–313, 2019.

[15] Bioinformatics Group at the Babraham Institute. FastQC: A quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`. Accessed: 2018-11-05.

[16] J. Bruijnesteijn, M. K. H. van der Wiel, N. de Groot, N. Otting, A. J. M. de Vos-Rouweler, N. M. Lardy, N. G. de Groot, and R. E. Bontrop. Extensive Alternative Splicing of KIR Transcripts. *Frontiers in Immunology*, 9:2846, 2018.

[17] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

[18] D. Carvalho-Silva, A. Pierleoni, M. Pignatelli, C. Ong, L. Fumis, N. Karamanis, M. Carmona, A. Faulconbridge, A. Hercules, E. McAuley, A. Miranda, G. Peat, M. Spitzer, J. Barrett, D. G. Hulcoop, E. Papa, G. Koscielny, and I. Dunham. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065, 2019.

[19] X. Chen and J. Zhang. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. *PLoS Computational Biology*, 8(11):e1002784, 2012.

[20] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016.

[21] D. Cook, D. Brown, R. Alexander, R. March, P. Morgan, G. Satterthwaite, and M. N. Pangalos. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6):419–431, 2014.

[22] D. E. Cook, J. E. Valle-Inclan, A. Pajoro, H. Rovenich, B. P. H. J. Thomma, and L. Faino. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. *Plant Physiology*, 179:38–54, 2019.

[23] P. Dehal, P. Predki, A. S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C. L. Ecale Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M. J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb, and L. Stubbs. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science*, 293(5527):104–111, 2001.

[24] D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012.

[25] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[26] A. Dobin, C. A. Davis, C. Zaleski, F. Schlesinger, J. Drenkow, M. Chaisson, P. Batut, S. Jha, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2012.

[27] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. D. Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological

databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

[28] S. Durnick, P. T. Spellman, E. Birney, and W. Huber. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191, 2009.

[29] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[30] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009.

[31] EMBL-EBI. Orthology quality-controls. `https://www.ensembl.org/info/genome/compara/Ortholog_qc_manual.html#wga`. Accessed: 2019-07-29.

[32] R. D. Emes, S. A. Beatson, C. P. Ponting, and L. Goodstadt. Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Research*, 14(4):591–602, 2004.

[33] P. A. Ewels, A. Peltzer, S. Fillinger, J. Alneberg, H. Patel, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen. nf-core: Community curated bioinformatics pipelines. *bioRxiv*, 2019.

[34] M. Failli, J. Paananen, and V. Fortino. Prioritizing target-disease associations with novel safety and efficacy scoring methods. *Scientific Reports*, 9(1):9852, 2019.

[35] W. M. Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19(2):99–113, 1970.

[36] W. M. Fitch. Homology a personal view on some of the problems. *Trends in Genetics*, 16(5):227–231, 2000.

[37] A. A. Fushan, A. A. Turanov, S. G. Lee, E. B. Kim, A. V. Lobanov, S. H. Yim, R. Buffenstein, S. R. Lee, K. T. Chang, H. Rhee, J. S. Kim, K. S. Yang, and V. N. Gladyshev. Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, 14(3):352–365, 2015.

[38] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, 2011.

[39] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333–351, 2016.

[40] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644–652, 2011.

[41] Z. R. Gregorich and Y. Ge. Top-down proteomics in health and disease: Challenges and opportunities. *Proteomics*, 14(10):1195–1210, 2014.

[42] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, D. Philip, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. D. Macmanes, M. Ott, J. Orvis, and N. Pochet. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, 8(8):1–43, 2013.

[43] S. Hansen and R. G. Q. Leslie. TGN1412: scrutinizing preclinical trials of antibody-based medicines. *Nature*, 441:282, 2006.

[44] J. Herrero, M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. J. Searle, R. Amode, S. Brent, W. Spooner, E. Kulesha, A. Yates, and P. Flicek. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016.

[45] K. Herrmann, F. Pistollato, and M. Stephens. Beyond the 3rs: Expanding the use of human-relevant replacement methods in biomedical research. *ALTEX - Alternatives to animal experimentation*, 36(3):343–352, Jul. 2019.

[46] M. Hölzer and M. Marz. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience*, 8(5), 2019.

[47] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, and L. Gatto. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015.

[48] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork. eggNOG 4.5: a hierarchical orthology framework

with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1):D286–D293, 2015.

[49] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162:1239–1249, 2011.

[50] Illumina. *De Novo* Assembly Using Illumina Reads. `https://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf`. Accessed: 2019-06-14.

[51] Illumina. Illumina sequencing platforms. `https://www.illumina.com/systems/sequencing-platforms.html`. 2019-04-28.

[52] Illumina. Introduction to SBS Technology. `https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html`. Accessed: 2019-05-20.

[53] P. Jaccard. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bulletin de la Société vaudoise des Sciences Naturelles*, 37:241–272, 1901.

[54] G. Jun, M. K. Wing, G. R. Abecasis, and H. M. Kang. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research*, 25(6):918–925, 2015.

[55] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

[56] O. Keller, F. Odronitz, M. Stanke, M. Kollmar, and S. Waack. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, 9, 2008.

[57] E. H. Kerns and L. Di. *Drug-like Properties: Concepts, Structure Design and Methods*. Elsevier, 2008.

[58] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, and L. M. Schriml. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, 43:D1071–D1078, 2015.

[59] J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[60] N. Kryuchkova-Mostacci and M. Robinson-Rechavi. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs , and Rapidly between Paralogs. *PLOS Computational Biology*, 12(12):1–13, 2016.

[61] S. Kumar, G. Stecher, M. Suleski, and S. B. Hedges. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7):1812–1819, 2017.

[62] R. Kuo. tama. `https://github.com/GenomeRIK/tama`. Accessed: 2019-08-08.

[63] G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PLoS One*, 12(5):e0177459, 2017.

[64] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[65] J. Liu, G. Li, Z. Chang, T. Yu, B. Liu, R. McMullen, P. Chen, and X. Huang. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. *PLoS Computational Biology*, 12(2):1–15, 2016.

[66] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014.

[67] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue):D52–D57, 2011.

[68] J. A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011.

[69] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

[70] B. Meibohm and H. Derendorf. Basic concepts of pharmacokinetic/pharmacodynamic (PK/PD) modelling. *International Journal of Clinical Pharmacology and Therapeutics*, 35(10):401–413, 1997.

[71] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, T. G. GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[72] D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.

[73] J. Mestas and C. C. W. Hughes. Of Mice and Not Men: Differences between Mouse and Human Immunology. *Journal of Immunology*, 2004.

[74] A. L. Mitchell, A. Sangrador-Vegas, A. Luciani, F. Madeira, G. Nuka, G. A. Salazar, H.-Y. Chang, L. J. Richardson, M. A. Qureshi, M. I. Fraser, M. Blum, N. D. Rawlings, R. Lopez, S. El-Gebali, S. Pesseat, S.-Y. Yong, S. C. Potter, T. Paysan-Lafosse, R. D. Finn, A. Marchler-Bauer, N. Thanki, H. Mi, P. D. Thomas, D. A. Natale, S. C. Tosatto, M. Necci, C. Orengo, I. Sillitoe, T. K. Attwood, P. C. Babbitt, S. D. Brown, P. Bork, A. Bridge, C. Rivoire, C. J. Sigrist, N. Redaschi, A. P. Pandurangan, J. Gough, D. R. Haft, G. G. Sutton, H. Huang, and I. Letunic. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2018.

[75] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

[76] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[77] R. J. Mural, M. D. Adams, E. W. Myers, H. O. Smith, G. L. G. Miklos, R. Wides, A. Halpern, P. W. Li, G. G. Sutton, J. Nadeau, S. L. Salzberg, R. A. Holt, C. D. Kodira, F. Lu, L. Chen, Z. Deng, C. C. Evangelista, W. Gan, T. J. Heiman, J. Li, Z. Li, G. V. Merkulov, N. V. Milshina, A. K. Naik, R. Qi, B. C. Shue, A. Wang, J. Wang, X. Wang, X. Yan, J. Ye, S. Yooseph, Q. Zhao, L. Zheng, S. C. Zhu, K. Biddick, R. Bolanos, A. L. Delcher, I. M. Dew, D. Fasulo, M. J. Flanigan, D. H. Huson, S. A. Kravitz, J. R. Miller, C. M. Mobarry, K. Reinert, K. A. Remington, Q. Zhang, X. H. Zheng, D. R. Nusskern, Z. Lai, Y. Lei, W. Zhong, A. Yao, P. Guan, R.-R. Ji, Z. Gu, Z.-Y. Wang, F. Zhong, C. Xiao, C.-C. Chiang, M. Yandell, J. R. Wortman, P. G. Amanatides, S. L. Hladun, E. C. Pratts, J. E. Johnson, K. L. Dodson, K. J. Woodford, C. A. Evans, B. Gropman, D. B. Rusch, E. Venter, M. Wang, T. J. Smith, J. T. Houck, D. E. Tompkins, C. Haynes, D. Jacob, S. H. Chin, D. R. Allen, C. E. Dahlke, R. Sanders, K. Li, X. Liu, A. A. Levitsky, W. H. Majoros, Q. Chen, A. C. Xia, J. R. Lopez, M. T. Donnelly, M. H. Newman, A. Glodek, C. L. Kraft, M. Nodell, F. Ali, H.-J. An, D. Baldwin-Pitts, K. Y. Beeson, S. Cai, M. Carnes, A. Carver, P. M. Caulk, A. Center, Y.-H. Chen, M.-L. Cheng, M. D. Coyne, M. Crowder, S. Danaher, L. B. Davenport, R. Desilets, S. M. Dietz, L. Doup, P. Dullaghan, S. Ferriera, C. R. Fosler, H. C. Gire, A. Gluecksmann, J. D. Gocayne, J. Gray, B. Hart, J. Haynes, J. Hoover, T. Howland, C. Ibegwam, M. Jalali, D. Johns, L. Kline, D. S. Ma, S. MacCawley, A. Magoon, F. Mann, D. May, T. C. McIntosh, S. Mehta, L. Moy, M. C. Moy, B. J. Murphy, S. D. Murphy, K. A. Nelson, Z. Nuri, K. A. Parker, A. C. Prudhomme, V. N. Puri, H. Qureshi, J. C. Raley, M. S. Reardon, M. A. Regier, Y.-H. C. Rogers, D. L. Romblad, J. Schutz, J. L. Scott, R. Scott, C. D. Sitter, M. Smallwood, A. C. Sprague, E. Stewart, R. V. Strong, E. Suh, K. Sylvester, R. Thomas, N. N. Tint,

C. Tsonis, G. Wang, G. Wang, M. S. Williams, S. M. Williams, S. M. Windsor, K. Wolfe, M. M. Wu, J. Zaveri, K. Chaturvedi, A. E. Gabrielian, Z. Ke, J. Sun, G. Subramanian, J. C. Venter, C. M. Pfannkoch, M. Barnstead, and L. D. Stephenson. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573):1661–1671, 2002.

[78] S. S. Negi, C. H. Schein, G. S. Ladics, H. Mirsky, P. Chang, J.-B. Rascle, J. Kough, L. Sterck, S. Papineni, J. M. Jez, L. Pereira Mouriès, and W. Braun. Functional classification of protein toxins as a basis for bioinformatic screening. *Scientific Reports*, 7(1):13940, 2017.

[79] N. L. Nehrt, W. T. Clark, P. Radivojac, and M. W. Hahn. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Computational Biology*, 7(6), 2011.

[80] R. B. Nelsen. Kendall tau metric. `http://www.encyclopediaofmath.org/index.php?title=Kendall_tau_metric&oldid=12869`. Accessed: 2019-01-31.

[81] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A. D. Rouillard, S. Schürer, T. Sheils, A. Simeonov, L. A. Sklar, N. Southall, O. Ursu, D. Vidovic, A. Waller, J. Yang, A. Jadhav, T. I. Oprea, and R. Guha. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research*, 45(D1):D995–D1002, 2017.

[82] F. Noor. The Changing Paradigm in Preclinical Toxicology: in vitro and in silico Methods in Liver Toxicity Evaluations. In *Animal Experimentation: Working Towards a Paradigm Change*, chapter 25, pages 610–638. Brill, 2019.

[83] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 11 2015.

[84] Pacific Biosciences. ASHG PacBio Workshop: The Iso-Seq method for discovering alternative splicing in human diseases. `https://www.pacb.com/videos/ashg-pacbio-workshop-the-iso-seq-method-for-discovering-alternative-splicing-in-human-diseases/`. Accessed: 2019-04-29.

[85] Pacific Biosciences. Overview of SMRT Technology. `https://www.pacb.com/smrt-science/smrt-sequencing/`. Accessed: 2019-05-25.

[86] Pacific Biosciences. Pacific Biosciences Glossary of Terms. `https://www.pacb.com/wp-content/uploads/2015/09/Pacific-Biosciences-Glossary-of-Terms.pdf`. Accessed: 2019-05-28.

[87] W. R. Pearson. An introduction to sequence similarity ("Homology") searching. *Current Protocols in Bioinformatics*, 42:3.1.1–3.1.8, 2013.

[88] G. Pertea. gffcompare. `https://github.com/gpertea/gffcompare/releases/tag/v0.11.2`. Accessed: 2019-04-28.

[89] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[90] M. Pertea, A. Shumate, G. Pertea, A. Varabyou, Y.-c. Chang, A. K. Madugundu, A. Pandey, and S. L. Salzberg. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv*, 2018.

[91] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753, 2001.

[92] R. A. Prentis, Y. Lis, and S. R. Walker. Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964-1985). *British Journal of Clinical Pharmacology*, 25(3):387–396, 1988.

[93] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[94] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.

[95] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[96] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29:24, 2011.

[97] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[98] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[99] E. L. Roggen. In vitro Toxicity Testing in the Twenty-First Century. *Frontiers in Pharmacology*, 2:3, 2011.

[100] W. M. S. Russel and R. L. Burch. *The principles of humane experimental technique.* London, Methuen.

[101] S. P. Sadedin, B. Pope, and A. Oshlack. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, 28(11):1525–1526, 2012.

[102] J. A. Salon, D. T. Lodowski, and K. Palczewski. The Significance of G Protein-Coupled Receptor Crystallography for Drug Discovery. *Pharmacological Reviews*, 63(4):901–937, 2011.

[103] S. L. Salzberg. Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, 20(92), dec 2019.

[104] S. Sayols, D. Scherzinger, and H. Klein. dupradar: a bioconductor package for the assessment of pcr artifacts in rna-seq data. *BMC Bioinformatics*, 17(1):428, 2016.

[105] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[106] K. Slowikowski. picardmetrics. `https://github.com/slowkow/picardmetrics/archive/0.2.4.tar.gz`. Accessed: 2019-02-10.

[107] R. Smith-Unna, C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26:1134–1144, 2016.

[108] G. K. Smyth, W. Shi, and Y. Liao. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[109] J. F. Söllner, G. Leparc, T. Hildebrandt, H. Klein, L. Thomas, E. Stupka, and E. Simon. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Scientific Data*, 4, 2017.

[110] J. F. Söllner, G. Leparc, M. Zwick, T. Schönberger, T. Hildebrandt, K. Nieselt, and E. Simon. Exploiting orthology and de novo transcriptome assembly to refine target sequence information. *BMC Medical Genomics*, 12(1):69, 2019.

[111] E. L. Sonnhammer and G. Östlund. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(D1):D234–D239, 2015.

[112] R. Stebbings, L. Findlay, C. Edwards, D. Eastwood, C. Bird, D. North, Y. Mistry, P. Dilger, E. Liefooghe, I. Cludts, B. Fox, G. Tarrant, J. Robinson, T. Meager, C. Dolman, S. J. Thorpe, A. Bristow, M. Wadhwa, R. Thorpe, and S. Poole. "Cytokine Storm" in the Phase I Trial of Monoclonal Antibody TGN1412: Better Understanding the Causes to Improve PreClinical Testing of Immunotherapeutics. *Journal of Immunology*, 179(5):3325–3331, 2007.

[113] A. D. Strand, A. K. Aragaki, Z. C. Baquet, A. Hodges, P. Cunningham, P. Holmans, K. R. Jones, L. Jones, C. Kooperberg, and J. M. Olson. Conservation of Regional Gene Expression in Mouse and Human Brain. *PLoS Genetics*, 3(4):e59, 2007.

[114] M. Tardaguila, L. de la Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. del Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M. Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno-Manzano, and A. Conesa. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, 28:396–411, 2018.

[115] F. Tekaia. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights*, 9:17–28, 2016.

[116] T. A. Thanaraj, F. Clark, and J. Muilu. Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, 31(10):2544–2552, 2003.

[117] The GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[118] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45:158–169, 2017.

[119] F. Thibaud-Nissen, A. Souvorov, T. Murphy, M. DiCuccio, and P. Kitts. *Eukaryotic Genome Annotation Pipeline*. 2013. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK169439/`.

[120] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2012.

[121] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, 28(5):511–515, 2010.

[122] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[123] E. Tseng. `https://github.com/PacificBiosciences/stsPlots/issues/2#issuecomment-254292866`. Accessed: 2019-04-29.

[124] E. Tseng. cDNA_Cupcake. `https://github.com/Magdoll/cDNA_Cupcake`. Accessed: 2019-08-08.

[125] E. Tseng, Y. Li, and A. Töpfer. Iso-Seq Deep Dive. `https://www.dropbox.com/s/0hlqi7b79kzi8bh/20180517_AsiaUGM_BioinformaticsDeck_English_V5.pdf?dl=0`. Accessed: 2019-05-28.

[126] E. Tseng, Y. Li, and A. Töpfer. Isoseq deep dive. `https://www.dropbox.com/s/0hlqi7b79kzi8bh/20180517_AsiaUGM_BioinformaticsDeck_English_V5.pdf?dl=0`, 2018. Accessed: 29 April 2019.

[127] E. Tseng, H.-T. Tang, R. R. AlOlaby, L. Hickey, and F. Tassone. Altered expression of the FMR1 splicing variants landscape in premutation carriers. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1860(11):1117–1126, 2017.

[128] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-k. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-h. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. V. Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. V. Heijne, J. Nielsen, and F. Pontén. Tissue-based map of the human proteome. *Science*, 347(6220), 2015.

[129] M. Uhlén, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten. Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology*, 28(12), 2010.

[130] M. Vaudel, K. Verheggen, A. Csordas, H. Raeder, F. S. Berven, L. Martens, J. A. Vizcaíno, and H. Barsnes. Exploring the potential of public proteomics data. *Proteomics*, 16(2):214–225, 2016.

[131] L. Venturini, S. Caim, G. G. Kaithakottil, D. L. Mapleson, and D. Swarbreck. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *bioRxiv*, 2017.

[132] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009.

[133] B. Wang, M. Regulski, E. Tseng, A. Olson, S. Goodwin, W. R. McCombie, and D. Ware. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Research*, 28(6):921–932, 2018.

[134] B. Wang, E. Tseng, M. Regulski, T. A. Clark, T. Hon, Y. Jiao, Z. Lu, A. Olson, J. C. Stein, and D. Ware. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, 7(1):11708, 2016.

[135] J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, and K. F. Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6:100, 2017.

[136] K. Wethmar, A. Barbosa-Silva, M. A. Andrade-Navarro, and A. Leutz. uORFdb — a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Research*, 42:60–67, 2014.

[137] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[138] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 2005.

[139] J. M. Young, C. Friedman, E. M. Williams, J. A. Ross, L. Tonnes-Priddy, and B. J. Trask. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Human Molecular Genetics*, 11(5):535–546, 2002.

[140] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012.

[141] Y. Yu, J. C. Fuscoe, C. Zhao, C. Guo, M. Jia, T. Qing, D. I. Bannon, L. Lancashire, W. Bao, T. Du, H. Luo, Z. Su, W. D. Jones, C. L. Moland, W. S. Branham, F. Qian, B. Ning, Y. Li, H. Hong, L. Guo, N. Mei, T. Shi, K. Y. Wang, R. D. Wolfinger, Y. Nikolsky, S. J. Walker, P. Duerksen-Hughes, C. E. Mason, W. Tong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, and C. Wang. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nature Communications*, 5(1):3230, 2014.

[142] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. Mclaren, B. Moore, J. Mudge, N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. Ensembl 2018. *Nucleic Acids Research*, 46:754–761, 2018.