Investigating normal human gene expression in tissues with high-throughput transcriptomic and proteomic data.

Mitra Parissa Barzine

July 2020

Hughes Hall,
University of Cambridge
EMBL's European Bioinformatics Institute (EMBL-EBI)

This dissertation is submitted for the degree of
Doctor of Philosophy

DECLARATION

I hereby declare that this thesis

- is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specified in the text;
- is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution; I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.
- contains fewer than the prescribed limit of 60,000 words exclusive of tables, footnotes, bibliography, and appendices and has fewer than 150 figures.

Mitra Parissa Barzine
July 2020

SUMMARY

Investigating normal human gene expression in tissues with high-throughput transcriptomic and proteomic data. — By Mitra Parissa Barzine.

With the improvement of high-throughput technologies during the last decade, several studies exploring the normal gene expression in human tissues have been published. Many studies examine the transcriptome with RNA sequencing (RNA-Seq), and others probe the proteome with unlabelled bottom-up Mass Spectrometry. As the sampling of undiseased tissues is difficult, the community often refers to expression atlases, which are collating these studies, to support or validate new findings. Despite many overlapping tissues between the studies, few atlases attempt to integrate all the data.

In this thesis, I investigate the consistency of gene expression across tissues and studies in human with the help of transcriptomics captured with high-throughput sequencing (RNA-Seq) and proteomics generated with label-free bottom-up Mass Spectrometry (MS).

After describing the transcriptomic and proteomic data and their state-of-art processing (Chapter 2), I review several identified sources of biases and my approaches to limit their effects (Chapter 3).

The integration of the various transcriptomic datasets (Chapter 4) shows that the biological signal dominates the technical noise for RNA-Seq data. Tissue samples display higher levels of correlation for identical tissues in other studies than for other tissues in the same datasets. In other words, interstudy correlations for identical tissues are higher than correlations between different tissues within the same study. Globally, genes show similar expression profiles across studies for a given set of tissues. All genes categories are involved, including the tissue-specific genes and the ubiquitously expressed ones.

After briefly discussing comparisons of proteomic data, I introduce a new proteomic quantification method, PPKM (Chapter 5). The PPKM method allows me to quantify about twice as many proteins compared to usual methods.

Limited numbers of previous studies have shown various correlation levels between the expression of protein and mRNA in studies combining high-throughput transcriptomics and proteomics. I show that, for most tissues, we can observe quite good correlation levels (*i.e.* significantly better than expected by chance), even when the samples have different biological and technical backgrounds as they have been independently sourced. Many genes share similar patterns of expression between the two biological layers, *e.g.* genes that have a protein detected in a single tissue are more likely to have their mRNA showing specificity for the same tissue. Additionally, three groups of genes present functional enrichments of biological processes. Genes having highly correlated protein and mRNA expressions across tissues are enriched in catabolic processes. Genes having the most anticorrelated expressions are enriched for ribosomes and ncRNAs regulation. Genes with a protein detected in a single tissue are enriched in signalling processes.

Overall, this thesis describes a global picture of the current consolidated knowledge we can extract from the joint study of public transcriptomic and proteomic data. Beyond confirming or improving observations reported in the literature, this work provides new insights into the ubiquitous and tissue-specific genes. To the best of my knowledge, this work has also established the most extensive list of genes with robust transcriptomic and proteomic expression across tissues and studies. Furthermore, it shows that joint study approaches can help the development of new methods, like the new proteomic PPKM quantification method. Finally, the highlighting of distinct functional enrichment profiles for groups of genes across tissues and studies lays a framework for further research.

PREFACE

The old saying *where there's a will, there's a way* fails to mention that proper support and means are must-haves in the journey that one takes for a PhD. Many institutions and people have a direct hand in the successful completion of this thesis.

I thank the EMBL predoc program, the University of Cambridge and Hughes Hall college to have provided me with the working and living environment for one of the most formative periods of my life.

I am extremely grateful to my supervisor Dr Alvis Brazma without whom this adventure wouldn't have happened. Thank you, Alvis, for accepting me into the Functional Genomic group, which comprises many exceptional people that contributed to my day-to-day life and enriched me as a person. Thank you for your trust and support through this roller coaster that was my doctorate.

I want to express my sincere thanks to Prof. Kathryn Lilley and Prof. Jürg Bähler for making the viva an outstanding experience. I will forever remember with the utmost joy our in-depth and intellectually stimulating discussion. I deeply appreciate your brilliant comments and expert suggestions, which have further broadened my understanding of the biological world.

I am very thankful to Dr Jyoti Choudhary and Dr James Wright for their collaboration. I wouldn't have ever delved into the proteomics world if it wasn't for their guidance and discussions. I also would like to thank my thesis advisory committee, Dr Sarah Teichman, Dr Gos Micklem and Dr Wolfgang Huber, for their expert feedback and discussion. Many thanks to Lynn French, Lorraine McAlister, Anna Alasalmi and Clare Impey for helping with many administrative tasks and paperwork involved in my doctorate.

I can't thank my teammates enough for all the welcoming, exciting and fun environment. Thank you for all your help and the lunch conversations about anything from top-notch science to absurd elephant jokes. Thanks to Mar, Nuno, Johan, Wanseon, Sérgio, Claudia, Aida (Fatemeh) for their discussions and close friendship. Special thanks to Nuno and Claudia, who provided me with the support I needed at crucial times. Nuno, thank you for your invaluable work and help with processing the data. You are like a bigger brother with whom I could discuss anything, from the inner hell circle of R to SciFi, in the most in-depth tiny details. Obrigada Nuno for all the years we share at the EBI.

I am most grateful to the EBI community and particularly to the EBI predocs, who are amazing people. Thanks to our morning coffee discussions and debates, I learnt so much about bioinformatic fields that were not directly involved in my work. I am truly fortunate to have met all of you. Furthermore, I want to thank more specifically the fantastic people who have taken time to read and give me feedback on this thesis. Mar, Nuno, Juan, Myrto, Aida, Sarah, Julian, Konrad, Nils, Hannah, Atoussa, Isabelle and Steve, I thank all of you for your invaluable inputs and corrections. Steve, thank you also so much for all the discussions about the fine points in British culture and language.

Thank you Isabelle for our hour-long conversations, which, for the most part, consisted of me figuring out the best ways to explain my doctorate works to a layperson. Thank you, Philippe, for your unfailing support and invaluable assistance, particularly regarding the logistics. Finally, I want to thank my sister, Atoussa, and my mother, Faranak, to always be by my side, despite any possible physical distance. Thank you both for helping me to grow and develop, if not formalise, my ideas. Thank you for sharing with me your own scientific and life interests, which, beyond being exciting, sustain my joy and creativity.

CONTENTS

APPENDIX 183

TERMS AND ABBREVIATIONS

Notation	Description
2D-DIGE	2D-differential in-gel electrophoresis 26
aa	amino acid xxiii, 5, 25, 31, 39–41, 184, 192
ADP	adenosine diphosphate 38
Ampholyte	molecule which can both gives or accepts a proton (H^+). 32
ArrayExpress	EBI archive of Functional Genomics Data. It stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community. 7, 10, 56, 58–60, 63, 128, 169
AUC	area under the curve 27
BAM	BGZF-compressed binary file that can be converted into SAM format SAM 17, 22
Bioconductor	Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. 24, 162
Bottom-up approach	In proteomics, bottom-up approaches (in contrast to *top-down* approaches) involve a step of proteolytic digestion prior to the Mass Spectrometry analysis. 28
bp	base pair; unit of length for double-stranded nucleic acids 58
CAGE	cap analysis gene expression 111
cDNA	complementary DNA 8, 11–13, 24, 58, 185
CDS	coding DNA sequence 68, 71
CID	collision-induced dissociation 30, 34, 38, 61, 62, 70, 188
Computer cluster	set of connected computers that work together to improve performance over a single computer 65
cv	coefficient of variation 106, 108
Da	Dalton is a unified atomic mass unit. It may be also annotated as **u** 25, 31, 40, 70, 71
dbGaP	The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in humans. 60, 63
DC	direct current 188
DDA	data-dependant acquisition 25–27, 30, 35
DEA	differential expression analysis 24, 51, 52
DGEA	differential gene expression analysis 22, 113
DIA	data-independant acquisition 25, 35, 37, 49
DNA	deoxyribonucleic acid xiii, xiv, xvi, 3, 5–8, 13, 14, 27, 28, 54, 58, 63, 178, 190

Notation	**Description**
dNTP	deoxynucleoside triphosphate. It can be any nucleotides (usually either adenosine (A), cytosine (C), guanine (G) or thymine (T)) 14
EBI	European bioinformatics institute 58, 62, 63, 65, 67, 72, 91, 101, 112, 114, 127, 128, 167, 175
ECD	electron capture dissociation 39
EM	expectation-maximisation 21, 47
ENA	The European Nucleotide Archive provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. 7, 56, 58, 63
ENCODE	The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the national human genome research institute 70
Ensembl	database that is the joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. 10, 65, 70–72, 100, 124, 125
eQTL	expression quantitative trait locus 59
ESI	electron spray ionisation 33, 189
EST	expressed sequence tag 7, 100, 185
ETD	electron transfer dissociation 35, 39
EThcD	electron-transfer and higher-energy collision dissociation 38
FANTOM5	FANTOM5 is a consortium that systematically investigated the genes expressed in all cell types the human body and the genomic regions that contains the transcription starting site. 111
FASP	filter-aided sample preparation 31
FASTQ	text-based file format. For each cluster read, it records a unique identifier, a nucleotide sequence and the call accuracy for each base (Phred score). Optionally, there can be more information, *e.g.* the spatial position of the cluster on the flow cell. 15, 63
FDR	false discovery rate 43, 44, 47, 48, 50, 51, 63, 69, 71, 85, 192, 193
FF	fresh-frozen 27
FFPE	formalin-fixed paraffin-embedded 8, 27
Flow cell	support of Illumina sequencing. It enables the parallelisation of the sequencing of millions of DNA fragments together which are kept spatially separated in clusters. It is a glass slide with lanes and each lane is coated with short nucleotide sequences that are used to hybridise by complementarity adapters on the DNA that will be sequenced. 11, 15, 23

Notation	**Description**
FPKM	fragments per kilobase of a feature (*i.e.* transcript in most cases) per million mapped fragments xix, xxi, xxiii, 22–24, 64, 68, 74, 77, 85, 92, 93, 95–97, 99, 101, 107, 108, 110, 112–114, 143, 144, 146, 147, 149, 154–156, 166, 178, 188, 205, 206, 208, 209, 213, 214, 223, 226
FT	Fourier transform 34, 189
FTMS	Fourier transform mass spectrometer 61, 189
Fusion gene	gene which is the fusion product of the parts of two (or more) different genes. 14
GENCODE	project that produces high quality reference gene annotation for human and mouse genomes. 68, 70, 71
GEO	gene expression omnibus 58
GFF	general feature format. Tab-delimited file format that records gene information or other features as DNA, RNA or protein sequences. xv, 22
GO	gene ontology xiv, 52, 162–164, 167, 169–171, 174
GOA	GO enrichment analysis 52, 162, 163
GRCh	The Genome reference consortium human genome build. It is always followed by a version number. 10, 65, 66, 68, 70, 72, 100, 110–112, 114, 177
GSEA	gene set enrichment analysis 51, 52
GTEx	The Genotype-Tissue Expression project establishes a resource database and associated tissue bank for the scientific community to study the relationship between genetic variation and gene expression in human tissues. 57, 59, 62, 63, 65, 74, 78, 86, 87, 91, 92, 94–96, 98, 105, 108, 110, 112, 113, 131, 132, 134, 136, 137, 144, 154–157, 159, 160, 165, 168, 176, 177, 199, 211, 214, 218, 222, 256
GTF	gene transfer format. Tab-delimeted file format based on the format and hold information about gene structure. A main feature of this file is that it can be validatable which increases the data reliability. 22
HBB	hemoglobin subunit beta 143
HCD	higher-energy collisional dissociation 34, 38, 61, 70
HLA	human leukocyte antigen 70
HPA	The human protein atlas is a Swedish-based programme aiming to map all the human proteins in cells, tissues and organs using integration of various omics technologies 59
HPLC	high-performance liquid chromatography 32, 33
IBAQ	intensity based absolute quantification 50, 123
IBM	Illumina body map 2.0 xix, 63, 68, 70, 74, 86, 87, 94, 96, 109, 112, 113, 211, 218
ICAT	isotope-coded affinity tag 26
ICR	ion cyclotron 34
ID	identification number 56, 58–62, 71, 128, 169
iTRAQ	isobaric tag for relative and absolute quantification 26, 34

Notation	**Description**
laser	light amplification by stimulated emission of radiation xv, 33
LC	liquid chromatography xv, xvi, 30–32, 35
LC-MS	liquid chromatography (LC) followed by MS 32
LC-MS/MS	liquid chromatography (LC) followed by tandem MS 27, 28, 30, 31, 37, 41, 61, 71
LDS	lithium dodecyl sulfate xv
LDS-PAGE	lithium dodecyl sulfate-polyacrylamide gel electrophoresis 32, 61
LECA	last eukaryotic common ancestor 81
LIT	linear ion trap 34, 188
lncRNA	long non-coding RNA 10, 70
LTQ	linear trap quadrupole 34, 62, 188, 189
MAD	median absolute deviation 104, 224
MALDI	matrix-assisted laser desorption ionisation 33, 34
miRNA	microRNA 12
miRNA-Seq	miRNA sequencing, which can also be called miRNA shotgun sequencing 11
MRM	multiple reaction monitoring 25
mRNA	messenger RNA xv–xvii, xxii, 3, 5, 6, 10, 54, 58, 77, 79, 81, 83, 85, 86, 94, 112, 123, 126, 127, 131, 133–137, 139–141, 143, 144, 146–158, 160–171, 173–176, 178, 179, 185, 194, 223, 243, 251, 256
MS	mass spectrometry xv, 2, 25, 27, 28, 30–37, 43, 54, 55, 60, 68, 69, 72, 81, 85, 117, 118, 121, 128, 129, 131, 134, 136, 137, 161, 165, 170, 173, 174, 176, 194
MS/MS	tandem MS 25–27, 33–43, 50, 61, 189, 193, 194
mt-mRNA	mitochondrial mRNA 83
ncRNA	non-coding RNA 10, 56, 164, 175
NHGRI	national human genome research institute xiv
NIH	national institutes of health (USA) 59
NP	nondeterministic polynomial time 45, 193
nt	nucleotid; common unit of length for single-stranded nucleic acids 14, 56, 83
OR	olfactory receptor 123, 127, 128
ORA	over-representation analysis 51
PAGE	polyacrylamide gel electrophoresis xv, xvii, 32
PCAWG	pan-cancer analysis of whole genomes 112, 113
PCR	polymerase chain reaction 7, 12, 23, 56
PEP	posterior error probability 43, 44, 47, 69, 71, 192, 193
Perl	The Perl language is an open-source interpreted programming language developed by Larry Wall and first released in 1987 to easily handle textual information. 65
pH	potentiel Hydrogen 32

Notation	**Description**
phenotype	set of observable characteristics or traits of an individual. The traits can be inherited (genotype), due to the environment, or from the interaction of the environment with the genotype. xiii, 5, 7, 54
Phred	A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing. xiv, 15, 63, 186
PPKM	PSMs per kilobase of gene per million v, 125, 126, 128, 137, 141–144, 147, 151, 152, 154–157, 160, 162, 165, 166, 169, 171, 174
ppm	part per million 70
PRIDE	The proteomics identifications (PRIDE) database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence. PRIDE is a core member in the ProteomeXchange (PX) consortium, which provides a single point for submitting mass spectrometry based proteomics data to public-domain repositories. Datasets are submitted to PRIDE via ProteomeXchange and are handled by expert biocurators. 61, 68, 71, 169
PRM	parallel reaction monitoring 26
ProteomicsDB	ProteomicsDB is a joint effort of the Technische Universität München (TUM) and SAP SE. It is dedicated to expedite the identification of the human proteome and its use across the scientific community. 61, 62, 68
PSM	peptide spectrum match xvi, 42–44, 46–49, 71, 123–126, 137, 192–194
PTM	post-translational modification 5, 39, 40
PTR	protein-to-mRNA ratio 169
Python	The Python language is a high-level and interpreted programming language developed by Guido van Rossum and first released in 1991. The main purpose of this language is to be multivalent while enhancing code readibility. 22
RefSeq	The reference sequence database is an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein products. 100
RF	radio frequency 188, 189
RNA	ribonucleic acid ix, xiv–xvii, 1, 3, 5–8, 10–12, 20, 22–24, 27, 28, 54–56, 59, 67, 77, 81, 83, 85, 101, 127, 135, 139, 146, 170, 173, 178, 185, 190
RNA-Seq	RNA sequencing, which can also be called whole transcriptome shotgun sequencing ix, xix, xxiii, 1, 2, 7–9, 15, 16, 18, 20, 23, 27, 30, 36, 50, 55–60, 62–67, 70, 72, 78, 79, 81, 83, 85, 89, 91, 93–95, 99, 109–114, 124, 125, 129, 131, 132, 134, 136, 137, 167, 173, 174, 178, 199

Notation	**Description**
RPKM	reads per kilobase of a feature (*i.e.* transcript in most cases) per million mapped reads 23, 85, 125
RPLC	reversed-phase LC 32, 61
rPTR	relative protein-to-mRNA ratio 169
rRNA	ribosomal RNA 3, 5, 10
RT-qPCR	Reverse-transcription quantitative real-time polymerase chain reaction is a molecular biology technique to quantify the amount of ribonucleic acid in a given cell or sample. Cycles of monitored replications are used to robustly measure the gene expression. It is often considered to be the most powerful and sensitive of the quantitative assay for ribonucleic acid. However, this method requires to know in advance which are the genes of interest. 22
SAM	sequence alignment/map. Text based file format. xiii, 17, 22
SDS	sodium dodecyl sulfate xvii, 32
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis 32, 61
SILAC	stable isotope labeling by amino acids in cell culture 26
SNP	single nucleotide polymorphism 20, 59
SNR	signal-to-noise ratio 37
SRM	selected reaction monitoring 25, 26
SVM	support vector machine 44, 48
TB	terabyte 36
TCGA	the cancer genome atlas 112
TDA	target decoy search approach 43, 51, 192
TiGER	tissue-specific gene expression and regulation database created by the Bioinformatics lab at Wilmer Institute, Johns Hopkins University 100, 101, 174
TMM	trimmed mean of M values (M: log expression ratios) 111
TMT™	tandem mass tags 26
TOF	time-of-flight 34
TREP	tissue reference expression profile 87, 92, 94–99, 105, 114, 121, 136, 141, 142, 146, 147, 213, 219, 222, 236, 237
tRNA	transfer RNA 3, 5
tryptic	adjective qualifying peptides that have been generated by trypsin digestion. 31
TS	tissue-specific 99–101, 103–105, 110, 111, 113, 114, 121, 128, 146, 149–152, 157, 159, 160, 162, 164, 166, 167, 169–171, 174, 175, 179, 256
TSS	transcription starting site xiv
UniProt	The Universal Protein Resource provides the scientific community with a comprehensive high-quality and freely accessible resource for protein sequence and functional annotation data. 68, 71
UPLC	ultra performance liquid chromatography 32, 33
XIC	extracted-ion current 27, 50

LIST OF FIGURES

LIST OF TABLES

I am not accustomed to saying anything with certainty after only one or two observations.

Andreas Vesalius [O'Malley, 1964]

INTRODUCTION

Today, we have comprehensive knowledge about the structure and functioning of the human body at the macroscopic level[1]. At the microscopic level, however, the identification and mapping of macromolecules (*e.g.* RNAs, proteins), their function and whereabouts still need to be refined.

Beyond the invaluable addition to our knowledge, other more practical reasons are also sustaining the effort for human expression atlases. Genes with specific behaviour in particular conditions are a convenient starting point for designing new diagnostic tests and discovering new effective drug targets. Besides, a robust atlas for non-diseased tissue baseline expression will allow a better understanding of unperturbed physiology. It can also serve as a reference in studies where controls are unavailable or hard to sample, which is generally the case in cancer research.

The completed and annotated human genome and technological advances in high-throughput expression studies have opened the way towards this future new milestone. Evidence of the community shared interest is the recent explosion of high-throughput transcriptomic atlases in the literature. Examples include expression atlases of mouse [C. Wu, Orozco, et al., 2009; Ringwald et al., 2012], pig [Freeman et al., 2012], sheep [Clark et al., 2017], plants as maize [Stelpflug et al., 2016], vigna [S. Yao et al., 2016], pigeon pea [Pazhamala et al., 2017], parasites, *e.g. Schistosoma mansoni*. Many focus on mapping the human gene expression either as a whole, see *e.g.* [Krupp et al., 2012; Jiménez-Lozano et al., 2012; Uhlén, Fagerberg, et al., 2015; GTEx Consortium, 2013] or for specific aspects, *e.g.* the organogenesis in human embryos [Gerrard et al., 2016].

There are large incentives to develop these atlases with transcriptome shotgun sequencing (*i.e.* RNA-Seq). The technology involved in similar older projects[2] has demonstrated to generate highly variable data [Rung et al., 2013] that is challenging to integrate (even when produced by the same laboratory on the same platform) [Walsh et al., 2015]. When I started my doctorate, little was known about the interstudy robustness of RNA-Seq. However, it had already shown less background noise and a more extensive dynamic range of detection than the array-based technology (microarray) used in previous expression studies [Z. Wang et al., 2009]. RNA-Seq also has the added advantage to discover new transcripts as it does not rely on previous knowledge [Z. Wang et al., 2009].

[1] Even though human anatomy can still be refined, and new findings can happen [Kumar et al., 2019].

[2] E.g. Gene Expression Atlas for Human Embryogenesis [Yi, Xue, et al., 2010], Atlas of human primary cells [Mabbott et al., 2013], Gene atlas of mouse and human protein-encoding transcriptomes (now hosted by *BioGPS*) [Su et al., 2004], Allen Brain Atlas [Hawrylycz et al., 2012].

Considering the growing number of studies referring to these atlases[3] or the numerous efforts to compile them into new resources for the community — *e.g.* TISSUES[4] [Santos et al., 2015], Harmonizome[5] [Rouillard et al., 2016] and, after reprocessing the raw data, Expression Atlas[6] [Petryszak, Keays, et al., 2015] — assessing the consistency of the results from one study to another has become paramount. The recently reported reproducibility crisis in science [Begley et al., 2012; Fatovich et al., 2017; Lindner et al., 2018; Lyu et al., 2018] has only further underlined this need.

Given the previous context, the first aim of my doctorate was to examine the consistency of the (non-diseased) human tissues landscape of expression in independent large-scale transcriptomics. Then, with the publication of the first drafts of the human proteome [M.-S. Kim et al., 2014; Wilhelm et al., 2014], my aims have expanded to the integration of human tissue expression data across different datasets and biological layers.

OUTLOOK OF THIS THESIS

First, I review the biological, chemical, experimental and computational background of my doctorate works in Chapter 1.

Then, in Chapter 2, I present the five transcriptomic (RNA-Seq) and the three proteomic (MS) datasets that I have preselected for my analysis. Then I describe the bioinformatic pipelines that have automated the processing of these large-scale datasets.

After considering several possible sources of bias and strategies to limit them in Chapter 3, I compare and integrate the independent transcriptomic datasets in Chapter 4. Following an assessment of the findings' consistency in proteomics in Chapter 5, I then employ different approaches for integrating transcriptomics and proteomics in Chapter 6.

Finally, I close this thesis with a few remarks in Chapter 7.

[3] More than 2,800 papers for the five human primary studies (presented in Section 2.2) on 15 September 2019.
[4] TISSUES — https://tissues.jensenlab.org/Search
[5] Harmonizome — https://amp.pharm.mssm.edu/Harmonizome
[6] Expression Atlas — https://www.ebi.ac.uk/gxa/home

It would probably be oversimplifying the matter, but I am strongly tempted to say, 'All life is nucleic acid; the rest is commentary'.

Asimov (1989)

1 BIOLOGICAL AND TECHNOLOGICAL CONTEXT OF THIS THESIS

The following pages (pp. 3–54) present a summary of facts and techniques that form the biological and technological context of the work presented in this thesis. The different sections may be read on their own or skipped by informed readers without understandability issues.

1.1 UNIVERSALITY AND DIVERSITY OF LIFE

Every known form of life depends on a common set of molecule types, within which the DNAs, RNAs and proteins are arguably the most specific ones and have the widest variety. The other molecules are either inorganic (water and salts), small, or simple organic ones (sugars, organic or amino acids, nitrogenous bases, lipids or their precursors). [Callen, 2005]

The entirety of the DNA (protein-coding and non-coding) of a living organism constitutes its genetic material (or genome). The coding sections of the DNA, *i.e.* the protein-coding genes, contain the instructions for making (via mRNAs) the proteins, which are the main effectors supporting life functions. When a gene is switched on, it triggers a process, illustrated in Figure 1.1, in which the first step is called transcription and the second one translation. [Callen, 2005; Pierce, 2005]

The transcription initiation happens when an RNA polymerase attaches to the start of the gene and uses the DNA strand as a template to create a corresponding RNA from free RNA nucleotides [Alberts et al., 2002]. The transcription is a directional process and always happens from the gene 5' end towards its 3' end [Callen, 2005]. (For 5' and 3', see Figure 1.2 and related section.) Messenger RNAs (mRNAs) are RNAs that are produced from a gene and used as a template for translation to synthesise proteins. There are other kinds of RNAs, *e.g.* ribosomal RNAs (rRNAs) (which are the most numerous RNAs in the cell), transfer RNAs (tRNAs) and many others [Callen, 2005]. The entire repertoire of transcripts (*i.e.* RNA molecules) expressed in a cell or group of cells (such as in a tissue) is called transcriptome [Velculescu et al., 1997; Piétu et al., 1999; Z. Wang et al., 2009]. Unlike the genome which is roughly identical regardless of which cell of a particular individual is considered, the transcriptome may vary, sometimes dramatically, according to the biological context (different organs or tissues, or different conditions, *e.g.* healthy or in a reaction to a disease) and through time and life stages. [Alberts et al., 2002]

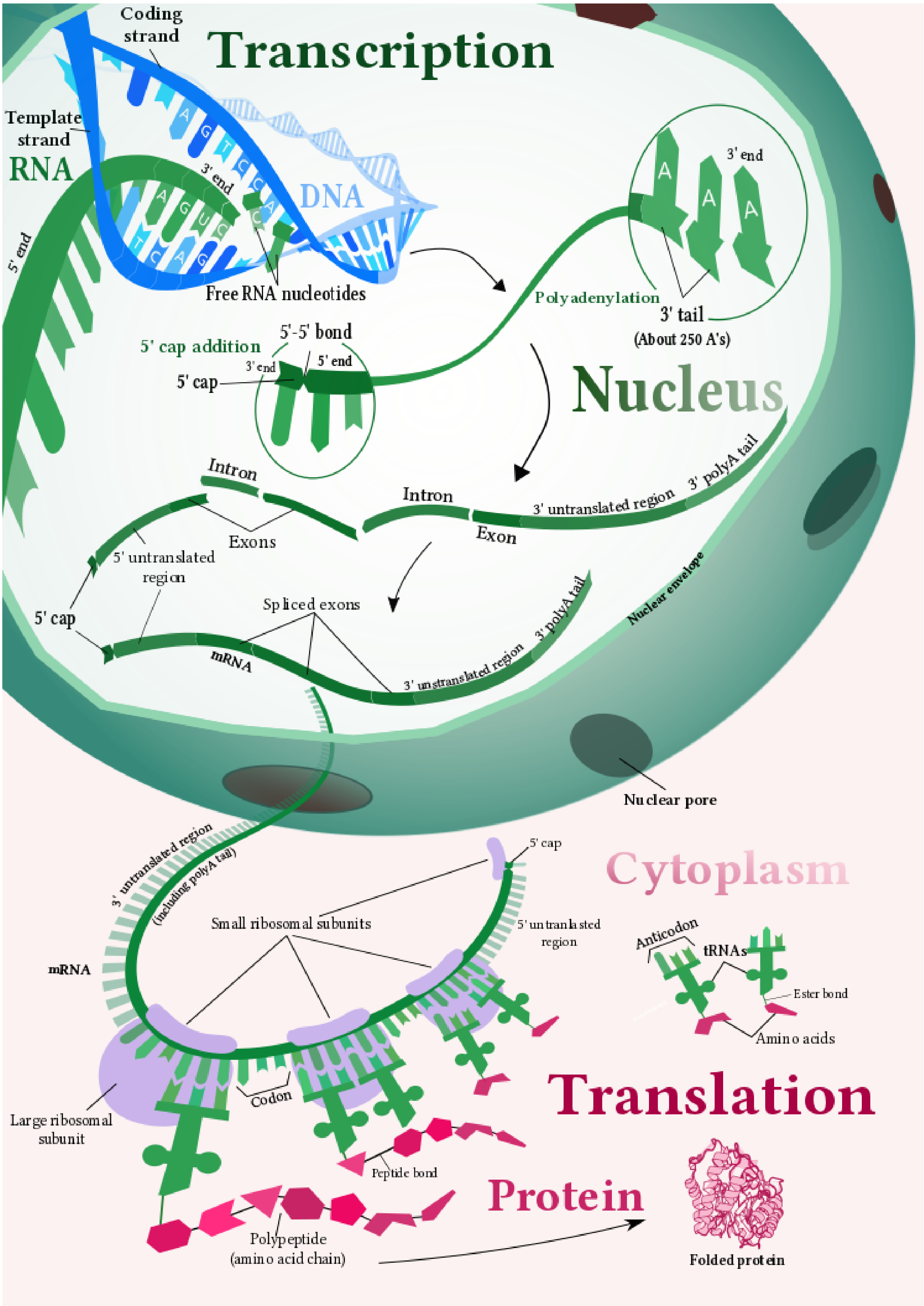


Figure 1.1. **Transcription and Translation: an overview.** This work, 'Transcription and Translation: an overview', is a derivative work of 'Simplified diagram of mRNA synthesis and processing. Enzymes not shown.' and 'Protein synthesis' both by Kelvinsong, used under [CC BY] (see Appendix A.2). 'Transcription and Translation: an overview' is licensed under [CC BY] by Mitra P. Barzine.

Before the mRNA is in turn used as a template to make the protein, a few modifications may occur. Among the typical post-transcriptional regulations, there are the capping and the addition of a polyadenylated (polyA) tail (both increasing the half-life of the mRNA), splicing (where internal parts of the mRNA — either non coding (*i.e. introns*) or coding (*i.e. exons*) — are removed), or RNA editing [Darnell, 2013]. For eukaryotic species (*e.g. Human*), the mRNAs have to be exported from the nucleus to the cytoplasm for the next step to happen [Callen, 2005].

The mRNAs initiate the translation, *i.e.* creation of new proteins, by binding to protein factories called ribosomes (complexes formed from rRNAs and ribosomal proteins). The ribosomes read the mRNA codons (*i.e.* groups of three consecutive bases) and produce the corresponding protein by synthesising a polypeptide chain from the free amino acids (aas) carried by tRNAs with the corresponding anticodons (*i.e.* complementary sequence to the mRNA codons). The aas are linked by peptide bonds formed through the reaction of functional groups on their primary chain: the carboxyl group (–COOH) of the first amino acid (aa) with the amine group (NH_2–) of the next one. This succession of primary chains and peptide bonds constitutes the protein backbone. This sequence is by convention described from the free amino group of the first aa (*i.e.* N-terminal) to the free carboxyl group of the last aa (*i.e.* C-terminal).

While the polypeptide chain is completed, it also folds into a three-dimensional structure, which is essential for fulfilling its role [Morris et al., 2016]. Many proteins need to undergo post-translational modifications (PTMs) before being functional. PTMs allow the regulation of the proteins' activity (activation and deactivation). They can involve proteolytic cleavages or the creation of new covalent bonds, as many proteins comprise more than one polypeptide chain to be functional. Other frequent PTMs include the phosphorylation, acetylation, or glycosylation of the aas (also called residues) [Alberts et al., 2002; Morris et al., 2016]. Besides the variety of possible PTMs and their combination, proteins can comprise twenty different aas (Appendix A.1), which have a vast range of physicochemical properties. Hence, in turn, the proteins have a wide physicochemical range too. [Morris et al., 2016; Callen, 2005].

Although the diversity of the proteome (*i.e.* entire repertoire of expressed proteins) is the primary contributor to the final phenotype[1] and functions of cells and tissues, its exhaustive study remains particularly challenging. Most proteins are quite stable, but their physicochemical diversity prevents the use of uniform and straightforward protocols that would encompass all the proteins in a given cell or tissue type. [Bruce et al., 2013]

On the other hand, all DNA and RNAs have very similar chemical properties, as their structures are very close as shown in Figure 1.2. They are polymeric molecules that are made by a chain of nucleotides. Nucleotides have three distinct chemical subunits: a phosphate group, a pentose (either ribose for RNA or deoxyribose for DNA) and a nitrogenous base, which can be a purine (adenine (A) or guanine (G)) or a pyrimidine

[1] Phenotype: set of observable characteristics or traits of an individual. The traits can be inherited (genotype), due to the environment, or from the interaction of the environment with the genotype.

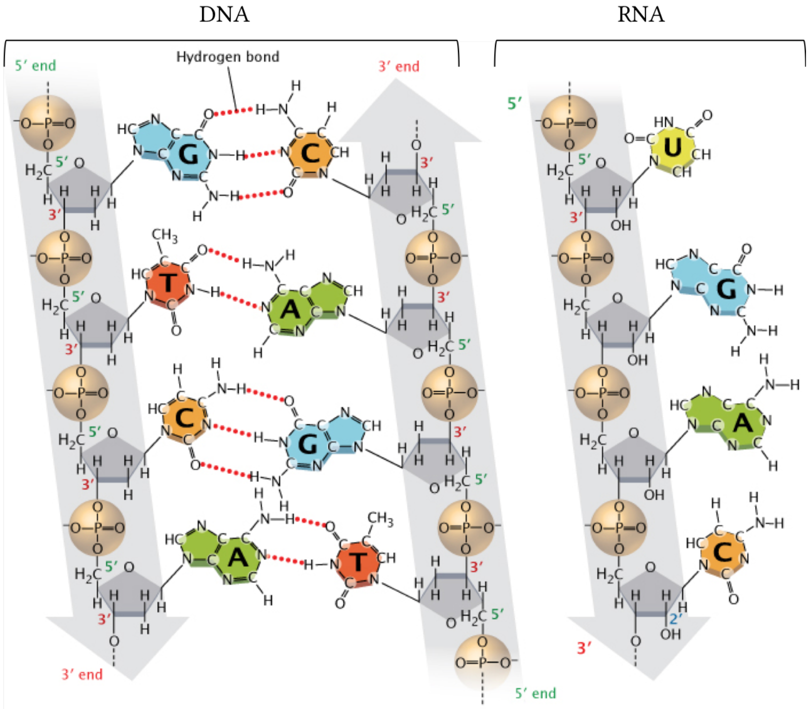


Figure 1.2. **DNA and RNA structures** © 2014 Nature Education — Adapted from Pierce (2005).

(cytosine (C) and either thymine (T) for DNA or uracil (U) for RNA). The alternation of phosphate groups and the pentoses create the biomolecules backbone, while the information is encoded into the nitrogenous bases sequence. DNA and RNAs all share the same directional reading frame, *i.e.* 5’ end to 3’ end, or in other words, from the phosphate group that is linked to the pentose carbon annotated 5’ to the phosphate group that is linked to the pentose carbon annotated 3’[2]. [Morris et al., 2016; Alberts et al., 2002; Callen, 2005]

The difference in physical properties between DNA and RNA molecules is primarily due to the predominant particular arrangement of DNA (double-stranded) and RNA (single-stranded). The double-stranded configuration of the DNA dramatically improves the stability of the DNA by involving many mechanisms, which includes many hydrogen bonds (three for each couple of G/C and two for each A/T). Besides, in eukaryotic cells, while the genome (DNA) is protected in organelles such as the nucleus, most of the mRNAs life is spent in the cytoplasm, which contains many enzymes (*e.g.* endonucleases and exonucleases) that cleave the nucleotide sequences and can ultimately degrade the mRNAs. Thus, even though mRNA half-lives are quite variable, mRNAs have generally the lowest stability when compared to the DNAs and proteins. [Alberts et al., 2002; Pierce, 2005]

[2] Pentoses are monosaccharides containing a chain of five carbons. When the pentose is part of a nucleotide, these carbons are annotated from 1’ to 5’ to avoid confusion with the carbons of nitrogenous cycles.

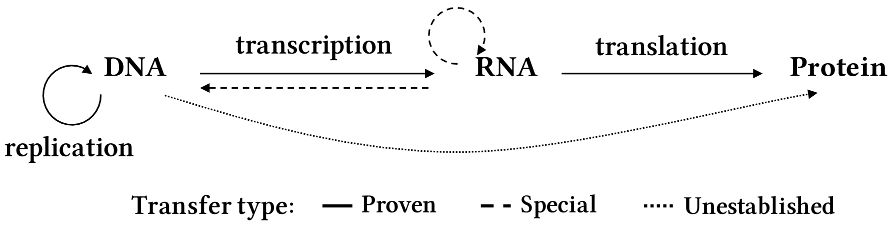


Figure 1.3. **The central dogma of molecular biology proposed by F. Crick.** The proven modes of information transfers are for the general ones in solid lines and in dashed lines for the special transfers (RNA to RNA or DNA) and the transfers that have yet to be established (DNA to protein).

This overall process, which allows the creation of the proteins and based on a flow of information initiated from the DNA, had been predicted by Francis Crick [Crick, 1958; Crick, 1970]. He stated what is now known as the core of the central dogma of molecular biology (shown in Figure 1.3): 'Once information has got into a protein it can't get out again'. In other words, the genome contains all the information needed to produce functional proteins, and, in theory, if we reach a total understanding of the information encoded into the DNA, we will be able to predict the phenotype due to the proteome. As DNA is static while the coding portion (about 2%) of the human genome [Venter et al., 2001] varies in expression (both in concentration and composition) depending on the tissue or cell type, genome studies are more established than transcriptomic or proteomic ones, but the latter ones are more phenotypically insightful.

1.2 TRANSCRIPTOME EXPLORATION WITH RNA SEQUENCING

With the recent era of short-sequencing technology and the completion of the Human genome [Venter et al., 2001; Lander et al., 2001; International Human Genome Sequencing Consortium, 2004], understanding the genome expression is increasingly a more reachable aim. From the early 1980s, technologies involved in transcriptome studies have substantially improved through many successive innovations [Lowe et al., 2017; Parkinson et al., 2009] that include Sanger sequencing [Sanger et al., 1975] or PCR (reviewed by VanGuilder et al. (2008)). Among the various transcriptomic study approaches, there are three key methods [Lowe et al., 2017]: EST sequencing (for gene discovery — see Appendix A.3), microarrays (for gene quantification — see Appendix A.4) and the one with which all the transcriptomic data used in this thesis have been generated: RNA-Seq (which is both used for gene discovery and gene quantification).

In the following section, I introduce the typical steps of the required workflow to study the transcriptome through sequencing on an Illumina platform. While not by conscious design, all the transcriptomes analysed in this thesis are the product of Illumina sequencing (see Sections 2.2.1 to 2.2.5). This is unsurprising as Illumina is by far the most popular platform for the last decade [McPherson, 2014] and has been used to generate most of the data in ENA and ArrayExpress. Indeed, Illumina sequencing offers a very good balance between accuracy and achieving the highest throughput for the lowest per-base cost [van Dijk et al., 2014]. I will emphasise the approaches and the tools I used to estimate the gene expression levels from raw nucleotide sequences. Figure 1.4 presents an overview of the typical steps of an RNA-Seq workflow from the libraries preparation to the sequencing.

Experimental protocols for other platforms may need various and specific modifications that are outside of the realm of this thesis and thus will not be covered here[3].

Although the collection and the conservation[4] of the samples before the RNA extraction most definitely affects the final estimations, I will set aside these steps from my review.

1.2.1 *Library preparation*

While there are sequencing technologies that can directly sequence RNAs (see Garalde et al., 2016), most of the technologies handle only DNA. Hence, the first step of a typical RNA-Seq workflow is the preparation of complementary DNA (cDNA) libraries from the starting material. This step and the sequencing itself are the most platform dependent parts of the overall protocol. Indeed, contingent upon which sequencing principle they rely, the sequencers need the libraries to be fixed and loaded differently.

1.2.1.1 *RNA extraction*

There are many methods to extract RNAs from the primary samples, and they are commonly standardised. Indeed, depending on the type of biological samples, the RNAs of interest, the aim of the study and the sequencing platform used, there is one (or more) available commercial kits. These are designed in a way to not interfere with any of the later steps of the library preparation or with the sequencing itself.

Unsurprisingly, the choice of one kit (and hence its method of extraction) over another can impact the final RNA-Seq data. The main difference between the most widespread methods being the quantity of non-mature RNAs (*i.e.* with longer intronic regions) detected according to which kit has been used. However, the relative gene expression levels are similar from one extraction protocol to another. [Sultan et al., 2014]

1.2.1.2 *RNA enrichment*

After extracting the RNA from the cells or tissues, the next step is to enrich the content of the samples with the RNAs of interest (*i.e.* the concentration of the RNAs of interest is increased either by specifically selecting it or by removing other RNAs). Indeed, the

[3] More details on the other main sequencing platforms and their relevant protocols may be found in Goodwin et al. (2016) review paper or at the online resource 'RNA-seqlopedia' (http://rnaseq.uoregon.edu/) [Cresko Lab, 2017].

[4] E.g. between fresh-frozen samples and FFPE samples see Esteve-Codina et al. (2017).

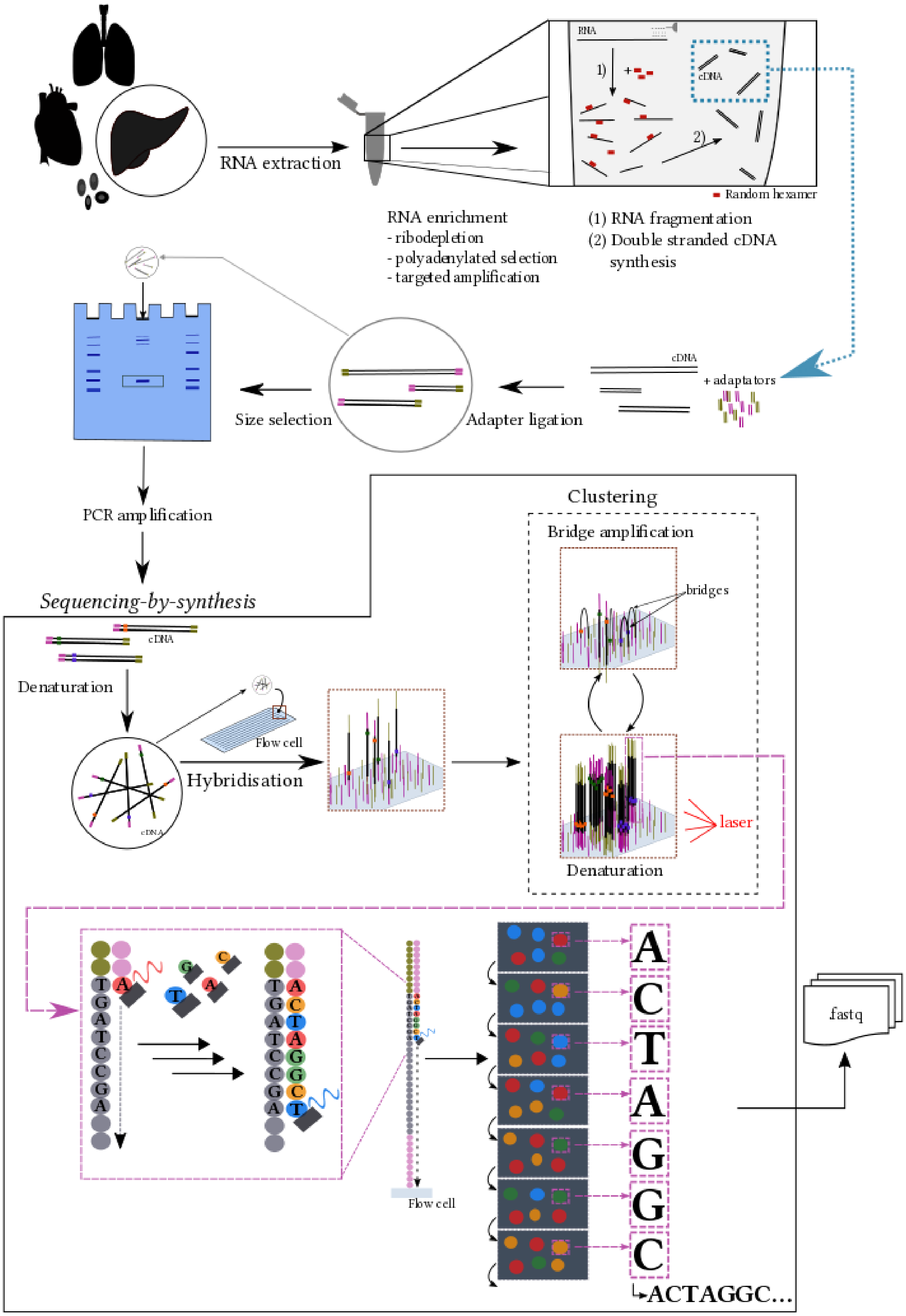


Figure 1.4. **Overview of a typical RNA-Seq workflow: library preparation and sequencing**

rRNAs are the most abundant type of RNA in any cell. Even though they account for a very small part of the genome[5], they represent by their number 70% or more of the total population of RNA [Davidson et al., 1999]. Although there are interests to study rRNAs (*e.g.* Pootakham et al., 2017), mRNAs studies are more popular, and they only constitute about 3 to 5% of the whole RNA population [Alberts et al., 2002]. Other studies research even scarcer kinds of RNA.[6]

There are typically three strategies to achieve RNA enrichment: either by polyA-selection, by ribodepletion or (more complex) by targeted amplification. While these approaches are insufficiently specific to select one particular kind of RNA or remove all rRNAs, it eases and improves the downstream analyses.

PolyA-selection

This strategy essentially targets the mRNAs. It exploits the polyadenylated tail at the 3' end of the mRNAs[7] that is added post-transcriptionally. Magnetic beads, supporting short strings of thymine (oligo-dT), capture these mRNAs efficiently while the others are washed away [Mortazavi et al., 2008].

This protocol is probably the most widespread one as it is the easiest and cheapest to set up. A dataset produced following this protocol is known as a *polyA-selected* dataset.

Ribodepletion

This strategy is preferred for the study of any non-coding RNA (ncRNA) or when researching the interaction of mRNAs with other RNAs [Morlan et al., 2012]. This strategy is in a way the reverse of the previous one as its also uses magnetic beads, but this time to efficiently[8] target the unwanted rRNAs as to remove them from the sample.

The ribodepletion can also be achieved through ribonucleases. These enzymes specifically digest rRNAs, and then RNAs of interest can be retrieved through size selection.

Datasets produced following a ribodepletion protocol are usually called *whole RNA* or *total RNA* in contrast to the *polyA-selected* ones.

Castle et al. (2010) created a total RNA dataset, but they use another approach where they amplify *every* other RNA with the help of specially designed probes (see Section 2.2.1). Hence, the protocol they have used is closer to the following one.

[5] For example, *Homo sapiens*, there are 568 genes (<1%) that are described as rRNA out of the 63,898 annotated genes of the Ensembl database (GRCh38.p10, Ensembl 89).

[6] Out of the 10,081 experiments tagged as `'rna assay'` and `'sequencing assay'` within ArrayExpress, 7,981 were also tagged as `'RNA-seq of coding RNA'`, 1,829 as `'RNA-seq of non coding RNA'` and 366 have both tags. 4 of them are only described as `'microRNA profiling by high-throughput sequencing'` — Query date: 22 June 2017.

[7] And a few other kinds of RNA, *e.g.* long non-coding RNAs (lncRNAs) [Cheng et al., 2005]

[8] ThermoFisher claims that its RiboMinus protocol can remove up to 99.99% of the rRNAs.

Targeted amplification

Targeted amplifications rely on primers that would be designed to target (or avoid as for Castle et al. (2010)) specific sequence motifs of the genome. Most studies based on this kind of approach are referred with a name based on the studied RNA type (*e.g.* miRNA-Seq) or emphasising the variation of the method (*e.g.* Capture-Seq [Bussotti et al., 2016]). Often, additional steps are required to prepare the libraries in comparison with a polyA-selected or ribodepleted dataset.

1.2.1.3 *RNA fragmentation*

Most sequencing platforms[9] require relatively short (*i.e.* 200 to 500 nt) length to sequence. Concomitantly, it also ensures a more uniform sampling along the RNA. This fragmentation can be carried out via divalent cations hydrolysis or nebulisation.

This step is performed on occasions after the cDNA synthesis (see next section). In those cases, the cDNAs are fragmented mostly by digestion with DNase I or by sonication.

1.2.1.4 *Double-stranded cDNA synthesis*

The RNA molecules are used as a template for a retro-transcription involving oligo-dTs or *random* hexamer primers, respectively only for polyA-selected datasets or any dataset (polyA-selected included). The set of random hexamer has been designed to cover the whole transcriptome. Unfortunately, these random hexamer primers have been proven to lack full randomness [Hansen, Brenner, et al., 2010].

At the end of the most common protocol, the order of synthesis of each cDNA strands is lost, *i.e.* it is impossible to distinguish which of the cDNA strands has the same sequence as the original RNA. Several techniques, called *strand-specific*, have been developed to compensate for this [Levin et al., 2010; Parkhomchuk et al., 2009].

1.2.1.5 *Adapter ligation, PCR amplification and size selection*

After generating blunt edges by restriction digest of the cDNAs, adapters (small known sequences of oligonucleotides) are ligated to both their ends. These adapters are constituted from several parts. A subset of them are later ensuring the hybridisation of the cDNAs with the flow cells[10] (based on sequence complementary), and another set of them are sequence binding sites that are used as primers for the following cluster amplification step occurring *in situ*. These adapters are also used to introduce additional motifs such as indexes.

[9] As for the Illumina platforms that have produced the transcriptomic datasets studied in this thesis.

[10] Flow cell: see Section 1.2.2.

The next two steps can be interchanged depending on the amount of starting material at disposition. PCR amplifies all the molecules before (or after) a size-selection is performed (per gel electrophoresis) to extract length-complying fragments (about 200 to 500 bp) to the sequencer machine requirements[11].

Unfortunately, the size-selection means that any transcript with an original length below the threshold used for the selection will be missed[12]. For example, microRNAs (miRNAs) are shorter than the general requirement of Illumina sequencers. Alternative protocols are addressing this issue [Zhuang et al., 2012].

1.2.1.6 *An example of alternative preparation strategy*

Along with the targeted, the strand-specific and small RNAs protocols, there are a few other variations to this typical protocol to handle other concerns. For example, it is occasionally necessary to sequence simultaneously (in a single run) multiple samples. This can either be motivated by practical reasons (to lower the experimental costs or hasten the overall processing time) [Hou et al., 2015] or be critical to the experimental design as a way to experimentally handle the *batch effects*[13] [Auer et al., 2010]. However, it is crucial to later extricate the several pooled samples from each other as a requirement to many downstream analyses.

Multiplexed protocols easily achieve the distinction between the multiple samples as they incorporate *barcodes* before ligating the adapters. These barcodes are also small sequences of nucleotides, and each sample has its unique associated barcode. In practice, each sample is prepared separately with the added extra step (before the adapters ligation) where the barcode is incorporated; then all the samples are pooled together before the next step, which consists of hybridising the cDNAs to the flow cell.

Other extra steps occur just after the sequencing and before any other data analysis: all the reads[14] are separated in files based on their barcodes and the barcode, along with the adapters, is trimmed from all the reads.

The main inconvenient of the multiplexing protocol is that the original sequenced length of the cDNAs are then shorter as the barcodes are also (and have to be) sequenced as well.

[11] Indeed, the previous fragmentation step creates a great length range of fragments.

[12] There is no problem for the greater length as they will statistically present fragments in the correct range.

[13] Batch effect: see Section 1.5.1

[14] Reads: see Section 1.2.3

1.2.2 *Clustering: Hybridisation and Bridge amplification* [Illumina, 2016]

Once the libraries are ready, they are loaded onto a *flow cell*[15].

The clustering step comprises two phases: hybridisation and bridge amplification of the cDNA fragments.

1.2.2.1 *Hybridisation*

The double-strand cDNAs are denatured, and then each fragment randomly hybridises across the flow cell surface with one of its small oligonucleotides. These are used as primers for polymerases which create a first complementary strand to the hybridised DNA fragments. The new double-strand molecule is denatured, and the original first template is washed away.

1.2.2.2 *Bridge amplification*

The strands then fold over and their (second) adapter hybridises with a complementary oligonucleotide sequence of the flow cell and thus creating a bridge. The flow cell complementary fragment is then used as the primer for a new strand. The new double-stranded DNA is then denatured (which dismantles the bridge). Each of the two tethered molecules creates a new bridge by hybridisation which are the templates for a new strand each. This process happens many times and simultaneously for millions of fragments. It creates clusters of clonal amplification of the original fragments of the library. After the bridge amplification, the reverse strands are cleaved and washed away. The 3' end primers are also blocked to avoid any unwanted priming.

1.2.3 *Sequencing-by-synthesis*

Illumina sequencers propose two approaches: single-end and paired-end. In *single-end sequencing*[16], the sequencing begins at one (and only one) of the fragment ends and progresses towards the second. In *paired-end sequencing*, once the first end has been sequenced, a bridge replication occurs, then the other end of the original fragment is also sequenced. Thus, in paired-end sequencing, the sequencing occurs at *both* ends of each original fragment.

[15] *Flow cells* are the support of Illumina sequencing. They parallelise through supported Chemistry the sequencing of millions of DNA fragments together which are kept spatially separated in clusters. Each flow cell is a glass slide with lanes. Each lane is coated with two short nucleotide sequences. One of these oligonucleotides is complementary to a region contained in the ligated adapters.

[16] Chronologically the oldest method

Though more expensive and more programmatically challenging, the paired-end approach facilitate the detection of genomic rearrangements[17] and repetitive sequence elements. It also helps to distinguish between a gene isoforms and provides greater support to identify novel transcripts (new isoform or gene) and fusion genes[18].

In both cases, Illumina's sequencing process, *sequencing-by-synthesis* [Bentley et al., 2008], is the same. It uses the DNA replication mechanism with modified deoxynucleoside triphosphates (dNTPs). Reversible fluorescent tagged dNTPs, which are protected at their 3'end to block any further elongation, allow a step-by-step incorporation. The product of this synthesis is called a *read*, and it supports the *base calling*[19].

The sequencing co-occurs on every identical fragment of every cluster of the flow cell. It begins with the hybridisation of a complementary 5' primer onto the 3' binding site of the tethered DNA template. This primer is then extended by replication through several sequencing cycles to create a new read.

A *sequencing cycle* starts with the addition of one complementary fluorescent dNTP to the new growing read, which stops the replication process as the dNTPs 3' end is blocked. A wash discards all the unlinked dNTPs away. Then, the clusters are excited by a light source, and the signal intensity and (characteristic) wavelength of each dNTPs are recorded since they allow the identification of the new nucleotide incorporated by each cluster and measure the accuracy of the base calling. Finally, the fluorescent tags and the 3' caps are cleaved and washed away before a new cycle begins. The number of cycles determines the final read length.

Unfortunately, as the sequencing proceeds, the error rate of the sequencers increases. This is due to the incomplete removal of the fluorescent signal which increases the background noise and thus reduces the signal-to-noise ratio.

Once the programmed read length is achieved (typically between 25 to 200 nt), the reads are washed away (after denaturation).

1.2.3.1 *Sequencing specificities for the* paired-end *protocol*

The paired-end protocol uses an additional primer. The first run is initiated by a single primer and follows the same steps of the single-end sequencing cycle. Once completed, the complimentary read is washed off, and the 3' end primer is deprotected. Then, the DNA fragment bends over and hybridises to a complementary oligonucleotide at the surface of the flow cell. Next, the second primer initiates a new sequencing cycle at the end of which the newly synthesised read is washed away. A single new bridge replication follows. The new double-stranded DNA fragment is denatured, and the 3' primers are protected before

[17] As indels or inversions

[18] A fusion gene is a gene that is the product of the fusion of parts of two (or more) different genes.

[19] Base calling: identification of the nucleosides in a sequence by assigning chromatogram peaks or another kind of signal variations to (nucleo)bases.

the original strand (that has been already read) is cleaved and washed away. Finally, the remaining strand is sequenced following the previously described sequencing cycle. Once the same number of sequencing cycles as the first strand is reached, the read product of the remaining strand is washed away. By convention, the first primer allows to read the forward strand and the second primer the reverse strand. Note that these forward and reverse concepts used in paired-end protocol have no relevance to the biological concepts of forward and reverse for genes.

1.2.4 *From analogous input to digital output*

At the end of the sequencing process, a set of images across the flow cell (one per sequencing cycle) is produced from the detected wavelengths. While it is possible to work with the images themselves, in most cases, the sequencing facilities will perform the base calling and other intermediate steps before providing the end-user with text files.

These files are usually distributed in the FASTQ format [Cock et al., 2010] which record for each cluster (read) a unique identifier, a nucleotide sequence and a Phred quality score for each base of the sequence. A few optional information can also be provided, *e.g.* the position of the read on the flow cell (See Appendix A.5 for a random read example). The Phred quality score (Q) measures the accuracy of the identification of the nucleobase to which it refers. These scores are set by the base calling program and are defined as $Q = -10\log_{10}(P)$ with P the probability of the base being called wrongly. There are several possible encoding formats (see Appendix A.6).

In single-end sequencing, there is one file per sample. In paired-end sequencing, the reads are usually separated based on their associated indexes into two ordered files: all the reads from the forward strands are grouped in one file, and the ones from the reverse strands in a second one.

1.2.5 *A typical bioinformatic workflow for RNA-Seq study*

From the reconstruction of the transcriptome to the normalisation of expression in each sample, the various steps may be addressed through many different algorithmic approaches. Often, the choice of a method at one stage implies a more limited number of alternatives from which to pick at later points in the pipeline. The choice is frequently driven by the kind of downstream analyses planned for the study. More than the practical format of the data for these, it is the assumptions and the methods used upstream that are critical for a rigorous investigation and, later, for an accurate interpretation of the results.

Figure 2.1 presents an example of the overall *in silico* process of raw RNA-Seq data. It summarises the steps and highlights the tools I used to process the data within this thesis.

Before any downstream analysis, for each read, the genomic region (or *locus*), from which it has been expressed initially, needs to be identified. Indeed, RNA-Seq main objective is to quantify the expression of genomic *features*[20]. In other words, the transcriptome needs to be reconstructed from the short reads and annotated (*i.e.* identify which features have been expressed in each library).

Two different main strategies (see further 'Reconstruction strategies' segment) manage to accomplish this identification step. Independently of the approach, this step is the most challenging and time-consuming of the workflow. Tools, which tackle the reconstruction, usually provide many tunable heuristic parameters (*e.g.* maximum number of allowed mismatches or indels per read before discarding a possible identification[21]) to speed up the task. Unfortunately, as on Illumina platforms, the base calling accuracy decreases along the read length, this may lead to an information loss [Minoche et al., 2011]. To prevent informative reads to be discarded, it is opportune to perform a quality check of the raw data before the identification step. Thus, reads with a drop of accuracy in their 3'end may be shortened (*i.e.* trimmed) and rescued for the next reconstruction step. Similarly, low-quality reads may be discarded hence lowering the complexity of the reconstruction task and hasten its accomplishment.

1.2.5.1 *Quality check, trimming and filtering*

The quality assessment allows removing any read (or part of it) that would increase the complexity of the reconstruction step or skew the downstream analyses.

It is wise to discard uninformative reads, *i.e.* reads with a low sequence complexity (*e.g.* poly-T or poly-A tails) or with ambiguous sequences (in other words with uncalled bases — reported as *N*). Indeed, these reads will hamper the processing time as they usually map to several parts of the genome while also decreasing the accuracy of the global gene expression estimations. For similar reasons, it is judicious to remove reads with a low overall quality score[22].

It is also prudent to check and remove any read that may map to possible contamination sources[23]. Indeed, as these reads are ambiguous, it is safer to discard them than skew the expression estimations.

Finally, as many tools (mappers in particular) require all the input reads to have the same length, the purity-length balance requires optimisation. Indeed, the trimming has to compromise between an approach too lenient (where the mappers discard many unfit

[20] These genomic features could be genes, isoforms, exons, novel genes, … In short, any genomic region with an annotated function.

[21] Indeed, many reads will have many identifications; these reads are defined as *ambiguous reads*.

[22] It may vary based on the complete set of reads to analyse.

[23] For example, for eukaryotes, by aligning (see next segment) every read to the *Escherichia coli* genome.

reads at a later step), and a too stringent one (where either the reads are shorter, which increases the overall complexity and therefore hinder the mapping both on time and accuracy [Williams et al., 2016], or too few are left for pertinent analyses). When the tools allow it, avoiding quality-based trimmings is probably a better practice.

Generally, after the sequencer calls the reads, a first trimming removes all the adaptors and barcodes needed by the sequencing protocols. Thus, in principle, they are not to be found in the 'raw data'. However, to avert any latter contingency, a search against a list of the most common adaptors and an over-representation assessment of small sequences (*k-mers*[24]) at each end of the reads is good practice.

1.2.5.2 *Reconstruction strategies*

Two main approaches can be used for the very computationally expensive step of identification. I will present them in decreasing order of complexity: the *de novo* assembly of the reads and then the reads alignment approach (to a genome reference or a transcriptome one).

Regardless of the approaches, the reconstructed transcriptome is usually reported as a *SAM* file [H. Li et al., 2009] (or one of its derivative formats: either *BGZF-compressed binary file that can be converted into SAM (BAM)* or more recently *CRAM*).

de novo Assembly

This approach is favoured when the reference genome of the species of interest is unavailable or of poor quality (*e.g.* many non-model organisms) or inadequate (*e.g.* cancer samples) for the samples of interest. However, if a reference already exists this strategy is avoided to the utmost.

It allows the unbiased discovery of novel exon-exon junctions [Robertson et al., 2010]. As none of the datasets I use in this thesis has been reconstructed through this approach, I briefly summarise the main points below as more in-depth reviews cover this strategy (see J. A. Martin et al., 2011).

In *de novo* assembly, the reconstruction of the transcriptome happens with the construction of the longest possible *contigs* (*i.e.* contiguously expressed regions) based on sets of overlapping reads (see also Figure 1.5). Shorter reads add to the overall complexity of this approach. While paired-end reads may help to solve many genomic regions, lowly expressed or repetitive regions remain challenging to determine. There are several algorithmic approaches for *de novo* transcriptome assembly [Wajid et al., 2012], though the most prevalent one is the de Bruijn representation [Robertson et al., 2010].

[24] *k-mer*: In the present context, all possible subsequences of length *k* of a *read*.

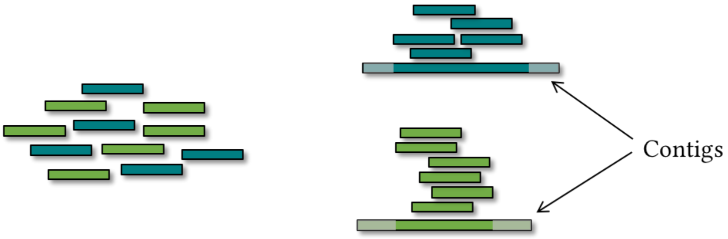
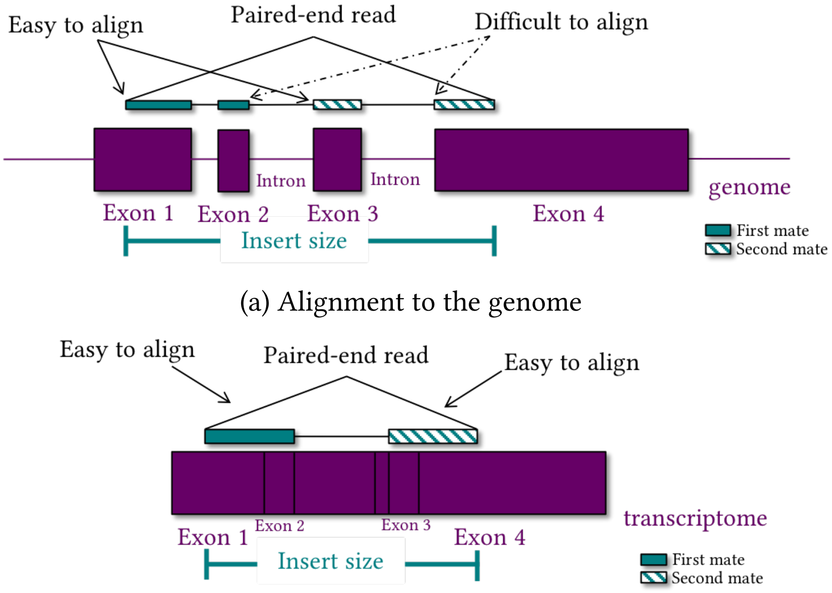


Figure 1.5. ***de novo* Assembly.** From overlapping regions of raw reads, *contigs* are created by integrating the reads sequences together.

Read alignment

This approach exploits prior knowledge. The reads are aligned to a reference to hasten the reconstruction process. The reference may be a genome or a transcriptome (provided that a good annotation is available).



(a) Alignment to the genome

(b) Alignment to the transcriptome

Figure 1.6. **Overview of main reconstruction strategies for an RNA-Seq transcriptome by alignment to a reference**

- **Genome reference**

Aligning to the genome allows discovering new genes or isoforms. However, it requires splice-aware algorithms, *i.e.* they need to align the reads across the splice-junctions (which is possible but non-trivial). As illustrated in Section 1.2.5.2, the reads might span many discontinued regions of the reference. While on the one hand, aligning to the genome avoids *multiple mapping issues*[25] for the same exon, this also implies that the genome needs to provide the coordinates for the different isoforms which will then require further analyses for accurate quantifications at that isomeric level. Indeed, irrespectively of the number of isoforms including a specific exon, the sequence of this exon is transcribed only once in the reference.

- **Transcriptome reference**

Using a transcriptome for reference instead of a genome reduces the complexity of the aligning step due to the lack of intronic sequences. However, it also limits the potential downstream analyses, *e.g.* any new (or unannotated) gene or isoform will be missed. This approach is the easiest, but a pre-existing accurate and well-annotated gene model is required. Section 1.2.5.2 shows in fact that this approach is simpler to the previous one as a direct read alignment is done against the transcriptome of reference. This enables the accurate gene isoforms expression quantification, provided that the gene model is correct and the reads may be attributed unambiguously to a single isoform for each gene. However, in practice, this approach produces many multimapped reads, particularly for shorter reads as many isoforms are very similar in sequence and vary only in the exons they retain. If the difference in the exon compositions is towards the end of the gene, reads from two isoforms may be indistinguishable. Paired-end sequencing helps to resolve part of the ambiguity encountered with single-end protocols.

To mitigate very computational greedy approaches and more constraining ones, several tools complement the previous strategies.

Hybrid approach between *de novo* and alignment

There are tools like *TopHat2*[26] (2.0.12) [D. Kim, Pertea, et al., 2013], that use a hybrid approach between a reference alignment and a *de novo* assembly.

Reads and fragments

In the case of paired-end data, each read of the pair is first processed separately. Then, in the final evaluation phase and with the help of additional information sources, they are used as a pair to infer among the many possibilities which are the most credible ones. Both parts of a paired-end data once aligned to a concordant region of the genome is then called a *fragment* instead of a *read*. Today, there is conflation between the '*read*' and the '*fragment*' terms. Even though the term *fragment* is more accurate and may be used in any situation (as it equals to one read for single-end data and a pair of related reads for paired-end data), it is frequent to see the term *read* instead (even for paired-end data).

[25] Due to sequence similarity, a same read or subpart of a read may be attributed to many different loci in the genome. As it is impossible to attribute the read to its original locus of expression directly, distribution models have to be pondered to avoid unnecessary skewness during the quantification step.

[26] *TopHat2* — https://ccb.jhu.edu/software/tophat/index.shtml

TopHat2 — along with *STAR*[27] [Dobin et al., 2013] — is the most popular splice-aware mapper for genomes with a near-complete annotation (*e.g.* for *Homo sapiens*) [Engström et al., 2013] despite being slower than the latter [D. Kim, Langmead, et al., 2015].

As many concepts or terms in Science, '*read mapping*' can have different (however very closely related) meanings. Hence, while for many people, *read mapping* will encompass any transcriptome reconstruction strategy (including *de novo* assembly) since the main point is to map the features to functional annotations, for others the term will only refer to 'read alignment' strategies specifically [Pachter, 2015]. Tools such as *TopHat2* contributes to this confusion.

1.2.5.3 *Quantification of* features

When working with RNA-Seq data, the typical next step after mapping the reads or fragments to the reference is to quantify the expression of the feature[28] of interest.[29] In the context of this thesis, I only consider gene expression (either as RNA or as protein). Hence, many subtleties required for isoforms or exons studies are here irrelevant and are left out of my overall review.

Several tools and algorithmic approaches are available. Indeed, for larger genomes, many regions may present high sequence similarity which results in many *ambiguous* (and challenging) reads as they mapped to many potential genomic sites. These reads are also called *multireads*. One early strategy to solve multireads is to discard them from later analyses; another one is to attribute them to the most credible locus based on the overall distribution of the reads for a given sample. [Mortazavi et al., 2008] Paired-end data help in many cases to discriminate between possible genomic original sites, thus decreasing the overall number of multimapped fragments.

Two popular tools were used in this thesis, *Cufflinks2*[30] (2.2.1) [Trapnell et al., 2010] and *HTSeq-count*[31] (0.6.1p1) [Anders et al., 2015], which are both compatible with *TopHat2* but rely on different concepts. I briefly present them below in their chronological release.

Cufflinks

Cufflinks2 is part of a collection of tools called *Tuxedo suite*[32] which also includes *TopHat2* and *Bowtie*. *Cufflinks2* can assemble *de novo* novel transcripts and isoforms following the same principles than *TopHat2*. Likewise, using good references is faster and more useful.

[27] *STAR* — https://github.com/alexdobin/STAR

[28] E.g. genes, isoforms, exons, splicing events.

[29] In fact, while genotyping, heredity studies and other genetically focused studies are possible in principle, the common main focus is centred on expression estimation. For example, instead of reporting a specific SNP, in an RNA-Seq study, the core interest is currently more about the specific allelic expression. Moreover, most of the RNA-Seq studies fail to provide the necessary sequencing depth and coverage for other kinds of study.

[30] *Cufflinks2* — http://cole-trapnell-lab.github.io/cufflinks/manual/

[31] *HTSeq-count* — http://www-huber.embl.de/HTSeq/doc/index.html#

[32] *Tuxedo suite* user group: https://groups.google.com/forum/#!forum/tuxedo-tools-users

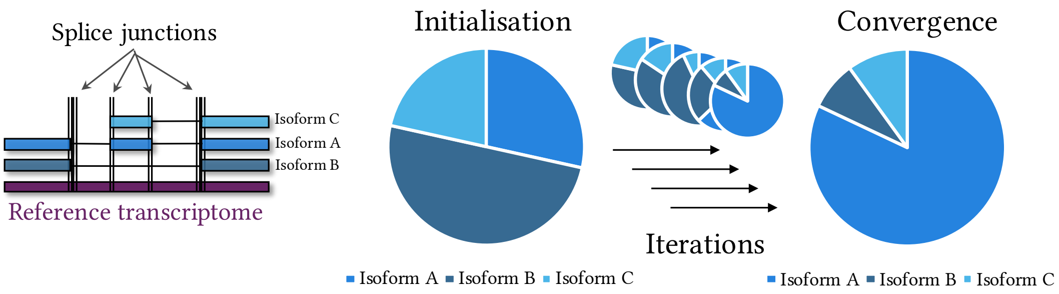


Figure 1.7. **Abundance estimation of isoforms by *Cufflinks2*** following an EM algorithmic approach. [Adapted from Turner (2015)]

In both cases, *Cufflinks2* infers the most parsimonious[33] and credible set of transcripts (and their isoforms) that can explain the complete set of observed fragments.

This task is challenging as many isoforms are sharing a common set of exons. While most genes present a dominant isoform for a specific condition, there are often a few other isoforms expressed along, even though their amount may be very limited [Gonzàlez-Porta et al., 2013].

Furthermore, *Cufflinks2* tries assigning the multimapped reads to one isoform only. First, sets of fragments are separated into sets of isoforms (all fragments that are likely produced from the same set of isoforms are regrouped together). Then to estimate the abundance of each isoform of one set, *Cufflinks2* integrates many information sources together. For example, the overall distribution of fragments (or reads), if these are spanning over known (or novel) splice-junctions. Particular attention is drawn to the fragments that map unambiguously to one unique isoform. When available, paired-end fragments are critical: as they cover longer regions, the probability that they span multiple adjacent exons is increased which helps to resolve the possible structure of the original isoforms. [Roberts et al., 2011]
The abundance is finally estimated through an EM algorithm [Do et al., 2008; Dempster et al., 1977], with the following main steps (see also Figure 1.7):

1. *Initialisation*: For each fragment, a rough estimation of the probability to be expressed from each isoform is computed based on the different piece of information cited previously.

2. *Iteration till convergence with the observed distribution of fragments*:

    a) Isoforms abundance are recomputed based on the updated fragment-to-isoform assignment

    b) Fragment-to-isoform assignment re-updated based on the isoforms abundance.

[33] As a requirement to Occam's razor, which may be a debatable strategy [Westerhoff et al., 2009]

Finally, to compute the gene expression levels, *Cufflinks2* aggregates *per* gene all the isoforms expression abundances together.

Cufflinks2 provides by default FPKM[34] *normalised* data.

HTSeq-count

The Python library *HTSeq* provides a stand-alone script (*HTSeq-count*) performing the feature quantification with a more conservative strategy. It discards all ambiguous reads from a SAM/BAM file and then only counts the *un*ambiguous reads that overlap with the features of interest for a given gene model[35].

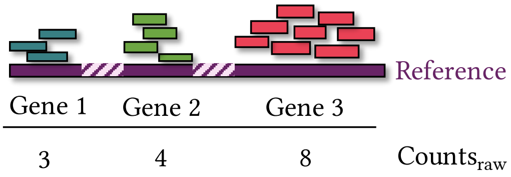


Figure 1.8. **Abundance estimation of genes by *HTSeq-count*.** The unambiguous reads (or fragments for paired-end data) overlapping locus annotated as gene are directly counted. [Adapted from Gonzàlez-Porta (2014)]

HTSeq-count deems as ambiguous any multiread or read that overlaps more than one annotation for the considered feature.[36] *HTSeq-count* provides three modes for fine-tuning the overlap definition. For this thesis, I used the '`intersection non-empty`' mode (see Figure A.4). This mode avoids discarding too many reads due to a too tolerant annotation (*i.e.* the annotation itself presents many overlapping definitions for a given pair of feature and chromosome region).

Initially, *HTSeq-count* [Anders et al., 2015] was designed for differential gene expression analysis (DGEA). This type of analysis compares expression profiles to highlight the genes (or transcripts) for which the expression is significantly different depending on the considered condition. As multireads are irrelevant for those studies, including or excluding them from the downstream analysis is insignificant.

Many papers (*e.g.* Fonseca, J. Marioni, et al. (2014), Robert et al. (2015), and Everaert et al. (2017)) have since shown that the gene expression estimation by *HTSeq-count*, while underestimated, is overall well-correlated with other RNA quantification methods (*e.g.* microarrays or RT-qPCR). Moreover, quantifications with *HTSeq-count* are highly correlated with *Cufflinks2* quantifications for most of the genes after proper

[34] See Section 1.2.5.4.

[35] Gene models are distributed as annotation file (usually either as GTF or GFF format) and refer to a specific reference (genome or transcriptome).

[36] In fact, reads might be discarded for one feature but kept for another one. For example, to quantify the expression of a given gene, *HTSeq-count* considers every read that unambiguously overlaps with any of its annotated exons — indeed, *HTSeq-count* defines a gene as the union of all its exons. However, while quantifying exon expression, many of these same reads may be discarded as they overlap several exon annotations with overlaying definitions.

normalisation [Everaert et al., 2017] as *HTSeq-count* provides *raw counts* (*i.e.* unnormalised counts).

1.2.5.4 *Normalisation*

Regardless of the quantification method, a normalisation is usually necessary to avoid a few statistical biases (mainly due to the sampling). The normalisation method, though, is generally determined based on the quantification (method or tool[37]) and they have to be suitable for the planned downstream analyses. As RNA-Seq fails to assess the absolute concentration of each expressed gene (or transcript) in a sample, each normalisation method is based on a specific set of assumptions that may be incompatible to the ones required by many investigative approaches.

Many papers review or compare normalisation methods (See Dillies et al. (2013), Zwiener et al. (2014), Zyprych-Walczak et al. (2015), and Peixoto et al. (2015)).

RPKM and FPKM

The first evident source of sampling bias is the total number of *mapped* reads (or fragments) between two RNA-Seq libraries (shortened as ‘libraries’ from now on). Indeed, there may be considerable discrepancies in their respective amount of starting material loaded on a flow cell[38] and, more importantly, *the number of mapped reads (or fragments)* to a reference[39].

The second source of bias arises when two genes have their expression level compared. Indeed, as a longer gene produces more reads (or fragments), it has a greater statistical chance to be sampled. Figure 1.8 illustrates this sample bias: Gene 3 is twice as long as Gene 2, and their raw counts also include this scaling factor. However, with proper normalisation, Gene 2 and Gene 3 are expressed in equal proportions.

To correct for these two biases, Mortazavi et al. (2008) introduced a new unit ‘RPKM’ which they first defined as *reads* per kilobase of *exon model* per million mapped reads. Since then, this unit has been redefined and replaced to avoid ambiguities in the case of paired-end data by another unit, ‘FPKM’, which stands for *fragments* per kilobase of *transcript* per million mapped *reads*[40].

The canonical formula for FPKM (or RPKM) is:

[37] Many quantification tools (*e.g. Cufflinks2*) perform the normalisation step automatically as well. They may also (or not) provide raw counts.

[38] In fact, this would involve the monitoring of many parameters or assessments of the samples before the library preparation. Moreover, while it *may* be possible to weight each sample before extracting the RNA, the many steps (involving the fragmentation of the RNAs, PCR syntheses or size-selection) and their associated biases overburden the tracking of the final amounts used for the sequencing.

[39] The quantification disregards the *unmapped* reads (or fragments) and so does the normalisation.

[40] As mentioned before, despite the inaccuracy, *read* and *fragment* are often used interchangeably; this is also the case for *RPKM* and *FPKM*.

$$\hat{\mu}_{ij} = \frac{f_i}{F_j \cdot 10^{-6} \cdot \ell_i \cdot 10^{-3}} = \frac{f_i}{F_j \cdot \ell_i} \cdot 10^9 \qquad \text{(Canonical F/RPKM formula)}$$

where:
$\hat{\mu}_{ij}$ is the normalised expression for the *feature* (*e.g.* gene or transcript) i in sample j,
f_i is the count number of the fragments (or reads) mapped to *feature* i in sample j,
F_j is the total count number of all the fragments (or reads) mapped in sample j,
ℓ_i is the length of *feature* i.

The scaling factor was introduced such as in most cases 1 FPKM is crudely equivalent to a single RNA molecule in the cell [Mortazavi et al., 2008]. This has been observed in other papers (see for example Hebenstreit et al., 2011) and also explains why 1 FPKM is a commonly used threshold.

This normalisation is quite intuitive and still largely used today. In fact, I use this normalisation through the thesis. Meanwhile, it is also unsuitable for a popular type of analysis, differential expression analysis (DEA), which seeks to highlight genes which expression varies between different biological conditions. However, if for any biological or technical reason, any set of genes detected in a specific condition and undetected in another, will affect the FPKM estimation of *every* RNA in both conditions (see Table A.3) and will entangle the interpretation.

Other normalisation approaches

Differential expression analyses (DEA) have led to the development of distinct models and methods. They generally involve a model where for most of the genes, the expression is assumed to be stable between conditions[41] (*e.g. edgeR*[42] [Robinson et al., 2010] or *DESeq2*[42] [Love et al., 2014]).

Also, many normalisation methods applied first to microarrays are used, *e.g.* the most common ones include a quantile normalisation method or a simple scaling normalisation.

Other normalisation methods try to correct *a priori* or *a posteriori* biases[43]. A few of them may correct *batch effect* (see Section 1.5) or other confounding factors. *RUVSeq*[42] [Risso et al., 2014] is one example.

Although FPKM normalisation is generally avoided in favour of another (more appropriate) method, this thesis aims to explore the baseline expression of the genes between and *within* tissues, and in this context, despite its biases, FPKM normalisation is better suited than any normalisation method designed for DEA. For this reason, the work based on transcriptomic data presented in this thesis is based on FPKMs (see Chapter 2).

[41] The comparison is usually between diseased (or treated) samples to control (healthy).

[42] Bioconductor package

[43] The Bioconductor package *CQN* [Hansen, Irizarry, et al., 2012] for example corrects the expression levels according to their length and their *GC* content before applying a quantile normalisation (as cDNAs enriched in GC bases are more stable and tend to a more optimal amplification).

1.3 PROTEOME EXPLORATION WITH MASS SPECTROMETRY

Through the last decade, the proteomics field has shifted from technical research on instruments and methods to the extensive and routine use of mass spectrometry (MS) as an analytical tool [Aebersold and Mann, 2016]. Many possible workflows for MS-based proteomic studies exist as MS is very versatile and supports many proteomic investigation approaches, such as protein characterisation, modification sites, structures, mechanism-oriented (interaction) studies [Aebersold and Mann, 2016]. In this regard, high-throughput protein identification and quantification have thoroughly developed when MS became the primary choice method since it allies a good dynamic range with high sensitivity and specificity [Aebersold and Mann, 2003; Brosh, 2009; Cox and Mann, 2011].

Depending on the study purpose, available time and money, the number of samples, the available instruments and the needed sensitivity and specificity, the choice will be based on one strategy rather than another. Although, *top-down* and, more recently, *middle-down* approaches exist, *bottom-up* approaches are the most favoured in the field by far.

Top-down approaches [Catherman et al., 2014] study the intact proteins, (*i.e.* as a whole without digesting them in smaller peptides). They are appealing, but still very challenging both experimentally and computationally [Aebersold and Mann, 2016]. They are more suitable for highly purified samples as the MS and fragmentation (tandem MS (MS/MS)) spectra are highly complex. On the other hand, digesting the proteins in smaller molecules allows them to ionise better and facilitates the spectra interpretation. *Middle-down* approaches [C. Wu, J. C. Tran, et al., 2012] produce large fragments (up to 20kDa). *Bottom-up* approaches generally use enzymatic digestion with trypsin to produce small peptides (about twelve aas on average). The cheapness, ease and reproducibility of the trypsin digestion[44] are the reason for the bottom-up approach popularity[45].

Bottom-up methods fall into two main types of strategies: *targeted* and *untargeted*. Another layer of complexity is added by the selected MS acquisition mode: data-dependent acquisition (DDA) or data-independent acquisition (DIA) (see Section 1.3.3.5). Since all the proteomic data presented in this thesis have been generated as part of global discovery studies through DDA bottom-up label free approaches (see Chapter 2), in the following section, I mainly focus on the obtention and processing of these types of proteomic data.

Figure 1.9 shows a summary of possible bottom-up approaches based on DDA methods. DDA targeted approaches [Domon et al., 2006; Shi et al., 2016] allow the absolute or relative quantification of a small preselected set of proteins. This strategy is favoured for example to validate possible biomarker candidates. Selected reaction monitoring (SRM) [Picotti et al., 2012] (also known as multiple reaction monitoring (MRM) [A. Hu

[44] See section 1.3.2.3

[45] The experimental spectra are compared to databases that collect theoretical spectra or experimental ones from prior studies to reconstruct the proteins from the peptidic fragments (hence *bottom-up*).

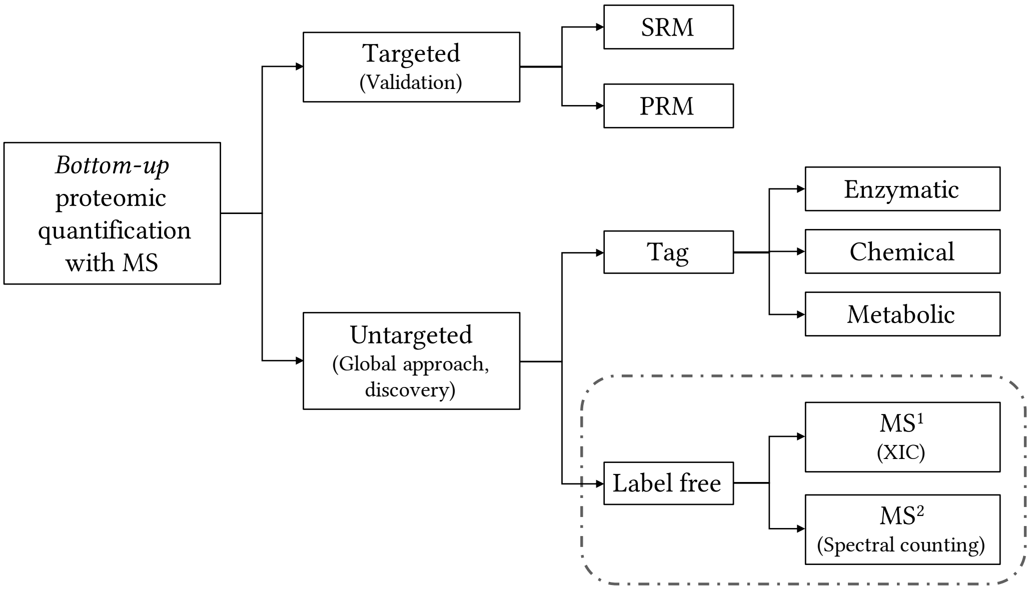


Figure 1.9. **Bottom-up quantification approaches**. The work presented in this thesis relies on proteomic data that has been acquired through a DDA bottom-up label-free MS/MS approach (dashed frame).

et al., 2016; Shi et al., 2016]) and its variant parallel reaction monitoring (PRM) [Gallien et al., 2012] are more sensitive and specific, and give more accurate quantification than untargeted methods: as the set of proteins and peptides to fragment is known, interpreting the spectra is much easier. However, SRM is the most sensitive, while PRM is the most specific and accurate [Benhaïm, 2017].

Among DDA methods, untargeted strategies are the most suitable for global approaches or discovery projects. Unless samples comprise a subset of proteins (or spike-ins) with known concentrations, these methods provide only relative protein quantification.

Tagged strategies [Zhou et al., 2014] label the proteins (before or after extraction) with stable isotopes (see Table A.4). For each condition, a specific isotope is used. Thus, the proteins and peptides have exactly identical physicochemical properties, have the same behaviour through the protocols, and only their mass can differentiate them. The labelling can be either enzymatic (*e.g.* ^{18}O-labelling [X. Ye et al., 2009]), chemical (*e.g.* isotope-coded affinity tag (ICAT) [Gygi, Rist, et al., 1999], isobaric tag for relative and absolute quantification (iTRAQ) [Ross et al., 2004], tandem mass tags (TMT™) [Thompson et al., 2003], dimethyl labelling [Hsu et al., 2003], or 2D-differential in-gel electrophoresis (2D-DIGE) [Unlü et al., 1997]), or metabolic (*e.g.* stable isotope labeling by amino acids in cell culture (SILAC) [X. Chen et al., 2015]). Some tagged approaches have multiplexing protocols. Note that other mass tags (*e.g.* metal coded affinity tag) attached to peptides or proteins are an alternative to isotopic labelling.

Label-free strategies [Hsu et al., 2003; Neilson et al., 2011; Sandin et al., 2014] analyse the proteins after their digestion with the trypsin. Since there is no marking in these methods, multiplexing is impossible. Two different methods can quantify the relative abundance of the proteins. Spectral counting (also called MS2 quantification) quantifies proteins based on the assumption that the more abundant a protein is, the more this protein is selected to be fragmented; thus the total number of MS/MS spectra that can be mapped back to the protein can be used to estimate it. Unfortunately, this technique is lacking accuracy and is highly criticised compared to the second method which is intensity based.

The second method, extracted-ion current (XIC) [Higgs et al., 2013], quantifies each peptide by first extracting its ion currents, its molecular mass and retention time (MS1) (or the ones of its fragments (MS2)), and then by integrating the area under the curve (AUC) of the peptide monoisotopic molecular mass (p), and the following isotopic molecular mass peaks (p+1) and (p+2). The XIC assumption is as follows: the more concentrated a peptide is, the greater is the AUC.

Note that since there are many possible MS approaches and protocols [Bantscheff et al., 2012], in the following section, I focus on the one that initially generated the proteomic datasets I am reusing for this thesis (see Section 2.3). Thus, I describe one widespread *bottom-up* approach, *i.e.* the label-free Liquid Chromatography (LC) followed by tandem Mass Spectrometry (LC-MS/MS) protocol (also known as *shotgun proteomics*) [Cox and Mann, 2011; Y. Zhang, Fonslow, et al., 2013]. The following segments may be suitable for other methods as well.

1.3.1 *Sample preparation*

DDA label free protocols to prepare samples for discovery proteomic analyses are generally much simpler to implement than the ones for RNA-Seq. However, as proteins are also more complex and heterogeneous than DNA and RNA [Bruce et al., 2013], there is a broader choice of them in order to adapt to any requirement [Feist et al., 2015].

1.3.1.1 *Sample collection and conservation*

Collection

Feist et al. (2015) report that traditional dissection, biopsies, blood draws and other methods can deliver adequate samples for proteome analysis.

Conservation

Recent developments have significantly improved proteome analysis from formalin-fixed paraffin-embedded (FFPE) samples [Steiner et al., 2014]. As they are still evolving, they may compare soon to fresh-frozen (FF) samples. However, for now, fresh or FF samples are remaining the best primary sources.

1.3.1.2 *Protein extraction and contaminant removal*

Protein extraction

Contrariwise to the collection step, Feist et al. (2015) explain that the crucial consideration is the cell lysis and the extraction approaches used on the protein as they may interact and disrupt the later characterisation step and thus require appropriate picking. Besides, the physicochemical properties of the proteins are far more heterogeneous than for DNA or RNA molecules [Bruce et al., 2013] and may be incompatible with many extraction protocols.

Contaminant removal

Gutstein et al. (2008), Bodzon-Kulakowska et al. (2007), Visser et al. (2005), and Hilbrig et al. (2003) review various examples of mechanical and chemical extraction and contaminant removal methods. Indeed, contaminants and detergents need to be eliminated from the samples before analysis as, in the typical bottom-up approach, they interfere with the digestion, the separation and fragmentations steps; precipitation and filtering (based on molecular weight cut-off) strategies are the best in the context of bottom-up MS analysis [Feist et al., 2015]. Indeed, many extraction solvents are inappropriate for MS. Precipitations may denature the proteins, but this is usually irrelevant in this situation. Feist et al. (2015) review different precipitation protocols and emphasise that caution is needed for the next re-suspension step to avoid missing a substantial part of the sample proteome. They list several techniques and approaches in this regard.

1.3.2 ***Reducing samples' complexity***

High-throughput bottom-up workflows (including LC-MS/MS) generally aim to decrease the sample complexity to analyse while increasing the depth of proteomic coverage [Z. Zhang et al., 2014; Bruce et al., 2013; Cox and Mann, 2011]. Indeed, many aspects may impede the characterisation of the proteins. For example, the broad physicochemical scope of the proteins will require various protocols to be handled efficiently. Their concentrations in a sample may saturate the MS characterisation capacity as the very abundant proteins are easily detected and quantified while rarer proteins may be missed entirely without specific precautions [K. Liu et al., 2009; Cappadona et al., 2012] or without unbearably increasing the analysis time per sample [Nilsson et al., 2010]. On the other hand, strategies to decrease saturation effects can also be used instead [Z. Zhang et al., 2014].

Hence, the usual workflow will usually involve the denaturation, the reducing, the alkylation and the digestion of the proteins. The peptide mixture products are then separated in smaller fractions before subjection to MS so as to increase the coverage

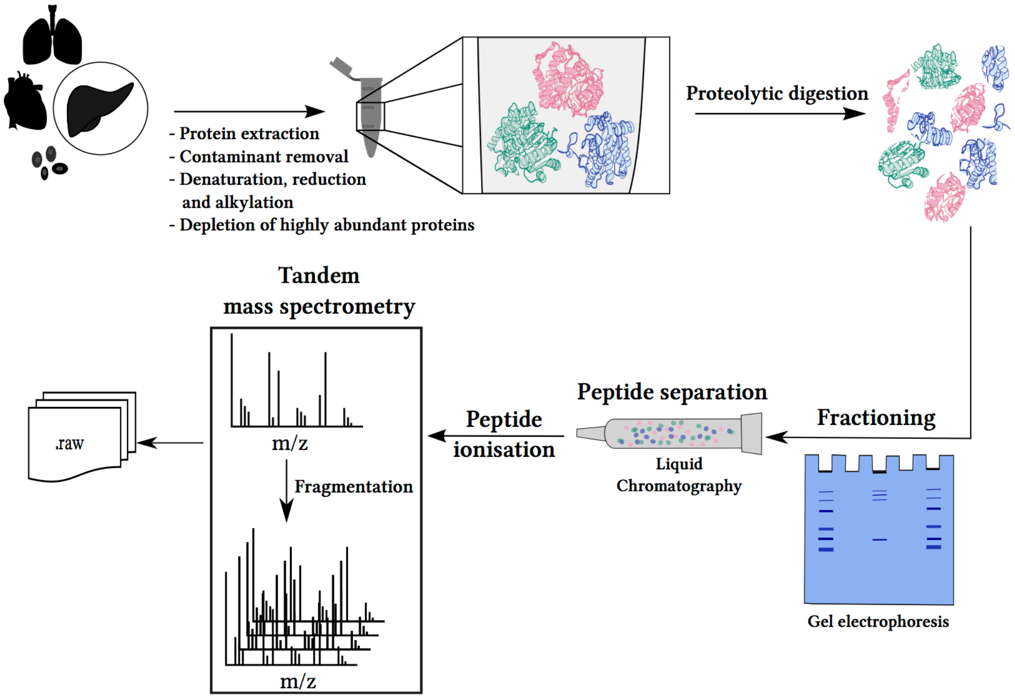


Figure 1.10. **Overview of proteomic data generation**.

depth while keeping a reasonable analysis time-frame [Aebersold and Mann, 2003; Cox and Mann, 2011; Y. Zhang, Fonslow, et al., 2013]. Figure 1.10 presents a general overview of one possible workflow. Indeed, the various complexity reducing steps may happen in a different order than I report hereinafter; they may also happen concomitantly. Additionally, protocols may skip or, conversely, perform the same type of complexity reducing step several times. For example, protocols may present *protein* fractionation as the first step and then *peptide* fractionation as a later one. Moreover, almost all protocols will only involve liquid chromatography for their fractioning steps.

1.3.2.1 *Denaturation, Reduction and alkylation*

These steps help the separation of the protein complexes, the relative linearisation of the proteins and, to some extent, the homogenisation of the crude mixtures [Feist et al., 2015; Bruce et al., 2013]. They may happen simultaneously with other steps. For example, they may occur during the extraction, the depletion or the digestion (where they facilitate the trypsin cleavage).

1.3.2.2 *Depletion of highly abundant proteins*

Regrettably, the proteomic field lacks amplification methods, and there are very few strategies to remove the highly abundant proteins which removal is required to capture scarcer proteins in untargeted DDA bottom-up protocols. Fortunately, these are very limited in number and can be precisely targeted. Z. Zhang et al. (2014) report two useful strategies. The first one aims to remove them entirely from the sample, either by selective precipitation or (more expensive) by affinity. The second approach aims to the equalisation of the proteome. It may be based either on combinatorial ligand libraries involving bead-supported ligands (on a similar model to RNA-Seq protocols) or on specific protease mixtures.

In general, the equalisation of the proteome improves the characterisation of the scarcer proteins [Z. Zhang et al., 2014] while they may introduce skewness to the analyses (due to the relative protein proportions).

1.3.2.3 *Proteolytic digestion*

It may seem contradictory to digest the proteins into peptide mixtures as a means of reducing the complexity. Nevertheless, while the main drawback is the inability to distinguish between proteoforms [Bruce et al., 2013], it improves the proteins characterisation on several points.

- It helps to homogenise the sample: peptides present closer physicochemical properties to each other than proteins [Z. Zhang et al., 2014]. Also, peptide separation (by gel or liquid chromatography (LC)) is easier than for proteins (see Section 1.3.2.4).

- MS is also more sensitive to peptides than to proteins due to being more sensitive towards lower molecular-weight molecules [Vitek, 2009; Cox and Mann, 2011]. Proteins may also be too large to be fragmented (*e.g.* with CID) [Bruce et al., 2013].

- It is also easier to accurately characterise (identification and quantification) smaller molecules. Large proteins with similar compositions present very similar molecular mass and may be impossible to discriminate. On the other hand, the sequence-specific enzymatic digestion gives hints on the protein sequence.

- Finally, it increases the coverage of the less abundant proteins [Z. Zhang et al., 2014]. Indeed, each protein is represented by multiple peptides hence increasing the sampling probability. Often, one or a small number of LC-MS/MS-characterised peptides is enough to identify the parent protein [Bruce et al., 2013].

The digestion may happen in-gel or in-solution. Although less commonly used as before in bottom-up approaches based on LC-MS/MS, in-gel digestion provides directly the fractioning and deals better with the more complex samples. Often, the gel will contain, in addition to the protease, a few chemical reagents that handle other steps (*e.g.* chaotropic reagents to denature the proteins at the same time). However, it requires greater time and amount of starting material than for the in-solution digestion. In-solution digestion, on the other hand, is the simplest and the most popular digestion approach among all the proteome studies in general. It usually precedes filtering and fractioning steps. An hybrid method combining both methods, filter-aided sample preparation (FASP) [Manza et al., 2005; Wiśniewski, Zougman, et al., 2009] exists.

Although there are other enzymes for restrictive digestion [Giansanti et al., 2016; Tsiatsiani et al., 2015], trypsin is the gold standard protease [Z. Zhang et al., 2014]. Trypsin is a serine protease that has a high proteolytic and highly specific activity: it cleaves the proteins at the carboxyl side of an arginine (R) or lysine (K) aas, when they are not followed by a proline (P). Trypsin's cheapness and ease of use can also explain its popularity. Besides, trypsin creates peptides that are in the ideal range for MS studies as it produces a large number of short (600 to $1,000$ Dalton (Da)) peptides that can be efficiently fragmented and identified, although inferring the parent proteins remains challenging [Laskay et al., 2013]. Circularity may also contribute to this trend: as more studies are using it, more data are available for comparison; hence more studies employ it.

Peptides produced by trypsin digestion are referred to as tryptic peptides.

1.3.2.4 *Separation methods (fractioning)*

Fractioning (or fractionation) is the principal method to simplify sample complexity. It also allows focusing selectively on subcellular fractions when needed. Indeed, one may want to study particular organelles, cell compartments or other kinds of the proteome (*e.g.* phosphorylated or glycosylated proteins) [Cox and Mann, 2011]. Besides, fractioning is also a good strategy to reduce the impact of undersampling and increase repeatability between analyses. With MS being prone to undersampling, repeated analyses may fail to yield the same protein identifications, as different sets of peptides may get sampled.

Several methods may fraction peptide mixture. Protocols may involve many of their various combinations. For example, M.-S. Kim et al. (2014) and Wilhelm et al. (2014) have used a gel and LC sequentially before MS analyses.

Precipitation

It is very easy to perform, and they usually involve solvent gradients. While their use is frequent for desalting crude mixtures, it is also quite limited for protein mixtures as other separation methods (based either on gel or capillarity such as LC) are more performant.

Gel electrophoresis based separation

Protocols may involve a first gel-based separation method before the liquid-based separation and fragmentation of LC-MS/MS [Feist et al., 2015].

The gel separation may comprise one step or two, which are respectively named one-dimension (1D) and two-dimensions (2D) gels.

1-D gel approach comprehends a denaturing alkylating gel (usually Sodium Dodecyl Sulphate-PAGE (SDS-PAGE) but others are possible, *e.g.* Lithium Dodecyl Sulphate-PAGE (LDS-PAGE)) [Shevchenko et al., 2006]. Thus, proteins lose all their quaternary, tertiary and secondary structures. The separation relies then on the length of the proteins. Indeed, SDS molecules carry negative charges, and they bind to the proteins proportionally to their length, and as an electric field is applied to the gel, the proteins migrate towards the positive side of the gel at different speeds due to their difference in their mass-charge ratio. The 1D protocol is faster than the 2D one, while the latter is more selective as it relies on very similar principles but in two separate steps[46].

Liquid chromatography (LC)

Chromatography is a technique of choice for the separation of mixtures into their components or — at least — in simpler mixtures. LC, as any chromatography, involves a *mobile* and a *stationary* phase. The mobile phase comprises an eluent (*i.e.* a solvent) and the mixture to be separated. A column plays the stationary phase part. High pressure is applied (*e.g.* HPLC (high-performance liquid chromatography) or UPLC (ultra performance liquid chromatography)) to improve and accelerate the process.

The separation relies on the difference of affinity of the mixture components between the mobile and stationary phases. Many combinations are possible. However, any poor choice may precipitate the mixture on the (*extremely*) expensive column which means that both the column and the mixture will be lost. Hence, in discovery mode and for complex mixtures (as for proteins extracts), instead of using a *normal phase*, *i.e.* crude silica-gel, column (which interacts tightly with polar molecules such as peptides and may prevent them to interact with the mobile phase afterwards), a *reversed phase* column is more common. In these columns, the silica-gel is modified and has been attached to long hydrocarbon chain. Thus, the column is unable to fix anything permanently. Along with a Reversed-phase LC (RPLC) column, polar eluants are used. These interact strongly with charged proteins and peptides. Hence, the separation occurs on the polarity of the molecules and the more hydrophobic a molecule, the longer it will remain in the column as it will present fewer interactions with the eluent.

The ease of coupling this powerful method to the (also powerful) MS explains why Liquid Chromatography (LC) followed by Mass Spectrometry (LC-MS) is so widespread today. Their combined use also allows high repeatability and optimised running time.

[46] Usually, the proteins are first separated based on their isoelectric point (or pI) in a *native* gel (as opposed to a denaturing one). An ampholyte reagent is added to the gel which ensures a stable gradient of pH through the gel. The proteins migrate until they reached a pH region where their overall charge is neutral. In a second time, the proteins are separated perpendicularly this time on their mass after the addition of SDS or an alike reagent.

1.3.3 *Characterisation through fragmentation profiles*

The first reported use of the principle underlying MS happened in 1913 when Sir Joseph John Thomson channelled a stream of neon ions through an electromagnetic field and captured its deflection on a photographic plate [Thomson, 1913]. Arthur Dempster and Francis W. Aston created the first mass spectrometers in 1918 and 1919 respectively [Aston, 1919]. In this context, the use of MS is quite recent in biology for proteomics as it has been developing from the 1980s onwards particularly with the development of soft ionisation methods [Papachristodoulou et al., 2014].

1.3.3.1 *General principle*

MS relays on the following principle: molecules of interest are ionised into charged particles, then the mass analyser separates them in the gas phase based on their total mass (m) to charge (z) ratio, *i.e.* m/z. A detector collects all these ions and translates their signal (intensity versus m/z) to an electric one which is the output serving as raw data for the later analyses. In the simplest case, the molecules are singly charged ($z = 1$), and only their molecular mass is recorded. However, it is quite common for the molecules to carry more than one electric charge and that if the energy used for the ionisation is substantial, they may also forego fragmentation and internal reactions that will result to the production of many spectra. The collection of spectra obtained from a single peptide ultimately increases the accuracy of the mass measurement and the identification of the peptide [Papachristodoulou et al., 2014]. Recent developments have shown that the use of two mass spectrometers in tandem (MS/MS) reduces the number of missed proteins significantly while keeping reasonable running times and it is more potent than excessive fractioning as is the improvement of HPLC [Cox and Mann, 2011].

1.3.3.2 *Ionisation*

While there are other methods (more adapted to small organic molecules) two *soft* ionisation methods are routinely used for proteomic samples: matrix-assisted laser desorption ionisation (MALDI) and electron spray ionisation (ESI).

MALDI is very useful for large molecules. It allows creating ions with a minimum of fragmentation (if any at all). The molecules are fixed onto a matrix and then a pulsed laser irradiates the sample which provokes the ablation and then the desorption from the matrix and ionisation of the molecules. However, it requires mass spectrometers compatible with this ionisation technique [Z. Zhang et al., 2014; Walther et al., 2010].

ESI is currently the most popular technique of ionisation, mainly because it may be used with a large panel of different analysers and it interfaces efficiently with HPLC or UPLC. An electrostatic method performs the ionisation. The application of a high electric potential on a needle through which a liquid (containing the dissolved analyte peptides) is passing provokes its dispersion into small and highly charged droplets. These droplets start to evaporate to the point where the charge on their surface is so high that the desorption of the analytes occurs. The analytes are then in an ionised form (often carrying many charges, contrary to MALDI). They are then released into the mass spectrometer. [Walther et al., 2010]

1.3.3.3 *Mass analyser*

Many different mass analyser designs are available. They all share two properties: they accelerate the ions (in a vacuum), so all the ions share the same kinetic energy, and then they deflect (and thus resolve) the ions based on their various m/z. Many are also capable of trapping and storing specific ranges of ions for more in-depth analyses until they release them based on their m/z. The most common commercial analysers include quadrupole (Q), linear ion trap (LIT), time-of-flight (TOF), ion traps and most recent Fourier transform (FT) analysers such as ion cyclotron (ICR) and Orbitrap™, which has become the most popular. The choice of one analyser over others or any combination of them depends on many factors (including availability). [Haag, 2016]

In Appendix A.7, I review the analysers involved in the production of the raw proteomic data used in this thesis. Indeed, all the datasets have been produced by a combination of linear trap quadrupole (LTQ) and Orbitrap™.

1.3.3.4 *Fragmentation techniques* [*Z. Zhang et al., 2014*]

The overall quality and success of peptide (and then protein) identification depends largely on the quality of ion fragmentation. Tandem-MS (MS/MS) improves the identification of the peptides by cleverly increasing the number of fragmentations. Indeed, the first MS is used to select specific ions (hence known m/z) to pass to a second MS. Thus for each MS1 spectrum, a collection of fragmentation mass spectra can be gathered. The peptide sequences are then more likely to be accurate, and the risk of false positive is significantly decreased.

While the experimenter may want to limit fragmentation, they may also often use dedicated techniques to increase it. There are schematically three different classes of fragmentation techniques: collisional, electron-based or photon-based. The collisional category, CID (see Appendix A.7), is widespread. The chosen ion is introduced in the collision cell, and its collision with an inert gas particle produces kinetic energy, which transforms into internal energy. The fragmentation happens when this internal energy is sufficient to activate the dissociation of the ion. Another related and slightly more effective method for higher charge state ions is the higher-energy collisional dissociation (HCD). This latest method (developed by ThermoFisher for Orbitrap™ analysers), has gained popularity with the development of quantitative proteomics (*e.g.* iTRAQ) as it provides in parallel peptide identification and quantitation. In the electron-based category, the most common is probably the electron transfer dissociation (ETD) technique: an anion donates an electron to a cationic peptide, and this transfer initiates the fragmentation of the peptide backbone [Syka et al., 2004]. For the other categories, see the review from Z. Zhang et al. (2014) and the included literature.

1.3.3.5 *Acquisition modes*

There are two possible acquisition modes: data-dependent acquisition (DDA) and data-independent acquisition (DIA).

The common DDA bottom-up label free protocol has many advantages [Aebersold and Mann, 2016]. It is untargeted and free from any hypothesis, and hence a great tool for global discovery study. It surveys the proteome all at once, and prior knowledge is unrequired. Its popularity increased concomitantly to the increased availability of high-quality genome and gene sequence databases and more recent technical advances in MS (development of new protein/peptide ionisation and fragmentation methods) [Aebersold and Mann, 2003; Cox and Mann, 2011; Z. Zhang et al., 2014]. The main DDA limitation [Guillaumot, 2017] is its inability to select all the existing peptidic ions for fragmentation by MS/MS. The stochastic selection is biased both by each peptide ionisation efficiency and the possible peptides coelution during the LC before being introduced in the MS. In practice, the instrument first quickly scans the current eluting peptides (MS1) before a few (the “top N”) precursors are selected one after the other for fragmentation (MS2). One MS1 spectrum with N MS2 spectra constitute one duty cycle, which usually lasts about 1s. Peptides elute over 30 to 40s, and software tools that drive the instruments can predict peptidic peaks and thus will increase the chance of high-quality MS2 spectra by selecting the peptides at their maximum abundance. A dynamic exclusion window avoids the same peptides being repeatedly selected and new targets to be fragmented. MS1 spectra allow selecting potential peptides to fragment, and their intensity. MS2 spectra are used to identify the peptides later.

DIA approaches are more complicated as no precursor is selected and all the coeluted peptides are fragmented at the same time. The sustained interest in DIA methods is because they provide a less biased overview of the proteomes, although generally more restricted. They can generate comprehensive fragment-ion maps for specific proteoforms [Chapman et al., 2014]. Often DIA is performed after a short DDA survey that establishes a small reference spectrum library to help with the analysis of the MS2 spectra. In the past decade, the number of DIA studies has expanded, particularly with methods such as SWATH-MS [Gillet et al., 2012] (for proteome quantification). DIA methods are likely to gain even more popularity as many efforts are put into their development [A. Hu et al., 2016].

1.3.4 *Bioinformatic strategies for proteomics studies*

MS-based proteomics is quite challenging, and, in most cases, the major bottleneck of proteomics pipelines remains the data analysis [Y. Chen et al., 2016; Tyanova, Temu, Sinitcyn, et al., 2016]. Since mass spectrometers' raw data output for proteomics is directly uninterpretable, it needs processing before being meaningful. Because the sheer amount of produced data can reach the terabyte (TB) range, it prohibits any manual handling and requires automation [Codrea et al., 2016], particularly for the peptide and protein identification steps [Nilsson et al., 2010].

The protein identification process leads to three computational challenges: peptide identification, protein inference and result validation [T. Huang et al., 2012]. Besides, shotgun proteomics produce highly redundant data: peptide subsets that ionise better than the rest are repeatedly and preferentially selected for fragmentation and thus will produce more MS/MS spectra [Eriksson et al., 2007; Koziol et al., 2013]. On the other hand, certain subsets of peptides can be undetectable with the currently available technology; they hardly ionise, or have a weak signal that is masked by other more abundant or more ionisable peptides. Thus, shotgun experiments are plagued by missing data [Stead et al., 2008; Lazar, Gatto, et al., 2016].

As for RNA-Seq, for each proteomics analysis step, there are many tools. Many integrate a few (if not all) of the steps described in the following pages, *e.g.* *MaxQuant*[47] [Tyanova, Temu, and Cox, 2016], *OpenMS*[48] [Pfeuffer, Sachsenberg, Alka, et al., 2017], *Skyline*[49] [Pino et al., 2020] or *Crux*[50] [Park et al., 2008; McIlwain et al., 2014]. From the raw data acquisition to downstream analyses (or any of the intermediate steps), pipelines integrate many different combinations of tools [Vitek, 2009].

Pipelines vary depending on experimental design, data type (*i.e.* expression, modification state or interactions map) and the data creation practicalities (*e.g.* method of separation, mass analyser kind, acquisition mode). Many factors are interconnected in MS and while individual effects have been extensively studied, designing a sound pipeline may be overlooked easily [Sun et al., 2012; Maes et al., 2016]. However, this risk is reduced as many tools imply working upstream or downstream with specific tools, *e.g.* the output of *MaxQuant* (raw data to protein quantification) is ready for use by *Perseus*[51] [Tyanova, Temu, Sinitcyn, et al., 2016] (for analyses such as pattern recognition, time-series analysis, cross-omics comparisons and multiple hypothesis testing). Besides, as files from mass spectrometers are usually encoded in proprietary (commercial) formats, there may be a limited choice of available tools or software for a specific combination of analysis and data. A few open file formats (and their appropriate converters) exist (*e.g.* mzML

[47] *MaxQuant* — http://www.maxquant.org
[48] *OpenMS* — http://www.openms.de/
[49] *Skyline* — https://skyline.ms/
[50] *Crux* — https://crux.ms/
[51] *Perseus* — http://www.perseus-framework.org

format [Martens et al., 2011] and *ProteoWizard*[52] [Chambers et al., 2012]), however, they may fail to record a few critical points of the raw data, and, so, they may also be unfit for particular analyses.

Spectronaut[53] [Bernhardt et al., 2014] is one example of specialised pipelines. *Spectronaut* handles the targeted analysis of DIA[54] data.

As bottom-up approaches are the most common, their associated tools also tend to dominate the MS-based proteomic bioinformatics [L. H. Lee, 2015]. In the following pages, I review the steps involved in the MS/MS pipeline, presented in Chapter 2, that has processed all the proteomic data presented in this thesis.

1.3.4.1 *Signal processing and peak-picking*

The signal processing step comprises the (noise) filtering (or *denoising*), the baseline correction (which eliminates systematic trends), the signal normalisation (or *centroiding*), and the peak picking (*i.e.* mass peaks detection) [Codrea et al., 2016; Nahnsen et al., 2013]. These steps are highly automatised and happen almost simultaneously to the signal acquisition.

Peak picking requires automation for proteomic experiments, notably as the number of spectra for one sample is in the order of 100,000. Even though instrument resolution has significantly improved over the last decade, peaks can be smooth (*i.e.* with a large signal-to-noise ratio (SNR)) and easy to pick, or they can still be noisy (*i.e.* barely distinct from the background) and trickier to identify. Peak width is related to the peptidic ion mass-to-charge (m/z), the mass analyser resolution and acquisition parameters. Algorithms may use this relation to detect peaks by scanning the mass spectra for local maxima of expected widths. [Bauer et al., 2011]. They can also rely on the associated isotopic peak clusters of the molecular mass peak (see Appendix A.8). These molecular peaks clusters (and their relative intensity) also provide information that can resolve the atomic composition of the peptides [Renard et al., 2008].

Optionally, a final molecular mass correction step can be applied to remove modifications due to the ionisation process.

1.3.4.2 *Peptide identification and validation*

An essential step of proteomic data processing is to identify and then validate a peptide sequence for each molecular mass that has been detected. Typically in LC-MS/MS, spectra from a first mass spectrometer (MS1) allow the selection of the ionised peptides (*i.e.* precursor ions) to be fragmented, while the spectra from the second one (MS2) allow

[52] *ProteoWizard*: http://proteowizard.sourceforge.net/
[53] *Spectronaut* — https://biognosys.com/spectronaut
[54] See Section 1.3.3.5.

their identification [X. Wang et al., 2019a; Codrea et al., 2016]. Identification robustness increases with the number of spectra associated with each peptide (or protein).

Matching spectra to peptide sequences

In MS/MS experiments, various fragments (product ions) may be produced through the cleavage of covalent bonds or internal rearrangements (*e.g.* loss of carbonyl group, a water or ammonium molecule) [Macias et al., 2020; Wysocki et al., 2005; Yagüe et al., 2003; Bythell et al., 2010].

While different protocols and parameters can produce similar fragments, the peptide MS/MS profiles (spectra) recorded by a given instrument with different fragmentation modes are distinct. Only fragments carrying at least one charge can be detected and measured. The product ions' nature and relative abundances depend on the initial peptide sequence, and the chosen fragmentation method[55] and energy of the fragmentation event [Révész et al., 2021]. Different fragmentation methods may produce different sets of diagnostic fragments, and hence it can be beneficial to use different methods in combination [Révész et al., 2021; Dupree et al., 2020; Tu, J. Li, Shen, et al., 2016; Diedrich et al., 2013].

The two most popular fragmentation approaches are collision-induced dissociations (CID or HCD) and produce many fragments that result from the peptide backbone fragmentation, of which peptide bonds (see Section 1.1) represent the predominant breakage pathway as they have the lowest energy [Dupree et al., 2020; Medzihradszky et al., 2015].

The Roepstorff–Fohlman–Biemann nomenclature [Roepstorff et al., 1984; Biemann, 1988], presented in Figure 1.11, has been widely accepted to designate the product ions generated by the backbone fragmentation [Medzihradszky et al., 2015].

When a peptide bond (in purple in Figure 1.11) break occurs, the precursor ion generates two complementary fragments. The products are named as:

- b-ions, for the fragments comprising the amino (or N-terminus) end of the precursor's sequence,
- y-ions, for the ones comprising the carboxyl (or C-terminus) end.

Before possible internal rearrangements[56], the sum of the mass of complementary ions equals the molecular mass of their precursor ion.

Other fragment types also exist, and they assist with a better peptide characterisation[57] [Noor et al., 2020; Steen et al., 2004; Wysocki et al., 2005]. For instance, one may use EThcD to generate more informative fragments to study peptides' phosphorylation or ADP-ribosylation [Penkert, Yates, et al., 2017; Penkert, Hauser, et al., 2019; Bilan et al., 2017]. Furthermore, other protocols fragment the side chain (R group —

[55] See Section 1.3.3.4.

[56] *e.g.*, water loss for fragments with Ser, Thr, Glu, or Asp [Medzihradszky et al., 2015]

[57] *i.e.* peptide identification

see Appendix A.1) and produce satellite ions (d-, v- and w-ions) that can help to differentiate between isomers (*e.g.* Leu and Ile) [Han et al., 2007; R. S. Johnson et al., 1987]. However, the PTMs' characterisation remains challenging [Dupree et al., 2020].

The predictability of peptide fragmentation[58] [Dupree et al., 2020; Medzihradszky et al., 2015] and the aas mass differences (see Table A.1) enable (at least partially) the peptide's sequence resolution. Overlaps of partial sequences obtained from different ion types allow stitching them together in the peptide sequence.

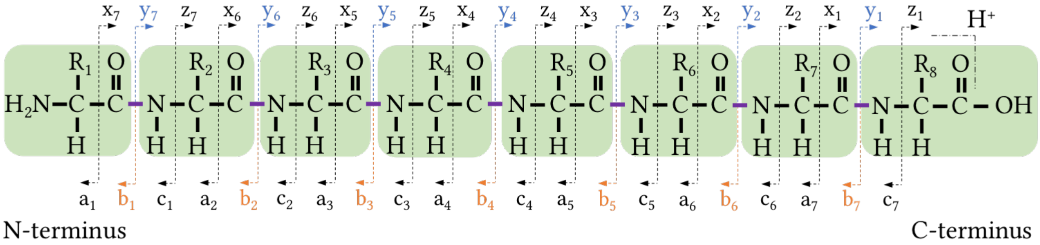

x7 y7 z7 x6 y6 z6 x5 y5 z5 x4 y4 z4 x3 y3 z3 x2 y2 z2 x1 y1 z1
H+
R1 O R2 O R3 O R4 O R5 O R6 O R7 O R8 O
H2N–C–C–N–C–C–N–C–C–N–C–C–N–C–C–N–C–C–N–C–C–N–C–C–OH
H H H H H H H H H H H H H H H
a1 b1 c1 a2 b2 c2 a3 b3 c3 a4 b4 c4 a5 b5 c5 a6 b6 c6 a7 b7 c7
N-terminus
C-terminus

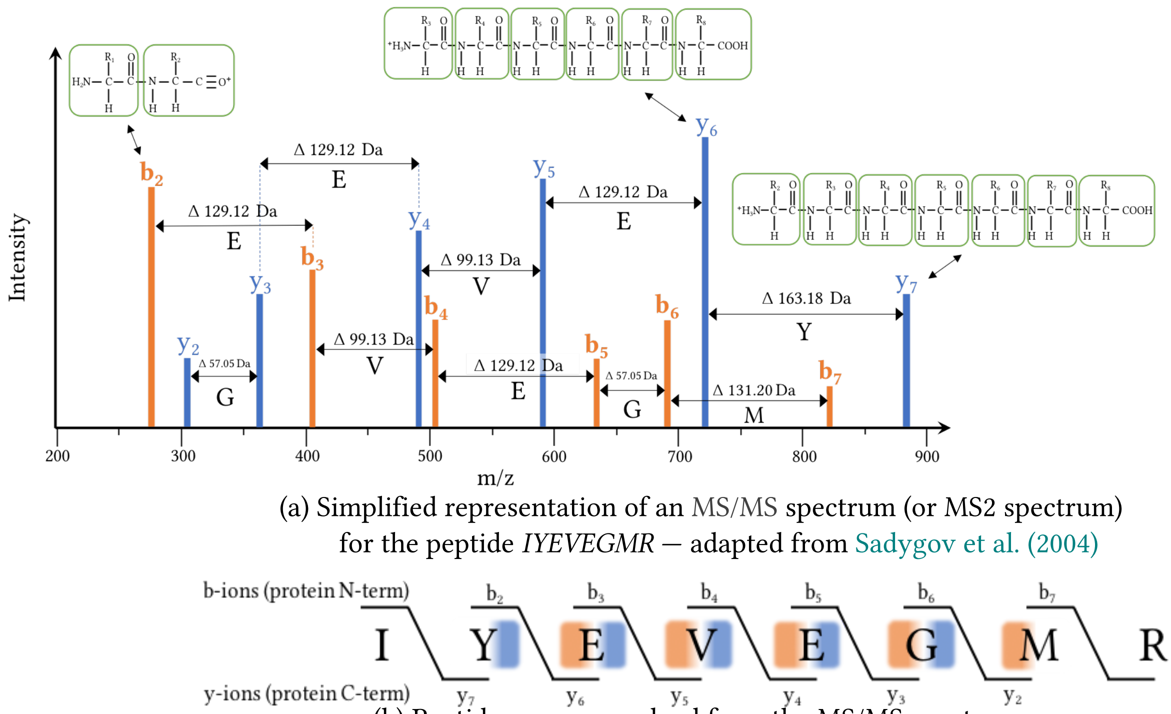
Figure 1.11. **The Roepstorff–Fohlman–Biemann nomenclature** [Roepstorff et al., 1984; Biemann, 1988] unambiguously designates the different ions generated from the peptide backbone fragmentation. Here is an illustration for a peptide formed by eight aas (one green box represents one aa or its residue — see Table A.1). The most common breakage occurring by collision is the cleavage of the peptide bond (amide bond, in purple), which can produce a y-ion (when the C-terminus part of the precursor ion has a charge) or a b-ion (the N-terminus part is charged). The peptide backbone fragmentation can generate other series of ion types, *e.g.* ETD and ECD can produce c- and z-ions (cleavage of the N-C$_\alpha$ bonds) [Han et al., 2007]. Like b-ions, a- and c-ions are produced when the N-terminal part of the peptide holds a charge, while, like the y-ions, x- and z-ions are created when the C-terminal part remains charged. An index notation distinguishes the ions from the same series: the index is the number of residues comprised (even partially) by the ions. When considering a couple of complementary ions, *e.g.*, b$_2$-ion and y$_6$-ion, the indices sum equals to the residues number of the precursor ion; in other terms, if n is the number of aas in the peptide, b$_m$-ion and y$_{(n-m)}$-ion are complementary.

Figure 1.12 shows an MS/MS spectrum, also called fragmentation spectrum or MS2 spectrum. When considering a series of either b- or y-ions, the mass difference between two (single charged) consecutive ions corresponds to the mass of the residue at the end[59] of the lengthiest (hence the heaviest) ion of the pair.

Although based on a real example, Figure 1.12 shows a simplified MS/MS spectrum to ease its interpretation. In practice, assigning peaks is more challenging, including assigning to one series over another one. The lower the mass accuracy, the more difficult it is to discriminate between possible isobaric combinations and different ion types [Medzihradszky et al., 2015].

[58] Both in the ions' nature and their relative intensities [Frank, 2009].

[59] either at the N-terminal end for a-,b- or c-ions or at the C-terminal end for the x-, y- and z-ions



(a) Simplified representation of an MS/MS spectrum (or MS2 spectrum) for the peptide *IYEVEGMR* — adapted from Sadygov et al. (2004)

(b) Peptide sequence solved from the MS/MS spectrum.

Figure 1.12. **MS/MS spectrum and peptide identification.** In (a), two consecutive ions of the same series have a mass difference that corresponds to the mass of one aa contained in one ion but missing in the other. The supplementary aa is found at the C-terminus of the heaviest ion. The respective *EVEGM* and *YEVEG* partial sequences can directly be solved from the b- and y-ion series. By combining the MS/MS spectrum presented in (a) and the precursor molecular mass given by the MS1 spectrum (997.16 Da — highlighted in red in (b)), one can deduce the remaining peptide sequence. Mass ions directly extracted from (a) are in orange for the b-ions and in blue for y-ions. Amino acids with supporting evidence in (a) are highlighted in (b) in orange when backed by b-ions and in blue by y-ions. The sum of the masses of complementary b- and y-ions equals the precursor's molecular mass. The complementary ion of y_7 has a mass of 114.16 Da, which corresponds to Leu's or Ile's acylium ion. Leu and Ile share the same mass and are undistinguishable from their b- or y-ions only. It may happen in the literature that a place holder is used to symbolise either of them, *e.g.* *J*, which refers to none of the standard aas. At the C-terminal end of the sequence, the complementary ion to b_7 has a mass of 175.20 Da that corresponds to Arg or the couple Val/Gly, an isobaric combination to Arg. However, the precursor peptide has been produced by trypsin digestion that specifically cleaves at the carboxyl side of Arg and Lys (when not followed by a Pro). Thus, it is more likely that the remaining sequence comprises Arg only. Theoretically, peptide sequences can be solved by collecting and identifying all the ions from the same series and correlating the mass differences between them with the residue masses of the amino acids — including any possible PTM or other modification. However, even when excluding Leu's and Ile's ambiguity, a direct resolution is hard to achieve in practice as series overlap and ions may not be detected or recognised (*e.g.* Ser and Glu can lose a water molecule through internal rearrangement [Medzihradszky et al., 2015]).

A complete ion series of a fragmented peptide is ideally required, but not all peptide bonds may break or be intense enough to be detected. In addition, water and ammonium losses and other modifications can shift the mass of the residues. However, there are also many rules of thumb (see Steen et al. (2004)) providing quick guidance about reasonable sequences. Furthermore, possible immonium ions ($+H_2N = CHR$) may help determine the aas composition of the precursor peptide, and past studies have highlighted many fragmentation rules that contribute to removing ambiguities. See Medzihradszky et al. (2015) for a more detailed discussion on fragmentation rules and MS/MS spectra interpretation refinements.

As modern LC-MS/MS experiments can produce tens of thousands of MS/MS spectra [O'Bryon et al., 2020], manual approaches are neglected in favour of algorithms to automate the peptide spectra matching.

There are three main possibilities for protein/peptide spectra matching: the sequence-based searching approach, the spectral library searching one and the de novo sequencing:

- The sequence-based approach matches the experimental spectra to theoretical ones. These theoretical spectra are created by adequate *in silico* digestion and fragmentation of proteins found in protein sequence databases such as UniProtKB/Swiss-Prot from *UniProt*[60] [The UniProt Consortium, 2017], which is manually annotated and reviewed, UniProtKB/TrEMBL (automatically annotated, unreviewed), or from genomic sequence databases (*e.g. NCBI*[61]) after *in silico translation* of the sequences. This type of matching requires database-dependent search engines, *e.g. Mascot*[62] [Perkins et al., 1999], *Sequest* [Eng, McCormack, et al., 1994; Tabb, 2015], *Andromeda* [Cox, Neuhauser, et al., 2011] from *MaxQuant*, *X!Tandem* [MacLean et al., 2006], *MS-GF+* [S. Kim et al., 2014], *MS Amanda* [Dorfer et al., 2014], *Open-pFind* [Chi et al., 2018], *MSFragger* [Kong et al., 2017], *Morpheus* [Wenger et al., 2013] or *Pulsar* included in Spectronaut®.
- The spectral library based approach matches the experimental spectra to other previously recorded experimental spectra. This approach relies on spectral matching engines, *e.g. HMMatch* [X. Wu et al., 2007], *SpectraST* [Lam et al., 2008] or *BiblioSpec* [Frewen et al., 2007] from *Skyline*, *M-Split* [J. Wang et al., 2010], *Pepitome* [Dasari et al., 2012], *QuickMod* [Ahrné, Nikitin, et al., 2011], *pMatch* [D. Ye et al., 2010], *COSS* [Shiferaw et al., 2020] or *ANN-SoLo* [Bittremieux et al., 2018].
- The *de novo sequencing* approach consists of comparing the observed mass data to the theoretical mass of every possible peptide sequence. This method is the closest to the manual approach described above. For example, see *PepNovo*[63] [Frank, 2009], *PEAKS*[64] [B. Ma et al., 2003], *pNovo* 3[65] [H. Yang et al., 2019], *DeepNovo* [N. H. Tran et

[60] *UniProt* — https://www.uniprot.org/
[61] *NCBI* — https://www.ncbi.nlm.nih.gov/protein/
[62] *Mascot* — http://www.matrixscience.com/
[63] *PepNovo* — http://proteomics.ucsd.edu/Software/PepNovo/
[64] *PEAKS* — https://www.bioinfor.com/peaks-studio/
[65] *pNovo* 3 — http://pfind.ict.ac.cn/software/pNovo/

al., 2017], *MetaSPS*[66] [Guthals et al., 2013], *Novor*[67] [B. Ma, 2015], *Lutefisk* [Taylor et al., 1997], *NovoHMM* [Fischer et al., 2005], *ANTILOPE* from *OpenMS*, *Twister* [Vyatkina et al., 2015] for topdown data, *UniNovo*[68] [Jeong et al., 2013], *UVNovo*[69] [Robotham et al., 2016].

The first two methods are favoured over *de novo* peptide sequencing, as the latter is very cumbersome and time-consuming [Codrea et al., 2016] and requires high quality data for best performance [Muth et al., 2018]. Note that there are also hybrid approaches based on *de novo* sequencing and database matching, such as *GutenTag* [Tabb, Saraf, et al., 2003], *InsPecT* [Tanner et al., 2005], *DirecTag* [Tabb, Z.-Q. Ma, et al., 2008], *ByOnic* [Bern et al., 2012], or *PEAKS DB* [J. Zhang et al., 2012].

Many search engines exist for matching MS/MS spectra, see Noor et al. (2020), C. Chen et al. (2020), Griss (2016), Shteynberg, A. I. Nesvizhskii, et al. (2013), and Eng, Searle, et al. (2011) for reviews. Furthermore, combining the results of several search algorithms when possible improves the final outcomes [Noor et al., 2020; C. Chen et al., 2020; Sadygov et al., 2004; Eng, Searle, et al., 2011; Shteynberg, A. I. Nesvizhskii, et al., 2013; Griss, 2016].

Sequence based algorithms are by far the most popular and many (*e.g. SEQUEST* and *Mascot*) rely on the same information and parameters choices to match (and score) the spectra to peptide sequences.

- Fragmentation mode
- Digestion enzyme and the number of possible omitted cleavages
- Mass tolerance for m/z ratio of peptidic and fragmented ions.
- Possible charges for the peptidic and fragmented ions.
- Knowledge database (the more complete, accurate and adequate a database is the better and more robust is the peptide and protein identification).

Scoring functions

While algorithms assign MS/MS spectra to peptide sequences, they simultaneously compute a corresponding score for each peptide-spectrum match (PSM). This score summarises the quality of the assignment and allows the selection of the best candidate-peptide for each spectrum. Only the matches with the best PSM scores (if not the very best one only) are reported and used in the next steps of the quantification.

Sadygov et al. (2004) classify the scoring functions in four classes: descriptive (*e.g. SEQUEST*[70] [Eng, McCormack, et al., 1994]) interpretative, stochastic and probability-based modelling (*e.g. Mascot*). While many of these functions return statistical scores (*e.g. Mascot* and *SEQUEST*), many other return non-statistical ones. For

[66] *MetaSPS* — http://proteomics.ucsd.edu/software-tools/metasps/
[67] *Novor* — rapidnovor.com
[68] *UniNovo* — http://proteomics.ucsd.edu/Software/UniNovo
[69] *UVNovo* — https://github.com/marcottelab/UVnovo
[70] *SEQUEST* — http://fields.scripps.edu/yates/wp/

these latter ones, tools such as *Percolator*[71] [Käll, Canterbury, et al., 2007; Spivak, Weston, Bottou, et al., 2009] or *PeptideProphet*[72] [Keller et al., 2002; K. Ma et al., 2012] allow transforming their scores into probabilities, and ease the application of threshold to remove unreliable matches [J. S. Cottrell, 2011]. Regardless of whether the score is either statistical (including probability-based) or not, searching the database to match spectra to peptide sequence is a statistical process. Most MS/MS spectra only partially cover a peptide sequence, which leads to many ambiguities. With the development of high-throughput MS-based proteomics, researchers dropped manual interpretation and validation and moved towards empirical score thresholds [Brosh, 2009]. Thresholds are a compromise between sensitivity and accuracy (or error rate), *i.e.* true positive (or correct) identifications proportion *versus* false positive (or incorrect) identifications proportion. A high threshold for the scores reduces the error rate, but decreases sensitivity. A low threshold accepts more PSMs, but also more incorrect matches.

Peptide validation

Today, among the many methods that validate the peptide assignment and estimate its error rate, the gold standard is the target-decoy search approach (TDA) [Perkins et al., 1999; Elias et al., 2007; Savitski et al., 2015].

Many search engines compute a p-value (see Appendix A.9.2) for each of the PSMs, but it is insufficient to determine if multiple PSMs are true matches. A low p-value PSM has a low probability of being incorrect. Because of the large number of PSMs per experiment, statistically, a portion of these *are* incorrect. Thus, other statistical measures adjusting for multi-testing (see Appendix A.9.3) are required. Corrections (such as Bonferroni [Shaffer, 1995]) can be applied though they are stringent and discard many correct PSMs. False discovery rate (FDR) [Benjamini and Hochberg (1995)] estimates the proportion of incorrect PSMs among all accepted PSMs [A. I. Nesvizhskii, 2010; Aggarwal et al., 2015], see Equation (FDR). Different approaches exist to estimate it. The most favoured is TDA as it is non-parametric, easy to apply, and it also has the advantage of working with search engines that have non-statistical scores.

Instead of figuring out which the correct and incorrect PSMs are, the target decoy search approach (TDA) aims to estimate the overall FDR associated with a specific collection of PSMs. In turn, this enables assessing the likelihood of each of the PSMs in the collection [Elias et al., 2007] (see following segments on *q-values* and PEP). The critical elements of this method are the creation of a *decoy* database with sequences that are incorrect but similar (while non-overlapping) to the *target* (*i.e.* true) ones. Consequently, any PSM found for a decoy sequence is by definition spurious. The known proportion of decoy versus target sequences in the search space allows the computation of the FDR [Elias et al., 2007; Elias et al., 2010] as shown in Equation (FDR). In Appendix A.10, I briefly discuss the decoy database.

[71] *Percolator* — http://percolator.ms/
[72] *PeptideProphet* — http://peptideprophet.sourceforge.net/

$$\begin{aligned}\text{FDR} &= \frac{\text{Number of accepted PSMs}_{\text{decoy}}}{\text{Number of all accepted PSMs}} \\ &= \frac{\text{Number of accepted PSMs}_{\text{decoy}}}{\text{Number of accepted PSMs}_{\text{decoy}} + \text{Number of accepted PSMs}_{\text{target}}}\end{aligned} \tag{FDR}$$

Most of the search engines return multiple scores. Hence, defining a proper threshold for each of them can be challenging or cumbersome. A possible solution is *Percolator*, which trains a support vector machine (SVM) [Boser et al., 1992] to distinguish the correct and incorrect PSMs [Käll, Canterbury, et al., 2007]. This machine learning based algorithm has several advantages. It can exploit many scores and other specific data features to automatically determine the best threshold without overfitting to a particular collection of PSMs. Thus, comparisons of results between studies and laboratories are facilitated. Moreover, many studies have reported that the use of *Percolator* improves the results in terms of both accuracy and sensitivity and increases the overall number of identified peptides [Granholm et al., 2014; Xu et al., 2013; Tu, Sheng, et al., 2015; The, MacCoss, et al., 2016; Wright, Collins, et al., 2012]. *Percolator* can either be used directly on the collection of target and decoy PSMs or as a post-processing step.

Protein inferring algorithms use two other significance measures for PSM validation: q-values and PEP, which are described in Appendix A.11.

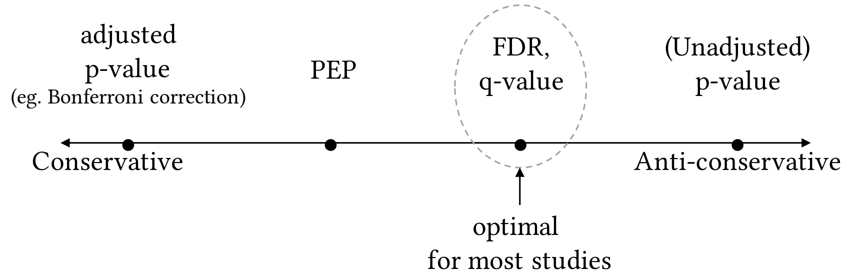


Figure 1.13. **Methods for assigning statistical significance to a collection of PSMs** — Adapted from [Käll, Storey, et al., 2008].

Methods based on PEP instead of q-values are more conservative, as PEPs are always greater than q-values [Käll, Storey, et al., 2008], see Figure 1.13. For an individual validation, *e.g.* checking the presence of a specific peptide in a particular condition, PEPs are more indicative. On the other hand, q-values are favoured where the whole collection of PSMs is considered, *e.g.* for an overview of the proteome landscape. [Käll, Storey, et al., 2008]

1.3.4.3 *Protein inference*

While identified as the key problem more than fifteen years ago, protein inference (*i.e.* identification) remains the main challenging issue in shotgun proteomics [A. I. Nesvizhskii and Aebersold, 2005; He et al., 2016]. Note that in the literature, *protein inference* often also encompasses the peptide identification as an intermediate step and the results validation as they influence the protein identification quality.

Protein inference consists in assembling peptides into sets of reliable proteins. Most assembly algorithms model the relationship between identified peptides and protein sequences as a bipartite graph [T. Huang et al., 2012], illustrated in Figure 1.14. Proteins identified by two or more (unique) peptides represent the best case: their identification is reliable and computationally lighter. However, the existence of 'degenerate peptides' and 'one-hit wonder' proteins [T. Huang et al., 2012; He et al., 2016] makes this task challenging and computationally intensive.

Degenerate peptides are ambiguous peptides that are shared by multiple protein sequence definitions. They create a computational challenge because it is difficult to resolve from which proteins they are derived and to select which of the two following options is true: either all the related proteins are expressed in the sample or only some of them. Some workflows cluster together proteins with homologous sequences to ease the process. More often, to lessen the computational and interpretation burden, these ambiguous peptides are discarded, and the inference relies solely on 'unique peptides', *i.e.* peptides that are attributable to one protein sequence only.

'One-hit wonders' are proteins that are identified by a single peptide only. They require careful handling regardless of their peptide uniqueness status, since if the peptide identification is a false positive (*i.e.* an artefact), then the protein is also a false positive. Shorter proteins are generally harder to identify (and quantify) for this reason.

As shown in Appendix A.12, the peptide assembly can be formulated as a *set covering problem* [Cormen et al., 2009; Hochbaum, 1997]. This problem is known to be NP-complete [van Leeuwen et al., 1990], and for which it is in practice impossible to calculate an optimal solution.

Inference algorithms seek a compromise between the minimal and exhaustive lists of possible proteins. Usually, algorithms approximate this solution through a parsimonious approach (see the below example based on *Maxquant*).

Regardless of the approach, all algorithms involve a bipartite graph (see Figure 1.14), even if they may also include other supplementary information to build their models (see Figure A.5, p. 195).

Figure 1.14 shows how degenerate peptides, one-hit wonder proteins and the validation quality of the peptide identification complicate the inference.

Note that if an edge connects a peptide i and a protein j, the peptide i is said to be covered by the protein j [T. Huang et al., 2012].

Protein 1 and *protein 2* share *peptide 1* and *peptide 2*. As both *protein 1* and *protein 2* are also covering another peptide (*peptide 3* for *protein 1* and *peptide 4* for *protein 2*), it seems

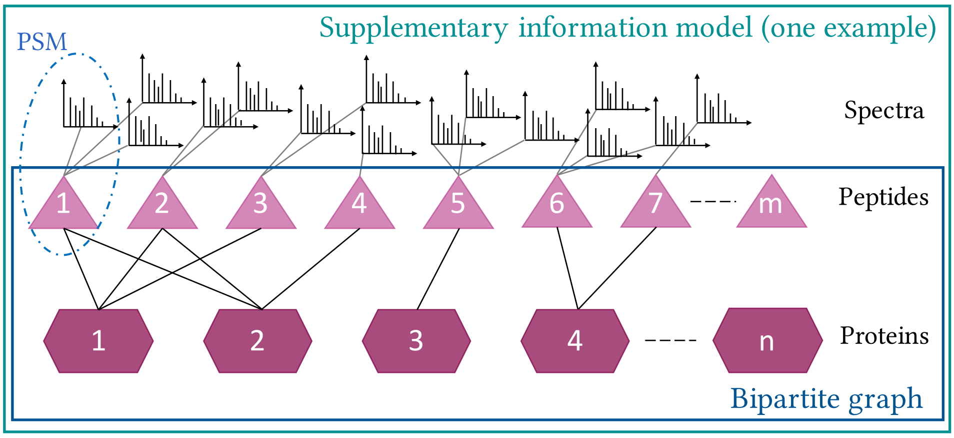


Figure 1.14. **Protein inference: the bipartite graph.** In order to infer proteins, algorithms attribute each validated peptide to possible proteins of origins. *Peptides 1 and 2* are both included in the definition of *proteins 1 and 2*; it is impossible to determine if both proteins are present in the sample or not based on these two peptides only. *Peptide 3* backs the existence of *protein 1* and *Peptide 4* backs the existence of *protein 2*; if either *peptide 3* or *peptide 4* is missed in detection, it is easy to conclude that only one of *proteins 1 and 2* is only present. On the other hand, *protein 3* is only identified by *peptide 5*. If the latter is an artefact, then *protein 3* is also an artefact. In order to achieve the inference, peptide assembly algorithms can rely on the bipartite graph as sole input (solid blue box), or they can include other data types as well, *e.g.* the score associated to each PSM (dashed blue circle) or the raw spectra itself (solid teal box).

reasonable to assume that both proteins are expressed without more information. If *peptide 4* is actually a false positive, would it mean then that only *protein 1* is expressed?

Now, what happens if *peptide 4* is hard to detect and is missing from the validate list of peptides? *Peptide 5* is the only one to identify *protein 3* while *peptide 6* and *peptide 7* are both backing the existence of *protein 4*. If one of the two latter peptides is an artefact, *protein 4* is still most likely a true positive. However, if *peptide 5* is a false positive, so is *protein 3*.

Many different approaches, algorithms and their related search engines for protein inference have been reviewed in the literature [T. Huang et al., 2012; Serang and Noble, 2012]. Despite the partial or lack of overlap between sets of confidently identified peptides, combining several search engines to infer the proteins has proven to yield better results than a single search [Searle et al., 2008]. At worst, it improves the confidence of the identification as more peptides are characterised per protein [T. Huang et al., 2012; Audain et al., 2017].

One possible approach for protein inference is a parsimonious approach, as implemented for example by *Maxquant* [Tyanova, Temu, and Cox, 2016; Cox and Mann, 2008].

After identifying all proteins covering a given peptide, *Maxquant* joins the proteins with the same set or subset of peptides in the same protein group. In other words, if the peptides set S_a, defining a protein P_a, is equal to or strictly included in the peptides set S_b, defining the protein P_b, then P_a and P_b are joined in the same group G_1. Then, in each group, the proteins are ordered by the decreasing number of peptides they cover. Hence, the protein sequence at the top of the group can explain all the group's peptides.

Maxquant refers to a peptide as 'unique' when found in only one group of proteins in contrast to degenerated peptides shared between two distinct protein groups that cannot be combined because their other peptides are unique to each group. Shared peptides are called 'razor' (referencing Occam's razor) in the protein group with the highest number of peptides since it is considered the simplest explanation.

The user can choose which peptides are included in the quantification: unique peptides only, both unique and razor peptides (default option[73]) or all the peptides (of each group). The following steps enable results filtering based on the protein groups PEPs (a protein group PEP equals the multiplication of its peptide PEPs), the spectra quality, the unique peptides number and the FDR threshold.

Many tools implement other approaches.

DTASelect [Tabb, McDonald, et al., 2002] sorts peptides by their identified locus (*i.e.* gene or protein identifier) and then by sequences. Next, *DTASelect* filters the results based on various criteria, including the spectra quality and user's inputs, to finally keep proteins supported by enough different peptides or by at least one peptide identified several times. The algorithm adopts an optimistic approach instead of a parsimonious one and only groups together proteins that have a strict identical sequence coverage.

ProteinProphet [A. I. Nesvizhskii, Keller, et al., 2003] (part of the *Trans-Proteomics Pipeline*) is one of the most widely used methods in the literature [Sikdar et al., 2016]. The tool computes in each sample the presence probability of a protein by combining the probabilities of its different identified peptides through an iterative process. An EM algorithm derives a mixture model of correct and incorrect peptide identifications from the observed data[74]. The following steps can summarise the inference[75]. *ProteinProphet* keeps the best spectrum matching any given assigned peptide. It retrieves all proteins that cover any identified peptide. The tool groups the peptides by proteins and computes the protein's presence probability based on the available peptide evidence. It readjusts the peptides' negative and positive distributions based on their sibling counts: proteins that have more than one assigned peptide are rewarded, while 'one-hit wonders' are penalised. *ProteinProphet* apportions the degenerate peptides across all their covering proteins before using a parsimonious approach to readjust their distributions through an EM algorithm. As a starting point, the sum of all the peptides' weights of one protein

[73] Considered by the authors as the best compromise between most accurate protein quantification and unequivocal peptide assignment [Cox and Mann, 2008].

[74] including the PSM results, their associated score or the properties of the peptide matched to the spectrum

[75] Based on A. Nesvizhskii (2006) and Serang and Noble (2012)

equals one. Through successive iterations, these weights are refined such that proteins with a higher number of peptides are rewarded. All redundant and indistinguishable protein entries are finally collapsed together.

The model learns directly from the observed data, which increases its robustness [T. Huang et al., 2012]. However, one identified issue with this award/penalty system is that it creates cases where proteins covering many low-scoring peptides can outrank proteins covering a smaller set of higher-scoring peptides [Serang and Noble, 2012]. As the size of the dataset to study grows, the problem worsens. A complimentary tool in the *Trans-Proteomics Pipeline* developed by the authors, *iProphet* [Shteynberg, Deutsch, et al., 2011], addresses this issue by considering other information levels and refining the computation of the posterior probabilities and protein FDR estimates.

Percolator[76] [Käll, Canterbury, et al., 2007; The, MacCoss, et al., 2016; Halloran et al., 2019] (part of both *The OpenMS proteomics pipeline (TOPP)* and the *Crux suite*[77]) is based on SVMs. It implements a generalised semi-supervised[78] learning approach distinguishing between target and decoy[79] PSMs for any shotgun dataset. First, a classifier is trained with a subset of the data, *i.e.* the algorithm knows the data labels. Second, after ranking targets and decoys according to a selection of features, the algorithm selects the target PSMs with a 1% FDR and then trains an SVM to discriminate between the kept targets and the full set of decoy PSMs, and induces a new ranking. These steps iterate until convergence of the ranking (*i.e.* the ranking remains the same from one iteration to the next).

Fido[80] [Serang, MacCoss, et al., 2010] (implemented first as a standalone tool, but now distributed as a part of *Percolator*) is based on Bayesian inference (see Appendix A.13). From a set of simple assumptions, the authors have developed a Bayesian model. This model estimates three parameters directly from the data: the probability of a present protein to generate peptides, the error probability of peptides to be detected from noise and the proteins' a priori probabilities to be present in a sample. These parameters are respectively annotated α, β and γ. *Fido* explicitly allows high ranking spurious PSMs. It uses their presence likelihood[81] and rewards proteins which include strong independent supporting evidence besides their degenerate peptides. *Fido* automatically apportions information from degenerate peptides while ensuring that each protein presence likelihood is unrestricted to their degenerate peptides. Compared to the initial approach of *ProteinProphet*, this method's accuracy is resistant to the dataset size. A related method to *Fido* is *EPIPHANY* [Pfeuffer, Sachsenberg, Dijkstra, et al., 2020].

[76] *Percolator* — http://percolator.ms

[77] *Crux suite* — http://crux.ms/

[78] *semi*-supervised since only decoy PSMs are labelled as 'incorrect' while target PSMs are an unlabelled mixture of correct and incorrect PSMs. [Halloran et al., 2019]

[79] shuffled or reverse peptide sequences

[80] *Fido* — https://noble.gs.washington.edu/proj/fido/

[81] By converting the PSMs prior probability estimates back into discriminant score-based likelihoods when needed

Another inference tool from *Percolator*'s authors is *Barista* [Spivak, Weston, Tomazela, et al., 2012], which is also part of the *Crux suite*. *Barista* combines the verification of the PSMs and the protein inference where *Percolator* handles only the second task. The authors advocate for a topdown approach to optimise the protein inference problem instead of subdividing the workflow into independent modules. Their tool builds a tripartite graph (spectra-peptide-protein) based on the results of a database search (target and decoy sequences) and a protein database. A learning algorithm infers the protein presence likelihood in each sample. This algorithm involves a similar set of features to *Percolator*. *Barista* iteratively refines the peptide identification ranks to better discriminate between correct and incorrect identification. In contrast to *Percolator*, the PSMs are not filtered in *Barista* and contribute to the results optimisations. The authors themselves state that assessing which approach is the best in practice can require more study [McIlwain et al., 2014].

Many other inference algorithms exist, including *PIA* [Uszkoreit et al., 2015], *MIPGEM* [Gerster et al., 2010], *PAnalyzer* [Prieto et al., 2012], *DBParser* [X. Yang et al., 2004] and *PANORAMICS* [Feng et al., 2007]. The field is steadily improving and adapting in response to the development of the measurement instruments and acquisition protocols (*e.g.* *IPF* [Rosenberger et al., 2017] for DIA data), preparation protocols (*e.g.* multiple proteolytic digestions in parallel, see [Miller et al., 2019]), increase of computing resources (for example *Percolator* optimisations [Halloran et al., 2019]), optimisation or implementation of new mathematical approaches and concepts (*e.g.* *DeepPep* [M. Kim et al., 2017] based on a deep-convolutional neural network framework or *gpGrouper* [Saltzman et al., 2018] and *MIPGEM* [Gerster et al., 2010] that implement a gene centric approach). Furthermore, the rise of metaproteomic, proteogenomic, metabolomic and multiomic studies has brought new perspectives and challenges [Gonnelli et al., 2015; Starr et al., 2018; Rechenberger et al., 2019; Menschaert et al., 2017; Liebal et al., 2020].

The, Edfors, et al. (2018) have designed an experiment simulating homologous proteins that can evaluate most protein inference algorithms. They observe better concordance between inferred proteins and the ground truth when excluding degenerate peptides rather than adopting a parsimonious approach. However, they report that results are different when considering protein groups instead of individually.

1.3.4.4 *Protein quantification (label-free)*

Label-free quantification methods will probably be more refined in the future. Over the last decade, along with the instruments' improvement and the increasing number of label-free bottom-up proteomics, there have already been many developments in quantification methods and tools. Blein-Nicolas et al. (2016), Y. Chen et al. (2016), and Lindemann et al. (2017) review many of the currently available ones. Note that methods created for label-free relative quantification experiment designs can be adapted to methods that allow absolute quantification [Pappireddi et al., 2019; Sinitcyn et al., 2018; Y. Chen et al., 2016]. Many normalisation methods have also been reviewed by Välikangas et al. (2018b).

As mentioned above, label-free quantification methods are either based on the number of identified peptides or MS2 (i.e. MS/MS) spectra matched to a protein or on the precursors' ion extracted current peak intensities (*i.e.* XIC) from the MS1 spectra.

Spectral/peptide counting is a widespread method in the literature. H. Liu, Sadygov, et al. (2004) present a linear correlation over two orders of magnitude between the relative protein abundance and the acquired spectra number. Spectral counting is simple but requires proper normalisation. Many other methods (*e.g.* *APEX* [Braisted et al., 2008], *emPAI* [Ishihama et al., 2005] or *MAI/PLGEM-STN/SC* [H. Y. Lee et al., 2019]) are derived from it. Cozzolino et al. (2020) have realised a comparative study between their method and two other spectral counting ones.

There are also peak intensity based methods such as intensity based absolute quantification (IBAQ), which are the most favoured today. Arike et al. (2012) report that this latter method is better than the previous ones based on spectral counting as the estimated quantification correlates better to the absolute abundance. Ahrné, Molzahn, et al. (2013) refine this statement as IBAQ has biases and quantification errors that make it unfit for direct use for a proportional assessment of the complete protein set; IBAQ dramatically underestimates low-abundant proteins. Instead, to improve the general quantification estimation, the authors advocate a *Top3* approach [Silva et al., 2006], which represents the abundance of each protein by the average or sum intensity of its three best ionised (unique) peptides. *Top3* allows better quantification of the proteome landscape in general, even if it is less accurate for the shortest proteins and saturates for the largest ones. MS1-based quantifications have been reviewed by X. Wang et al. (2019b).

Many normalisations include the protein sequences length (longer proteins are more likely to produce a greater number of sampled peptides) and the sum of ion intensity or the spectra number (for spectral count methods) that acquired for each protein within a sample [Blein-Nicolas et al., 2016]. Similar approaches to RNA-Seq can be used to allow protein comparison across different samples.

Most analyses benefit from validating the inferred proteins. Unfortunately, there is a lack of consensual methodology to determine a protein FDR and strong opinion divergences on its conceptual validity in the field [Savitski et al., 2015]. A few are even questioning its validity more generally [J. Cottrell, 2013]. Beyond the differences due to frequentist and Bayesian definitions and approaches, part of the pointed out discrepancies [The, Tasnim, et al., 2016] is caused by overlooking that protein inference tools are simultaneously using two different null hypotheses ($\mathcal{H}_0$ – see Appendix A.9.1). Two common $\mathcal{H}_0$ statements testing the validation of a protein identification are:

$\mathcal{H}'_0$ The best scoring peptide is incorrectly matched to the protein. (Often, the protein FDR is derived from its best scoring peptide.)

$\mathcal{H}''_0$ The protein is absent from the sample.

Nonetheless, the most widespread method is based on the TDA and the protein FDR is computed with Equation (FDR) (p. 44). Savitski et al. (2015) demonstrate that the *classic* TDA largely overestimates the FDR for large datasets. To overcome this, Savitski et al. (2015) propose a new '*picked*' TDA. This approach pairs together target and decoy sequences of each protein instead of treating them individually. For each pair, the target's and the decoy's protein scores are compared, and the highest is kept and the other one discarded. The, Tasnim, et al. (2016) encourage the use of the '*picked*' TDA when one's analysis is based on $\mathcal{H}'_0$, but recommend to keep the classical TDA for $\mathcal{H}''_0$ as there is a lack of a better method to date (and '*picked*' TDA actually underestimates FDR in that case).

Several benchmarking studies [Välikangas et al., 2018a; Al Shweiki et al., 2017; Bubis et al., 2017; Navarro et al., 2016] have been published recently that may help choose between the many existing quantification methods and the tools implementing them.

1.4 POSSIBLE DOWNSTREAM ANALYSES FOR EXPRESSION DATA

Over-representation analyses (ORAs) are a standard final stage analysis on expression data. They can provide biological insights and hint about mechanisms. ORAs highlight sets of gene (or protein or metabolite) categories that are overrepresented in a selected subset of the data compared to the expectation of a random category. Many expression studies aim to produce 'a list of "interesting" biomolecules' [Tipney et al., 2010], ORAs help to determine the most pertinent ones by providing biological context and increasing statistical power. Besides, such lists are often substantial so extracting common functional information from subsets of genes/proteins eases the interpretation of the data. Various tools and algorithms have been developed to overcome the daunting task of individually checking the genes/proteins. See Shi Jing et al. (2015) for some examples. While, the field has been reviewed a few times (see Khatri and Drăghici (2005), D. W. Huang et al. (2009), and Khatri, Sirota, et al. (2012)), it is still unfortunately lacking a gold standard and systematic comparative studies [Mathur et al., 2018].

Depending on the experimental design and study, the list of biomolecules can be associated with a rank or another form of a score. As differential expression analyses (DEAs) are the most widespread type of analyses for expression data, many tools and algorithms work solely with the corresponding outputs. Hence, those are unfit for other types of studies (such as the ones in this thesis). Available gene set enrichment analysis (GSEA) [Subramanian et al., 2005] tools are commonly inadequate for any other purpose than DEA studies. See Tamayo et al. (2012) and Irizarry, C. Wang, et al. (2009) and the included references for some examples of GSEA tools.

On the other hand, while still devised for DEA, other tools handle any rank or score and thus are more flexible; they can analyse data outside of their original scope. Many gene ontology (GO) analysis tools fall into this latter category.

1.4.1 *GO analysis (GOA)*

The GO is a collaborative and curated classification that describes the gene products following three hierarchically structured and controlled vocabularies (also known as ontologies): either based on the biological processes ('BP') to which they contribute, their position (when active) in the cell (cellular component ('CC')) or their biochemical activity, *i.e.* molecular function ('MF'). [Ashburner et al., 2000]

Generally, GO enrichment analyses (GOAs) compare a selected list (with the genes/proteins of interest) to a background list (*e.g.* all observed genes in the experiment or all existing genes in the annotation). For each GO term, this method computes the enrichment of the selected set based on the real fraction of the set for the considered GO term and its likelihood (computed on the background list). For example, if a GO term $\mathcal{A}$ is associated with 0.1% of all the background (list) genes, but then over 70% of the genes from the selected list are associated with GO term $\mathcal{A}$, one can safely accept that the selected genes list is enriched for this term. Various tools rely on different statistical tests to determine if the enrichments are significant. Investigating only a few GO terms of interest is common.

While ranked lists are unnecessary for GOAs, one can apply a cut-off before running this type of analysis. However, in those cases, GSEA will be generally favoured to a GOA.

1.5 REPRODUCIBILITY AND EXPERIMENTAL DESIGN

Science develops based on reproducible facts, as they help with drawing relevant and accurate conclusions. To increase reproducibility, it is essential that observations and measurements be (as much as possible) unbiased towards any parameter outside of the study focus. To this end, one of the critical issues that need to be tightly monitored is the presence of *batch effects*. Including adequately designed *replicates* in the study is the most effective way to control for the unwanted batch effects.

Another issue is due to high-throughput transcriptomics and proteomics still being evolving fields. For each new identified problem, researchers create new tools and algorithms. However, these are often aimed at one study only and their reuse in another study can be difficult, or even impossible (*e.g.* discontinued proprietary software). On the other hand, well established tools allow fine tuning many parameters, the impact of which can be overlooked while the reporting. Thus, it is unsurprising that result agreement between different tools is unsatisfactory [Conesa et al., 2016].

1.5.1 *Batch effects*

Batch effects are artefactual and due to all the variables that the investigator can not control (either by lack of technology, design or knowledge), for example, environmental conditions, reagent or sample lots, genetic population background or experimenters. They are often the source of complication for many studies (including high-throughput genomic ones) [Leek et al., 2010].

The danger lies in overlooking them and then confusing these artefacts with biological results, which will lead to flawed interpretation and conclusions. Several studies have been refuted in the past because of unaccounted batch effects. Usually, issues in the initial results are detected by others laboratories, which notice high correlations between the 'biological' findings of the original study and the running dates or processing groups. Thus, questions about the biological validity arise.

The first step to address them is through well-designed experiments [Leek et al., 2010], which include technical and biological replicates. These replicates are usually created and randomly processed as to avoid creating any artefactual link between them. A replicate is a set of measurements done in the same condition.

Correction can be applied and may help resolve batch effects in some cases. For examples, see Oytam et al. (2016), Gagnon-Bartsch et al. (2012), and Peixoto et al. (2015).

1.5.2 *Technical replicates*

Technical replicates are initially from the same sample, which has been tested multiple times through a given experimental protocol. It allows testing for the variability of the protocol itself. While using the same sample to test different protocols may also be referred to as 'technical', it is better to avoid this terminology as this creates confusion.

1.5.3 *Biological replicates*

Biological replicates are testing the same cells or tissues from different individuals through the same protocol. They allow assessing the biological variability (which is higher than the technical variability, since it also encompasses it).

1.5.4 *Study design example: meta-analyses*

Meta-analyses are studies that combine (or aggregate) the results of multiple analyses. One of their main weaknesses is that meta-analyses cannot control bias sources and correct for bad designs [Slavin, 1986]. Because results from smaller studies are more prone to 'play of chance', they are usually weighted less than bigger studies when they are directly combined across many studies [Egger et al., 1997], particularly when combining statistical results. However, it is not always true and bigger studies may be subject to greater uncontrolled variations. Egger et al. (1997) recommend testing the heterogeneity across the studies to be combined. Examining the studies outcomes' similarity degree allows figuring out if the variation between the studies is only due to sampling or to a distribution of different effects. In order to compare studies together, individual results are expressed in a standardised fashion (*e.g.* means, confidence interval).

1.6 DISCUSSION AND CONCLUSION

Over the years, the central dogma of molecular biology has been being refined and better understood. However, as the first apparent linearity of the theory tends to persist, so are many assumptions. For example, all (or almost all) information necessary to explain the phenotype is in the cell; in its genome and edits provided by its transcription and translation regulatory mechanisms that may be triggered by environmental stimuli. Another assumption is that mRNA and protein levels should share strong correlates. Alternatively, as our genomes are so similar, our transcriptomes and proteomes should also share many similarities. Although the truth seems more intricate, these assumptions remain as we still lack the technical means to test them to their fullest.

Due to the intrinsic nature of DNA, mRNAs and proteins, high-throughput DNA and RNA studies are more well-established and standardised than proteomic studies. Although, the study of the genome is the most mature, it fails to contextualise the phenotype, especially for non-disease cases (often referred to as healthy or normal conditions in this thesis).

Proteins are in theory the best candidates to study the phenotype. Unfortunately, high-throughput protein studies such as shotgun MS are particularly challenging. The proteins physicochemical diversity and the lack of technology to amplify proteomes mean that in order to optimise their studies many different protocols and experimental designs had to be developed to reach the different proteins in a sample.

Since the transcriptome has a dynamic dimension like the proteome and is technically easier to explore and quantify, its study has emerged as a reasonable trade-off strategy.

Data! Data! Data! I can't make bricks without clay!
Sherlock Homes [Doyle, 1892]

2 AVAILABLE HIGH-THROUGHPUT HUMAN DATASETS

In the past few years, many laboratories have studied the expression of human genes at the transcriptome and at the proteome levels by taking advantage of high-throughput techniques (*e.g.* Krupp et al. (2012), Brawand et al. (2011), Ramsköld et al. (2009), Fagerberg et al. (2014), Uhlén, Fagerberg, et al. (2015), Gremel et al. (2015), Melé et al. (2015), Desiere et al. (2006), M.-S. Kim et al. (2014), and Wilhelm et al. (2014)). In this chapter, I review the openly available (for research) data I use within my thesis to explore the gene expression in (undiseased) tissues and explain how they have been reprocessed. Besides the results I report in the subsequent chapters, the present chapter also provides the basis for work I have co-authored[1].

Unless otherwise stated, all the computational processing of the RNA-Seq part described here have been performed by myself under the supervision of Dr Alvis Brazma. I also received general feedback from Dr Mar Gonzàlez-Porta, Dr Johan Rung and Dr Nuno Fonseca. The proteome data has been processed by Dr James Wright.

2.1 INTRODUCTION

All the datasets were selected to fit three main criteria. Firstly, they comprise normal (*i.e.* reported as disease-free) human samples from at least three different tissue types. Secondly, gene expression quantifications are based on RNA-Seq for the transcriptome and on label-free MS for the proteome[2]. Finally, the *raw* data is openly available and reusable[3].

In the next section, I first describe the RNA-Seq and the MS data I use in my thesis; then I detail how these data have been processed to be employed in the next chapters for various analyses that explore the transcriptome, the proteome and finally, the comparison and integration of these two biological layers.

[1] J. C. Wright, J. Mudge, et al. (2016). 'Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow'. *Nat. Commun.* 7, p. 11778.

[2] These technologies are non-targeted high-throughput and allow one, in theory, to study the whole repertoire of RNAs or proteins in a sample.

[3] In this context, *reusable* means that the data can be processed as accurately by third-party researchers than the original authors, and this without the need to access additional information that have been not openly released

2.2 TRANSCRIPTOME RNA-SEQ STUDIES

I describe hereinafter the five transcriptomic datasets I used in the chronological order of their first public release. Table 2.1 summarises the main characteristics of these datasets.

2.2.1 *Castle et al. dataset*

Castle et al. (2010) released this dataset along with their study: *'Digital Genome-Wide ncRNA Expression, Including SnoRNAs, across 11 Human Tissues Using PolyA-Neutral Amplification'*. The authors were interested in exploring the whole RNA repertoire with sequencing-based technology and they primarily focused their study on the non-coding part.

Purchased RNA extracts were used to create multiple-donors pooled samples for 11 tissues from which total RNA libraries were prepared following a total transcriptomic protocol where nonribosomal RNA transcripts are amplified specifically by PCR [Armour et al., 2009].

For each library (tissue), an average of 50 million sequence reads were sequenced using an Illumina Genome Analyser-II sequencer (single-end). The original reads were trimmed to 28 nt before being released through EMBL archives (ENA ID: ERP000257 and ArrayExpress ID: E-MTAB-305).

Despite several limitations, such as the lack of replicates, the old technology and the short reads, I have included this dataset for two main reasons. Firstly, it is the oldest available RNA-Seq data I found that was performed on normal human tissues. Thus, the congruence of results for this dataset with the following ones gives a rough idea on the extent of RNA-Seq datasets that may be integrated together. Secondly, as RNA-Seq studies are prepared mainly with polyA-selected protocols today, I was interested in gauging how the library preparation protocols — and the presence of ncRNAs — can affect the quantifications and then any final observation.

2.2.2 *Brawand et al. dataset*

In the article entitled *'The evolution of gene expression levels in mammalian organs'*, Brawand et al. (2011) focused their interest on the evolution of the mammalian transcriptomes[4].

They collected 6 organs from 10 different vertebrates: 9 mammalians (including human) and a bird. There are no technical replicates, but two biological replicates per tissue: one male and one female for every tissue except the testis (two males). The 131 libraries (including 23 for *Homo sapiens*) were prepared with a polyA-selected protocol. Hence,

[4] While there were existing studies on the matter, the sequencing approach was then creating new perspectives.

Table 2.1. **General description of the five transcriptomic datasets (RNA-Seq) used for this study**
Illumina Body Map (IBM) has no ‘regular’ technical replicates as the ‘replicates’ are the product of different protocols, thus are unfit to estimate the specific noise of either protocol (single-end or paired-end). **N.B.**: The protocols used for GTEx and Castle datasets are not the same: GTEx is following the most common ribodepletion protocol, while Castle is based on a targeted amplification protocol.

ArrayExpress ID	Data ID	Library Preparation		Sequencing		Replicates		Number of Tissue Types	Multi-sampling from the same individual
		Total RNA	PolyA selected	Single end	Paired end	Biological	Technical		
E-MTAB-305	Castle	✓		✓				11	
E-GEOD-30352	Brawand		✓	✓		✓		8	
E-MTAB-513	IBM		✓	✓	✓		(✓)	16	
E-MTAB-2836 (and E-MTAB-1733)	Uhlén		✓		✓	✓	✓	32	
E-MTAB-2919	Gtex (v4)		✓		✓	✓		54	✓

✓ indicates that the dataset presents the characteristic, and
(✓) that one (or more) of the required criteria of the characteristic is lacking.

they are largely enriched in protein-coding genes.

An average of 3.2 billion 76 bp-long single-end reads were generated per sample using an Illumina Genome Analyser IIx (single-end) and they released them through GEO (accession number: GSE30352). I personally retrieved the human data from ArrayExpress ID: E-GEOD-30352[5].

2.2.3 *Illumina Body Map 2.0 (IBM)*

This dataset, created in 2010, has been released in 2011[6] by Illumina and it used its most recent technology at that time: the paired-end sequencing. Until then, all the sequencing was done from only one end of the DNA or cDNA fragments. From that date, most of the following transcriptome studies based on RNA-Seq use paired-end sequencing.

The dataset covers 16 tissues (one donor per tissue) and the libraries were prepared following a polyA-selected and are enriched in protein coding genes.

Although each sample has been sequenced twice and despite having in principle *technical* replicates, these are "non-regular" technical replicates. *Technical* replicates, by contrast to *biological* replicates, usually imply that their processing uses the same sample source and protocols. Thus, the error and noise due to a specific technique could be determined. Here, however, each tissue has been sequenced once with a single-end protocol and once with a paired-end one to compare their ability to discriminate between mRNA isoforms. Indeed, Illumina's main incentive to develop its paired-end technology was to improve the accurate identification of spliced mRNAs.

The sequencing was performed with an Illumina HiSeq 2000, and the reads were released through ArrayExpress ID: E-MTAB-503 (ENA ID: ERP000546), from where I have retrieved both the single-end and paired-end mono-tissue samples (the original dataset includes raw data files for mixtures that have been created with the tissue samples).

Despite the lack of biological replicates, it was for an extended time the most extensive freely available RNA-Seq dataset of human tissues. Hence, it has been referenced many times (*e.g.* Asmann et al. (2012), Barbosa-Morais et al. (2012), Smith et al. (2012), Derrien et al. (2012), Florea et al. (2013), D. Kim, Pertea, et al. (2013), Kechavarzi et al. (2014), Zhao (2014), Pasquali et al. (2014), Corpas et al. (2014), Petryszak, Burdett, et al. (2014), Brown et al. (2015), Jänes et al. (2015), De Simone et al. (2016), Kern et al. (2016), Iwakiri et al. (2016), L. Yao et al. (2017), and Akers et al. (2018)) in the literature since its release. In fact, this dataset is the most viewed one in ArrayExpress (with 68,020 views on 31 May 2018 — the second most viewed dataset (46,247 views) being ArrayExpress ID: E-MTAB-62[Lukk et al., 2010]).

[5] ArrayExpress was routinely importing datasets from GEO on a weekly basis until very recently. While not automatically, GEO data are still included in EBI Gene Expression Atlas.

[6] See *Human BodyMap 2.0 data from Illumina* - Ensembl Blog, 2011

2.2.4 *Uhlén et al. dataset*

Uhlén et al. have created the Human Protein Atlas[7] (often referred as HPA in the literature). This atlas revolves mostly around the spatial distribution of the proteins through the human body. Using diverse approaches and techniques, including RNA-Seq, they first released RNA-Seq data for 27 normal tissues as part of their study: *'Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics'* [Fagerberg et al., 2014]. Later, they extended the dataset with new samples and 5 new tissues. The latest version was published within *'Tissue-based map of the human proteome'* [Uhlén, Fagerberg, et al., 2015] in *Science*.

For each of the 32 tissues, there are (at least two) biological replicates. With a few exceptions, the tissues have both male and female donors. Many of the tissues present also technical replicates. The total set comprises 200 samples, which have been picked by pathologists based on the screening of frozen biopsy samples.

The polyA-selected libraries were paired-end sequenced with an Illumina HiSeq 2000 or 2500. I first started to work with the early version of this dataset (ArrayExpress ID: E-MTAB-1733 — 171 samples for 27 tissues), and then I upgraded my work with the extended more recent version (ArrayExpress ID: E-MTAB-2836).

At the preparation time of this thesis, this normal human dataset is the most comprehensive, freely and publicly available dataset: either regarding the number of tissues (see Table 2.1) or the number of samples (see Table 2.2). Therefore, its growing number of references is unsurprising.

2.2.5 *GTEx dataset*

The Genotype-Tissue Expression (GTEx) project is funded by the NIH Common Fund and aims to establish, in its authors' own words, 'a resource database and associated tissue bank for the study of the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues'. The project was first introduced in GTEx Consortium (2013). It aims to quickly collect various tissues from postmortem donors for genotype-tissue expression analyses (notably eQTL studies, which study the function of SNPs in the modulation of RNA expression). The results of the analyses are released through the GTEx portal[8].

As the project is quite ambitious and the collection and sequencing of the samples spreads over a long period of time, several intermediate data '*freezes*' have been released[9]. My analyses include samples up to the fourth release of the pilot phase (v4). This release

[7] Human Protein Atlas — https://www.proteinatlas.org/

[8] GTEx portal — https://gtexportal.org

[9] Many groups are involved in collecting, producing or processing the data. To ease the communication and work coordination, many time points are used to reference each a specific state (version) of the data. Each version of the data is called a *freeze*.

covers 54 tissues/cell types (53 normal and 1 tumoral) collected on from individuals for a total of 3,276 samples.

The RNA-Seq libraries were prepared following a polyA-selected protocols and have been paired-end sequenced on an Illumina HiSeq 2000/2500. There is an average of 80 million reads per sample.

For privacy reasons, the raw data is available only through controlled access via dbGaP ID: phs000424.v4.p1 (access number specific to the version of the data I used in my study). Unfortunately, this translates to a slow access time to the raw data.

During the data selection process, I had to disregard a few studies as the raw data was not fitting the reusability criterion [E. T. Wang et al., 2008; Pan et al., 2008]. Many times I came across studies with ambiguous encoding format for the raw data such as the ArrayExpress ID: E-GEOD-41637 dataset [E. T. Wang et al., 2008]. Despite my best efforts, I was unable to resolve this issue by contacting the respective authors. E. T. Wang et al. (2008) (ArrayExpress ID: E-GEOD-41637) is one example of study that I unfortunately had to dismiss.

2.3 PROTEOME MASS SPECTROMETRY BOTTOM-UP STUDIES

As mentioned earlier, the proteomic data have been selected and handled by Dr Jyoti Choudhary and Dr James Wright.

Until recently, compared to the transcriptome, the proteome world was lacking in normal human tissues expression quantification experiments. In fact, while there were human protein maps available (*e.g.* the Human Protein Atlas[10]), these are mostly reporting the spatial expression of proteins (as they are based on immunohistochemistry or other means of identification) than quantifying their (non-targeted) abundance in each tissue.

In 2014, two independent groups of authors [M.-S. Kim et al., 2014; Wilhelm et al., 2014] published (in *Nature*, issue 7502) their own '*draft of the human proteome*' based on the study of tissues with MS. These two datasets complement a previous smaller one that was publicly released but was never the object of a publication.

Hereinafter, I present these three datasets that I use in my thesis. See Figure 2.3A (p. 69) for a short summary.

[10] Human Protein Atlas — https://www.proteinatlas.org

2.3.1 *Pandey Lab dataset*

The Pandey Lab [M.-S. Kim et al., 2014] created the Human Proteome Map[11] which they released alongside *'A draft map of the human proteome'* in *Nature*.

For their study, they processed 30 kinds of histological normal human tissues and cell line samples (17 adult tissues, 7 foetal tissues and 6 haematopoietic cell types). Each sample was created from pooling samples of three individuals (generally two males and one female).

Their proteomic libraries were prepared with a label-free method to quantify as many proteins as possible. The samples were fractionated to protein level through SDS-PAGE, and then at peptide level after trypsin digestion by RPLC to create 85 experimental samples. Finally, state-of-art MS/MS protocols (with high-resolution and high accuracy FTMSs Thermo Scientific Orbitrap™ instruments) was used to generate about 25 million of (HCD) high-resolution mass spectra which account for 2,212 LC-MS/MS profiles. The raw spectra were retrieved from ProteomeXchange via the repository PRIDE ID: PXD000561.

While the authors' effort to generate technical high quality raw data was highly appraised by the scientific community, their processing (identification and quantification) methods were criticised (see Ezkurdia et al. (2014) and Deutsch et al. (2015)). Thus, for this thesis I have relied only on quantifications provided by Dr James Wright who reprocessed the raw spectra.

2.3.2 *Kuster Lab dataset*

In their approach of the human proteome map, the Kuster Lab [Wilhelm et al., 2014] combined newly generated LC-MS/MS spectrum data (about 40% of their complete working set) with already publicly available data (either from their colleagues or accessible through repositories — for the remaining 60%). They reprocessed the whole collection of spectra to maximise proteome coverage and make it available through their own repository: ProteomicsDB[12].

The subset of data considered in my thesis is also known as the [protein] Human BodyMap which is the part that the Kuster Lab primary generated for their own study. They collected 48 experiments covering 36 tissues (adult and foetal) and cell lines. After LDS-PAGE fractionation and digestion into peptides with trypsin, they processed the samples with LC-MS/MS to create 1,087 profiles. Overall, that represents about 14 million of HCD/CID spectra from Thermo Scientific instruments (including an Orbitrap™). This specific raw data subpart was downloaded from ProteomicsDB ID: PRDB000042.

[11] Human Proteome Map — http://www.humanproteomemap.org

[12] ProteomicsDB — https://www.proteomicsdb.org/

2.3.3 *Cutler Lab dataset*

This dataset was generated prior to the Pandey Lab and the Kuster Lab data as it was released in 2011 through PeptideAtlas[13] [Desiere et al., 2006], IDs: [PAe001768 — PAe001778].

It was created by Paul Cutler at Roche Pharmaceuticals. It comprises 10 different tissues (and one sample per tissue) that after trypsin digestion, were analysed through Thermo Scientific LTQ-Orbitrap™. In total, there are 1,618 CID profiles which accounts for 13 million raw CID spectra from a LTQ-Orbitrap™ instrument.

While this dataset was never published on its own, it has been used in different studies (*e.g.* Wilhelm et al., 2014). The raw files were accessed and downloaded from ProteomicsDB ID: PRDB000012.

2.4 CONSISTENT PROCESSING PIPELINES

The authors of these five transcriptomic and three proteomic studies have, in most cases, released the quantification of the expression values either directly (*e.g.* Krupp et al., 2012) or upon requests (*e.g.* M.-S. Kim et al., 2014). Third-parties also distribute quantification for these studies either retrieved from the original studies, such as BioGPS[14] [C. Wu, Macleod, et al., 2013] or Harmonizome[15] [Rouillard et al., 2016], or after reprocessing the raw data as the *EBI Gene Expression Atlas*[16] [Petryszak, Keays, et al., 2015] does.

To primarily reduce for avoidable technical variability, and despite readily available quantifications for most of the datasets, I only used data reprocessed from raw files by myself or Dr Nuno Fonseca (GTEx dataset) for the transcriptomic data and by Dr James Wright for the proteomic data as already mentioned.

In fact, each study has been originally processed with different protocols, *e.g.* GTEx [Melé et al., 2015] and Castle [Krupp et al., 2012]. While the EBI Gene Expression Atlas reprocesses raw data through the same methods and has quantification for most of the aforementioned datasets (it is still lacking the Castle et al. one.), these were still the products of different protocols when I started my work. Indeed, the datasets were processed with different versions of reference (Human genome build and annotation) and tools.

Intuitively, we expect that different processing protocols produce different results. As I started to work with RNA-Seq data, I noticed many potential analysis variables that impact at various levels the resulting gene expression values. Indeed, many of these have

[13] PeptideAtlas — http://www.peptideatlas.org/
[14] BioGPS — http://biogps.org/
[15] Harmonizome — https://amp.pharm.mssm.edu/Harmonizome
[16] *EBI Gene Expression Atlas* — https://www.ebi.ac.uk/gxa/home

been reported in the literature since then; in fact, annotation versions [Frankish et al., 2015], contamination (from viruses or bacteria DNA) [Cantalupo et al., 2015], quality controls (and subsequent reads filtering choices) [Kroll et al., 2014], mapping and quantifications pipelines [Fonseca, J. Marioni, et al., 2014] have considerable effects on the final quantification. Lastly, normalisation methods also greatly impact the final expression figures [Dillies et al., 2013; Zwiener et al., 2014]. For all these reasons, I decided to reprocess all transcriptomic datasets with the same exact protocol as the first step of my study. Recently, Danielsson et al. (2015) compare results based on prepublished data and reprocessed ones and conclude that using a single processing pipeline ensures better results.

Likewise, there are various tools and many parameters for each processing step needed to quantify the proteomic data that may impact the final expression values [Aebersold, 2011]. For example, various search engines allow detecting different sets of peptides [Griss, 2016]. Mackay (2015) has reviewed and analysed the impact of many of these variables more specifically for label-free proteomics, such as the effect of FDR, protein inference tools or normalisation methods. Therefore, the three datasets were reprocessed uniformly from the raw spectra up to the normalisation of the protein expression values.

2.4.1 *RNA-Seq raw data processing*

As presented in Section 1.2.5, there are many steps from the raw data files to the quantification matrices on which this thesis' analyses are based. Figure 2.1 presents a general overview of the RNA-Seq processing protocol I used.

I downloaded and entirely processed four of the transcriptomic datasets myself (Castle, Brawand, IBM and Uhlén data) and Dr Nuno Fonseca retrieved and processed the GTEx dataset[17]. In this thesis, I present results computed on the quantification of these five datasets which have been processed through the same identical pipeline.

2.4.1.1 *Data retrieval and preparation*

I retrieved the human raw data of each dataset from ArrayExpress and ENA through their identifier (see section 2.2) (p. 56). After we received our access approval, Dr Nuno Fonseca retrieved GTEx data from dbGaP.

While most of the raw files can be used as they are, an additional step is needed for the Castle files. Indeed these files are using an older FASTQ format that is non-compliant to the most accurate and recent tools used for this thesis. As it is a simple matter of changing the quality score scale (see appendix A.6), I converted these files to Phred+33 FASTQ files

[17] As the GTEx data is involved in many projects within the EBI and due to its huge amount of files (number and size — see Table 2.2), it was agreed that this would be processed centrally by one person and then redistributed to all the other interested parties. Dr Nuno Fonseca had this tremendous task.

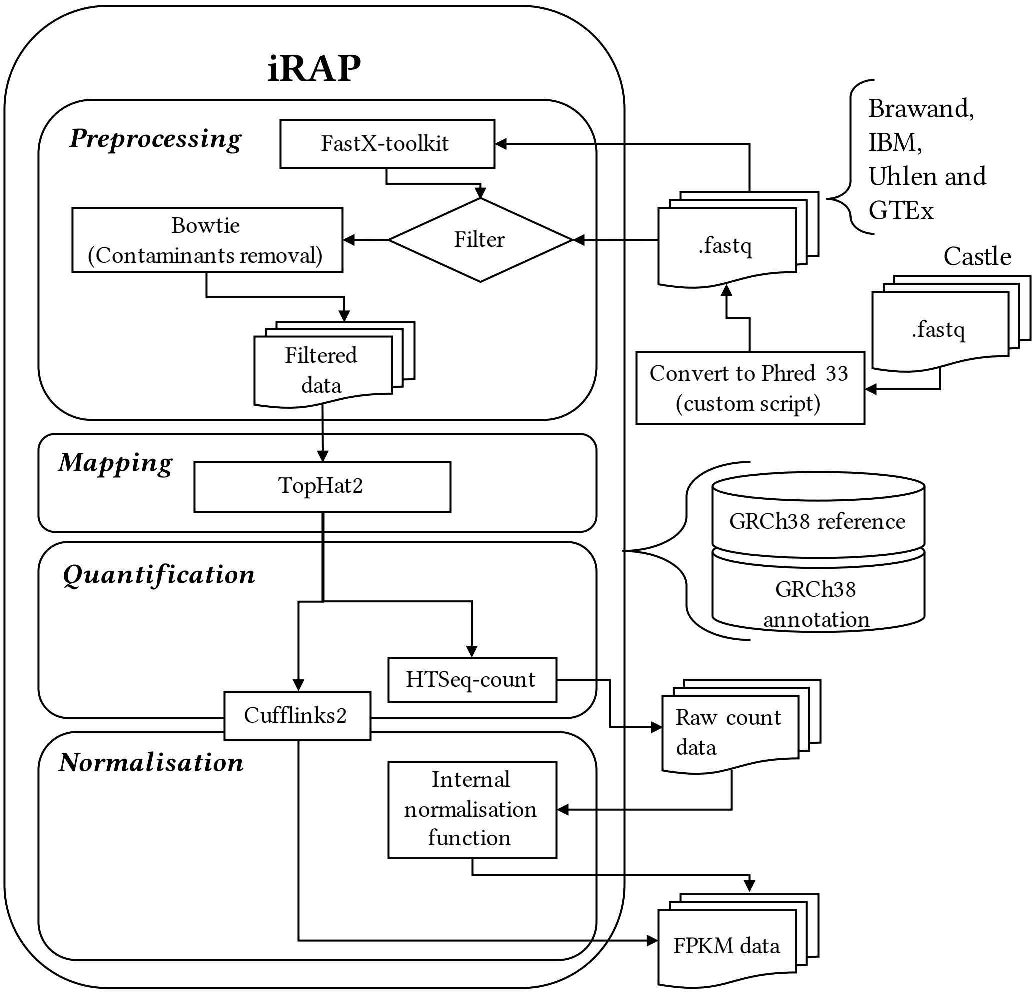


Figure 2.1. **General steps for processing the transcriptome.** The pipeline *iRAP* integrates all the tools needed for the state-of-art processing of RNA-Seq data. The quality of the reads is checked and they are trimmed if needed. After removal of possible contaminant reads (such as *E. coli*), the reads are aligned with *TopHat2*. The gene expression is then quantified with two different approaches: based on the aggregation of isomers for each gene or simply based on the number of aligned fragment on the gene locus defined in the reference. *Cufflinks2* provides directly FPKM values. *HTSeq-count* provides raw counts which were normalised by an *iRAP* function into FPKM.

Table 2.2. **Technical description of the five transcriptomic datasets**
I processed all the datasets except the one in *italic*.
For the Brawand dataset, I only included and processed the *Homo sapiens* part.

Dataset	Participant number	Library number	File number	Total size of the fastq raw files (GB)	Mean number of biologic samples per tissues [min;max]
Castle	10	11	11	58	10 (mixture)
Brawand	18	21	23	111	2.8 [2;3]
Illumina Body Map	16	36	48	1,004	1
Uhlén	122	200	400	1,851	3.81 [2;11]
GTEx (v4)	*551*	*3,276*	*6,552*	$\sim$ *50,000*	*60.67 [4;214]*

with a *Perl* script (provided digitally as supplementary data).

2.4.1.2 *Genome and annotation reference*

I collected and processed the datasets through an extended period of time. Hence, for a subset of them, I produced many intermediate sets of results based on the GRCh37.p12 (and later GRCh37.p13) human reference genome and the latest available Ensembl gene set annotation (73, 74 or 75) at that time. In fact, the quality of each new annotation update is generally greater than its predecessor[18].

As the GTEx data was processed with GRCh38.p1 and Ensembl 76, that led me to reprocess all the other four RNA-Seq datasets for the sake of consistency and to avoid more biases [Guo et al., 2017]. Thus, unless indicated otherwise, the results presented in the current work are based on the GRCh38.p1 human genome reference and the Ensembl 76 gene set annotation.

2.4.1.3 *Data processing*

In the early stages of my research, I was processing each of the different steps sequentially and semi-manually with the help of custom made scripts. While the EBI computer cluster greatly facilitated the handling of the numerous files, the task remained quite tedious. Additionally, the scripts I wrote would need a fair amount of work to achieve general reproducibility on other platforms.

Fortunately, Dr Nuno Fonseca developed an 'integrated RNA-seq analysis Pipeline': iRAP[19] [Fonseca, Petryszak, et al., 2014]. This tool allows the automation of the typical state-of-the-art and optimised workflow to study RNA-Seq. It takes full advantages of the capacities provided by computer clusters. Thus, I switched from my original set of

[18] Although, it is not unusual to have gene or transcript additions based on new studies that are then removed (or fused to another) in a later version.

[19] iRAP — https://nunofonseca.github.io/irap/

scripts to *iRAP* to improve my workflow without changing any step or parameter.

Besides the usual input files (raw RNA-Seq files and genome/annotation references), *iRAP* needs a configuration file that precisely describes the dataset (its design and technical features) and, if needed, specific parameters to use. To provide full reproducibility, each version of *iRAP* is shipped with its own set of third-party version-defined tools and default parameters. Thus, apart from remarkably speeding up the data processing, *iRAP* also ensures the protocol integrity across the five transcriptomic datasets I use in my thesis regardless of who runs the pipeline.

Each of the transcriptomic datasets is the product of the same version of the *iRAP* pipeline (development version 0.6.3b) and set of parameters. As the default parameters of *iRAP* are tuned for human Illumina paired-end data, I only have to define a few of them. Hence, the quality and contamination checks, and the filtering and trimming of the reads are done following the default options of *iRAP*.

Quality assessment, trimming and filtering

iRAP uses internally *FastX toolkit*[20] (0.0.13) to perform the assessment and the trimming. The usual uninformative and ambiguous reads (see Section 1.2.5.1: Quality check, trimming and filtering (p. 16)) have been discarded as were any with an overall quality score below a threshold of 10.

The quality of the call decreases while the base calling progresses — see Section 1.2.3: Sequencing-by-synthesis (p. 13). On another note, some tools (mappers in particular) need all the reads to be trimmed to the same length. *iRAP* optimises the compromise between the purity and the length of the reads to avoid more errors or biases due to smaller reads [Williams et al., 2016] by trimming at most 15% of the original length while discarding more reads if necessary to maximise the length.

Reads that could be assigned to a likely contamination source, here *Escherichia coli* (as I work with *Homo sapiens*), are also discarded. A non-splice aware mapper, *Bowtie*[21] (1.1.1) [Langmead et al., 2009] maps all the reads to the contaminant genome and all the reads mapping perfectly and unambiguously are discarded.

Mapping

I mapped the reads to the genome (GRCh38.p1) and the transcriptome (Ensembl 76 gene set annotation) with *iRAP*'s (0.6.3b) proposed default splice-aware mapper *TopHat2*[22] (2.0.12) [D. Kim, Pertea, et al., 2013] with its set of default predefined arguments. Indeed, *TopHat2* can handle reads from many organisms by fine-tuning the parameters (*e.g.* number of mismatches or indels to tolerate), but the default parameters are adjusted for

[20] *FastX toolkit* — http://hannonlab.cshl.edu/fastx_toolkit/

[21] *Bowtie* — http://bowtie-bio.sourceforge.net/

[22] *TopHat2* — https://ccb.jhu.edu/software/tophat/index.shtml

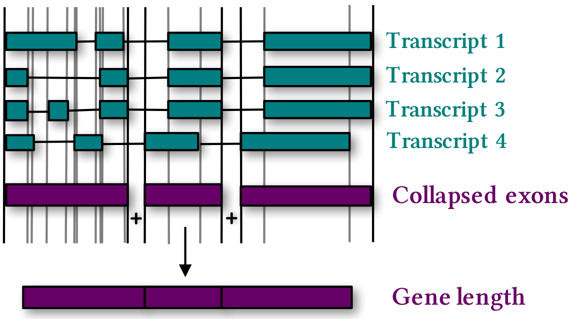


Figure 2.2. **Gene length is equal to the sum of the lengths of all its collapsed exons.** Though, this method lacks complete accuracy, it provides a sufficient estimation of the gene length for an efficient normalisation regarding the length bias. The coordinates for the 5' and 3' ends of each exon is extracted from the annotation and they are collapsed together. This *gene length* is unaffected by incorrect attribution of a fragment to a specific transcript when there are many possible options.

normal human.

Quantification and Normalisation

While RNA-Seq can be used to identify (and discover) RNA isoforms, I have focused my thesis on the gene level expression. Indeed, current annotations and knowledge are still lacking in the reasons and external conditions that impact the expression of a specific isoform over the others. In addition, criticisms have been raised on the accuracy of distinction between them [Engström et al., 2013; Jänes et al., 2015; Dapas et al., 2017].

However, normalising gene expression presents more challenges than specific transcript expression. For instance, the definition of the gene length may be different from one laboratory to another. In this thesis' framework, when I have to use a gene length for a computation, I use the identical gene length definition as found in *iRAP* and EBI Gene Expression Atlas. Thus, as shown on Figure 2.2 the gene length is defined as the sum of the lengths of all its *collapsed* exons.

As mentioned in Section 1.2.5.4: Normalisation (p. 23), I used two different popular tools based on different strategies to estimate gene expression levels: *Cufflinks2*[23] (2.2.1) [Trapnell et al., 2010] and *HTSeq-count*[24] (0.6.1p1) [Anders et al., 2015] (with the `intersection non-empty` mode). These tools are also integrated in *iRAP*.

For *Cufflinks2*, I used the mode where the multi-mapped reads are probabilistically assigned depending on the coverage of each mapped locus. In addition, *Cufflinks2* provides normalised gene expression levels by aggregating their corresponding normalised isoform expression levels. *Cufflinks2* uses the equation (Canonical F/RPKM formula) to normalise isoform expression levels. The length of the isoforms are extracted

[23] *Cufflinks2* — https://cole-trapnell-lab.github.io/cufflinks/manual/
[24] *HTSeq-count* — https://htseq.readthedocs.io/

from the reference.

On the other hand, *HTSeq-count* provides only *raw* counts for the feature of interest. *iRAP* provides an internal FPKM normalisation function that is an implementation of the equation (Canonical F/RPKM formula). As I requested *HTSeq-count* to work at gene level, this formula requires gene lengths which are computed with the aforementioned method.

All the configuration files I created for this thesis may be found at my personal Github repository[25].

As the paired-end set of the IBM data was presenting an overall better quality than its single-end counterpart, I only include IBM's paired-end data for the remaining of the thesis.

2.4.2 *MS data processing*

After retrieval of the data from PRIDE and ProteomicsDB, Dr James Wright reprocessed the three proteome MS-based datasets. Figure 2.3 illustrates the pipeline that processed the three datasets in a consistent and optimal manner. I summarise this protocol in Figure 2.3 (p. 69) and in the following sections 2.4.2.1 to 2.4.2.4. See Wright, Mudge, et al. (2016) and Weisser et al. (2016) for more details.

2.4.2.1 *Spectral processing*

The *msconvert* module of *ProteoWizard*[26] (v3.0.6485) [Holman et al., 2014] converted all the files to the standard format mzML. *TOPP* [Kohlbacher et al., 2007] from *OpenMS*[27] (pre-v2.0 development build) [Röst et al., 2016], processed the raw spectra. Notably, *PeakPickerHiRes* which centroids them and *FileMerger* that merges the ones from the same fractionated experiments.

2.4.2.2 *Sequence database creation and searching preparation*

The target sequence database is a critical element of the MS pipeline and thus, Dr James Wright has carefully designed it. It combines six different parts, three based on known protein sequences and three other covering possible new protein candidates.

The known sources include the complete human GRCh38 (v.20) coding DNA sequence (CDS) translated sequences from GENCODE; the human reference proteome from UniProt[28] [The UniProt Consortium, 2017] (in its May 2014 version); common

[25] https://github.com/barzine/phd-analyses/tree/master/chapter2/irap-configuration-files
[26] *ProteoWizard* — http://proteowizard.sourceforge.net/
[27] *OpenMS* — https://www.openms.de/
[28] UniProt — http://www.uniprot.org/

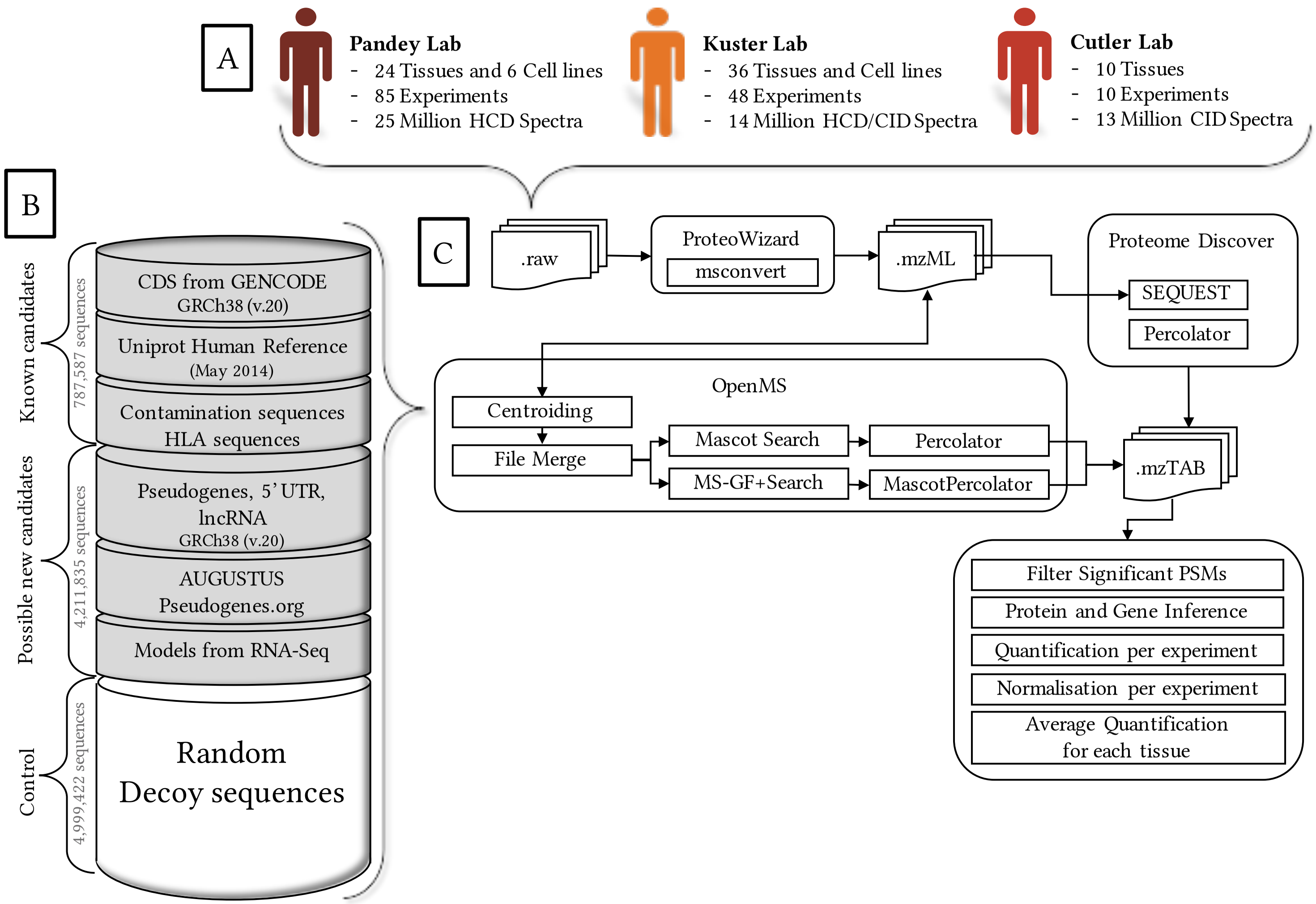


Figure 2.3. **General steps for processing the proteome.** [Adaptation of courtesy materials from Dr James Wright]. (**A**) The three datasets have been processed through the same pipeline. In this thesis, I only use the samples from adult tissues. (**B**) Extensive sources of protein sequences were used for the search database, including prediction of novel proteins. Contamination and decoy sequences were also included to allow for FDR estimation. (**C**) State of the art workflow was used to process the MS data from raw files. This workflow combines multiple MS search engines and post-search evaluation tools. Results were filtered by peptide length, FDR, PEP and agreement between the multiple search algorithms. Note that there is no relation between the real size of the database parts and their representation; the decoy sequences are as numerous as the sum of the known and possible candidates.

contamination protein sequences[29] and HLA sequences[30]. This *known* portion of the target sequence database represents 787,587 tryptic peptide sequences.

The sources for potential novel proteins included a selection of non-coding gene sequences (including pseudogenes, lncRNA and untranslated regions) from GENCODE GRCh38 (v.20); prediction of novel sequences with *AUGUSTUS*[31] [Stanke et al., 2004]; a set of two-consensus predictions (December 2013) from *Pseudogene.org*[32] [Karro et al., 2007] and three-frame translated RNA-Seq transcript sequences. These translated sequences include models built on IBM by Ensembl and by the Kellis lab in addition with models built on different ENCODE cell lines by Caltech and CSHL. This *novel* portion of the target sequence database provides an addition of 4,211,835 tryptic peptide sequences.

Mimic[33] generated 4,999,422 $(787,587 + 4,211,835)$ randomised decoy sequences, *i.e.* the decoy database and the target database have an equal size of peptide sequences. The different databases were then merged together. It is represented on Figure 2.3B.

To account for the isobaric peptides, all isoleucine (I) residues were converted to leucine (L) before the search and then after the search all leucine (L) residues were converted to (J)[34] to avoid later misconceptions.

2.4.2.3 *Spectral identification and database search pipeline*

Figure 2.3C describes the overall workflow used by Dr James Wright to quantify the protein abundance in each tissue. As mentioned in Section 1.3.4.2, workflows involving several algorithms produce better results. *Mascot* Server (v 2.4— Matrix Science) cluster produced a first search on the mzML files submitted through *MascotAdapterOnline* (part of *TOPP*). In parallel, Dr James Wright also used *MS-GF+ Search*, which involves the run of *MS-GF+* (v. 10089) [S. Kim et al., 2014]. *MascotPercolator* (v 2.08) [Brosch et al., 2009; Wright, Collins, et al., 2012] optimised and rescored the results from *Mascot* and *msgf2pin/Percolator* (v 2.08−1) [Granholm et al., 2014] optimised the results from *MS-GF+*. Finally, *SEQUEST* [Eng, McCormack, et al., 1994] and *Percolator* [Spivak, Weston, Bottou, et al., 2009] performed a search in a *Proteome Discoverer* (v 1.4 — Thermo Scientific) workflow.

The different workflows used common stringent parameters for all the database searches: the precursor tolerance was set to 10 ppm; fragment tolerance for HCD spectra to 0.02 Da and to 0.5 Da for CID spectra; the allowed missed cleavages were limited to 3. As described in Wright, Mudge, et al. (2016), the research also accounted for several amino acid modifications by including (known) mass tolerances. The fixed modification

[29] Contamination sequences — http://maxquant.org/contaminants.zip
[30] HLA sequences — https://www.ebi.ac.uk/ipd/imgt/hla/download.html
[31] *AUGUSTUS* — http://bioinf.uni-greifswald.de/augustus/
[32] *Pseudogene.org* — http://pseudogene.org/
[33] *Mimic* — https://github.com/percolator/mimic
[34] As J is one of the letter from the Latin alphabet that do not map to any amino acid.

carbamidomethyl (+57.0214 Da) was specified for all cysteine residues. The searches also comprised the following variable modifications: N-terminal acetylation (+42.01056 Da), N-terminal carbamidomethyl (+57.0214 Da), deamidation of asparagine and glutamine residues (+0.984 Da), oxidation of methionine residues (+15.9949 Da), and the possible N-terminal conversion to pyro-glutamine of glutamine (−17.0265 Da) and glutamic acid (−18.0106 Da) residues.

The search results were converted into mzTab formatted files and uploaded along with the mzML spectra and FASTA search database to PRIDE ID: PXD002967.

2.4.2.4 *Results processing and filtering*

Custom Perl scripts parsed, merged and filtered the results of each search engine so that every PSM had the same identification in at least two of the three search engines. In each case, the least confident PEP (*i.e.* the highest, see Appendix A.11) was retained.

The PSMs were then filtered to keep matches only to the three following criteria: *q*-value (see Appendix A.11 and Appendix A.9.3) less than or equal to 0.01 (*i.e.* 1% FDR); a PEP inferior or equal to 0.05 and a peptide length superior or equal to seven amino acids. PSMs matching contaminant or decoy sequences were also removed.

The resulting list of peptides was then used to infer the proteins with a simple approach. Protein clusters were created based on the common matching non-null set of peptides, *i.e.* each protein cluster has at least one unique peptide. Then, the GENCODE CDS and UniProt accession were mapped back to Ensembl identifiers. Proteins with a gene (or gene clusters) definition matching at least three unique peptides were kept for the remaining of the analysis while the others were discarded.

The quantification of the retained proteins was computed for each experiment with an approach close to the Top3 method [Silva et al., 2006]. The precursor intensities of the three most intense *unique* peptides per gene identifier (or for gene cluster) were summed, before being divided by the total summed quantification of all proteins in each sample to provide the '*within sample abundance*'. Then, these abundance values were normalised by the ten genes displaying the lowest coefficient of variation across all tissues. When there was more than one experiment per tissue, the final quantification values are the median value across all the replicates of each tissue.

Protein clusters matching several Ensembl gene identifiers or failing the *unique peptide* rule are discarded from the presented further analyses. The list of the discarded clusters is different for each of the proteome datasets.

Compared to the original Pandey Lab study [M.-S. Kim et al., 2014], fewer proteins were quantified, but the results are congruent to other previous studies on the range of detection and quantification of LC-MS/MS. The quantifications were released along with our paper [Wright, Mudge, et al., 2016] and the reanalysis of the Pandey data was also released through EBI Gene Expression Atlas under the accession: E–PROT–1 and described in Petryszak, Keays, et al. (2015) and Wright, Mudge, et al. (2016).

2.5 DISCUSSION AND CONCLUSION

In this chapter, I introduced the five transcriptomic and three proteomic normal human tissues datasets on which I based my thesis. I described how both the transcriptomic and the proteomic datasets have been reprocessed from raw files with state-of-the-art unified pipelines which are also using the same genome build and annotation references in the final processed version.

As mentioned before, I have produced a subset of the transcriptomic datasets with the previous human reference genome (GRCh37) and three different Ensembl gene set annotations (73, 74 and 75). I have run many of the analyses of Chapters 3 and 4 on these data. While the results may vary for individual genes, the overall outcomes are congruent hence supporting the robustness of the findings presented in this thesis. In addition, all the products of the RNA-Seq pipeline are in agreement with the original studies findings. The MS pipeline also produces similar results to the original studies — except for M.-S. Kim et al. (2014), which original processing and results raised many criticisms [Ezkurdia et al., 2014].

While we are in the era of *data deluge* and *big data*, the number of tissue overlaps for independent normal human studies is surprisingly small — see Figure 4.1. (p. 92) Most of these datasets have been (and will be) referenced through many papers for comparison (or as control) purposes; hence, it is essential to assess the soundness of these practices by assessing the consistency between these datasets.

See first, think later, then test. But always see first.
Otherwise you will only see what you were expecting.
Douglas Adams

3 ABOUT EXPRESSION, VISUALISATION, CORRELATION AND CLUSTERING

As a first step towards the different meta-analyses presented in this thesis, I have opted for a largely empirical approach to determine a consensus set of methods and parameters on each individual study before applying them across all the datasets in the further chapters. This strategy has also allowed me to estimate the overall data quality per dataset and to structure them appropriately for the upcoming analyses.

I mentioned in Chapter 2 quality checks that happened before the processing of the data. Those quality assessments are rather technical[1]. In the present chapter, I describe post-processing quality (or sanity) checks that examine higher (and fuzzier) aspects, *e.g.*:

- Possible outliers in the data
- Systematic and unsystematic *batch effect* within each study
- Adequacy of data, concepts and statistical models

Even if every of these aspects may not be addressed or corrected, the final results and interpretations of this study are then more solid.

3.1 VISUALISATION OF EXPRESSION DATA

Data visualisation is a simple, but very effective method towards adequate analyses and thus more pertinent results. It allows uncovering the detection of underlying structures and possible unwanted artefacts.

3.1.1 *Distribution plots*

In the literature, expression values are frequently visualised on a log-scale ($\log_2(x)$). Figure 3.1 illustrates how this scaling improves the readability of the plot. To overcome the lack of definition of $\log_2(0)$, I have added a common *pseudocount* (equal to 1) to all the observations. However, in a few cases and only for visualisation purposes, I have removed the null values to avoid misinterpretations; for examples the expression distribution (per tissue) plots Figures 3.1 to 3.3. When I remove the null values I clearly state it in the plot legend as the norm is the pseudocount addition.

[1] Is it true signal or noise? Are all the nucleotides called? Is it a true identification or a false positive? ...

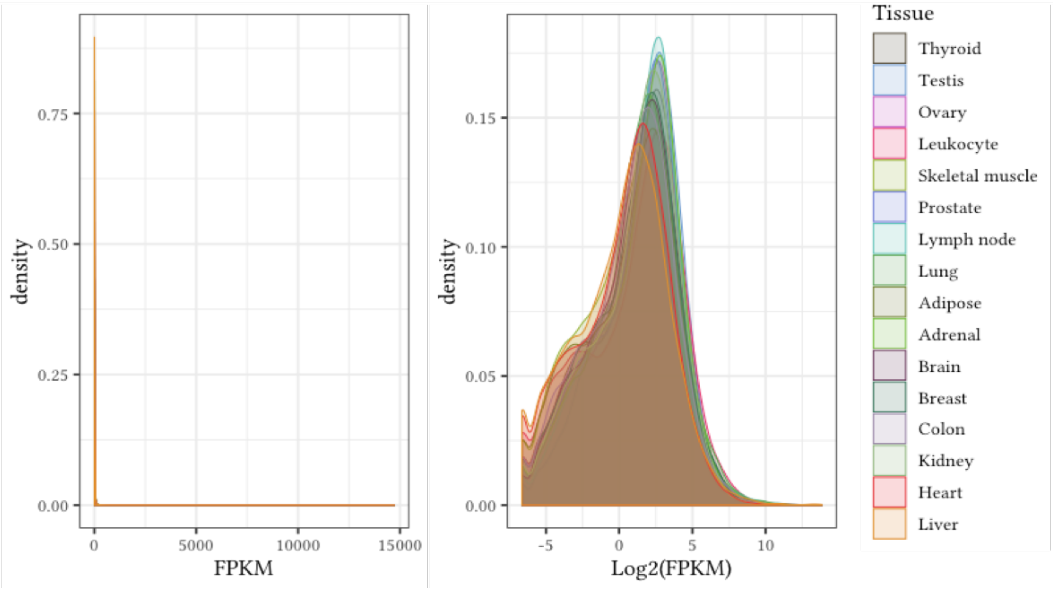


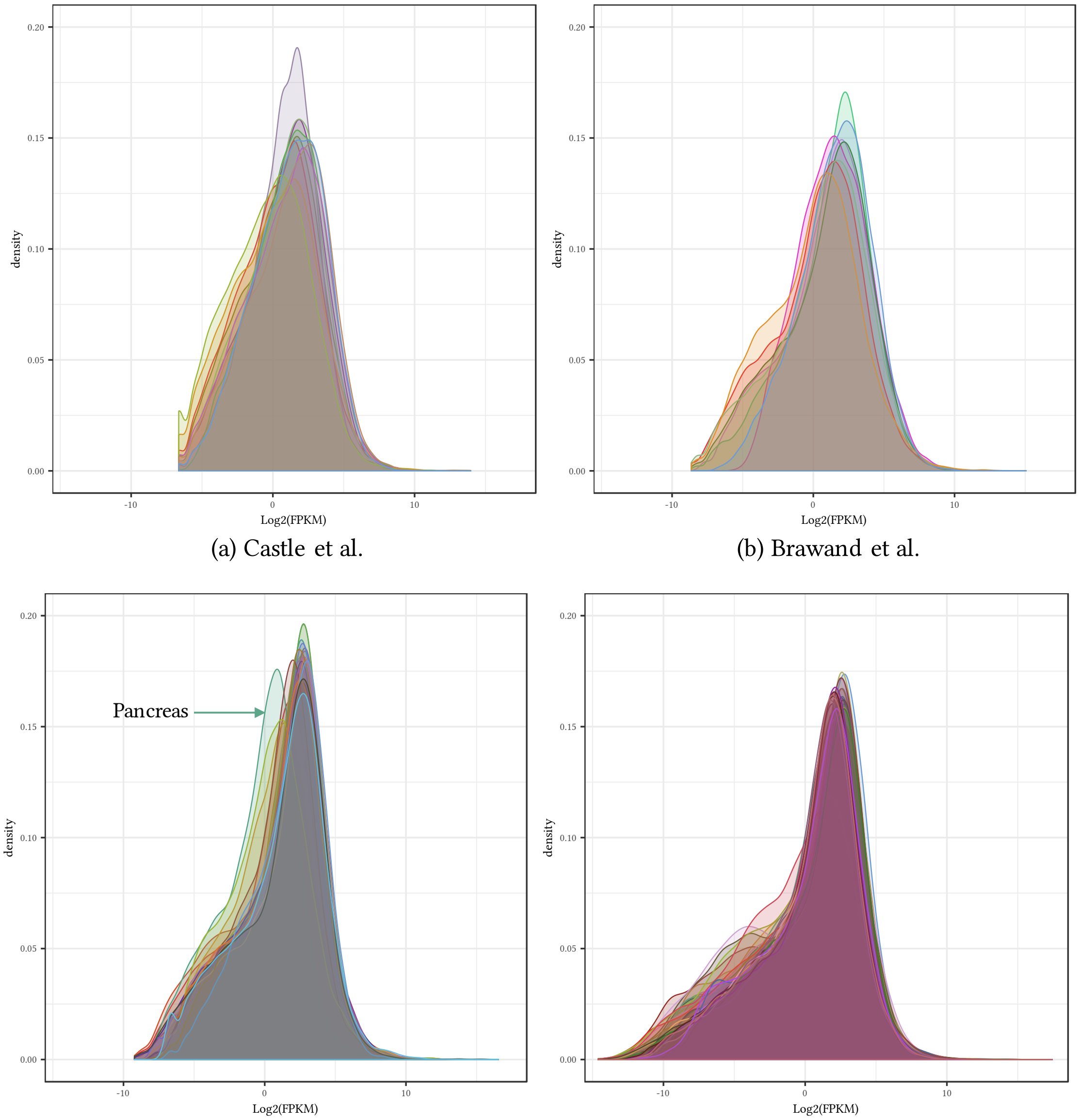
Figure 3.1. Untransformed (left) and Log$_2$-transformed (null values removed) (right) profile of expression levels (FPKM, protein-coding genes only and all null values excluded) for the IBM dataset

Figure 3.2 shows all the remaining transcriptome datasets. Overall, on this $\log_2(x)$ scale (and with all null values excluded): all the samples present a similar shape; a peak near 0 for the lowly expressed and undetected genes and a long-trailing tail. The bulk of the expressed genes on this scale is below 6 (*i.e.* below 63 FPKM). In Figure 3.2c, we can observe that the general expression of the pancreas is shifted towards the left in comparison to the other tissues. This may be an artefact as this shift of the values distribution is absent in the pancreas of the other transcriptomic studies (Figure 3.1b and Figure 3.2d). Moreover, as highlighted by the next chapter analyses, Uhlén's and GTEx's pancreas are strongly correlated ($r = 0.83$;$\rho = 0.96$).

Aside from the Pandey data (Figure 3.3c), the expression of the proteins is more heterogeneous (in particular Cutler data, see Figure 3.3a). This is concordant to the more disparate and variable techniques involved in the proteomic sample preparation (see Section 1.3.1: Sample preparation).

3.1.2 *Scatter plots*

Anscombe (1973) created four datasets (see Figure B.1) which share similar descriptive statistics to show the importance of data visualisation even through a simple scatter plot. He demonstrated that checking the datasets graphically with scatter plots allows one to quickly detect outliers and roughly estimate the relationship between two variables. Even a non-linear but strong relationship is promptly highlighted (*e.g.* top right corner of Figure B.1).



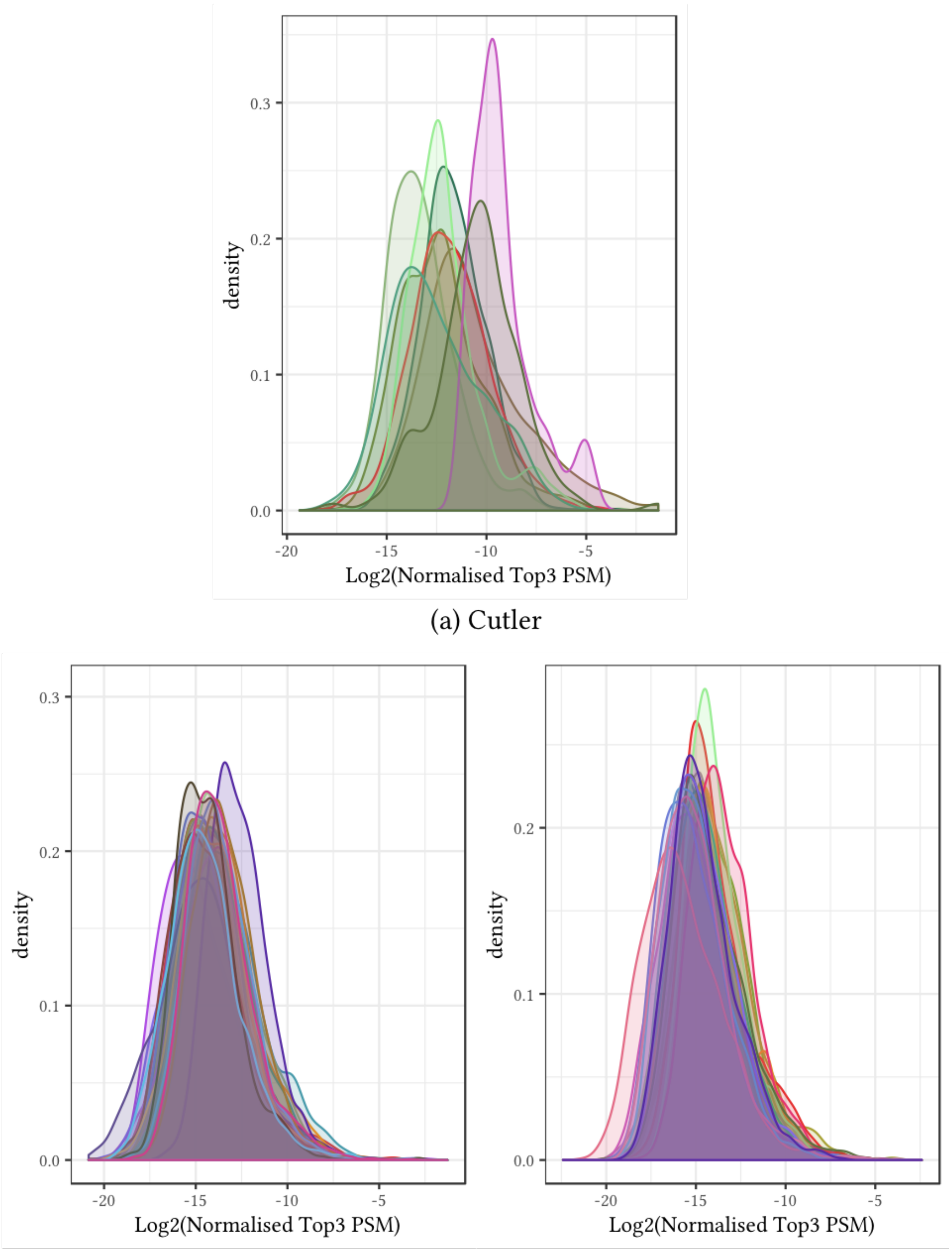
(a) Castle et al.

(b) Brawand et al.

(c) Uhlén et al.

(d) GTEx

Figure 3.2. Profile of expression levels across the transcriptomic (protein-coding genes only) studies (null values removed)

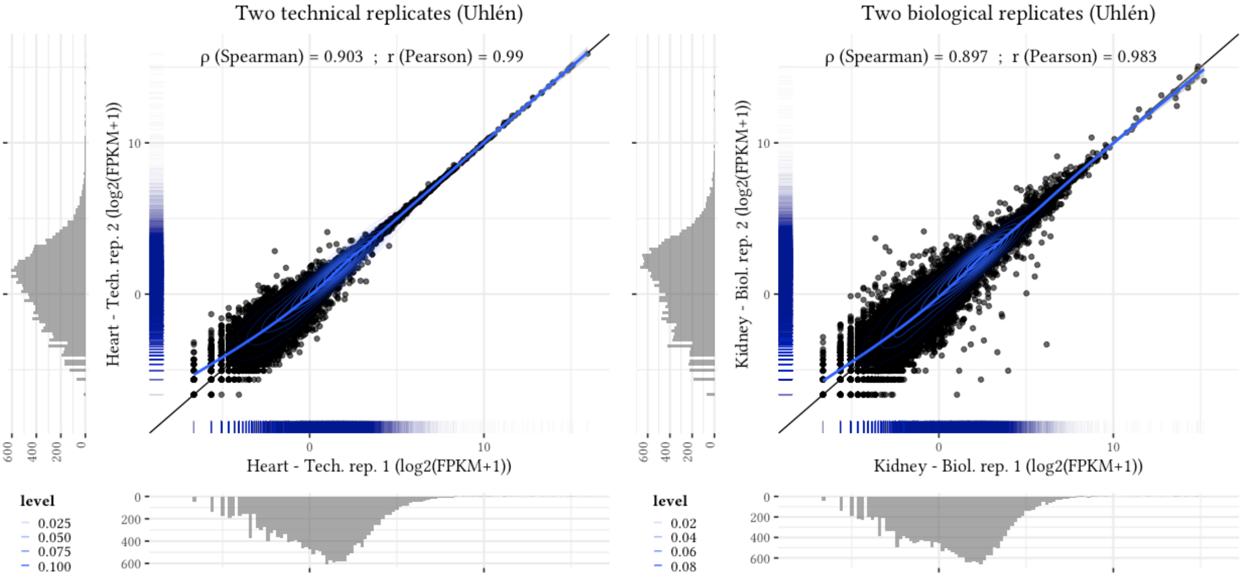


(a) Cutler

(b) Kuster

(c) Pandey

Figure 3.3. Profile of expression levels across the proteomic studies (null values removed)



(a) Technical replicates (Heart)

(b) Biological replicates (Kidney)

Figure 3.4. **Examples of scatter plot for replicates from** Uhlén **(transcriptome)** Technical replicates present very strong correlations particularly for higher expressed mRNAs (≥ 32 FPKM). Biological replicates present lower but still strong correlations within the same dataset.

Figure 3.4a illustrates how very lowly detected RNAs diverge even in technical replicates. Biological replicates, within a same dataset, may present very close profiles even if the spread for the lowly detected genes is even greater as showed in Figure 3.4b.

3.2 MAIN STATISTICAL APPROACHES

As the general normal distribution shape of the gene expression levels on log-scale are similar, I have also computed Pearson correlation coefficients (in addition to Spearman ones) to assess the similarity of the replicates within (intra) and between (inter) studies.

3.2.1 *Correlation*

Correlation coefficients are a measure of the statistic dependence between two continuous variables[2] (*e.g.* X and Y) and always ranges within $[-1, 1]$ (see also appendix B.1: Correlation).

[2] In the context of this study, the variables are either expression levels of a given gene across samples/tissues or expression levels of all genes *between* two samples or tissues

The correlation coefficient is computed by the pairwise comparison of observations between two variables. Most implementation methods will manage an unbalanced number of observations by excluding the incomplete pairs. To ease the interpretation I filtered the data *a priori*; I only kept expression values *effectively observed* in all the datasets (as I explain in section 3.3.3: Expressed or not expressed).

From the several methods available to compute the correlation coefficient, I chose both the Spearman and the Pearson correlations. As Spearman correlations are computed on ranks, they report any kind of relationship, while Pearson correlations are computed on the values and report only linear relationships. However, Pearson correlations are easier to interpret and can be used with one of the variable to predict the other one. (See also Appendix B.1.1: Spearman correlation and Appendix B.1.2: Pearson correlation).

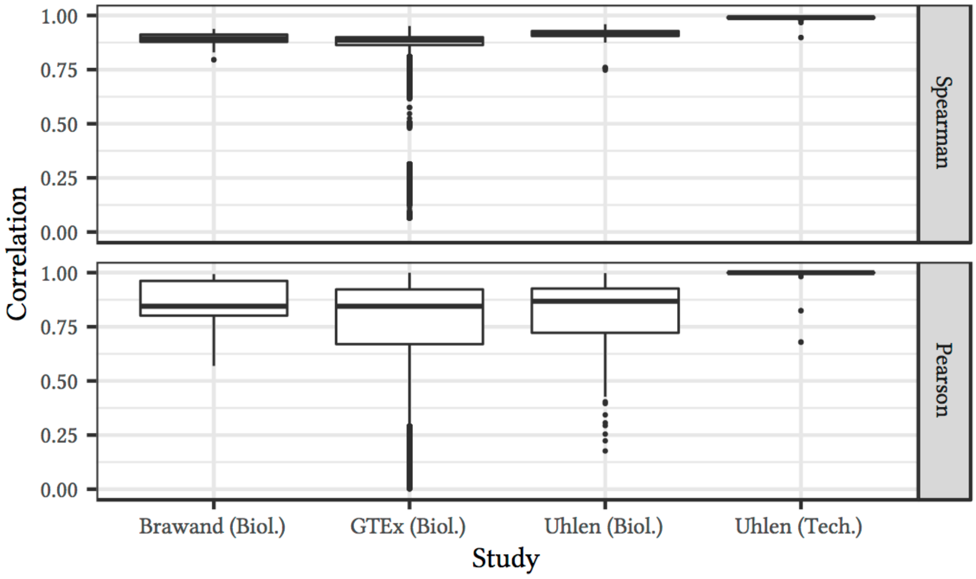


Figure 3.5. **Correlation coefficients between RNA-Seq replicates.** The correlation means and medians are high across the studies replicates. However, the range of the correlations are quite extreme in a few case. Spearman correlations are higher than the Pearson ones. See also Appendix B.1.3.

Figure 3.5 (and Table B.1) presents the Pearson and Spearman correlation coefficients for the technical replicates for Uhlén study and the biological replicates within Brawand, GTEx and Uhlén studies. On average the correlation coefficients are high both for the technical or the biological replicates. The GTEx study presents the same average correlation coefficients but a more extreme range. This may be explained by a strong batch effect as the samples were collected and sequenced at different times by different laboratories.

3.2.2 *Clustering analysis*

As we know the tissue type for each sample of each dataset, we may debate that supervised analyses can be more informative than unsupervised ones. However, they would involve proper corrections for batch effects and other technical biases for each dataset. This is challenging as it often requires more knowledge than what is available through the repositories. In Chapter 2, we have seen that is also unwise to rely solely on the normalised data provided by the original authors when working with various datasets that are non-uniformly processed[3].

To assess the consistency of the quantification across the different datasets, in particular for RNA-Seq, I picked a widely used unsupervised method for exploratory analysis in gene expression studies: clustering analysis. There are many available approaches and algorithms from which to pick; I chose a (bottom-up) *hierarchical* clustering (a.k.a. *connectivity-based* clustering). This sort of clustering is widely used in gene expression studies. Broadly speaking, this method groups samples by similarity in an extensive hierarchy, which allows uncovering possible hidden structures within the data; thus establishing if samples are more alike biologically or by study origin for instance.

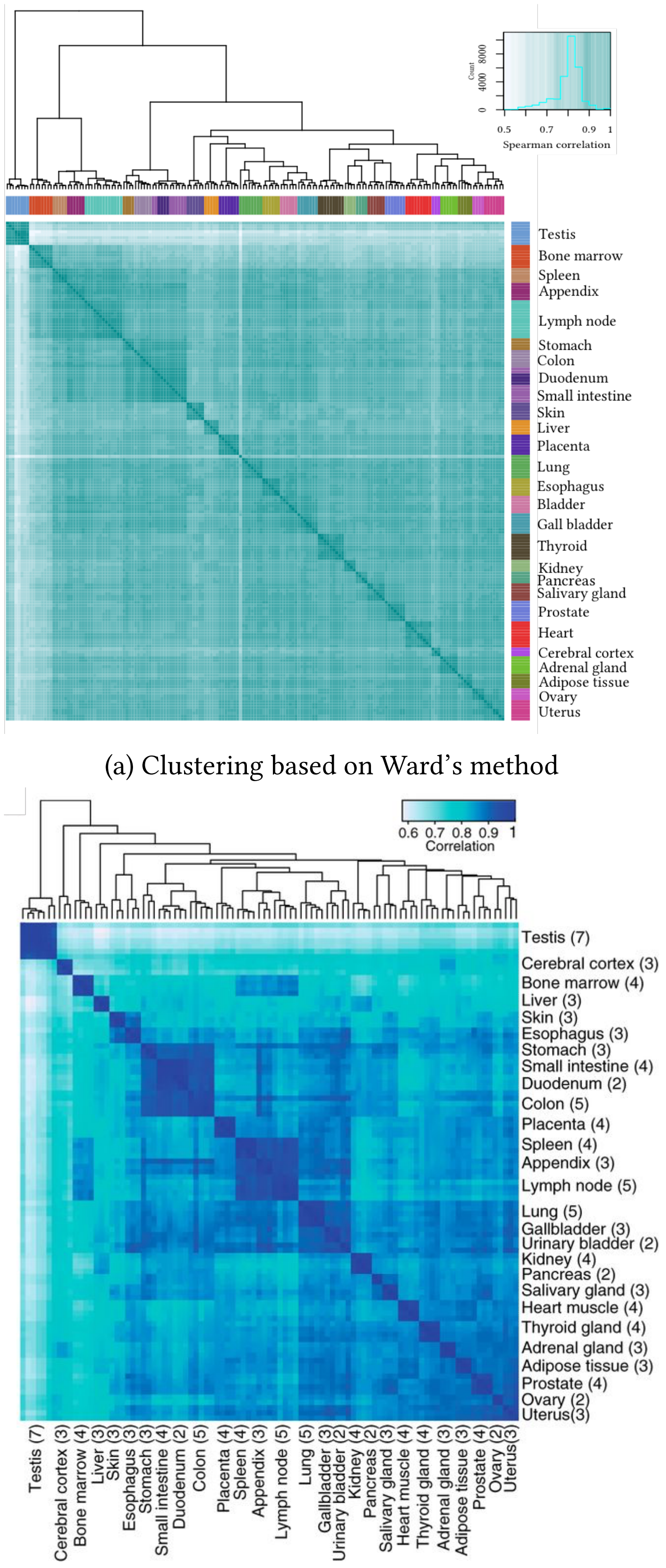
In general, we expect biology to be a better predictor when we only consider data from either transcriptome or proteome. Even more so if the identification technology and the quantification workflows are consistent. Yet, a technical predictor can not be directly excluded. Indeed, most transcripts (in particular mRNAs) are expressed in many tissues. Two tissues chosen at random share about 60 to 90% of their pool of mRNAs [Ramsköld et al., 2009; Gremel et al., 2015]. Parallelly, on the proteome side, M.-S. Kim et al. (2014) estimate that 75% of the mass of a cell is due to ubiquitous proteins and Wilhelm et al. (2014) estimate that about 10,000 to 12,000 proteins are ubiquitously detected, which represent about 60 to 75% of the proteins that they identified per tissue. Thus, if the variation of expression are too subtle from one tissue to another, a strong sample collection or data processing bias may hide any relevant biological signal.

In practice, each sample starts in its own cluster and then iteratively, each cluster is merged with its nearest one. The method has two parameters: the distance and the linkage method. Debate is still going on how to pick these parameters among the many possible choices (for more details see [Jaskowiak et al., 2014; Guinand et al., 2002]).

The distance measures the dissimilarity between two samples and one common approach is to calculate the subtraction result of the correlation coefficient from 1 (hence, a greater similarity between the two samples means a smaller distance). In analogy to previous analyses, I have also used both Spearman and Pearson correlation methods.

The linkage parameter specifies which part of each cluster is used as reference for computing the distance between the clusters. There are many methods and after trying

[3] Eventual bias corrections in RNA-Seq vary according to planed downstream analyses and proteomic data is hard to handle and two processing pipelines may rather give quite different results (see Section 5.2).



(a) Clustering based on Ward's method

(b) Clustering directly extracted from the original study [Fagerberg et al., 2014]

Figure 3.6. **Comparison of two clustering methods on a subset of the Uhlén study**. ((a)) includes all the samples and we observe that only a few of them are mixed with other samples from other tissues. This mixture is only observed between *Small intestine* and *Duodenum*.

several, I have arbitrarily selected the one that divides most accurately the samples by their tissue source across the different datasets. In fact, I noticed that Ward's method [Ward, 1963] was the best for this task and was outperforming the complete-linkage method[4].

Indeed, this latter method was used in [Fagerberg et al., 2014] (first release of the Uhlén dataset) where the authors have discarded a few samples as they were clustering incoherently in regards of their biological nature. Figure 3.6b presents the effect of the different clustering methods. Notice that all the tissue clusters are better defined when Ward's method is used: this method allows conserving all the samples for the analysis as long as other bias sources are corrected (see Section 3.3.1: Mitochondria issue).

3.3 reducing sources of bias

Many non-trivial methods correct for the skewness present in RNA-Seq and MS-based proteome global expression distributions and for other possible bias sources. For some examples, see Leek et al. (2010), Leek (2014), Yi, Raman, et al. (2018), S. Li et al. (2014), and Stegle et al. (2012).

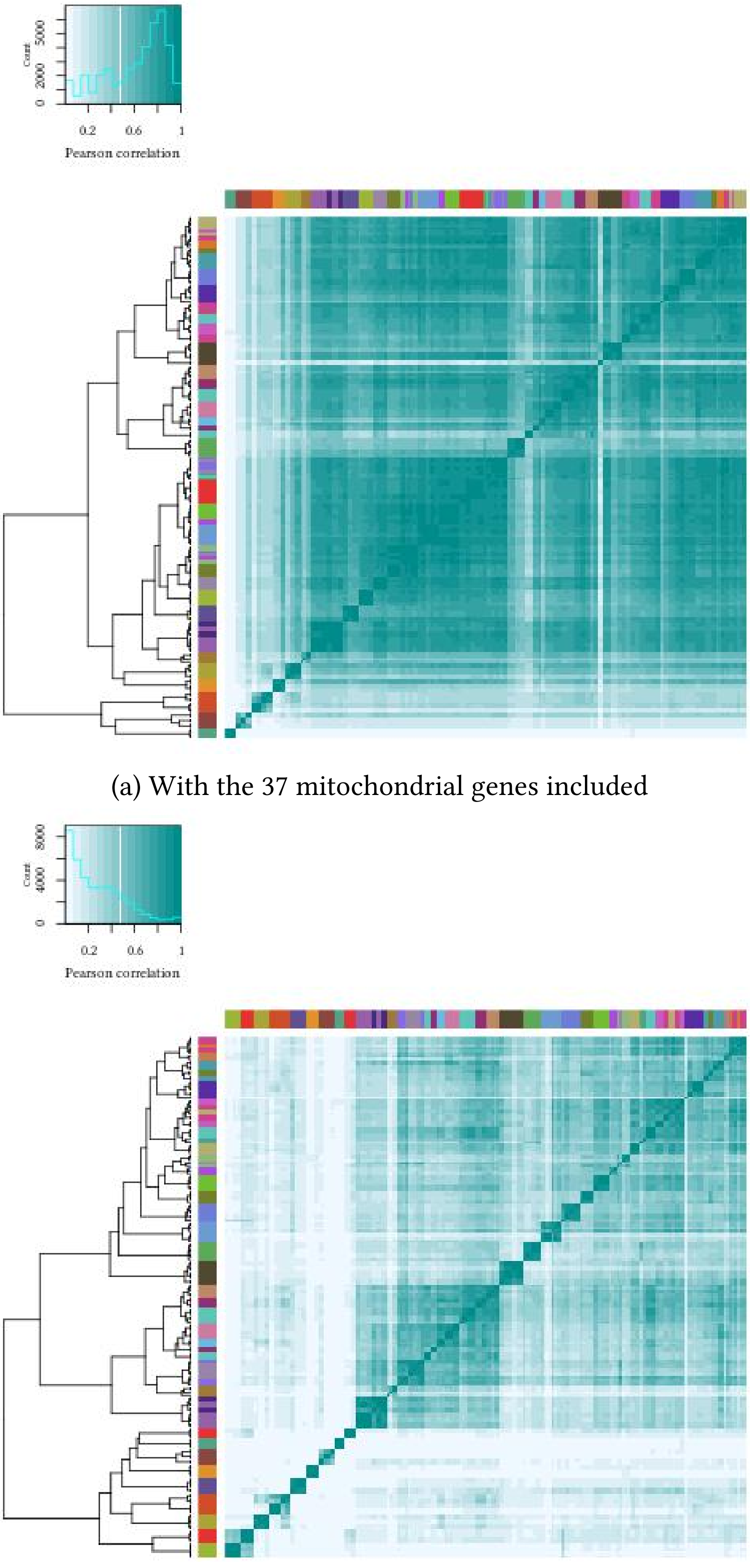
However, it may be complex to assess the biological relevance of those corrections. Moreover, many require more metadata that is often available in the public repositories. In the context of this thesis, I am interested in consistent traits across the datasets that may be consolidated into a reference. Thus, biases that are common to all the different included studies are in practice negligible. On the other hand, I have adjusted for a few easily avertible biases that I describe hereinafter.

3.3.1 *Mitochondria issue*

Mitochondria are organelles that can be found in eukaryotic cells. They have a central role in many essential processes [Kotrys et al., 2019]. While mitochondria share many similarities with bacteria[5], substantial divergences, notably for mammals, have been discussed in many reviews such as Boguszewska et al. (2020), Barshad et al. (2018), Hillen et al. (2018), Al-Faresi et al. (2019), Ladoukakis et al. (2017), and Shokolenko et al. (2017). One remarkable key difference is the polyadenylation of the RNAs. In bacteria, the polyadenylation of mRNAs prompts their degradation [Hajnsdorf et al., 2018; Rorbach et al., 2014]. On the other hand, the entire range of the polyadenylation effects is still

[4] Ward's method minimises at each step the variance within each cluster; the complete-linkage method (or *farthest neighbour clustering*) uses the maximum distance between the two farthest elements of each pair of clusters and merges the pair with the smaller inter-cluster distance.

[5] Despite the many debates and investigation going on about the lineages and mechanisms, current research accepts that mitochondria have evolved from an α-proteobacterial endosymbiont of a host cell prior to the last eukaryotic common ancestor (LECA) [W. F. Martin et al., 2015; Stairs et al., 2015].



(a) With the 37 mitochondrial genes included

(b) Without the mitochondrial genes

Figure 3.7. **Clustering of the biological samples of Uhlén study based on the Pearson correlation** — all expressed genes are included.

elusive for the human mitochondrial RNAs (mt-mRNAs) [Al-Faresi et al., 2019]. The polyadenylated tail has only a relative effect on the mt-mRNAs' stability [Bratic et al., 2016]. Besides, it is highly transcript-specific and can either stabilise them or flag them for degradation [Kotrys et al., 2019]. Most likely, its prime roles are the creation of a functional stop codon and the protection of the 3' side of the mt-mRNA against degradation [Bratic et al., 2016]. Except for *MT-ND6*, a polyadenylated tail has been observed for the twelve other mt-mRNAs. The extent of polyadenylation can vary across cell types [Kotrys et al., 2019]. However, the polyadenylated tail has an average length of 45 nt [Rorbach et al., 2014], which explains its possible captures by RNA-Seq (see Section 1.2.1).

Gene expression levels of mitochondria can report very useful information, *e.g.* the stress level of a cell in a single cell experiment [Ilicic et al., 2016]. However, it is unwise to keep them for a bulk analysis, particularly when comparing different biological sources. Indeed, it is very hard to properly normalise their expression; it involves knowing the amount of mitochondrial genome copies in the studied samples, while the mitochondria are from an unknown polyploidy and RNA-Seq protocols are badly suited for polyploid organisms [Pearce et al., 2015]. Thus, I have decided to remove them from the analysis as they skew anything relying on correlation. In fact, there are always mitochondrial genes among the highest expressed genes and they usually dominate manifolds the expression of the other genes.

Removing the (37) mitochondrial genes from the bulk of expressed genes (more than 10,000 protein-coding genes) produces more defined clusters as showed on Figure 3.7 (see also, in the next chapter analysis, Figure 4.3 in comparison of Figure C.4, where the simple exclusion of the mitochondrial genes allows all tissues to cluster by biological origin rather than the mixtures observed when they are kept). Figure C.14 illustrates furthermore the distinctiveness of the mitochondrial genes expression levels.

3.3.2 *Protein-coding genes only*

I have focused my analyses on the mRNAs (*i.e.* RNAs that have a biotype described as *protein-coding* in Ensembl 76).

In addition to the obvious reason to match with the proteomic data, most of the transcriptomic data is the product of poly-A selected protocols (see Section 1.2.1.2: RNA enrichment). Thus, aside from the mRNAs, all the other RNAs are off-target and, for many of them, their expression levels estimations may be highly imprecise.

3.3.3 *Expressed or not expressed*

While it can seem a trivial concept and might be overlooked, whether a specific molecule is truly expressed — or not — in a given condition, can actually have an extensive impact on the results of the analyses, particularly when integrating proteome and transcriptome

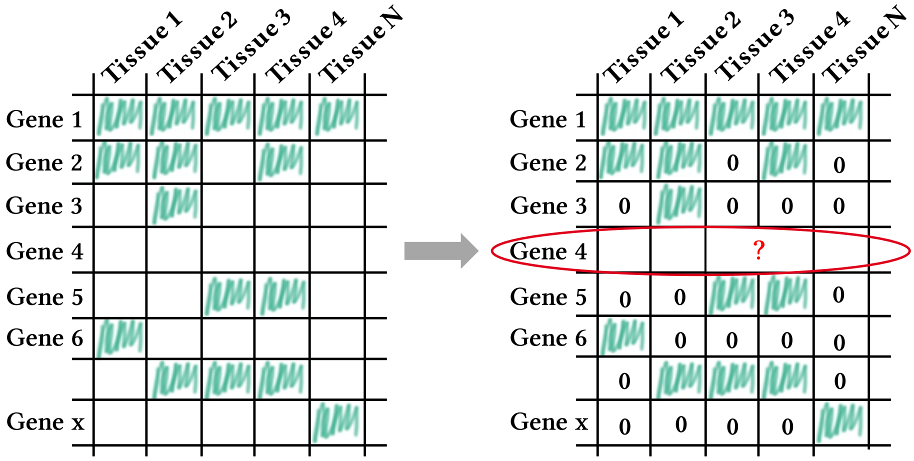


Figure 3.8. **Expressed or not: several cases illustrated.**

Genes like *Gene 1* are unequivocal: they have been detected in all the different tissues. Genes that have been quantified in *some* of the conditions are, in principle, detectable with the protocol of sampling and quantification used for the assay. For these genes, when no signal is collected, I assume this is a true 0 signal. In contrast, genes without any quantification in any tissue, *e.g. Gene 4*, are discarded from the remaining analysis as it is impossible to state whether they are truly absent from the biological sample or if it is due to the protocol used; they are *undefined*. The same approach is used for the transcriptome and the proteome.

together.

For example, the Pearson correlation coefficient is very sensitive to outliers and null values. If for both samples, a vast number of null values are recorded, this will lead to a greater similarity. Hence, it is important that the data used for the analysis is meaningful in its entirety, *i.e.* a null value is a truly an observation and translates to a lack of expression, rather than a lack of observation.

- **The undefined**: If a protein or transcript is never found in any of the samples of a dataset, then I considered that we can not determine if the protein or transcript was either truly not expressed or, for any reason, was not captured during the library preparation or the identification/quantification steps. Hence, those are excluded from the analyses as I can not resolve precisely if this is a technical artefact or a biological truth. An example is illustrated by the row circled in red in Figure 3.8.

- **Expression in a dataset**: By contrast, if a protein or a transcript is expressed in some samples of the dataset, then, whenever no expression was recorded in the other samples, I consider that the expression of the considered macromolecule is truly null for those samples.

- **Expression within a sample**: Due to the technical (and biological) differences between proteomics and transcriptomics, I use different thresholds to classify the presence of a protein or a transcript.

  - Expressed protein: On the proteomic side, I consider that a protein is expressed if it has been identified and quantified. In other words, a protein is expressed if the expression value is greater than zero in a sample. As described in Section 1.3.4.3, a protein identification and expression are inferred on a set of selected peptides. Thus, if the peptide selection changes, the identified proteins and their level of expression as well.

  - Expressed transcript: On the transcriptomic side, while the identification is direct, we have to account for technical noise, but also for 'transcriptional noise' [Z. Wang et al., 2009; Dar et al., 2015]. Indeed, SEQC/MAQC-III Consortium (2014) reports that excluding low-expression measurements reduce the FDR of RNAs considerably.
    While we can empirically evaluate it for each RNA-Seq dataset [Ramsköld et al., 2009], there is a widespread threshold used in the literature: 1 FPKM (or RPKM) — *e.g.* Fagerberg et al. (2014) and Uhlén, Fagerberg, et al. (2015). In fact, Hebenstreit et al. (2011) showed in their study *'RNA sequencing reveals two major classes of gene expression levels in metazoan cells'*, that to be detected and quantified at protein level, an mRNA should at least present an expression equals to 1 RPKM.
    As an important part of my thesis focuses on the comparison of proteomic and transcriptomic data (see Chapter 6), I have conducted all the analyses at least once with this threshold of 1 FPKM (I have also used 0 (*i.e.* using the same threshold for mRNAs for the proteins) and 5 FPKM as other thresholds for a few specific analyses).

Limitations of the study

While I have chosen to define proteins as *expressed* if only they present a non-null value, the truth is more complex.

One major challenge of bottom-up proteomics is the high rate of missing values. The detection of 10 to 50% of the expressed proteins can fail in a given study, and the proportion of a peptide/protein exhibiting a missing value at least once within the same study can reach 90% [Lazar, Gatto, et al., 2016]. As presented in Section 1.3.3, the detection of a protein is affected by the expression ranges of all the other proteins in the mixture. Thus, due to its relative abundance to other proteins in two given samples or tissues, the same protein can be detected in one and missed in another one while present in both; it can reach the MS detection limit in the first case but not in the second. Imputing the missing values is a widespread handling approach, for which, many algorithms have been developed and reviewed [Webb-Robertson et al., 2015; Välikangas et al., 2018a; Gardner et al., 2020].

For this thesis, I have chosen to not impute the missing data and exclude any mRNA or protein that is not expressed in at least one sample in *each and every* dataset used for the analyses. This led me to define different working sets for the following chapters analyses to limit the number of omitted mRNAs/proteins.

I have compared the list of undefined, expressed and unexpressed molecules. However, the bulk of the analyses has been done on the common expressed genes across the datasets.

3.3.4 *Aggregating tissue expression*

To deal with an unbalanced number of biological replicates across the datasets (see Section 1.5: Reproducibility and Experimental design), I computed a *'virtual' reference* for each tissue within the datasets that present more than one biological sample per tissue, *i.e.* Brawand, Uhlén and GTEx datasets. Note that, Castle, Cutler, Kuster and Pandey datasets present by design only one measure of expression per gene per tissue.

To compute the 'virtual' references for Brawand, Uhlén and GTEx datasets, I have taken the median value of each gene across all the biological replicates for each of their tissues.

Notes:

- For IBM dataset, I have discarded the single-end sequenced samples (as already mentioned in Chapter 2).
- The Uhlén dataset required an extra *a priori* step to the averaging of the biological replicates for some of the tissues as they present technical replicates. For these, I have first averaged the gene expression levels for each subject-tissue pair before computing the gene expression level medians of each tissue.
- The GTEx dataset required another *post*-processing step after the averaging of the biological replicates. Indeed, the samples are described based on their body site sources while the other datasets describe their samples only based on their tissue origin. Thus, while there is only *Heart* samples in Castle, Brawand, IBM, Uhlén, Cutler, Kuster and Pandey, GTEx has samples from the *left ventricle of the Heart* and from the *Atrial appendage of the Heart.* For this case and other similar ones, I have average the virtual reference of the body sites in GTEx that I considered relevant for comparison with tissues found in the other datasets.
- While I have detected many samples that seem outliers to their biological replicates within the same datasets, I have decided to keep all the samples for the tissues expression averaging step. On this matter, in many scientific exchanges, I was repetitively asked about the inclusion of all the GTEx dataset samples related to *Oesophagus* for the averaging step. Indeed, while two of the three GTEx's body sites (*i.e. Gastro oesophageal junction* and *Oesophageal muscularis*) present great similarities together ($r = 0.99$) and more modest correlations to Uhlén's *Oesophagus* samples ($r < 0.80$), the last body site (*i.e. Oesophageal muscularis*) expression, very dissimilar to the two former ones ($r < 0.64$), presents higher correlation to Uhlén's ($r = 0.94$). Thus, while only considering this latter body site

significantly improves the overall *Oesophagus* correlation between GTEx and Uhlén, I have decided to include all three body sites (as one tissue) in my study. Indeed, there are no suitable reasons or information that allows excluding any of the body site *prior* to the analyses; gene expression scatter plots of the Uhlén's samples versus any of these three body sites present two trends which suggest that Uhlén's *Oesophagus* samples are composite.

The meta-analyses of the following chapters use a TREP (tissue reference expression profile) for each tissue of each study, *i.e.* there is only one measure per gene for each tissue. These measures are either the primary sample expressions for Castle and IBM studies or these *'virtual'* constructed *references* for Brawand, Uhlén and GTEx.

3.4 DISCUSSION AND CONCLUSION

In this chapter, I have reviewed many fine quality control points that may be (and often are) overlooked. These details have critical impact on the results of the analyses I perform and more gravely on their interpretations[6].

Aside the datasets selection criteria, this phase is by far the most subjective one of the whole thesis. Hence, to avoid excessive *data cleaning* and possible cognitive biases, I have formulated the aforementioned filtering rules that are important but simple.

Overall, I am quite stringent and I have preferred to keep more data unless there is a strong rationale to discard them. Therefore, there are sharper filters and samples exclusion options that may easily improve the results I present in this thesis.

[6] For more, see *'The devil in the details of RNA-seq'* [Kratz et al., 2014] and the included references.

So far, I think it's been working. But who knows?
Mark Watney [Weir, 2014]

4 INTEGRATING GENE EXPRESSION DATA FROM UNDISEASED TISSUES ACROSS RNA-SEQ STUDIES

To pave the way towards a generalised baseline expression reference for the *normal* (*i.e.* non-disease) human, I assess in this chapter the similarity of the tissues sourced from different RNA-Seq studies and the general profiles of their expressed genes. When I started this project in 2013, little was known of either the robustness or the shortcomings and pitfalls of RNA-Seq and its related processing of output data. Since then, several studies were published assessing RNA-Seq robustness (See Appendix C.4). A few are close in scope to my own investigations. Thus whenever relevant, I introduce and discuss my results in relation to the published ones.

In the first part of this chapter, I introduce the datasets based on the transcriptome studies (described in Chapter 2) that I use in the different meta-analyses. In the second part, I appraise the congruence of the interstudy tissue expression profiles. Then, I examine different components that may contribute the most to (and thus explain) the overall strong biological correlations that are observed between the studies' tissues. Finally, I explore the interstudy consistency of the gene expression profiles.

All the work presented in this chapter was performed by myself under the supervision of Dr Alvis Brazma. I received invaluable advice and help from my discussions with Dr Nuno Fonseca. I also received general feedback and comments from Dr Mar Gonzàlez-Porta, Dr Johan Rung, Dr Sarah Teichmann, Dr Gos Micklem and Dr Wolfgang Huber.

Communication to the community derived from this chapter

- (paper) R. Petryszak, M. Keays, et al. (2015). 'Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants'. *Nucleic Acids Res.* 44 (D1), pp. D746–52
- (short talk) Quantitative Genomics 2015 — Integration of independent human RNA-seq datasets: a feasibility study
- (poster) ECCB 2014 — A feasibility study: Integration of independent RNAseq datasets
- (poster) SymBLS 2014 — Integration of independent human RNAseq datasets, a feasibility study
- (invited short talk) GM^2 2013 — Baseline Gene expression Atlas
- (flash talk) CSAMA 2013 — How quantitative is RNA-seq?

In the past years, RNA-Seq rapidly gained popularity for human gene expression studies due to a broader dynamic range than previous technologies and the promise to enable quantitative profiling [J. C. Marioni et al., 2008]. That technology was an advancement with respect to microarray assays that are semiquantitative [M.-L. Lee, 2006] and very prone to batch effects [Irizarry, Warren, et al., 2005]. However, RNA-Seq studies had shown variation in their conclusions on various occasions for similar research topics [SEQC/MAQC-III Consortium, 2014]. At the time that I started my Ph.D., it appeared that RNA-Seq might share at least partially the problems encountered with microarray assays. In fact, *batch effects* restrain the use of direct approaches for the comparison of independent microarray data, and the resulting insights are usually limited [Walsh et al., 2015; Chrominski et al., 2015; Rung et al., 2013; Lazar, Meganck, et al., 2013].

The following meta-analyses attempt to provide more insights into the interstudy RNA-Seq robustness for tissues expression as a supporting exploratory study to the EBI Gene Expression Atlas [Petryszak, Keays, et al., 2015].

4.1 META-ANALYSES' COMBINED DATASETS

Through this chapter meta-analyses, I use two sets based on combined subsets of the transcriptomic studies introduced in Chapter 2.

The following Sections 4.1.1 and 4.1.2 illustrate the construction of these sets.

While many approaches exist, I usually consider the most stringent routes, *i.e.* I rather exclude part of the data to infer conclusions than keep wider datasets and more partial, biased or ambiguous results. Thus, I identified the identical core of explored tissues and expressed genes across the studies. From this base, I created two more robust combined datasets ($\mathcal{W}_1$ and $\mathcal{W}_2$) for the meta-analyses.

4.1.1 *Tissue overlaps across available normal human RNA-Seq studies*

Figure 4.1 presents the tissue overlap between the five studies. All studies share at least four tissues: *Heart*, *Kidney*, *Liver* and *Testis*. This 4-tissue set is the base of the first combined dataset ($\mathcal{W}_1$).

The greatest number of shared tissues occurs between the two most recent studies, Uhlén [Uhlén, Fagerberg, et al., 2015] and GTEx [Melé et al., 2015]. This 23-tissue set is the base of the second combined dataset ($\mathcal{W}_2$) and includes *Adipose tissue*, *Adrenal gland*, *Bladder*[1], *Cerebral cortex*, *Colon*, *Oesophagus*, *Fallopian tube*, *Heart*, *Kidney*, *Liver*, *Lung*, *Ovary*, *Pancreas*, *Prostate*, *Salivary gland*, *Skeletal muscle*, *Skin*, *Small intestine*, *Spleen*, *Stomach*, *Testis*, *Thyroid* and *Uterus*.

[1] May also be referred to as *Urinarybladder*

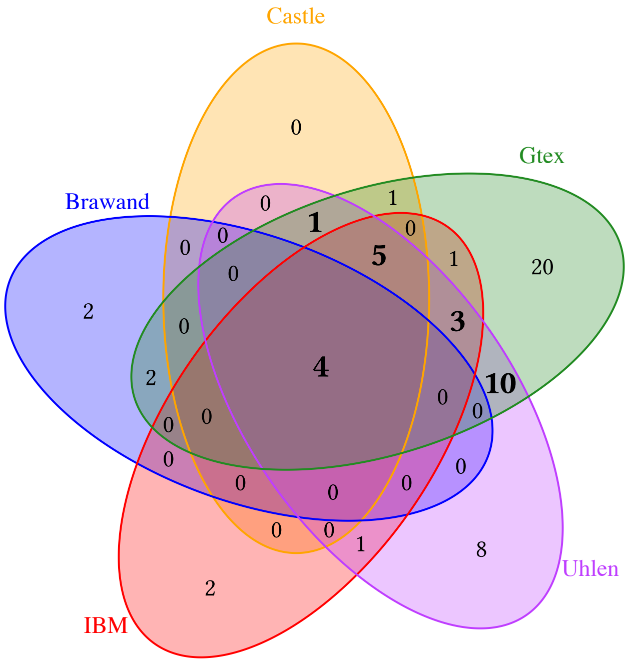


Figure 4.1. **Distribution of unique and shared tissues between the transcriptomic studies.** The five studies share four common tissues: *Heart*, *Kidney*, *Liver* and *Testis*. The most prominent overlap of tissues (23) is between Uhlén and GTEx.

4.1.2 *Common measured genes for each of the main shared-tissue sets*

In the following sections of the thesis, I only present the results based on the *HTSeq-count* quantification.

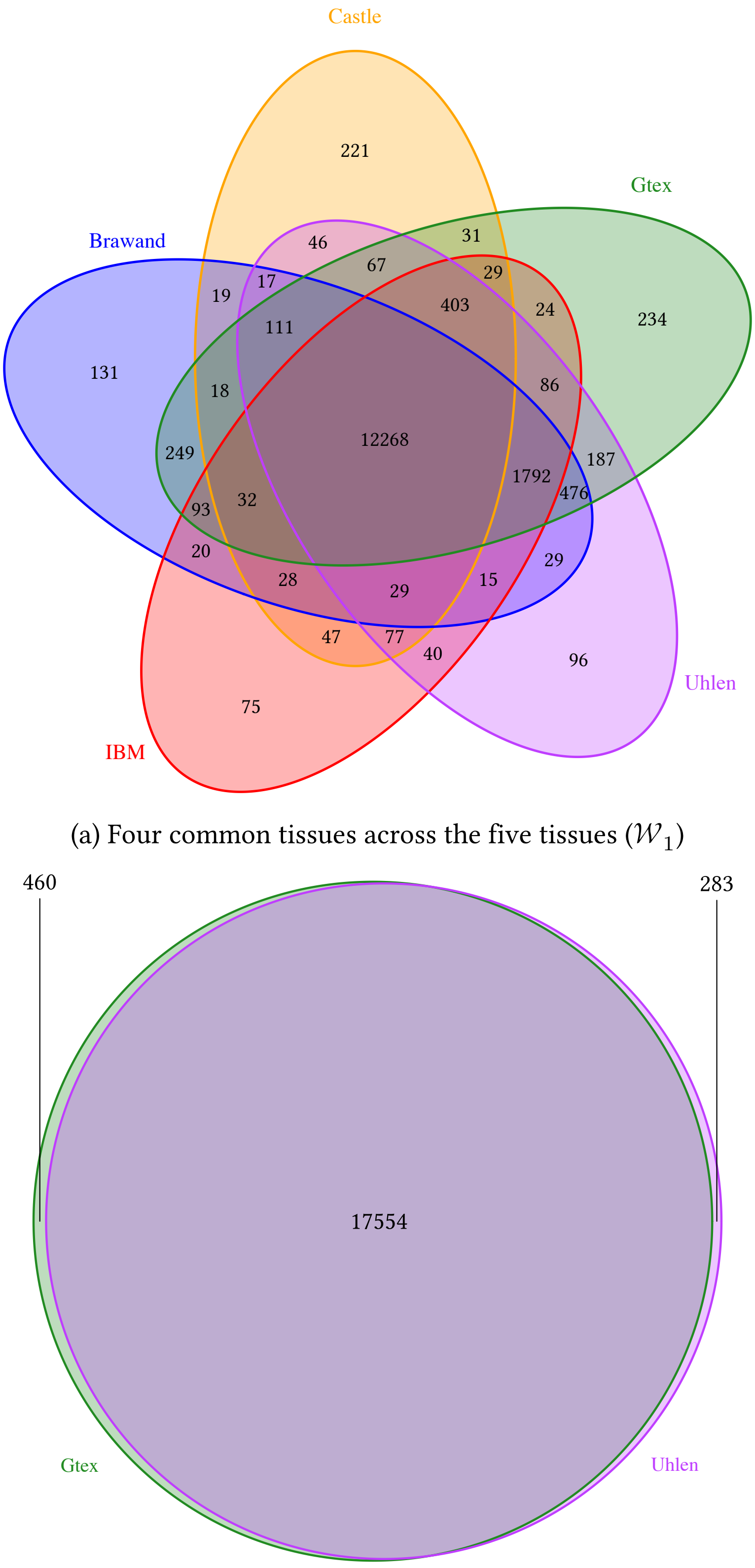
As shown in Table 2.1 (p. 57), many of the transcriptomic studies I use were produced through polyA-selected library protocols. Hence, to avoid unnecessary biases[2], I have limited my analyses to protein-coding genes (Ensembl 76). All mitochondrial genes have been filtered out before any TREPs analysis (as specified in Section 3.3.1).

The Venn diagram presented in Figure 4.2a only includes protein-coding genes that are observed[3] at least once at 1 FPKM for one of the four shared tissues. As mentioned in the previous chapter (Section 3.3.4 p. 87), the bulk of expressed genes at this threshold is common to all five studies. While each study presents a tiny portion of genes that are unique, overall most genes are detected in at least two studies. The most considerable contingent of shared gene expression is observed between Uhlén and GTEx.

Figure 4.2b presents a similar Venn diagram which focuses on the set of twenty-three tissues ($\mathcal{W}_2$) between Uhlén and GTEx studies. The uniquely expressed genes in each study are negligible compared to the overlap. They represent at most 0.03% of the measured

[2] See Section 3.3.2: Protein-coding genes only.

[3] See Section 3.3.3: Expressed or not expressed.



(a) Four common tissues across the five tissues ($\mathcal{W}_1$)

(b) Twenty-three common tissues between Uhlén et al. and GTEx studies ($\mathcal{W}_2$)

Figure 4.2. **Unique and shared protein-coding genes expressed (≥ 1 FPKM) across the RNA-Seq studies for $\mathcal{W}_1$ and $\mathcal{W}_2$**

genes in each of the studies.

I have also analysed all the other subgroups of genes (*i.e.* unique to each study or shared only between two to four studies) for any functional annotation enrichment (see Section 1.4). No analysis provided any conclusive result.

4.1.3 *Combined datasets summary*

I have based all the meta-analyses of this chapter on the two $\mathcal{W}_1$ and $\mathcal{W}_2$ datasets, which are defined as follow:

$$\mathcal{W}_1 : \mathcal{D}_{\text{Trans}_1} \times \mathcal{G}_{\text{protein coding}_1} \times \mathcal{T}_1$$

and

$$\mathcal{W}_2 : \mathcal{D}_{\text{Trans}_2} \times \mathcal{G}_{\text{protein coding}_2} \times \mathcal{T}_2$$

where:

- $\mathcal{D}_{\text{Trans}_i}$ is a set of mRNA expression studies (presented in Section 2.2). With $\mathcal{D}_{\text{Trans}_1}$ = { Castle, Brawand, IBM, Uhlén, GTEx} and $\mathcal{D}_{\text{Trans}_2}$ = { Uhlén, GTEx}
- $\mathcal{G}_{\text{protein coding}_i}$ is a set of genes g_{pc} that are shared by all elements of $\mathcal{D}_{\text{Trans}_i}$ and have a biotype described as *protein coding* in Ensembl 76. $\mathcal{G}_{\text{protein coding}_1}$ comprises 12,268 protein-coding genes. $\mathcal{G}_{\text{protein coding}_2}$ comprises 17,551 protein-coding genes.
- $\mathcal{T}_i$ is a set of tissues that are shared by all elements of $\mathcal{D}_{\text{Trans}_i}$. $\mathcal{T}_1$ includes four tissues and $\mathcal{T}_2$ twenty-three tissues.

Note that as stated in Section 3.3.4, to avoid an unbalanced number of samples per tissues across studies, I aggregate into a single virtual reference, *i.e.* TREP, the gene expression measured for each gene for each tissue in each study, regardless of the number of replicates. Thus, $\mathcal{W}_1$ comprises 20 TREPs, and $\mathcal{W}_2$ 46 TREPs.

4.2 PREVALENCE OF BIOLOGICAL SIGNAL OVER TECHNICAL VARIABILITIES AT TISSUE-LEVEL

As shown in Chapter 3, the expression levels of biological replicates (*i.e.* identical tissue samples) are highly correlated within the same study and allow one to group the samples based on their biological source. Thus, clustering the samples across studies should offer a rough assessment of the underlying driving forces for the observed gene expression levels. A clustering by study would mean that technical variabilities are stronger than any biological expression signature (which is an actual recurrent observation with microarray studies due to their strong batch effects [Sudmant et al., 2015]). On the other hand, an interstudy sample clustering by tissue would imply that RNA-Seq measurements demonstrate a good (biological) signal over (technical) noise ratio. In other words, RNA-Seq would be then less prone to batch effects and more robust than microarray assays [Taminau et al., 2014; Walsh et al., 2015].

The heatmaps of the hierarchical clustering of the TREPs[4] for $\mathcal{W}_1$ and $\mathcal{W}_2$ are respectively presented in Figures 4.3 and 4.4. They are based on clustering (Ward's method linkage [Ward, 1963]) the TREPs' Pearson correlation coefficients (protein-coding genes expressed at least at 1 FPKM).

Both heatmaps show that the overall biological signal measured in the tissues is stronger than the noise generated by any technical variation or batch effect.

In Figure 4.3, each cluster corresponds to a tissue. The clustering signal by tissue dominates over the signal from the dataset. It highlights a greater biological similarity of the TREPs due to their sampling origins rather than any possible technical similarity due to laboratory protocol variations.

One may object that the very different gene expression landscapes of *Heart*, *Kidney*, *Liver* and *Testis* [Ramsköld et al., 2009; Lukk et al., 2010; Danielsson et al., 2015; Sudmant et al., 2015; Melé et al., 2015; Uhlén, Fagerberg, et al., 2015] may drive this result and lesser differentiated tissues may exhibit more mitigated correlations. Figure 4.4 ($\mathcal{W}_2$) confirms that the biological origin of the tissues is the dominant criterion for the clustering of the TREPs.

Moreover, in many cases, TREPs mixtures occur in close biologically related tissues, *e.g.* *Fallopian tube* and *Ovary* or *Salivary gland* with *Oesophagus* and *Stomach* TREPs. The functional proximity of these tissues is likely supported by an overall similarity in their gene expression. Thus, even though there are clear biological substructures emerging like for *Heart* and *Skeletal muscle*, without correction, the biological signal to technical noise ratio for close tissues may be insufficient to discriminate them accurately in every case.

The general observed biological prevalence holds when I extend the analysis to include all the available tissue samples (see Figure C.5 and Figure C.6). See also Section 2.2: Transcriptome RNA-Seq studies (p. 56) and Table B.1: Correlation coefficients between RNA-Seq replicates (p. 198).

Figure 4.5 shows the distribution of the Pearson correlation coefficients for the pairs of the profiles (TREPs) of tissues with the same name sourced from the different studies for both of the combined datasets $\mathcal{W}_1$ (4-tissue set) and $\mathcal{W}_2$ (23-tissue set). Even with the lack of any batch effect correction, most of the Pearson correlations are above 0.5. There are two exceptions: the correlation between the *Testis* TREPs of Castle and Brawand ($r = 0.42$) from $\mathcal{W}_1$ and *Salivary gland* TREPs of Uhlén and GTEx ($r = 0.2$) from $\mathcal{W}_2$. The median correlation for $\mathcal{W}_1$ is about 0.7 and 0.84 for $\mathcal{W}_2$. Spearman correlation gives even better

[4] Tissue reference expression profile. See Section 3.3.4: Aggregating tissue expression.

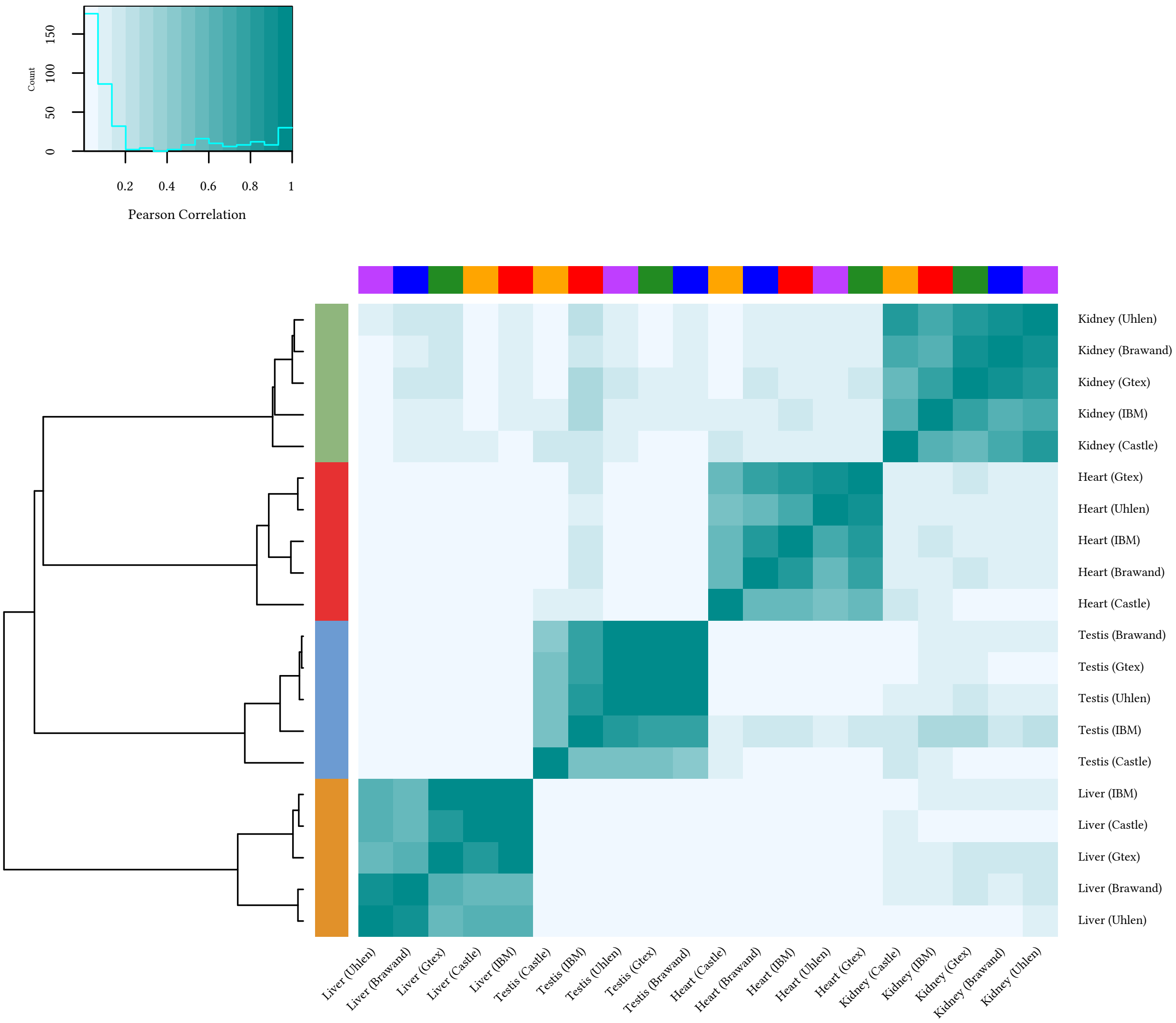


Figure 4.3. **Heatmap of the four common tissues across the five studies.**
All protein-coding genes (except the mitochondrial ones) at least expressed at 1 FPKM are included.
All the different TREPs cluster by tissue of origin regardless of their study sources. Each of the colours on the top bar following the x-axis is associated to one of the study (purple for Uhlén, blue for Brawand, green for GTEx, orange for Castle and red for IBM), and the colours on the side bar following the y-axis are associated to the tissues (green for *Kidney*, red for *Heart*, blue for *Testis* and orange *Liver*).

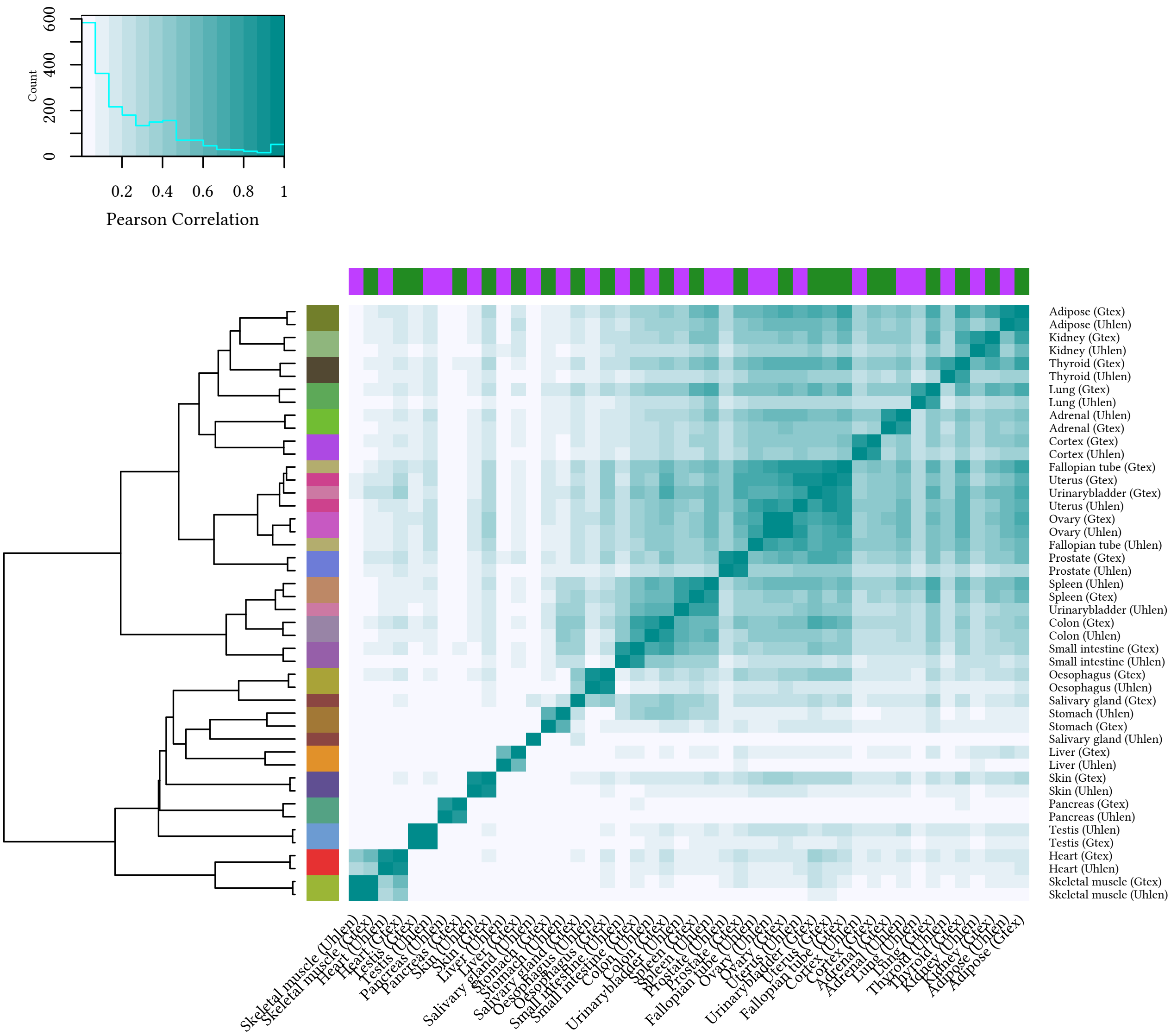


Figure 4.4. **Heatmap of twenty-three common tissues between Uhlén and GTEx studies.** All protein-coding genes (≥ 1 FPKM with the exclusion of the mitochondrial genes) are included.
Most TREPs cluster by tissues (y-axis colour bar) rather than by study of origin (x-axis colour bar) with a few exceptions: there is a mixture of the *Fallopian tube* and *Ovary* TREPs. In addition, *Salivary gland* TREPs is more correlated to *Oesophagus* or *Stomach* regarding the original study. *Urinarybladder* TREPs seem to cluster randomly with the others. However, these TREPs are in singleton groups.

results: average correlation coefficients are $\rho = 0.49$ for $\mathcal{W}_1$ and $\rho_{\mathcal{W}_2} = 0.9$; the median correlations are $\rho_{\mathcal{W}_1} = 0.88$ and $\rho_{\mathcal{W}_2} = 0.93$.

Both the Pearson[5] and the Spearman correlation coefficients for the more exhaustive $\mathcal{W}_2$ set, which comprises the two most recent studies, are higher than the observed correlations for $\mathcal{W}_1$. Three main reasons may explain this situation as they contribute to lower the technical variations:

- In addition to using paired-end sequencing, the library preparation protocols were better established for these two more recent studies;
- The instrument used for the sequencing were from the same series (HiSeq 2000 and HiSeq 2500); and
- These studies present a higher number of replicates per tissue.

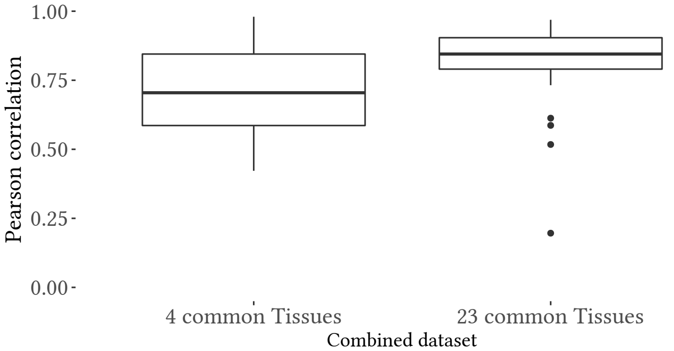


Figure 4.5. **Distribution of the Pearson correlation coefficients of same tissues pairs for the four and the twenty-three tissues combined datasets.** In general, the Pearson correlations are high when we are *directly* comparing TREPs from different studies.
The same-tissue pairs in the 23-tissues combined dataset ($\mathcal{W}_2$) present a higher median correlation (0.85) and narrower distribution than in the 4-tissues combined dataset ($\mathcal{W}_1$) (median= 0.74). However, $\mathcal{W}_2$ displays one outlier with a very low Pearson correlation (0.2: *Salivary gland* tissue). Sampling, processing differences or biological reasons may just as well explain this outlier.

On the other hand, the pairs comprising different tissues are very lowly correlated in general (see Figure C.7). Although, in a few cases of $\mathcal{W}_2$ (23-tissues combined dataset), high correlations are also observed for different-tissues pairs, *e.g. Fallopian tube* and *Uterus* from GTEx study (see also Figure 4.4). It is rather hard to decipher if this may be due to a technical issue (*e.g.* at the collection or library preparation stage) or because these tissues are biologically very close.

As the exclusion of the *undefined*[6] genes from the analyses handles one possible source of spurious Pearson correlation (due to null values), I have then checked, and, discarded,

[5] Despite one major outlier in the second combined dataset (*Salivary gland* — Pearson correlation: r=0.2)

[6] I.e. *unobserved* — See also Section 3.3.3: Expressed or not expressed

another possible source that is the skewed distributions; highest expressed genes may be technical artefacts, but they fail to show any significant correlation (see Appendix C.1).

The high correlations are the results of the overall similarity of genes expression patterns across tissues and studies.

4.3 global stability of gene expression profiles across studies

After validating that RNA-Seq allows distinguishing the shared biological origin of most tissues (TREPs) across different studies, the question then arises as to how consistent is the expression of each gene for a given tissue between studies.

To this aim, I first assess the expression variability of the genes across the studies. Then, I explore the interstudy coherence of several gene categories. I focus on the tissue-specific (TS) genes, and on a larger number of categories developed upon classifications from Uhlén's laboratory.

4.3.1 *Genes with tissue-specific (TS) expression*

Tissue-specific (TS) genes are arguably the genes that ought to present a robust expression profile across studies.

Tissue specificity definition

The definition of tissue specificity varies from one study to another. See also Santos et al. (2015). Liang et al. (2006) that define *tissue specificity* only for genes expressed solely in one tissue, and then *tissue selectivity* for genes expressed in more than one tissue with an expression enriched in one or a subset of tissues. Other studies have a broader definition of tissue specificity. They identify genes above a given threshold of tissue selectivity (or enrichment) as tissue-specific genes (*e.g.* Fagerberg et al. (2014) and C. Jiang et al. (2016)). In this second case, genes with a single-tissue expression are an extreme case of Tissue-Specific (TS) genes.

Within this thesis, I use the second definition, *i.e.* I consider genes as TS as long as they display a higher tissue selectivity than a preset threshold, regardless of how many tissues express them. Indeed, every study presents a subset of genes that are expressed above 1 FPKM in a sole tissue (see Figure C.2). However, the decreasing number of these genes when increasing the number of considered tissues highlights the arbitrary relativity introduced by the study design.

For this definition, there are many methods to characterise genes tissue-specificity (*e.g.* Cavalli et al. (2011), Xiao et al. (2010), Karthik et al. (2016), P. Kim et al. (2017), Kryuchkova-Mostacci et al. (2017), Kadota et al. (2006), X. Yu et al. (2006), and Martínez et al. (2008)).

There are also databases that record previously identified TS genes, for normal conditions, *e.g.* *TiGER*[7] [X. Liu et al., 2008] or *TiSGeD*[8] [Xiao et al., 2010] and more specialised ones, *e.g.* for cancer *TissGDB*[9] [P. Kim et al., 2017].

TS genes characterisation approaches used in this thesis

From the possible approaches to characterise the TS protein-coding genes, I detail three that I used in the following subsections. First, I have queried TiGER to capitalise on previous knowledge. Then, to derive the TS genes directly from $\mathcal{W}_1$ and $\mathcal{W}_2$, I have used a published method, that uses the gene expression *fold change* (FC) ratio across the tissues. Finally, I have employed a robust method designed to detect outliers in data, *i.e.* Hampel's test [Hampel, 1974], to identify genes which present an unusual expression level in a single tissue. In fact, as gene tissue selectivity and tissue specificity definitions are relative to a context, if the latter changes, the genes attributes may change as well (*e.g.* one gene that is non-specific in $\mathcal{W}_2$ may be *Heart*-specific in $\mathcal{W}_1$).

4.3.1.1 *Use of prior knowledge: TiGER database*

TiGER [X. Liu et al., 2008] reports TS genes for thirty independent tissues (based on ESTs experiments).

After retrieving the list of genes for all reported tissues, I have mapped the RefSeq identifiers provided by TiGER to Ensembl gene identifiers (GRCh38, Ensembl 76). Then, I removed all duplicates due to the identifier translation within each tissue, and I also filtered out all the genes identifiers that I found in more than one tissue: TiGER lists a subset of the same genes in many tissues, but the modification in the annotation may also explain part of the repetitive genes. Thus, for each tissue, I have a list of identifiers that are specific to that tissue only.

Figure 4.6 is an expression heatmap based on a subset (*i.e.* 916) of protein-coding genes that are present in this final list of translated TiGER genes for the *Heart*, *Kidney*, *Liver* and *Testis*. There are three main types of genes. The largest group comprises the genes with a corroborating profile between the TiGER definition and the real data. Then, a second smaller group encompasses genes listed as TS in TiGER, but fails to demonstrate expression specificity towards any tissue in the real data. Finally, the third group includes a very tiny subset of genes which are more specific to another tissue than the initially stated one.

Thus, without any additional knowledge, it is difficult to predict beforehand which original TiGER definitions will be confirmed or rejected by expression data. Remarkably, most of the genes present the same expression pattern through the tissues across each of the studies and thus regardless of their TiGER category. Once again, Castle expression data is

[7] *TiGER* — http://bioinfo.wilmer.jhu.edu/tiger/
[8] *TiSGeD* — http://bioinf.xmu.edu.cn:8080/databases/TiSGeD/index.html
[9] *TissGDB* — https://bioinfo.uth.edu/TissGDB/index.html

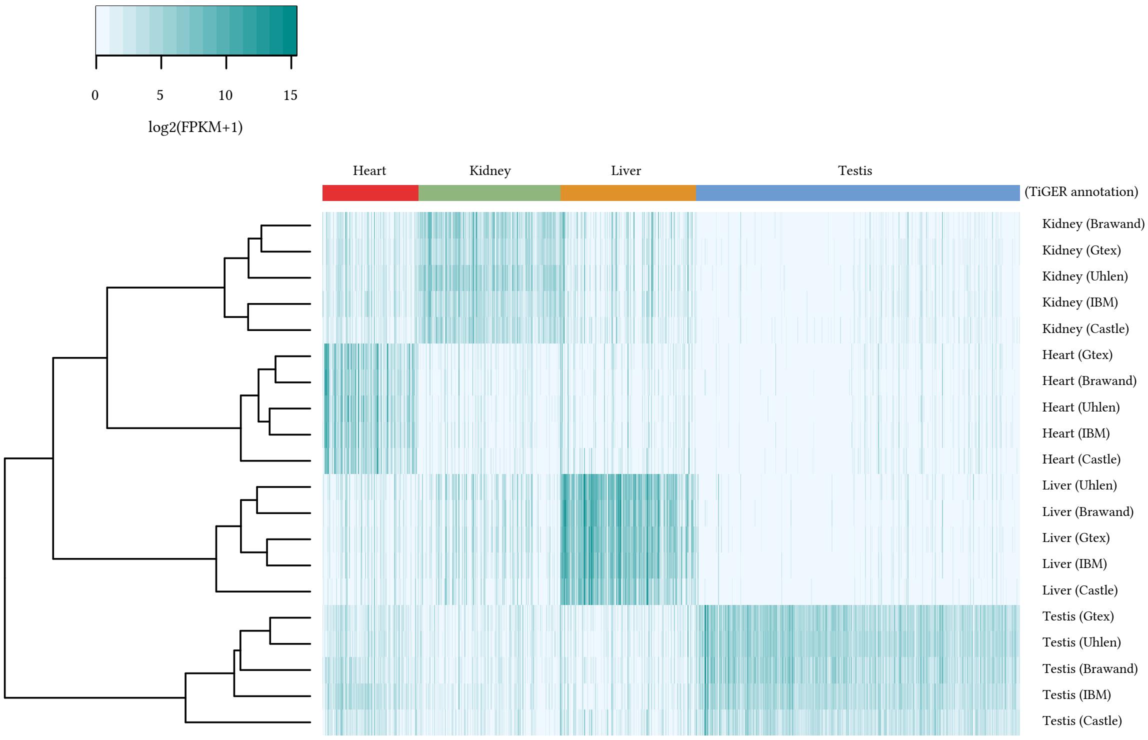


Figure 4.6. **Expression heatmap of the four tissues across the five datasets based on TiGER information.** This heatmap illustrates three subsets of genes: genes for which real expression data confirm their TiGER definition; genes failing to show any TS profile in their expression data; and genes with mismatching tissue specificity between TiGER definition and the expression data. The colourbar above the heatmap is representing the tissue for which TiGER annotates the genes (presented as columns) as TS (red for *Heart*, green for *Kidney*, light orange for *Liver* and blue for *Testis*). TiGER definitions are of variable accuracy.

exhibiting the only few observed discrepancies[10].

4.3.1.2 *Fold change method*

As Love et al. (2014) noted the most common approach for detecting a gene expression difference between two conditions is to study the expression fold change (FC) ratio between these conditions. This method is still broadly present in the literature, especially for studies other than differential gene expression analyses[11]; as examples, see Uhlén, Fagerberg, et al. (2015), Zhu et al. (2016), and N. Y.-L. Yu et al. (2015). Besides, EBI Gene

[10] Reminder: the FPKM quantification (used here) is sensitive to the number of identified genes (see Equation (Canonical F/RPKM formula) equation (Canonical F/RPKM formula) on page 24) and Castle study identifies and quantifies many more RNAs than the other studies as it uses a whole RNA protocol while the others are using polyA-enrichment (see Chapter 2).

[11] Studies comparing gene expression of a treated or diseased condition to control samples

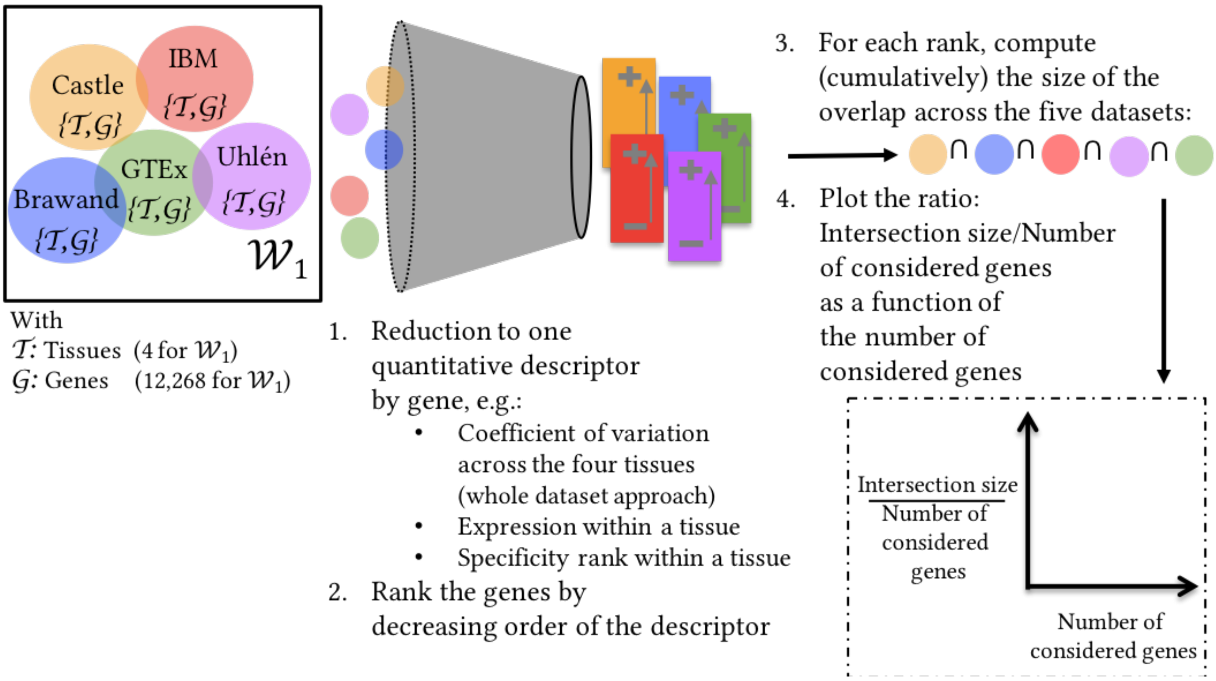

Castle {T,G}
IBM {T,G}
GTEx {T,G}
Uhlén {T,G}
Brawand {T,G}
$\mathcal{W}_1$
With
$\mathcal{T}$: Tissues (4 for $\mathcal{W}_1$)
$\mathcal{G}$: Genes (12,268 for $\mathcal{W}_1$)
1. Reduction to one quantitative descriptor by gene, e.g.:
• Coefficient of variation across the four tissues (whole dataset approach)
• Expression within a tissue
• Specificity rank within a tissue
2. Rank the genes by decreasing order of the descriptor
3. For each rank, compute (cumulatively) the size of the overlap across the five datasets:
● ∩ ● ∩ ● ∩ ● ∩ ●
4. Plot the ratio: Intersection size/Number of considered genes as a function of the number of considered genes
Intersection size / Number of considered genes
Number of considered genes

Figure 4.7. **Overview for the comparison of the genes across the five studies based on a ranked descriptor.** The first step applies individually to each of the studies within the combined dataset (*i.e.* here $\mathcal{W}_1$). It consists in extracting a single value per gene (*e.g.* a statistic or any other quantitative descriptor) either for the entire *d*ataset (referred thereafter as *D-approach*) or for each *t*issue in each dataset (referred as *T-approach*). The next steps include computing (cumulatively) the intersection size number for each rank and plotting this number divided by the rank as a function of the number of considered genes (*i.e.* rank).

Expression Atlas [Petryszak, Keays, et al., 2015] relies on this method to select the most specific genes for *baseline* studies[12] (see Figure C.22). There are also a few variations on how to compute this ratio; see Zhu et al. (2016) and Uhlén, Fagerberg, et al. (2015).

In this thesis, I compute the FC ratio by dividing the expression of each gene in a given tissue by the average expression of this gene across all the other tissues of that study in the combined dataset.

$$\mathcal{FC}_{g,t,d} = \frac{x_{g,t,d}}{\frac{1}{n}\sum_{i=1}^{n} x_{g,i,d}} \qquad \text{(Fold change (FC) ratio)}$$

where:
- x is the expression value of the gene g in the tissue t in a study d
- n is the number of tissues t

[12] In contrast to differential gene studies, the baseline studies focus on depicting the expression landscape of each covered condition instead of focusing on the gene expression through these conditions.

Studies usually pick arbitrary cut-offs to characterise the specific genes. Uhlén, Fagerberg, et al. (2015) uses two-fold and five-fold cut-offs to determine *enriched* and *enhanced* tissue genes. Zhu et al. (2016) set their cut-off at 2 to characterise TS protein-coding and noncoding transcripts. However, I avoid arbitrary cut-offs, and I use the FC ratio to rank the protein-coding genes of my combined datasets according to their specificity within each tissue: higher FC ratios indicate genes with higher specificity. I then assess the consistency of the tissue specificity of the genes through the various studies. For that, I have followed the *T-approach* overviewed in Figure 4.7.

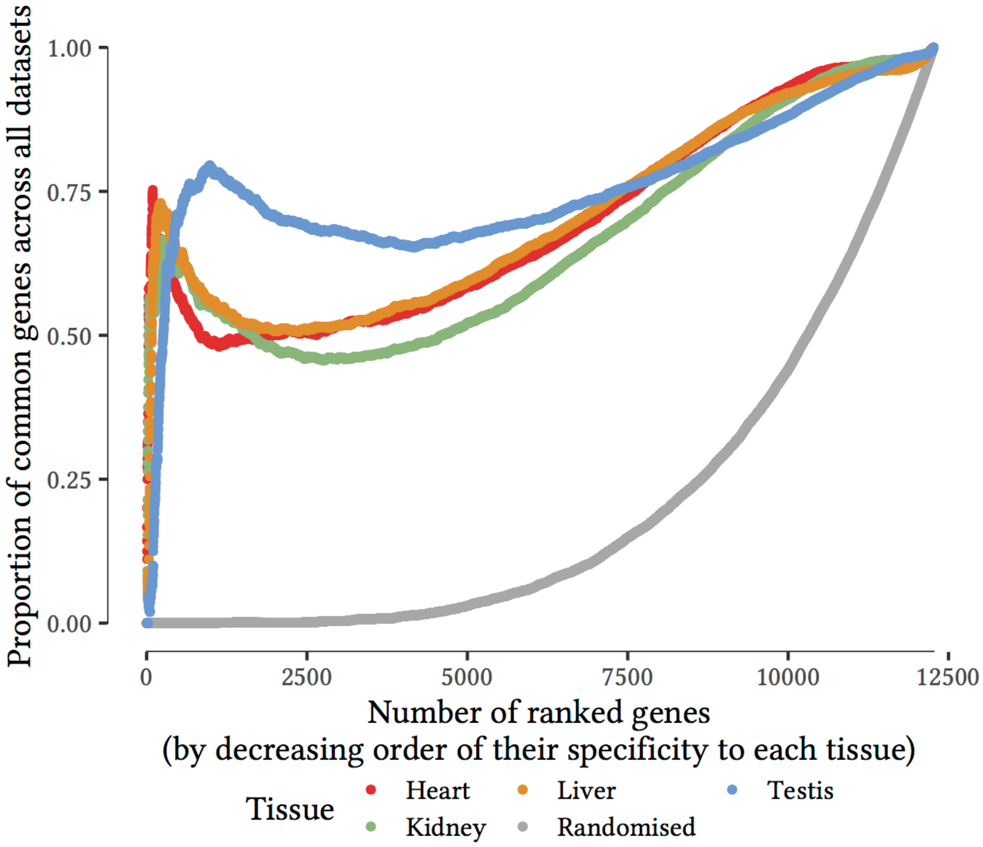


Figure 4.8. **Intersection size curve of $\mathcal{W}_1$ genes based on their specificity (FC ratio rank) in each tissue across the five studies.** When ranked by specificity in *Heart*, *Kidney* and *Liver*, one fortieth of $\mathcal{W}_1$ protein-coding genes are commonly shared between the five studies. For *Testis*, the shared amount of genes reaches more than one-tenth of $\mathcal{W}_1$ whole set of genes. Compared to the most variable genes (Figure C.19), the most specific genes seem to be more consistent across the studies.

Figure 4.8 presents the shifts in the intersection size curves of the four tissues of $\mathcal{W}_1$. The most specific genes in each tissue of $\mathcal{W}_1$ are shared between the five studies. Indeed, the slopes are very sharp before reaching a peak and dropping as sharply for the first fortieth genes in *Heart*, *Kidney* and *Liver*. The intersection of the most specific genes is even greater for *Testis* as it concerns more than a tenth of $\mathcal{W}_1$ genes. Figure 4.9 shows that this observation holds true for $\mathcal{W}_2$ when the number of tissues and genes is increased.

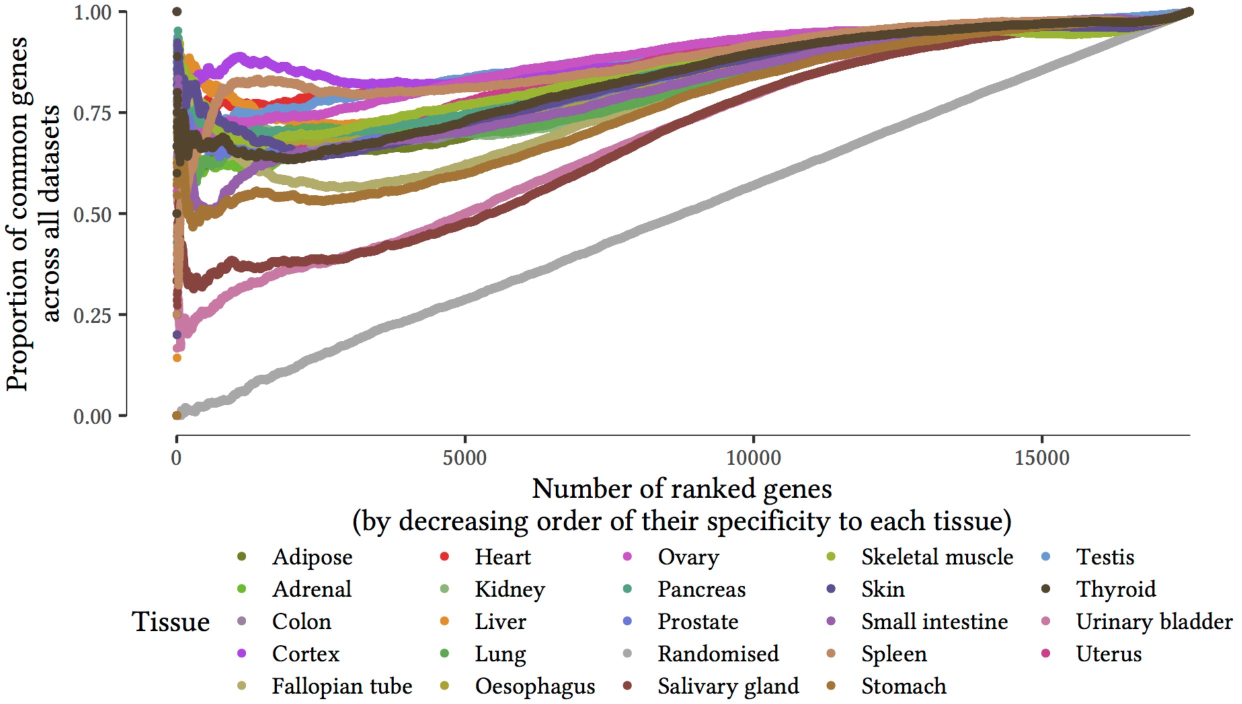


Figure 4.9. **Intersection size curve of $\mathcal{W}_2$ genes based on their specificity (FC ratio rank) in each of the twenty-three tissues across the two studies.**

4.3.1.3 *Hampel's test: detection of* atypical *expression*

The last method I used to characterise the TS genes is the Hampel's test. This test is a robust method for detecting outliers [Davies et al., 1993; Pearson, 2002] in data that are identically and independently distributed (i.i.d.) [H. Liu, Shah, et al., 2004], while easy to implement and use [Linsinger et al., 1998]. Much interlaboratory or interstudy research in the literature *e.g.* Linsinger et al. (1998), Lewczuk et al. (2006), Rocke (1983), and Apfalter et al. (1999) use the Hampel's test to detect outliers. The method uses the median and the MAD[13] to estimate the location and the spread, and a cut-off to define the observations that stand apart.

For this thesis, I have derived this test to identify conditions (*i.e.* tissues) where the gene expression is *atypical*. I rely on the two facts that most genes are expressed everywhere [Ramsköld et al., 2009; Uhlén, Fagerberg, et al., 2015; Melé et al., 2015] (see also Figure C.2), and they mostly present a limited variation in their expression through the various tissues (see Figures 4.11 and 4.12). There is a tissue specificity for a gene when its expression to this tissue is atypical, *i.e.* the expression in this tissue is an outlier to the average expression profile across the other tissues. Besides, this test allows detecting genes that are over- or under-expressed in specific tissues, whereas the other methods are only detecting the

[13] MAD: median absolute deviation

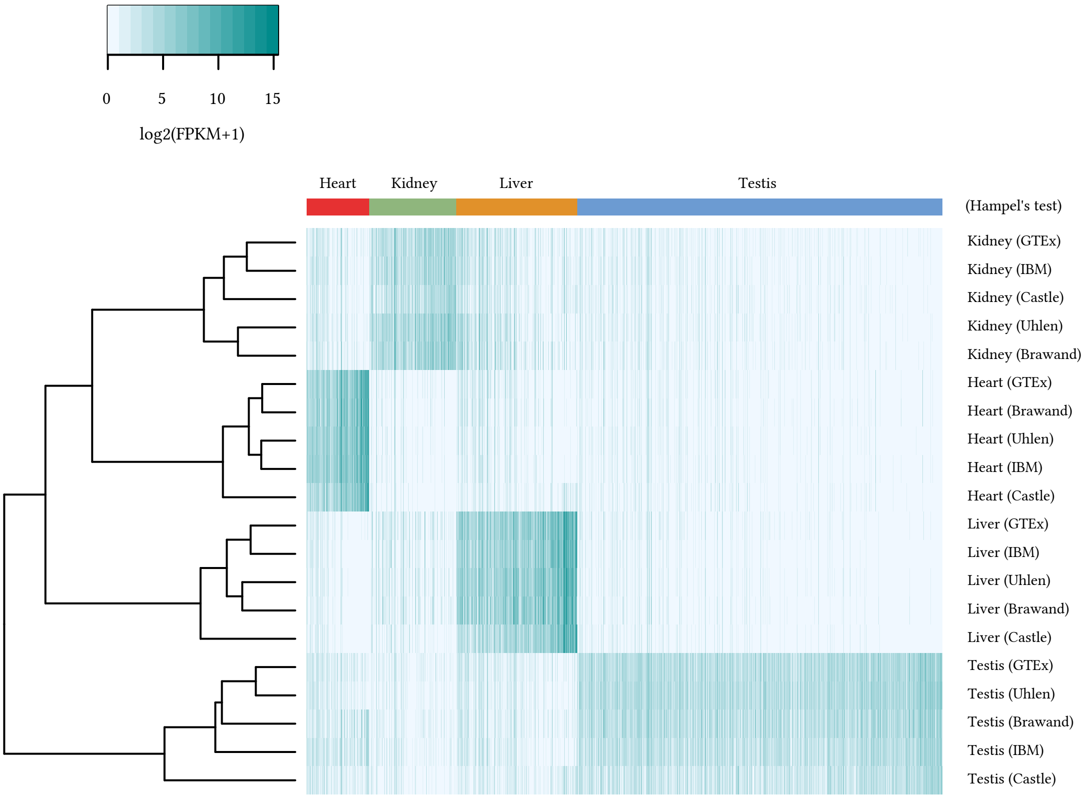


Figure 4.10. **Expression of the genes picked consistently with the Hampel method in each study solely in one tissue.**

overexpressed genes.

After implementing the method (see Algorithm 1, p. 224) with a (widespread) adimensional cut-off of 5.2, I have applied it to $\mathcal{W}_2$ and the whole original datasets. $\mathcal{W}_1$ comprises too few tissues to allow detecting *atypical* expression. Overall, there are always more than 60% of congruence between the genes tagged as atypical in a specific tissue for $\mathcal{W}_2$ and the whole dataset. The proportion of agreement between the partial and whole datasets increases when I filter the results to keep only the genes that are recurrently picked for both Uhlén et al. and GTEx data.

Figure 4.10 regroups the genes that the Hampel test detects as outliers for the five studies for their four shared tissues. All corresponding-tissue TREPs present similar patterns of expression regardless of their original study, although the Castle TREPs have overall lower expression values than the others. Other filters may improve the results.

Overall, the TS genes, identified separately in each dataset and with several methods, are showing a cleaner biological interstudy signal over possible technical intrastudy noise and are contributing to the high interstudy tissue correlations presented above.

4.3.2 *Uhlén categories*

Uhlén laboratory publications [Fagerberg et al., 2014; Uhlén, Fagerberg, et al., 2015; Uhlén, Hallström, et al., 2016] use different categories of genes to describe the normal human transcriptome. As their classification changes between these related papers, I have redefined a classification based on them (presented in Table 4.1) before applying it to $\mathcal{W}_1$ and $\mathcal{W}_2$ (Table 4.2).

The following classification considers the breadth[14], the level and the specificity of the gene expression.

Table 4.1. **Gene classification**
adaptation of Uhlén et al. classification [Fagerberg et al., 2014; Uhlén, Fagerberg, et al., 2015; Uhlén, Hallström, et al., 2016]

Category	Definition
Not detected	Never detected above 0 FPKM
Not expressed	Never detected above 1 FPKM
Mixed high	Expressed in a subset of tissues and always ≥ 10 FPKM
Mixed Low	Expressed in a subset of tissues and always < 10 FPKM
Ubiquitous High	Expressed in all the tissues and always ≥ 10 FPKM
Ubiquitous Low	Expressed in all the tissues and always < 10 FPKM
Group enhanced	Expressed in a subset of tissues with an expression $\geq 5 * \text{mean}_{\text{all the tissues}}$
Tissue enhanced	Expressed in a single tissue with an expression $\geq 5 * \text{mean}_{\text{all the tissues}}$
Tissue enriched	Expressed in a single tissue with an expression $\geq 5 * \text{Max}_{\text{all the other tissues}}$

Table 4.2 shows that for many of these categories, the number of shared protein-coding genes is high between the different studies of the two combined datasets $\mathcal{W}_1$ and $\mathcal{W}_2$. It supports the previous results that protein-coding genes present in general a similar gene expression profile for a same set of tissues across studies.

4.3.3 *Similar expression variability of the genes across studies*

To further appraise the robustness of gene expression, I have studied more globally their variability across studies.

There are several available estimators to describe the gene expression variability, *e.g.* the standard deviation (sd) the variance (sd^2) or the coefficient of variation ($\frac{sd}{mean}$). I only report here the results based on the coefficients of variation. The coefficient of variation (cv) allows assessing the dispersion of the gene expression values across the tissues within each dataset. As it adjusts for the mean, it is a more straightforward estimator to interpret

[14] The breadth of expression of a gene is the number of tissues (or cell lines) in which it is expressed.

Table 4.2. **Uhlén et al. gene categories**
Apart from the undetected genes and the ones expressed below 1 FPKM, a gene may be referenced in several categories.

Ensembl 76 (∼22,500 protein coding genes)		Not detected	Not expressed at 1 FPKM cut-off	Mixed expression		Ubiquitous expression		Group Enhanced	Tissue Enhanced	Tissue Enriched
				Low (< 10 FPKM)	High (≥ 10 FPKM)	Low (< 10 FPKM)	High (≥ 10 FPKM)			
Whole dataset	Castle	3,403	3,268	8,773	1,033	1,399	634	11	3,664	1,975
	Brawand	2,964	3,095	8,034	1,788	1,760	958	0	2,729	2,548
	IBM	2,693	2,605	7,325	1,406	1,135	858	322	5,248	2,453
	Uhlén	2,662	1,747	5,769	1,053	456	406	2,511	5,201	2,333
	GTEx	2,197	1,886	5,556	1,117	687	698	3,859	4,356	1,919
	Consensus	2,197	486	1,749	221	33	161	0	677	531
Common 4-tissues combined datasets	Castle	19,066	2,994	8,589	1,513	2,994	1094	—	—	2,185
	Brawand	19,505	2,962	8,626	2,228	2,962	1251	—	—	3,672
	IBM	19,776	2,989	8,534	1,954	2,989	1212	—	—	2,824
	Uhlén	19,807	2,917	8,367	2,227	2,917	1190	—	—	3,730
	GTEx	20,272	3,870	8,988	2,312	3,870	1427	—	—	3,554
	Consensus	1,973	550	3,351	649	550	439	—	—	1,412
Common 23-tissues combined datasets	Uhlén	2,662	1,970	6,160	1,135	594	427	1,285	5,776	2,518
	GTEx	2,197	2,258	6,966	1,540	1,822	997	1,048	5,496	2,460
	Consensus	2,197	1,544	4,936	791	423	417	558	4,223	1,885

than the variance itself, in particular for interstudy comparisons.

As depicted in Figure 4.11 (and Figure 4.12 for $\mathcal{W}_2$), the distribution of gene expression coefficients of variation presents a similar pattern across the five studies of $\mathcal{W}_1$.

The five datasets present two peaks. One at approximately 0.5 which characterises genes that are quite invariant in their expression across the four tissues within each study. Another subset of genes forms a peak for coefficients of variation equal to 2. This last group of genes are the most variable ones in each dataset. There is an overlap of the most variable genes between the five datasets (as shown in Figure C.19).

While Figure 4.4 has already established that expression profiles for each tissue are very similar across the studies, Figure 4.12 highlights that most genes seem to share the same intertissue expression profile variability as the distribution of the coefficients of variation across Uhlén et al. and GTEx studies are very alike.

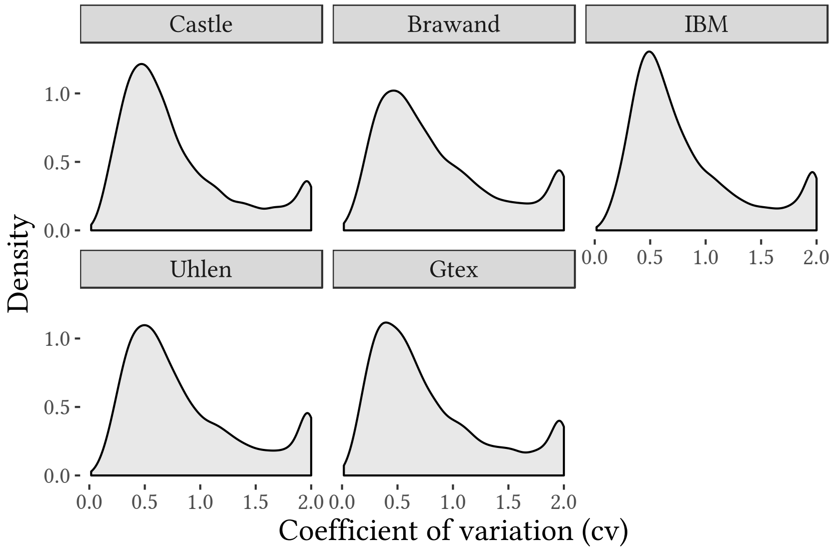


Figure 4.11. **Distribution of the coefficients of variation (cv) across $\mathcal{W}_1$ (common set of expressed protein-coding genes across the four common tissues): {*Heart, Kidney, Liver, Testis*} across the five studies.**
The coefficients of variation (cv) of the protein-coding genes (12,268) of the four tissues present the same bimodal distribution profile across the five studies.
These profiles present two peaks: at 0.5 and 2.
Genes with a cv less than or equal to 0.5 have a similar expression profile to a left-truncated version of the complete gene set ones (due to the 1 FPKM cut-off) as in Figure 3.1. On the other hand, the protein-coding genes with a coefficient of variation equal to or greater than 1.5 have two kinds of distinct profiles:

- The gene expression is low across the four tissues, and it is above the cut-off of 1 FPKM only once (see Figure C.21); or
- The gene expression is specifically high for one single tissue relative to the three others (See Figure C.18).

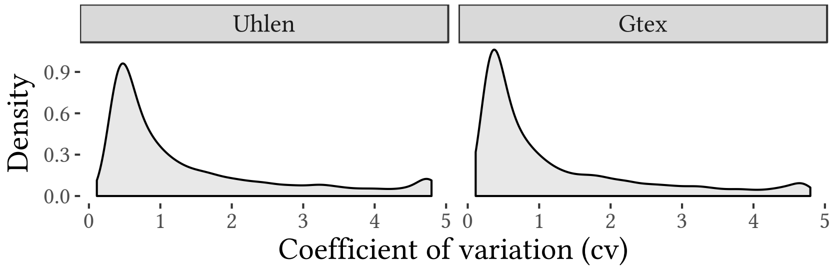


Figure 4.12. **Distribution of the coefficients of variation across $\mathcal{W}_2$.** The bimodal distribution is more unbalanced than in Figure 4.11. Indeed, as more tissues are included for the calculation of the coefficients of variation, the second peak is found around 5. This peak has a smaller amplitude than the peaks at 2 in Figure 4.12. There are still many genes that have a coefficient of variation around 2. However, the overall distribution of the higher coefficients of variation is smoother than for $\mathcal{W}_1$. Hence, most genes present a similar profile of expression through the various tissues.

More in-depth analyses confirmed that overall the genes present an equivalent coefficient of variation from one study to another for the same tissue set. See Figures C.19 and C.20.

4.3.4 *Curated sets*

Together the results from the previous sections indicate that many genes categories (if not all of them) have an equivalent (*i.e.* stable) expression profile across studies for the same tissues.

Protein-coding genes that were characterised consistently as any of the aforementioned categories across the five datasets of $\mathcal{W}_1$ or the two of $\mathcal{W}_2$ are provided digitally as supplementary data. They can be found at `http://www.barzine.net/~mitra/thesis`.
The code required to produce these results can be found at https://github.com/barzine/phd-analyses.

4.4 DISCUSSION

In this chapter, I have directly compared and integrated the human tissue transcriptome from five RNA-Seq studies. The meta-analyses are based on the largest number of undiseased human tissue studies to date. I have constructed two combined datasets of protein-coding genes. The first one ($\mathcal{W}_1$) comprises four tissues and 12,268 shared genes extracted from five independent studies (Krupp et al., 2012; Brawand et al., 2011; Uhlén, Fagerberg, et al., 2015; Melé et al., 2015 and IBM) and the second one ($\mathcal{W}_2$) comprises twenty-three tissues and 17,554 shared genes from two studies (Uhlén, Fagerberg, et al., 2015; Melé et al., 2015).

Clustering analyses (and Welch's Two Sample *t*-test) of these two sets confirm that RNA-Seq technical noise is lower than relevant biological signals present in the data. Indeed, Sudmant et al., 2015; Danielsson et al., 2015; N. Y.-L. Yu et al., 2015 and Uhlén, Hallström, et al., 2016 also observe that interstudy corresponding tissues pairs are more related than intrastudy non-corresponding tissue ones (average correlation for corresponding tissue-pairs $r = 0.75$, $\rho = 0.88$; average for non-corresponding tissue-pairs $r = 0.20$, $\rho = 0.75$).

I have then shown that overall genes present similar interstudy expression variability profiles for the same tissue set. I have considered different gene groups to examine their coherence of expression profiles more closely.

Since there is no generally accepted definition of a TS gene (see Section 4.3.1), I have relied on three different methods to study them, including extracting TS gene definitions from an existent resource *TiGER* [X. Liu et al., 2008] that I have updated to the current human genome build (GRCh38). Mining the experimental data with this updated list highlights the need for caution when dealing with older resources. While the congruence of the three methods is partial, the TS genes show distinct expression profiles across tissues that are rather consistent through the different studies.

I have also explored the congruence of several other genes categories across the studies following a classification inspired by Uhlén et al. publications [Fagerberg et al., 2014; Uhlén, Fagerberg, et al., 2015; Uhlén, Hallström, et al., 2016]. These gene categories are based on the level and breadth of expression ('Not detected', 'Not expressed at 1 FPKM', 'Ubiquitous low expression ($<$ 10 FPKM)', 'Ubiquitous high expression ($\geq$ 10 FPKM)', 'Mixed low expression (when expressed, $<$ 10 FPKM)', 'Mixed high expression (when expressed, $\geq$ 10 FPKM)', 'Group Enhanced', 'Tissue Enhanced', and 'Tissue Enriched').

Finally, I have compiled all the genes showing a consistent pattern of expression through the meta-analyses across the studies into curated sets.

Since I started this project, other research groups have published similar studies. However, at the time of writing this thesis, all the other studies (including the aforementioned ones) were still based on the human genome build GRCh37 (or hg19), while I am using the more recent GRCh38 one. These studies either have different focus, aims, approaches or more limited scopes. Below, I outline how my work expands or completes theirs.

- M. Uhlén, B. M. Hallström, et al. (2016). 'Transcriptomics resources of human tissues and organs'. *Mol. Syst. Biol.* 12 (4) This review presents the results of N. Y.-L. Yu et al. (2015) and Danielsson et al. (2015) discussed in the following points. It also compares data released by the GTEx consortium [Bahcall, 2015; GTEx Consortium, 2015; Gibson, 2015] to its authors' own dataset (Uhlén data [Fagerberg et al., 2014; Uhlén, Fagerberg, et al., 2015]). As many findings from other studies (that I discuss hereafter) are reviewed, the comparison of the two studies is limited to the examination of the proportion of genes in each category of a simplified classification for nineteen tissues. We reach the same conclusions: overall, there are significant overlaps across the datasets for each of the categories they have considered in their study, *i.e.* 'Expressed in all', 'Not detected', 'Tissue enriched', 'Group enriched', 'Enhanced' and 'Mixed'.

- P. H. Sudmant et al. (2015). 'Meta-analysis of RNA-seq expression data across species, tissues and studies'. *Genome Biol.* 16, p. 287. The authors focus on interspecies, intertissue and interstudy comparisons. The major issue of the study is its use of TMM as a gene expression unit. TMM normalisation has the assumption that genes have a stable expression across conditions while the different analyses I have presented indicate that many protein-coding genes expression profiles show tissue specificity while they have a stable expression across the studies. They also limit their scope to very specific orthologs since they explore RNA-Seq expression data across species, tissues and studies. They confirm that with RNA-Seq expression profiling, the interstudy technical variation is generally lower than the intrastudy biological one, *i.e.* interstudy homologous tissues of the same species are usually closer in similarity than different tissues of the same species (or matched tissues of different species) of the same study. They found that interstudy comparisons are more variable for *human* than other species. They finally note that this kind of meta-analysis is dependent on the choice of tissues to be studied.
- N. Y.-L. Yu et al. (2015). 'Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium'. *Nucleic Acids Res.* 43 (14), pp. 6787–6798, integrate Uhlén et al. data [Fagerberg et al., 2014] with CAGE peak expression data from the FANTOM5 consortium [FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014]. Overall, their analyses are very similar to the ones I have presented in this chapter. We also reach similar conclusions as well:
  - Overall gene expression is comparable through their two datasets.
  - Tissue expression signatures are independent of the data set (and profiling method).
  - Global comparison of ubiquitously expressed and TS genes are comparable across the two studies
  - They compare the two datasets at gene levels because of the lack of accuracy of the current RNA-Seq protocols and algorithms and focus on the protein-coding genes as the level of agreement between the two studies is low (which they attribute to the polyA-selected protocol of Uhlén et al. data).

  Besides the choice of the original studies, the few differences are (1) the version of the annotation (they use the previous human genome build (GRCh37)) and (2) they apply more simplified classification for ubiquitous and TS genes. In addition, they are also more lenient to assess the congruence (*e.g.* expressed in all tissues in one dataset and 95% of the tissues in the other datasets is considered as concordant). Strikingly, they also found a discrepancy with *Salivary gland* compared to the other tissues which I have noticed in this study[15].

- F. Danielsson et al. (2015). 'Assessing the consistency of public human tissue RNA-seq data sets'. *Briefings Bioinf.* 16 (6), pp. 941–949, covers three different tissues (*Brain*[16], *Heart*, *Kidney*) extracted from five projects (E. T. Wang et al., 2008; Brawand et al., 2011; Uhlén, Fagerberg, et al., 2015; Krupp et al., 2012 and IBM). Their study is limited to the comparison of precomputed data (from the original laboratories) versus uniformly reprocessed data (by themselves). They are exploring experimental variation factors and possible correction strategies. They reveal that original precomputed data have considerable study-specific biases. Their results on interstudy tissue similarities are superficial. One of their most 'fine-grained' results is the very low number of shared genes amongst the hundred most expressed genes for the three tissues across their five datasets.
- A. Santos et al. (2015). 'Comprehensive comparison of large-scale tissue expression datasets'. *PeerJ* 3, e1054, focuses on the congruence of gene—tissue association through different types of expression data (transcriptome and proteome) and resources. They report that most genes are either expressed in every considered tissue or in small subsets in their constructed working dataset. They also find that tissue specificity trends are globally similar, even though there are many differences in the identified genes set across studies. They have integrated together five tissues (*Heart*, *Kidney*, *Liver*, *Nervous system* and *Small intestine*) from five transcriptome datasets that they use 'as-is'[17], before refining a complementary set that comprises only the three highest-quality datasets: *UniGene database*[18] [Wheeler et al., 2003; Pontius, Joan U. and Wagner, Lukas and Schuler, Gregory D., 2002], Uhlén et al. data (HPA RNA-Seq) [Fagerberg et al., 2014; Uhlén, Fagerberg, et al., 2015] and Castle et al. (RNA-Seq atlas) [Krupp et al., 2012] for which they provide association data for 14,722 genes. They forsake the direct quantitative exploration and comparison of gene expression between the studies, but examine the tissue association enrichment through the gene expression fold change (FC) for a qualitative cross-study. The main issue with their transcriptomic study is their assumption that higher expression[19] means more robust gene—tissue association which may often (but wrongly) be translated to a greater tissue-specificity.
- Q. Wang et al. (2017). 'Enabling cross-study analysis of RNA-Sequencing data'. *bioRxiv* (110734), have used subsets of GTEx and the cancer genome atlas (TCGA) raw data that they have quantified mRNAs at transcript levels with GRCh37 (hg19). They have corrected for the study effect with *ComBat* [W. E. Johnson et al., 2007] and have released the normalised data to the community. Note that EBI Gene Expression Atlas provides more recent versions of the GTEx and TCGA (as a part of the pan-

[15] *Salivary gland* is the only tissue for which Uhlén and GTEx show a Pearson correlation coefficient $r < 0.65$.

[16] Either *Cerebral cortex* or *Hypothalamus*

[17] Even in the follow-up paper, where they reprocess all the RNA-Seq data for *mouse*, *rat*, *pig*, Palasca et al. (2018) do not mention any improvement for the *human* RNA-Seq data (either in the results or methods and data).

[18] *UniGene database* — https://www.ncbi.nlm.nih.gov/unigene

[19] FPKM values are directly used as score for true presence and selectivity to a tissue

cancer analysis of whole genomes (PCAWG) project[20]) data.

- SEQC/MAQC-III Consortium (2014). 'A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium'. *Nat. Biotechnol.* 32 (9), pp. 903–914, have found that relative expression measurements by RNA-Seq are accurate and reproducible across sites. The authors also showed that the overlap of identified and characterised genes is imperfect (91%) even when the design includes the same two well-characterised reference RNA samples across all sites. Also, this specific design prevents inferring how the biological signal may compare to the individual variations and the possible noise introduced by collection, storage and extraction protocols.
- Several papers (Khang et al., 2015; Peixoto et al., 2015; Rau et al., 2014) explore the reliability of RNA-Seq in the context of DGEA which is outside the scope of this thesis.
- Zhuo et al., 2016 explore the stable expressed genes across multiple (24) RNA-Seq studies for *Arabidopsis* which is why I will not discuss it further.

In summary, while the expression levels are hard to translate directly from one study to another [N. Y.-L. Yu et al., 2015; Santos et al., 2015], many facts have been highlighted in this thesis with a subset of them confirmed by the studies mentioned before:

- Tissues are clustering preferably with corresponding (or closely biologically related) tissues even across studies rather than clustering with different tissues from the same study (*i.e.* biological signal >>> technical noise due to RNA-Seq protocols).
- More recent transcriptome studies are more congruent than previous ones.
- *Testis* presents the highest number of TS protein-coding genes (see Figure C.1). It also presents the most variety of expressed protein-coding genes ($\geq$ 1 FPKM) Castle, Brawand, IBM and Uhlén studies. This extends to GTEx study if all detected genes (*i.e.* above 0 FPKM) are considered.
- *Liver* has the most robust TS protein-coding genes. It may be explained either by a robuster gene expression, its greater homogeneity than most tissues, or a greater knowledge and better annotation than the other tissues.
- Most genes are ubiquitously expressed while a small proportion are expressed in a very limited set of tissues.
- Well-differentiated tissues have specific expression profiles that allow using processed data 'as-is' for rough comparisons such as sample swap checks or quality controls, *e.g.* *Salivary gland* in Uhlén et al. data which presents low correlation with GTEx data (see Figure 4.5, p. 98) and with FANTOM5 data [N. Y.-L. Yu et al., 2015].
- Annotations have an essential effect on the final results. Thus, whenever possible, we ought to keep the resources up-to-date.

[20] PCAWG project analyses conjointly all available kind of research data related to cancer

Unsurprisingly, updating the human genome version from GRCh37 to GRCh38 for the reconstruction step[21] enhances the results significantly.

Besides, as my analyses were incorporating more studies, the results were supporting more similarity in the gene expression levels across the tissues and studies. Indeed, as I focus my analyses on the common set of genes across the studies, I remove the most interstudy variant genes (*i.e.* that are probably more sensitive to technical factors), and I bias the analyses towards the genes for which RNA-Seq is more robust to quantify their expression profiles. Hence, it may be interesting to relax the filters by including genes that are found in any two or more datasets as a follow-up study.

4.5 CONCLUSION

The meta-analyses show that RNA-Seq captures a strong biological signal for tissue gene expression despite any noise created by batch effect or technical variations.

Tissues reference expression profiles (TREPs) are well correlated across independent studies and are the sum of the overall genes contributions.

While highest expressed genes fail to show significant correlation between the different studies, the analyses show that most gene expression profiles are comparable from one study to another. The *gene centred* heatmap (Figure 4.13) available now in *EBI Gene Expression Atlas*[22] [Petryszak, Keays, et al., 2015] and its corresponding widget are a direct translation of this observation.

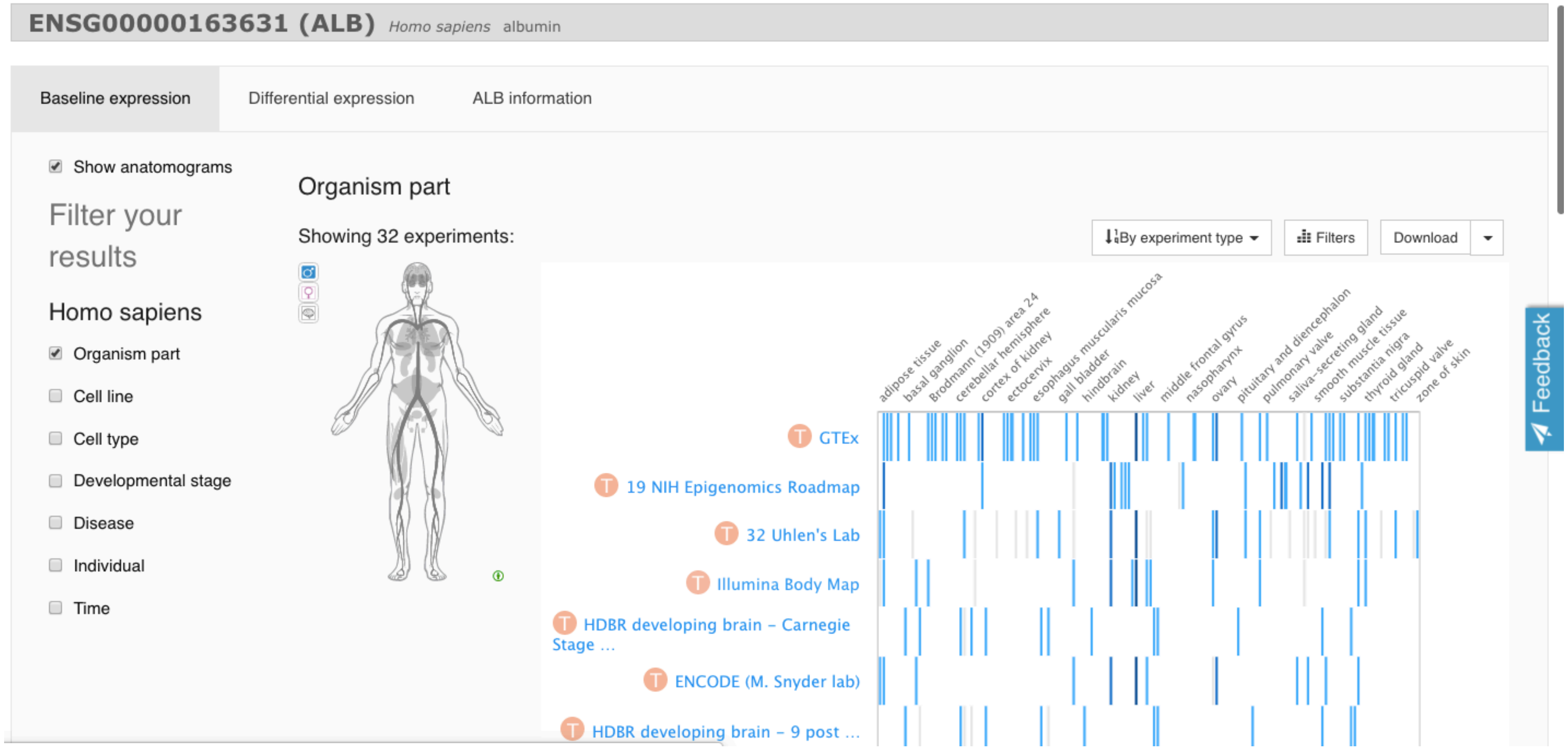
To assist further research, I provide extensive curated and consolidated gene sets for the different categories reviewed in the above analyses, *i.e.* for the TS genes, and the categories derived from Uhlén publications: 'Not detected', 'Not expressed at 1 FPKM', 'Ubiquitous low expression ($<$ 10 FPKM)', 'Ubiquitous high expression ($\geq$ 10 FPKM)', 'Mixed low expression (when expressed, $<$ 10 FPKM)', 'Mixed high expression (when expressed, $\geq$ 10 FPKM)', 'Group Enhanced', 'Tissue Enhanced', and 'Tissue Enriched'.

There is a need for more multi-tissue studies with biological replicates to refine and complete the above findings for other tissues and extend them to the transcript isoform level.

New strategies, notably normalisation methods, have to be also developed to allow the easy reuse of uniformly processed and quantified data by the community. Ideally, the final aim should be to provide a general human transcriptome build as it already exists for the genome. Finally, as long as the annotations are redefined and refined, it also means that periodic resources reprocessing may be inevitable.

[21] See Section 1.2.5.2: Reconstruction strategies

[22] *EBI Gene Expression Atlas* — https://www.ebi.ac.uk/gxa/



ENSG00000163631 (ALB) *Homo sapiens* albumin

Baseline expression | Differential expression | ALB information

☑ Show anatomograms

Filter your results

Homo sapiens

☑ Organism part

□ Cell line

□ Cell type

□ Developmental stage

□ Disease

□ Individual

□ Time

Organism part

Showing 32 experiments:

Figure 4.13. **Example of EBI gene expression atlas gene centric heatmap**. This heatmap shows the relative expression of the Albumine (ENSG00000163631) across the tissues and studies. Note that the expression is calculated within each study library before being aggregated by identical condition or tissue.

I was taught that the way of progress is neither swift nor easy.

Marie Curie

5 HUMAN MS-BASED PROTEIN EXPRESSION LANDSCAPE

After exploring the high-throughput human transcriptomic studies in Chapter 4 and before integrating them with proteomic data in Chapter 6, I present in this chapter the comparison of the three proteomic datasets introduced in Chapter 2. Ezkurdia et al. (2014) and Deutsch et al. (2015) have partially reviewed these data. However, we have reprocessed them starting from the raw data for this thesis. In this context, reassessing the quantified processed proteomic data before any integration is pertinent.

The work presented in this chapter was done in collaboration with Dr James Wright who has implemented the new protein quantification method (presented in Section 5.2). I have received general feedback from Dr Alvis Brazma, Dr Mar Gonzàlez-Porta, Dr Sarah Teichmann.

5.1 AN OVERALL FRAGMENTED AND DISPARATE UNIVERSE TO EXPLORE

As I have described in Chapter 1, proteins present a wide range of physicochemical properties (see Section 1.1 and Appendix A.1) and are challenging to identify and quantify in high-throughput studies (see Section 1.3). Thus, it comes as no real surprise that while the use of MS for proteomics has been developing since the 1980s [Papachristodoulou et al., 2014], the first notable attempts to draft the human proteome occurred only recently in 2014 [M.-S. Kim et al., 2014; Wilhelm et al., 2014], or that the oldest (unpublished) available multi-tissue Cutler dataset is from 2010 (see Section 2.3.3). Till early 2019, Cutler Lab, Kuster Lab and Pandey Lab datasets are the only ones that explore concurrently several non-diseased human tissues. See Section 2.3 for more details and the processing pipeline designed and implemented by Dr James Wright to handle them.

As presented in Figure 5.1, they share four tissues: *Heart*, *Lung*, *Ovary* and *Pancreas*. The protein overlap of this four-tissues set between these three datasets is rather narrow as shown in Figure 5.2. The Cutler Lab dataset shares the smallest number of tissues with the two other studies; Pandey Lab and Kuster Lab datasets share over twice as many proteins that they share with Cutler (3,338 instead of 1,384). The number of shared proteins between Pandey Lab and Kuster Lab data rise to 4,172 when all their fourteen common tissues are considered.

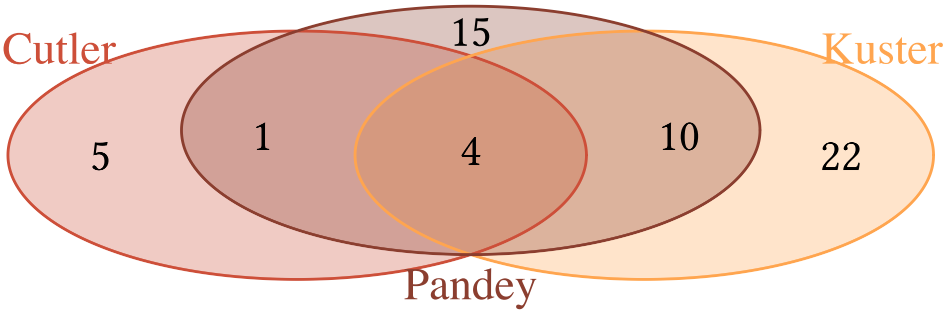


Figure 5.1. **Distribution of unique and shared tissues between the three MS-based proteomic studies.** The three datasets share together: *Heart*, *Lung*, *Ovary* and *Pancreas*. The two most recent studies share fourteen tissues in total; the additional ten tissues are: *Adrenal gland*, *Colon*, *Gallbladder*, *Oesophagus*, *Kidney*, *Liver*, *Placenta*, *Prostate*, *Rectum* and *Testis*.

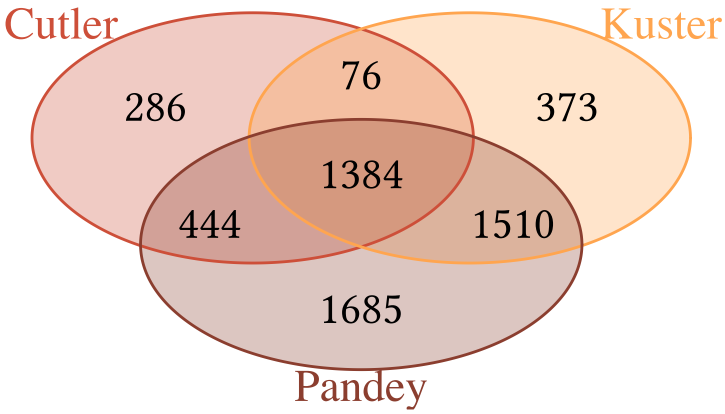


Figure 5.2. **Proteins overlap between the common four tissues of the three proteomic studies.** Unique and shared proteins detected and quantified across the three MS studies for their four shared tissues: *Heart*, *Lung*, *Ovary* and *Pancreas*.

5.1.1 *MS proteomic data has high detection variability*

Figure 5.3 illustrates the number of proteins identified in each of these four tissues. The colours indicate in which dataset (or group of datasets) the proteins have been identified. See Figure D.1 for the precise numbers of each set.

The tissue with the highest number of identified proteins, regardless of which dataset, is the *Ovary*.

The highest number of proteins identified in all three datasets at once (600) is in the *Heart*. The other tissues have the *Kuster and Pandey* set as their largest protein group formed by

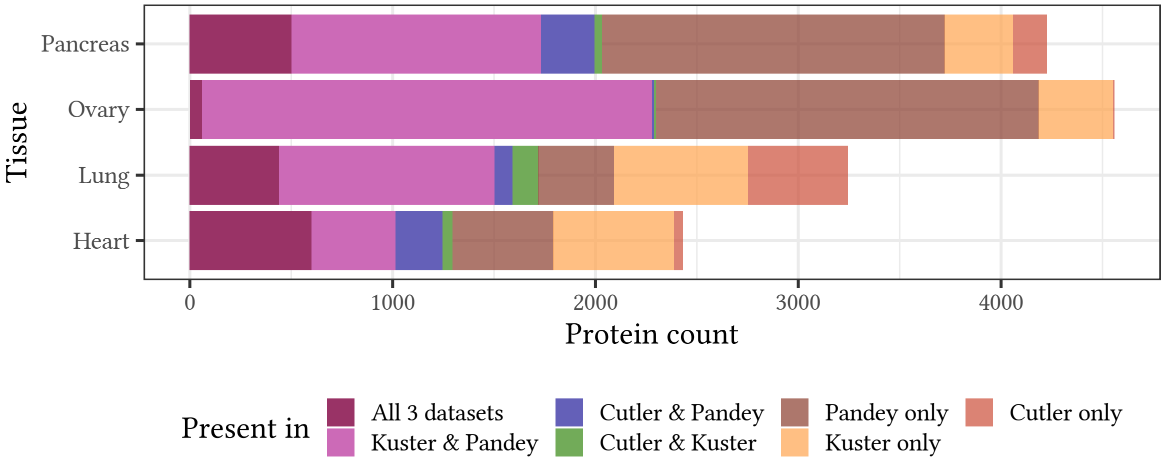


Figure 5.3. **Number of identified proteins in each of the four common tissues for the three proteomic data.** Proteins found in more than one dataset are most likely true (in red, light and darker green or purple — the most validated ones in red as they are found in all three datasets). See Figure D.1 for the precise numbers in each set.

more than one dataset.

The largest set of identified proteins in *Pancreas* and the second one in *Ovary* are proteins only found in the Pandey Lab dataset (*Pandey only*). As shown in Figure 5.4, our state-of-the-art pipeline (see Section 2.4.2) has identified the highest number of proteins in the Pandey Lab dataset. Thus, it is coherent that Pandey Lab proteins represent a large part of the identified proteins in each tissue (either as *Pandey only* set or in agreement with the other datasets: *All 3 datasets*, *Kuster & Pandey* or *Cutler & Pandey*). More surprising is that the Cutler dataset comprises a notable amount of proteins in *Lung* that are missing in the other two. A few of these proteins (82) are missing altogether, but a subset of them (410) is still found in (at least) another tissue of the other datasets.

While proteins found in more than one dataset are more likely true positives, it is impossible to exclude without risks the ones that are identified in one dataset only. Whether an identified protein in one dataset is an artefact (*i.e.* false positive) or a miss (false negative) in the other datasets is a challenging question; the diversified nature of proteins involves many sample preparation and simplification methods (see Sections 1.3.1 and 1.3.2).

5.1.2 *Overall about half of the proteins identified in each study for any given tissue are validated in a different study.*

As shown in Figures 5.3, D.1 and D.3, besides a few exceptions (*Oesophagus*, *Gallbladder* and *Testis*), more than half of the proteins are identified in the same tissue in more than one

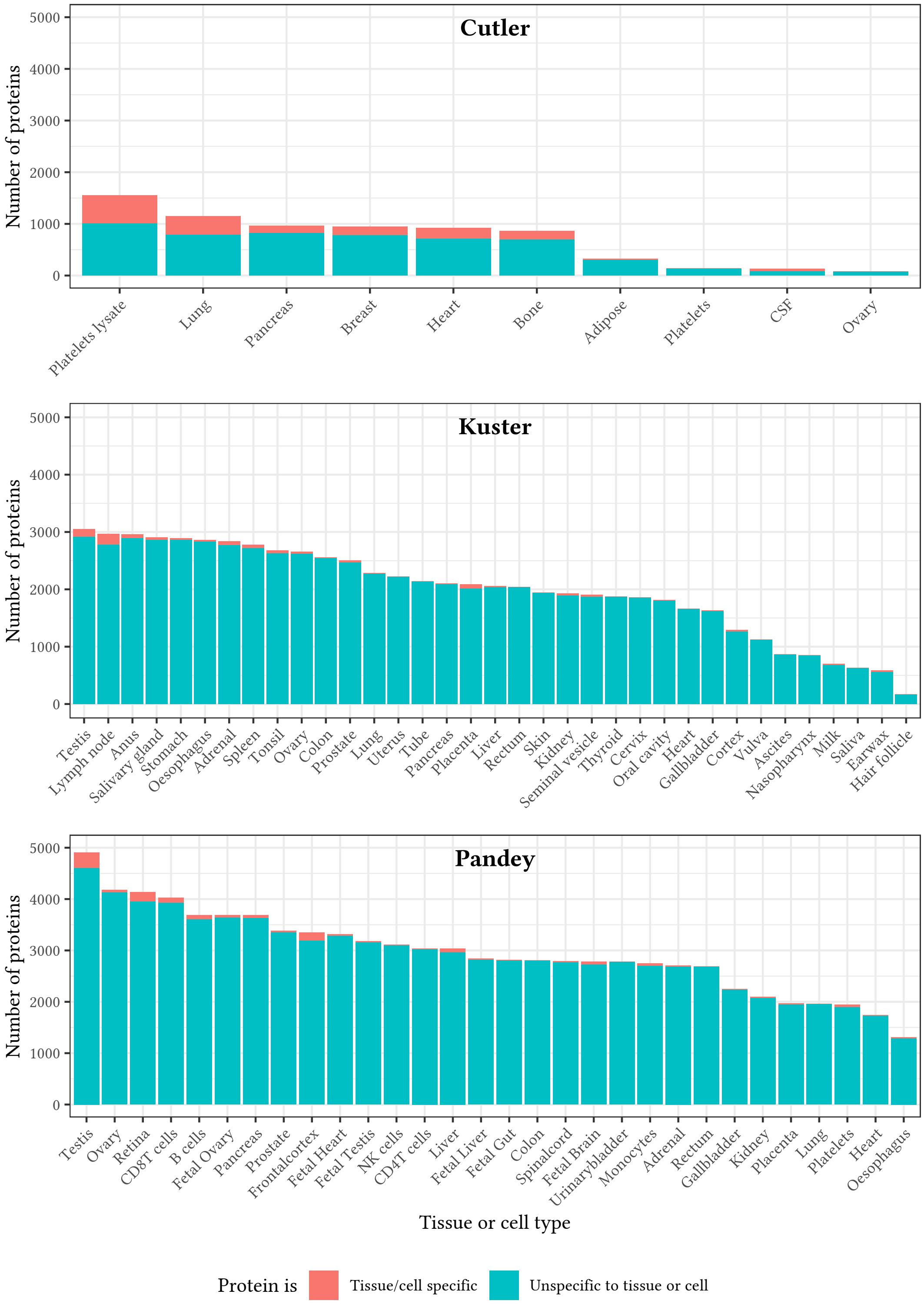


Figure 5.4. **Distribution of the proteins per tissue across the three datasets.** Cutler Lab dataset has the smallest and Pandey Lab the highest number of proteins per tissue. Coloured in red are the proteins that are specific to one tissue (or cell type); these proteins have been identified in one tissue solely within each dataset. Proteins in turquoise have been identified in several tissues of the dataset.

dataset. Three proteins are found in every tissue of every dataset: ALB, KRT9, KRT10. This number rises to forty when only the Pandey Lab and the Kuster Lab data are considered (see Table D.2). I have also investigated tissue-specific (TS) proteins (in red in Figure 5.4) that are also identified in more than one dataset. While the three datasets lacked to identify any TS protein at once, Pandey Lab and Kuster Lab datasets share a few (44 across eight tissues) — see Table D.3 for the complete list.

TS proteins are more difficult to confirm through different datasets, but one needs to be careful with the ubiquitous proteins as well. The latter may be present in the samples due to contamination: none of the three ubiquitous proteins (ALB, KRT9 and KRT10) is detected in *Heart* by the Human Protein Atlas[1] [Uhlén, Fagerberg, et al., 2015] while they are found in epithelial cells. One hypothesis is that contamination occurred when sampling from the donor or during the preparation or MS analysis. ALB found in the tissues is more likely coming from the blood supply (where ALB is abundant); KRT9 and KRT10 are environmental contaminants.

Lists for ubiquitous and TS proteins of each dataset separately and across the three (when consistent) are given as digital supporting data.

5.1.3 *Technical variability prevails over biological signal: intrastudy correlations of different tissues are globally stronger than same-tissue interstudy correlations.*

After defining the (1,384) protein set that is consistently detected in the four common tissues of the three datasets, I have assessed how consistent is their expression quantification across tissues and studies.

Following a similar approach to Figure 4.3 (p. 96), I cluster the twelve proteomic TREPs[2]. I have used Ward's method to link the TREPs based on their similarity that I have computed by subtracting from 1 the pairwise Spearman correlation of the expression levels of the 1,384 common proteins. As shown in Figure 5.5, the technical variability overcomes the biological signal as the proteomic TREPs cluster according to their original laboratory/study rather than their biological source, except for Cutler Lab and Pandey Lab *Heart*. However, Cutler Lab and Pandey Lab share the same organisation of their remaining tissues: *Pancreas* and *Lung* are the most correlated, and their pair is in turn most correlated to *Ovary*. Kuster Lab TREPs display the greatest amount of study bias (probably due to stronger batch effects — see Section 1.5.1).

Removing the proteins translated from the mitochondrial genes slightly improves the results but excluding the three ubiquitous (likely contaminants) proteins, presented in Section 5.1.2, is impactless.

[1] Human Protein Atlas — https://www.proteinatlas.org/

[2] Tissue reference expression profile. See Section 3.3.4 on page 87 for more details on TREPs.

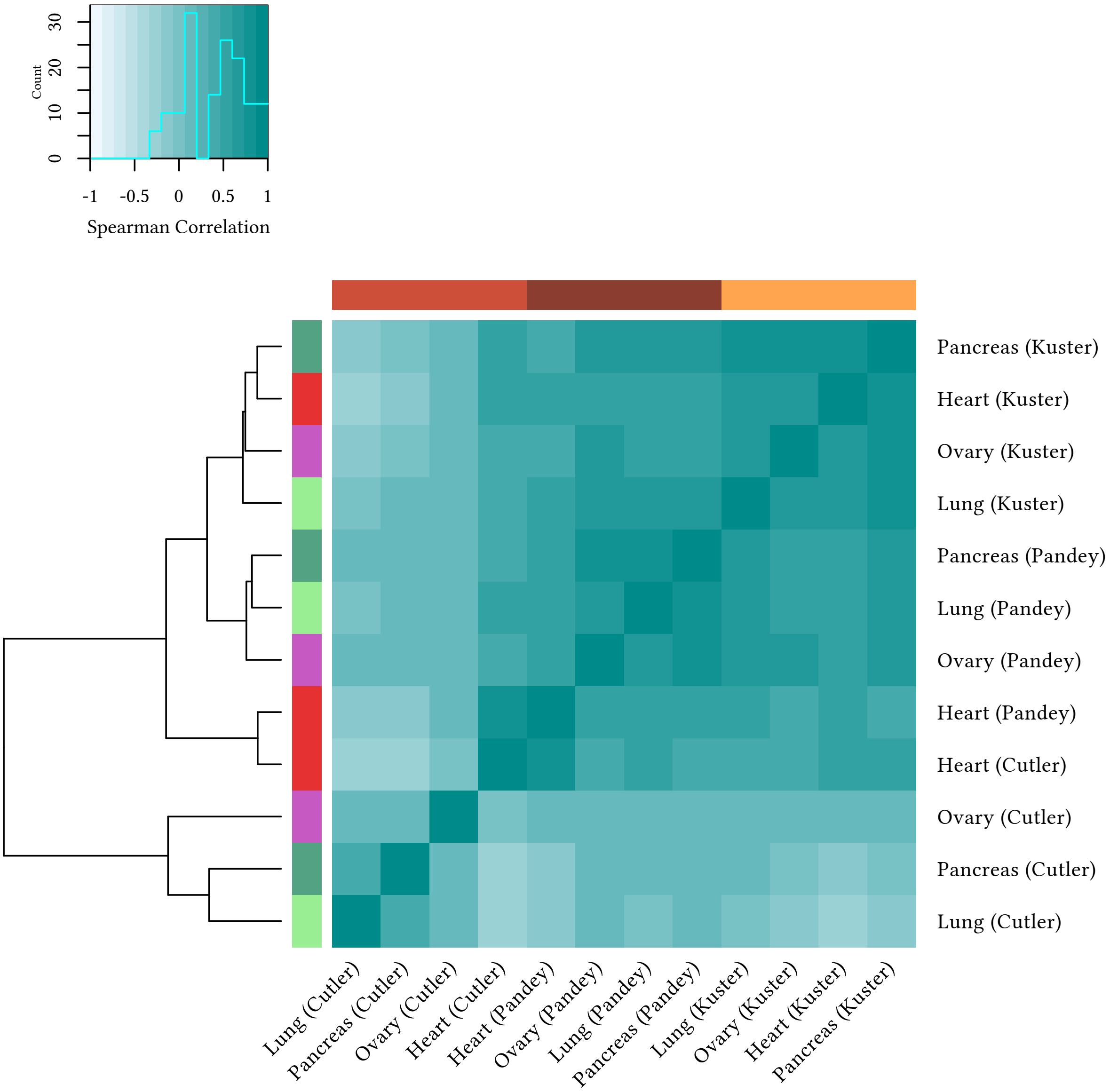


Figure 5.5. **Heatmap of the four common tissues between the three proteome datasets** based on the pairwise Spearman correlations clustering of the expression levels of 1,384 shared proteins. The samples mostly cluster by laboratory. Only *Heart* from Cutler Lab and Pandey Lab have a stronger intratissue correlation than an intrastudy one. Both for Pandey Lab and Cutler Lab, *Pancreas* and *Lung* are more correlated to each other and their pair to *Ovary*. The heatmap (Figure D.4) based on Pearson correlation instead prompts globally identical observations. See also Figure D.5 that shows (as a scatterplot) the relationship for *Heart* between Pandey Lab and Cutler Lab, and then, Figure D.9 between Pandey Lab and Kuster Lab.

Neither applying quantile normalisation or widespread scaling methods on top of this quantification allowed correcting for the technical variability.

Because of the limited number of proteins included in this analysis, it is unwise to draw definite conclusions except that there is an extensive need for new quantification normalisation methods that can help with protein expression meta-analyses and, as for now, one has to be very cautious when comparing proteome samples.

I attempted to expand this analysis by comparing the fourteen common tissues of Pandey Lab and Kuster Lab datasets, but the results are as inconclusive (see Figures D.7 and D.8).

5.2 NEW QUANTIFICATION METHOD

In this thesis context where I aim to integrate the mRNA expression levels to the protein ones (presented in Chapter 6), I have developed with the help of Dr James Wright a new method to infer and quantify the proteins.

Our original processing workflow (detailed in Section 2.4.2) intends to provide a reliable, state-of-the-art, protein quantification. Our method seems more rigorous than M.-S. Kim et al. (2014) original paper; Ezkurdia et al. (2014) challenge the correctness of their quantification by highlighting the disputable presence of olfactory receptors (ORs) in many tissues. On the other hand, our workflow lacks to detect or quantify any possible OR in the Pandey Lab data. The left part of Figure 5.6 summarises our first method of quantification.

While probably more accurate, our state-of-the-art quantification method is also more stringent than the original authors' one. Thus, the total number of quantified proteins is more limited.

As presented in Chapter 1, protein quantification methods rely on PSMs identification (see Section 1.3.4.2) and a chosen approach for protein inference (see Section 1.3.4.3). I realised that our main limiting factor is that we get quantification only for proteins that have at least three unique peptides since this first method is based on the Top3 approach [Silva et al., 2006] and uses the three *unique*[3] most expressed peptides of a protein to estimate its overall expression (see Equation (Top3 IBAQ) on p. 123 and Section 1.3.4.4).

$$\hat{\mu}_{ij} = \frac{\sum \text{Intensity of Top 3 unique peptides}_{ij}}{\text{Total Intensity in experiment } j} \tag{Top3 IBAQ}$$

where:

- $\hat{\mu}_{ij}$ is the normalised expression for the protein i in experiment j (normalised PSM),
- $\sum$ Intensity of Top 3 unique peptides$_{ij}$ is the sum of the intensity of the three most intense unique peptides of the protein i,
- Total Intensity in experiment j is the total sum of the intensity of all the peptides identified in the experiment j.

[3] I.e. exclusive to a single protein

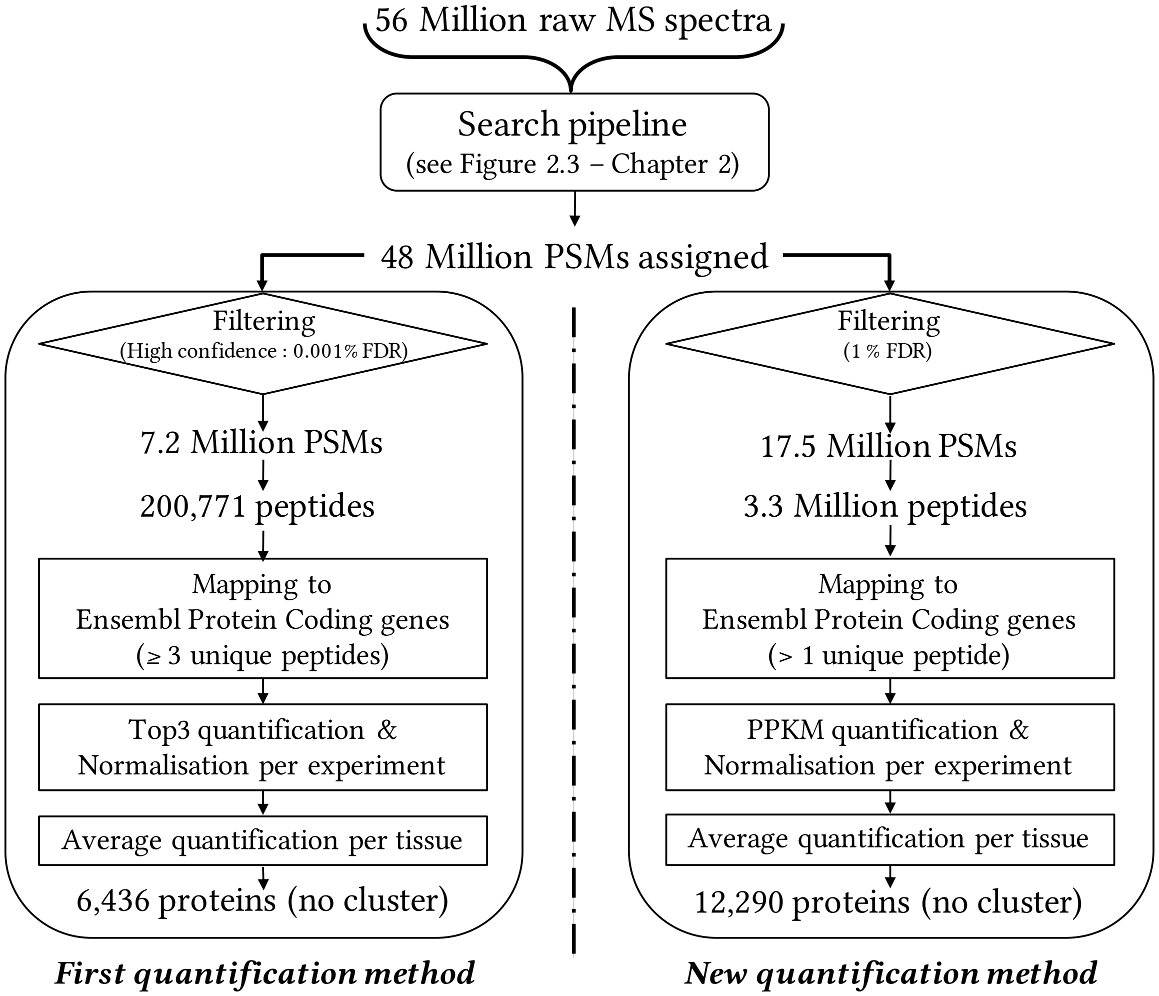

56 Million raw MS spectra
Search pipeline
(see Figure 2.3 – Chapter 2)
48 Million PSMs assigned
Filtering
(High confidence : 0.001% FDR)
7.2 Million PSMs
200,771 peptides
Mapping to
Ensembl Protein Coding genes
(≥ 3 unique peptides)
Top3 quantification &
Normalisation per experiment
Average quantification per tissue
6,436 proteins (no cluster)
First quantification method
Filtering
(1 % FDR)
17.5 Million PSMs
3.3 Million peptides
Mapping to
Ensembl Protein Coding genes
(> 1 unique peptide)
PPKM quantification &
Normalisation per experiment
Average quantification per tissue
12,290 proteins (no cluster)
New quantification method

Figure 5.6. **Two quantification methods applied to Pandey Lab data.** For both approaches, the search pipeline assigning PSMs remains identical (see Section 2.4.2). The first quantification method, which follows a robust inference method involving at least three unique peptides, relies on the intensity of the three most intense unique peptides. The new quantification method that I have devised allows more relaxed inference parameters as it also uses the non-unique peptides for the quantification. See Equation (PPKM definition), which is similar to Equation (Canonical F/RPKM formula) on page 24. After averaging per tissue and removing the clusters, the number of quantified proteins with the new method is close to twice the number provided by the first described method. The new quantification method was designed for the analyses in Chapter 6 in particular; this is why the filtering is also less strict than for the first quantification method since my main focus is the integration of the proteomic data with the RNA-Seq data. Clusters are protein groups that can be mapped to more than one Ensembl gene identifier. Note that the final proteins numbers include only the fifteen tissues used for the integration in the following Chapter 6.

For the following chapter analyses, I requested Dr James Wright to provide me a new quantification for the Pandey Lab data where both unique and the *degenerate* (*i.e.* non-unique) peptides are involved in the estimation of the protein expression.

The new method I have devised allocates the degenerate peptides in proportion to the distribution of unique peptides per protein following a similar approach to *Cufflinks2* for RNA-Seq data (see Section 1.2.5.3). As three unique peptides are no longer required for the quantification, it allows relaxing the inference parameters to two unique peptides (in order to still avoid *one-hit wonders*, see Section 1.3.4.3) for the identification of the proteins.

Once the identification is done, all the unique and degenerate peptides are mapped to the identified proteins. For the unique peptides, their quantification is directly linked to one and only protein. However, for the degenerate peptides, it is necessary to gauge the likely amount provided by each of their matching proteins.

For this purpose, the distribution coefficient of the degenerate peptide d to the protein p, $C_{d,p}$, is defined as below:

$$C_{d,p} = \frac{\sum_{i=1}^{Nu_p} \frac{\text{Number of PSM}(u_{p,i})}{Nu_p}}{\sum_{q \in P_d} \left(\frac{\sum_{j=1}^{Nu_q} \text{Number of PSM}(u_{q,j})}{Nu_q} \right)} \qquad \text{(Distrib. coeff. of the degenerate peptide)}$$

where:

- P_d is the set of identified proteins that include the degenerate peptide d
- Nu_p is the number of unique peptides of the protein p, $\forall p \in P_d$
- $u_{p,i}$ is the ith unique peptide of the protein p, $\forall p \in P_d$
- Number of PSM(u) is the number of PSMs of the peptide u

Then, the contribution of the degenerate peptide d to the quantification of a protein p is computed as:

$$\mathrm{Q}_{d,p} = C_{d,p} \cdot \mathrm{Q}_d \qquad \text{(Distribution of the degenerate peptide quantification to a protein)}$$

where:

- $C_{d,p}$ is the distribution coefficient defined above
- Q_d is the total quantification of the degenerate peptide d

The new quantification follows a similar approach to the F/RPKM normalisation — see Equation (Canonical F/RPKM formula). Protein expression levels are expressed in PPKMs, which stands for PSMs Per Kilobase of gene per Million. As shown in Equation (PPKM definition), this method counts the number of PSMs that can be mapped to the corresponding Ensembl gene identifier of a protein. Then, this raw count is normalised by dividing it by the product of the longest transcript length and the total number of PSMs assigned in that experiment; the result is finally multiplied by a factor (10^6) to facilitate reading. Using the longest *transcript* length instead of the longest protein isomer allows avoiding issues due to annotation differences between gene and protein levels in the analyses of the following Chapter 6.

$$\hat{\mu}_{ij} = \frac{\text{Number of PSMs matching } G_i \cdot 10^{-6}}{\ell_{G_i} \cdot \text{Number of PSM}_j} \qquad \text{(PPKM definition)}$$

where:

- $\hat{\mu}_{ij}$ is the normalised expression of the protein i in the experiment j,
- G_i is the gene that corresponds to the protein i,
- Number of PSMs matching G_i is the number of PSMs mapped to G_i,
- ℓ_{G_i} is the length of the longest transcript of G_i,
- Number of PSM$_j$ is the total number of PSMs identified in the experiment j.

Dr James Wright has implemented this new method and provided me PPKM quantifications for the Pandey Lab and Kuster Lab data.

Although smaller, another limiting factor is the filtering threshold of the selected PSMs to be inferred in proteins. We have chosen a conservative (and state-of-the-art) threshold prior to the first quantification filtering and a less strict one for the new method. As my aim is to compare and integrate proteomic and transcriptomic data together, the primary purpose of the new quantification is to provide a number of proteins that is roughly similar to the number of mRNAs species'. Thus, we also have had to relax parameters and allow an increased number of false positives among the identified proteins (see Section 1.3.4.2). I have included proteins quantified with both methods in the analyses of the next chapter (Chapter 6).

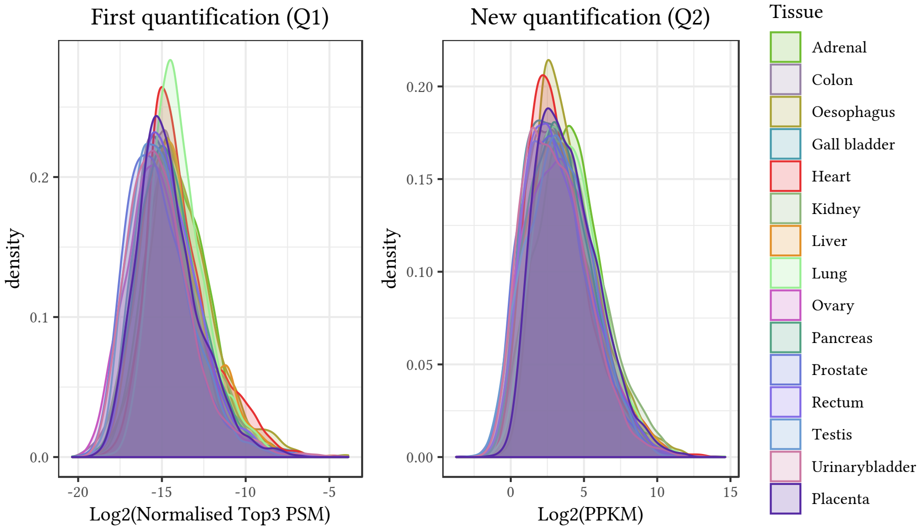


Figure 5.7. **Distribution of the protein expression levels with two different methods.** On the left, the protein expression levels distribution for the adult tissues of the Pandey Lab data with our first quantification method (described in Section 2.4.2). This figure is similar to Figure 3.3(c) but without the fetal tissues. On the right are the protein expression levels of the same tissues that have been computed with the new quantification method. The overall shape of the density plots is very similar between the two approaches.

Before moving on to the integration of the proteomic and transcriptomic data in the next chapter, I have carried out a few comparisons between the first and the new quantification methods. As shown in Figure 5.7, the densities of distribution of protein expression levels per tissues have similar shapes between the two methods.

Figure 5.8 presents the number of quantified proteins per tissue. The new method allows quantifying for some tissues up to more than twice the number quantified by the first method. This proportion is consistent with the total number of proteins identified across the fifteen adult tissues by the first method (6,436) and the new one (12,290).

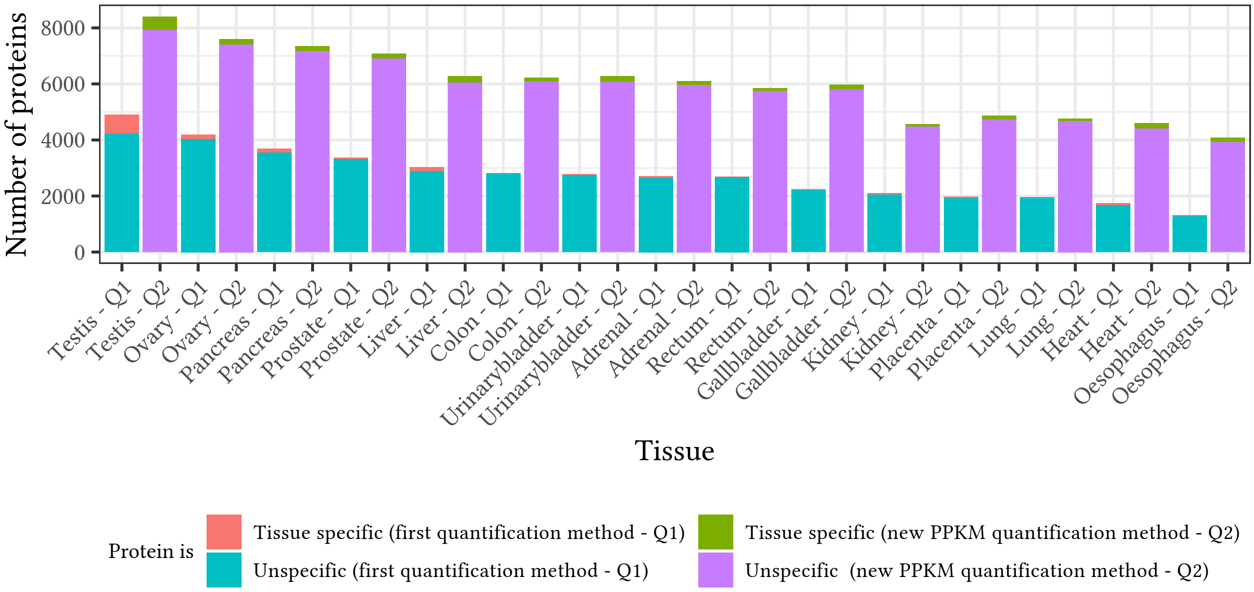


Figure 5.8. **Comparison of the impact of the quantification method on the protein distribution per tissue for the Pandey Lab data.** This figure partly reproduces the *Pandey Lab* part of Figure 5.4. Indeed, the turquoise/red $Q1$ bars are the adult tissues of the Pandey Lab data quantified with the first quantification method (described in Figure 5.4). The purple/green $Q2$ bars are their equivalent to the new quantification method. The new method allows a considerable increase of the identified and quantified proteins number, sometimes more than twice as many as the first quantification. On the other hand, the rank orders of the total and tissue-specific protein numbers per tissue are quite similar between the two methods.

I have also checked the presence of OR in the Pandey Lab data with the new quantification method; only two ORs are present: OR1M1 in *Kidney* and OR13C4 in *Liver*. Their presence may be artefactual, or it may be an issue with the annotation. Indeed, while these two proteins are missing from all tissues — either at RNA or protein levels — in *The Human Protein Atlas*[4] [Uhlén, Fagerberg, et al., 2015], their mRNAs are present in the *Baseline expression* of *EBI Gene Expression Atlas*[5] [Petryszak, Keays, et al., 2015] in the human

[4] *The Human Protein Atlas* — https://www.proteinatlas.org/

[5] *EBI Gene Expression Atlas* — https://www.ebi.ac.uk/gxa/

Chloroid plexus at 10 post-conception weeks (HDBR developing brain — ArrayExpress ID: E-MTAB-4840) and in one sheep *Testis* sample (ArrayExpress ID: E-MTAB-3838). In addition, OR1M1 seems to be expressed in the *Blood* of the green monkey (*Chlorocebus sabaeus* — ArrayExpress ID: E-MTAB-4404). Both proteins are also found to be up or down regulated at transcript level in different tumoral samples (*Differential expression* tab of EBI Gene Expression Atlas). See also the digital supporting data.

I have also assessed the consistency of expression measurements across Pandey Lab and Kuster Lab data and their common tissues as I have done for the three datasets with the first quantification (presented in Section 5.1.3). The new quantification gives similar results as shown in Figure D.12. Besides, as shown in Figure D.13, more than half of the proteins quantified in each dataset within a given tissue is also quantified in the other dataset.

5.3 UBIQUITOUS AND TS PROTEINS

Previously, W. Liu et al. (2014) had compiled a list of 627 TS and 1,093 housekeeping proteins from the expression data released originally by Pandey Lab [M.-S. Kim et al., 2014]. The data processing has a significant impact on protein identification; in our first version of Pandey Lab data, I have found 534 housekeeping (ubiquitous) proteins and 1,491 TS proteins, and for the PPKM quantification: 2,057 ubiquitous and 2,640 TS proteins. I provide as digital supporting data the list of the TS and ubiquitous proteins.

5.4 DISCUSSION AND CONCLUSION

In this chapter, I have reviewed human proteome data from three projects presented in Chapter 2, that have been reprocessed by Dr James Wright with two pipelines for the two largest studies Pandey Lab data [M.-S. Kim et al., 2014] and Kuster Lab data [Wilhelm et al., 2014]. Currently, state-of-art bottom-up label-free MS proteomics captures human tissues expression as a fragmented and disparate universe. Our first processing pipeline, based on the Top3 quantification method presented in Section 2.4.2, appears more reliable than some of the original authors'. The original Pandey Lab data was disputably quantifying ORs in many tissues [Ezkurdia et al., 2014]. No trace of ORs was found in any of our reprocessed data. I have also described our new quantification method (PPKM), which allows us to estimate the expression of nearly twice as many proteins than the first one we used.

For both quantification methods, the technical variability is generally stronger than the biological interstudy signal even for similar tissues. Besides, across the different tissues, about half of the proteins are consistently observed in the same tissues at least in two datasets.

Even when limited to the protein identification only, the general lack of repeatability and reproducibility has been well reported and described for technical and biological replicates in MS proteomics (*e.g.* Tu, J. Li, Sheng, et al. (2014) and Tabb, Vega-Montoto, et al. (2010)). Canterbury et al. (2014) report that the intra-assay variation between two technical replicates for complex mixtures can be at least 50%; different runs of the same sample or experiment are often produced to raise the interstudy results repeatability and confidence.

Thus, beyond the quantification method I have developed with the help of Dr James Wright by drawing on RNA-Seq ones, there is a definite need for new MS protocols and quantification methods for baseline expression[6] to correct the extreme variability and ease the integration of proteomics data across studies.

[6] Normalisation methods for differential expression analysis are usually unsuited for baseline expression studies. See Välikangas et al. (2018b) for possible differential expression quantification methods.

Scientists like ripping problems apart, collecting as much data as possible and then assembling the parts back together to make a decision.

Shirley M. Tilghman

6 INTEGRATION OF TRANSCRIPTOMIC WITH PROTEOMIC DATA

After assessing the similarity of the human gene expression profiles across various tissues at transcriptomic level (with RNA-Seq studies in Chapter 4) and proteomic level (with *bottom-up* MS studies in Chapter 5), my next step is to examine how these gene expression profiles compare between these two different biological layers.

One major aim of this study is to assess how the correlations between the transcriptome and proteome described in the literature, mostly measured in cells, hold at the tissue level. Moreover, good correlations may potentially lead to the development of new strategies. These may use the expression levels of mRNA as proxies to estimate protein expression, which is generally difficult to measure directly (see Section 1.3).

I have performed the integration and all the analyses presented in this chapter under the supervision of Dr Alvis Brazma and Dr Jyoti Choudhary.

A few closely related studies [Kosti et al., 2016; Franks et al., 2017; D. Wang et al., 2019] have been published while I was working on the integration of the non-diseased human transcriptome and proteome. As their analyses rely on the same data sets (*i.e.* Uhlén, GTEx, Pandey Lab data) that I include in my work, I describe and discuss together my results and theirs whenever relevant.

Communication to the community derived from this chapter

- (paper) Mitra P. Barzine, Kārlis Freivalds, James Wright et al. (2020). 'Using Deep Learning to Extrapolate Protein Expression Measurements'. *Proteomics* 20 (21–22), e2000009
- (submitted paper) Andrew F. Jarnuczak; Hanna Najgebauer; Mitra Barzine; Deepti J. Kundu; Fatemeh Ghavidel; Yasset Perez-Riverol; Irene Papatheodorou; Alvis Brazma; Juan Antonio Vizcaíno An integrated landscape of protein expression in human cancer
- (poster) CSHL Biology of Genomes 2015 — A feasibility study: Integration of independent human RNA-Seq and proteomic datasets
- (talk) GTEx meeting 2017 — A. Brazma Correlating transcriptome and proteome in human tissues
- (poster) HUPO 2018 — Jarnuczak et al. An integrated atlas of protein expression in human cancer derived from publicly available
- (poster) ECCB 2018 — Viksna et al. An integrated approach to missing data imputation in quantitative proteomics experiments
- (poster) RECOMB 2018 — Viksna et al. Deep learning for protein abundance prediction using Gene Ontology and RNA abundance information

An on-going debate in the literature is whether good correlations of expression levels prevail between mRNAs and proteins [Uhlén, Hallström, et al., 2016]. The implicit assumption of a proportional relationship is persisting as the many remaining technological limitations prevent rigorous testing [Vogel and Marcotte, 2012]. To date, the existence or concentration of a given mRNA transcript is usually insufficient to ensure detection of the protein in a sample.

On the one hand, Ramakrishnan et al. (2009) report that mRNAs abundance are roughly sufficient to predict the protein presence or absence from a sample and Vogel, Abreu, et al. (2010) that mRNA level estimations and sequence features are enough to predict two-thirds of the human protein abundance variation.

On the other hand, the literature fails to report any high correlation between the transcriptome and the proteome for any organism. Previous investigations found low or no correlation between the measured expression profiles of the mRNAs and proteins in human [Anderson et al., 1997; G. Chen, Gharib, et al., 2002; Tian et al., 2004; Pascal et al., 2008; Gry et al., 2009; Lundberg et al., 2010], other mammals [Ghazalpour et al., 2011], and across many other species [Gygi, Rochon, et al., 1999; Maier, Güell, et al., 2009; Maier, Schmidt, et al., 2011; Yeung, 2011; Palmblad et al., 2013; Freiberg et al., 2016].

In their encompassing reference experiment, Schwanhäusser et al. [Schwanhäusser et al., 2011; Schwanhäusser et al., 2013] present rather moderate correlations ($r^2 \leq 0.41$, *i.e.* $r < 0.64$) and highlight that mRNA levels explain only about 40% of protein variations they have observed.

Other studies have explored the mRNAs and proteins relationship in answer to stimuli [Marguerat et al., 2012] or with an increased focus to post-transcriptional regulations (including degradation rates) [Jovanovic et al., 2015]. While many other regulatory processes may occur (*e.g.* translation rates), post-transcriptional modifications and technical noise are (still) perceived as the probable primary sources of mRNA/protein concentration discrepancies [Vogel and Marcotte, 2012; Plotkin, 2010].

Joint studies of transcriptome and proteome have already helped to highlight links between genotype and phenotype [Vogel and Marcotte, 2012]. However, the mitigated results reported above may explain the focus shift of many subsequent studies. While previous efforts were about linking the actual expression levels, more recent studies primarily have mostly compared qualitative attributes of given proteins and related mRNAs. Examples include the comparison of the presence or absence of mRNAs and their proteins in specific conditions or tissues [Santos et al., 2015; Freiberg et al., 2016; Uhlén, Fagerberg, et al., 2015] or the comparison of their differential expression profiles across identical sets of conditions [Väremo et al., 2015].

All (or almost all) aforementioned studies have turned to cells for their joint analyses of transcriptome and proteome. In contrast, the analyses and integration I present in this chapter are based on tissue studies.

6.1 DATA AND PRINCIPAL ANALYTICAL APPROACHES

Since the human proteome drafts [M.-S. Kim et al., 2014; Wilhelm et al., 2014] in 2014, we have an unparalleled availability of large-scale tissue studies both at the transcriptomic and proteomic layers to explore and integrate together (see Chapter 2). While these data are independent (collected from various individuals, prepared, and characterised by different laboratories), their combined study may help to shed light on the relationship between the transcriptome and proteome at the tissue level. Using different sources for the transcriptome and proteome increases the overall technical noise, but it may also help to highlight relevant biological signals (as they need to be stronger than the noise and batch effects to be captured).

In Chapter 4, I show that the transcriptome RNA-Seq datasets present high interstudy tissue correlations (median value for Pearson: $r_{\mathcal{W}_1} = 0.75$; $r_{\mathcal{W}_2} = 0.85$ – Spearman: $\rho_{\mathcal{W}_1} = 0.88$; $\rho_{\mathcal{W}_2} = 0.93$). For this chapter analyses, I only consider the datasets with the highest similarity (highest correlations) that incidentally comprise the greatest number of tissues and are the two most recent studies, *i.e.* Uhlén et al. [Uhlén, Fagerberg, et al., 2015] and GTEx [Melé et al., 2015] data.

To compensate for the shortfalls in the study design implied by the reuse of published data[1], I use both Uhlén et al. and GTEx data to filter out mRNAs with high interstudy variability for identical tissues. Whether this variability is technical or biological is irrelevant; in both cases, interpreting the relationship between a highly variable mRNAs and its protein from another dataset remains hard to interpret. For these mRNAs, it is impossible to explain the observed variability between the two transcriptomic datasets. Indeed, any result is subject to the transcriptomic dataset chosen for the comparison with the proteomic one. Furthermore, the comparison of the two transcriptomic data may give a reference, *i.e.* an ideal case scenario, for the proteomic/transcriptomic one.

On the other hand, as shown in Section 5.1.3, the technical variability prevails over the biological signal of same-tissue samples for the available high-throughput proteomics. With the current technological state, different tissues from the same proteomic study are more likely to present a higher correlation than the same tissues from two different studies.

To avoid an overly restricted protein set for the following analyses, I only include one proteomic study: Pandey Lab [M.-S. Kim et al., 2014]. All its samples have been run through the same MS platform and with the same protocol. Moreover, it presents more homogeneous protein distributions (see Figure 3.3 and Figure 5.7) and quantifies more proteins per tissue (Figure 5.4) than the two other datasets. Since a current major limitation of bottom-up MS proteomic studies is the possible lack of detection of proteins for various

[1] Independent data also means different collection and sampling processing methods and lack of information on the samples population background.

reasons (see Section 1.3.2), the higher number of detected proteins in Pandey Lab data suggests that this dataset has a higher quality than the two others.

Though I include one proteomic dataset only, as the literature reports that the proteome is more conserved than the transcriptome (across individuals and species) [Laurent et al., 2010; Y. Liu et al., 2016], this data collection ought to provide a crude estimate of the extent of observations that hold from cell to tissue level.

This chapter integrates and analyses the matching pairs of mRNA/proteins of the common set of tissues between Pandey Lab and the two transcriptomic datasets.

6.1.1 *Overlapping set of tissues for the three datasets*

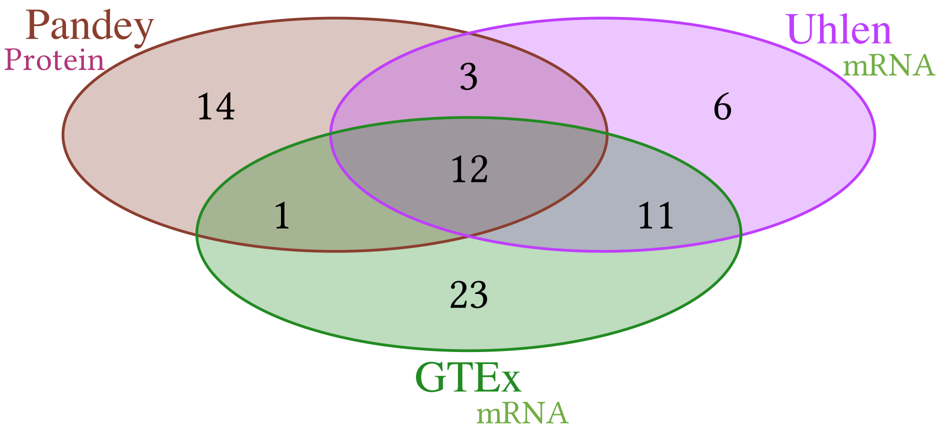


Figure 6.1. **Number of shared and unique tissues between the proteomic (Pandey Lab) and the transcriptomic (Uhlén et al. and GTEx) data.**

All analyses include the twelve tissues shared between the three datasets (*Adrenal gland*, *Urinarybladder*[2], *Colon*, *Oesophagus*, *Heart*, *Kidney*, *Liver*, *Lung*, *Ovary*, *Pancreas*, *Prostate* and *Testis*).

In a few cases, I have also extended the analyses to three additional tissues (*i.e. Gallbladder*, *Placenta* and *Rectum*) by including the Uhlén et al. data on the transcriptomic side only.

6.1.2 *Matching pairs of mRNAs and proteins*

To avoid unnecessary biases (described in Section 3.3), I only consider the mRNAs (*i.e.* RNAs with a *protein-coding* biotype — Ensembl 76) for the following analyses. Moreover, since missing data is common for proteomics [Lazar, Gatto, et al., 2016], only proteins that are detected in each dataset in at least one of the included tissues are considered for further

[2] May also be referred to as *Urinary Bladder*

analyses.

Besides, while in the transcriptomics studies biological replicates of each tissue have been processed as individual RNA-Seq libraries, in the proteomic one, the biological replicates have been pooled per tissue before any MS profiling. Thus, to prevent an unbalanced number of samples biasing the integration analyses (see Chapter 3), I use 'virtual references', *i.e.* TREPs[3] that I computed for each tissue by taking the median values of each gene across the biological replicates (see Section 3.3.4).

As exposed in Chapters 2 and 5, all the proteomic quantifications have been provided by Dr James Wright.

The first quantification follows state-of-the-art practices with stringent parameters (described in Section 2.4.2) since accurate protein identification is paramount for reliable proteome exploration. The protein levels are the intensity of their top three unique peptides normalised within-sample. Figure 6.2 presents the genes overlap across twelve shared tissues between the Pandey Lab's proteins quantified through this first method and Uhlén et al.'s and GTEx's mRNAs quantified with *HTSeq-count* (see Section 2.4.1.3). Figure 6.3 is the same analysis across the fifteen shared tissues between Pandey Lab and Uhlén et al. data.

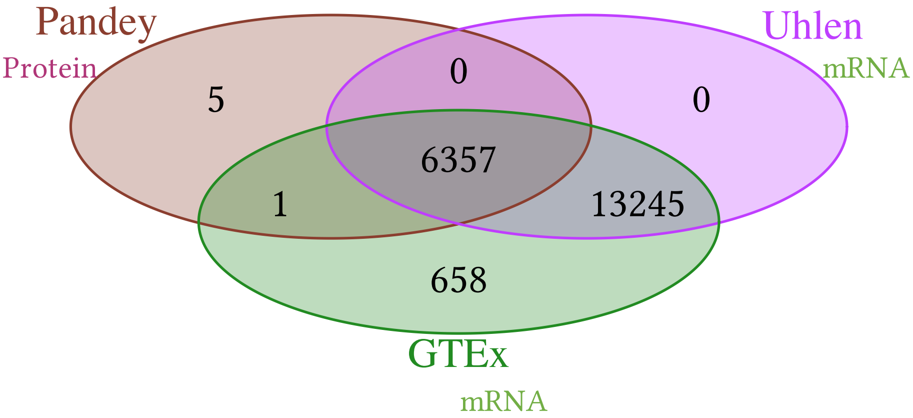


Figure 6.2. **Distribution of the unique and shared proteins of Pandey Lab data and mRNAs from Uhlén *et al.* and GTEx ones across their twelve shared tissues.** There are 6,357 matching gene products between the three datasets. Only 5 proteins have apparently no matching partners in the Uhlén et al. or GTEx data.

This first proteomic quantification is following robust guidelines, and both figures show that almost all the genes with an observed protein also have an observed mRNA. However, only about 32% of the quantified mRNAs in the Uhlén et al. and GTEx data have a corresponding protein detected in the Pandey Lab data.

[3] TREP: tissue reference expression profile

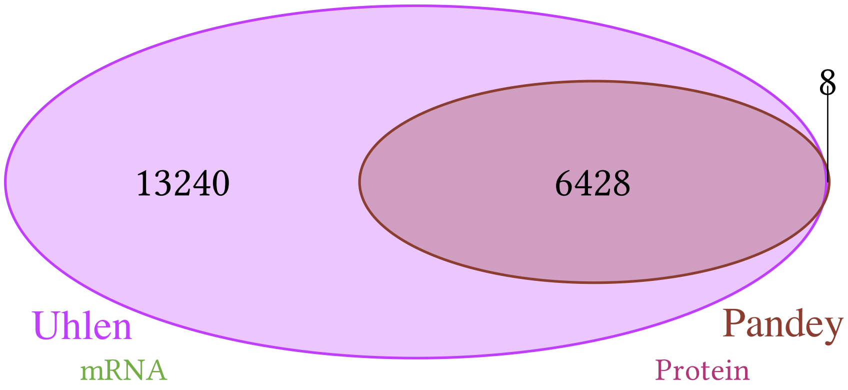


Figure 6.3. **Distribution of the unique and shared proteins/mRNAs for Pandey Lab and Uhlén *et al.* across their fifteen shared tissues.** The number of matching pairs (6,428) and proteins that lack a counterpart in the transcriptomic data (8) are similar regardless of how many different transcriptomic data is included (see Figure 6.2).

Once I learned more about the bioinformatic challenges of bottom-up proteomics (described in Section 1.3.4), I chose to be more flexible with the identification and quantification methods to increase the number of proteins included in my analyses. As I aim to integrate independent proteomics with transcriptomics, I mostly focus on robust expression between the two biological layers since discrepancies in this study context are hard to interpret. While artefacts may persist, further analyses with targeted proteomics (see Section 1.3) can help prune or validate the results.

I have drawn on RNA-Seq transcriptomic approaches to devise a new quantification method, which is described in Section 5.2 and implemented by Dr James Wright. The method takes advantage of the *degenerate* peptides[4] that are distributed across possible protein parents in proportion to their *unique* peptides. The method produces normalised values of the protein expression levels (whose unit is the PPKM, *i.e.* PSMs Per Kilobase of gene per Million).

As shown in Figures 6.4 and 6.5, while the number of quantified proteins with our new method covers about 62% of Uhlén et al.'s and GTEx's quantified mRNAs, the number of proteins for which no mRNA was detected in the transcriptomic data remains marginal.

Whether it reflects the biological reality or is solely due to RNA-Seq technology being more sensitive than bottom-up MS alone, current techniques detect more individual mRNAs than proteins as confirmed by Figures 5.4 and C.1. Thus, it may be surprising that a few proteins lack a match in the transcriptome data. Several possible explanations exist.

[4] See Section 1.3.4.3.

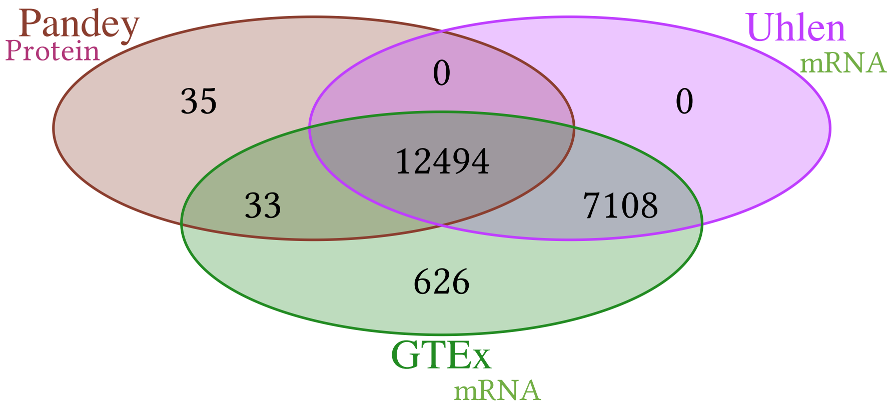


Figure 6.4. **Distribution of the unique and shared proteins/mRNAs across twelve shared tissues** between Pandey Lab (**new quantification method**), Uhlén et al. and GTEx data.

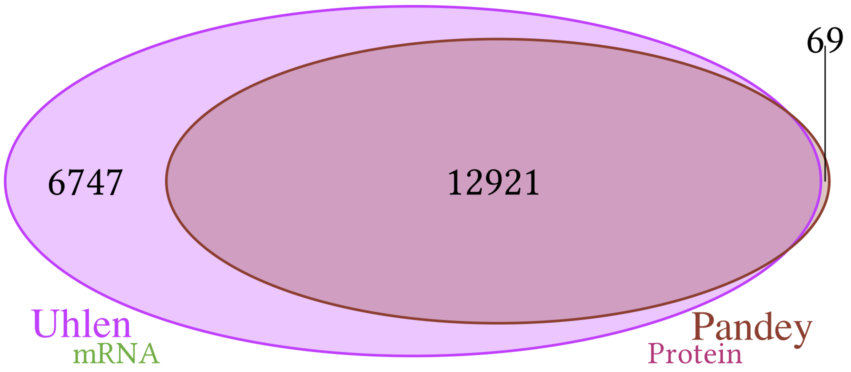


Figure 6.5. **Distribution of the unique and shared proteins/mRNAs across fifteen tissues between the Pandey Lab (new quantification method) and Uhlén et al. data.**

Artefacts or technical issues are the most likely. For example, the annotation might miss the matching RNAs definitions or defines them with another biotype than *protein-coding*[5]. Or, peptides and mRNA reads may be assigned to different gene IDs. Alternatively, the mRNAs are present in the sample, but the library preparation has missed their capture (see Section 1.2.1). Or even, the presence of proteins in the sample is a false positive or the result of contamination.

However, biological processes might also explain the mismatches. One example is the case of mRNAs with short half-lives while their proteins are very stable. Another possible explanation is that the original location of the proteins is different from the tissue in which they were detected (like hormones or cytokines).

Lastly, as the transcriptomic and proteomic samples are independently sourced, a protein may be specific to an individual or a population. This last hypothesis is the most unlikely as there are several biological replicates on the transcriptomic side. A mixture of the previous causes is also plausible.

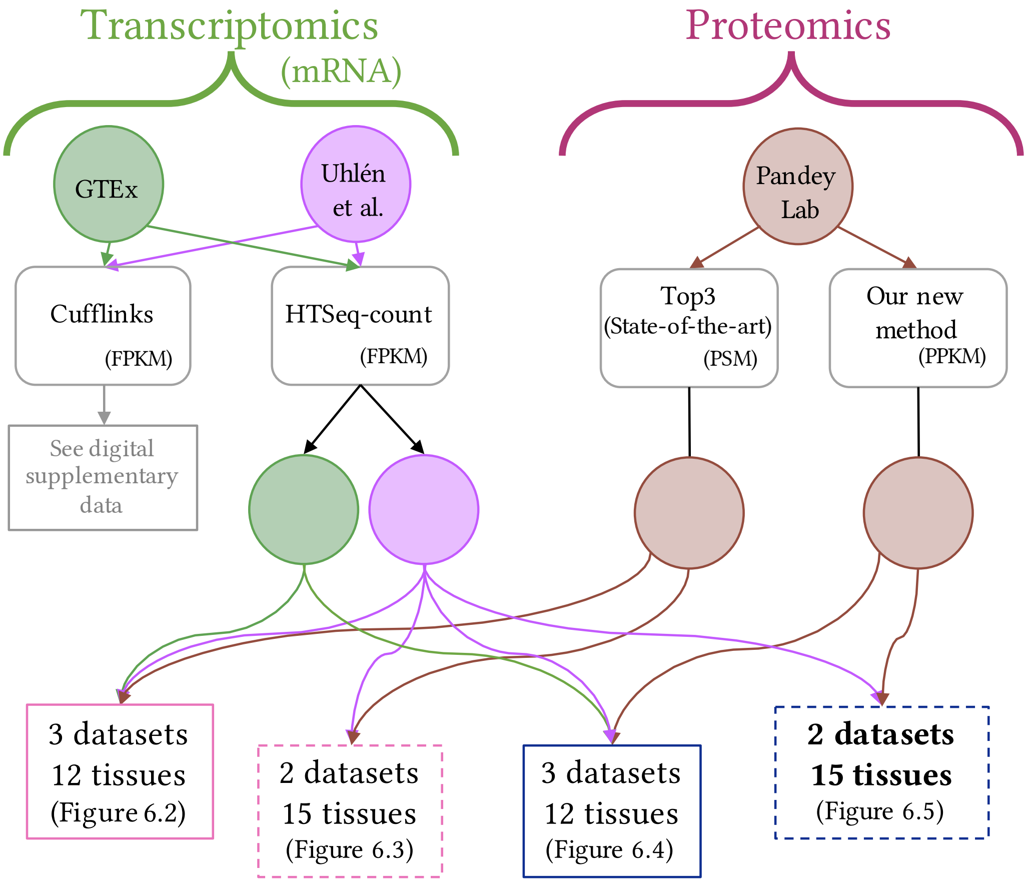

Transcriptomics
(mRNA)
Proteomics
GTEx
Uhlén et al.
Pandey Lab
Cufflinks (FPKM)
HTSeq-count (FPKM)
Top3 (State-of-the-art) (PSM)
Our new method (PPKM)
See digital supplementary data
3 datasets 12 tissues (Figure 6.2)
2 datasets 15 tissues (Figure 6.3)
3 datasets 12 tissues (Figure 6.4)
2 datasets 15 tissues (Figure 6.5)

Figure 6.6. **Overview of different studied datasets combinations.**

[5] E.g. *XXYAC-YRM2039.2* annotated as *unprocessed pseudogene* and now known as *WASH1* since Ensembl 77 (October 2014) or *TRAJ61* which is annotated as *TR J gene*.

I exclude the unmatched proteins and mRNAs from further analyses. Table E.1 provides the unmatched protein lists for the Ensembl 76 annotation.

Unless otherwise stated, to avoid issues exposed in Section 3.3.1, I also remove all the proteins and mRNAs of the mitochondrial genome from the subsequent analyses.

Note that Figure 6.6 presents an overview of the various datasets combinations presented in Figures 6.2 to 6.5.

6.1.3 *Tissue-centric and gene-centric approaches*

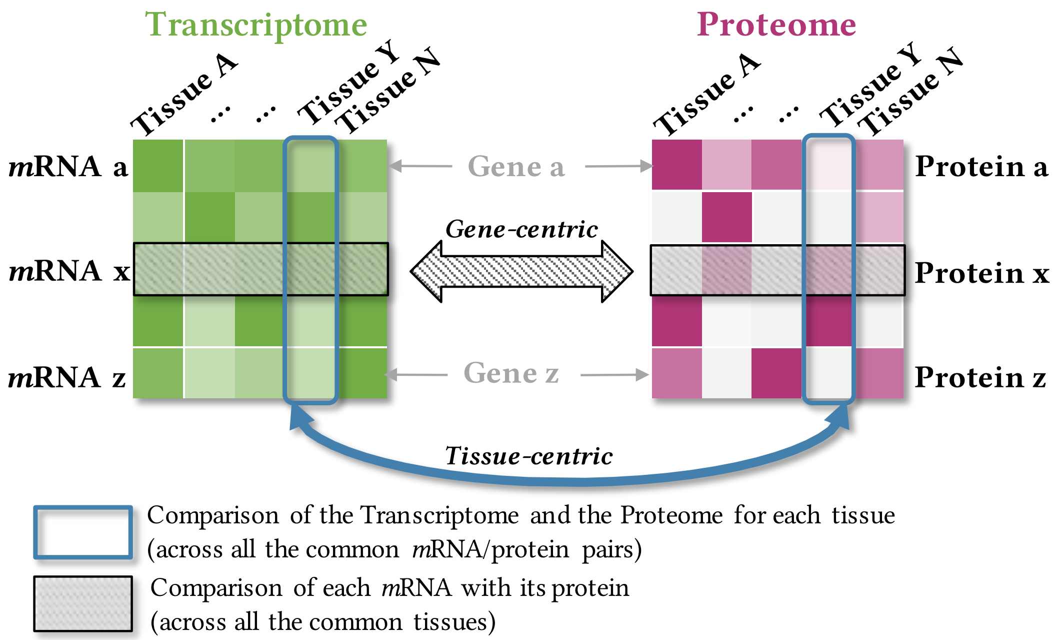

Transcriptome
Proteome
Tissue A
...
...
Tissue Y
Tissue N
Tissue A
...
...
Tissue Y
Tissue N
mRNA a
mRNA x
mRNA z
Gene a
Gene-centric
Gene z
Protein a
Protein x
Protein z
Tissue-centric
Comparison of the Transcriptome and the Proteome for each tissue (across all the common mRNA/protein pairs)
Comparison of each mRNA with its protein (across all the common tissues)

Figure 6.7. **Approaches summary of the expression comparison between the transcriptome and proteome.** *Tissue-centric* analyses focus on how the transcriptome and proteome relate to each other within the same tissue. *Gene-centric* analyses study for each gene how its mRNA expression levels across all (or a subset of) the tissues may relate to the quantified expression levels of its corresponding protein.

Figure 6.7 summarises the two analytical approaches I use to compare transcriptomic and proteomic data. The *tissue-centric* approach compares for each tissue the global expression of its transcriptomic landscape to its proteomic one. In contrast, the *gene-centric* approach compares for each gene its expression levels in mRNA and protein across all the tissues.

Confusion can arise when integrating proteomics and transcriptomics. Hence, it is essential to define the taken approach clearly [Y. Liu et al., 2016].

6.2 FAIR CORRELATIONS BETWEEN INDEPENDENTLY SOURCED PROTEOMICS AND TRANSCRIPTOMICS OF HUMAN TISSUES

For the first tissue-centric analysis, I assess for each tissue the relationship between the expression of its proteome and transcriptome through the correlation of the protein expression values with their corresponding mRNA ones.

After scaling with $\log_2(x+1)$, I compare proteomic and transcriptomic TREPs from identical and random tissue pairs, which are similar and roughly correspond to Gaussian distributions as illustrated by Figures 3.2 and 5.7.

Figure 6.8 presents the correlation distribution range of transcriptomic and proteomic TREPs from identical and random pairs of tissues, both with Spearman and Pearson correlation methods (see Appendix B.1).

Although transcriptomics and proteomics have independent sources, the Spearman correlations of the same tissues TREPs are equivalent to correlations in cell studies [Lundberg et al., 2010; Schwanhäusser et al., 2011] where the same sample provides mRNAs and proteins. Regardless of the protein quantification method (Top3 [Silva et al., 2006] or PPKM — equation (PPKM definition) on page 126), the median Spearman correlation coefficients are above 0.5 for matched proteomic and transcriptomic TREPs (also referred to as *same-tissue pairs*). The unscaled data presents identical outcomes (see Table E.2 and Figure E.4).

The Pearson correlation is closer to the literature for our new PPKM quantification than for the Top3 quantification. The PPKM Pearson correlation averages above 0.5 [min: 0.38 (*Oesophagus*) ; max: 0.61 (*Liver*)] (and is within [min: 0.45 (*Oesophagus*) ; max: 0.67 (*Liver*)] for the untransformed data).

As tissue proteomic samples can present high correlation without being related in any manner (see Chapter 5 and figure D.10), a Welch t-test [Welch, 1951] allows assessing the significance of the correlation for the same-tissue pairs by comparison to random tissue pairs. The one-sided Welch's Two Sample *t*-test[6] allows rejecting the null hypothesis H_0 (the means of the correlation coefficients for same-tissues pairs are identical or lower to random tissues pairs). Irrespective of the protein quantification or computational methods, all the same-tissue pairs correlations are significant (p-value $< 5.10^{-5}$, except for Pearson correlation with Top3 quantification where p-value < 0.05 — see Table E.2).

The previous correlation distribution may imply a modest relationship between these independent proteomics and transcriptomics, but the same-tissue pairs scatterplots (*e.g.* Figure 6.9) show tighter links than first suggested. Besides, these scatterplots share a coarse profile despite the wide correlation ranges.

Figure 6.9 illustrates the comparison of expression for *Kidney* between transcriptomics (Uhlén et al.) on the x-axis and proteomics (Pandey Lab — PPKM) on the y-axis. *Kidney*'s

[6] See Appendix C

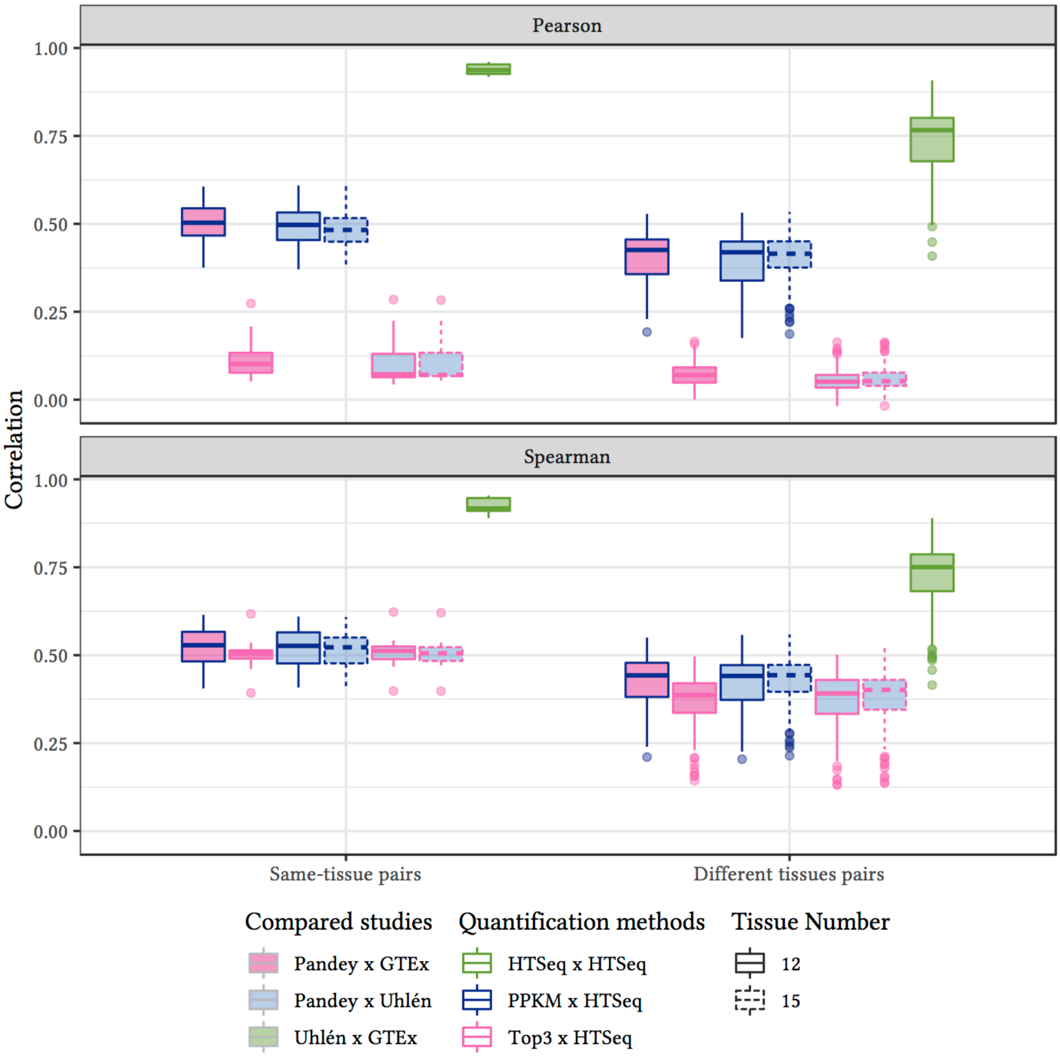


Figure 6.8. **Distribution of Pearson and Spearman correlation coefficients for same-tissue proteomic and transcriptomic pairs versus random tissue pairs (log$_2$-scaled data).** Depending on the protein quantification method, there are two types of distribution ranges for the Pearson correlations. Top3 quantification method provides a lower correlation (mean $\approx$ 0.11). The PPKM method (Section 5.2) produces higher correlations (mean $\approx$ 0.5). All the Spearman correlation ranges between same-tissue proteomic and transcriptomic TREPs are quite similar, regardless of the method quantifying the proteins. The median of Spearman correlation is 0.52. With the Top3 quantification (*i.e.* pink countered boxes — Top3 x HTSeq), two outliers are noticeable, and they are common to the three comparisons, Pandey x Uhlén (12 tissues and 15 tissues) and Pandey x GTEx (12 tissues): the lowest Spearman correlation is *Oesophagus* ($\rho = 0.39$) and the highest *Liver* ($\rho = 0.62$). Both for the Pearson and Spearman correlations, even when the correlations are very low, same-tissue pairs always have higher correlations than different (random) tissues pairs (all p-values computed with Welch t-test <0.05 — see Table E.2). Thus, even the lowest same-tissue correlations are significant. The green boxplots, comparing the two transcriptomic datasets, are only represented for reference purposes.

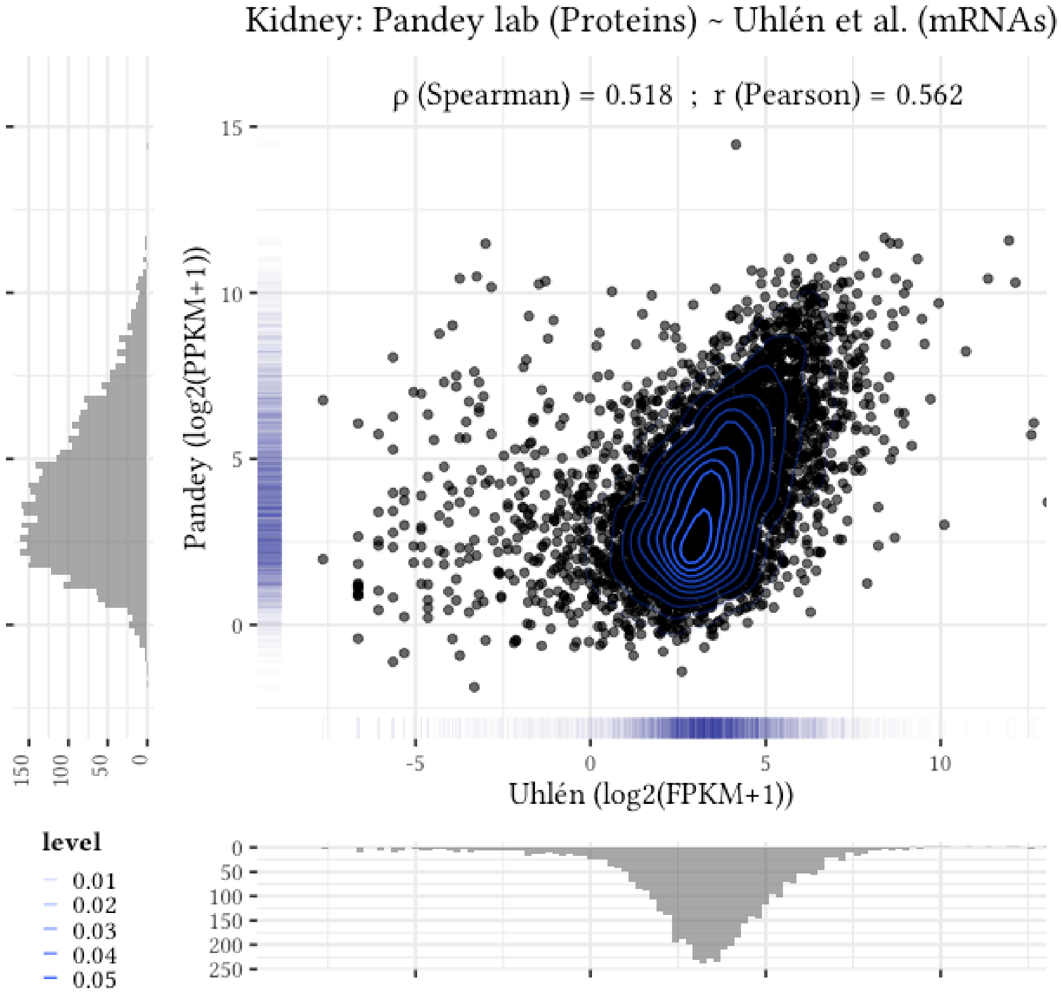


Figure 6.9. **Scatterplot of protein (Pandey Lab — PPKM quantification) and mRNAs (Uhlén *et al.*) expression for Kidney.** Each point of this scatterplot represents a gene; it has the $\log_2$-transformed expression value of the corresponding Uhlén et al. mRNA (FPKM) on the x-axis and the $\log_2$-transformed expression value of the Pandey Lab protein (PPKM) on the y-axis. Most of the mRNA/protein pairs are distributed in an area that can be fitted by a linear function with a positive slope, which indicates a high correlation between mRNAs and proteins expression levels. However, genes with lower expressed mRNAs have a less associated expression between their protein and mRNA, in particular, mRNAs that are expressed below 1 FPKM (*i.e.* below 0 on the x-axis). On the other side, genes with the highest expressed mRNAs may present a saturation effect (Section 1.3.2) in the quantification of the protein expression. The highest expressed protein is HBB (*i.e.* Hemoglobin Subunit Beta), which is also found in the five highest expressed proteins in all the other tissues. Possibly, its presence is due to remaining erythrocytes in the samples. On the outer parts of the scatterplot, there are the respective distribution densities of the proteins and the mRNAs. Whilst the correlation calculation includes every pair of mRNA and protein, the plot excludes any pair with an unexpressed mRNA or protein to optimise the visualisation. Figure E.2 presents an overview of the other tissue scatterplots.

correlation coefficients stand in the middle of the range regardless of the considered studies, protein quantification or correlation methods involved in the comparison.

A linear function with a positive slope (not drawn) can fit the bulk of the points. Indeed, the expression of most mRNAs and proteins in a tissue are highly associated with the exception of the lowest (< 1 FPKM) and a number of the highest measured mRNAs.

Besides the mismatching sampling sources, other possible explanations for the observed divergences are technical limitations (such as protein saturation effect, see Section 1.3.2), translational noise (see Section 3.3.3) or a consequent half-life difference between the mRNA and its protein.

Although the number of genes presenting lowly associated levels of mRNA/protein expression is rather limited, it is enough to impair the Pearson and Spearman correlation coefficients.

Systematic exclusion of lowly associated pairs of mRNAs and proteins is impractical and arguable as they are inconsistent from one tissue to another. Case-by-case treatment will be necessary.

Removing the lowly expressed mRNAs (< 1 FPKM) only marginally changes the correlation coefficients, *e.g.* for *Kidney*, when considering the PPKM quantification for the proteins, the Pearson correlation increases from 0.56 to 0.58, while the Spearman correlation is relatively unchanged (0.51 instead of 0.52). There are similar changes observed when considering the more conservative Top3 protein quantification. The Pearson correlation $r = 0.18$ increases to 0.21. The Spearman correlation remains unchanged ($\rho = 0.52$).

Both transcriptomic studies (Uhlén et al. or GTEx) providing alike results, I describe for most of the following analyses the data combination that provides the greatest number of tissues and genes to study, *i.e.* the fifteen-tissue set between Uhlén et al. and Pandey Lab data quantified with the PPKM method.

The other combinations (provided in Appendix E or electronic format) may diverge for individual genes through the various combinations, but the general trends are identical.

I focus on Pearson correlation over Spearman correlation in the following parts since the results for the PPKM quantification are globally similar for both.

6.2.1 *Mixed biological signal between the proteome and transcriptome across the tissues*

As shown in Figure 6.10, for nine tissues (in yellow) transcriptomic and proteomic expressions correlate better in matching tissues. For four other tissues (*Colon*, *Lung*, *Oesophagus* and *Urinarybladder* — in dark pink), only the proteomics correlate the best with the matching transcriptomics, while the transcriptomics correlates better with other proteomics tissues. The remaining two tissues have their proteomics correlating as much (*e.g. Gallbladder*) to other tissues or more to transcriptomics from other tissues (*Rectum*).

While the different correlation methods lead to similar result trends, individual differences persist. In a few cases, *e.g. Heart*, these slight differences may considerably change the

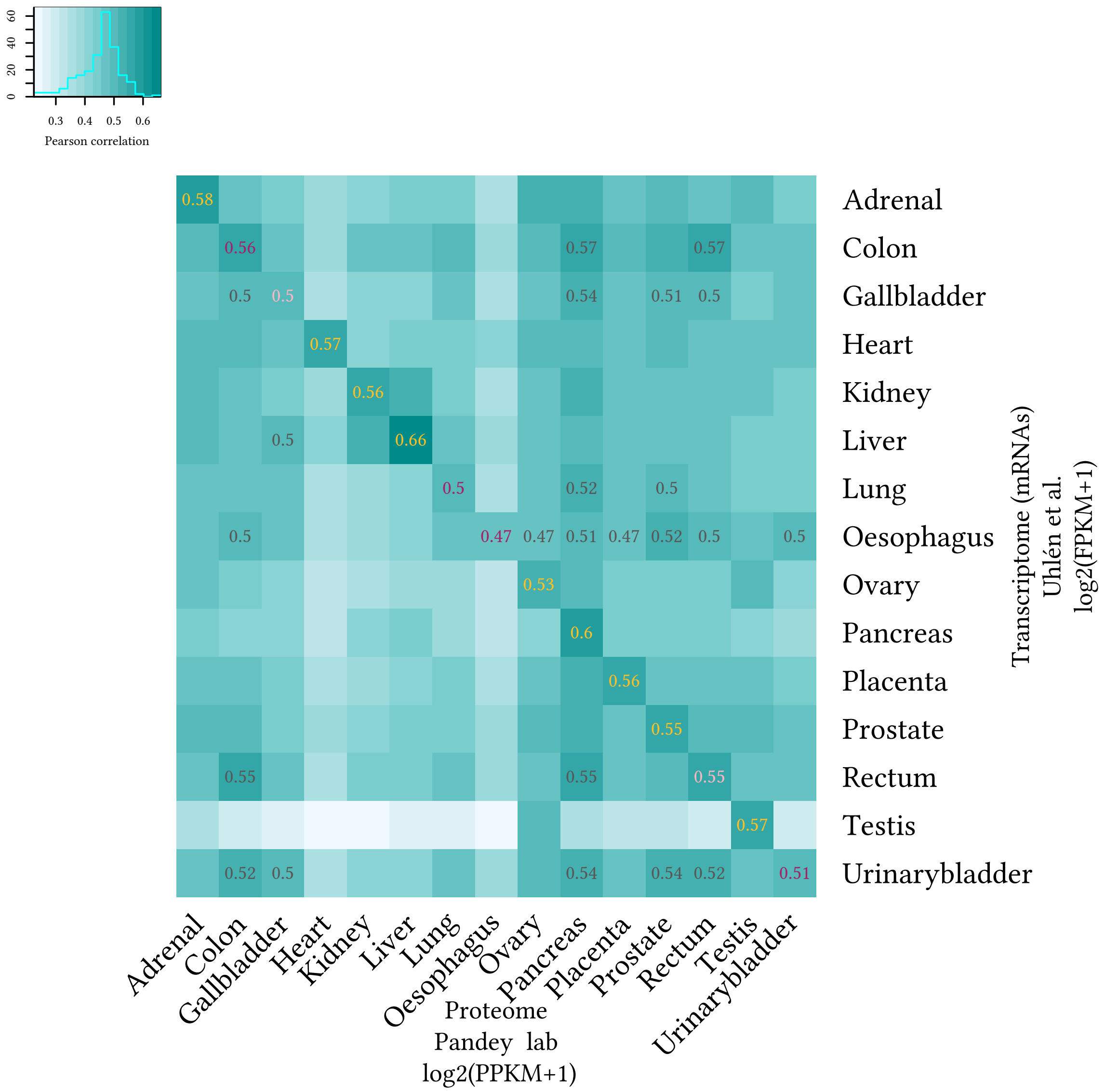


Figure 6.10. **Heatmap based on the Pearson correlation between protein and mRNAs expression (alphabetically ordered tissue).** Correlations for same tissue pairs (diagonal) are highlighted in yellow when the highest observed correlations are between the matching proteomics and transcriptomics pairs; in dark pink, when the proteomics correlates the best with the matching transcriptomics. When other higher correlations are observed for a tissue proteomics or transcriptomics they are in given grey.

relative correlation ranking order of the TREPs (see Figure E.5).

In the following sections, I explore several avenues to identify possible factors that influence the association strength between the proteome and transcriptome.

I first study the effect of tissue composition (in proteins and mRNAs) on the correlations. I begin with the assessment of the impact of the proteins and mRNAs that are found in one tissue only, before looking into the tissue-specific (TS) proteins and mRNAs.

Then, in a more quantitative approach, I examine more closely how the mRNA expression profiles relate to their respective protein ones.

6.2.2 *Influence of the expression breadth on the tissue mRNAs/proteins correlation*

In Chapter 5, I have shown that the protein expression of both different tissues and same-tissue pairs are sharing a similar correlation range (see Figure D.11). In this context, genes expressed in a small number of tissues (both as a protein and mRNA) can have a significant impact on the correlation and may explain the mitigated results.

The expression breadth of a gene is the number of tissues and cell lines within which the gene is expressed at a given threshold[7]. Figure 6.11 allows visualising the distribution of the expression breadth of the mRNAs (Uhlén et al.) and the proteins (Pandey Lab data) across their fifteen common tissues. In the following sections, I may refer to a (TS) gene as a *unique gene* when it is only expressed in a single tissue.

Figure 6.11a shows that the distribution of the protein expression breadth is bimodal. Either due to technical limitations or biological reasons, proteins detected in a sole tissue form the most numerous class and represent 20 % of the overall number. Proteins expressed in all tissues are the second most numerous class (about 16 %); the third largest class (12 %) comprises the proteins expressed in two tissues.

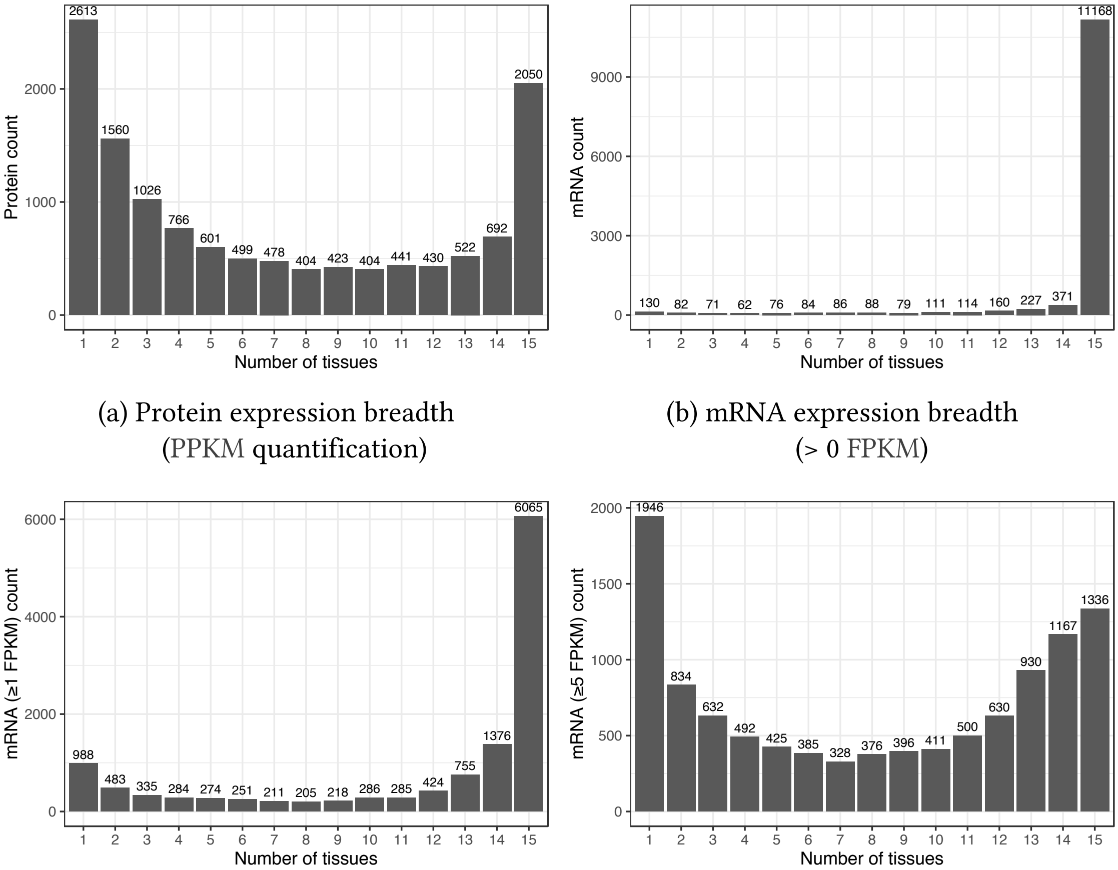
On the other hand, almost all mRNAs are expressed in every tissue (Figure 6.11b). One hypothesis is that mRNAs levels have to exceed a sufficient threshold for their proteins to be detected. Thus, I also studied the effect of two additional minimum expression thresholds for the mRNAs on the expression breadth.

The two new expression breadth profiles are more alike to the proteomic one. As shown in Figure 6.11c, the number of transcripts only found in one tissue increases at the widespread 1 FPKM threshold, which roughly equates to one RNA in the cell [Mortazavi et al., 2008; Hebenstreit et al., 2011].

The expression breadth profile of the mRNAs expressed at or above 5 FPKM present a similar bimodal distribution (Figure 6.11d) to the protein one. While arbitrary, 5 FPKM

[7] If a gene is expressed below the considered threshold in all the tissues, its expression breadth is null.

is a threshold commonly found in the literature [Uhlén, Fagerberg, et al., 2015; Gonzàlez-Porta et al., 2013; J. Chen et al., 2018].



(a) Protein expression breadth (PPKM quantification)

(b) mRNA expression breadth (> 0 FPKM)

(c) mRNA expression breadth (≥1 FPKM)

(d) mRNA expression breadth (≥5 FPKM)

Figure 6.11. **Expression breadth of the proteins and mRNAs.** The expression breadth of the proteins has a bimodal distribution. Many proteins are detected either in a single tissue or in all of them. Almost every mRNA is detected in every tissue. Their breadth becomes bimodal when their expression threshold is increased to 5 FPKM. To ease the general visualisation, I have omitted to plot the mRNAs for which the expression remains below the threshold for all tissues (*i.e.* expression breadth=0 for the considered threshold).

Figure 6.12 displays the fraction of unique genes (*i.e.* only expressed in a single tissue) detected as a protein or an mRNA at a considered threshold for each tissue as the analysis is seeking a possible link between the number of uniquely detected genes and the correlation strength between the proteomic and transcriptomic TREPs. Hence, these fractions are computed by dividing the number of unique genes (proteins or mRNAs) of each tissue by the total amount of uniquely detected genes across all tissues. The tissues are ordered by increasing order of their fraction.

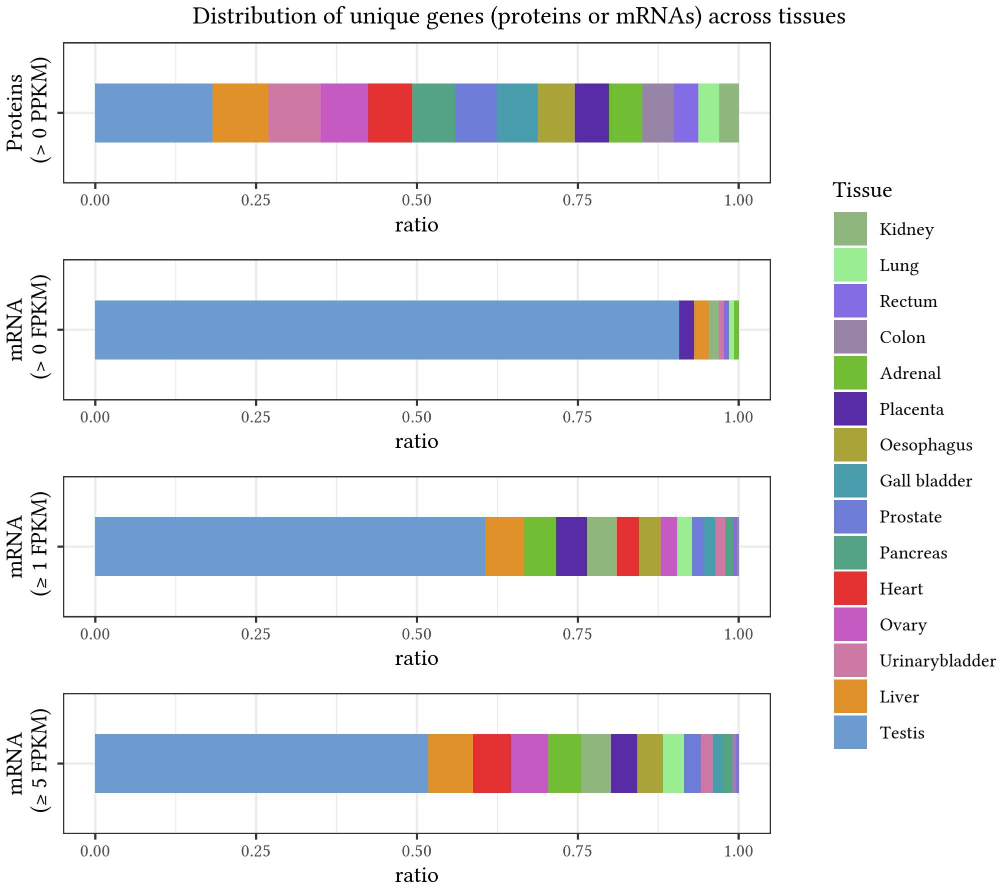


Figure 6.12. **Unique proteins or mRNAs fractions across tissues.**

Although their proportion varies from one tissue to another, all fifteen tissues have proteins that are specifically detected in each tissue solely, as shown in the top plot in Figure 6.12. In contrast, unique mRNAs are detected in a more limited number of tissues (see the three bottom plots of Figure 6.12). Besides, the unique proteins are more evenly distributed between the fifteen tissues than the unique mRNAs.

Except for *Testis* and *Liver*, which are consistently expressing the highest number of uniquely detected genes, the other tissues fail to present any similarity between the available proteomic and transcriptomic data.

Liver is the most correlated tissue (Figure 6.10) and comprises the second-highest number of unique genes. *Testis* is the third-best correlated tissue despite having the highest fractions of unique proteins and mRNAs regardless of any threshold. It may be tempting to hypothesise that the number of unique genes relate to correlation levels, but the other tissues fail to show any relationship between the number of unique mRNAs and proteins they expressed and the strength of the correlations.

Put together, these results suggest that the number of proteins and mRNAs uniquely expressed in these tissues play a minor role at best in the mRNA/protein correlation computed for each tissue. The lack of relation between the proteomic and transcriptomic observations is confirmed by a more refined analysis of the expression breadth.

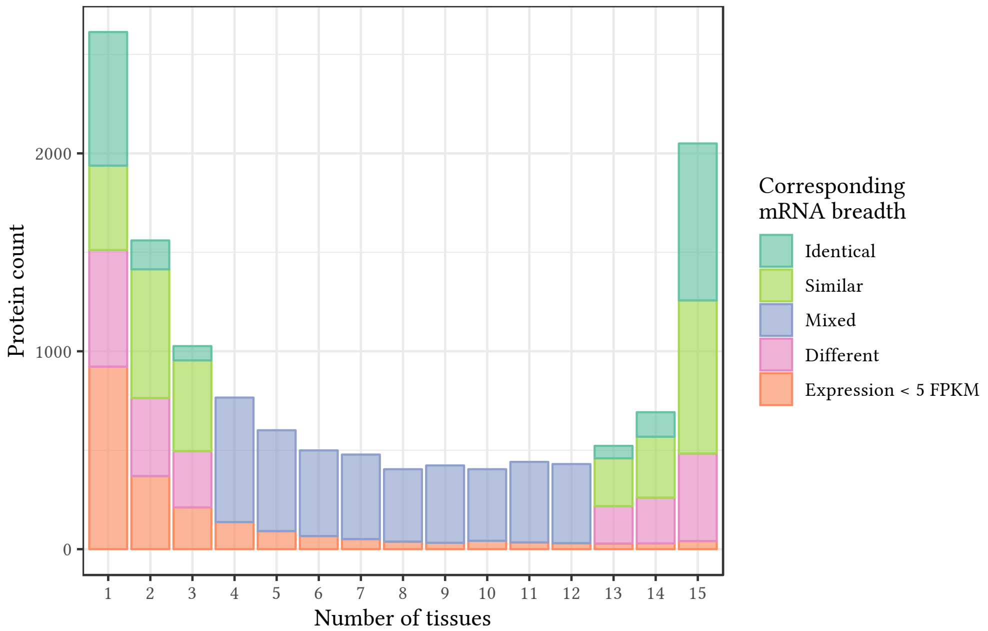


Figure 6.13. **Comparison of proteins expression breadth to their corresponding mRNA.** The proteins' expression breadth (Figure 6.11a) is coloured according to their corresponding mRNA expression breadth at 5 FPKM (Figure 6.11d). About one-fifth of the uniquely detected proteins have their corresponding mRNA identically expressed once at or above 5 FPKM. The number of proteins classified as *Identical* decreases significantly through other breadths until it raises again from thirteen tissues to reach about one-third of the ubiquitous proteins. Proteins and mRNAs with mismatching expression breadth are split into several categories. Proteins and mRNAs that are both detected within four to twelve tissues are described as *Mixed*. If the expression breadths of the remaining pairs are close (± 2), they are identified as *Similar* otherwise as *Different*. Finally, many genes detected at least once as a protein have an mRNA expression that never reaches 5 FPKM (*Expression* < 5 *FPKM*).

Figure 6.13 shows that the expression breadth of mRNAs (expressed ≥ 5 FPKM or even smaller threshold) concurs in very few cases to their corresponding protein breadth. Thus, the mRNA expression breadth is insufficient to predict the expression breadth of the corresponding protein. Even for the two extreme cases where the protein is unique to a tissue or ubiquitous (found in all fifteen tissues), there are differences between the expression breadths of the mRNA and the protein of the same gene.

All the expression breadth analyses of the transcriptome rely on expression levels. However, Chapter 4 underlines that high mRNA expression levels are unrelated to high interstudy correlation of same-tissue pairs while TS mRNAs present a rather strong relation with it. For this reason, the following analysis examines the relationship between TS mRNAs and TS proteins.

6.2.3 *Tissue-specific mRNAs have significant overlap with tissue-specific proteins*

Unlike mRNAs, many proteins are only expressed in one unique tissue. These are the ones I refer to as TS proteins in the remainder of this thesis.

To enable the comparison of these TS proteins with possible transcript partners, I first need to define a set of TS mRNAs. To find the latter, I choose the n mRNAs most specific to a tissue based on the Fold change method (Section 4.3.1.2) where n is the number of TS proteins of that tissue. Then, as detailed in Figure 6.14, I examine for each tissue the overlap between its n TS proteins with its n mRNAs with the highest tissue-specific ranks. Figure 6.15 illustrates the *Heart* example.

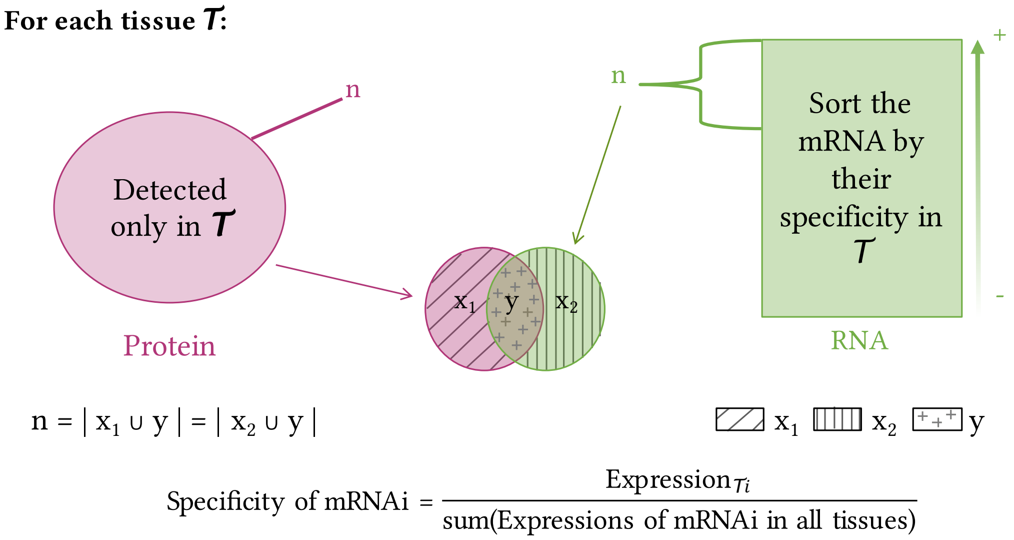

For each tissue $\mathcal{T}$:
n
Detected only in $\mathcal{T}$
Protein
n
Sort the mRNA by their specificity in $\mathcal{T}$
+
-
RNA
x_1
y
x_2
$n = |x_1 \cup y| = |x_2 \cup y|$
▨ x_1 ▥ x_2 ⊞ y
$\text{Specificity of mRNAi} = \frac{\text{Expression}_{\mathcal{T}i}}{\text{sum(Expressions of mRNAi in all tissues)}}$

Figure 6.14. **Overview of the comparison of the TS proteins and TS mRNAs.** TS proteins are the n proteins only expressed in one tissue. Once the mRNAs have been sorted by decreasing order of their relative specificity to a given tissue, the first n mRNAs identities are compared to the ones of the n TS proteins present in the same tissue.

Each tissue has a different number of TS proteins. I thus refine this analysis by computing Jaccard similarity coefficients (or Jaccard indices) [Jaccard, 1901; Lin et al., 2008], see Equation (Jaccard similarity coefficient). The Jaccard indices allow assessing the relationship between TS proteins and mRNAs across all the tissues at the same time and ease the result interpretation in contrast to the raw overlap numbers.

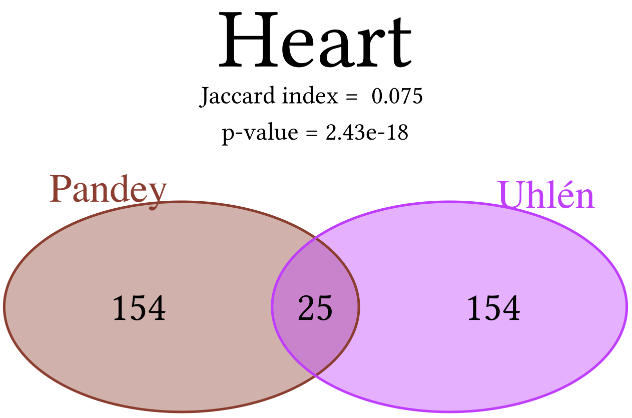


Figure 6.15. **Example of overlap of TS proteins and TS mRNAs.**

The Jaccard index is computed as follow:

$$\begin{aligned} J(x_1, x_2) &= \frac{|x_1 \cap x_2|}{|x_1 \cup x_2|} \\ &= \frac{|x_1 \cap x_2|}{|x_1| + |x_2| - |x_1 \cap x_2|} \end{aligned} \qquad \text{(Jaccard similarity coefficient)}$$

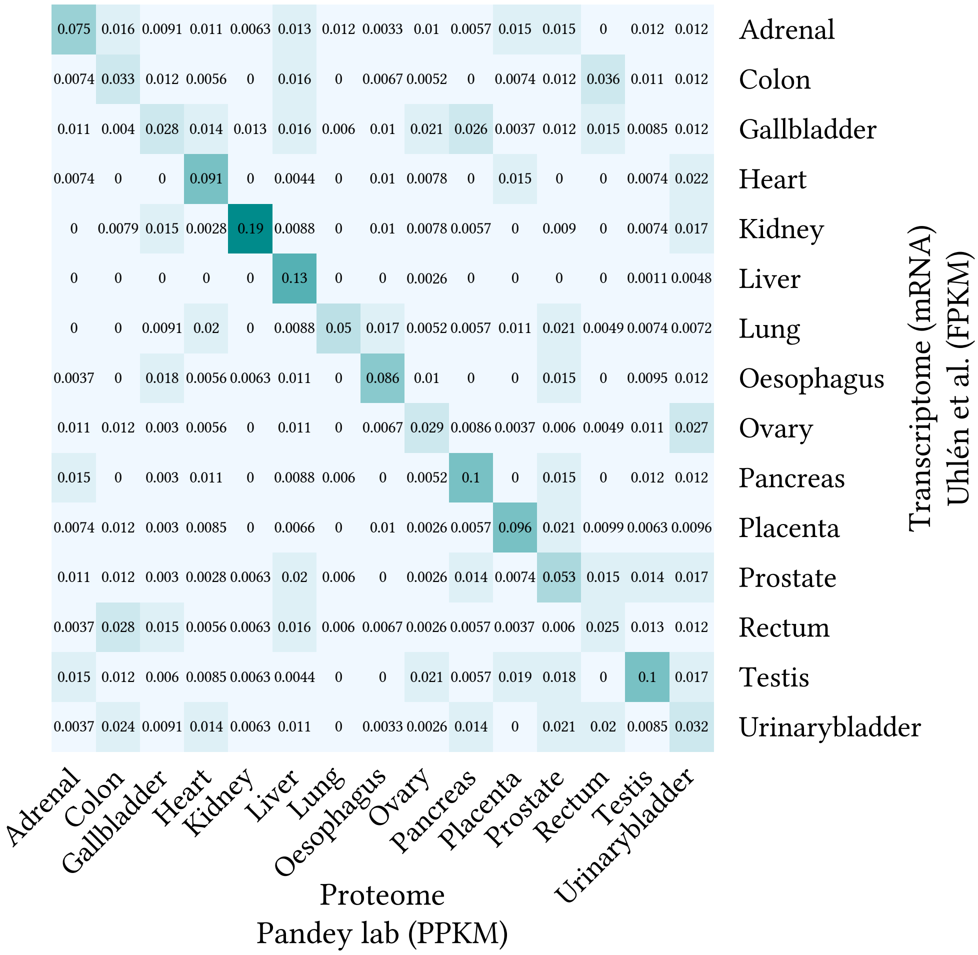
When applied specifically to Figure 6.14, we get: $J(\text{●}, \text{●}) = \frac{k}{2n-k}$, with n the number of proteins (●) that are only found in a given tissue and k is the number of common genes between these n unique proteins and the n most specific mRNAs of the tissue (●).

To measure the Jaccard indices significance, I use the hypergeometric test [Field et al., 2012] (see Appendix E.1). In the current analysis, I consider as 'success' when a TS mRNA is among the n TS proteins and test if the number of observed successes is greater than the expected number for random sampling.

The Jaccard indices for all pairs of the fifteen shared tissues between the Pandey Lab (PPKM quantification) and Uhlén et al. are summarised in Figure 6.16, while Figure 6.17 displays their respective p-values (hypergeometric test).

I have rerun these analyses with different sets of parameters and I have consistently observed statistically significant overlaps except in rare cases, which include the comparison of TS genes for *Urinarybladder* between Pandey Lab (PPKM quantification) and Uhlén et al. (*HTSeq-count*) based on the fifteen-tissue set and where the TS mRNAs are selected with the fold change method. Overall, the Jaccard indices remain within the same ranges for different sets of parameters.

If ranked by their Jaccard indices, several tissues fall within a range similar to their correlation coefficient, while others not — as shown by Figure E.5. For instance, *Liver*,

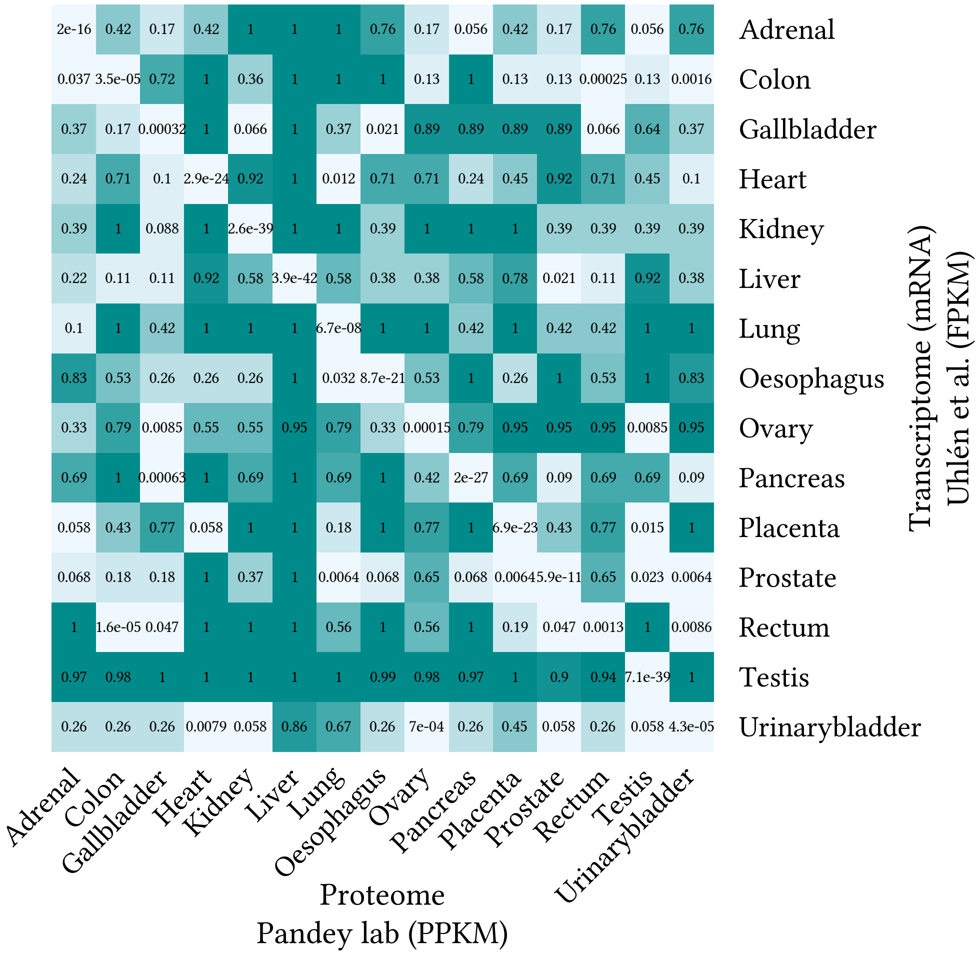


	Adrenal	Colon	Gallbladder	Heart	Kidney	Liver	Lung	Oesophagus	Ovary	Pancreas	Placenta	Prostate	Rectum	Testis	Urinarybladder
Adrenal	0.075	0.016	0.0091	0.011	0.0063	0.013	0.012	0.0033	0.01	0.0057	0.015	0.015	0	0.012	0.012
Colon	0.0074	0.033	0.012	0.0056	0	0.016	0	0.0067	0.0052	0	0.0074	0.012	0.036	0.011	0.012
Gallbladder	0.011	0.004	0.028	0.014	0.013	0.016	0.006	0.01	0.021	0.026	0.0037	0.012	0.015	0.0085	0.012
Heart	0.0074	0	0	0.091	0	0.0044	0	0.01	0.0078	0	0.015	0	0	0.0074	0.022
Kidney	0	0.0079	0.015	0.0028	0.19	0.0088	0	0.01	0.0078	0.0057	0	0.009	0	0.0074	0.017
Liver	0	0	0	0	0	0.13	0	0	0.0026	0	0	0	0	0.0011	0.0048
Lung	0	0	0.0091	0.02	0	0.0088	0.05	0.017	0.0052	0.0057	0.011	0.021	0.0049	0.0074	0.0072
Oesophagus	0.0037	0	0.018	0.0056	0.0063	0.011	0	0.086	0.01	0	0	0.015	0	0.0095	0.012
Ovary	0.011	0.012	0.003	0.0056	0	0.011	0	0.0067	0.029	0.0086	0.0037	0.006	0.0049	0.011	0.027
Pancreas	0.015	0	0.003	0.011	0	0.0088	0.006	0	0.0052	0.1	0	0.015	0	0.012	0.012
Placenta	0.0074	0.012	0.003	0.0085	0	0.0066	0	0.01	0.0026	0.0057	0.096	0.021	0.0099	0.0063	0.0096
Prostate	0.011	0.012	0.003	0.0028	0.0063	0.02	0.006	0	0.0026	0.014	0.0074	0.053	0.015	0.014	0.017
Rectum	0.0037	0.028	0.015	0.0056	0.0063	0.016	0.006	0.0067	0.0026	0.0057	0.0037	0.006	0.025	0.013	0.012
Testis	0.015	0.012	0.006	0.0085	0.0063	0.0044	0	0	0.021	0.0057	0.019	0.018	0	0.1	0.017
Urinarybladder	0.0037	0.024	0.0091	0.014	0.0063	0.011	0	0.0033	0.0026	0.014	0	0.021	0.02	0.0085	0.032

Figure 6.16. **Heatmap of Jaccard indices across the common fifteen tissues between Uhlén *et al.* and Pandey Lab data.** For each tissue, the TS proteins are the proteins (quantified with PPKM method) that are expressed only in that tissue. The TS mRNAs are the mRNAs with the highest specific coefficients in that tissue.

Testis and *Pancreas* have high ranks and *Urinarybladder* and *Gallbladder* low ones for both their correlation and their Jaccard indices. On the other hand, *Kidney* ranks first for the Jaccard index, but only reaches the seventh rank for Pearson correlation. While the *Rectum* has the smallest Jaccard index and thus ranks last (*i.e.* fifteenth), it gets ranking number nine for Pearson correlation. These results suggest that TS proteins and TS mRNAs are unrelated to the tissue correlation levels.

Overall, the above direct approaches (based on the gene expression breadth across tissues and their tissue-specificity) fail to show a prominent (if any) contribution or association to the correlation levels between the proteome and transcriptome. These are most likely resulting from multiple subtle similarities based on identical differential expression within



	Adrenal	Colon	Gallbladder	Heart	Kidney	Liver	Lung	Oesophagus	Ovary	Pancreas	Placenta	Prostate	Rectum	Testis	Urinarybladder
Adrenal	2e-16	0.42	0.17	0.42	1	1	1	0.76	0.17	0.056	0.42	0.17	0.76	0.056	0.76
Colon	0.037	3.5e-05	0.72	1	0.36	1	1	1	0.13	1	0.13	0.13	0.00025	0.13	0.0016
Gallbladder	0.37	0.17	0.00032	1	0.066	1	0.37	0.021	0.89	0.89	0.89	0.89	0.066	0.64	0.37
Heart	0.24	0.71	0.1	2.9e-24	0.92	1	0.012	0.71	0.71	0.24	0.45	0.92	0.71	0.45	0.1
Kidney	0.39	1	0.088	1	2.6e-39	1	1	0.39	1	1	1	0.39	0.39	0.39	0.39
Liver	0.22	0.11	0.11	0.92	0.58	3.9e-42	0.58	0.38	0.38	0.58	0.78	0.021	0.11	0.92	0.38
Lung	0.1	1	0.42	1	1	1	6.7e-08	1	1	0.42	1	0.42	0.42	1	1
Oesophagus	0.83	0.53	0.26	0.26	0.26	1	0.032	8.7e-21	0.53	1	0.26	1	0.53	1	0.83
Ovary	0.33	0.79	0.0085	0.55	0.55	0.95	0.79	0.33	0.00015	0.79	0.95	0.95	0.95	0.0085	0.95
Pancreas	0.69	1	0.00063	1	0.69	1	0.69	1	0.42	2e-27	0.69	0.09	0.69	0.69	0.09
Placenta	0.058	0.43	0.77	0.058	1	1	0.18	1	0.77	1	6.9e-23	0.43	0.77	0.015	1
Prostate	0.068	0.18	0.18	1	0.37	1	0.0064	0.068	0.65	0.068	0.0064	5.9e-11	0.65	0.023	0.0064
Rectum	1	1.6e-05	0.047	1	1	1	0.56	1	0.56	1	0.19	0.047	0.0013	1	0.0086
Testis	0.97	0.98	1	1	1	1	1	0.99	0.98	0.97	1	0.9	0.94	7.1e-39	1
Urinarybladder	0.26	0.26	0.26	0.0079	0.058	0.86	0.67	0.26	7e-04	0.26	0.45	0.058	0.26	0.058	4.3e-05

Figure 6.17. **p-values associated with the Jaccard indices** of Figure 6.16. These p-values have been computed with the hypergeometric test.

many clusters of the proteins and mRNAs.

Given the mixed results of the direct approaches built on uniquely detected genes, I next examine whether or not an indirect method may be more appropriate. For this new analysis, I build hierarchical cluster trees that try to translate the tissues' expression 'closeness' or differentiation distances. Then, I compare the proteins' and transcripts' trees.

6.2.4 *Proteins and mRNAs tissue trees present partial concordant results*

As presented in Section 3.2.2, a hierarchical clustering analysis requires a linkage method and a distance between each element that is included in the analysis.

I have built the tree of each dataset by linking the tissues with Ward's method [Ward, 1963] like in the previous analyses.

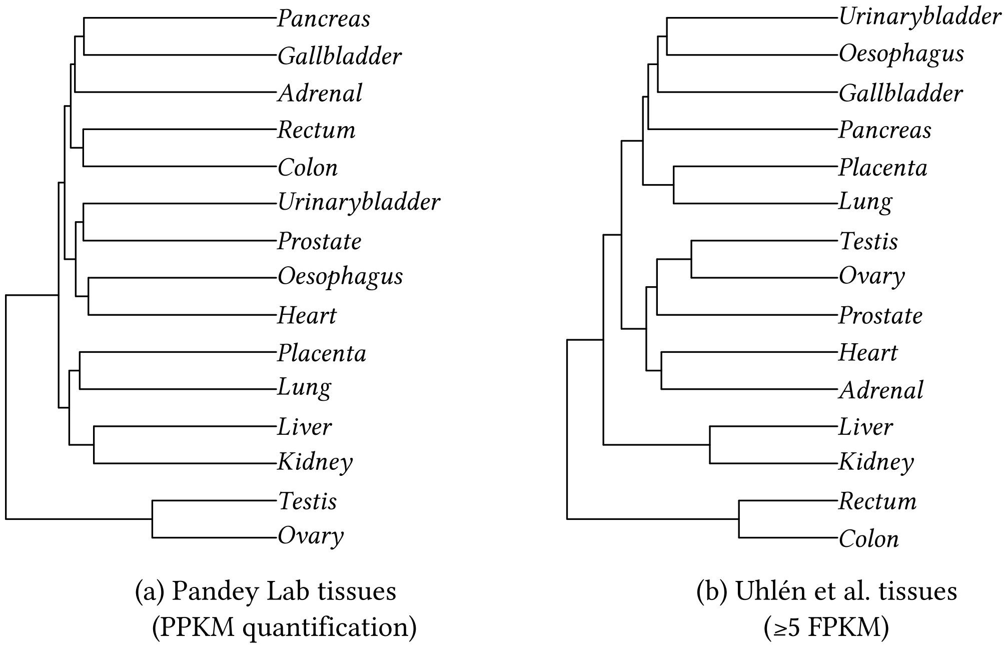
I have performed an initial analysis that used the tissue correlations for the distance, but it did not highlight any similarity between the proteomics and transcriptomics trees of hierarchical clusters.

The distance used in the following analysis reflects the difference in composition of gene populations expressed by the tissues. It is based on the Jaccard index (see Equation (Jaccard similarity coefficient)) of the tissues where I only consider genes (proteins or mRNAs) that are detected in two tissues strictly.

Making the hypothesis that the closeness of two tissues increases with the number of genes they share, I compute the distance between two tissues, t_1 and t_2 using the following formula: distance$(t_1, t_2) = 1 - J(t_1, t_2)$.

Note that for this analysis, I present the results for Pandey Lab (PPKM quantification) and Uhlén et al. for their fifteen shared tissues as for the other analyses, but I also present and compare with the GTEx data and their twelve shared tissues.

Figure 6.18 shows the hierarchical clustering of the fifteen tissues for Pandey Lab (PPKM quantification) and Uhlén et al. (≥ 5 FPKM) studies.



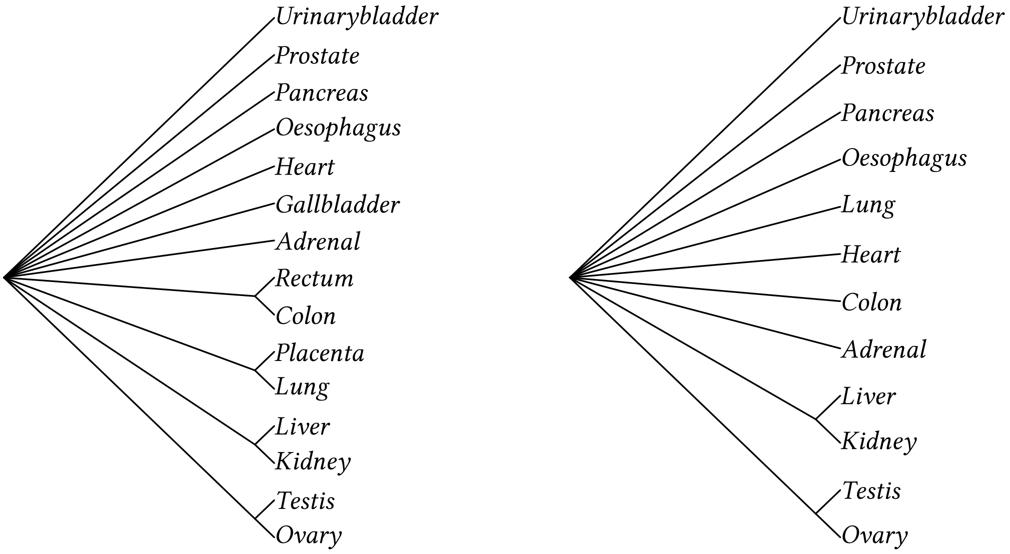
(a) Pandey Lab tissues (PPKM quantification)

(b) Uhlén et al. tissues (≥5 FPKM)

Figure 6.18. **Hierarchical clustering for the fifteen tissues of Pandey Lab and Uhlén *et al.* studies.**

Both Pandey Lab and Uhlén et al. trees, respectively Figures 6.18a and 6.18b display the same four pairs of tissues the most closely related: *Rectum* and *Colon*, *Placenta* and *Lung*, *Liver* and *Kidney*, and *Testis* and *Ovary*.

Comparing more than two hierarchical trees for possible finding congruence is cumbersome manually; thus, methods exist to create consensus trees [Felsenstein et al., 2004]. The methods can be strict or create a consensus based on the majority. Since there is a maximum of three trees (one for each dataset) to compare at a time, all the consensus trees within this thesis are strict, *i.e.* all the trees must be in agreement on a hierarchical organisation for the consensus tree to include it. I use one of the possible implementations of these methods from the R package *ape* (v5.3) [Paradis et al., 2019].



(a) Consensus tree of the fifteen shared tissues between Pandey Lab and Uhlén et al. (≥ 5 FPKM) data

(b) Consensus tree of the twelve shared tissues between Pandey Lab data and Uhlén et al. and GTEx data (≥ 5 FPKM)

Figure 6.19. **Consensus of the hierarchical clustering of the tissues across the different studies.**

Figure 6.19 shows two consensus trees. Figure 6.19a is the consensus tree built on the previous trees of Pandey Lab and Uhlén et al. fifteen shared tissues. The tree groups are clearly featured. To assay if these groups may still be found beyond these two datasets, I extend the analysis with the GTEx data.

Figure 6.19b relies on the set of twelve shared tissues between Pandey Lab data (quantified by the PPKM method) and the two transcriptomic datasets (≥ 5 FPKM): Uhlén et al. and GTEx data. Compared to Figure 6.19a, only two tissue-sets are consistently observed as most closely related: *Liver* and *Kidney*, and *Testis* and *Ovary*.

Note that the results seem unaffected by the protein quantification method as PPKM and Top3 methods give identical results. However, the threshold, above which the mRNAs are considered, influences the analysis' outcomes. As very few mRNAs are found uniquely in two tissues at 0 FPKM, this threshold is insufficient to identify any hierarchical organisation in the transcriptomic data. Increasing the threshold, for instance to 1 FPKM, allows exposing clusters of tissues, *e.g. Liver* and *Kidney*.

However, different thresholds can also highlight different tissue clusters, including *Rectum* and *Colon* at 5 FPKM or *Pancreas* and *Gallbladder* at 1 FPKM. The influence of the thresholds on the results suggests that genes have their expression levels varying depending on their tissue context.

In summary, even indirect analyses based on the proteins and mRNAs breadth expression have some degree of similarity in their results and seem to partially capture a biological signal.

Further more in-depth (direct or indirect) analyses may clarify the relationship between the proteome and transcriptome. For example, equivalent correlation levels between specific gene groups (gene co-expression correlation) may be highlighted for each tissue at both biological layers. However, this kind of analysis requiring a case-by-case approach will greatly benefit from well-established and proven mRNA and protein expression baselines for each tissue.

The following section Section 6.3 analyses precisely whether genes have comparable expression profiles as a protein than as an mRNA.

6.3 WIDE CORRELATION RANGE FOR PROTEIN/MRNA PAIRS

As previously reported, the expression levels of mRNA/protein pairs can present a tight relationship in one tissue while being seemingly unrelated in another. The first gene-centric analysis explores for each gene the relationship between the expression levels of its mRNAs and proteins across all available tissues. This analysis helps one to determine whether any intrinsic trend structures the gene expression or if it is only subject to the environment.

Figure 6.20 displays the Pearson correlation between the matching pairs of mRNAs from Uhlén et al. and the proteins from Pandey Lab (quantified with the PPKM method) across their fifteen common tissues. The observed levels of correlation (in pink) are higher than the expected levels computed by random permutation (the grey line is showing the average of 10,000 permutations).

I also compare the Pearson correlation of the matching mRNAs/protein pairs with the ones (in green) of the mRNAs/mRNAs pairs from Uhlén et al. and GTEx data to provide more context. About two-thirds of the genes (8,550) present a strong correlation ($r > 0.8$) between the mRNA expression levels of Uhlén et al. and GTEx data, but only about 6% (775) of the mRNA/proteins pairs are above this limit (in dark blue).

The Pearson correlation between the mRNA/protein pairs ranges rather widely from 1 (for 32 pairs, which are detected as mRNA and protein in the same single tissue only) to below

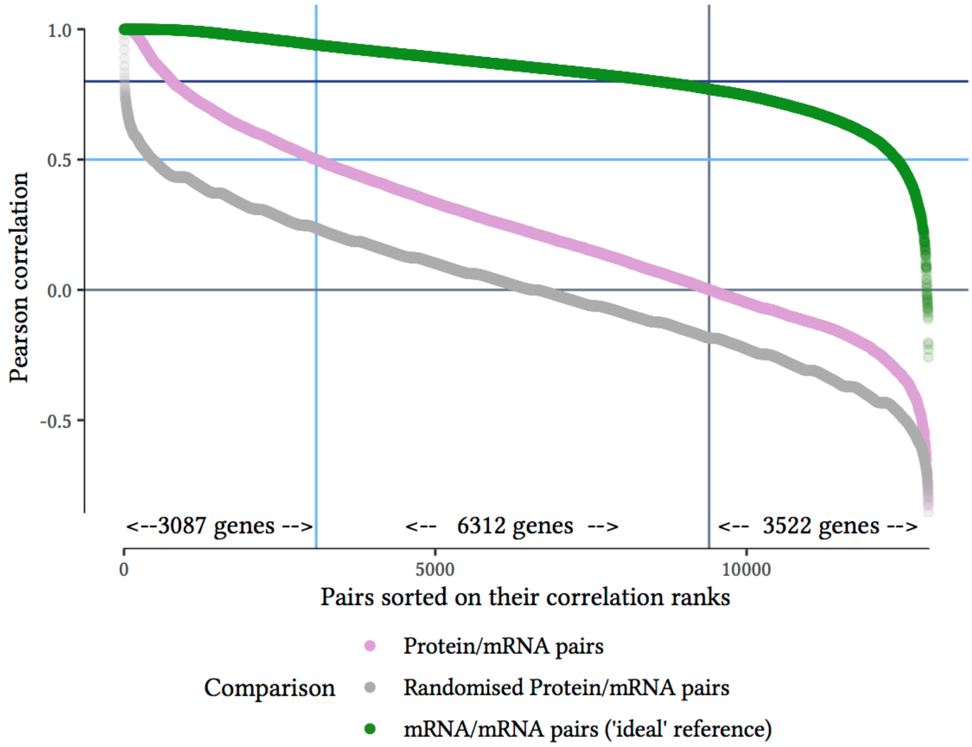
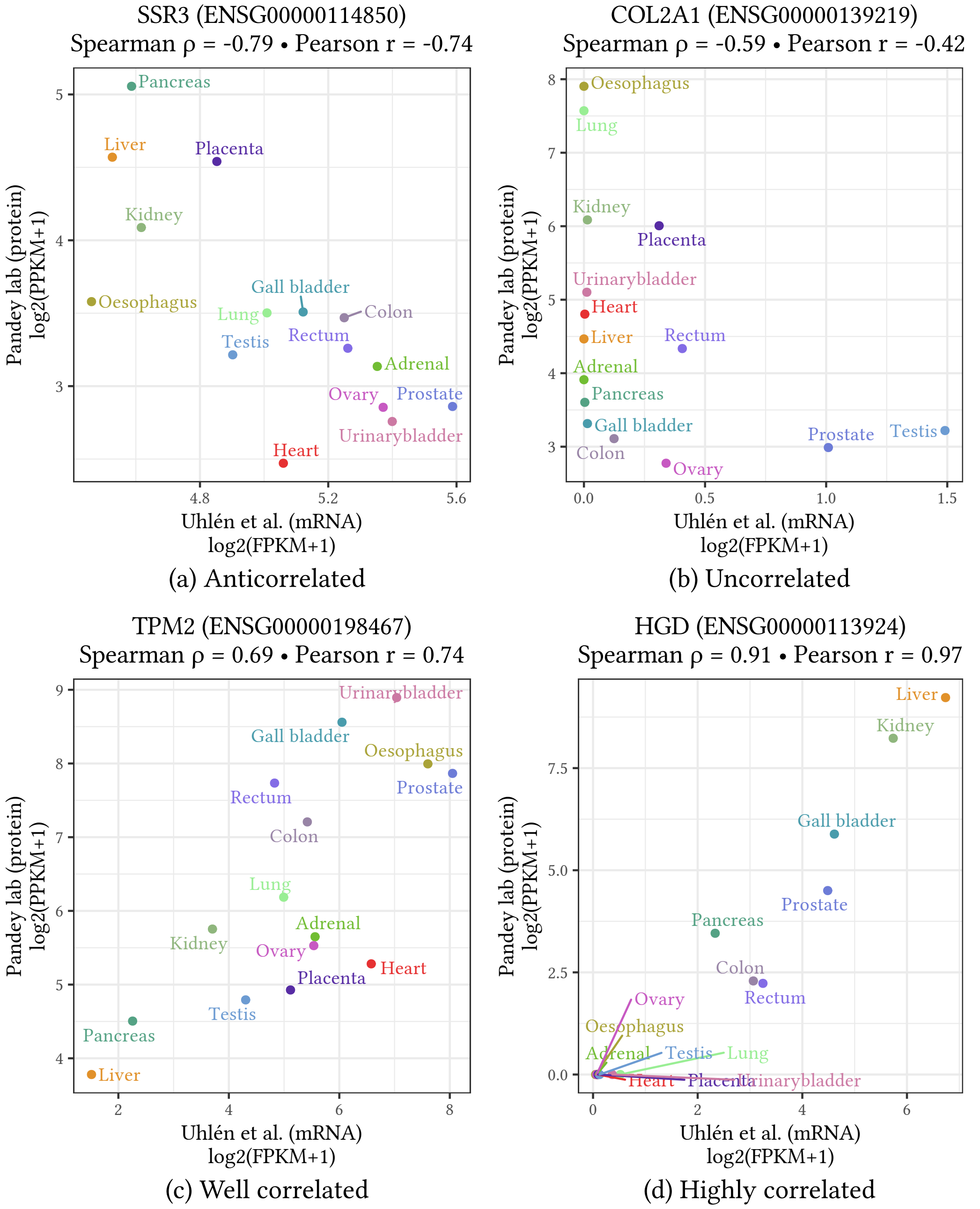


Figure 6.20. **Pearson correlation of gene expression levels between studies across the shared tissues in descending order.** For each mRNA and its corresponding protein, I have computed their Pearson correlation (in pink) across the fifteen common tissues between Pandey Lab (PPKM) and Uhlén et al. data and then ordered them in decreasing order. The x-axis shows the rank of each pair and the y-axis its correlation coefficient (computed with $\log_2(\text{level} + 1)$). The grey line represents the mean of the 10,000 randomisations (by pair composition permutation). The permutation confirms that the observed correlation coefficients are significantly higher than expected by chance. The green line serves as the most *ideal* comparison case: it represents the Pearson correlation of mRNAs pairs between Uhlén et al. and GTEx data.

-0.5 (for 105 pairs, with $r = -0.83$ for the most anticorrelated one).

A closer look at both extremes (Figure 6.21) reveals several possible relationship profiles between the expression of the mRNAs and proteins. The negatively correlated genes can present overall anticorrelated (Figure 6.21a) or rather unrelated (Figure 6.21b) expression levels of mRNAs and proteins. On the other end, the highest correlated genes present genes with a tissue-specific (TS) protein or whose mRNA and protein expression levels are tightly related (Figures 6.21c and 6.21d).

I only present herein (and in the following sections) the set of results based on the Pearson correlation of the gene expression levels between the Pandey Lab data quantified with the PPKM method and Uhlén et al. data across their fifteen shared tissues since the different sets share similar results trends. Furthermore, this data combination provides the highest



SSR3 (ENSG00000114850)
Spearman ρ = -0.79 • Pearson r = -0.74

(a) Anticorrelated

COL2A1 (ENSG00000139219)
Spearman ρ = -0.59 • Pearson r = -0.42

(b) Uncorrelated

TPM2 (ENSG00000198467)
Spearman ρ = 0.69 • Pearson r = 0.74

(c) Well correlated

HGD (ENSG00000113924)
Spearman ρ = 0.91 • Pearson r = 0.97

(d) Highly correlated

Figure 6.21. **Different cases of correlation for protein/mRNA pairs.** The scatterplots show the expression of the genes as mRNA in Uhlén et al. data on the x-axis and as a protein in Pandey Lab data on the y-axis across their common fifteen tissues. Figure 6.21a shows that *SSR3* apparently has a protein expression anticorrelated to its mRNA's one: when the mRNA expression is low (*e.g. Pancreas*, *Liver*), the protein expression is high and when the mRNA expression is high (*e.g. Adrenal gland*, *Prostate*), the protein expression is low. Figure 6.21b features *COL2A1* and for which protein expression is observed for many tissues (*e.g. Oesophagus*, *Kidney*, *Pancreas*) while no or very low mRNA expression has been captured. Figure 6.21c shows that *TPM2* is expressed in all tissues and the expression of the protein is well correlated to the mRNA's one. Figure 6.21d shows that *HGD* has highly correlated protein and mRNA expression when found in a tissue. Note that true perfect correlation can be observed for pairs where both mRNA and protein are tissue-specific (TS).

number of pairs to be studied across the highest number of tissues. It also allows continuity with the above tissue-centric analyses. Complementary results for Spearman correlation and sets that include GTEx data can be found in digital format at http://www.barzine.net/~mitra/thesis.

6.3.1 *TS protein enrichment for the most correlated pairs*

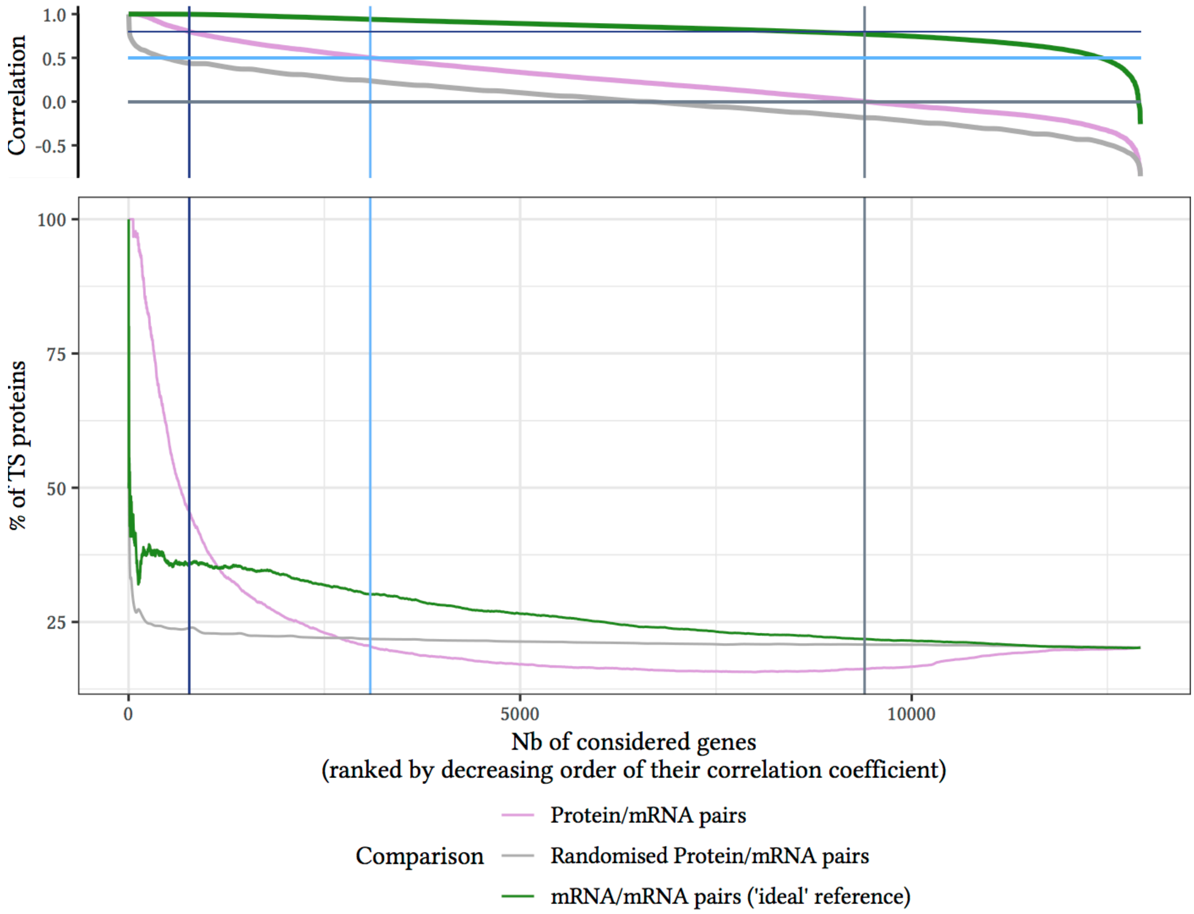


Figure 6.22. **TS proteins percent as a function of the considered number of genes (ranked by Pearson correlation).** Before being plotted, the genes are first ranked in decreasing order of Pearson correlation between the expression levels of the two considered datasets across their shared tissues. The top plot, which is a reproduction of Figure 6.20, is for interpretive convenience only. The two parts of the figure have corresponding x-axes. The x-axis of the top plot presents the rank associated with the Pearson correlation coefficients of the genes on the y-axis. The x-axis of the bottom plot represents the upper limit rank up to which are considered the genes (for the calculus). The lower limit rank is 1. The y-axis of the bottom plot displays the percentage TS protein for a set of (ordered) genes.

Here, I investigate the incidence of the TS proteins on the level of correlation between mRNAs and their proteins.

I first compute for each gene the correlation between the expression levels of Uhlén et al. data and Pandey Lab data. Then, I organise the genes in a sequence[8] by ranking them by decreasing order of correlation. Thus, the first most correlated gene's rank is 1. Finally, for each rank k, I calculate the number of genes with TS proteins among the first k ranks before converting it into percentage.

Equation (TS protein percentage) on page 256 presents the formula with which I compute the percentage of TS proteins. Figure 6.22 illustrates the percentage of TS proteins for a given number of considered genes, which have been ranked in decreasing order of Pearson correlation coefficients. Similarly to Figure 6.20, randomised protein/mRNA pairs (average for 10,000 permutations) in grey and mRNA/mRNA pairs (Uhlén et al. data and GTEx data) in green are providing some context.

The TS protein percentage in pink is extremely high for the highest range of observed correlation between the expression of the Pandey Lab proteins (quantified with PPKM) and Uhlén et al. mRNAs pairs across their shared fifteen tissues. The TS protein percentage then decreases quickly before finally increasing slowly for the lower range of correlation. Thus, the genes identified as TS proteins in Pandey Lab data enrich the most correlated and most anticorrelated mRNA/protein pairs clearly. Most of the pairs have a Pearson correlation above 0.5, although they show a wide range of Pearson correlation, from $r = -0.77$ (for *ZNF770*) to $r = 1$ (for several genes).

However, the corresponding mRNA/mRNA pairs between Uhlén et al. and GTEx data show a more evenly distribution of these genes through the 10,000 highest gene correlations after an initial peak. The average of the 10,000 permutations (in grey) of the Pandey Lab with the Uhlén et al. data also shows an initial high TS protein percentage that drops to the global amount of TS proteins among the complete set of shared genes. Overall, TS proteins represent about one-fifth of all the genes.

6.3.2 *Gene expression profiles clue about biological and technical differences*

As mentioned previously, either artefacts or biology may explain the observed low correlations. It is rather difficult to identify which artefacts are specifically impacting each protein/mRNA pair as there are many and can occur in combination. However, two major technical sources of artefacts, which have been reported, are dispersion for the lowly expressed mRNAs and saturation for the highly expressed proteins. See Sections 1.2 and 1.3 (from p. 7) for more details.

Figure 6.23 shows possible profiles of relationships between the protein and mRNA pairs. Well-correlated pairs are in the grey area delimited by the green line. For these genes,

[8] A sequence is an ordered set.

the expression levels of the protein observed in a tissue is tightly related to the levels of the mRNA. Although the data have been sampled from different sources, the stronger associations suggest a similar translation process across the tissues and may also imply less post-translational modifications that can hinder protein identification and quantification than for other genes.

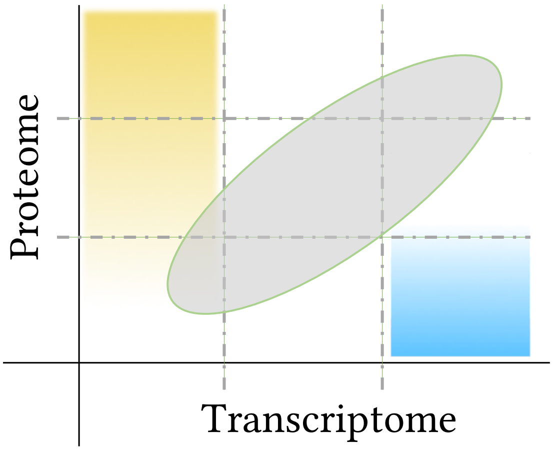


Figure 6.23. **Possible mRNA/protein expression profiles due to biological reasons.** Genes with well-correlated transcriptomic and proteomic expression are found in the grey area delimited by the green line. The genes in the yellow area, *i.e.* which present a high protein concentration and a low mRNA concentration, may have stable proteins and mRNAs with short half-lives. The genes in the blue area, which present a low protein concentration and high mRNA concentration, may have a highly regulated translation, a protein challenging to capture with MS or a misfit between the annotation definition of their mRNA and protein.

Genes in the yellow area present a high protein concentration and low mRNA concentration. One possible cause may be that these genes have stable proteins and mRNAs with short half-lives. On the other hand, genes in the blue area present a low protein concentration high mRNA concentration. These latter genes may have a protein that is more challenging to capture (either because of unsuited protocols as described in Section 1.3 or annotation misfits, *e.g. STAU2* as shown in Figure E.3), may forego through higher regulation through their translation or may be actively exported outside of the tissue that synthesises them. Anticorrelated genes are another category that will likely require further analysis for a better overall understanding. The observed anticorrelation between the proteins and mRNAs expression can be caused by various elements, which may include tissue-dependent isomers (either for the mRNA or protein) expression, self-regulation or a variable secretion rate of the protein.

At the time of writing, the most accurate way to classify the protein/mRNA pairs remains empirical, *i.e.* human interpretation based on the joint visualisation of protein and mRNA expression across the different tissues (and dataset combination).

As empirical approaches are better designed to examine a few genes of interest per study than to give a broader view of the expression landscape, in the next section I favour a more general strategy instead. I study the three gene groups of interest highlighted above with a GO enrichment analysis (see Section 1.4) to find possible biological factors that may differentiate them.

As the pairs with a TS protein enrich both the most correlated and the most anticorrelated genes, I also choose to study them but separately, and thus, I remove them from the most correlated and anticorrelated gene lists.

6.3.3 *Distinct functional enrichment profiles for pairs with a TS protein, and for the best correlated and most anticorrelated ones*

A GO enrichment analysis (GOA — see Section 1.4.1) uses gene ontologies that provide defined terms covering the gene product (*i.e.* mRNA and protein) properties. These terms are structured into categories, which helps to assess whether a gene set is associated with a biological process (BP), a molecular function (MF) or a cellular component (CC). The three ontologies are included in the Bioconductor package *org.Hs.eg.db* [Carlson, 2019] for analysis in *R*[9] [R Core Team, 2019].

The enrichment computation requires the comparison between the GO terms set associated with each studied list of genes to a reference. I consider three gene lists: the three hundred best correlated and the three hundred most anticorrelated protein/mRNA pairs and all the pairs (2,613) with a TS protein. As a reference, I use the 12,921 matching protein/mRNA pairs between Pandey Lab (PPKM quantification) and Uhlén et al. data.

The Bioconductor package *clusterProfiler* (v3.12) [G. Yu et al., 2012] provides a function enrichGO, which implements the GOA as the over-representation test described by Boyle et al. (2004) and handles all the required statistical testing and (Benjamini and Hochberg [Benjamini et al., 1995]) correction.

The lists of the best correlated pairs and the ones with a TS protein present distinctive enrichment profiles through the three ontologies. However, the most anticorrelated pairs list presents an enrichment only for biological process (BP) terms. All the individual enrichment GO analyses along with the comparison of the enrichment for the CC and MF ontologies are provided as digital supplementary material.

Figure 6.24 presents the results for the comparison of the enrichment of the three considered gene lists for the BP ontology.

The left side of the figure is a heatmap that marks the pairs' associations with the GO categories on the y-axis. It includes all protein/mRNA pairs of the three studied gene lists. They are sorted on the x-axis in decreasing order of their Pearson correlation.

[9] *R* — https://cran.r-project.org/

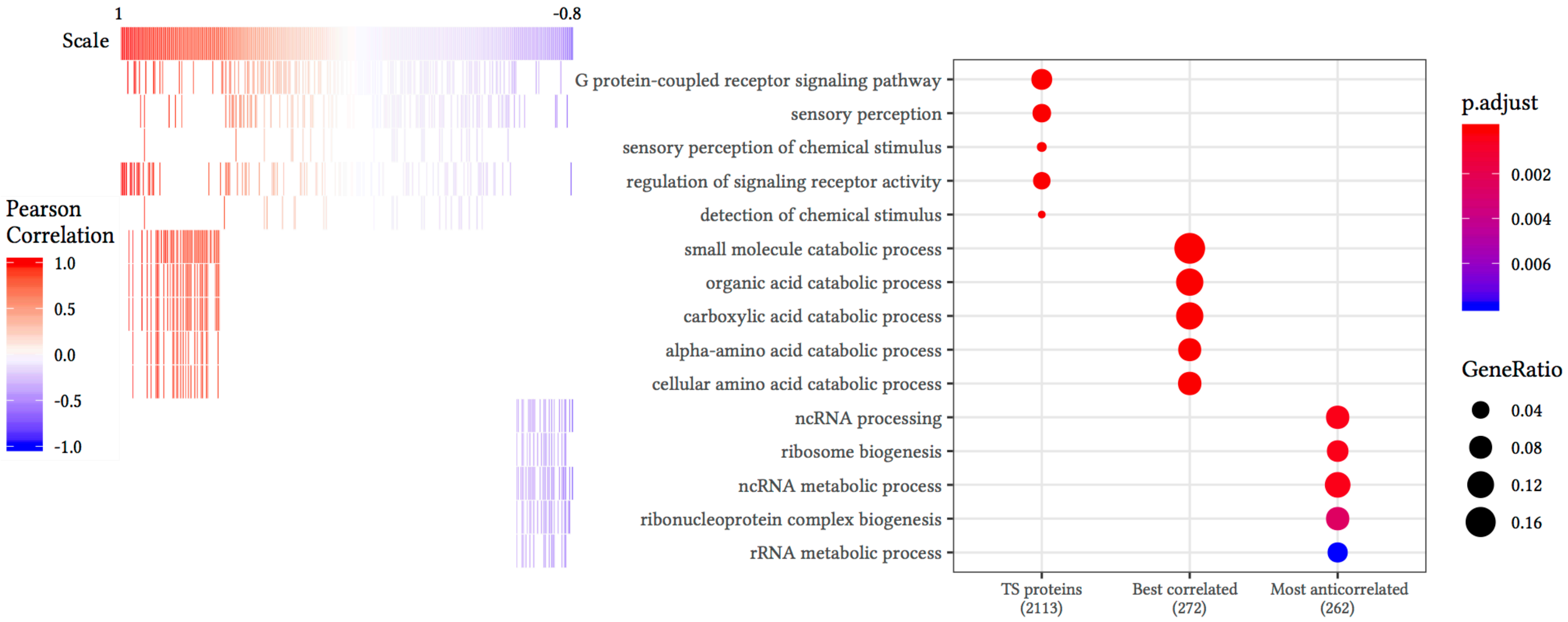


Figure 6.24. **Enriched GO categories for the genes with a TS protein, the three hundred with the highest correlations and the three hundred with the highest anticorrelations.** The shared y-axis of the two parts includes the enriched GO categories (for any of the three groups). The left part of the figure shows a heatmap where all the included protein/mRNA pairs (*i.e.* 3,213) are sorted by their Pearson correlation on the x-axis and that each association of a pair with a GO category is marked. The right part shows the results of the BP GOA analysis with *clusterProfiler* (reference: the complete set of 12,921 genes); the three groups are showed on the x-axis with their number of genes annotated in the considered ontology. For each dot, the size represents the ratio of pairs within each group contributing to each category enrichment, and the colour indicates their significance.

These GO categories, shared between both sides of the figure, combine the five most enriched categories for each of the three lists (TS proteins, best correlated and most anticorrelated pairs). The categories enrichment is provided by another *clusterProfiler* function, compareCluster (which internally invokes enrichGO), which has also produced the right side plot. No cross-enrichment of GO category exists between the three groups (ensured by the 'includeAll=TRUE' option).

Notably, the GO categories associated with each gene list create coherent groups of similar biological processes. The genes with a TS protein are enriched in terms for specific signalling, either for its detection ('*detection of chemical stimulus*', '*sensory perception of chemical stimulus*', '*sensory perception*'), as a response to a signal ('*G protein-coupled receptor signalling pathway*') or as a regulation ('*regulation of signalling receptor activity*').

Concurrently, genes with the best correlated mRNA and protein pairs of expression are associated with catabolic processes[10], and genes with the most anticorrelated pairs are related to ribosomes and ncRNAs regulation, thus by extension to the translation and its regulation.

The GO terms enrichment analysis with the CC ontology shows that the best correlated genes are the most enriched for the following categories: the '*postsynaptic membrane*', '*apical plasma membrane*', '*apical part of cell*', '*cluster of actin-based cell projections*', '*brush border*' and the '*cornfield envelope*'.

On the other hand, the pairs presenting a TS protein show a slight enrichment for '*ion channel complex*', '*transmembrane transporter complex*', '*transporter complex*', and '*cation channel complex*'. Whereas the categories associated with the TS proteins are more ubiquitous and can concern every cell type, the enriched categories for the best correlated genes are referring more specifically to subsets of cells. The localisation of the best correlated pairs probably suffers from their overall ubiquity. Thus, the results for the best correlated genes are probably an artefact of annotation even though they comparatively rely on more genes.

The enrichment analysis with the MF ontology for the best correlated pairs points to different activities: oxidoreductase, cofactor and transmembrane transporter or signalling activities ('*anion transmembrane transporter activity*', '*oxidoreductase activity, acting on CH-OH group of donors*', '*oxidoreductase activity, acting on the CH-OH group of donors NAD or NADP as acceptor*', '*cofactor binding*', and '*coenzyme binding*').

The pairs with a TS protein are also associated transporter and signalling activities (with the following five categories: '*transmembrane signalling receptor activity*', '*signalling receptor activity*', '*molecular transducer activity*', '*channel activity*', and '*passive transmembrane transporter activity*').

[10] Catabolic processes are an energy release source and depend on molecules requiring to be break down [Alberts et al., 2002].

Put together, these results suggest that when there is a high correlation or anticorrelation between an mRNA and its protein, biological processes play a more likely role than possible technical confounding.

The anticorrelated pairs fail to present any enrichment for a specific cell compartment or a molecular function. Thus, it implies that regardless of their localisation within the cell or their chemical properties (which relate to their molecular function), the bottom-up MS studies manage to capture most of the proteins with variable effectiveness.

Nevertheless, bottom-up MS studies favour some proteins, and missing proteins are a primary source of ambiguities [Poverennaya et al., 2017]. Hence, comparing the relative expression levels of proteins within a tissue may lead to misinterpretations. On the other hand, while it requires caution, the relative expression levels of each protein across tissues ought to provide biological insights.

Finally, running all the gene-centric analyses presented above with the other possible combinations of parameters mentioned for the tissue-centric analyses gives similar results.

6.4 DISCUSSION

In this chapter, I describe the integration and comparison of independent large-scale proteomic and transcriptomic expression datasets of undiseased human tissues. After assessing the range of correlation between the two biological layers, I have tried to identify possible factors that may influence the association between the expression of the mRNAs and proteins. I have employed both tissue- and gene-centric approaches.

Building on insights gained in previous chapters (particularly Chapters 4 and 5), I have restricted the integration study to the three following independent sources: Uhlén et al. [Uhlén, Fagerberg, et al., 2015] and GTEx [Melé et al., 2015] data for the transcriptomics and Pandey Lab data [M.-S. Kim et al., 2014] for the proteomics. The three datasets share twelve tissues, while the combined datasets based only on Uhlén et al. and Pandey Lab data present three additional tissues.

The above analyses provide the comparison of mRNA and protein expression across fifteen tissues for 12, 921 pairs as they include Uhlén et al. data and Pandey Lab data quantified with our PPKM method (see Chapter 5). This new quantification allows encompassing about twice as many proteins than with the standard state-of-art method (see Chapter 2), which identifies 6,428 proteins only.

The tissue-centric analyses show that even independently sourced proteomics and transcriptomics of similar tissues present reasonable correlation coefficients. For instance, the range of Spearman correlation ($\rho_{Oesophagus} = 0.39 \leq \rho_i \leq \rho_{Liver} = 0.62$) is consistent with the literature, either published before or during this study (see examples further below).

Besides, the new PPKM quantification method for proteins provides similar Pearson correlation ranges ($r_{Oesophagus} = 0.38 \leq r_i \leq r_{Liver} = 0.61$) to ones previously described for cell studies specifically designed for the joint integration of *same-sourced* proteomics and transcriptomics (*e.g.*, Marguerat et al. (2012), Schwanhäusser et al. (2011), Schwanhäusser et al. (2013), and J. J. Li et al. (2014)).

Most remarkably, all the same-tissue pairs of transcriptome and proteome have a statistically significant correlation despite the tissue proteomic expression profiles closeness.

I have based my following considerations and discussion on the Pearson correlation results even though Spearman correlation is often used in the literature when comparing independent sources of data. The use of the Spearman method is regularly motivated by the lack of data distribution normality. As shown in Figure 5.7, the PPKM quantification produces protein expression levels that share the same logit-normal profile of distribution observed for the mRNAs (see Chapter 3), thus allowing an appropriate use of the Pearson method.

Next, I have considered several possible properties that might be factors influencing the mixed correlation levels. I have compared the expression breadth of the mRNAs and the proteins before comparing the most specific proteins and mRNAs of each tissue to one another.

The mRNAs and proteins expression breadths (*i.e.* the number of tissues within which an mRNA or protein is expressed) share the same overall shape for their distribution, but are only partially concordant. For example, at 5 FPKM, only about 26% of the proteins that are expressed in one tissue and about 40% that are expressed in fifteen tissues have an mRNA with an identical breadth of expression.

The analysis of expression breadth highlights noteworthy facts — including a few recently reported in the literature; *Testis* displays the most unique and diverse expression both at transcriptomic and proteomic levels (see also D. Wang et al. (2019) and Y. Zhang, Q. Li, et al. (2015)); *Liver* (the most correlated tissue) presents the second-highest number of mRNAs and proteins with a unique expression breadth. Besides, when the expression breadth of a protein is unique, the expression breadth of its related mRNA is more likely to be unique at the threshold of 1 or 5 FPKM as well. However, the expression breadth of mRNAs gives no indication of the proteins' one.

Nonetheless, mRNAs' and proteins' expression breadths convey part of the biological signal consistently. From the Jaccard indices of the proteins and mRNAs solely expressed in two tissues, I have built hierarchical trees, and their consensus tree outlines the *Ovary*/*Testis* and the *Kidney*/*Liver* clusters at both proteomic and transcriptomic layers.

I have then compared the TS proteins with the TS mRNAs. The overlaps between the n TS proteins (expressed in a single tissue only) and the n most TS mRNAs of each tissue are non-empty, and except for one tissue (*Urinarybladder*), statistically significant with an α level of 0.01. While most tissues have similar ranking trends for their overlaps of TS genes and correlation between their proteomics and transcriptomics, either high levels (*e.g. Liver*, *Testis* or *Pancreas*), medium (*e.g. Prostate*) or low ones (*Urinarybladder*, *Lung* or *Gallbladder*), other tissues have not.

Regarding the gene-centric analyses, they show that about 6% of the mRNAs/proteins are highly correlated ($r \geq 0.8$), about 18% of them are well correlated ($0.8 > r \geq 0.5$), about 75% of them are poorly correlated ($0.5 > r \geq -0.5$), and less than 1% of the pairs are anticorrelated ($-0.5 > r \geq -0.83$).

A fourth of the genes included in this study have a TS protein. Most of them have a Pearson correlation above 0.5, and they considerably enrich the set of most correlated pairs of mRNAs and proteins. However, their correlation range remains rather wide ($-0.77 \leq r \leq 1$).

Finally, using a GO enrichment analysis, I have investigated whether the three groups of mRNA/protein pairs (the ones with a TS protein, the best correlated pairs and the most anticorrelated ones) are related to any biological or technical reason.

While the most correlated pairs and the ones including a TS protein present an enrichment in biological process (BP), molecular function (MF) and cellular component (CC) terms, the anticorrelated pairs present an enrichment only in BP terms. Overall, the most correlated mRNA/protein pairs seem highly associated with catabolic processes; the pairs with a TS protein are most likely involved with specific signalling (its detection, transduction or answer) and more concentrated in the transmembrane area. The most anticorrelated pairs are enriched in regulation processes.

On top of the true relationship between the expression of the mRNAs and proteins, correlations are dependent on the quantification of identified molecules. In general, while many biological reasons (*e.g.* protein degradation) can lead to midrange or low correlation levels, other technical artefacts may also be at play: saturation effect of more abundant proteins (see Section 1.3.2); degenerate peptides (see Section 1.3.4.3); quantification inconsistencies between methods; platforms and studies [Dapas et al., 2017; Aebersold, Agar, et al., 2018]; annotation, or other oft-neglected sources, *e.g.* gene or isoform length.

One example where the annotation might be at fault is *STAU2* (Pearson $r = -0.59$; Spearman $\rho = -0.65$), which appears to be known to have different annotations in the proteomic and transcriptomic communities. Thus, *STAU2*'s anticorrelation observed between its mRNA and protein expressions might predominantly result from artefacts rather than any biological cause.

Normalisation methods for RNA-Seq require the genes or transcripts length (see Section 1.2.5.4). To keep congruity with the gene length employed by *EBI Gene Expression Atlas*[11] [Petryszak, Keays, et al., 2015], I use the sum of the lengths of all its

[11] *EBI Gene Expression Atlas* — https://www.ebi.ac.uk/gxa/home/

collapsed exons, which is graciously provided by the metapipeline iRAP[12] [Fonseca, Petryszak, et al., 2014].

Although, each gene has various mRNA isoforms [Gonzàlez-Porta, 2014] and proteoforms [Aebersold, Agar, et al., 2018] I chose to perform all the analyses at gene-level expression (see p. 67) due to the many existing criticisms about current algorithms' accuracy with isoforms [Engström et al., 2013; Jänes et al., 2015; Dapas et al., 2017].

Yet, most genes present one dominant transcript [Gonzàlez-Porta et al., 2013]. One possible method to improve the mRNA/protein correlations is to identify the most dominant transcript isoform of each gene and use their length for the normalisation. This method ought to be rather easy to implement, either with the help of resources like *APPRIS*[13] [Rodriguez et al., 2018] or through a direct data analysis. More generally, besides improving current results, (even partial) better identification of the mRNAs and proteins isoforms will most likely unveil still undetected divergences.

In terms of the true relationship between the expression of mRNAs and proteins, it is imperative to remember that the proteome and the transcriptome I use in the analyses are independently sourced and aggregated over several individuals. Therefore, some genes displaying mixed correlation in this thesis may present high correlation or anticorrelation in matched samples. The most affected genes are the most sensitive ones to inter-individual variation and batch effects. On the other hand, the highly correlated (or anticorrelated) pairs highlighted above are more likely having a robust expression across individuals and time, while being less subject to technical noise.

Among the studies published during my thesis' works, three are most notably pertinent to the integration and comparison analyses I have outlined above.

A first published comparison [Kosti et al., 2016] of the expression data of GTEx [Melé et al., 2015] and the Pandey Lab [M.-S. Kim et al., 2014] provides weaker results than the above-presented ones as the authors have based their complete study on data provided as-is by the primary studies. Overall, the range of Spearman correlation they present for the tissues does not exceed $\rho = 0.5$. Additionally, this study also fails to show any functional enrichment in their selection of matched pairs of mRNA and protein. Hence, although these primary data resources are useful for others to appraise the presence of a given gene in a particular tissue, more thorough uses need more considerations and probably preliminary treatments.

Franks et al. (2017) integrate a subset of the Uhlén data included in this chapter analyses as they have extracted their transcriptomic data as-is from Fagerberg et al. (2014) to integrate it with the data they have reprocessed from Pandey Lab data [M.-S. Kim et al., 2014] and Wilhelm et al. (2014). The study's primary aim is to quantify the post-transcriptional regulation of the genes through their translational efficiency. To this end, they compute

[12] iRAP — https://nunofonseca.github.io/irap/
[13] *APPRIS* — http://appris-tools.org/

for each gene a *protein-to-mRNA (PTR)*. PTRs ease the assessment of the gene expression variability across their set of tissues. This study underlines the utmost caution required when one considers the transcriptome as a possible proxy for the proteome. It also reminds that data quality and reliability are primary caveats for possible analyses. Franks et al. (2017) also show that genes from the same set of GO terms display similar relative protein-to-mRNA (rPTR) profile across tissues, *i.e.* they have higher and lower rPTR in the same tissue sets. This observation suggests concerted functional regulations.

Overall, the results presented in this thesis are confirmed and unsurprisingly improved by D. Wang et al. (2019) since they have matching samples (same sources) for the proteomics and transcriptomics. Moreover, there are many biological replicates per tissue. In that follow-up study, the authors have generated proteomics data[14] from the original samples they had used to produce transcriptomic data[15] [Uhlén, Fagerberg, et al., 2015], which I am using in this thesis, including this chapter. While the (Spearman) correlation between the transcriptome and proteome of each tissue has a similar range globally, the expression of an mRNA and its protein across the tissues have a positive correlation in about 90% of all cases and half of them are statistically significant. They also report that there is a core set of ubiquitously expressed pairs of mRNA/protein and that key differences between tissues are more characterised by the level of expression of the molecules rather than their presence or absence. When they compare the highest expressed proteins and mRNAs, they observe a limited overlap that, in my opinion, further confirms that TS genes are more pertinent for integration than the highest expressed genes. They observed that while disease-associated genes are expressed more globally, G protein-coupled receptors are mostly restricted to specific tissues, and are often identified drug targets. Finally, the authors point out that part of the observed discrepancies in the proteomic and transcriptomic expression may be due to the difference in strategies of each community on how to handle the degenerate peptides or multireads. Our PPKM quantification presented in Section 5.2 is one solution to this issue, and it proves to improve the results, particularly for the Pearson correlation.

A companion study, by Eraslan et al. (2019), reports that mRNA expression variation across a set of tissues is a better predictor for protein levels than considering all mRNAs expression levels in a tissue. The authors have thus computed a protein-to-mRNA (PTR) for over 11,500 genes. The study uses these PTR to model and probe possible regulatory mechanisms. However, while Franks et al. (2017)'s PTRs may be partially miscalculated because of the independent sampling sources, for Eraslan et al. (2019), there might be an overfitting problem. Further analyses based on other datasets are required to ensure the reproducibility of these results. They warn that for many genes, PTRs is only useful as a gauge of the protein abundance magnitude order.

Considering the previous papers together with my results, I have similar correlation levels for same-tissue pairs of independent proteomics and transcriptomics to the ones

[14] The proteomic data can be retrieved in PRIDE ID: PXD010154 — https://www.ebi.ac.uk/pride/archive/projects/PXD010154

[15] ArrayExpress ID: E-MTAB-2836 — https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/.

for matched samples (*i.e.* collected from the same source) as I have processed the raw data with consistency.

Despite many deployed efforts, Franks et al. (2017) and Eraslan et al. (2019) suggest that predicting protein levels directly from mRNA levels is still unlikely at the moment.

However, GO information can be useful to appraise protein expression as confirmed in the recent collaborative work I was involved. With the help of Dr Kārlis Freivalds, we have shown that deep learning algorithms that include GO information can reasonably predict the order of expression magnitude of missing proteins in a given label-free MS proteomic study from RNA expression data. While our approach, described in [Barzine, Freivalds, Wright et al. — Barzine et al., 2020], is not an imputation method, it can be used in complement or instead of one.

Beyond any predictive interest, GO annotation, through enrichment analyses for instance, can provide biological or mechanism insights and help the design of new research avenues. Previous studies have reported similar results to the highest correlated pairs of mRNA/protein that are enriched with catabolic genes, and the most anticorrelated ones that are enriched for regulatory processes.

Vogel, Abreu, et al. (2010) have observed the highest protein-per-mRNA ratios for mammalian metabolic genes. Several papers [Vogel and Marcotte, 2012; Schwanhäusser et al., 2011] observe higher expression stability for RNAs and proteins related to mammalian metabolism and a more rapid degradation tendency for proteins involved in transcriptional regulation (and chromatin organisation). Furthermore, organs appear to present different metabolic profiles [Berg et al., 2002], and, regardless of each individual's particulars (*e.g.* sex, alimentary diet, age, physical level), the expression of many catabolic genes only varies according to the tissue.

Accepting the premise that, regardless of the tissue, a similar sequence of regulatory steps apply to a gene (from its transcription to possible post-translational modifications), taken together the above facts may suggest that the catabolic genes may have a more straightforward modulation of their expression than the other genes. In organic chemistry, it is well-known that the more a molecule requires steps to be produced, lesser is its yield (*i.e.* final amount). One possible way to test this hypothesis is to explore if the amount of substrate can regulate these genes' expression levels. An indirect approach can be the comparison across tissues of the co-expression profiles of these catabolic genes with others involved with the active or passive transport of the molecules to be degraded (*e.g.* transmembrane channel proteins).

Another research avenue can be the exploration of possible links between the tissue-specific (TS) genes and the G protein-coupled receptors.

6.5 CONCLUSION

Despite possible batch effects and technical noise, a part of the biological signal is strong enough to be consistently captured by the transcriptome and proteome through direct and indirect analyses.

At tissue level, the independent transcriptomics and proteomics included in these analyses give similar Spearman or Pearson correlation to samples produced from the same biological sources. The signal appears stronger for more homogeneous tissues, *e.g. Liver*, or which expression is more distinctive, *e.g. Testis.*

In any cases, a significant number of tissue-specific (TS) genes are consistently shared between proteome and transcriptome. Besides, even indirect analyses, such as hierarchical clustering trees created with genes that are only shared by two tissues, can highlight identical structures between the two biological layers.

On the other hand, the gene-centric analyses have shown that while only 24% of the mRNA/protein pairs are well correlated (r>0.8), most (about 73%) have a positive correlation for their expression. While the highest correlated genes are enriched in TS proteins, many of them are expressed ubiquitously.

The GO enrichment analysis highlights that genes presenting a TS protein are enriched in specific signalling, genes with the highest correlated mRNA/protein pairs in catabolic processes and genes with the most anticorrelated ones in regulatory processes.

Providing proper care and consistent processing, one may use these independent resources as part of their study to achieve lower but still significant results.

Results can improve from better identification and quantification alone. A possible approach can be the standardisation of annotations between communities. Optimisation or new algorithmic strategies are other ones as illustrated by our PPKM quantification applied for this thesis' works.

Has been done. Can be done. Must be done...
Fandarel [McCaffrey, 1964]

7 CONCLUDING REMARKS

At the time I started my doctorate, an increasing number of gene expression datasets assaying undiseased human tissues for RNA expression were published. In addition, the first genome wide MS-proteomics studies were performed soon after and raw data made available. My primary aim was to integrate and compare these data, first, on RNA level, and second, to compare RNA and protein expression. I concluded that the published RNA measurements were robust and different datasets were highly consistent when processed uniformly. I also found that correlation between RNA and protein expression was typically higher than 0.5, though different groups of genes behaved differently and correlation in some tissues was better than others. Lately, the focus of RNA gene expression studies has been shifting towards single cell level, however genome wide proteomics studies are still in their infancy. Therefore, comparative studies of genome wide transcriptomics and proteomics data on whole tissue level are of significant interest.

SUMMARY

In Chapter 1, I reviewed the biological, chemical and bioinformatic aspects and challenges involved in expression studies based on the high-throughput technologies of RNA-Seq for mRNAs and MS for proteins.

Then in Chapter 2, I presented the five transcriptomic and three proteomics studies that I have considered for this thesis. I also described the pipelines with which I have processed them. Since for each dataset the number and size of files are extremely large, automation is paramount to ensure consistency and minimise errors.

I detailed in Chapter 3 various data quality controls and statistical approaches. I also discussed possible biases and how the contextual scope of the tissues and genes considered for analyses can influence results. To minimise errors due to context issues, I limited most of my further investigations to a common subset of tissues and expressed genes. Normalisation methods are still inadequate to treat mitochondria genes accurately. Therefore, I chose to remove them from most analyses, which led to better results as it allowed me to integrate samples that the original authors had to discard because of their lack of congruency with the other samples of the same tissue.

In Chapter 4, I integrated the five independent transcriptomic datasets and showed that the biological signal dominates the technical noise. All datasets have a higher interstudy correlation for the same tissues than any intrastudy correlation for different tissues. This trend is stronger for more recent studies. I tested various criteria as possible driving forces of the high interstudy tissue correlations. I found that the most variable genes and tissue-specific (TS) genes are more robustly identified across studies than the highest expressed ones. Besides, I also noted that the inclusion of external resources, especially when outdated like TiGER [X. Liu et al., 2008], requires caution. Many listed genes are either wrongly attributed to a tissue or lack to display any specificity and many TS genes highlighted by RNA-Seq are missing. Overall, the integration has revealed that genes present identical general profiles across the studies, even though direct comparisons of independent data may be still impossible. By repeatedly showing that genes show a similar expression profile for a tissue across studies, my analyses prompted the creation of the heatmap visualising expression data and its associated widget for baseline expression data in EBI Expression Atlas[1] [Petryszak, Keays, et al., 2015]. Finally, I provided a core set of genes that are expressed ubiquitously or as TS consistently across all studies.

The comparison of three available proteomic datasets in Chapter 5 illustrates the fragmentation and the disparity of the high-throughput MS-based proteomics. MS detection variability induces considerable technological noise, which explains the intrastudy correlations I observed between different tissues are higher than the interstudy correlation for the same tissues. I provided curated sets of the TS and ubiquitous detected proteins. Finally, with the help of Dr James Wright, I have devised the PPKM quantification method, which quantifies more proteins by also accounting for degenerated peptides.

Chapter 6 reported the integration of the independent proteomic and transcriptomic data. Regardless of the quantification method, I found similar Spearman correlation levels for the included independent studies to those typically observed in the literature for same-sourced proteome and transcriptome source. While proteomic standard state-of-art processing leads to very low Pearson correlation: $0.04 \leq r \leq 0.28$, our PPKM quantification broadly improves this range ($0.38 \leq r \leq 0.61$). Two tissues, *Testis* and *Liver*, have exhibited distinct characteristics across the various analyses that are supported by the literature. *Testis* has the most diverse and specific expression at both transcriptomic and proteomic levels. On the other hand, *Liver* has the most robust expression across studies and the highest correlation between its mRNAs and proteins expression levels. However, there are shared coherent gene signatures across tissues between the proteome and transcriptome. Furthermore, even indirect analyses can capture part of the biological signal. The tissue specificity of proteins and mRNAs are globally more relevant than their expression levels. In addition to the significant overlaps of TS proteins with the most TS mRNAs, most genes with a TS protein have a high correlation between their mRNA and protein expression across tissues. GO analyses show distinct profiles for three gene lists of interest. First, the genes with a TS protein are enriched for specific signalling, including signal detection,

[1] EBI Expression Atlas — https://www.ebi.ac.uk/gxa/home

response pathway and regulation. Second, the genes with the highest correlations between their mRNA and protein expression are enriched for catabolic processes. Thirdly, the genes with the highest anticorrelation between their mRNA and protein expression have shown enrichment for ribosome complexes and ncRNAs regulation. I provide (digitally) the complete set of mRNA/protein pairs with their correlation across the common set of tissues. I also supply the list of overlapping TS proteins and mRNAs.

PRACTICAL CHALLENGES

Throughout this thesis' analyses, I had to overcome many practical challenges. While most of the difficulties encountered ordinarily pertain to Big data projects, one unexpected issue was the current global complexity state of the proteomic world.

Proteomics: a hard field to grasp by newcomers

While constituting a technical obstacle, the characterisation of proteins (or assimilated complexes) is a strong interest for many fields (*e.g.* molecular biology, medicine, drug design, green chemistry). As shown in Section 1.3, the physicochemical properties of the proteins make them intrinsically complex to study, which can partly explain the complexity of the theoretical approaches. Understanding high-throughput proteomics requires many prerequisites. However, there is a lack of a clearly identified entry-level document reviewing the field from the bench to normalised protein levels. The available teaching materials are mostly practical or experimental oriented [Y. Zhang, Fonslow, et al., 2013; Z. Zhang et al., 2014; Domon et al., 2010] or on particular steps, *e.g.* protein inference [He et al., 2016]. The scattered information hampers the ability of newcomers to achieve a global vision of high-throughput proteomics.

Big data Challenges

Big data is often characterised through, what was first defined by IBM[2], the *4 V's*: volume, variety, veracity and velocity. Each of them can entail issues at different project levels.

Volume

The volume of files (see Table 2.2) to handle and process just for the transcriptomics is overwhelming. It requires appropriately dimensioned infrastructure like in the EBI, which can provide high-throughput computing. It is in practice impossible to reproduce the complete work underlying this thesis in a personal computer within a reasonable time. Although (commercial) solutions are increasingly in use for academic projects, dedicated

[2] https://www.ibmbigdatahub.com/infographic/four-vs-big-data

storage, and high computing facilities ease the analyses considerably and allow more in-depth testing. Even if best practices are continuously refined for transcriptomics, there are still many factors that can be improved and tuned. Besides the raw data, storage capacity is also required for the intermediate and final files. Organising such a large amount of data was time-consuming and challenging at times.

Variety

The variety of the type of input data files is kept to a minimum as the data was retrieved from public or academic repositories that follow community guidelines[3]. Issues still ensued from the matching of samples or tissues across the studies. For many tissues, I chose to mix several 'body parts' from the same tissue or organ (*e.g.* 'Left Ventricle' and 'Atrial Appendage' for *Heart*) in GTEx to match them to the other studies' tissues. Perhaps, in some case, one 'body part' is perfectly matched to the samples from another study where the authors have only reported the tissue instead. Hence, for these cases, keeping only one of the 'body parts' would have been a better choice if additional data had been available.

Another source of variety that I have limited in the above work is the diverse annotation versions. Even when mapped to the same genome and annotation versions, discrepancies persist between how transcriptomics and proteomics are defined and assigned to the genes. A possible improvement of the present work will be to develop new quantification tools that use the chromosome coordinates to which mRNAs and proteins map instead of using their gene identifiers.

Veracity

With all possible tunable parameters for the raw data processing, the data to be integrated can vary widely. Although I have tried a limited number of combinations, my study shows that the results' trend remains the same regardless of the chosen settings. Moreover, as the number of datasets I included in my analyses grows, the results became increasingly more stable. Many of them are also confirmed by the literature, adding credibility and confidence to the findings, especially considering the recent discussions on reproducibility crisis [Morrison, 2014; Glenn Begley et al., 2015; Goodman et al., 2016; Fatovich et al., 2017; Coiera et al., 2018; Lindner et al., 2018]. However, results for individual genes can vary from one set of settings to another one, and need to be considered with more caution.

Velocity

Finally, the velocity of new data availability and the required preparation time made the prospect of including all the latest studies in my analyses impractical. Unfortunately, although new GTEx samples or other tissue studies (*e.g.* Oncobox Atlas of Normal Tissue Expression (ANTE) [Suntsova et al., 2019]) kept being released, I had to stop including

[3] ENA for the sequencing data and ProteomeXchange for MS/MS proteomics.

them in my study. Likewise, I ceased updating the genome and annotation and settled for GRCh38.p1 and Ensembl 76.

My resolving approach

To minimise errors, assure consistency and ease future reiterations or extension of these analyses, I provide script files that can reproduce the whole study and its results. I have also automated and structured the analyses through modular functions as much as possible. I have avoided any manual change and have documented all the name change and sample pairings in the scripts. I provide the necessary code to replicate all of the above (and complementary) results as supplementary material[4].

I chose to develop the analyses with open-source software around the programming language *R*[5] [R Core Team, 2019]. See Appendix F for the complete list of R packages involved in this work. This language provides statistical and visualisation functions and is easily expanded through packages developed by the community. While packages are allowing to easily built on previous work, they can be highly interdependent and may evolve rapidly. To draw from an extra package (or fix an identified bug), a comprehensive and time-consuming update of the working environment will often ensue. In turn, I had also to update or rewrite my analyses' code on many occasions. Furthermore, new software installations or updates are more complicated for distributed computing facilities than on a personal computer. Today, new solutions are developed to facilitate these tasks (*e.g. Packrat* [Ushey et al., 2018] that create isolated environment) and, hopefully, this burden will be significantly lowered.

Note that Dr Nuno Fonseca, who provided me with the quantification of the GTEx data, and Dr James Wright, who provided me with the proteomics ones, have also both developed their processing pipelines with open-source software. Hence, the entirety of the thesis can be repeated (conditionally upon access to GTEx data and high-throughput computing facilities).

FUTURE WORKS

Many improvements are conceivable:

- The inclusion of new samples and dataset of transcriptomics (preferably with biological replicates), *e.g.* extend to the last version of GTEx and the ANTE dataset [Suntsova et al., 2019].
- Add the matching proteomics of D. Wang et al. (2019) to the transcriptomic Uhlén data [Uhlén, Fagerberg, et al., 2015] and then compare the results to the unmatched samples.

[4] https://github.com/barzine/BaselineAtlas/tree/thesis

[5] *R* — https://cran.r-project.org

- Work on new models of annotation or build a consensus between the current transcriptomic and proteomic annotations. Today, it is difficult to determine for some genes if the observed anticorrelation or lack of correlation between the mRNA and protein expression levels is due to the biology, batch effects or divergences between transcriptomic and proteomic annotations.
- Changing the quantification (parameters or methods) may also give better results.

New quantification proposal for baseline studies

Most normalised quantification methods are designed for differential expression studies, particularly in transcriptomics [Dillies et al., 2013]. Most of these methods are built with the premise that only a limited number of genes present a differential expression across conditions while most gene expressions remain unaffected by the context [P. Li et al., 2015]. As a result, most normalisations are ill-suited for comparing independent samples across multiple studies. Two normalisation methods do not imply any preconception on the study design: FPKM [Mortazavi et al., 2008; Trapnell et al., 2010] (see Section 1.2.5.4), which I use in this thesis, and TPM (Transcript per Million)[6] [Wagner et al., 2012]. Both normalisation methods account for the global library size of each sample. While the motivation is sound, the quantification is thus contextual to how many and how much RNAs are detected.

Previous efforts to improve the quantification have focused on differential expression studies. Synthetic spike-in molecules [L. Jiang et al., 2011] ensure more reliable quality controls. However, while theoretically plausible, accurate absolute quantification for the whole dynamic expression range has yet to be reached [L. Jiang et al., 2011; SEQC/MAQC-III Consortium, 2014; Hardwick et al., 2017]. Furthermore, spike-ins fail to permit the absolute quantification of the other molecules in the sample. Incidentally, Rudnick et al. (2014) find that spike-in proteins and peptides lack effectiveness for proteomic studies. For proteomics, Wiśniewski, Hein, et al. (2014) propose to use the histone to create a '*proteomic ruler*'. They can assess through this proteomic ruler the amount of DNA in the sample, and thus, give an estimate on the cell number, which provides some context to interpret the quantification of the proteins. Histone genes are, however, ill-suited for bulk RNA-Seq studies [Zhao et al., 2018].

I firmly believe that using internal standards chosen within the naturally expressed population of the studied macromolecules is more appropriate than any external additive. I expect that giving each gene (RNA or protein) expression as a ratio of the expression of a reference gene will be more robust. It will free the expression from the influence of the presence of quantification of any other gene while it will still account for the difference of sequencing depth between samples. Then, the next question is: 'Which gene (or set of genes) will possibly enable the best normalisation?'

[6] FPKM is easily converted to TPM by scaling with a constant to correct the sum of all values in a library to 1 million.

Through this thesis' various analyses, I have highlighted genes that have a robust and ubiquitous expression across all the studied tissues both at transcriptomic and proteomic levels. Moreover, many genes present high correlation coefficients between the expression of their mRNA and protein. These genes are the best potential candidates as reference. With Dr Nuno Fonseca, we have performed preliminary analyses in this direction across other studies to reduce the list of these candidates.

However, considering the best practices in analytical chemistry [Arvid, 1997], a set of standards that can cover the complete dynamic expression ranges of the considered molecules and adjust for the saturation effects, may resolve the abundances better, and thus, be better suited than a single reference. On the other hand, studies focusing on a single tissue may better benefit from a reference built on TS genes.

The recent development of single-cell transcriptomics (scRNA-Seq) [G. Chen, Ning, et al., 2019], and the probable feasibility of single cells proteomics [Marx, 2019], may ease refining which genes are the most suitable as universal references, or for specific conditions.

Ongoing implementation of an application

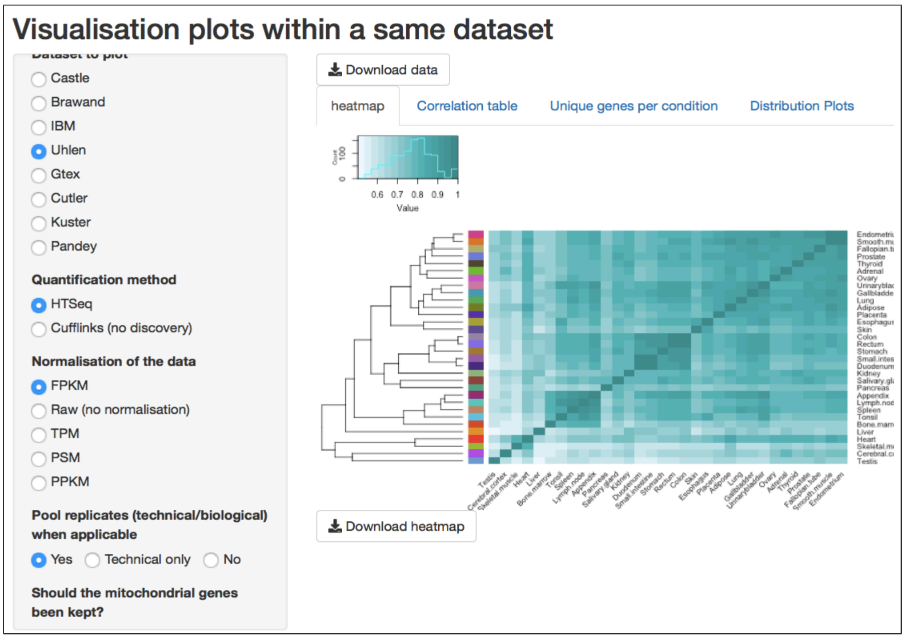
Integrating proteomics and transcriptomics remains laborious, and results can be mixed. Nonetheless, even simple comparisons can help to improve our general knowledge and current biological models. For instance, in one of our papers [Wright, Mudge, et al., 2016], we have confirmed the existence of putative proteins by observing coverage of the genome both by transcriptomics and proteomics.

To help further possible projects, I am currently compiling the different analyses into a set of interactive applications that can replicate all the results and figures presented in this thesis without requiring any programming skill as a prerequisite. Figure 7.1 covers Chapter 3[7].

Once completed, one will be able to analyse and compare their own data to the different datasets I have presented in this thesis. Furthermore, I share all my code (including for the application) under a creative commons license, *Attribution 4.0 International (CC BY 4.0)* 🅭🅯[8]. Anyone can thus use, adapt or build upon this work as they wish.

[7] For a live demo, see http://barzine.net/shiny/mitra/thesis/chapter3/.

[8] *Attribution 4.0 International (CC BY 4.0)* 🅭🅯 — https://creativecommons.org/licenses/by/4.0



Visualisation plots within a same dataset

Dataset to plot

○ Castle
○ Brawand
○ IBM
● Uhlen
○ Gtex
○ Cutler
○ Kuster
○ Pandey

Quantification method

● HTSeq
○ Cufflinks (no discovery)

Normalisation of the data

● FPKM
○ Raw (no normalisation)
○ TPM
○ PSM
○ PPKM

Pool replicates (technical/biological) when applicable

● Yes ○ Technical only ○ No

Should the mitochondrial genes been kept?

Figure 7.1. **Preview of the application** developed with R and served through a shiny server [Chang et al., 2019].

APPENDIX

A SUPPLEMENTARY MATERIAL FOR CHAPTER 1

A.1 AMINO ACIDS

As shown in Figure A.1, the amino acids have different chemical properties. Their primary and side chains are respectively shown in black and green. The amino acids all share the same primary chain.

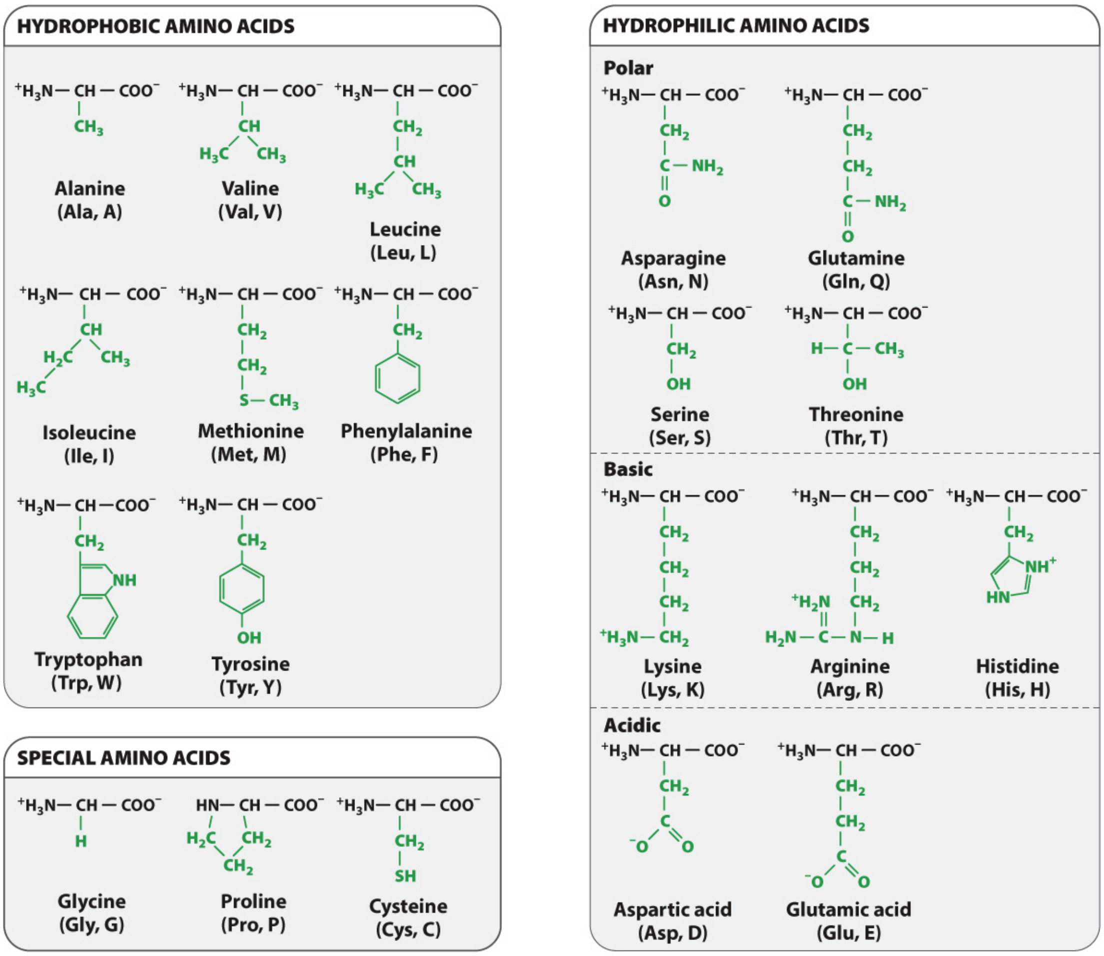


Figure A.1. **Amino acids formulas** — from Morris et al. (2016)

Table A.1. Molecular weight of the most common aas and their residues (from Lide, 2005)

Name	Abbr.		Molecular Formula	Molecular Weight	Residue Formula	Residue Weight (-H2O)
Alanine	Ala	A	$C_3H_7NO_2$	89.10	C_3H_5NO	71.08
Arginine	Arg	R	$C_6H_{14}N_4O_2$	174.20	$C_6H_{12}N_4O$	156.19
Asparagine	Asn	N	$C_4H_8N_2O_3$	132.12	$C_4H_6N_2O_2$	114.11
Aspartic acid	Asp	D	$C_4H_7NO_4$	133.11	$C_4H_5NO_3$	115.09
Cysteine	Cys	C	$C_3H_7NO_2S$	121.16	C_3H_5NOS	103.15
Glutamic acid	Glu	E	$C_5H_9NO_4$	147.13	$C_5H_7NO_3$	129.12
Glutamine	Gln	Q	$C_5H_{10}N_2O_3$	146.15	$C_5H_8N_2O_2$	128.13
Glycine	Gly	G	$C_2H_5NO_2$	75.07	C_2H_3NO	57.05
Histidine	His	H	$C_6H_9N_3O_2$	155.16	$C_6H_7N_3O$	137.14
Hydroxyproline	Hyp	O	$C_5H_9NO_3$	131.13	$C_5H_7NO_2$	113.11
Isoleucine	Ile	I	$C_6H_{13}NO_2$	131.18	$C_6H_{11}NO$	113.16
Leucine	Leu	L	$C_6H_{13}NO_2$	131.18	$C_6H_{11}NO$	113.16
Lysine	Lys	K	$C_6H_{14}N_2O_2$	146.19	$C_6H_{12}N_2O$	128.18
Methionine	Met	M	$C_5H_{11}NO_2S$	149.21	C_5H_9NOS	131.20
Phenylalanine	Phe	F	$C_9H_{11}NO_2$	165.19	C_9H_9NO	147.18
Proline	Pro	P	$C_5H_9NO_2$	115.13	C_5H_7NO	97.12
Pyroglutamatic	Glp	U	$C_5H_7NO_3$	139.11	$C_5H_5NO_2$	121.09
Serine	Ser	S	$C_3H_7NO_3$	105.09	$C_3H_5NO_2$	87.08
Threonine	Thr	T	$C_4H_9NO_3$	119.12	$C_4H_7NO_2$	101.11
Tryptophan	Trp	W	$C_{11}H_{12}N_2O_2$	204.23	$C_{11}H_{10}N_2O$	186.22
Tyrosine	Tyr	Y	$C_9H_{11}NO_3$	181.19	$C_9H_9NO_2$	163.18
Valine	Val	V	$C_5H_{11}NO_2$	117.15	C_5H_9NO	99.13

A.2 ORIGINAL MATERIAL

To create Figure 1.1, I used original material by Kelvinsong (https://commons.wikimedia.org/wiki/User:Kelvinsong): 'Simplified diagram of mRNA synthesis and processing. Enzymes not shown.' (https://commons.wikimedia.org/wiki/File:MRNA.svg) and 'Protein synthesis' (https://commons.wikimedia.org/wiki/File:Protein_synthesis.svg).

A.3 EXPRESSED SEQUENCE TAG (EST) SEQUENCING

ESTs are short nucleotide sequence generated from randomly selected RNA transcript [Parkinson et al., 2009]. mRNAs are reverse transcribed into double-stranded cDNAs (either from the 5' or 3' end of the transcript) [Lowe et al., 2017]. These cDNAs are cloned to create libraries [Harbers, 2008] and then sequenced either by Sanger method [Sanger et al., 1975] or a more high-throughput one such as the sequencing-by-synthesis (Section 1.2.3). Although this technique is subject to sampling bias [Nagaraj et al., 2007] and often account for only 60% of an organism expressed genes [Bonaldo et al., 1996], it remains a relatively low cost alternative approach to study the transcriptome (gene discovery).

A.4 MICROARRAYS

Microarrays require prior knowledge (*e.g.* annotated genome or ESTs libraries) of the organism of interest as they exploit it to design *probes* (short nucleotide oligomers) that are arrayed on a solid support (*e.g.* a glass or silicon thin film cell) [Lowe et al., 2017; Schena et al., 1995; Bumgarner, 2013]. For transcriptome profiling, the expressed RNAs are first reverse transcribed into cDNAs (also referred as *targets*) and then, after being fluorescently labelled, they are complimentary hybridised to the microarray probes; the relative abundance of the transcripts is assessed by measuring the intensity of the fluorescence after the excess of unhybridised cDNAs is washed away [Lowe et al., 2017]. This technology is extremely powerful and popular as it allows global and parallel analyses of cellular activity. Microarray technology also has many variations [Hoheisel, 2006] in addition to its original cDNA version for transcriptional profiling [Schena et al., 1995], *e.g.* for genotyping [D. G. Wang et al., 1998; Gunderson et al., 2006], protein profiling [Hall et al., 2007; Sutandy et al., 2013; Duarte et al., 2017], splice-variant analysis [Cuperlovic-Culf et al., 2006] or transcription factor binding [Bulyk et al., 2002; Bulyk, 2007] studies.

A.5 FASTQ FORMAT

```
@ERR030856.1 HWI-BRUNOP16X_0001:1:1:2669:1073#0/1
AAAGGATTATGCAGANGTAGGGCGTGTNNNNNNNNNNNNNGGCTGGGGNNNNNNNNNNNNNNNNNNATNNNCTGACCANCTGAAGTATGTCANGCTGCCT
+
HHHHHHHIHHFFFFF#>>@>GGGFG###########################################################################
@ERR030856.2 HWI-BRUNOP16X_0001:1:1:4476:1072#0/1
GATAGATTATCAGAANGACAGTTACTTNNNNNNNNNNNNNGGGCACTTNNNNNNNNNNNNNNNNNNATNNNTCATAAGNNCTGTTGCCAAATNAGTGATA
+
HHHHHHHHHHDDDDD#@@AAGGGGG###########################################################################
```

Legend:
- Read identifier
- Optional information (here flow cell lane:tile number:x:y:z)
- First member of pair (here) or single-end
- Nucleotide sequence of the read
- Separator (+ or any string of character)
- Phred score (here Phred 33)

Figure A.2. FASTQ format

A.6 PHRED SCORE

Table A.2. Phred quality score to accuracy significance

Phred quality score (Q)	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

The Phred quality score can be encoded in several standards as shown in Figure A.3.

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...............................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL...............................................
!"#\$\%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                        |    |        |                              |                    |
33                        59   64       73                             104                  126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0........................26...31........41

S - Sanger        Phred+33, raw reads scores between  0 and 40
X - Solexa        Phred+64, raw reads scores between -5 and 40
I - Illumina 1.3+ Phred+64, raw reads scores between  0 and 40
J - Illumina 1.5+ Phred+64, raw reads scores between  3 and 40
        with 0=unused, 1=unused, 2=Read segment Quality Control Indicator
L - Illumina 1.8+ Phred+33, raw reads scores between  0 and 41
```

Figure A.3. The available Phred score quality score encoding formats

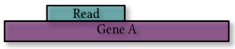
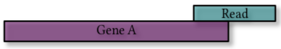
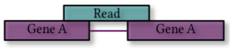
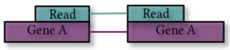
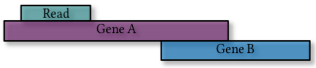
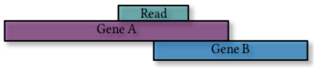
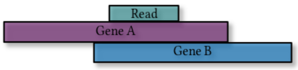
	Union	Intersection_strict	Intersection_nonempty
	Gene A	Gene A	Gene A
	Gene A	No feature	Gene A
	Gene A	No feature	Gene A
	Gene A	Gene A	Gene A
	Gene A	Gene A	Gene A
	Ambiguous	Gene A	Gene A
	Ambiguous	Ambiguous	Ambiguous

Figure A.4. **Overlap resolution effects for each *HTSeq-count* mode.** Each mode resolves a number of overlap situations differently. The mode used in this thesis is the `intersection non-empty` mode. This specific mode resolves more situations than the two others. Hence, the loss of ambiguous reads is reduced in this mode. [Adaptated from HTseq documentation: http://www-huber.embl.de/HTSeq/doc/count.html]

Table A.3. FPKM are unsuitable for differential expression analysis

	Sample 1		Sample 2	
	raw counts	normalised counts	raw counts	normalised counts
$Gene_1$	100	0.010	80	0.008
$Gene_2$	100	0.010	80	0.008
...	...	...	...	...
$Gene_i$	100	0.010	80	0.008
$Gene_{i+1}$	0	0	2000	0.2
Total number of fragments (F)	10,000	1	10,000	1

A.7 MASS ANALYSERS

See Haag (2016) for more details on other types of analysers.

Quadrupole analyser

It is one of the most popular analysers as they are cheap compared to the others. They are also compact, durable and reliable. The quadrupole analyser can filter the ions based on their difference of m/z. They are adequately named quadrupole as they comprise four cylindrical or hyperbolic rods in parallel to each other. Opposite rods are connected together electrically and radio frequency (RF) potential is applied. A direct current (DC) potential is superimposed on the RF one. These combinations of RF and DC potentials constrain the ions to oscillate between the rods as they pass through them. Hence, by tuning the RF and DC, it is easy to select for which range of m/z ions will have a stable trajectory and thus the only one detected. Indeed, the ions with unstable trajectory will collide with the rods and be 'filtered' out. If used in 'RF-only' mode (DC reduced to a minimum), the quadrupole may have other applications. For example, it can guide specific m/z ions to other areas (while the bulk of ions will remain trapped). It may also be used as collision cells for CID: by introducing an inert gas and tuning with the RF-energy, the amount of fragmentation undergone by the targeted ions can be precisely controlled. [Haag, 2016]
The quadrupole analyser is also qualified as the *mass filter*.

Linear trap quadrupole (LTQ)

LTQ is a particular kind of linear ion trap (LIT) which is in principle a sort of a quadrupole mass analyser [Z. Zhang et al., 2014]. A LTQ uses a set of quadrupole rods and a two-dimensional RF field confines the ions radially. In addition, a static electrical potential is applied to end electrodes which forbid the ions to escape axially. However, the quadrupole is commonly segmented into three parts which ensure a perfect homogeneity of the electric field of the trap area and thus avoiding ion loss when the trapping is done. While they may be used as an ion trap, they may be also used as a simple mass filter. RF voltage is tuned to produce multi-frequency resonance ejection waveforms are applied as to eliminate all the undesirable ions in the trap before the fragmentation and mass analysis of the remaining ones. Frequently, these LTQs are used as a front-end to other mass analysers as they have high injection efficiencies and high ion storage capacities. They may be equipped then with two biased radial ejection slits and then be used with two detectors hence the signal-to-noise ratio may be doubled.

Compared with other traps, linear ion traps provide an enhanced dynamic range with a reduced low mass cut-off as the ion cloud is spatially distributed on a linear axis and not a 3D centre which improves the sensitivity. And then, for example, the ions may then be accumulated before being released into another mass analyser [Madalinski et al., 2008].

Orbitrap™

It is a very recent analyser and it relies on FT. Recently, there is increasing use of FTMSs for proteomic studies. Indeed, these FTMSs are more precise than previous analysers and allow the detection of a greater range of ions in very short lapses of time [Scigelova et al., 2011]. In this kind of analyser, ions are trapped and both orbit around and oscillate in an electrostatic field between an inner and outer part of a central electrode shaped as a spindle. The ions can only move following the spindle long axis [Makarov, 2000]. While moving around the spindle the ions create a current. The outer part of the spindle records images of this current. Fourier transformation of these images allows obtaining very highly accurate and sensitive mass spectra for a greater dynamic range than most of the other analysers. [Q. Hu et al., 2005]

LTQ-Orbitrap™

It is a hybrid (tandem) mass spectrometer that uses ESI for the ionisation step and has an LTQ as a first analyser (MS^1) and an Orbitrap™ as a second one (MS^2). This MS/MS enables multiple levels of fragmentation for the elucidation of a wide range of peptides and can be coupled with an ESI which is a continuous source of ionisation. This instrument allows analysing proteomic samples optimally both in terms of starting material, time [Scigelova et al., 2011] and provides ‘ultrahigh’ mass resolution, high mass accuracy and enhanced dynamic range with respect to mass accuracy [Madalinski et al., 2008].

A.8 ISOTOPES OF COMMON ELEMENTS AND THEIR NATURAL FREQUENCY

Table A.4 lists the mass [Audi et al., 1993; Audi et al., 1995] and the percent natural abundance [Rosman et al., 1998] for stable nuclides (*i.e.* atom distinctly characterised by its number of protons (Z) and number of neutrons (N)) that may be found in DNAs, RNAs and proteins.

Table A.4. **Most common constitutive elements and their stable isotopes found in DNAs, RNAs and proteins.** Asterisks (*) mark abundances that are not available. Adapted from [Audi et al., 1993; Audi et al., 1995; Rosman et al., 1998]

z (Atomic number)	Name	Isotope	Mass atomic (u)	Natural frequency (%)
1	Hydrogen	^{1}H	1.007825	99.9885
	Deuterium	^{2}H	2.014102	0.0115
	Tritium	^{3}H	3.016049	*
6	Carbon	^{12}C	12.000000	98.93
		^{13}C	13.003355	1.07
		^{14}C	14.003242	*
7	Nitrogen	^{14}N	14.003074	99.632
		^{15}N	15.000109	0.368
8	Oxygen	^{16}0	15.994915	99.757
		^{17}0	16.999132	0.038
		^{18}0	17.999160	0.205
15	Phosphorus	^{31}P	30.973762	100
16	Sulphur	^{32}S	31.972071	94.93
		^{33}S	32.971458	0.76
		^{34}S	33.967867	4.29
		^{35}S	35.967081	0.02
53	Iodine	^{127}I	126.904468	100

A.9 HYPOTHESIS TESTING

A.9.1 $\mathcal{H}_0$

In statistical testing, the null hypothesis $\mathcal{H}_0$ is an answer to the intrinsic nature of statistical calculation: the smaller a given interval is, the lower the probability of a simple random draw in that interval. The null hypothesis can be of different natures. It is generally formulated as an absence of difference between two objects to be compared, or as an absence of relationship between two variables of a same population; its purpose is to be rejected. It is always opposed to another *alternative* hypothesis ($\mathcal{H}_1$), which is accepted when $\mathcal{H}_0$ is rejected.

To test an hypothesis, one needs to construct a statistical model that can represent an ideal form of the data if it were to be generated by random processes alone. This model is also referred as the *distribution under the null hypothesis*. Then, the likelihood of the collected (observed) data is computed. Finally, it is compared to the (random) probability determined by the model to either accept $\mathcal{H}_0$ or reject it if the observed data is very unlikely under the null hypothesis. Usually a test statistic (*i.e.* quantity derived from the sample used for the hypothesis testing) that measures the apparent departure from the null hypothesis is compared to a value defined such as the probability of a 'more extreme value' is even smaller under the null hypothesis. Prior to the analysis, an arbitrary level of significance (or α) is set either to 0.1, 0.05, 0.01, 0.005 or 0.001, *i.e.* 10%, 5%, 1%, 0.5% or 0.1% risk to reject $\mathcal{H}_0$ by mistake.

Depending on whether the observed data is tested, case (1): in both direction, *i.e.* the data is either *greater or equal* to the critical value (x) or *lesser or equal* to the additive inverse of the critical value ($-x$), or, case (2) in one direction only, *i.e.*, (for example) the data is (only) *greater or equal* to the critical value (or the data is (only) *lesser or equal* to the critical value), the statistical test is two-tailed (case 1) or one-tailed.

A.9.2 *p-value*

In statistical hypothesis testing, the p-value quantifies the statistical significance of results, under the null hypothesis $\mathcal{H}_0$ (see Appendix A.9.1). It allows rejecting (or not) $\mathcal{H}_0$. The p-value is the probability for a given statistical model of obtaining an equal value or an even more extreme value than what has been observed when $\mathcal{H}_0$ is true. Depending on the situation, the more extreme value can mean:

One-tail event	Left tail event	$\Pr(X \leq x \mid \mathcal{H}_0)$
	Right tail event	$\Pr(X \geq x \mid \mathcal{H}_0)$
Two-tail event	$2\min(\Pr(X \leq x \mid \mathcal{H}_0), \Pr(X \geq x \mid \mathcal{H}_0))$	

The smaller is a p-value, the higher the significance, *i.e.* the stronger the evidence that $\mathcal{H}_0$ has to be rejected. $\mathcal{H}_0$ is rejected if the adequate probability is less than or equal to an arbitrary pre-defined (*i.e.* prior to the analysis) threshold value α.

Under $\mathcal{H}_0$, the assumption is that the p-values are uniformly distributed.

A.9.3 *q-value*

A q-value is an adjusted p-value (which the calculation may or may not be based on the p-value). In the context of multi-testing, *i.e.* when multiple simultaneous statistical tests occur, the likelihood of rejecting $\mathcal{H}_0$ due to a random sampling increases. To avoid accepting the alternative hypothesis by mistake, the whole p-values collection is tested and adjusted for false discovery rate. A q-value of 5% means that 5% of all the significant results are actually false positives.

A.10 TARGET DECOY SEARCH DATABASE

For best effectiveness, decoy and target peptide sequence databases are searched with the same parameters. Furthermore, to ensure that a wrong hit in the target database and a hit in the decoy one are equally likely, the decoy sequences have to be as similar as possible to the target ones (concerning aa frequencies and composition, length, mass, charges, assigned scores). There are different ways to design the decoy sequences. For example, by reversing the peptide or protein sequences, either with complete or pseudo-reversion (where the last aa is kept in place). Alternatively, by using stochastic methods on the target database such as the randomisation of the sequences or through the creation of new ones based on aa frequencies, their length distribution, and their number in the original database; Markov models [Gagniuc, 2017] are often used to mimic the closest target sequences. Many studies explore and compare the different decoy creation methods [Elias et al., 2007; G. Wang et al., 2009; Elias et al., 2010; Wright and Choudhary, 2016]. As for the target database, the decoy sequences are digested *in silico* before the search. While the search can be done independently on the target and the decoy databases [Blanco et al., 2009], Elias and Gygi (2007) report that searching their resulting concatenation gives better results.

Besides, the TDA approach can guide the selection of sensitive PSM attributes (*e.g.*, elution time, charge, peptide length, score) as filtering criteria to discern correct identifications [Elias et al., 2010].

A.11 PSM VALIDATION WITH Q-VALUE AND PEP

A possible definition of PSM's *q-value* is the minimal FDR threshold for which the PSM is accepted as correct. As the q-values are derived from the FDR, which is specific to a PSM collection, they are also (solely) specific to this collection. For example, *Percolator* estimates the q-value by using the score distribution from the TDA. On the other hand, posterior error probability (PEP) (also known as *local FDR* [Efron et al., 2001]) is the probability of a PSM being incorrect; a PSM's PEP is independent of the PSM collection.

A classical approach to estimating PEPs uses training sets of target and decoy PSMs to learn the parameters of a probability model (indispensable to compute the PEPs). Thus, for each given score (of any collection), a specific PEP is associated. Choi et al. (2008) showed that for a given collection, the sum of the PEPs is equal to the expected number of incorrect PSMs, which allows calculating the (global) FDR.

A.12 PROTEIN INFERENCE: COMPUTATIONAL CHALLENGING STEP

To explain the computational challenges of the peptide assembly, T. Huang et al. (2012) propose to start with two assumptions: (1) all (m) peptides are true positive, and (2) peptides have an equal probability of detectability. A first assumption corollary is the presence of many homologous proteins in the sample.

Besides, one can derive from the first assumption that there are a minimal and a maximal value for the number n of proteins that can be identified from the set of m peptides. Returning the exhaustive list of proteins (*i.e.* n_{Max}) that comprise all m peptides (*e.g.* Tabb, McDonald, et al. (2002)) is one possible solution, but it is much more difficult to calculate the minimal list (*i.e.* n_{min}) that does the same. As all the peptides m are supposed to be true, they have to be included in any of the final minimal list proteins. Therefore inferring this protein list can be formulated as a *set covering problem* [Cormen et al., 2009; Hochbaum, 1997]. The set covering problem is known to be NP-complete [van Leeuwen et al., 1990], and for which it is in practice impossible to calculate an optimal solution. Usually, algorithms approximate this solution through a parsimonious approach.

Many inference algorithms seek a compromise between the minimal and exhaustive lists of possible proteins. While the minimal list probably excludes many true positives (but can still include false positives), the exhaustive list is indisputably comprising a large number of false positives as the parameters are set to maximise the number of peptide/proteins matches; statistically, a subset of these matches are random. In the sequence database, there are many proteins with the same peptidic sequence, *e.g.* a protein A, expressed in one set of cells, and another, B, expressed only in another non-overlapping set of cells. While a sample is only expressing A, an exhaustive solution will also report B as one of the proteins expressed in the sample. Statistically, the greater the size of the reference database and the expression complexity of the sample, the greater is the number of false positives in the results because of degenerate peptides.

On the other hand, if a peptide is associated to one unique protein in a database when this peptide is identified with high confidence in a sample, it is extremely probable that the protein is truly present. However, these one-hit wonders are also trickier because the protein presence is reduced to the probability of the peptide to be a true positive instead of an artefact. Even a greater number of MS/MS spectra supporting a peptide existence is only the reflection of a remarkably low probability of the protein being absent in the sample.

In this hypothetical setting where all peptides are true positives and equally likely to be detected, the inference is already challenging; it becomes even more complex with real data. The minimal list can be shorter than the theoretical one as the identified peptides can be false positives. On the other hand, proteomic platforms and pipelines tend to repeatedly and consistently detect and quantify particular sets of peptides (*proteotypic* peptides) [Mallick et al., 2007; Bergeron et al., 2007; Fusaro et al., 2009]. Thus, many peptides are difficult to capture with MS. This has two implications.

First, many algorithms associate additional information to the bipartite peptide/protein graph (shown in Figure 1.14) to improve the identification coverage. The algorithms can exploit different data sources, *e.g.* raw and corrected PSM scores, single stage MS or raw MS/MS spectra, peptide expression profiles, mRNA expression data, protein-protein interaction network or gene model. T. Huang et al. (2012) propose that additional information can further extend the exhaustive list of possible proteins.

Secondly, proteotypic peptides have led to the development of *peptide detectability* [Tang et al., 2006; Alves et al., 2007], which can help to deal with degenerate peptides by attributing probabilities to each peptide/protein assignment. Peptide detectability is considered as a intrinsic peptide property. It is only determined by the peptide primary sequence and its location within the protein.

Many different algorithms tackle this peptide assembly key step. T. Huang et al. (2012) organise them in two categorisation frameworks: one based on the needed search engine for the list of PSMs, the second one (presented in Figure A.5) based on the underlying algorithmic technique.

T. Huang et al. (2012) describe parametric approaches as those that request prior knowledge to estimate the peptides' distribution. When there is no need for prior knowledge, they describe the approach as non-parametric, even when the tool assesses the peptide distribution by extracting information from the MS or MS/MS spectra.

A.13 BAYESIAN INFERENCE

Bayesian inference is developed upon Bayes' theorem, which allows the computation of conditional probabilities, *i.e.* computation of the likelihood of an event happening given prior known conditions, including the likelihood of another event being true.

Bayesian statistics focuses on the credibility of events happening rather than their occurrence frequencies (as in frequentist statistics). See B. Li et al. (2019) for general mathematical definitions of Bayesian inference models and Kurt (2019) for a layman's guide to Bayesian statistics.

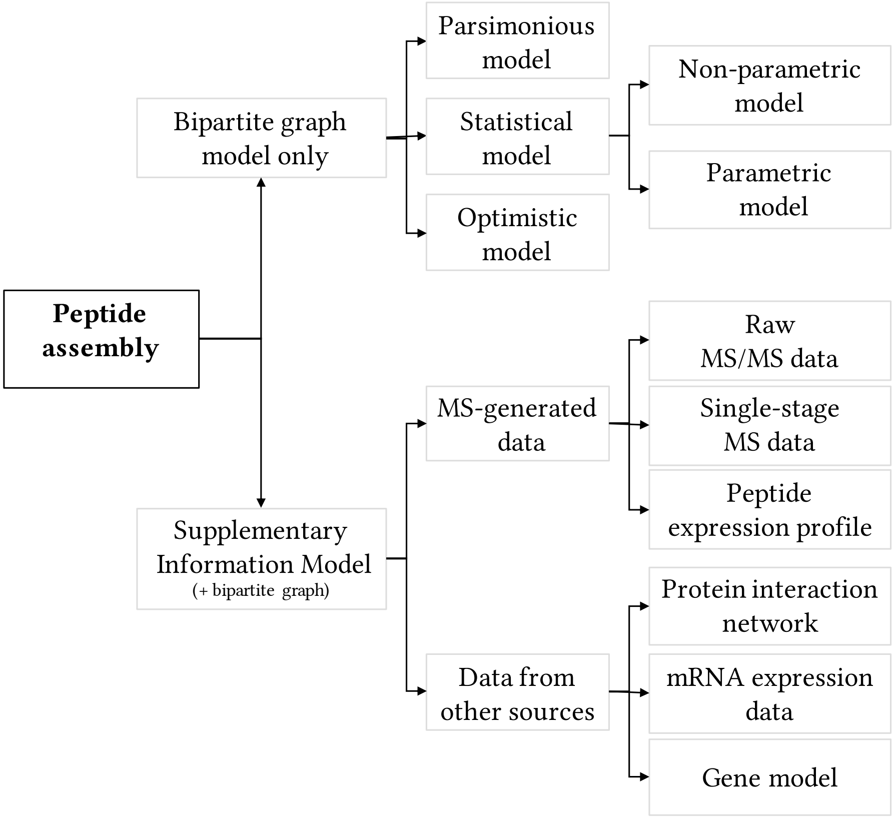


Figure A.5. **T. Huang et al. (2012) peptide assembly models classification.**

B | SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 CORRELATION

Correlation can be considered as a scaled version of the covariance (Equation (Covariance)) of two random variables. Correlation coefficients are adimensional and varie in a restricted range $[-1, 1]$. While 1 and -1 mean a perfect correlation (either positive or negative), a value equal to 0 expresses that the two variables are not sharing any linear relationship. A value within $(-1, 0)$ or $(0, 1)$ needs more interpretation. In gene expression studies, if the coefficient is within $[-0.5, 0.5]$, the variables are generally considered as independent.

Spearman and *Pearson* are only two methods to compute correlations among other ones.

B.1.1 *Spearman correlation*

The Spearman correlation coefficient (usually noted as ρ) is more robust than the Pearson correlation. However, it only assesses the monotonic dependence between two variables. The Spearman correlation coefficient is defined as the Pearson correlation of the *ranked values* of two variables. Spearman correlations are widely used within the literature for biological studies [Brawand et al., 2011; Fagerberg et al., 2014; Danielsson et al., 2015; N. Y.-L. Yu et al., 2015].

B.1.2 *Pearson correlation*

The Pearson correlation coefficient (usually noted as r) assesses the linear dependence between two variables. It is invariant to systematic addition of a constant or to simple scaling factors between the two variables.

The correlation coefficients computed for this thesis rely on the *Sample correlation coefficient* (as opposed to the *Population* formula — see equation (Population correlation coefficient)).

The (sample) Pearson correlation coefficient can be defined as the following equation (indeed, many rearrangements are possible):

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{((Sample) Pearson correlation coefficient)}$$
$$= corr(x, y)$$

where:

• x, y	are observed values of two random variables X and Y
• n	is the sample size of x and y
• i	is the index of the current observed value x or y
• $\bar{x}, \bar{y}$	are respectively the sample (x and y) means (see Equation (Mean))
• $corr(X, Y)$	is another notation of $r_{x,y}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Mean)}$$

where:

• x	is the possible observed values of X
• n	is the sample size of x
• i	is the index of the current observed value of x

B.1.3 *Different advantages of Pearson and Spearman correlations*

Pearson correlations are easier to understand, interpret and then to use as predictor while Spearman correlations are more robust and thus better fitted for interstudy comparisons. Computationally, correlations (Spearman in particular) can be challenging to compute, especially for large matrices such as gene expression matrices [S. Wang et al., 2014].

de Siqueira Santos et al. (2014) review Spearman and Pearson correlations along with six other statistical methods. They also summarise many use cases of each of these methods in the general context of gene expression study.

Table B.1. **Correlation coefficients between RNA-Seq replicates** Numeric summary of Figure 3.5 — The correlation means are high across the studies replicates. However, the range of the correlations (in brackets) are quite extreme in a few case.

Tissue	Replicates type	Pearson Correlation	Spearman Correlation
Brawand	Biological	0.90 [0.45;1]	0.93 [0.80;0.99]
GTEx	Biological	0.75 [0.01;1]	0.93 [0.06;0.99]
Uhlén	Biological	0.81 [0.15;1]	0.95 [0.70;0.99]
	Technical	0.99 [0.68;1]	0.99 [0.92;1]

B.2 OTHER COMMON MATHEMATICAL DEFINITIONS

The population correlation coefficient (ρ) of two random variable X and Y is defined as:

$$\begin{aligned}\rho_{X,Y} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\ &= corr(X,Y)\end{aligned} \qquad \text{(Population correlation coefficient)}$$

where:

- X, Y are two random variables
- $cov(X,Y)$ is the covariance of X and Y (see Equation (Covariance))
- σ_X, σ_Y are the standard deviations of X and Y (see Equation (Standard deviation))
- $corr(X,Y)$ is another notation of $\rho_{X,Y}$

The covariance is the measure of the joint variability of two random variables, *e.g.* X and Y. Specifically, it allows quantifying the degree to which two variables are linearly associated.

$$cov(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \qquad \text{(Covariance)}$$

where:

- X, Y are random variables
- x, y are respectively one observation of X and Y
- $\bar{x}, \bar{y}$ are the means of all observed values of X and Y
- N is the number of observations of X and Y

The standard deviation (sd or σ) measures the amount of dispersion of the possible values of a random variable around its expected value (E) (theoretical average).

$$\begin{aligned}sd(X) &= \sqrt{E[X^2] - (E[X])^2} \\ &= \sqrt{Var(X)}\end{aligned} \qquad \text{(Standard deviation)}$$

where:

- X is a random variable
- $E[X], E[X^2]$ are respectively the expected values (or theoretical averages) of X and X^2 (see equation (Expectation))

$$\begin{aligned}E[X] &= x_1 p_1 + x_2 p_2 + \cdot + x_k p_k \\ &= \text{weighted average}(X) \\ &= \mu_X\end{aligned} \qquad \text{(Expectation)}$$

where:

• E	is the expectation
• X	is a random variable
• $x_1, x_2, ..., x_k$	are possible value of X
• $p_1, p_2, ..., p_k$	are the probabilities of the different values of X and their sum is equal to 1.
• μ_X	is the theoretical average of X

$$\begin{aligned} Var(X) &= \frac{\sum (x_i - \bar{x})^2}{N - 1} \\ &= sd^2(X) \end{aligned} \qquad \text{(Variance)}$$

where:

• X	is a random variable
• x	is one observation of X
• $\bar{x}$	is the mean of all observed values of X
• N	is the number of observations of X
• sd^2	is another notation of the variance as the standard deviation is equal to the square root of the variance.

B.3 DATA VISUALISATION

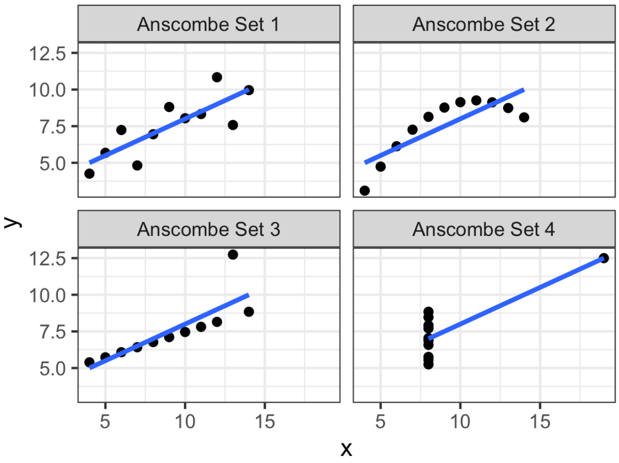
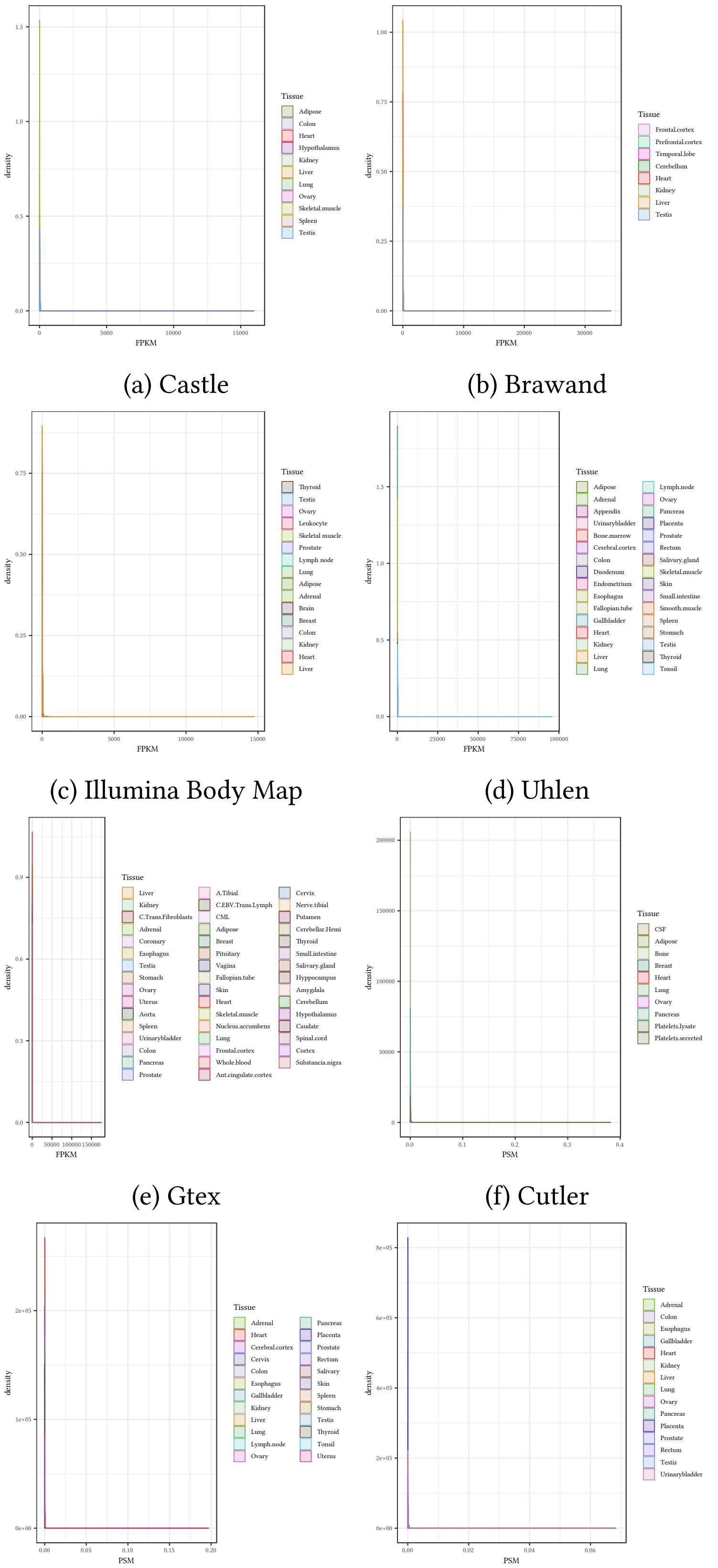


Figure B.1. **Anscombe quartet — why data should always visually checked.** All the datasets, while presenting different distributions, have equal or very similar descriptive statistic indicators; the means and variances for both x and y variables and the Pearson correlation between x and y, and their linear regressions are very similar.



(a) Castle

(b) Brawand

(c) Illumina Body Map

(d) Uhlen

(e) Gtex

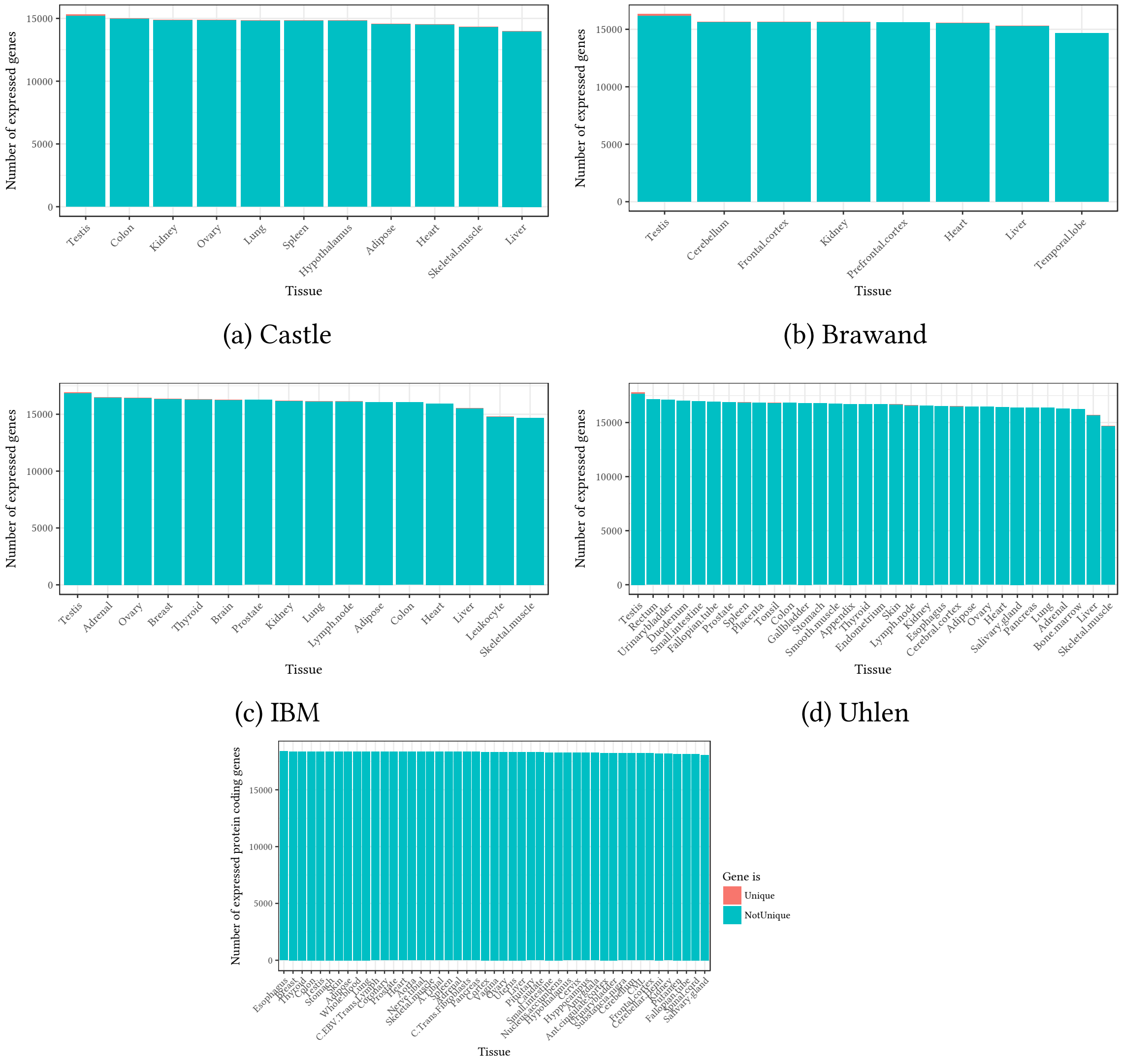
(f) Cutler

(g) Kuster

(h) Pandey

Figure B.2. Profile of expression across the transcriptome (protein coding genes only) and proteome datasets

C SUPPLEMENTARY MATERIAL FOR CHAPTER 4



(a) Castle

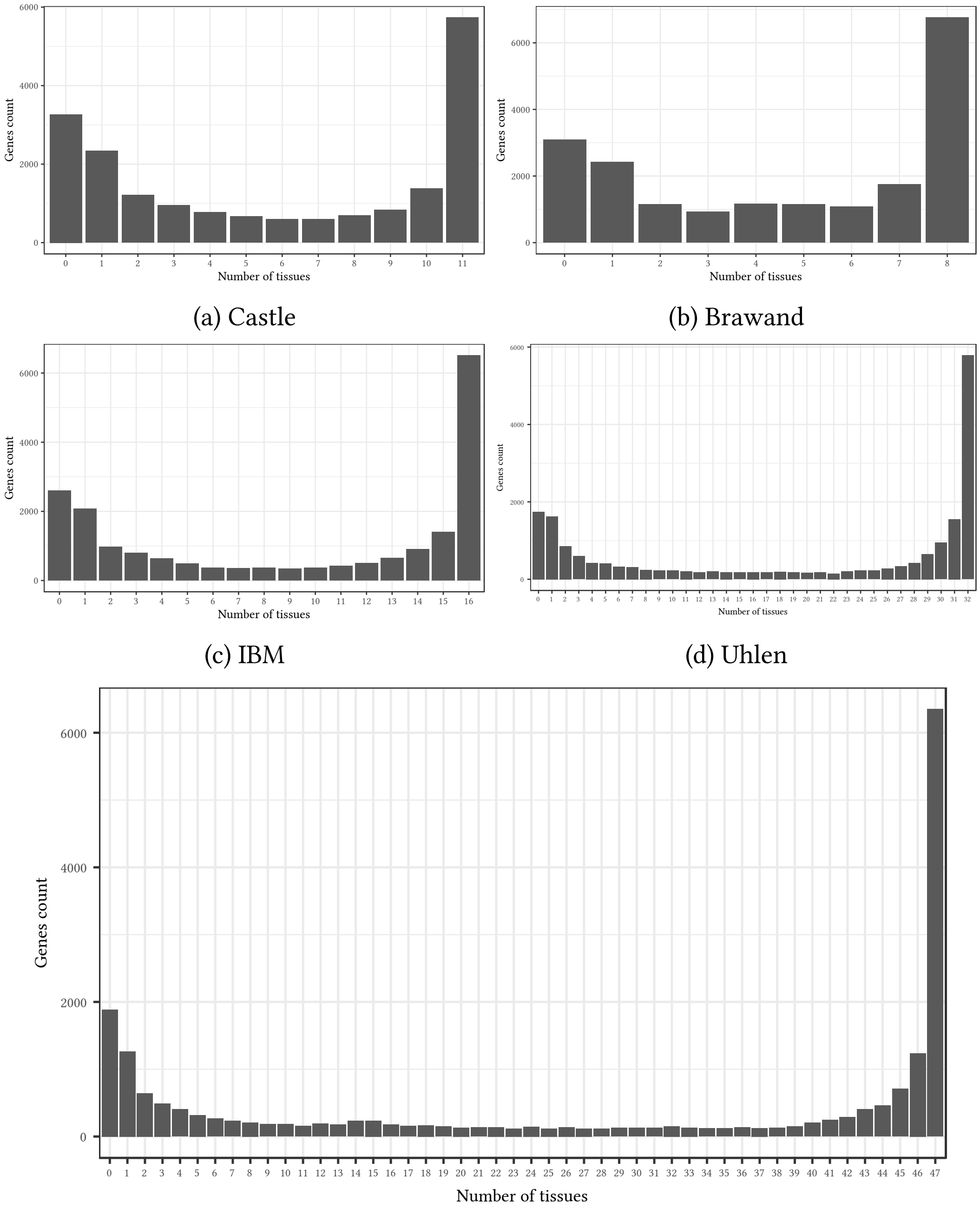
(b) Brawand

(c) IBM

(d) Uhlen

(e) Gtex

Figure C.1. Number of protein-coding genes expressed per tissue



(a) Castle

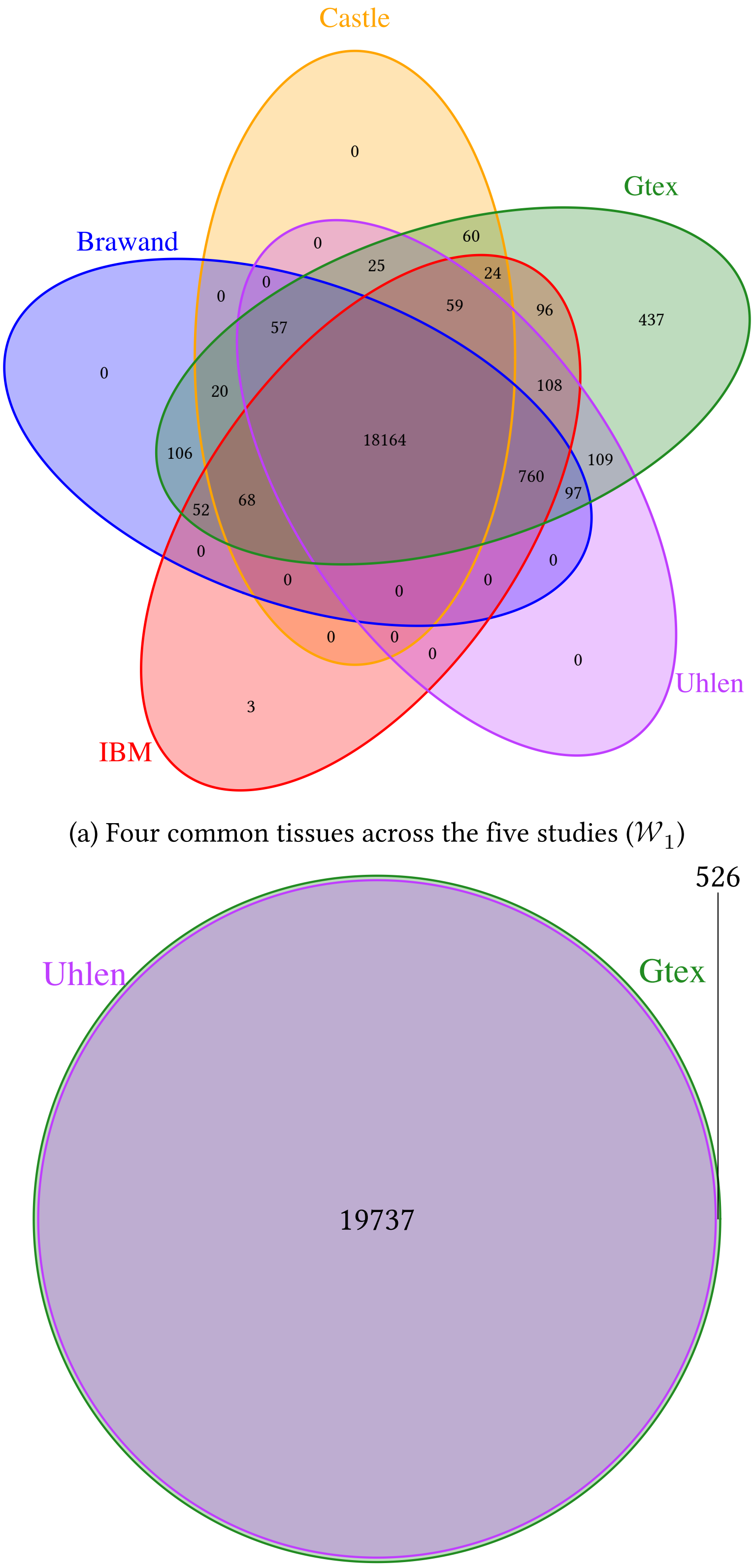
(b) Brawand

(c) IBM

(d) Uhlen

(e) Gtex

Figure C.2. Breadth of expression of the protein-coding genes expressed above 1 FPKM



(a) Four common tissues across the five studies ($\mathcal{W}_1$)

(b) Twenty-three tissues across Uhlén et al. and GTEx studies.

Figure C.3. **Unique and shared protein coding genes expressed at any level (> 0 FPKM) in $\mathcal{W}_1$ and $\mathcal{W}_2$.**

Table C.1. **Expressed protein-coding genes.**
In Ensembl 76, there are 22,469 genes that have a biotype annotated as '*protein-coding*'.

Dataset	Number of Tissues	Number of mRNAs expressed across all tissue		Number of mRNAs expressed at least once			
				4 common tissues		23 common tissues	
		>0 FPKM	≥ 1FPKM	>0 FPKM	≥ 1FPKM	>0 FPKM	≥ 1FPKM
Castle	11	19,066	15,798	18,477	13,443	—	—
Brawand	8	19,505	16,410	19,324	15,327	—	—
IBM	16	19,776	17,171	19,334	15,058	—	—
Uhlén	32	19,807	18,060	19,379	15,739	19,737	17,832
GTEx	47	20,272	18,386	20,242	16,100	20,263	18,013

Two-sample test

Welch's test [Welch, 1947], also known as the unequal variances *t*-test, is an adaptation of the Student *t*-test [Student (Gosset, 1908] and it is better fitted for groups that have different variance and sample sizes. Except in the case of (true) paired data (sampled on the same source), it gives better or at least equal results than the traditional Student *t*-test [Fagerland, 2012; Derrick et al., 2016; Delacre et al., 2017].

Student's two-sample location test (*i.e.* *t*-test) is a test where the null hypothesis is defines as the means of the two populations which have been sampled are equal. Student's *t*-test relies on the assumption that the variance of the population is also equal.

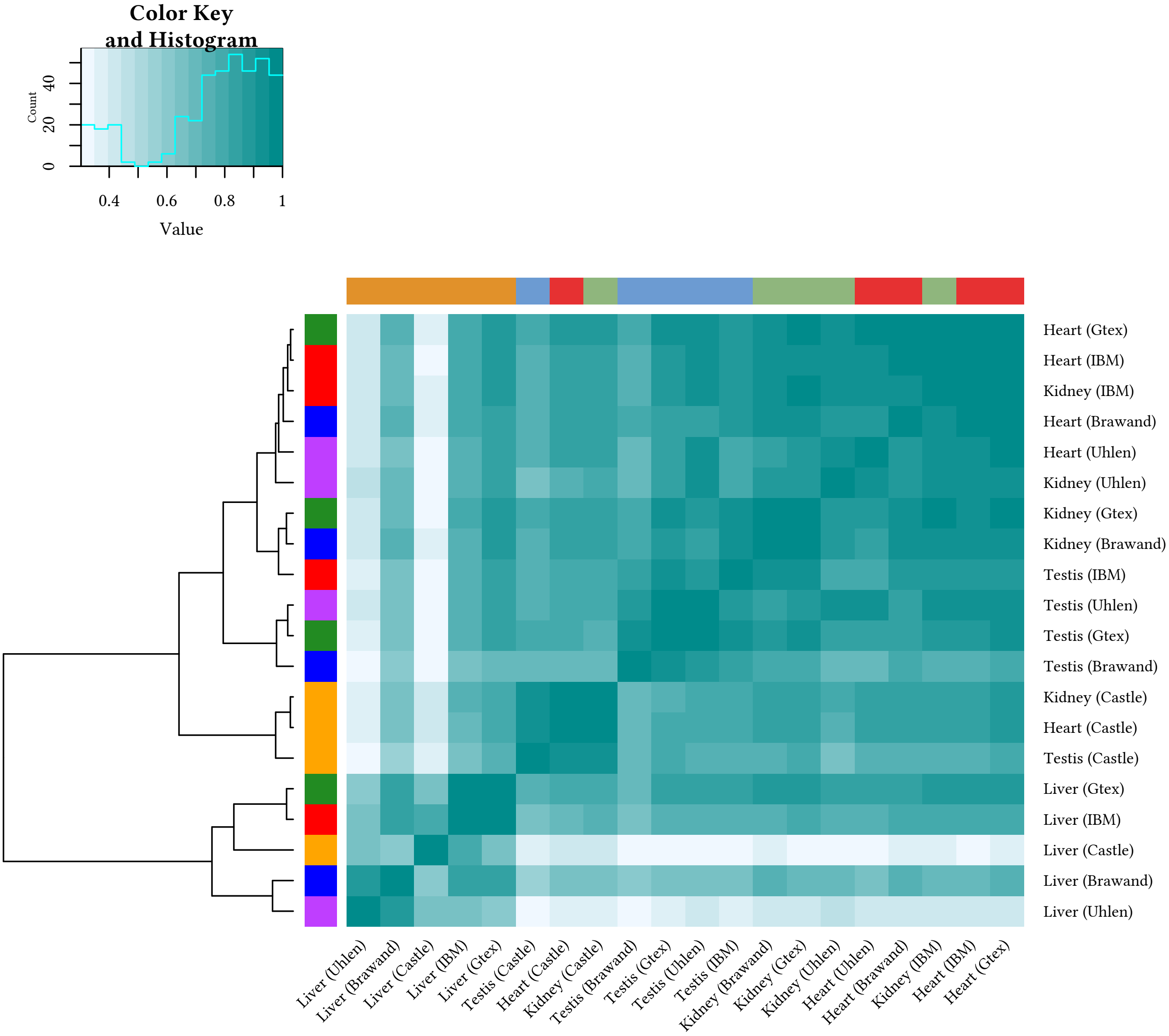


Figure C.4. **Comparison of profiles across the 5 studies for their 4 common tissues — including the 37 mitochondrial genes.**

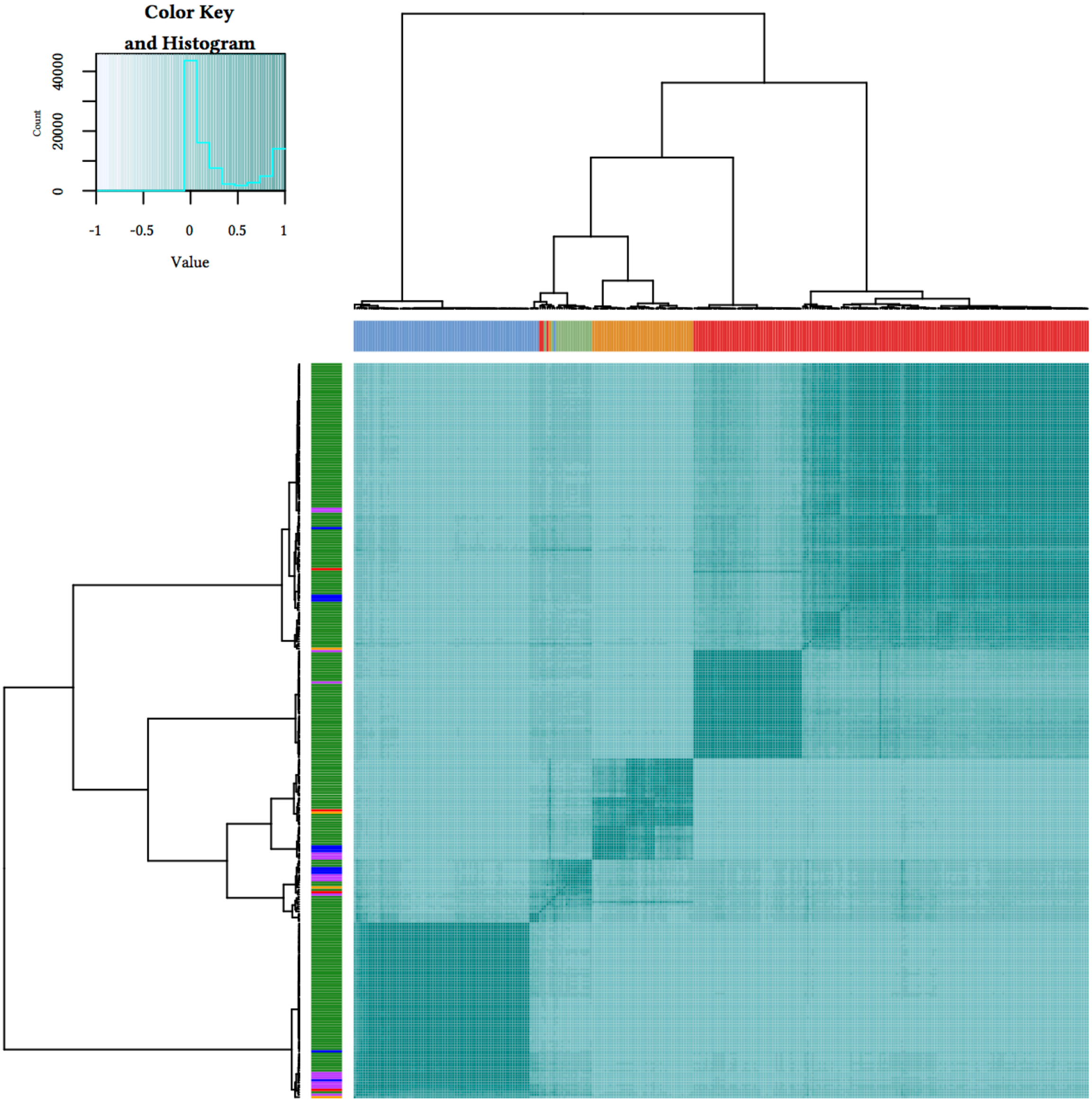


Figure C.5. **Heatmap including all the replicates of the four common tissues across the five studies.** All protein-coding genes (except the mitochondrial ones) at least expressed at 1 FPKM are included. All the samples, except from the Castle study are clustering by their tissue of origin. Remarkably, while the replicates may cluster by their study in each of the tissue groups, many pairs with higher correlations are involving replicates from different studies. Castle study is not a polyA-selected study, hence, its samples clustering may be due entirely to the effect size bias of the FPKM normalisation method.

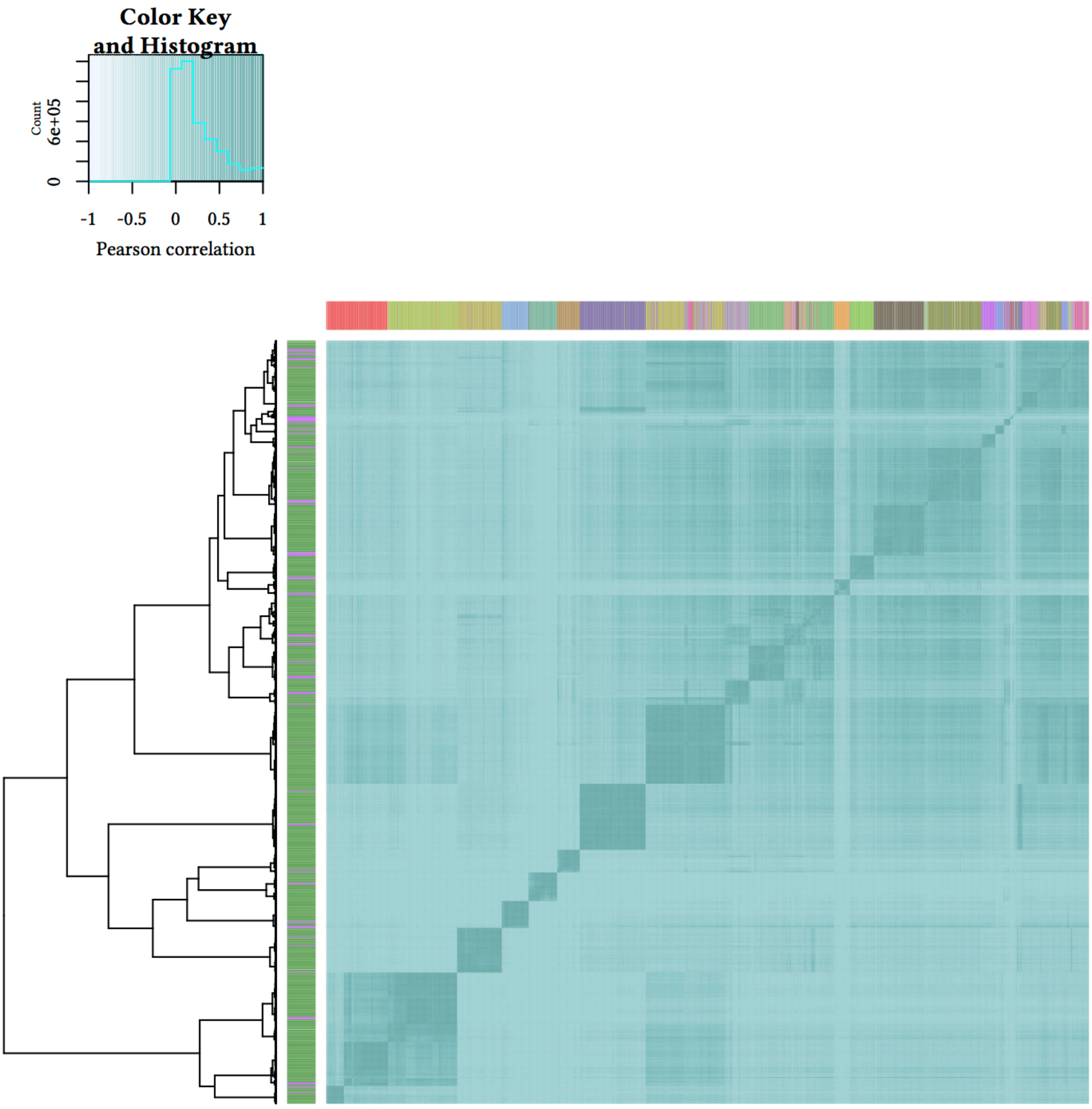
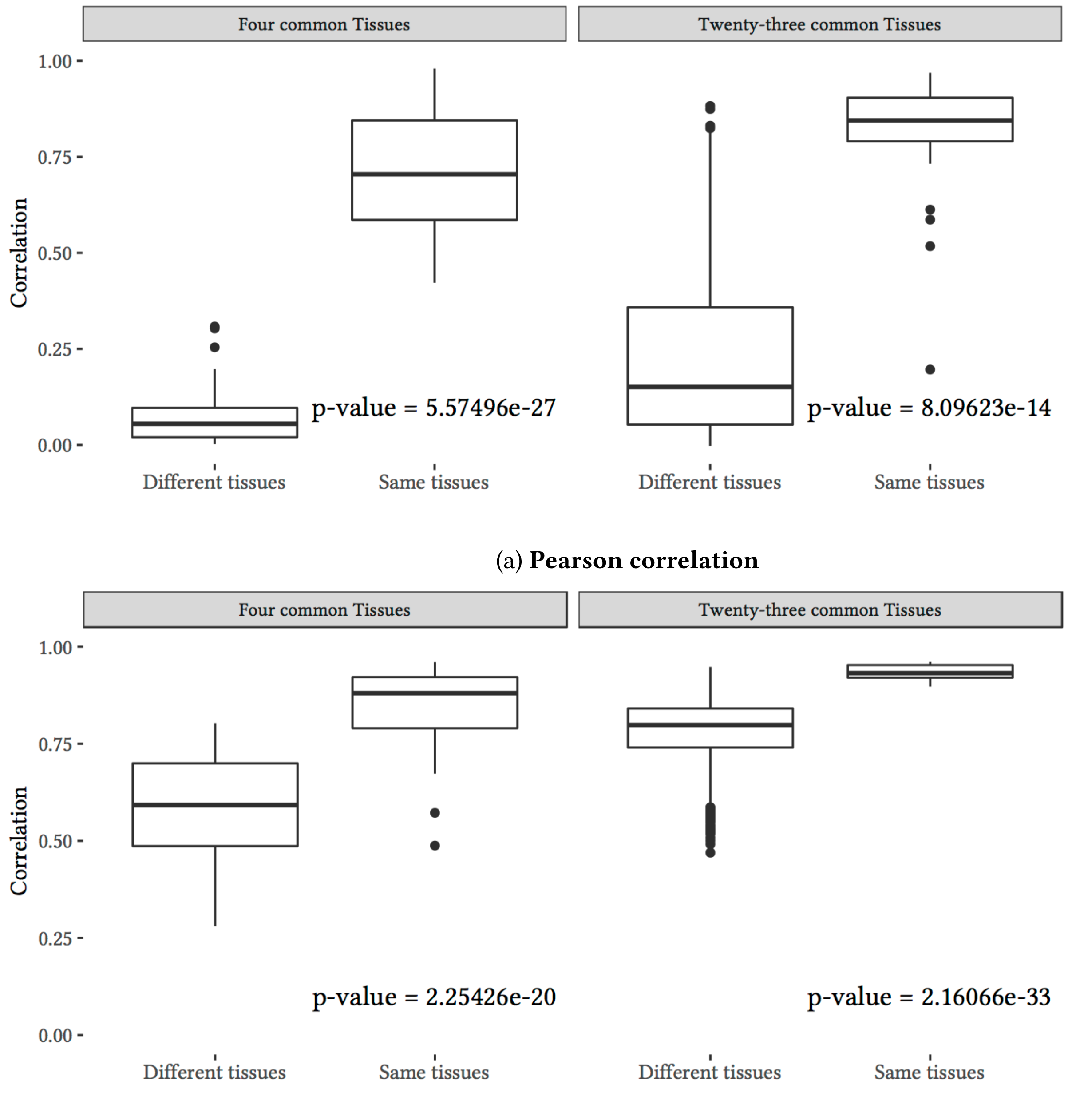


Figure C.6. **Heatmap including all the replicates of the twenty-three common tissues between Uhlén et al. and GTEx studies.** All protein-coding genes (except the mitochondrial ones) at least expressed at 1 FPKM are included. Most samples are clustering by their tissue of origin while we can observe than many single replicates may cluster less expectedly. Many small mixtures are observed; often they involve closely related tissues, *i.e. Heart* and *Skeletal muscle* or *Ovary* and *Fallopian tube*.



(a) **Pearson correlation**

(b) **Spearman correlation**

Figure C.7. **Distribution of the correlation of matched and unmatched tissues pairs for the two working sets.** The displayed p-values[a] have been computed with a Welch two-sample t-test.

[a] Thresholds above which the H_0 hypothesis is safe to be rejected.
H_0: The correlations of same tissue pairs and different tissues pairs are similar.

C.1 HIGHEST EXPRESSED GENES

Note: the cut-off used in Figure C.9 and Figure C.8 is a range (step 10) of possible values (integers) of gene expression.

Here below, the few exceptions grouped by tissue for $\mathcal{W}_1$:

Heart Uhlén-GTEx pair

Kidney Uhlén-GTEx, Castle-Uhlén and Castle-GTEx pairs,

Liver Brawand-IBM, IBM-Uhlén and IBM-Uhlén pairs;

Testis IBM-Uhlén, Brawand-GTEx, Brawand-Uhlén and Uhlén-GTEx pairs

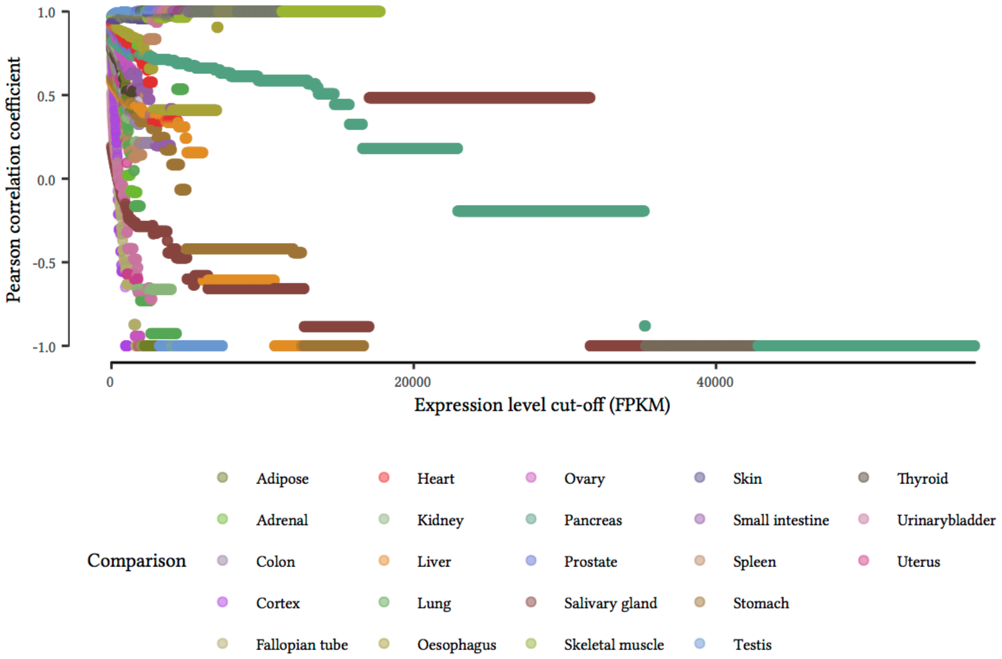
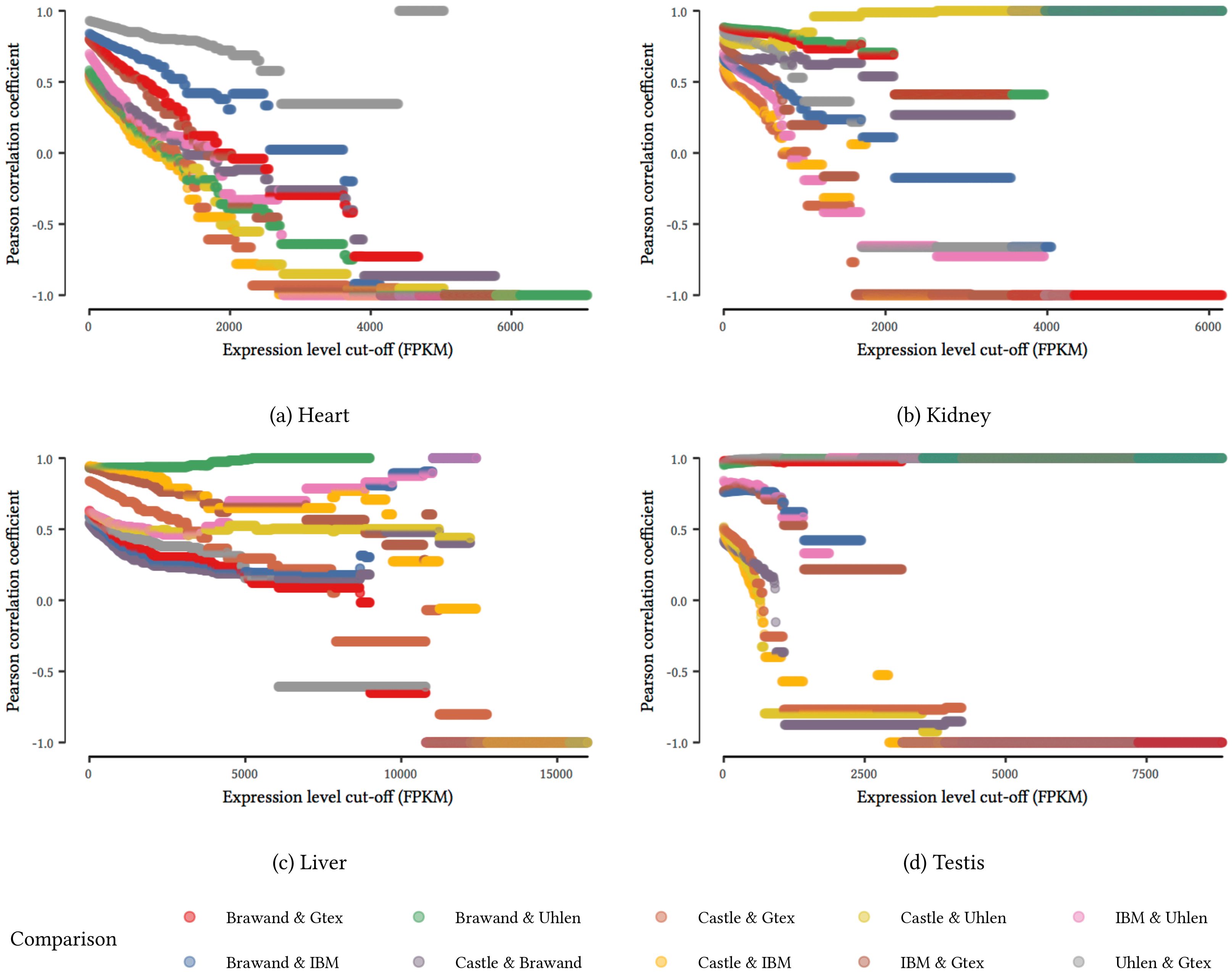


Figure C.8. **Pearson correlation coefficient trend based on the expression levels of the genes considered for each of the twenty-three common tissues between** Uhlén **and** GTEx**.** In almost every case the complete set of common expressed protein-coding genes of each tissue gives the highest correlations.



(a) Heart

(b) Kidney

(c) Liver

(d) Testis

Figure C.9. **Pearson correlation coefficient trends based on the expression levels of the genes considered for each of $\mathcal{W}_1$'s tissues.**

W_2's results should be interpreted more carefully as there are only two studies involved and there are no actual means to distinguish between an artefact or a true biological reason that may drive the higher correlations.

Both for Figures C.8 and C.9, it seems that a couple of TREPs are perfectly anticorrelated for subsets of highest expressed genes, These are most likely mathematical artefacts. The correlation calculations are involve very few genes and as such the correlations are more sensitive to any change.

As very few genes are involved, the slightest changes in their respective order of magnitude may imply reversed trends.

Table C.2. Example of gene subsets for a two studies (A and B) for a tissue

	$Gene_a$	$Gene_b$	$Gene_c$	$Gene_d$	$Gene_e$
Study A, TREP T_1 (T_1^{StudyA})	1000	2000	3000	4000	5000
Study B, TREP T_1 (T_1^{StudyB})	500	2800	6000	5000	4000

For example, if we consider T_1^{StudyA} and T_1^{StudyB} for a set of genes $a, \ldots, e$ (see Table C.2) and a cut-off at 3,000 FPKMs, then the correlation will only involve $Gene_c$, $Gene_d$ and $Gene_e$. Thus,
while $corr_{T_1^{StudyA},T_1^{StudyB}}(c, d, e) = -1$,
$corr_{T_1^{StudyA},T_1^{StudyB}}(a, b, c, d, e) = 0.6836$

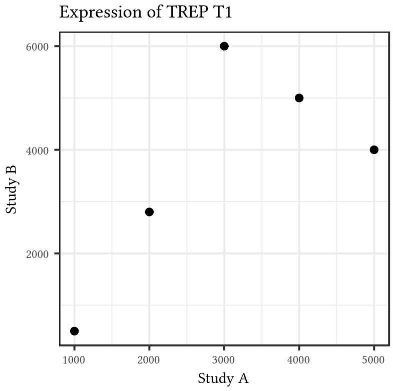


Figure C.10. Example on how correlation may change to cut-offs

C.1.1 *Overlap of the top high expressed genes between the five datasets*

Figure C.11 shows the ratios of the number of common protein-coding genes for a given amount of highest expressed protein-coding genes to that very number across the studies and for each tissue. Among these (cumulative) proportions, a few present very high value for a minimal subset of genes (below 10 FPKM across all the tissues) which then drop

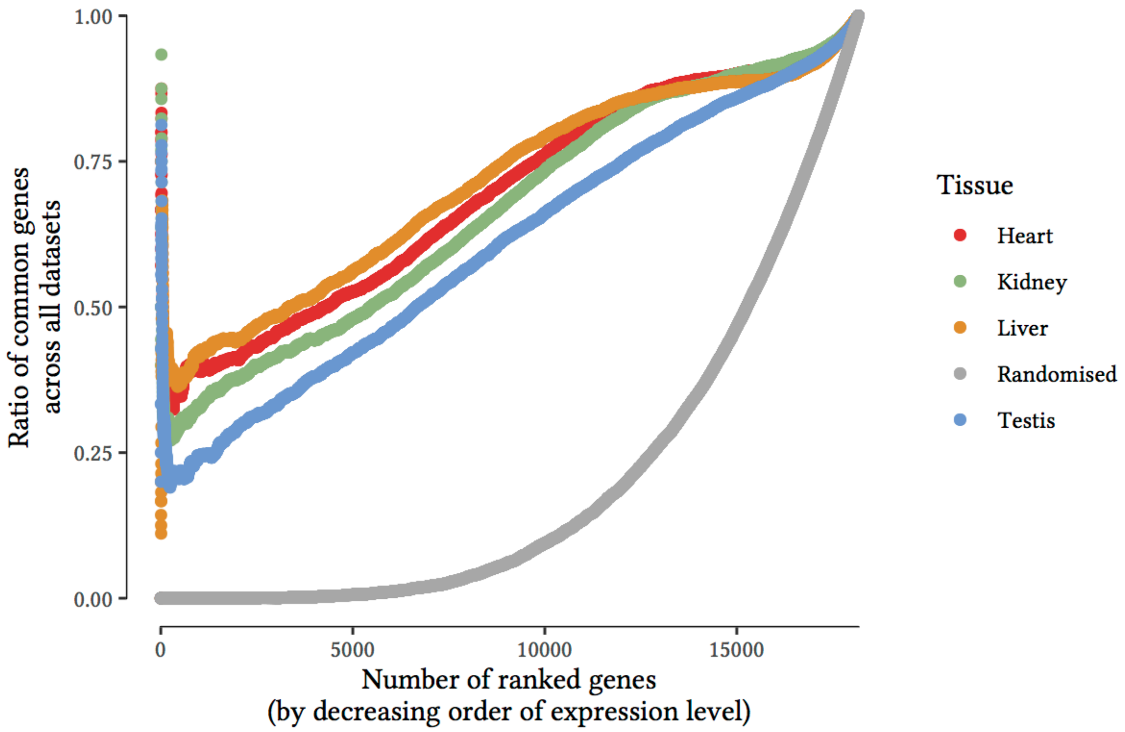


Figure C.11. **Cumulative shared set of genes ranked by their decreasing order of expression across the five studies.** Apart from a very small subset for the highest expressed protein coding genes, the overlap of the gene ranks across the 5 studies is rather small. The grey line presents the evolution of the ratios for the randomly permuted data which highlights that there is an underlying structure.

dramatically to finally increase slowly to reach the expected ratio of 1 FPKM.

Figure C.12 presents the same kind of ratios, however, limited to Uhlén and GTEx only. Here as well, aside from the expected perfect ratio for the complete set of protein-coding genes, only a tiny subset of the highest expressed genes produce high rates of highly expressed common genes to the number of considered ranked genes. These results are quite unsurprising as they involve only two studies. In addition to increasing the number of overlaps probabilistically, these two studies are also the two most recent ones and comprise a higher number of replicates per tissues; thus the measurements for each gene in each condition are likely more robust.

Comparing the calculated ratios between the real (colour) and the randomly permuted data (grey) on Figures C.11 and C.12 plainly show that there are common (biological) structures that are nonfortuitous across studies.

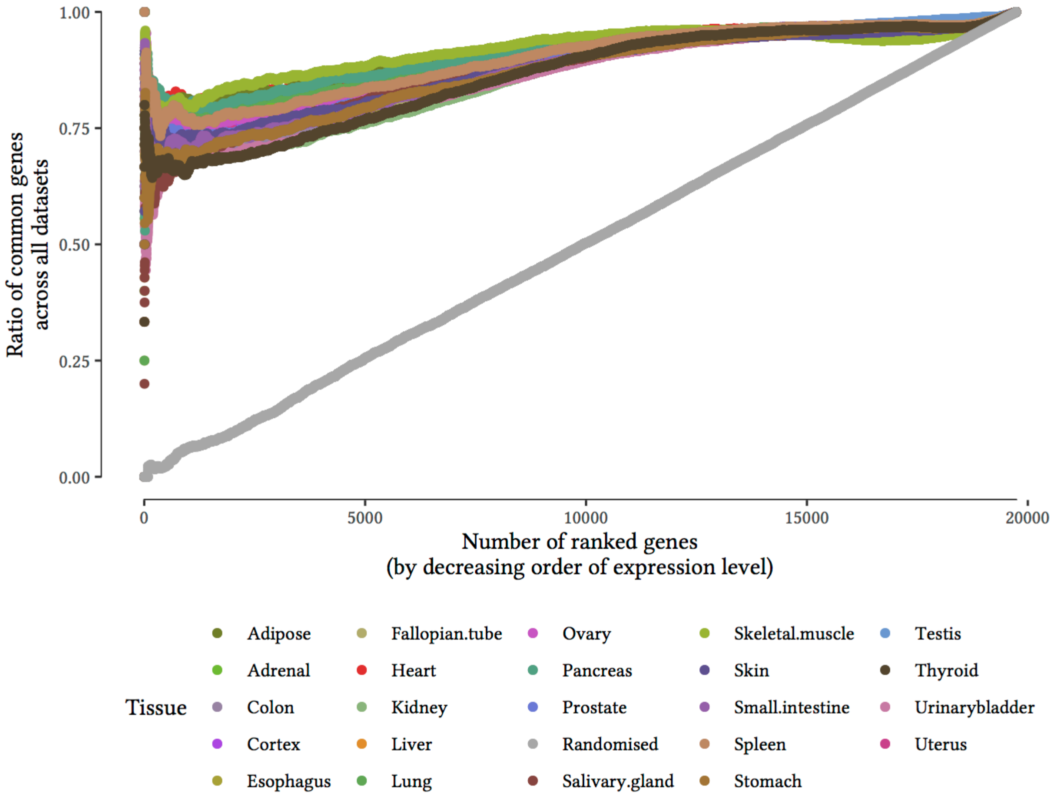


Figure C.12. **Cumulative shared set of genes, ranked by their decreasing order of expression, between Uhlén et al. and GTEx studies.** The highest expressed genes present greater ratios than when the 5 studies are considered (see Figure C.11). The fewer number of studies considered, which, additionally to be the most recent studies, are the ones to comprise greater number of biological replicates per tissues may be the sole reasons for the improved results.

C.2 MOST VARIABLE GENES

C.2.1 *Validation of the association of the most variable genes with the highest correlations*

After ordering the genes by decreasing order of their coefficient of variation within each of the datasets comprised in $\mathcal{W}_1$, I have calculated the size of overlap for each rank (*i.e.* from 1 to 12,268) between the five datasets. To help with the interpretation, I finally divide the previous number by the rank.

Figure C.19 and Figure C.20 present the result for $\mathcal{W}_1$ and $\mathcal{W}_2$. Many of the most variable genes are commonly present in the top tier of the five studies, though they have different individual rank. There is a strong growth for about the first 1,250 genes that then settles

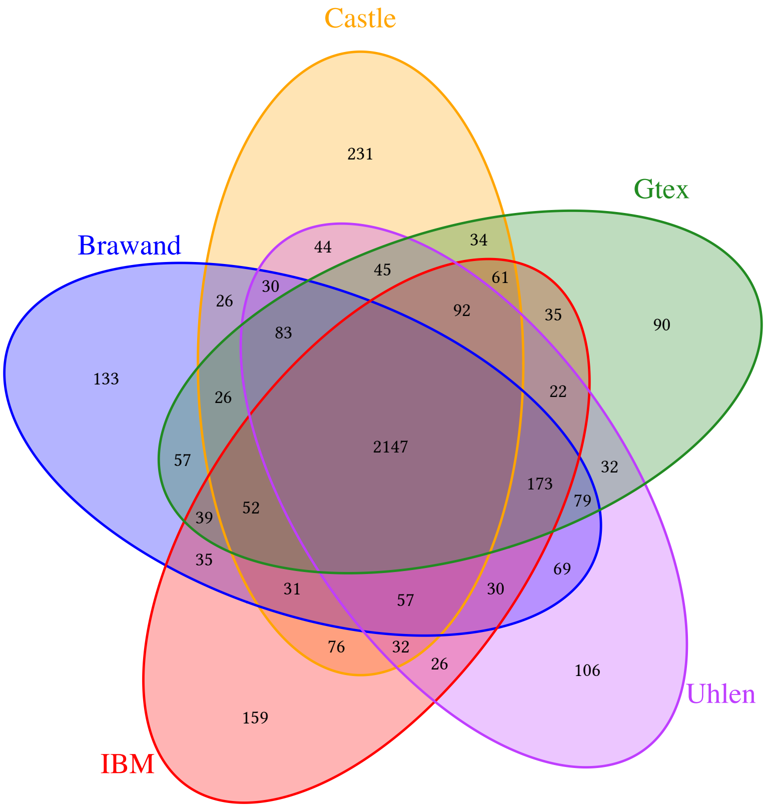


Figure C.13. **Overlap of the most variable genes across the five studies for the set of the four common tissues.** In each study, I rank the protein-coding genes in decreasing order. This Venn diagram presents the shared and unique protein-coding genes in the top quarter of the most variable genes.

a plateau which increases toward the final ratio (1). Using the first quarter of the most variable genes as a cut-off appears to be an acceptable threshold as it comprises the initial growth and part of the plateau.

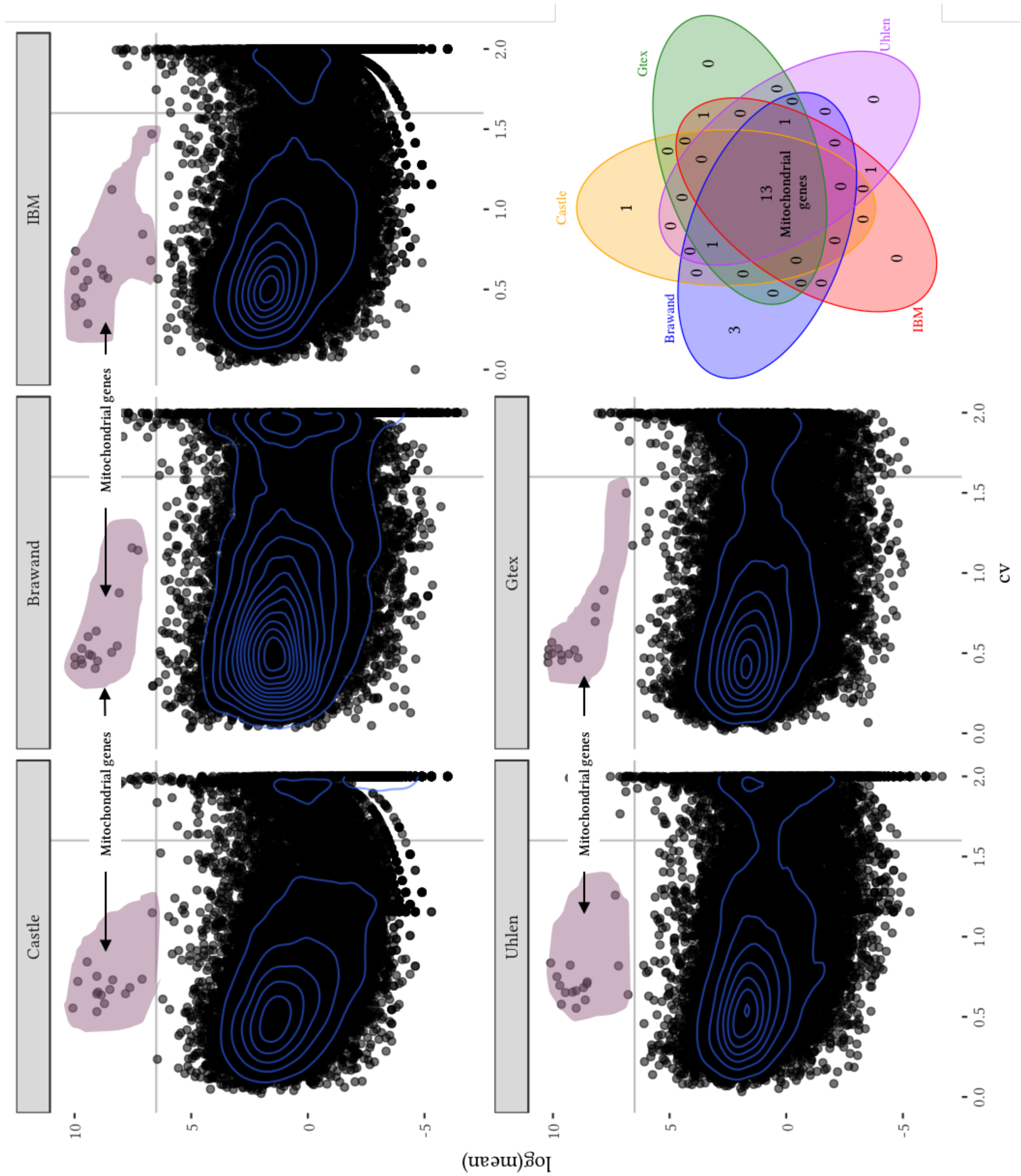


Figure C.14. **Mean expression of genes compared to their coefficient of variation.**

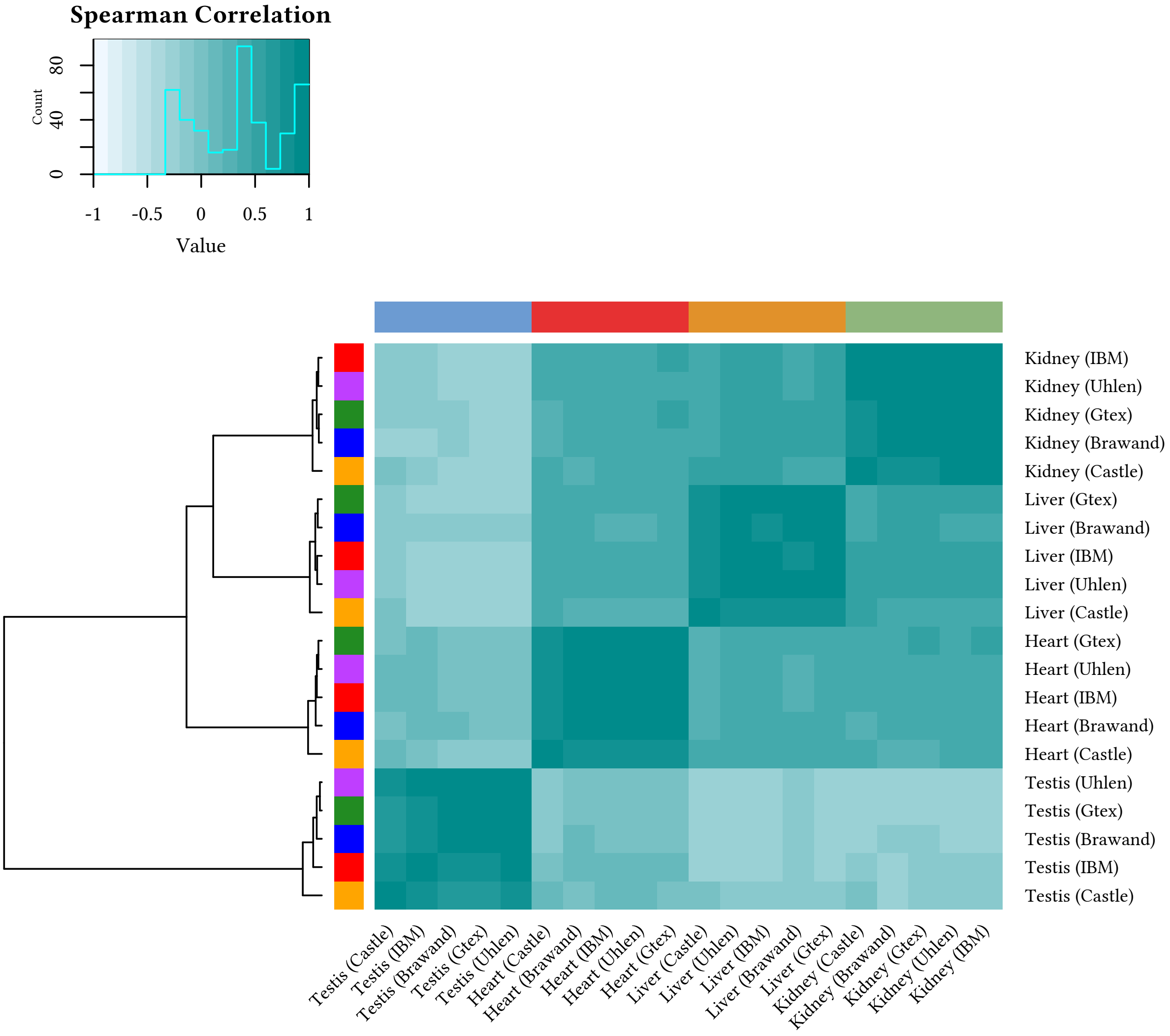


Figure C.15. **Clustering of the four common tissues across the five studies for the most common variable genes.** The samples cluster by tissue of origin rather than by original study. Each cluster of tissue presents a different hierarchy of study: for *Kidney*, Uhlén sample is closer to the IBM sample, while for *Testis*, Uhlén sample is closer to the GTEx one.

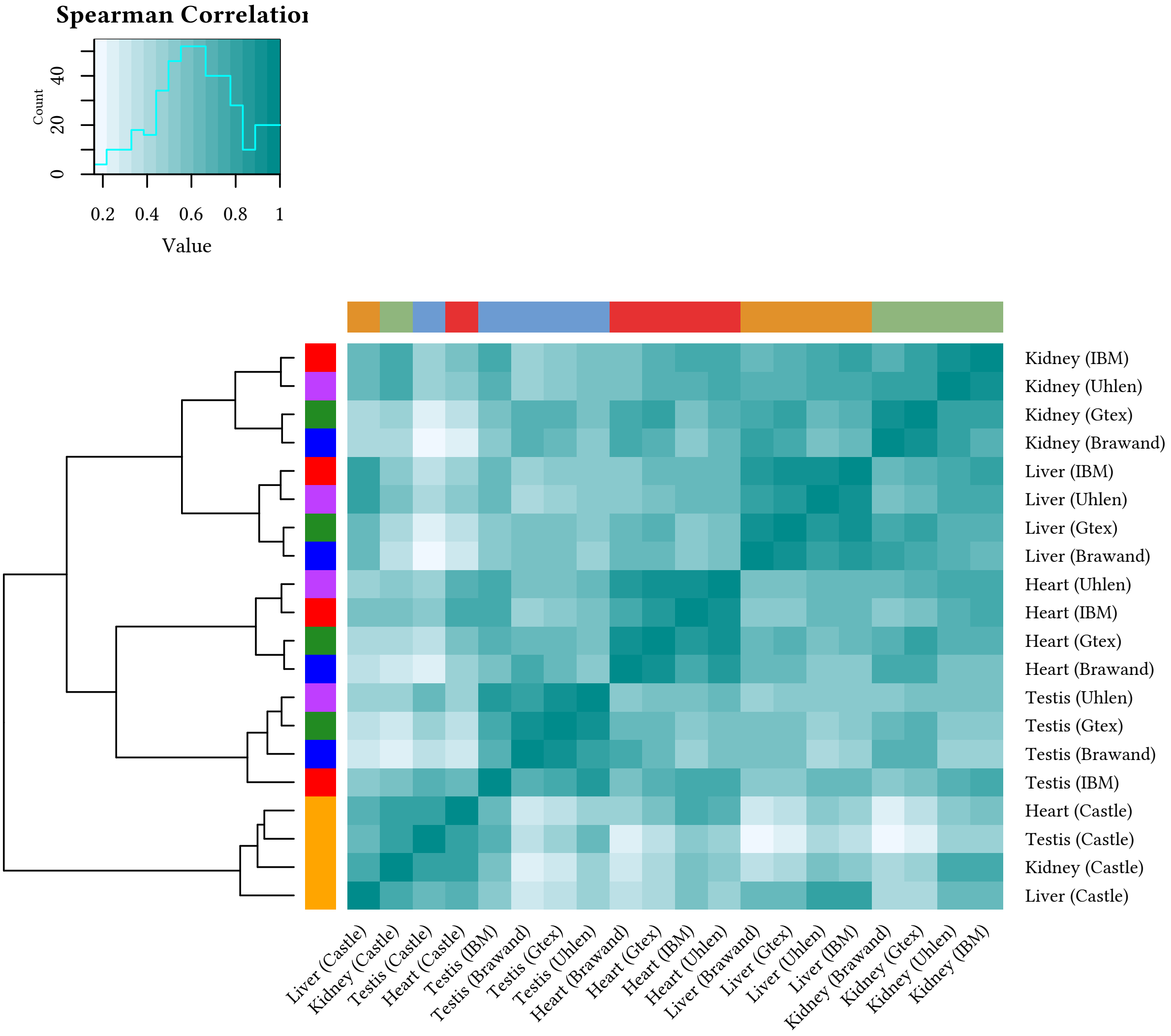


Figure C.16. **Clustering of the four common tissues across the five studies (excluding the most variable genes).** Apart from the Castle samples, the samples cluster by tissue of origin rather than by original study. (Note that Pearson correlations give stronger clustering results towards the biological origin of the TREPs.)

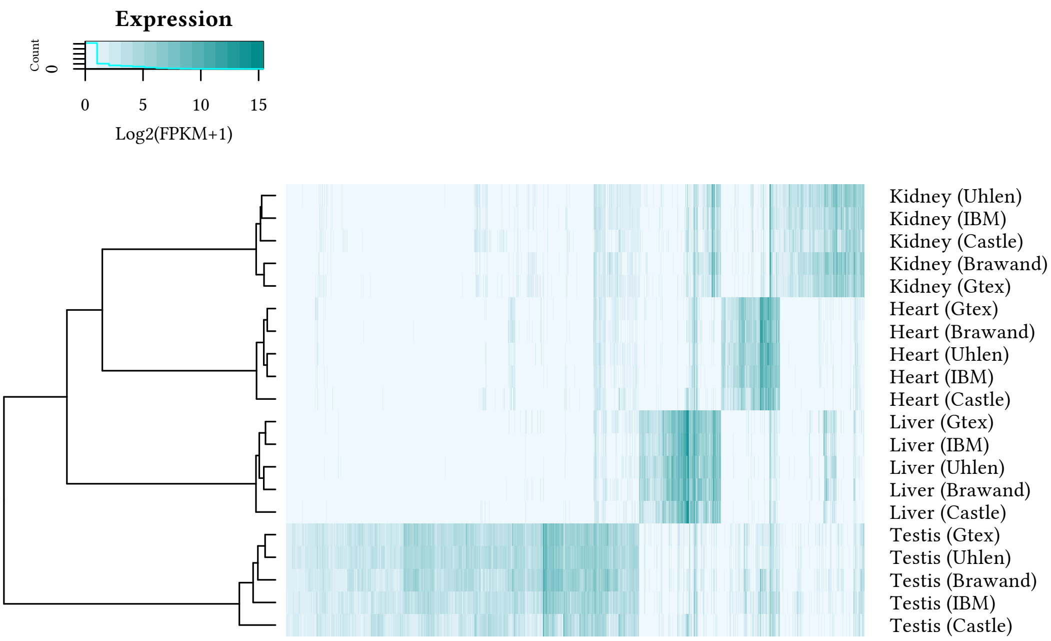


Figure C.17. **Expression of the most common variable genes.**

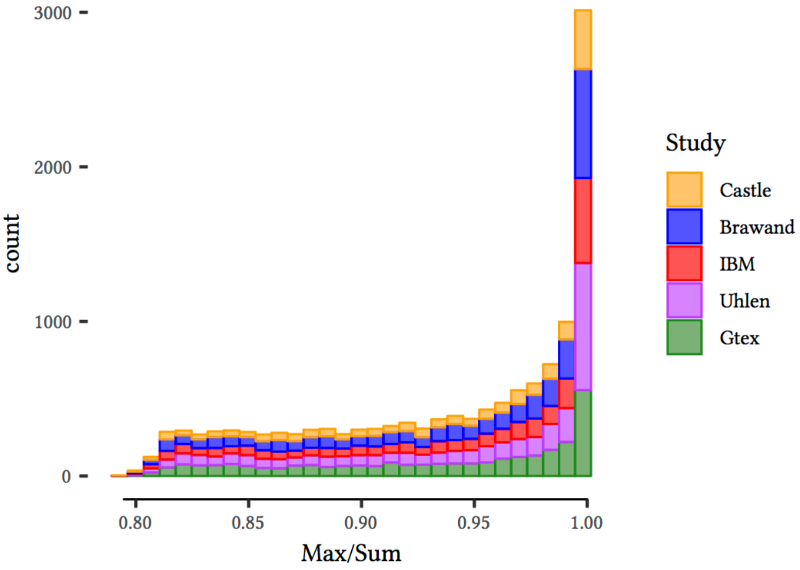


Figure C.18. **Ratio of Maximum of expression/Sum of expression for the most variable genes (cv≥1.5) in $\mathcal{W}_1$ that are expressed at least in two different tissues at 1 FPKM.** The lowest ratio is above 0.79 and the highest ratio is close to 1. This range shows that the most variable genes are expressed in one tissue more specifically than the three others as the tissue where they are the highest expressed accounts for more than 79% of the sum of expression across the four tissues.

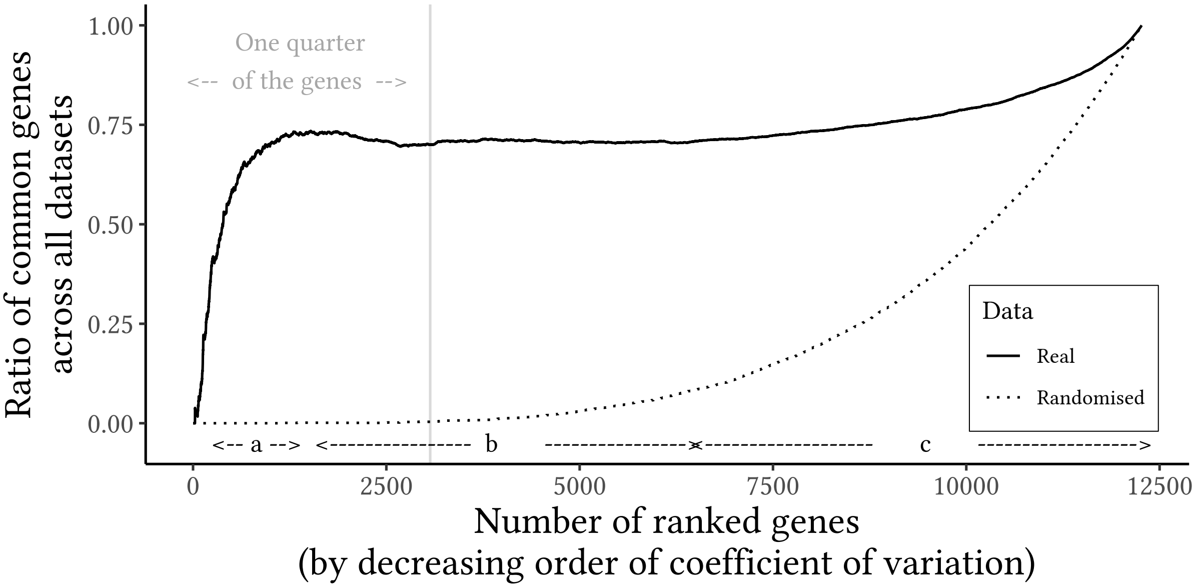


Figure C.19. **Intersection size course of $\mathcal{W}_1$ genes (based on their coefficient of variation rank in each of the five studies).** There are three main parts. There is an initial strong growth (a) which then settles a plateau (b). Eventually, the ratio increases slowly again until reaching the expected ratio of 1 once all the genes from $\mathcal{W}_1$ are included (c). The first quarter of the genes covers (a) and a part of (b). Apart from (a), the overlap of shared genes between the five datasets when ranked on their coefficient of variation is above 70%. The sigmoid curve (dashed line) is based on randomised data where permutations break the original order of the genes. (Within each dataset, all the gene expression levels are permuted within each tissue, *i.e.* the overall pattern of expression of each tissue is conserved. This operation is performed 10,000 times. The dashed line is a summary of all these permutations). There is a distinct dissimilarity between the real and the randomised data.

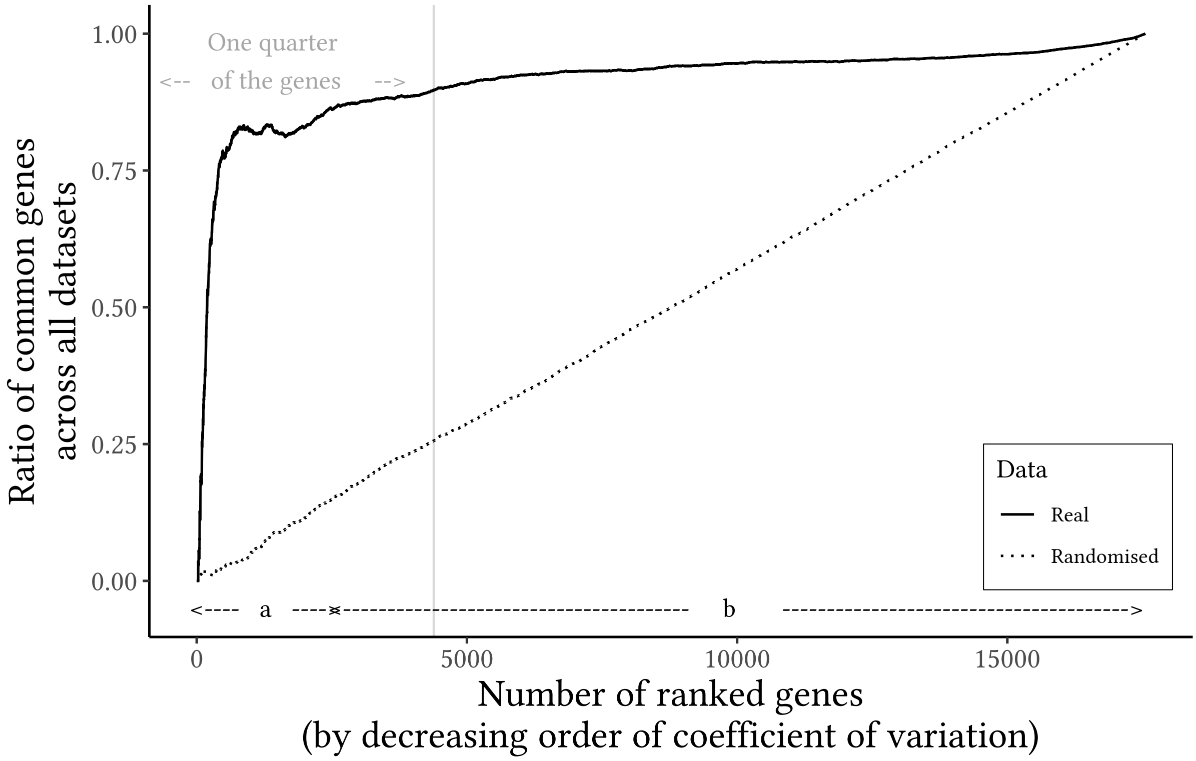
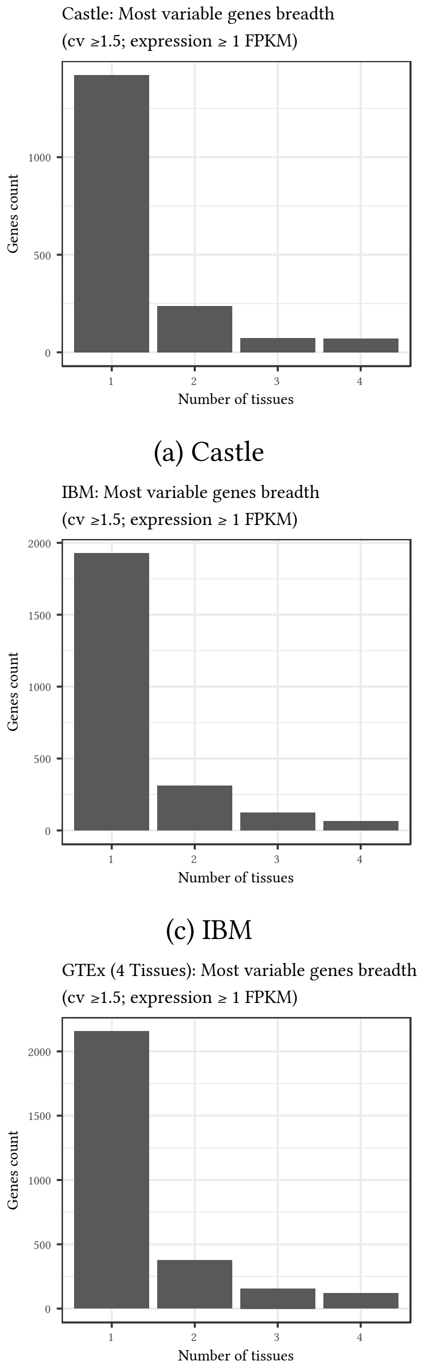


Figure C.20. **Intersection size course of $\mathcal{W}_2$ genes (based on their coefficient of variation rank in each of the two studies).** Globally, there are two parts: one initial strong growth (a) and a second part (b) where the curve shifts shallowly towards the expected final ratio of 1. While the number of genes involved in $\mathcal{W}_2$ is higher than in $\mathcal{W}_1$, the ratio of common genes that are the twentieth most variable in each study is above 75%. There are three main reasons that may explain this improved result compared to Figure C.19.

- $\mathcal{W}_2$ involves a smaller degree (*i.e.* number of studies) than $\mathcal{W}_1$ (respectively 2 and 5). Hence, bigger intersection sizes are easier to occur.
- As previously mentioned, GTEx and Uhlén studies provide probably more accurate TREPs than the other studies (see section 4.2 on page 98).
- The greater number of tissues induces a wider range of coefficients of variation, which allows picking up genes with more subtle variations.



(a) Castle

(c) IBM

(e) GTEx

Figure C.21. **Breadth of expression (≥1 FPKM) for the most variable mRNAs (cv≥1.5) across $\mathcal{W}_1$.** Most of these mRNAs are expressed only at 1 FPKM or above.

C.3 TISSUE SPECIFIC (TS) GENES

C.3.1 *Hampel method*

The Hampel method allows detecting outliers and it relies on the median and the MAD (median absolute deviation) as robust estimate of the location and spread (instead of the more commonly used mean and standard deviation). [Hampel, 1971; Hampel, 1974]

Algorithm 1: Hampel method

```
Data: Expression matrix; Genes as rows and conditions (tissues) as columns
Input: threshold: numeric
Input: bool: boolean
Result: Indicates if a gene presents an atypical (outlier) expression for any condition
        (as a boolean or a numeric ratio)

foreach Gene g (i.e. row) of the input matrix do
    med=compute median(g);
    /* compute the M.A.D. (median absolute deviation)                    */
    mad=median(absolute(g-med));
    if !bool then
        /* Return boolean answer                                         */
        newg=absolute(g-med) > threshold*mad;
    else
        /* Return ratios that can be later sorted                        */
        newg=(absolute(g-med))/mad;
    end
    return(newg);
end
```

C.3.1.1 *Median*

Median can be found by listing all values from smallest to greatest.
If the number of values is odd, the median is the middle value.
If the number of values is even, the median is the mean of the two middle values.

C.3.1.2 *Median absolute deviation (M.A.D.)*

For a give variable X, M.A.D. is defined as follow:

$$MAD = median(|X_i - median(X)|) \tag{M.A.D.}$$

C.3.2 *List of the tissues available in TiGER*

Thirty tissues (or equivalent) are available through this database: *Bladder*, *Blood*, *Bone*, *Bone marrow*, *Brain*, *Cervix*, *Colon*, *Eye*, *Heart*, *Kidney*, *Larynx*, *Liver*, *Lung*, *Lymph node*, *Mammary gland*, *Muscle*, *Ovary*, *Pancreas*, *Peripheral nervous system*, *Placenta*, *Prostate*, *Skin*, *Small intestine*, *Soft tissue*, *Spleen*, *Stomach*, *Testis*, *Thymus*, *Tongue* and *Uterus*.

C.3.3 *Uhlén categories*

C.4 LIST OF PUBLICATIONS BASED ON RNA-SEQ AND COVERING AT LEAST PARTIALLY ITS ROBUSTNESS

- SEQC/MAQC-III Consortium (2014). ‘A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium’. *Nat. Biotechnol.* 32 (9), pp. 903–914
- A. Santos et al. (2015). ‘Comprehensive comparison of large-scale tissue expression datasets’. *PeerJ* 3, e1054
- P. H. Sudmant et al. (2015). ‘Meta-analysis of RNA-seq expression data across species, tissues and studies’. *Genome Biol.* 16, p. 287
- F. Danielsson et al. (2015). ‘Assessing the consistency of public human tissue RNA-seq data sets’. *Briefings Bioinf.* 16 (6), pp. 941–949
- M. Uhlén, B. M. Hallström, et al. (2016). ‘Transcriptomics resources of human tissues and organs’. *Mol. Syst. Biol.* 12 (4)
- L. Peixoto et al. (2015). ‘How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets’. *Nucleic Acids Res.* 43 (16), pp. 7664–7674
- Q. Wang et al. (2017). ‘Enabling cross-study analysis of RNA-Sequencing data’. *bioRxiv* (110734)

Table C.3. Uhlén et al. gene categories for all genes (*i.e.* unrestricted to protein-coding genes)

Ensembl 76 (62,757 gene definitions)		Not detected	Not expressed at 1 FPKM cut-off	Mixed expression		Ubiquitous expression		Group Enhanced	Tissue Enhanced	Tissue Enriched
				Low (< 10 FPKM)	High (≥ 10 FPKM)	Low (< 10 FPKM)	High (≥ 10 FPKM)			
Whole dataset	Castle	18,836	16,258	19,079	1,203	1,456	703	77	8,319	3,896
	Brawand	18,278	20,173	15,254	2,057	1,873	977	0	6,180	5,442
	IBM	14,494	20,858	16,633	1,582	1,194	926	733	10,042	4,453
	Uhlen	17,345	16,548	15,372	1,351	467	419	4,615	10,644	5,498
	Gtex	5,755	25,138	17,172	1,464	775	713	7,164	10,032	5,117
	Consensus	5,747	4,231	4,121	230	33	166	0	1,073	531 [518]
Common 4 tissues Working datasets	Castle	43,921	3,267	14,850	1,735	3,267	1,181	—	—	4,645
	Brawand	44,479	3,193	13,975	2,541	3,193	1,282	—	—	7,002
	IBM	48,263	3,262	13,672	2,160	3,262	1,299	—	—	5,242
	Uhlen	45,412	3,146	14,332	2,546	3,146	1,213	—	—	7,665
	Gtex	57,002	4,516	16,652	2,771	4,516	1,459	—	—	8,155
	Consensus	9,655	557	4,366	675	557	448	—	—	1,960
Common 23 tissues Working datasets	Uhlen	17,345	27,575	14,981	1,427	611	440	2,203	11,252	5,678
	Gtex	5,755	38,988	16,982	1,909	2,122	1,021	1,746	11,236	5,971
	Consensus	5,755	27,149	12,250	973	433	430	797	8,030	4,281

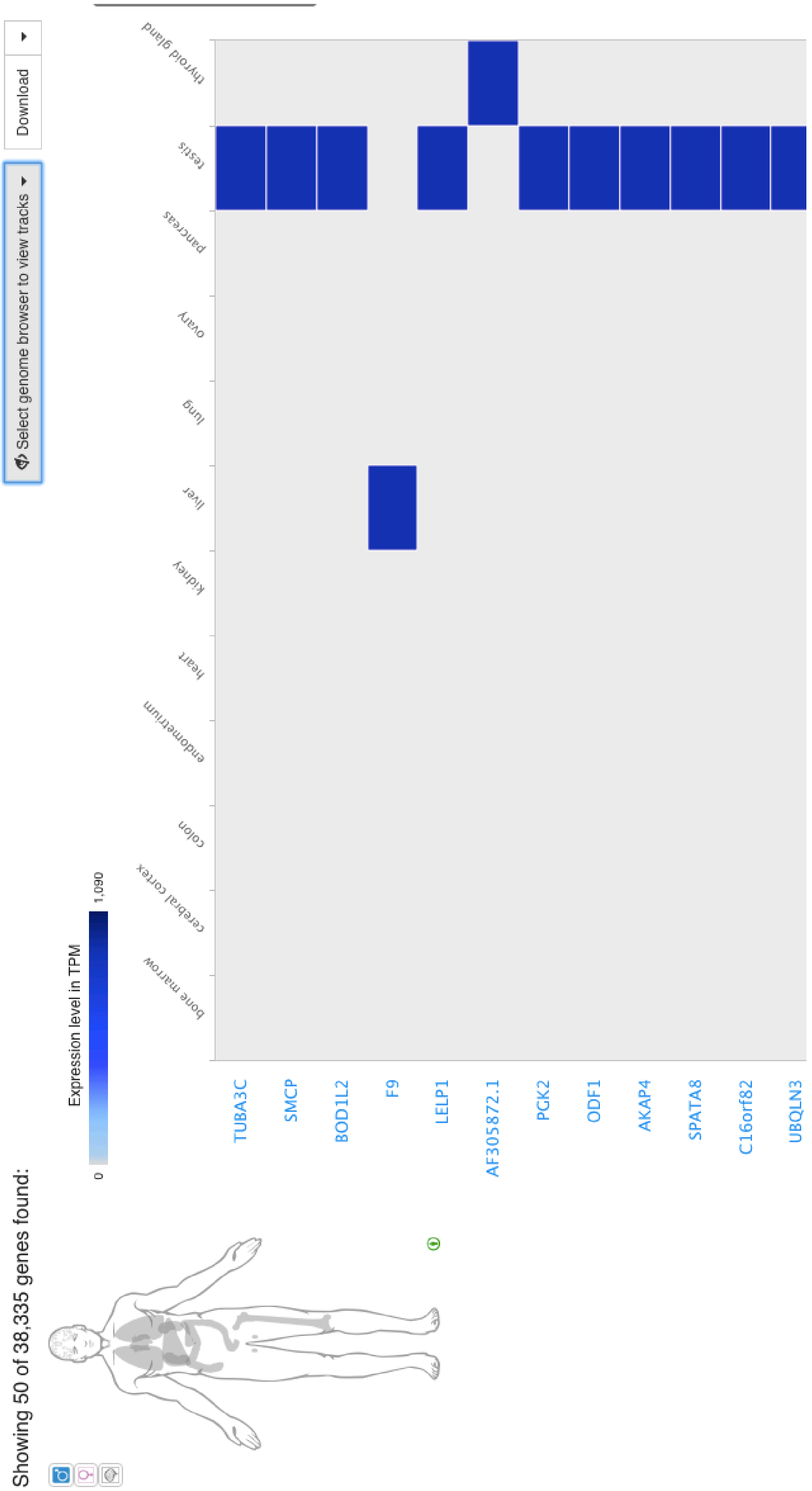
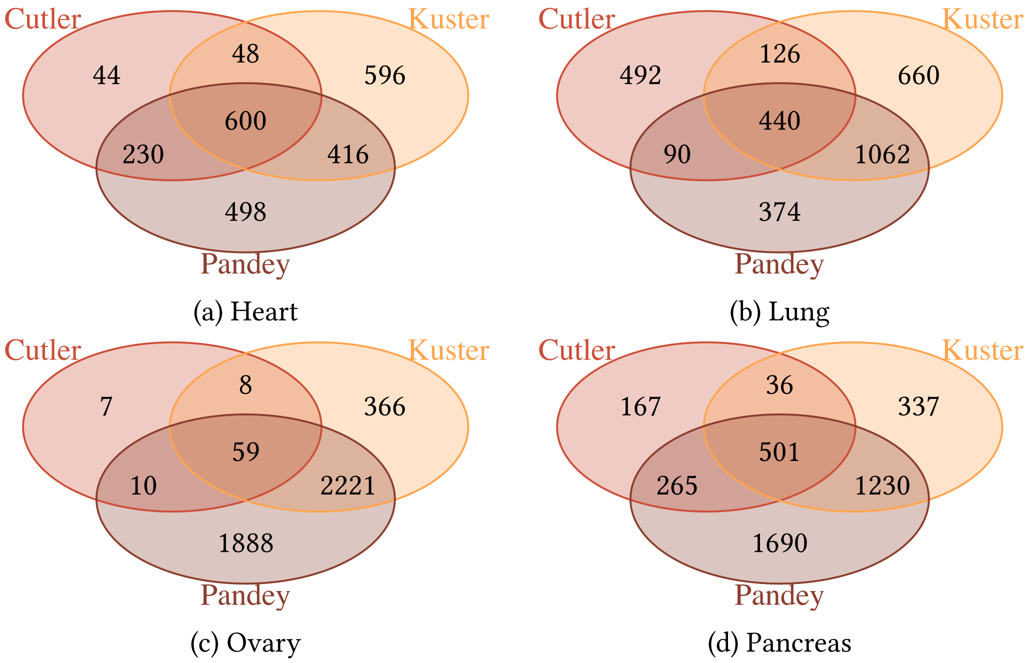


Figure C.22. **Most specific genes highlighted in EBI gene expressio atlas.**

D | SUPPLEMENTARY MATERIAL FOR CHAPTER 5



(a) Heart

(b) Lung

(c) Ovary

(d) Pancreas

Figure D.1. **Unique and shared proteins across the proteomic studies**

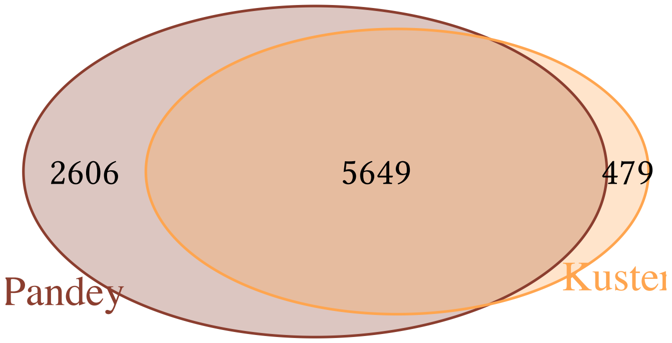
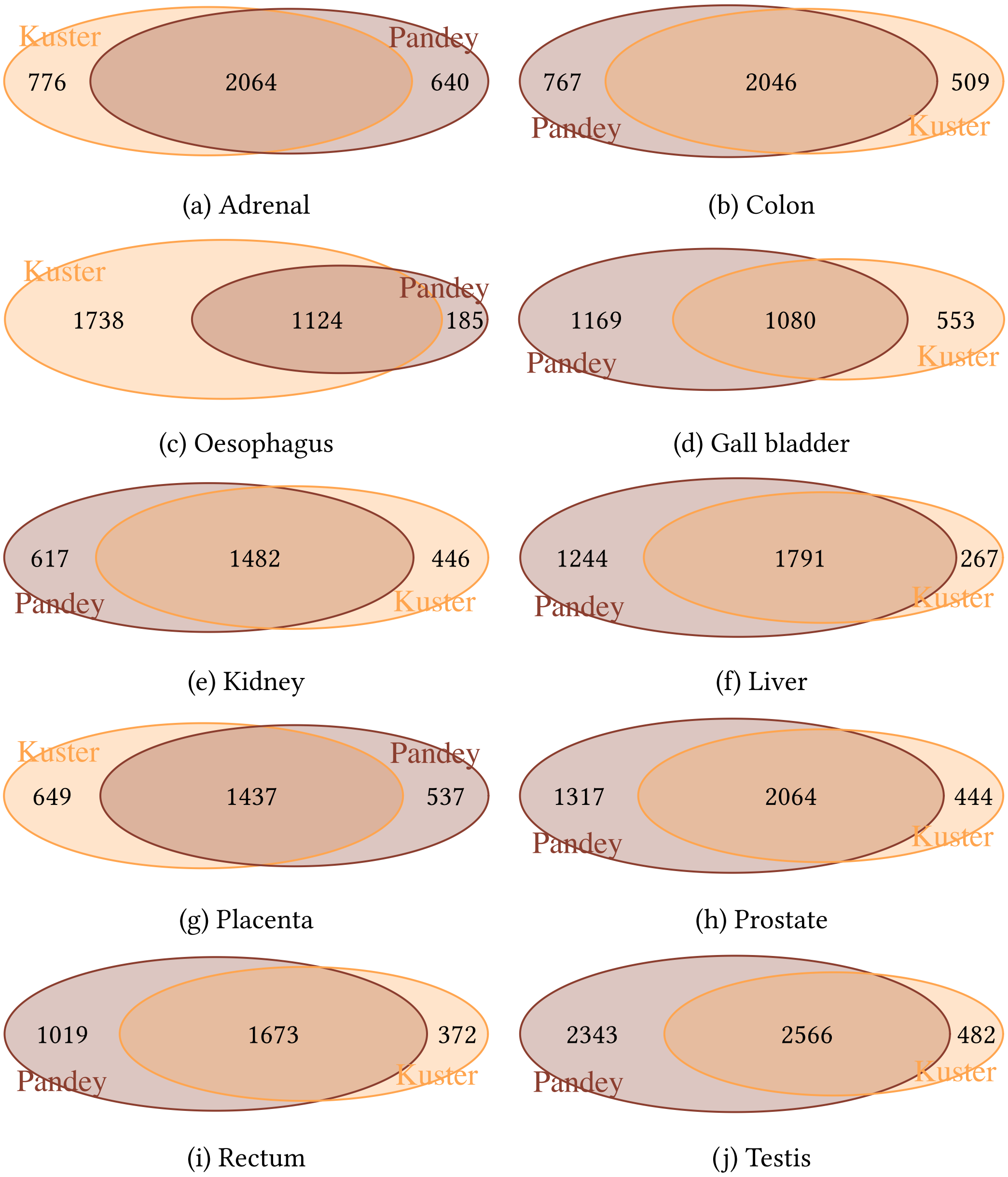


Figure D.2. **Proteins overlap between the fourteen common tissues between Pandey and Kuster proteome data.**



(a) Adrenal

(b) Colon

(c) Oesophagus

(d) Gall bladder

(e) Kidney

(f) Liver

(g) Placenta

(h) Prostate

(i) Rectum

(j) Testis

Figure D.3. **Unique and shared proteins across the other ten common tissues between Pandey and Kuster proteomic studies**

Table D.1. Proteins found in every tissue in all three datasets

Ensembl (76) gene ID	Gene symbol
ENSG00000163631	*ALB*
ENSG00000171403	*KRT9*
ENSG00000186395	*KRT10*

Table D.2. Proteins found in every tissue in Pandey and Kuster datasets

Ensembl (76) ID	Gene symbol	ENSEMBL (76) ID	Gene symbol
ENSG00000023191	*RNH1*	ENSG00000134308	*YWHAQ*
ENSG00000044574	*HSPA5*	ENSG00000134333	*LDHA*
ENSG00000067225	*PKM*	ENSG00000140575	*IQGAP1*
ENSG00000071127	*WDR1*	ENSG00000148180	*GSN*
ENSG00000074800	*ENO1*	ENSG00000149925	*ALDOA*
ENSG00000080824	*HSP90AA1*	ENSG00000160752	*FDPS*
ENSG00000089220	*PEBP1*	ENSG00000163631	*ALB*
ENSG00000089597	*GANAB*	ENSG00000164924	*YWHAZ*
ENSG00000092820	*EZR*	ENSG00000165280	*VCP*
ENSG00000096384	*HSP90AB1*	ENSG00000166598	*HSP90B1*
ENSG00000100345	*MYH9*	ENSG00000166794	*PPIB*
ENSG00000102144	*PGK1*	ENSG00000167658	*EEF2*
ENSG00000108518	*PFN1*	ENSG00000170027	*YWHAG*
ENSG00000108953	*YWHAE*	ENSG00000170248	*PDCD6IP*
ENSG00000111530	*CAND1*	ENSG00000171403	*KRT9*
ENSG00000111640	*GAPDH*	ENSG00000178209	*PLEC*
ENSG00000111669	*TPI1*	ENSG00000179218	*CALR*
ENSG00000111716	*LDHB*	ENSG00000182718	*ANXA2*
ENSG00000117450	*PRDX1*	ENSG00000186395	*KRT10*
ENSG00000130985	*UBA1*	ENSG00000204628	*GNB2L1*

Table D.3. Tissue specific proteins found both in Pandey et al. and Kuster et al. datasets

Tissue	Ensembl (76) ID	Gene symbol	Tissue	Ensembl (76) ID	Gene symbol
Adrenal gland	ENSG00000141744	*PNMT*	*Pancreas*	ENSG00000114204	*SERPINI2*
Adrenal gland	ENSG00000148655	*C10orf11*	*Pancreas*	ENSG00000141086	*CTRL*
Adrenal gland	ENSG00000160882	*CYP11B1*	*Pancreas*	ENSG00000143954	*REG3G*
Adrenal gland	ENSG00000163428	*LRRC58*	*Pancreas*	ENSG00000187021	*PNLIPRP1*
Adrenal gland	ENSG00000163626	*COX18*	*Pancreas*	ENSG00000215704	*CELA2B*
Kidney	ENSG00000074803	*SLC12A1*	*Pancreas*	ENSG00000266200	*PNLIPRP2*
Kidney	ENSG00000100253	*MIOX*	*Placenta*	ENSG00000105825	*TFPI2*
Kidney	ENSG00000112499	*SLC22A2*	*Placenta*	ENSG00000116183	*PAPPA2*
Kidney	ENSG00000113361	*CDH6*	*Placenta*	ENSG00000137868	*STRA6*
Kidney	ENSG00000148942	*SLC5A12*	*Placenta*	ENSG00000148848	*ADAM12*
Kidney	ENSG00000149452	*SLC22A8*	*Placenta*	ENSG00000163283	*ALPP*
Kidney	ENSG00000154025	*SLC5A10*	*Placenta*	ENSG00000172296	*SPTLC3*
Kidney	ENSG00000158296	*SLC13A3*	*Placenta*	ENSG00000172901	
Kidney	ENSG00000169344	*UMOD*	*Placenta*	ENSG00000183668	*PSG9*
Kidney	ENSG00000170482	*SLC23A1*	*Placenta*	ENSG00000243137	*PSG4*
Kidney	ENSG00000186335	*SLC36A2*	*Prostate*	ENSG00000044524	*EPHA3*
Kidney	ENSG00000197901	*SLC22A6*	*Prostate*	ENSG00000103710	*RASL12*
Liver	ENSG00000084734	*GCKR*	*Rectum*	ENSG00000205277	*MUC12*
Liver	ENSG00000100197	*CYP2D6*	*Testis*	ENSG00000052841	*TTC17*
Liver	ENSG00000135094	*SDS*	*Testis*	ENSG00000109762	*SNX25*
Liver	ENSG00000172497	*ACOT12*	*Testis*	ENSG00000130948	*HSD17B3*
Liver	ENSG00000198650	*TAT*	*Testis*	ENSG00000160310	*PRMT2*
Pancreas	ENSG00000010438	*PRSS3*			

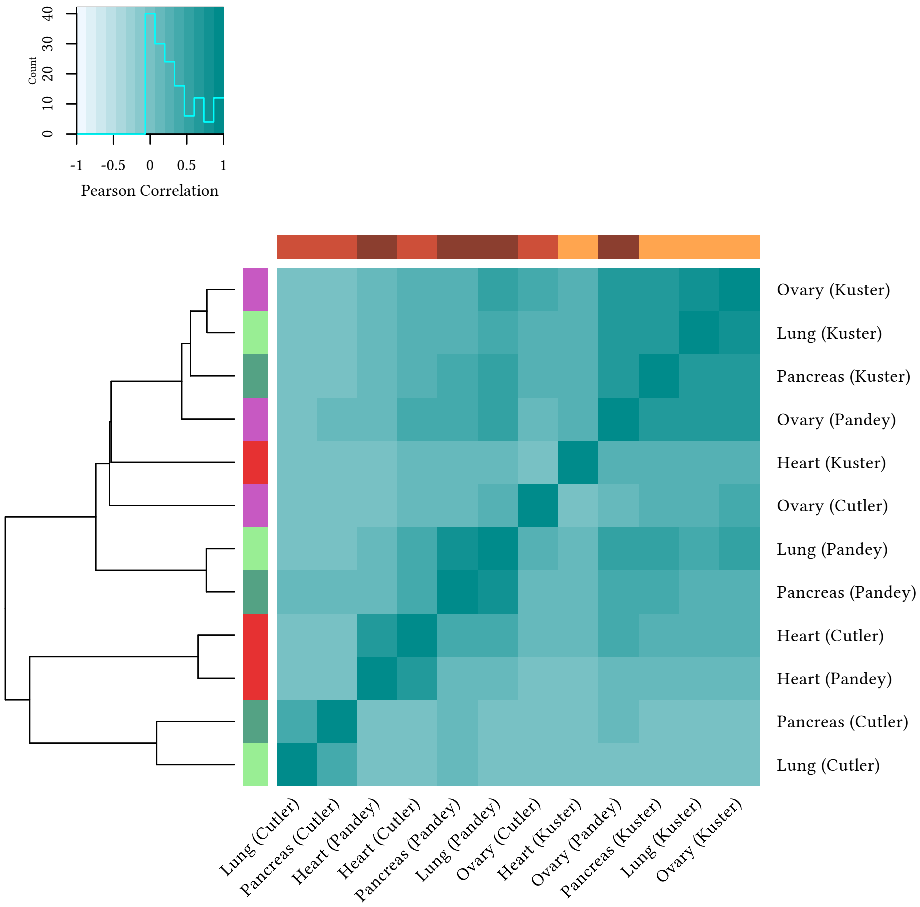


Figure D.4. **Heatmap of the four common tissues between the three proteome datasets** based on the pairwise **Pearson correlations** clustering of 1,384 proteins expression levels. See Figure 5.5 for the heatmap based on Spearman correlation.

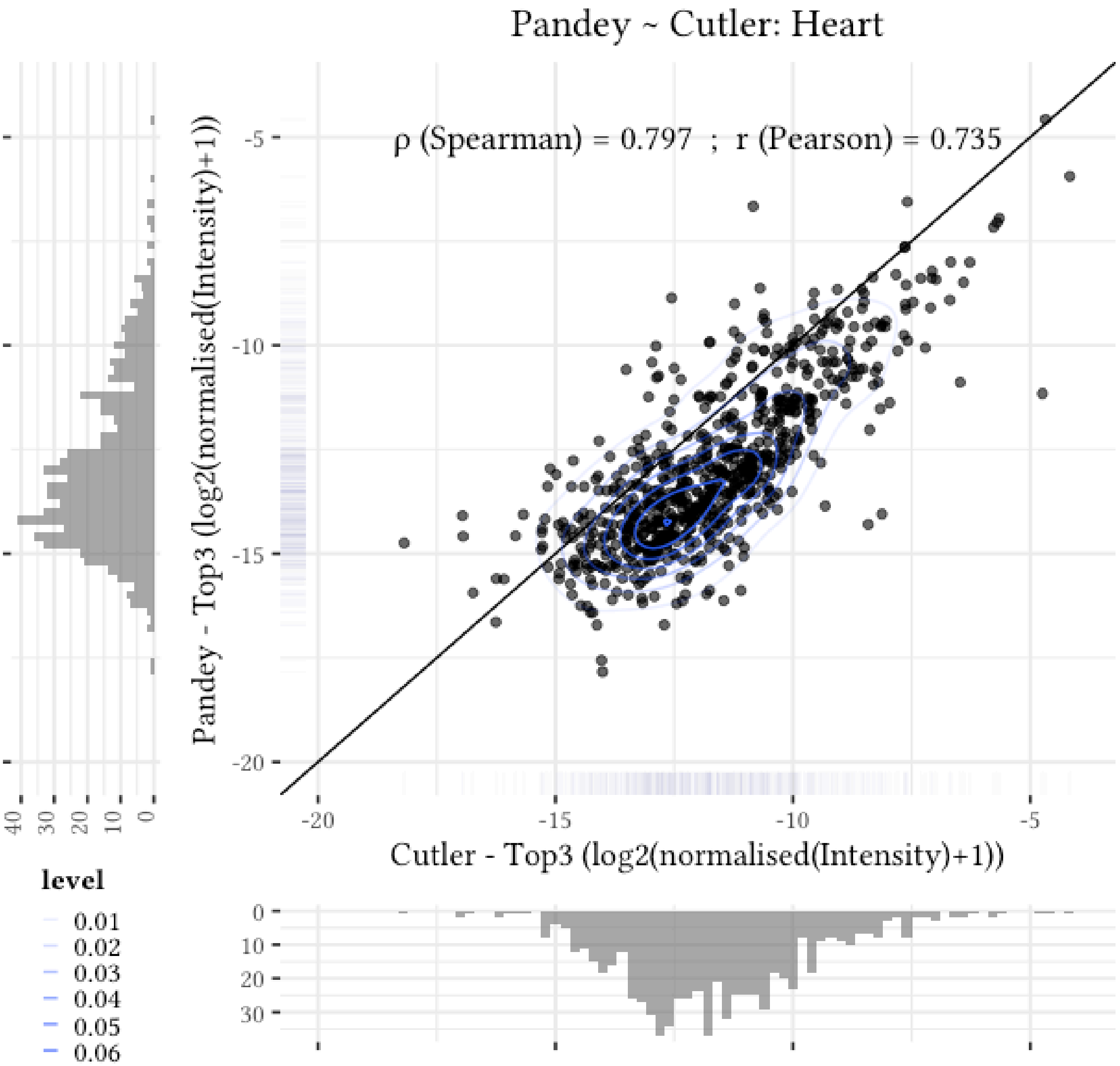


Figure D.5. **Scatterplot of Heart for Cutler and Pandey data.** There is an overall strong correlation between the expression of the *Heart* proteins between Pandey (y-axis) and Cutler (x-axis) even though the dispersion is quite substantial. The black line $y = x$ is only present as a visual reference.

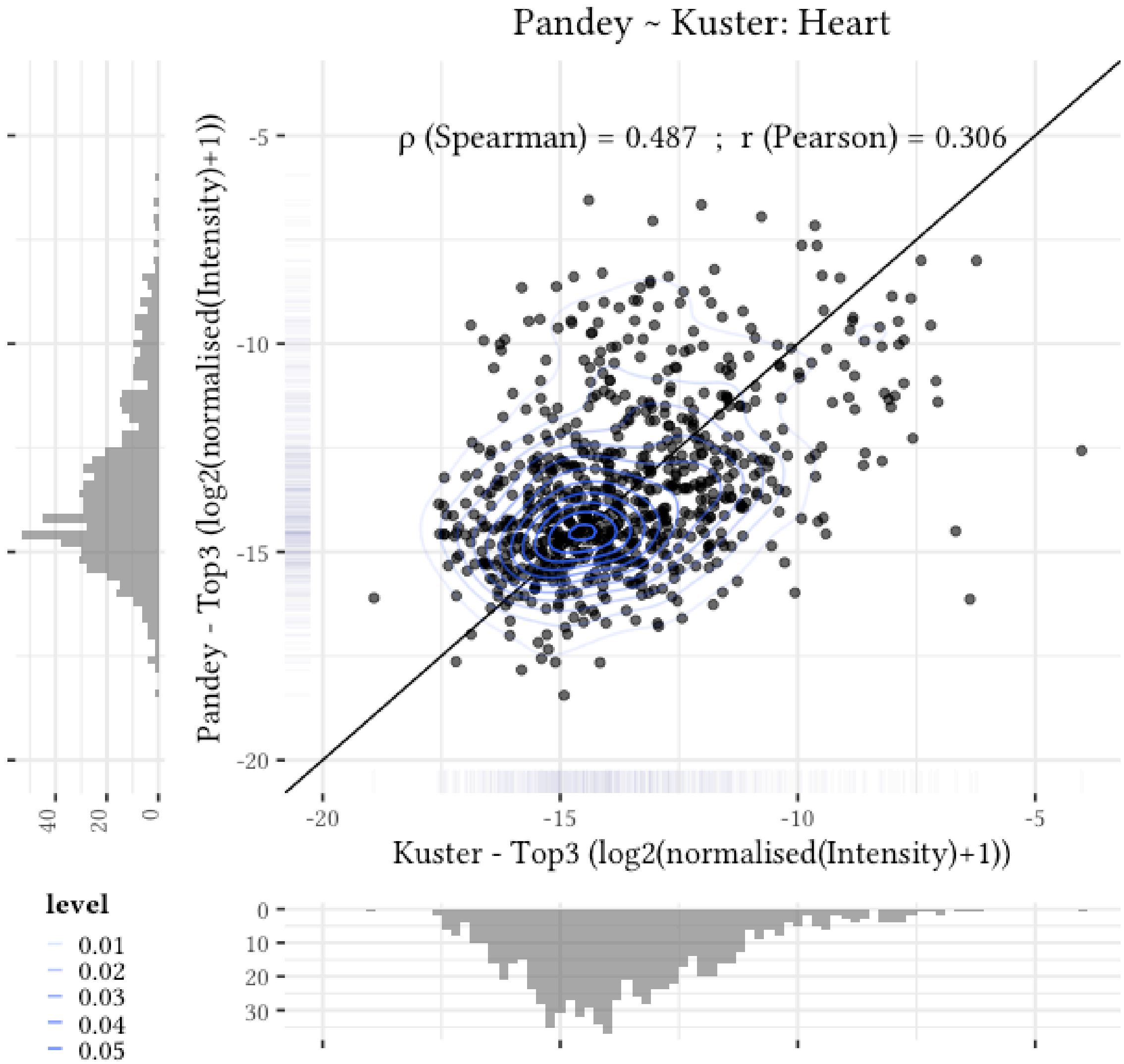


Figure D.6. **Scatterplot of Heart for Kuster and Pandey data.** The dispersion of expression is more significant than between Pandey and Cutler (see Figure D.5). It is rather difficult to assess the protein expression in *Heart* from Pandey (or Kuster) based on the other study. The black line $y = x$ is only present as a visual reference.

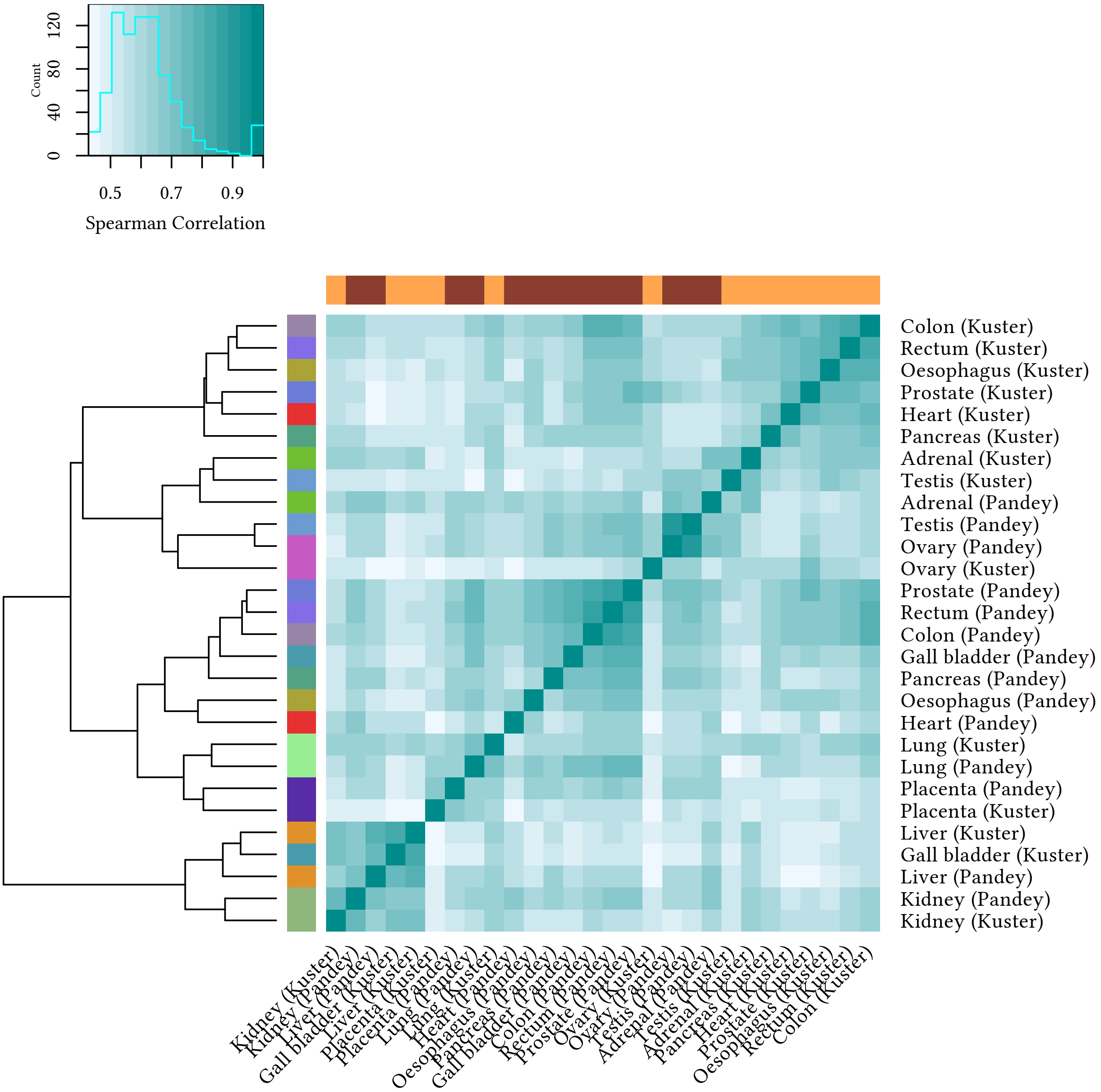


Figure D.7. **Heatmap of the fourteen common tissues between Pandey and Kuster datasets** based on the pairwise **Spearman correlations** of the expression levels of their 4,172 common proteins. *Placenta*, *Lung* and *Kidney* TREPs between Pandey and Kuster show an overall higher biological similarity than technical variability to the other tissues from the same study source. See also Figures D.8 to D.11.

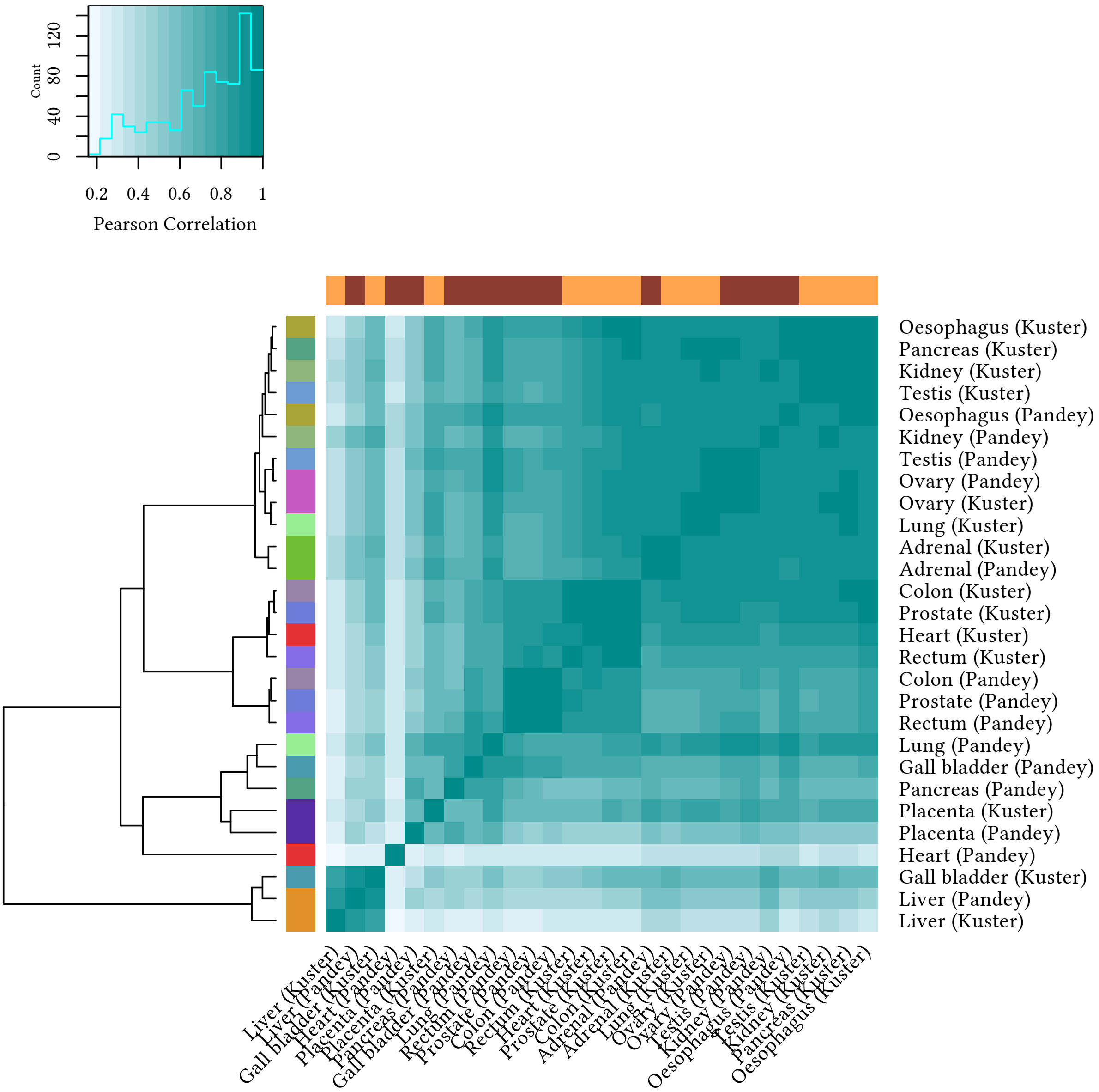


Figure D.8. **Heatmap of the fourteen common tissues between Pandey and Kuster datasets** based on the pairwise **Pearson correlations** of the expression levels of their 4,172 common proteins. Only *Placenta* and *Adrenal gland* TREPs between Pandey and Kuster show a greater biological similarity than technical one. See also Figures D.7 and D.9 to D.11.

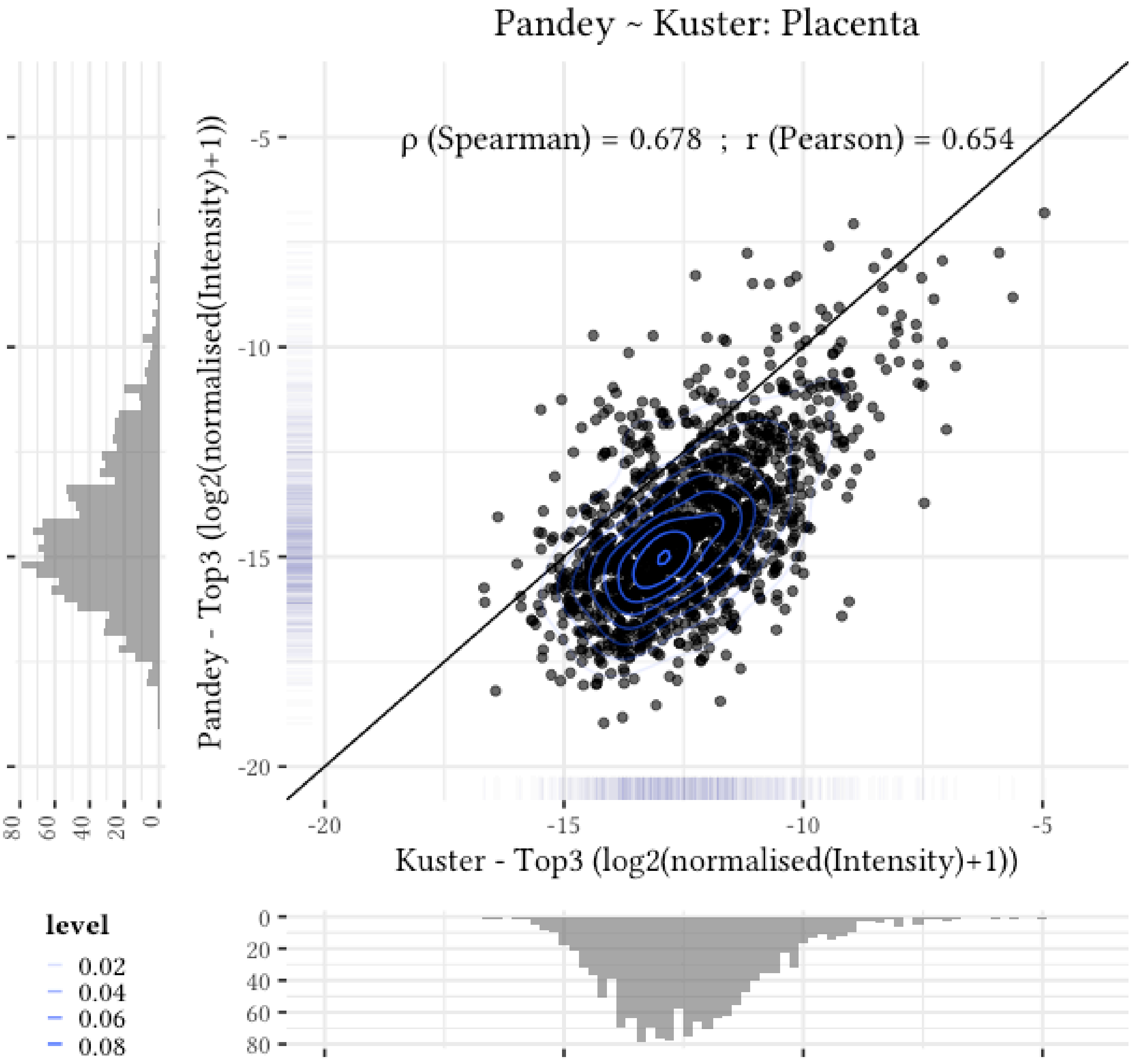


Figure D.9. **Scatterplot of Placenta for Kuster and Pandey data.** While the expression of some proteins are more spread between both datasets, there is an overall strong linear correlation between Pandey and Kuster for their *Placenta* tissue. Besides a few exception, protein expression levels in Pandey seem to be underestimated compared to Kuster. This is most probably due to the normalisation method. The black line $y = x$ is only present as a visual reference.

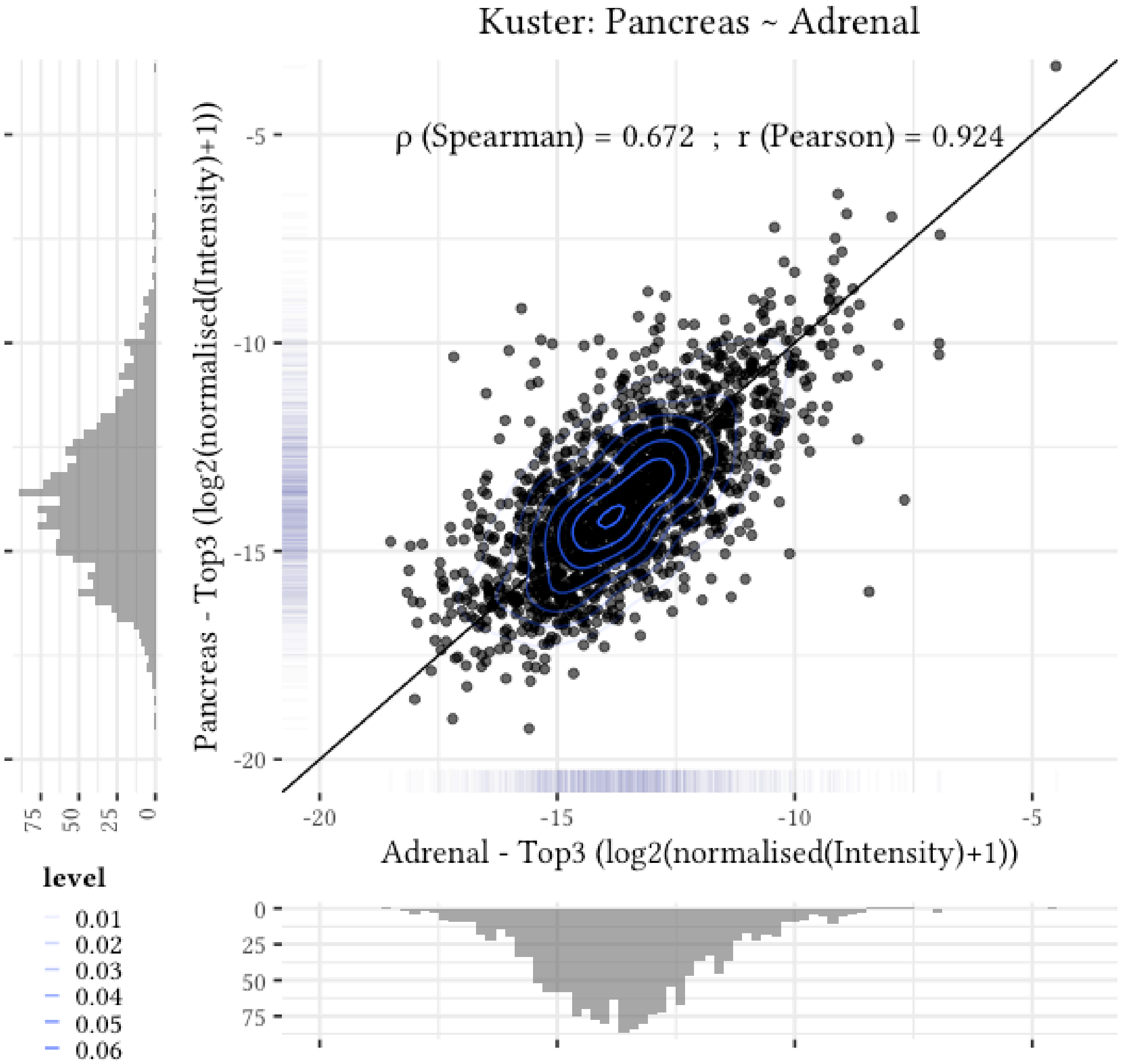


Figure D.10. **Scatterplot of Pancreas and Adrenal (from Kuster).** Although Kuster's *Pancreas* and *Adrenal gland* are never found in Figures D.7 and D.8 as the most similar to each other, their expression levels present a strong linear relationship to each other. The dispersion and outliers seem insuffisant to unquestionably distinguish different tissues. Figure D.11 is even more compelling. The black line $y = x$ is only present as a visual reference.

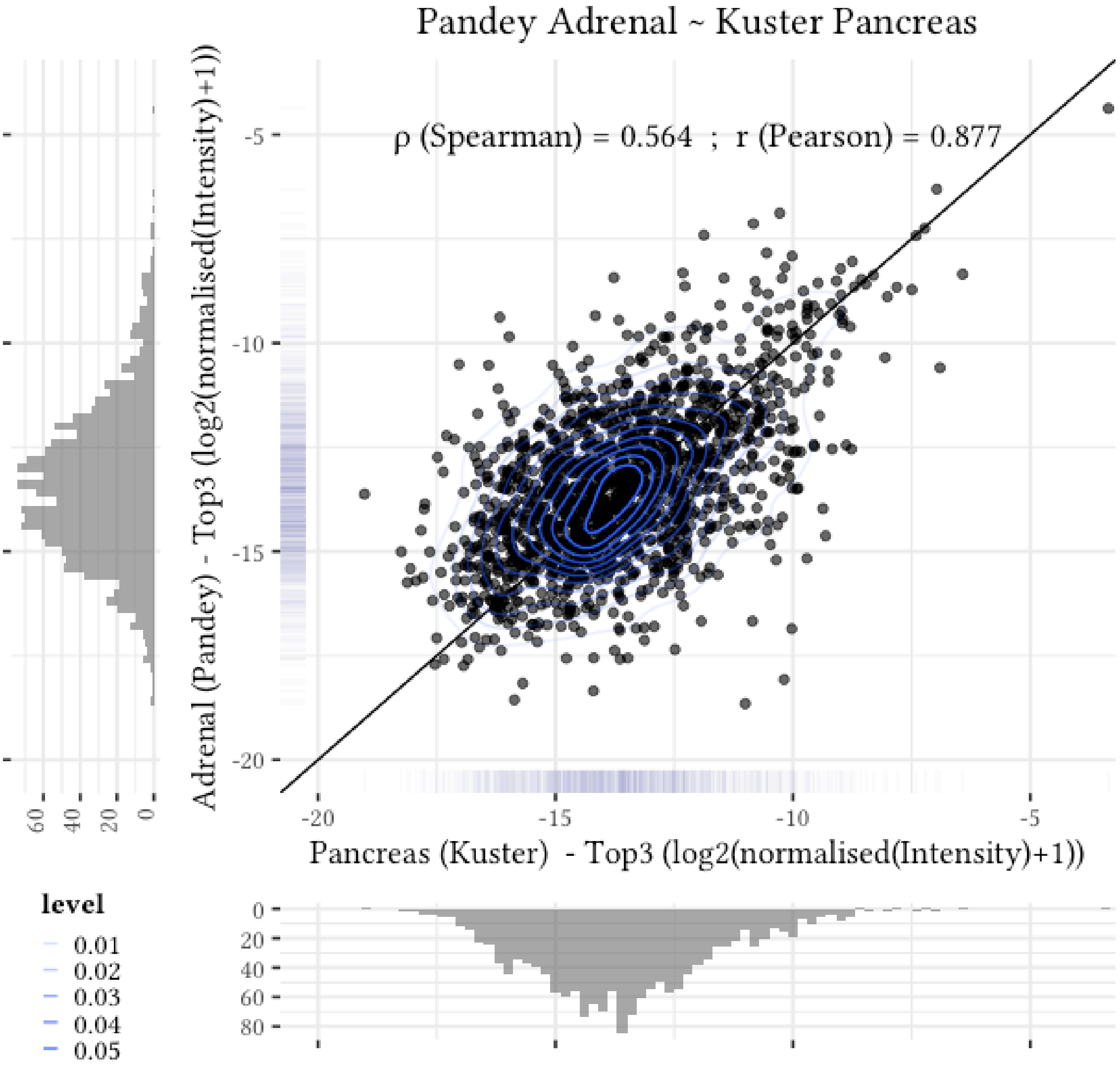


Figure D.11. **Scatterplot of Kuster Pancreas and Pandey Adrenal.** As in Figure D.10, Kuster's *Pancreas* and Pandey's *Adrenal gland* are never found as the most similar to each other. Once again, there is a strong linear relationship between their protein expression levels, even though the dispersion is greater and the outliers more numerous. The black line $y = x$ is only present as a visual reference.

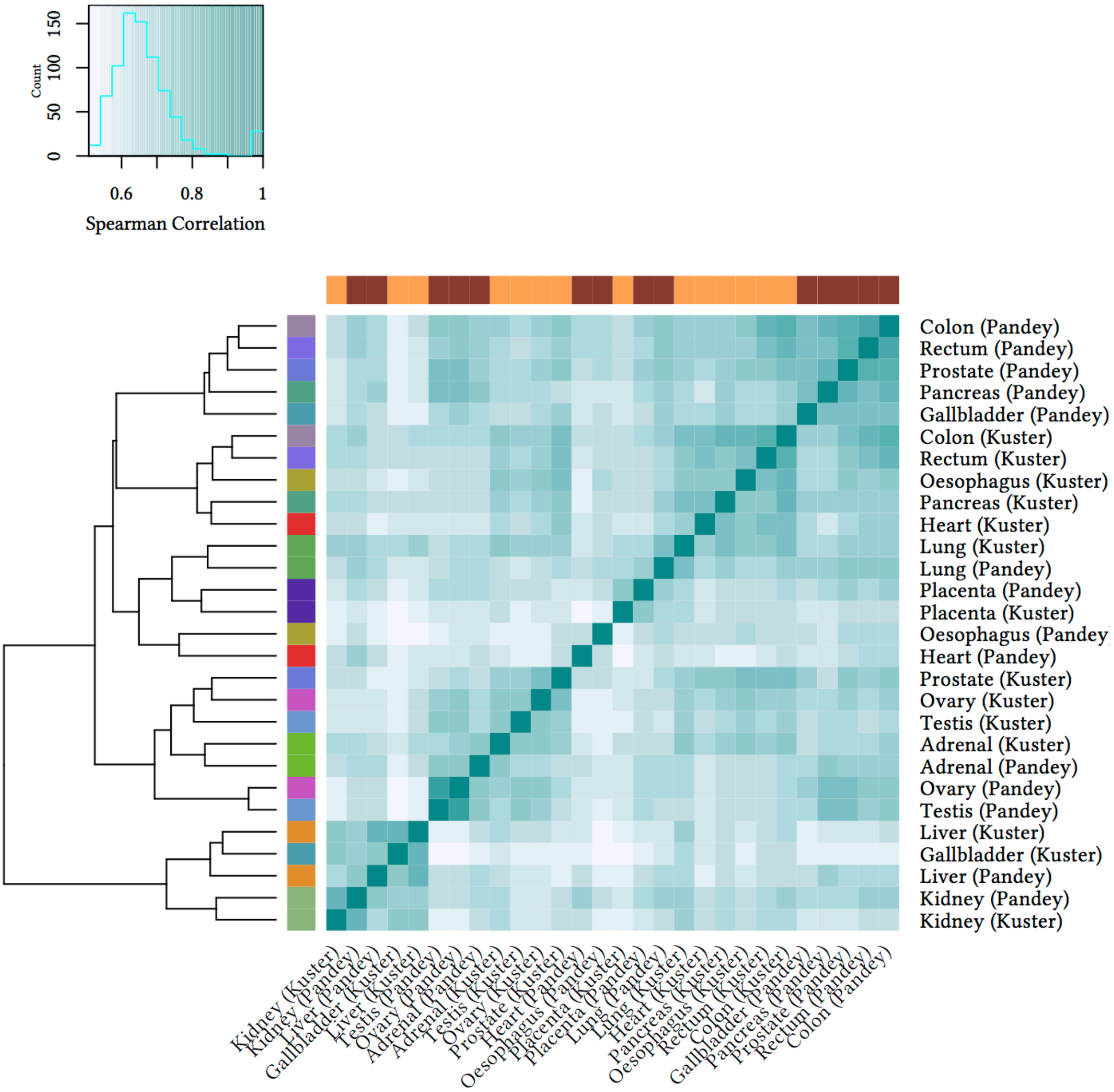


Figure D.12. **Heatmap of the fourteen common tissues between Pandey and Kuster (PPKM) datasets** based on the pairwise Spearman correlations of the expression levels of their 8,680 common proteins. Compared to Figure D.7, once again *Placenta*, *Lung*, *Kidney* are displaying a higher biological signal than technical variability. This is also the case of *Adrenal gland* tissue (as in Figure D.8). Thus, the results are similar to the ones from the first quantification method.

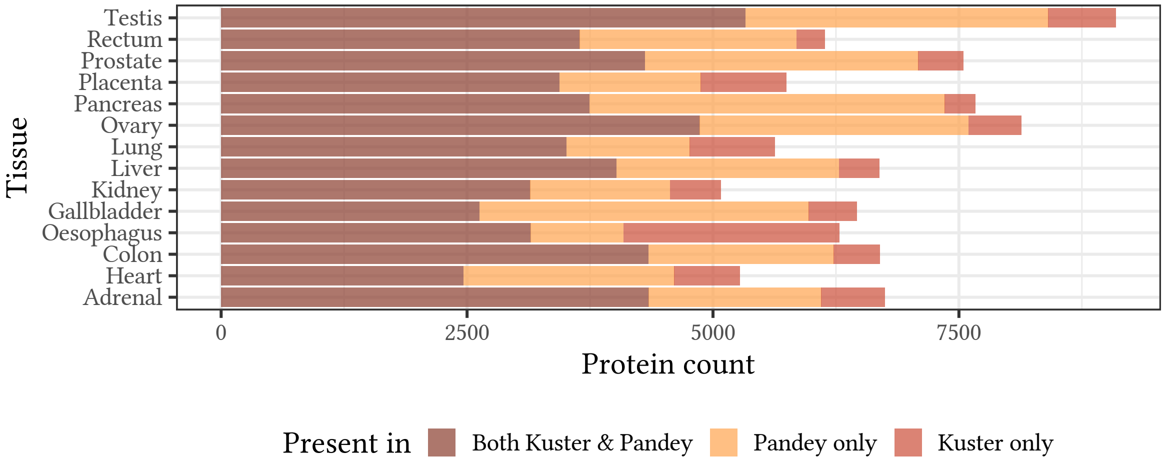


Figure D.13. **Number of identified proteins in each of the fourteen common tissues for Kuster and Pandey proteomic data with our new PPKM quantification method.** Although the number of proteins quantified by each method is different, this figure is very similar to Figure D.14.

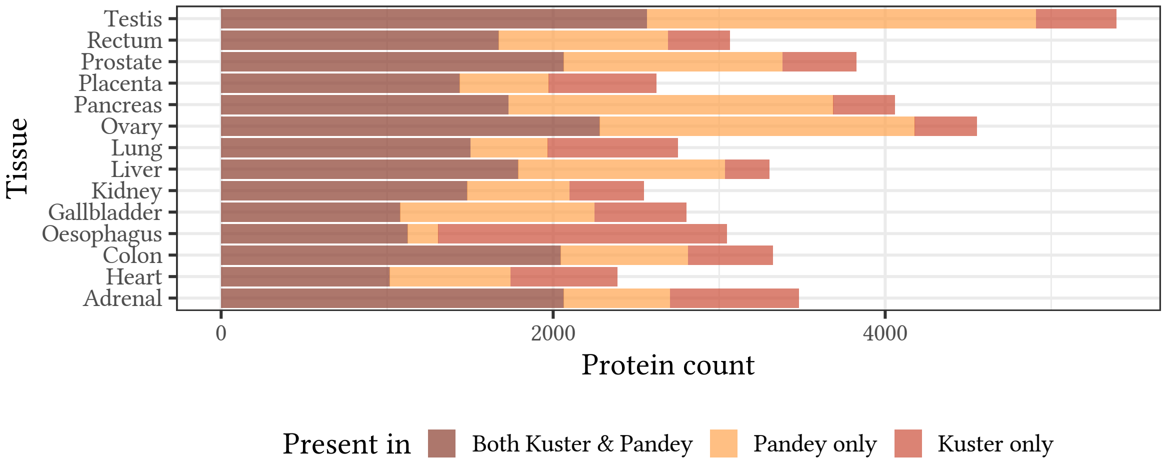


Figure D.14. **Number of identified proteins in each of the fourteen common tissues for Kuster and Pandey proteomic data quantified with the quantification described in Chapter 2.**

E SUPPLEMENTARY MATERIAL FOR CHAPTER 6

E.1 HYPERGEOMETRIC TEST

The hypergeometric test uses the hypergeometric distribution and equates to the one-sided Fisher's exact test. It allows measuring the statistical significance of randomly sampling k successes out of n draws, without replacement, from a population of N that contains K successes. Depending on whether the test is about an over or under-representation, the p-value is the probability of drawing respectively a minimum or a maximum of k successes.

See also N. L. Johnson et al. (2005) for more examples using this test.

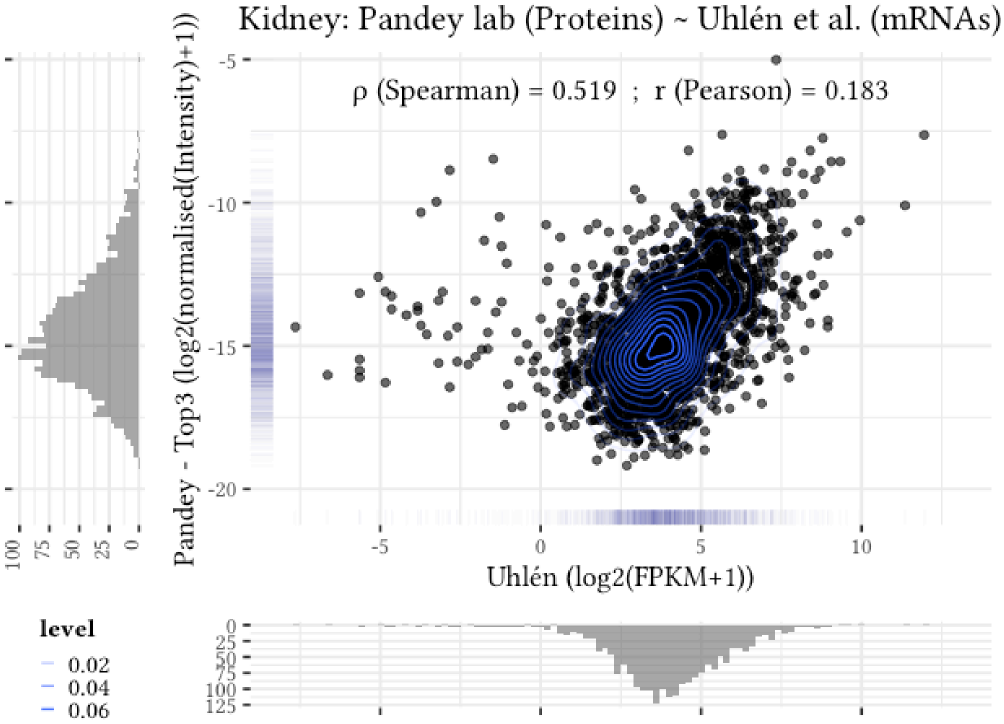


Figure E.1. **Scatterplot of protein (Pandey et al. — Top3 quantification) and mRNA (Uhlén et al.) expression for Kidney.**

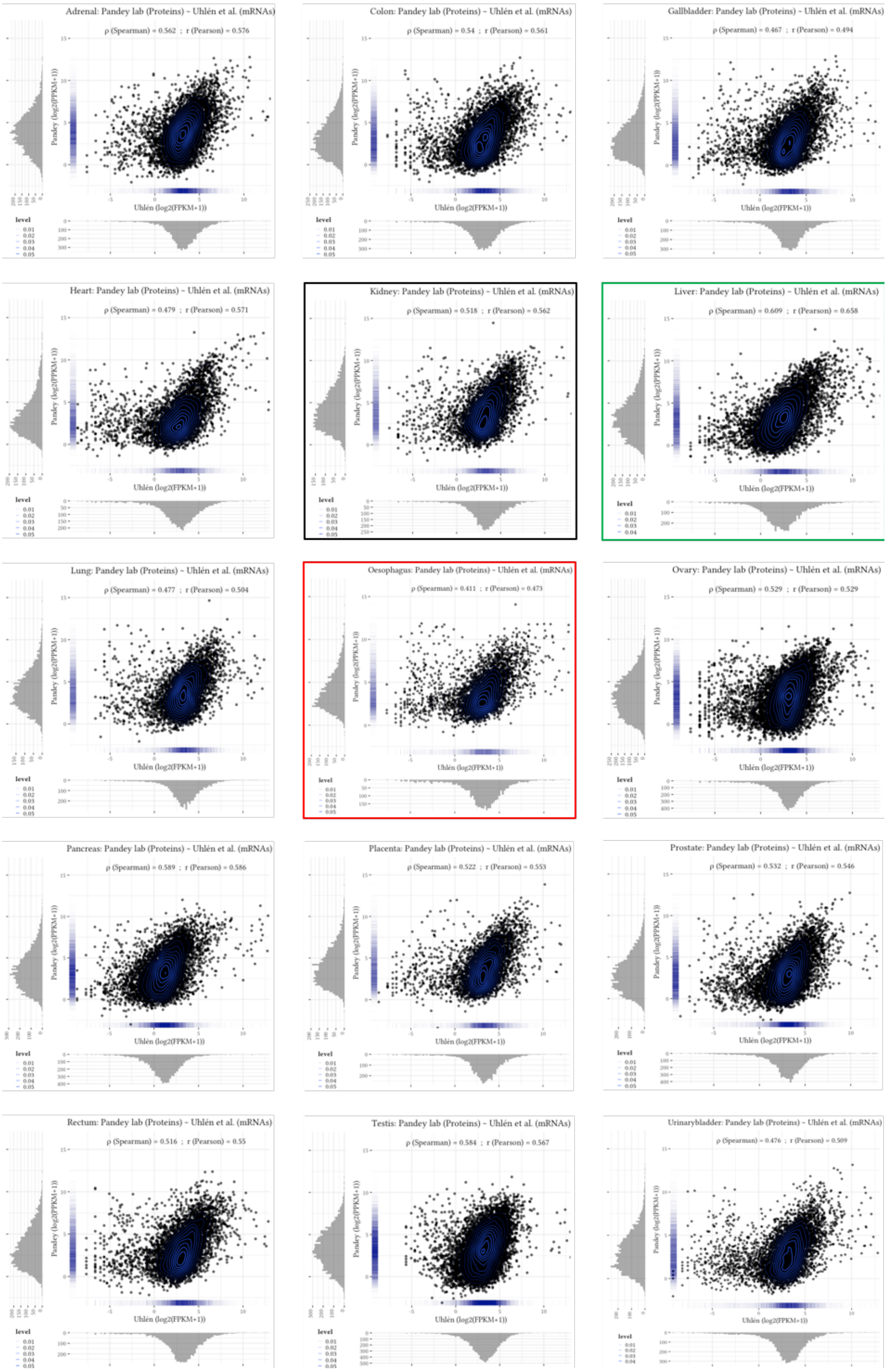


Figure E.2. **Overview of the tissue scatterplots between Uhlén and Pandey data.** The *Liver* presents the highest correlation and the *Oesophagus* the lowest one.

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
a, b, c, d	ENSG00000173349	SFT2D3	p. coding	SFT2 domain containing 3	HGNC Symbol Acc: 28767
a, b	ENSG00000198788	MUC2	processed transcript	mucin 2, oligomeric mucus/gel-forming	HGNC Symbol Acc: 7512
a, b, c, d	ENSG00000223953	C1QTNF5	p. coding	C1q and tumor necrosis factor related protein 5	HGNC Symbol Acc:14344
a, b	ENSG00000256453	DND1	p. coding	DND microRNA-mediated repression inhibitor	HGNC Symbol Acc:23799
a, b, c, d	ENSG00000262664	OVCA2	p. coding	ovarian tumor suppressor candidate 2	HGNC Symbol Acc:24203
b	ENSG00000163157	TMOD4	p. coding	tropomodulin 4 (muscle)	HGNC Symbol Acc:11874
b	ENSG00000203618	GP1BB	p. coding	glycoprotein Ib (platelet), beta polypeptide	HGNC Symbol Acc:4440
b	ENSG00000251322	SHANK3	processed transcript	SH3 and multiple ankyrin repeat domains 3	HGNC Symbol Acc:14294
c, d	ENSG00000105371	ICAM4	p. coding	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	HGNC Symbol Acc:5347
c	ENSG00000164708	PGAM2	p. coding	phosphoglycerate mutase 2 (muscle)	HGNC Symbol Acc:8889
c, d	ENSG00000181404	XXyac-YRM2039.2	unprocesssed pseudogene		
c, d	ENSG00000183336	BOLA2	p. coding	bolA family member 2	HGNC Symbol Acc:29488

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
c	ENSG00000196101	HLA-DRB3	p. coding	major histocompatibility complex, class II, DR beta 3	HGNC Symbol Acc:4951
c	ENSG00000203618	GP1BB		glycoprotein Ib (platelet), beta polypeptide	HGNC Symbol Acc:4440
c, d	ENSG00000206203	TSSK2	p. coding	testis-specific serine kinase 2	HGNC Symbol Acc:1140
c, d	ENSG00000206240, ENSG00000206306	HLA-DRB1	p. coding	major histocompatibility complex, class II, DR beta 1	HGNC Symbol Acc:4948
c, d	ENSG00000206305	HLA-DQA1		major histocompatibility complex, class II, DQ alpha 1	HGNC Symbol Acc:4942
c, d	ENSG00000206450, ENSG00000223532	HLA-B	p. coding	major histocompatibility complex, class I, B	HGNC Symbol Acc:4932
c, d	ENSG00000225691	HLA-C	p. coding	major histocompatibility complex, class I, C	HGNC Symbol Acc:4933
c, d	ENSG00000206505, ENSG00000224320, ENSG00000227715, ENSG00000235657, ENSG00000223980, ENSG00000229215	HLA-A	p. coding	major histocompatibility complex, class I, A	HGNC Symbol Acc:4931
c, d	ENSG00000211594	IGKJ4	IG J gene	immunoglobulin kappa joining 4	HGNC Symbol Acc:5722
c, d	ENSG00000211595	IGKJ3	IG J gene	immunoglobulin kappa joining 3	HGNC Symbol Acc:5721

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
c	ENSG00000213402	PTPRCAP	p. coding	protein tyrosine phosphatase, receptor type, C-associated protein	HGNC Symbol Acc:9667
c	ENSG00000215695	RSC1A1	p. coding	regulatory solute carrier protein, family 1, member 1	HGNC Symbol Acc:10458
c, d	ENSG00000227357	HLA-DRB4	p. coding	major histocompatibility complex, class II, DR beta 4	HGNC Symbol Acc:4952
c, d	ENSG00000231021	HLA-DRB4	p. coding	major histocompatibility complex, class II, DR beta 4	RefSeq mRNA Acc:NM_021983
c, d	ENSG00000231286	HLA-DQB1	p. coding	major histocompatibility complex, class II, DQ beta 1	HGNC Symbol Acc:4944
c, d	ENSG00000231679	HLA-DRB3	p. coding	major histocompatibility complex, class II, DR beta 3	RefSeq mRNA Acc:NM_022555
c, d	ENSG00000256453	DND1	p. coding	DND microRNA-mediated repression inhibitor	HGNC Symbol Acc:23799
c	ENSG00000263353	CH17-118O6.1	processed transcript		
c, d	ENSG00000276938	FAM157A	p. coding	Homo sapiens family with sequence similarity 157, member A	RefSeq mRNA Acc:NM_001145248
c, d	ENSG00000277656	GSTT1	p. coding	glutathione S-transferase theta 1	HGNC Symbol Acc:4641
c, d	ENSG00000277897	GSTT2	p. coding		
d	ENSG00000105507	CABP5	p. coding	calcium binding protein 5	HGNC Symbol Acc:3714
d	ENSG00000105954	NPVF	p. coding	neuropeptide VF precursor	HGNC Symbol Acc:13782

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
d	ENSG00000142539	CTD-2545M3.6	p. coding		
d	ENSG00000147896	IFNK	p. coding	interferon, kappa	HGNC Symbol Acc:21714
d	ENSG00000148136	OR13C4	p. coding	olfactory receptor, family 13, subfamily C, member 4	HGNC Symbol Acc:4722
d	ENSG00000163157	TMOD	p. coding	tropomodulin 4 (muscle)	HGNC Symbol Acc:11874
d	ENSG00000164708	PGAM2	p. coding	phosphoglycerate mutase 2 (muscle)	HGNC Symbol Acc:8889
d	ENSG00000166884	OR4D6	p. coding	olfactory receptor, family 4, subfamily D, member 6	HGNC Symbol Acc:15175
d	ENSG00000169840	GSX1	p. coding	GS homeobox 1	HGNC Symbol Acc:20374
d	ENSG00000170929	OR1M1	p. coding	olfactory receptor, family 1, subfamily M, member 1	HGNC Symbol Acc:8220
d	ENSG00000171053	PATE1	p. coding	prostate and testis expressed 1	HGNC Symbol Acc:24664
d	ENSG00000171396	KRTAP4-4	p. coding	keratin associated protein 4-4	HGNC Symbol Acc:16928
d	ENSG00000172155	LCE1D	p. coding	late cornified envelope 1D	HGNC Symbol Acc:29465
d	ENSG00000176239	OR51B6	p. coding	olfactory receptor, family 51, subfamily B, member 6	HGNC Symbol Acc:19600
d	ENSG00000182346	DAOA	p. coding	D-amino acid oxidase activator	HGNC Symbol Acc:21191

Table E.1. Found proteins without a counterpart in the transcriptomic data

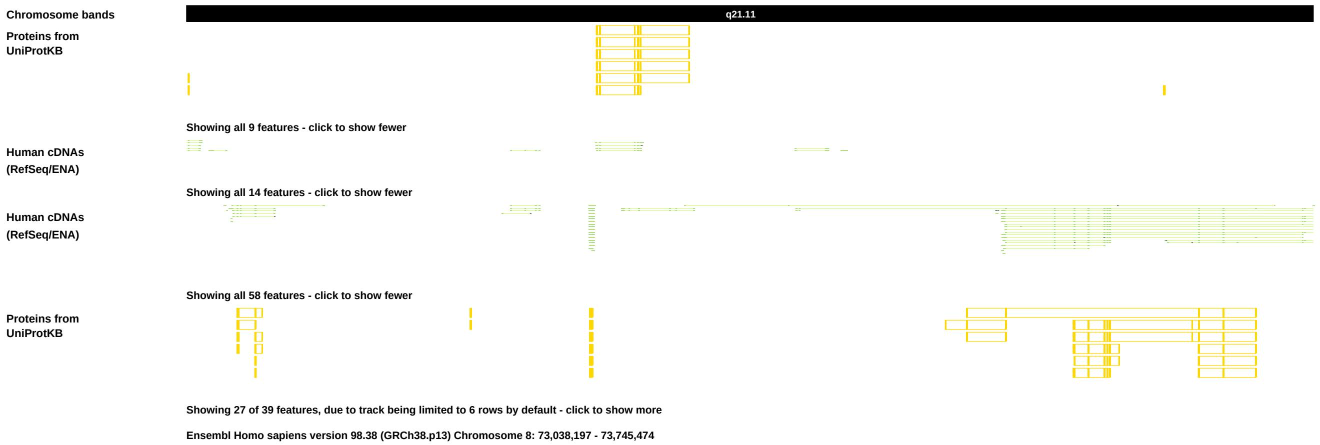
Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
d	ENSG00000182591	KRTAP11-1	p. coding	keratin associated protein 11-1	HGNC Symbol Acc:18922
d	ENSG00000184321	OR51J1	p. coding	olfactory receptor, family 51, subfamily J, member 1 (gene/pseudogene)	HGNC Symbol Acc:14856
d	ENSG00000187173	LCE2A	p. coding	late cornified envelope 2A	HGNC Symbol Acc:29469
d	ENSG00000187766	KRTAP10-8	p. coding	keratin associated protein 10-8	HGNC Symbol Acc:20525
d	ENSG00000196101	HLA-DRB3	p. coding	major histocompatibility complex, class II, DR beta 3	HGNC Symbol Acc:4951
d	ENSG00000203618	GP1BB	p. coding	glycoprotein Ib (platelet), beta polypeptide	HGNC Symbol Acc:4440
d	ENSG00000203818	HIST2H3PS2	p. coding	histone cluster 2, H3, pseudogene 2	HGNC Symbol Acc:32060
d	ENSG00000205883	DEFB135	p. coding	defensin, beta 135	HGNC Symbol Acc:32400
d	ENSG00000206452	HLA-C	p. coding	major histocompatibility complex, class I, C	HGNC Symbol Acc:4933
d	ENSG00000211831	TRAJ61	TR J gene	T cell receptor alpha joining 61 (non-functional)	HGNC Symbol Acc:12094
d	ENSG00000211835	TRAJ56	TR J gene	T cell receptor alpha joining 56	HGNC Symbol Acc:12088
d	ENSG00000213316	LTC4S	p. coding	leukotriene C4 synthase	HGNC Symbol Acc:6719

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
d	ENSG00000213402	PTPRCAP	p. coding	protein tyrosine phosphatase, receptor type, C-associated protein	HGNC Symbol Acc:9667
d	ENSG00000215695	RSC1A1	p. coding	regulatory solute carrier protein, family 1, member 1	HGNC Symbol Acc:10458
d	ENSG00000224902	GAGE12H	p. coding	G antigen 12H	HGNC Symbol Acc:31908
d	ENSG00000233732	IGHV3OR16-10	IG V gene	immunoglobulin heavy variable 3 OR16-10 (non-functional)	HGNC Symbol Acc:5634
d	ENSG00000249209	AP000304.12	p. coding		
d	ENSG00000249730	OR10J4	polymorphic pseudogene	olfactory receptor, family 10, subfamily J, member 4 (gene/pseudogene)	HGNC Symbol Acc:15408
d	ENSG00000253148	RGS21	p. coding	regulator of G-protein signaling 21	HGNC Symbol Acc:26839
d	ENSG00000255009	UBTFL1	processed pseudogene	upstream binding transcription factor, RNA polymerase I-like 1	HGNC Symbol Acc:14533
d	ENSG00000255472	RP11-998D10.1	p. coding	uncharacterized protein	UniProtKB/TrEMBL Acc:E9PR74
d	ENSG00000259490	IGHV3OR15-7	IG V gene	immunoglobulin heavy variable 3 OR15-7 (pseudogene)	HGNC Symbol Acc:5633
d	ENSG00000263353	CH17-118O6.1	processed transcript		
d	ENSG00000270467	IGHV3OR16-12	IG V gene	immunoglobulin heavy variable 3 OR16-12 (non-functional)	HGNC Symbol Acc:5636

Table E.1. Found proteins without a counterpart in the transcriptomic data

Set	ENSEMBL (76) ID	Gene name	Biotype	Description	Source and Accessing number
d	ENSG00000270472	IGHV3OR16-9	IG V gene	immunoglobulin heavy variable 3 OR16-9 (non-functional)	HGNC Symbol Acc:5644



Showing all 9 features - click to show fewer

Showing all 14 features - click to show fewer

Showing all 58 features - click to show fewer

Showing 27 of 39 features, due to track being limited to 6 rows by default - click to show more

Ensembl Homo sapiens version 98.38 (GRCh38.p13) Chromosome 8: 73,038,197 - 73,745,474

Figure E.3. ***STAU2* definition** The chromosome annotations for the mRNA and the protein of *STAU2* are different.

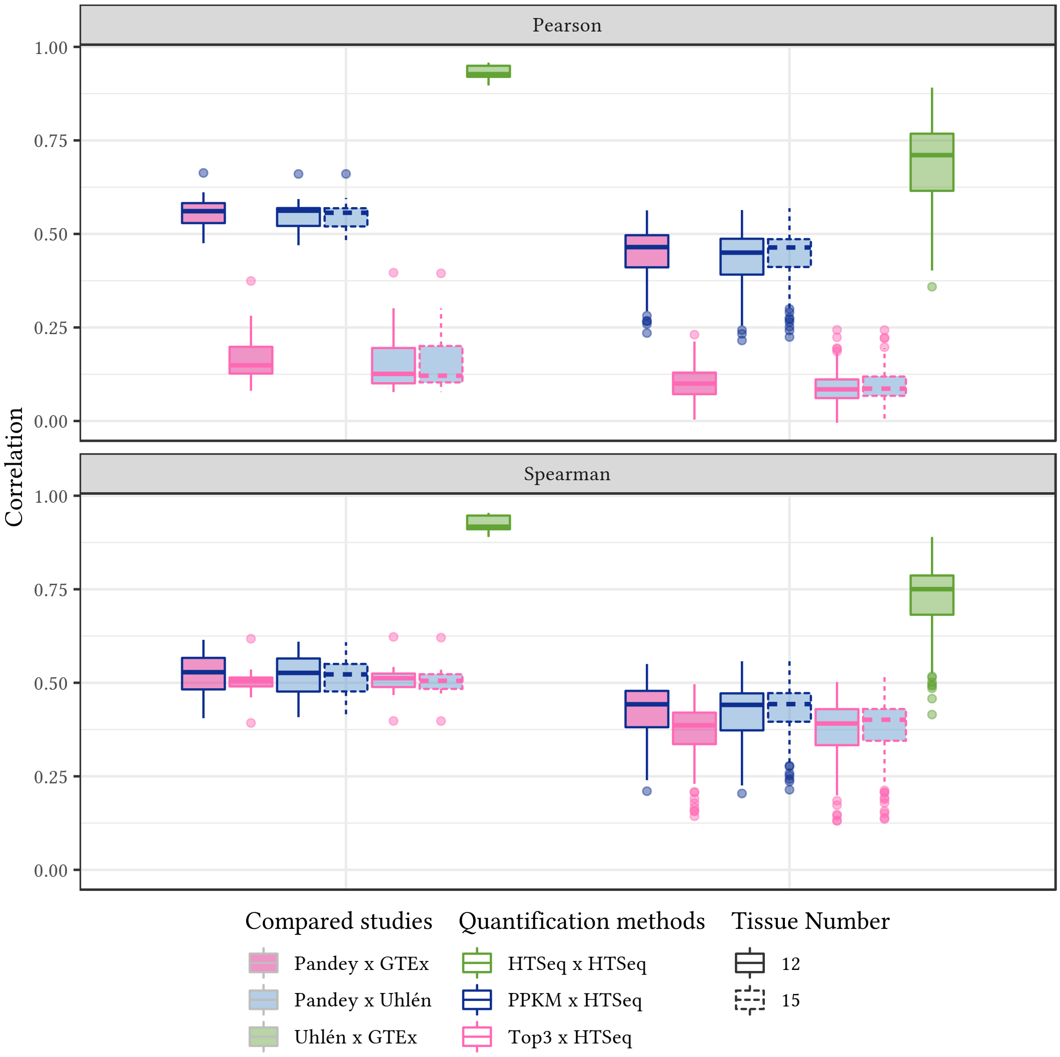


Figure E.4. **Distribution of Pearson and Spearman correlation coefficients for same-tissue proteomic and transcriptomic pairs versus random tissue pairs (untransformed data).**

Table E.2. **Summary of Pearson and Spearman correlation coefficients between proteomics and transcriptomics** across several data combinations. See also Figure 6.8.

Datasets		Number of tissues	Quantification methods	Scaled data $\log_2(x+1)$	Correlation method	Mean correlation of		p-value
						Same-tissue pairs	Different tissues pairs	
Pandey et al.	Uhlén et al.	12	Top3 x HTSeq	True	Spearman	0.51	0.37	4.66e-07
Pandey et al.	GTEx	12	Top3 x HTSeq	True	Spearman	0.5	0.37	7.379e-07
Uhlén et al.	GTEx	12	HTSeq x HTSeq	True	Spearman	0.91	0.66	< 2.2e-16
Pandey et al.	Uhlén et al.	15	Top3 x HTSeq	True	Spearman	0.5	0.38	2.659e-08
Pandey et al.	Uhlén et al.	12	Top3 x HTSeq	True	Pearson	0.11	0.06	0.03696
Pandey et al.	GTEx	12	Top3 x HTSeq	True	Pearson	0.12	0.07	0.02895
Uhlén et al.	GTEx	12	HTSeq x HTSeq	True	Pearson	0.93	0.68	< 2.2e-16
Pandey et al.	Uhlén et al.	15	Top3 x HTSeq	True	Pearson	0.1	0.06	0.02271
Pandey et al.	Uhlén et al.	12	PPKM x HTSeq	True	Spearman	0.52	0.42	4.795e-05
Pandey et al.	GTEx	12	PPKM x HTSeq	True	Spearman	0.52	0.43	8.475e-05
Uhlén et al.	GTEx	12	HTSeq x HTSeq	True	Spearman	0.92	0.72	< 2.2e-16
Pandey et al.	Uhlén et al.	15	PPKM x HTSeq	True	Spearman	0.52	0.43	8.422e-06
Pandey et al.	Uhlén et al.	12	PPKM x HTSeq	True	Pearson	0.5	0.37	0.0004002
Pandey et al.	GTEx	12	PPKM x HTSeq	True	Pearson	0.5	0.41	0.0003306
Uhlén et al.	GTEx	12	HTSeq x HTSeq	True	Pearson	0.94	0.73	< 2.2e-16
Pandey et al.	Uhlén et al.	15	PPKM x HTSeq	True	Pearson	0.49	0.4	9.941e-05
Pandey et al.	Uhlén et al.	12	Top3 x HTSeq	False	Spearman	0.51	0.37	4.66e-07
Pandey et al.	GTEx	12	Top3 x HTSeq	False	Spearman	0.5	0.37	7.379e-07
Uhlén et al.	GTEx	12	HTSeq x HTSeq	False	Spearman	0.91	0.66	< 2.2e-16
Pandey et al.	Uhlén et al.	15	Top3 x HTSeq	False	Spearman	0.5	0.38	2.66e-08

Table E.2. **Summary of Pearson and Spearman correlation coefficients between proteomics and transcriptomics** across several data combinations. See also Figure 6.8.

Datasets		Number of tissues	Quantification methods	Scaled data $\log_2(x+1)$	Correlation method	Mean correlation of		p-value
						Same-tissue pairs	Different tissues pairs	
Pandey et al.	Uhlén et al.	12	Top3 x HTSeq	False	Pearson	0.17	0.09	0.022
Pandey et al.	GTEx	12	Top3 x HTSeq	False	Pearson	0.17	0.1	0.015
Uhlén et al.	GTEx	12	HTSeq x HTSeq	False	Pearson	0.92	0.64	< 2.2e-16
Pandey et al.	Uhlén et al.	15	Top3 x HTSeq	False	Pearson	0.16	0.1	0.012
Pandey et al.	Uhlén et al.	12	PPKM x HTSeq	False	Spearman	0.52	0.42	4.795e-05
Pandey et al.	GTEx	12	PPKM x HTSeq	False	Spearman	0.52	0.43	8.475e-05
Uhlén et al.	GTEx	12	HTSeq x HTSeq	False	Spearman	0.92	0.72	< 2.2e-16
Pandey et al.	Uhlén et al.	15	PPKM x HTSeq	False	Spearman	0.52	0.43	8.422e-06
Pandey et al.	Uhlén et al.	12	PPKM x HTSeq	False	Pearson	0.55	0.43	1.059e-06
Pandey et al.	GTEx	12	PPKM x HTSeq	False	Pearson	0.56	0.45	2.026e-06
Uhlén et al.	GTEx	12	HTSeq x HTSeq	False	Pearson	0.93	0.69	< 2.2e-16
Pandey et al.	Uhlén et al.	15	PPKM x HTSeq	False	Pearson	0.55	0.45	1.061e-07

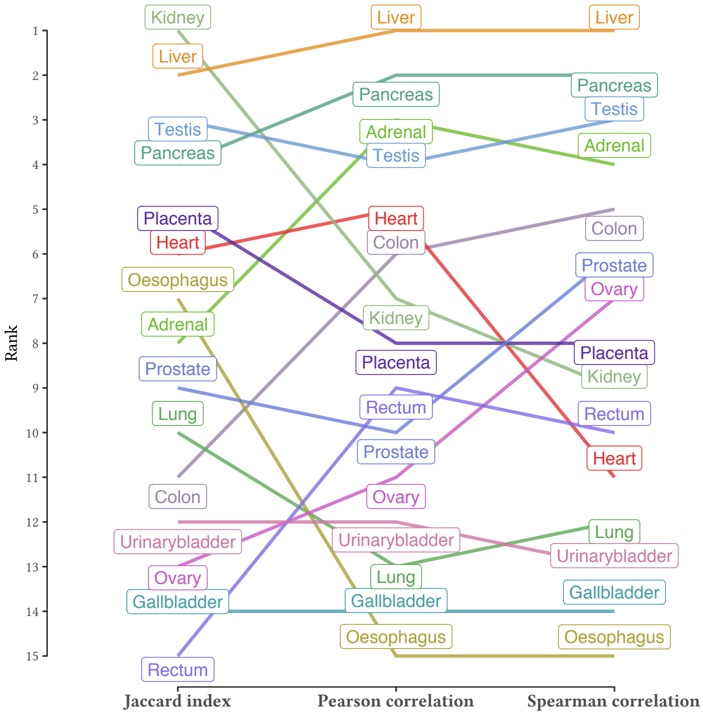


Figure E.5. **Rank comparison between the Pearson/Spearman correlation and the Jaccard indices computed for matching proteomics and transcriptomics.**

E.2 TS PROTEIN PERCENT

The percentage of TS proteins is calculated as follow:

$$\forall a \in \mathcal{A}, \forall n \in [1, \mathcal{N}] \quad p_{\mathcal{TS}}(n,a) = \sum_{k=1}^{n} \delta_{g_{a,k}} \cdot \frac{1}{n} \cdot 100 \qquad \text{(TS protein percentage)}$$

where:

- $\mathcal{S}$ is the set of 10,000 randomised expression datasets based on Pandey Lab data. These simulated datasets are created by random permutation of the gene labels and their associated vector of expression values across the tissues.
- $\mathcal{D}$ is the set of expression datasets; $\mathcal{D} = \mathcal{S} \cup \{$ protein expression for Pandey Lab data; mRNA expression for Uhlén et al. data; mRNA expression for GTEx data $\}$.
- $\mathcal{G}$ is the set of genes g that are shared by all elements of $\mathcal{D}$.
- $\mathcal{N}$ is the number of elements in $\mathcal{G}$.
- $\mathcal{TS}$ is the set of genes g for which the protein is TS (tissue-specific) in Pandey Lab data. $\mathcal{TS} \subset \mathcal{G}$.
- $\forall g \in \mathcal{G}, \delta_g = \begin{cases} 1 & \text{if } g \in \mathcal{TS} \\ 0 & \text{if } g \notin \mathcal{TS} \end{cases}$
- $\mathcal{A}$ is a set of unordered 2-tuples of elements from $\mathcal{D}$; $\mathcal{A} = \{($protein expression for Pandey Lab data, mRNA expression for Uhlén et al. data$)$; $($mRNA expression for Uhlén et al. data, mRNA expression for GTEx data$)$; $(s,$ mRNA expression for Uhlén et al. data$)\}$. $\forall s \in \mathcal{S}$.
- $\mathcal{C}$ is a correlation function such that: $\forall g \in \mathcal{G}$, $\forall a = (d_1, d_2) \in \mathcal{A}\ \mathcal{C}(g,a) \longmapsto$ correlation coefficient of g for its expression across tissues shared by d_1 and d_2.
- $(g_{a,k})$ is the sequence of genes $g_{a,k}$ of $\mathcal{G}$ such that: $\forall k \in [1; \mathcal{N}-1]\ \mathcal{C}(g_{a,k}, a) \geq \mathcal{C}(g_{a,k+1}, a)$

F LIST OF R PACKAGES

R [R Core Team, 2019] packages versions are only given as an indication. Most of the code can be run with older or newer versions.

- extrafont (0.17) [Chang, 2014]
- RColorBrewer (1.1) [Neuwirth, 2014]
- Cairo (1.50) [Urbanek et al., 2019]
- reshape2 (1.4.3) [Wickham, 2007]
- scales (1.0) [Wickham, 2018]
- MASS (7.3) [Venables et al., 2002]
- data.table (1.12.2) [Dowle et al., 2019]
- ggplot2 (3.1.1) [Wickham, 2016]
- gridExtra (2.3) [Auguie, 2017]
- gridBase (0.4) [Murrell, 2014]
- ggthemes (4.1.1) [Arnold, 2019]
- devtools (2.1.0) [Wickham et al., 2019]
- modules [Schubert and Rudolph, 2014]
- ebits [Rudolph, 2014]
- VennDiagram (1.6.20) [H. Chen, 2018]
- gplots (3.0.1.1) [Warnes et al., 2019]
- Bioconductor (2.44) [Huber et al., 2015]
- ape (5.3) [Paradis et al., 2019]
- biomaRt (2.40) [Durinck et al., 2005]
- clusterProfiler (3.12) [G. Yu et al., 2012]
- org.Hs.eg.db (3.8.2) [Carlson, 2019]
- WGCNA (1.67) [Langfelder et al., 2008]
- mgcv (1.8) [Wood, 2004]
- europepmc (0.3) [Jahn, 2018]
- rmarkdown (1.12) [Xie, Allaire, et al., 2018]
- DT (0.6) [Xie, Cheng, et al., 2019]
- shiny (1.3.2) [Chang et al., 2019]
- clustermq (0.8.5) [Schubert, 2019]

I have created two packages to help reproduce the different analyses presented in this thesis:

- barzinePhdData for the data (https://github.com/barzine/barzinePhdData), and
- barzinePhdR (https://github.com/barzine/barzinePhdR).

G LIST OF PUBLICATIONS

Barzine, M. P., K. Freivalds, J. C. Wright, M. Opmanis, D. Rituma, F. Z. Ghavidel, A. F. Jarnuczak, E. Celms, K. Čerāns, I. Jonassen, L. Lace, J. Antonio Vizcaíno, J. S. Choudhary, A. Brazma, and J. Viksna (2020). 'Using Deep Learning to Extrapolate Protein Expression Measurements'. *Proteomics* 20 (21-22), e2000009.

Jarnuczak, A. F., H. Najgebauer, M. Barzine, D. J. Kundu, F. Ghavidel, Y. Perez-Riverol, I. Papatheodorou, A. Brazma, and J. A. Vizcaíno (2019). 'An integrated landscape of protein expression in human cancer'. (under review).

Petryszak, R., M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, A. Füllgrabe, A. M.-P. Fuentes, S. Jupp, S. Koskinen, O. Mannion, L. Huerta, K. Megy, C. Snow, E. Williams, M. Barzine, E. Hastings, H. Weisser, J. Wright, P. Jaiswal, W. Huber, J. Choudhary, H. E. Parkinson, and A. Brazma (2015). 'Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants'. *Nucleic Acids Research* 44.D1, pp. D746–52.

Rustici, G., E. Williams, M. Barzine, A. Brazma, R. Bumgarner, M. Chierici, C. Furlanello, L. Greger, G. Jurman, M. Miller, B. F. Francis Ouellette, J. Quackenbush, M. Reich, C. J. Stoeckert, R. C. Taylor, S. C. Trutane, J. Weller, B. Wilhelm, and N. Winegarden (2021). 'Transcriptomics data availability and reusability in the transition from microarray to next-generation sequencing'.

Wright, J. C., J. Mudge, H. Weisser, M. P. Barzine, J. M. Gonzalez, A. Brazma, J. S. Choudhary, and J. Harrow (2016). 'Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow'. *Nature Communications* 7, p. 11778.

REFERENCES

Aebersold, R. (2011). 'Editorial: from data to results'. *Mol. Cell. Proteom.* 10 (11), E111.014787.

Aebersold, R., J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlén, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, and B. Zhang (2018). 'How many human proteoforms are there?' *Nat. Chem. Biol.* 14 (3), pp. 206–214.

Aebersold, R. and M. Mann (2003). 'Mass spectrometry-based proteomics'. *Nature* 422 (6928), pp. 198–207.

Aebersold, R. and M. Mann (2016). 'Mass-spectrometric exploration of proteome structure and function'. *Nature* 537 (7620), pp. 347–355.

Aggarwal, S. and A. K. Yadav (2015). 'False Discovery Rate Estimation in Proteomics'. *Statistical Analysis in Proteomics.* 1362 of the series Methods in Molecular Biology. Vol. 1362. New York, NY, US: Springer, pp. 119–128.

Ahrné, E., L. Molzahn, T. Glatter, and A. Schmidt (2013). 'Critical assessment of proteome-wide label-free absolute abundance estimation strategies'. *Proteomics* 13 (17), pp. 2567–2578.

Ahrné, E., F. Nikitin, F. Lisacek, and M. Müller (2011). 'QuickMod: A tool for open modification spectrum library searches'. *J. Proteome Res.* 10 (7), pp. 2913–2921.

Akers, N. K., E. E. Schadt, and B. Losic (2018). 'STAR Chimeric Post For Rapid Detection of Circular RNA and Fusion Transcripts'. *Bioinformatics* 34 (14).

Al Shweiki, M. R., S. Mönchgesang, P. Majovsky, D. Thieme, D. Trutschel, and W. Hoehenwarter (2017). 'Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance'. *J. Proteome Res.* 16 (4), pp. 1410–1424.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2002). *Molecular Biology of the Cell.* Oxford, UK: Garland Science.

Alves, P., R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly, and H. Tang (2007). 'Advancement in protein inference from shotgun proteomics using peptide detectability'. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 409–420.

Anders, S., P. T. Pyl, and W. Huber (2015). 'HTSeq–a Python framework to work with high-throughput sequencing data'. *Bioinformatics* 31 (2), pp. 166–169.

Anderson, L. and J. Seilhamer (1997). 'A comparison of selected mRNA and protein abundances in human liver'. *Electrophoresis* 18 (3-4), pp. 533–537.

Anscombe, F. J. (1973). 'Graphs in Statistical Analysis'. *Am. Stat.* 27 (1), pp. 17–21.

Apfalter, S., R. Krska, T. Linsinger, A. Oberhauser, W. Kandler, and M. Grasserbauer (1999). 'Interlaboratory comparison study for the determination of halogenated hydrocarbons in water'. *Fresenius J. Anal. Chem.* 364 (7), pp. 660–665.

Arike, L., K. Valgepea, L. Peil, R. Nahku, K. Adamberg, and R. Vilu (2012). 'Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli'. *J. Proteom.* 75 (17), pp. 5437–5448.

Armour, C. D., J. C. Castle, R. Chen, T. Babak, P. Loerch, S. Jackson, J. K. Shah, J. Dey, C. A. Rohl, J. M. Johnson, and C. K. Raymond (2009). 'Digital transcriptome profiling using selective hexamer priming for cDNA synthesis'. *Nat. Methods* 6 (9), pp. 647–649.

Arnold, J. B. (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'.* R package version 4.1.1.

Arvid, S. D. (1997). *Chimie analytique; Trad. et révision scientifique de la 7e éd. américaine par C. Buess-Herman, J. Dauchot-Weymeers et F. Dumont.* Bruxelles, BE: DeBoeck Université.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium'. *Nat. Genet.* 25 (1), pp. 25–29.

Asimov, I. (1989). *The relativity of wrong. Essays on the Solar System and Beyond.* Guernsey, Channels Islands, GB: Oxford University Press. Chap. Beginning with Bone.

Asmann, Y. W., B. M. Necela, K. R. Kalari, A. Hossain, T. R. Baker, J. M. Carr, C. Davis, J. E. Getz, G. Hostetter, X. Li, S. A. McLaughlin, D. C. Radisky, G. P. Schroth, H. E. Cunliffe, E. A. Perez, and E. A. Thompson (2012). 'Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer'. *Cancer Res.* 72 (8), pp. 1921–1928.

Aston, F. W. (1919). 'A positive ray spectrograph'. *Philosophical Magazine* 38 (228), pp. 707–714.

Audain, E., J. Uszkoreit, T. Sachsenberg, J. Pfeuffer, X. Liang, H. Hermjakob, A. Sanchez, M. Eisenacher, K. Reinert, D. L. Tabb, O. Kohlbacher, and Y. Perez-Riverol (2017). 'In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics'. *J. Proteom.* 150, pp. 170–182.

Audi, G. and A. H. Wapstra (1993). 'The 1993 atomic mass evaluation (I) Atomic mass table'. *Nuclear Physics A* 565 (1), pp. 1–65.

Audi, G. and A. H. Wapstra (1995). 'The 1995 update to the atomic mass evaluation'. *Nuclear Physics A* 595 (4), pp. 409–480.

Auer, P. L. and R. W. Doerge (2010). 'Statistical design and analysis of RNA sequencing data'. *Genetics* 185 (2), pp. 405–416.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics.* R package version 2.3.

Bahcall, O. G. (2015). 'Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals'. *Nat. Rev. Genet.* 16 (7), p. 375.

Bantscheff, M., S. Lemeer, M. M. Savitski, and B. Kuster (2012). 'Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present'. *Anal. Bioanal. Chem.* 404 (4), pp. 939–965.

Barbosa-Morais, N. L., M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe (2012). 'The evolutionary landscape of alternative splicing in vertebrate species.' *Science* 338 (6114), pp. 1587–93.

Barshad, G., S. Marom, T. Cohen, and D. Mishmar (2018). 'Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective'. *Trends Genet.* 34 (9), pp. 682–692.

Barzine, M. P., K. Freivalds, J. C. Wright, M. Opmanis, D. Rituma, F. Z. Ghavidel, A. F. Jarnuczak, E. Celms, K. Čerāns, I. Jonassen, L. Lace, J. Antonio Vizcaíno, J. S. Choudhary, A. Brazma, and J. Viksna (2020). 'Using Deep Learning to Extrapolate Protein Expression Measurements'. *Proteomics* 20 (21-22), e2000009.

Bauer, C., R. Cramer, and J. Schuchhardt (2011). 'Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry'. *Data Mining in Proteomics: From Standards to Applications.* Ed. by M. Hamacher, M. Eisenacher, and C. Stephan. Totowa, NJ, US: Humana Press, pp. 341–352.

Begley, C. G. and L. M. Ellis (2012). 'Drug development: Raise standards for preclinical cancer research'. *Nature* 483 (7391), pp. 531–533.

Benhaïm, M. (2017). 'Développements méthodologiques en protéomique quantitatives pour mieux comprendre la biologie évolutive d'espèces non séquencées'. PhD thesis. Université de Strasbourg.

Benjamini, Y. and Y. Hochberg (1995). 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. *J. R. Stat. Soc. B* 57 (1), pp. 289–300.

Bentley, D. R. et al. (2008). 'Accurate whole human genome sequencing using reversible terminator chemistry'. *Nature* 456 (7218), pp. 53–59.

Berg, J. M. and L. Stryer (2002). *Biochemistry.* New York, NY, US: W.H. Freeman.

Bergeron, J. J. M. and M. Hallett (2007). 'Peptides you can count on'. *Nat. Biotechnol.* 25 (1), pp. 61–62.

Bern, M., Y. J. Kil, and C. Becker (2012). 'Byonic: advanced peptide and protein identification software'. *Curr. Protoc.Bioinform.* Chapter 13, Unit13.20.

Bernhardt, O., N. Selevsek, L. Gillet, O. Rinner, P. Picotti, R. Aebersold, and L. Reiter (2014). *Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data.*

Biemann, K. (1988). 'Contributions of mass spectrometry to peptide and protein structure'. *Biomed. Environ. Mass Spectrom.* 16 (1-12), pp. 99–111.

Bilan, V., M. Leutert, P. Nanni, C. Panse, and M. O. Hottiger (2017). 'Combining Higher-Energy Collision Dissociation and Electron-Transfer/Higher-Energy Collision Dissociation Fragmentation in a Product-Dependent Manner Confidently Assigns Proteomewide ADP-Ribose Acceptor Sites'. *Anal. Chem.* 89 (3), pp. 1523–1530.

Bittremieux, W., P. Meysman, W. S. Noble, and K. Laukens (2018). 'Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing'. *J. Proteome Res.* 17 (10), pp. 3463–3474.

Blanco, L., J. A. Mead, and C. Bessant (2009). 'Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets'. *J. Proteome Res.* 8 (4), pp. 1782–1791.

Blein-Nicolas, M. and M. Zivy (2016). 'Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics'. *BBA* 1864 (8), pp. 883–895.

Bodzon-Kulakowska, A., A. Bierczynska-Krzysik, T. Dylag, A. Drabik, P. Suder, M. Noga, J. Jarzebinska, and J. Silberring (2007). 'Methods for samples preparation in proteomic research'. *J. Chromatogr. B* 849 (1-2), pp. 1–31.

Boguszewska, K., M. Szewczuk, J. Kaźmierczak-Barańska, and B. T. Karwowski (2020). 'The Similarities between Human Mitochondria and Bacteria in the Context of Structure, Genome, and Base Excision Repair System'. *Molecules* 25 (12).

Bonaldo, M. F., G. Lennon, and M. B. Soares (1996). 'Normalization and subtraction: two approaches to facilitate gene discovery'. *Genome Res.* 6 (9), pp. 791–806.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). 'A Training Algorithm for Optimal Margin Classifiers'. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory.* COLT '92. Pittsburgh, Pennsylvania, USA: ACM, pp. 144–152.

Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock (2004). 'GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes'. *Bioinformatics* 20 (18), pp. 3710–3715.

Braisted, J. C., S. Kuntumalla, C. Vogel, E. M. Marcotte, A. R. Rodrigues, R. Wang, S.-T. Huang, E. S. Ferlanti, A. I. Saeed, R. D. Fleischmann, S. N. Peterson, and R. Pieper (2008). 'The APEX

Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results'. *BMC Bioinf.* 9, p. 529.

Bratic, A., P. Clemente, J. Calvo-Garrido, C. Maffezzini, A. Felser, R. Wibom, A. Wedell, C. Freyer, and A. Wredenberg (2016). 'Mitochondrial Polyadenylation Is a One-Step Process Required for mRNA Integrity and tRNA Maturation'. *PLOS Genet.* 12 (5), e1006028.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, and H. Kaessmann (2011). 'The evolution of gene expression levels in mammalian organs'. *Nature* 478 (7369), pp. 343–348.

Brosch, M., L. Yu, T. Hubbard, and J. Choudhary (2009). 'Accurate and sensitive peptide identification with Mascot Percolator'. *J. Proteome Res.* 8 (6), pp. 3176–3181.

Brosh, M. (2009). 'Development of computational methods for analysing proteomic data for genome annotation'. PhD thesis. University of Cambridge.

Brown, S. D., L. A. Raeburn, and R. A. Holt (2015). 'Profiling tissue-resident T cell repertoires by RNA sequencing'. *Genome Med.* 7, p. 125.

Bruce, C., K. Stone, E. Gulcicek, and K. Williams (2013). 'Proteomics and the analysis of proteomic data: 2013 overview of current protein-profiling technologies'. *Curr. Protoc. Bioinformatics* S41 (13.21), pp. 1–17.

Bubis, J. A., L. I. Levitsky, M. V. Ivanov, I. A. Tarasova, and M. V. Gorshkov (2017). 'Comparative evaluation of label-free quantification methods for shotgun proteomics'. *Rapid Commun. Mass Spectrom.* 31 (7), pp. 606–612.

Bulyk, M. L. (2007). 'Protein binding microarrays for the characterization of DNA-protein interactions'. *Adv. Biochem. Eng. Biotechnol.* 104, pp. 65–85.

Bulyk, M. L., P. L. F. Johnson, and G. M. Church (2002). 'Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors'. *Nucleic Acids Res.* 30 (5), pp. 1255–1261.

Bumgarner, R. (2013). 'Overview of DNA microarrays: types, applications, and their future'. *Curr. Protoc. Mol. Biol.* Chapter 22, Unit 22.1.

Bussotti, G., T. Leonardi, M. B. Clark, T. R. Mercer, J. Crawford, L. Malquori, C. Notredame, M. E. Dinger, J. S. Mattick, and A. J. Enright (2016). 'Improved definition of the mouse transcriptome via targeted RNA sequencing'. *Genome Res.* 26 (5), pp. 705–716.

Bythell, B. J., P. Maître, and B. Paizs (2010). 'Cyclization and rearrangement reactions of a(n) fragment ions of protonated peptides'. *J. Am. Chem. Soc* 132 (42), pp. 14766–14779.

Callen, J.-C. (2005). *Biologie cellulaire — Des molécules aux organismes (2eme edition).* Paris, FR: Dunod.

Cantalupo, P. G., J. P. Katz, and J. M. Pipas (2015). 'HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18'. *J. Virol.* 89 (8), pp. 4051–4057.

Canterbury, J. D., G. E. Merrihew, M. J. MacCoss, D. R. Goodlett, and S. A. Shaffer (2014). 'Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics'. *J. Am. Soc. Mass Spectrom.* 25 (12), pp. 2048–2059.

Cappadona, S., P. R. Baker, P. R. Cutillas, A. J. R. Heck, and B. van Breukelen (2012). 'Current challenges in software solutions for mass spectrometry-based quantitative proteomics'. *Amino Acids* 43 (3), pp. 1087–1108.

Carlson, M. (2019). *org.Hs.eg.db: Genome wide annotation for Human.* R package version 3.8.2.

Castle, J. C., C. D. Armour, M. Löwer, D. Haynor, M. Biery, H. Bouzek, R. Chen, S. Jackson, J. M. Johnson, C. A. Rohl, and C. K. Raymond (2010). 'Digital Genome-Wide ncRNA Expression, Including SnoRNAs, across 11 Human Tissues Using PolyA-Neutral Amplification'. *PLOS ONE* 5 (7), pp. 1–9.

Catherman, A. D., O. S. Skinner, and N. L. Kelleher (2014). ‘Top Down proteomics: facts and perspectives’. *BBRC* 445 (4), pp. 683–693.

Cavalli, F. M. G., R. Bourgon, W. Huber, J. M. Vaquerizas, and N. M. Luscombe (2011). ‘SpeCond: a method to detect condition-specific gene expression’. *Genome Biol.* 12 (10), R101.

Chambers, M. C., B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick (2012). ‘A cross-platform toolkit for mass spectrometry and proteomics’. *Nat. Biotechnol.* 30 (10), pp. 918–920.

Chang, W. (2014). *extrafont: Tools for using fonts.* R package version 0.17.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2019). *shiny: Web Application Framework for R.* R package version 1.3.2.

Chapman, J. D., D. R. Goodlett, and C. D. Masselon (2014). ‘Multiplexed and data-independent tandem mass spectrometry for global proteome profiling’. *Mass Spectrom. Rev.* 33 (6), pp. 452–470.

Chen, C., J. Hou, J. J. Tanner, and J. Cheng (2020). ‘Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis’. *Int. J. Mol. Sci.* 21 (8).

Chen, G., B. Ning, and T. Shi (2019). ‘Single-Cell RNA-Seq Technologies and Related Computational Data Analysis’. *Front. Genet.* 10, p. 317.

Chen, G., T. G. Gharib, C.-C. Huang, J. M. G. Taylor, D. E. Misek, S. L. R. Kardia, T. J. Giordano, M. D. Iannettoni, M. B. Orringer, S. M. Hanash, and D. G. Beer (2002). ‘Discordant protein and mRNA expression in lung adenocarcinomas’. *Mol. Cell. Proteom.* 1 (4), pp. 304–313.

Chen, H. (2018). *VennDiagram: Generate High-Resolution Venn and Euler Plots.* R package version 1.6.20.

Chen, J., K. Sathiyamoorthy, X. Zhang, S. Schaller, B. E. Perez White, T. S. Jardetzky, and R. Longnecker (2018). ‘Ephrin receptor A2 is a functional entry receptor for Epstein-Barr virus’. *Nat. Microbiol.* 3 (2), pp. 172–180.

Chen, X., S. Wei, Y. Ji, X. Guo, and F. Yang (2015). ‘Quantitative proteomics using SILAC: Principles, applications, and developments’. *Proteomics* 15 (18), pp. 3175–3192.

Chen, Y., F. Wang, F. Xu, and T. Yang (2016). ‘Mass Spectrometry-Based Protein Quantification’. *Modern Proteomics – Sample Preparation, Analysis and Practical Applications.* Adv. Exp. Med. Biol. Cham, CH: Springer, Cham, pp. 255–279.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras (2005). ‘Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution’. *Science* 308 (5725), pp. 1149–1154.

Chi, H., C. Liu, H. Yang, W.-F. Zeng, L. Wu, W.-J. Zhou, R.-M. Wang, X.-N. Niu, Y.-H. Ding, Y. Zhang, Z.-W. Wang, Z.-L. Chen, R.-X. Sun, T. Liu, G.-M. Tan, M.-Q. Dong, P. Xu, P.-H. Zhang, and S.-M. He (2018). ‘Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine’. *Nat. Biotechnol.*

Choi, H., D. Ghosh, and A. I. Nesvizhskii (2008). ‘Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling’. *J. Proteome Res.* 7 (1), pp. 286–292.

Chrominski, K. and M. Tkacz (2015). ‘Comparison of High-Level Microarray Analysis Methods in the Context of Result Consistency’. *PLOS ONE* 10 (6), e0128845.

Clark, E. L., S. J. Bush, M. E. B. McCulloch, I. L. Farquhar, R. Young, L. Lefevre, C. Pridans, H. Tsang, C. Wu, C. Afrasiabi, M. Watson, C. B. Whitelaw, T. C. Freeman, K. M. Summers, A. L. Archibald, and D. A. Hume (2017). 'A high resolution atlas of gene expression in the domestic sheep (Ovis aries)'. *PLOS Genet.* 13 (9), e1006997.

Cock, P. J. A., C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice (2010). 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants'. *Nucleic Acids Res.* 38 (6), pp. 1767–1771.

Codrea, M. C. and S. Nahnsen (2016). 'Platforms and Pipelines for Proteomics Data Analysis and Management'. *Modern Proteomics – Sample Preparation, Analysis and Practical Applications.* Adv. Exp. Med. Biol. Cham, CH: Springer, Cham, pp. 203–215.

Coiera, E., E. Ammenwerth, A. Georgiou, and F. Magrabi (2018). 'Does health informatics have a replication crisis?' *JAMIA* 25 (8), pp. 963–968.

Conesa, A., 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 (2016). 'A survey of best practices for RNA-seq data analysis'. *Genome Biol.* 17, p. 13.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2009). *Introduction to Algorithms.* Cambridge, MA, US: MIT Press.

Corpas, M., R. Jimenez, S. J. Carbon, A. García, L. Garcia, T. Goldberg, J. Gomez, A. Kalderimis, S. E. Lewis, I. Mulvany, A. Pawlik, F. Rowland, G. Salazar, F. Schreiber, I. Sillitoe, W. H. Spooner, A. S. Thanki, J. M. Villaveces, G. Yachdav, and H. Hermjakob (2014). 'BioJS: an open source standard for biological visualisation - its status in 2014'. *F1000Research* 3, p. 55.

Cottrell, J. (2013). *Does protein FDR have any meaning?* http://www.matrixscience.com/blog/does-protein-fdr-have-any-meaning.html. Accessed: 2018-10-30.

Cottrell, J. S. (2011). 'Protein identification using MS/MS data'. *J. Proteom.* 74 (10), pp. 1842–1851.

Cox, J. and M. Mann (2008). 'MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification'. *Nat. Biotechnol.* 26 (12), pp. 1367–1372.

Cox, J. and M. Mann (2011). 'Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology'. *Annu. Rev. Biochem.* 80 (1), pp. 273–299.

Cox, J., N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann (2011). 'Andromeda: a peptide search engine integrated into the MaxQuant environment'. *J. Proteome Res.* 10 (4), pp. 1794–1805.

Cozzolino, F., A. Landolfi, I. Iacobucci, V. Monaco, M. Caterino, S. Celentano, C. Zuccato, E. Cattaneo, and M. Monti (2020). 'New label-free methods for protein relative quantification applied to the investigation of an animal model of Huntington Disease'. *PLOS ONE* 15 (9), e0238037.

Cresko Lab (2017). *RNA-seqlopedia.* URL: http://rnaseq.uoregon.edu.

Crick, F. (1958). 'On protein synthesis'. *Symposia of the Society for Experimental Biology* 12, pp. 138–163.

Crick, F. (1970). 'Central Dogma of Molecular Biology'. *Nature* 227 (5258), pp. 561–563.

Cuperlovic-Culf, M., N. Belacel, A. S. Culf, and R. J. Ouellette (2006). 'Microarray analysis of alternative splicing'. *OMICS* 10 (3), pp. 344–357.

Danielsson, F., T. James, D. Gomez-Cabrero, and M. Huss (2015). 'Assessing the consistency of public human tissue RNA-seq data sets'. *Briefings Bioinf.* 16 (6), pp. 941–949.

Dapas, M., M. Kandpal, Y. Bi, and R. V. Davuluri (2017). 'Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms'. *Briefings Bioinf.* 18 (2), pp. 260–269.

Dar, R. D., B. S. Razooky, L. S. Weinberger, C. D. Cox, and M. L. Simpson (2015). 'The Low Noise Limit in Gene Expression'. *PLOS ONE* 10 (10), e0140969.

Darnell Jr, J. E. (2013). 'Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture'. *RNA* 19 (4), pp. 443–460.

Dasari, S., M. C. Chambers, M. A. Martinez, K. L. Carpenter, A.-J. L. Ham, L. J. Vega-Montoto, and D. L. Tabb (2012). 'Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment'. *J. Proteome Res.* 11 (3), pp. 1686–1695.

Davidson, V. L. and D. B. Sittman (1999). *Biochemistry.* Philadelphia, PA ,US: Lippincott Williams & Wilkins.

Davies, L. and U. Gather (1993). 'The Identification of Multiple Outliers'. *J. Am. Stat. Assoc.* 88 (423), pp. 782–792.

De Simone, M., A. Arrigoni, G. Rossetti, P. Gruarin, V. Ranzani, C. Politano, R. J. P. Bonnal, E. Provasi, M. L. Sarnicola, I. Panzeri, M. Moro, M. Crosti, S. Mazzara, V. Vaira, S. Bosari, A. Palleschi, L. Santambrogio, G. Bovo, N. Zucchini, M. Totis, L. Gianotti, G. Cesana, R. A. Perego, N. Maroni, A. Pisani Ceretti, E. Opocher, R. De Francesco, J. Geginat, H. G. Stunnenberg, S. Abrignani, and M. Pagani (2016). 'Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells'. *Immunity* 45 (5), pp. 1135–1147.

De Siqueira Santos, S., D. Y. Takahashi, A. Nakata, and A. Fujita (2014). 'A comparative study of statistical methods used to identify dependencies between gene expression signals'. *Briefings Bioinf.* 15 (6), pp. 906–918.

Delacre, M., D. Lakens, and C. Leys (2017). *Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test (in press for the International Review of Social Psychology).*

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *J. R. Stat. Soc. B* 39 (1), pp. 1–38.

Derrick, B., D. Toher, and P. White (2016). 'Why Welch's test is Type I error robust'. *TQMP* 12.1, pp. 30–38.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigó (2012). 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression'. *Genome Res.* 22 (9), pp. 1775–1789.

Desiere, F., E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, and R. Aebersold (2006). 'The PeptideAtlas project'. *Nucleic Acids Res.* 34 (Suppl. 1), p. D655.

Deutsch, E. W., Z. Sun, D. Campbell, U. Kusebauch, C. S. Chu, L. Mendoza, D. Shteynberg, G. S. Omenn, and R. L. Moritz (2015). 'State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet'. *J. Proteome Res.* 14 (9), pp. 3461–3473.

Diedrich, J. K., A. F. M. Pinto, and J. R. Yates 3rd (2013). 'Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot'. *J. Am. Soc. Mass Spectrom.* 24 (11), pp. 1690–1699.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and French StatOmique Consortium (2013). 'A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis'. *Briefings Bioinf.* 14 (6), pp. 671–683.

Do, C. B. and S. Batzoglou (2008). 'What is the expectation maximization algorithm?' *Nat. Biotechnol.* 26 (8), pp. 897–899.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013). 'STAR: ultrafast universal RNA-seq aligner'. *Bioinformatics* 29 (1), pp. 15–21.

Domon, B. and R. Aebersold (2006). 'Mass spectrometry and protein analysis'. *Science* 312 (5771), pp. 212–217.

Domon, B. and R. Aebersold (2010). ‘Options and considerations when selecting a quantitative proteomics strategy’. *Nat. Biotechnol.* 28 (7), pp. 710–721.

Dorfer, V., P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler, and K. Mechtler (2014). ‘MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra’. *J. Proteome Res.* 13 (8), pp. 3679–3684.

Dowle, M. and A. Srinivasan (2019). *data.table: Extension of ‘data.frame‘*. R package version 1.12.2.

Doyle, A. C. (1892). ‘The Adventure of the Copper Beechees’. *The Strand Magazine.*

Duarte, J. G. and J. M. Blackburn (2017). ‘Advances in the development of human protein microarrays’. *Expert Rev. Proteomics* 14 (7), pp. 627–641.

Dupree, E. J., M. Jayathirtha, H. Yorkey, M. Mihasan, B. A. Petre, and C. C. Darie (2020). ‘A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of this Field’. *Proteomes* 8 (3).

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005). ‘BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis’. *Bioinformatics* 21 (16), pp. 3439–3440.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). ‘Empirical Bayes Analysis of a Microarray Experiment’. *J. Am. Stat. Assoc.* 96 (456), pp. 1151–1160.

Egger, M., G. D. Smith, and A. N. Phillips (1997). ‘Meta-analysis: principles and procedures’. *BMJ* 315 (7121), pp. 1533–1537.

Elias, J. E. and S. P. Gygi (2007). ‘Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry’. *Nat. Methods* 4 (3), pp. 207–214.

Elias, J. E. and S. P. Gygi (2010). ‘Target-decoy search strategy for mass spectrometry-based proteomics’. *Methods Mol. Biol.* 604, pp. 55–71.

Eng, J. K., A. L. McCormack, and J. R. Yates (1994). ‘An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database’. *J. Am. Soc. Mass Spectrom.* 5 (11), pp. 976–989.

Eng, J. K., B. C. Searle, K. R. Clauser, and D. L. Tabb (2011). ‘A face in the crowd: recognizing peptides through database search’. *Mol. Cell. Proteom.* 10 (11), R111.009522.

Engström, P. G., T. Steijger, B. Sipos, G. R. Grant, A. Kahles, G. Rätsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigó, P. Bertone, and RGASP Consortium (2013). ‘Systematic evaluation of spliced alignment programs for RNA-seq data’. *Nat. Methods* 10 (12), pp. 1185–1191.

Ensembl Blog (2011). *Human BodyMap 2.0 data from Illumina.* URL: http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/ (visited on 03/11/2013).

Eraslan, B., D. Wang, M. Gusic, H. Prokisch, B. M. Hallström, M. Uhlén, A. Asplund, F. Pontén, T. Wieland, T. Hopf, H. Hahne, B. Kuster, and J. Gagneur (2019). ‘Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues’. *Mol. Syst. Biol.* 15 (2), e8513.

Eriksson, J. and D. Fenyö (2007). ‘Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs’. *Nat. Biotechnol.* 25 (6), pp. 651–655.

Esteve-Codina, A., O. Arpi, M. Martinez-García, E. Pineda, M. Mallo, M. Gut, C. Carrato, A. Rovira, R. Lopez, A. Tortosa, M. Dabad, S. Del Barco, S. Heath, S. Bagué, T. Ribalta, F. Alameda, N. de la Iglesia, C. Balaña, and GLIOCAT Group (2017). ‘A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples’. *PLOS ONE* 12 (1), e0170632.

Everaert, C., M. Luypaert, J. L. V. Maag, Q. X. Cheng, M. E. Dinger, J. Hellemans, and P. Mestdagh (2017). ‘Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data’. *Sci. Rep.* 7, p. 1559.

Ezkurdia, I., J. Vázquez, A. Valencia, and M. Tress (2014). ‘Analyzing the first drafts of the human proteome’. *J. Proteome Res.* 13 (8), pp. 3854–3855.

Fagerberg, L., B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C. A.-K. Szigyarto, M. Skogs, J. O. Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J. M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, A. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Ponten, and M. Uhlén (2014). 'Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics'. *Mol. Cell. Proteom.* 13 (2), pp. 397–406.

Fagerland, M. W. (2012). 't-tests, non-parametric tests, and large studies–a paradox of statistical practice?' *BMC Med. Res. Methodol.* 12, p. 78.

FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. (2014). 'A promoter-level mammalian expression atlas'. *Nature* 507 (7493), pp. 462–470.

Al-Faresi, R. A. Z., R. N. Lightowlers, and Z. M. A. Chrzanowska-Lightowlers (2019). 'Mammalian mitochondrial translation - revealing consequences of divergent evolution'. *Biochem. Soc. Trans.* 47 (5), pp. 1429–1436.

Fatovich, D. M. and M. Phillips (2017). 'The probability of probability and research truths'. *EMA* 29 (2), pp. 242–244.

Feist, P. and A. B. Hummon (2015). 'Proteomic challenges: sample preparation techniques for microgram-quantity protein analysis from biological samples'. *Int. J. Mol. Sci.* 16 (2), pp. 3537–3563.

Felsenstein, J. and J. Felenstein (2004). *Inferring phylogenies.* Vol. 2. Sunderland, MA, US: Sinauer associates.

Feng, J., D. Q. Naiman, and B. Cooper (2007). 'Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data'. *Anal. Chem.* 79 (10), pp. 3901–3911.

Field, A., J. Miles, and Z. Field (2012). *Discovering Statistics Using R.* London, UK: SAGE Publications.

Fischer, B., V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann (2005). 'NovoHMM: a hidden Markov model for de novo peptide sequencing'. *Anal. Chem.* 77 (22), pp. 7265–7273.

Florea, L., L. Song, and S. L. Salzberg (2013). 'Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues'. *F1000Research* 2.

Fonseca, N. A., R. Petryszak, J. Marioni, and A. Brazma (2014). 'iRAP - an integrated RNA-seq Analysis Pipeline'. *bioRxiv* (005991).

Fonseca, N. A., J. Marioni, and A. Brazma (2014). 'RNA-Seq gene profiling–a systematic empirical comparison'. *PLOS ONE* 9 (9), e107026.

Frank, A. M. (2009). 'Predicting intensity ranks of peptide fragment ions'. *J. Proteome Res.* 8 (5), pp. 2226–2240.

Frankish, A., B. Uszczynska, G. R. S. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo, and J. Harrow (2015). 'Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction'. *BMC Genom.* 16 Suppl 8, S2.

Franks, A., E. Airoldi, and N. Slavov (2017). 'Post-transcriptional regulation across human tissues'. *PLOS Comput. Biol.* 13 (5), e1005535.

Freeman, T. C., A. Ivens, J. K. Baillie, D. Beraldi, M. W. Barnett, D. Dorward, A. Downing, L. Fairbairn, R. Kapetanovic, S. Raza, A. Tomoiu, R. Alberio, C. Wu, A. I. Su, K. M. Summers, C. K. Tuggle, A. L. Archibald, and D. A. Hume (2012). 'A gene expression atlas of the domestic pig'. *BMC Biol.* 10, p. 90.

Freiberg, J. A., Y. Le Breton, B. Q. Tran, A. J. Scott, J. M. Harro, R. K. Ernst, Y. A. Goo, E. F. Mongodin, D. R. Goodlett, K. S. McIver, and M. E. Shirtliff (2016). 'Global Analysis and Comparison of the Transcriptomes and Proteomes of Group AStreptococcusBiofilms'. *mSystems* 1 (6).

Frewen, B. and M. J. MacCoss (2007). 'Using BiblioSpec for creating and searching tandem MS peptide libraries'. *Curr. Protoc. Bioinformatics* Chapter 13, Unit 13.7.

Fusaro, V. A., D. R. Mani, J. P. Mesirov, and S. A. Carr (2009). 'Prediction of high-responding peptides for targeted protein assays by mass spectrometry'. *Nat. Biotechnol.* 27 (2), pp. 190–198.

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation.* Hoboken, NJ, US: Wiley.

Gagnon-Bartsch, J. A. and T. P. Speed (2012). 'Using control genes to correct for unwanted variation in microarray data'. *Biostatistics* 13 (3), pp. 539–552.

Gallien, S., E. Duriez, C. Crone, M. Kellmann, T. Moehring, and B. Domon (2012). 'Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer'. *Mol. Cell. Proteom.* 11 (12), pp. 1709–1723.

Garalde, D. R., E. A. Snell, D. Jachimowicz, A. J. Heron, M. Bruce, J. Lloyd, A. Warland, N. Pantic, T. Admassu, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, B. Sipos, S. Young, S. Juul, J. Clarke, and D. J. Turner (2016). 'Highly parallel direct RNA sequencing on an array of nanopores'. *bioRxiv* (068809).

Gardner, M. L. and M. A. Freitas (2020). 'Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-up Proteomics'.

Gerrard, D. T., A. A. Berry, R. E. Jennings, K. Piper Hanley, N. Bobola, and N. A. Hanley (2016). 'An integrative transcriptomic atlas of organogenesis in human embryos'. *eLife* 5.

Gerster, S., E. Qeli, C. H. Ahrens, and P. Bühlmann (2010). 'Protein and gene model inference based on statistical modeling in k-partite graphs'. *PNAS* 107 (27), pp. 12101–12106.

Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P.-Z. Wen, H. Brewer, K. Weitz, D. G. Camp II, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, and A. J. Lusis (2011). 'Comparative Analysis of Proteome and Transcriptome Variation in Mouse'. *PLOS Genet.* 7 (6), e1001393.

Giansanti, P., L. Tsiatsiani, T. Y. Low, and A. J. R. Heck (2016). 'Six alternative proteases for mass spectrometry-based proteomics beyond trypsin'. *Nat. Protoc.* 11 (5), pp. 993–1006.

Gibson, G. (2015). 'Human genetics. GTEx detects genetic effects'. *Science* 348 (6235), pp. 640–641.

Gillet, L. C., P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold (2012). 'Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis'. *Mol. Cell. Proteom.* 11 (6), O111.016717.

Glenn Begley, C. and J. P. A. Ioannidis (2015). 'Reproducibility in Science'. *Circ. Res.* 116 (1), pp. 116–126.

Gonnelli, G., M. Stock, J. Verwaeren, D. Maddelein, B. De Baets, L. Martens, and S. Degroeve (2015). 'A decoy-free approach to the identification of peptides'. *J. Proteome Res.* 14 (4), pp. 1792–1798.

Gonzàlez-Porta, M. (2014). 'RNA sequencing for the study of splicing'. PhD thesis. University of Cambridge.

Gonzàlez-Porta, M., A. Frankish, J. Rung, J. Harrow, and A. Brazma (2013). 'Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene'. *Genome Biol.* 14 (7), R70.

Goodman, S. N., D. Fanelli, and J. P. A. Ioannidis (2016). 'What does research reproducibility mean?' *Sci. Transl. Med.* 8 (341), 341ps12.

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). 'Coming of age: ten years of next-generation sequencing technologies'. *Nat. Rev. Genet.* 17 (6), pp. 333–351.

Granholm, V., S. Kim, J. C. F. Navarro, E. Sjölund, R. D. Smith, and L. Käll (2014). 'Fast and accurate database searches with MS-GF+Percolator'. *J. Proteome Res.* 13 (2), pp. 890–897.

Gremel, G., A. Wanders, J. Cedernaes, L. Fagerberg, B. Hallström, K. Edlund, E. Sjöstedt, M. Uhlén, and F. Pontén (2015). 'The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling'. *J. Gastroenterol.* 50 (1), pp. 46–57.

Griss, J. (2016). 'Spectral library searching in proteomics'. *Proteomics* 16 (5), pp. 729–740.

Gry, M., R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson (2009). 'Correlations between RNA and protein expression profiles in 23 human cell lines'. *BMC Genom.* 10, p. 365.

GTEx Consortium (2013). 'The Genotype-Tissue Expression (GTEx) project'. *Nat. Genet.* 45 (6), pp. 580–585.

GTEx Consortium (2015). 'Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans'. *Science* 348 (6235), pp. 648–660.

Guillaumot, N. (2017). 'Nouvelles applications et opportunités en protéomique'. PhD thesis. Université de Strasbourg.

Guinand, B., A. Topchy, K. S. Page, M. K. Burnham-Curtis, W. F. Punch, and K. T. Scribner (2002). 'Comparisons of likelihood and machine learning methods of individual classification'. *J. Hered.* 93 (4), pp. 260–269.

Gunderson, K. L., F. J. Steemers, H. Ren, P. Ng, L. Zhou, C. Tsan, W. Chang, D. Bullis, J. Musmacker, C. King, L. L. Lebruska, D. Barker, A. Oliphant, K. M. Kuhn, and R. Shen (2006). 'Whole-genome genotyping'. *Meth. Enzymol.* 410, pp. 359–376.

Guo, Y., Y. Dai, H. Yu, S. Zhao, D. C. Samuels, and Y. Shyr (2017). 'Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis'. *Genomics* 109 (2), pp. 83–90.

Guthals, A., K. R. Clauser, A. M. Frank, and N. Bandeira (2013). 'Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides'. *J. Proteome Res.* 12 (6), pp. 2846–2857.

Gutstein, H. B., J. S. Morris, S. P. Annangudi, and J. V. Sweedler (2008). 'Microproteomics: analysis of protein diversity in small samples'. *Mass Spectrom. Rev.* 27 (4), pp. 316–330.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold (1999). 'Quantitative analysis of complex protein mixtures using isotope-coded affinity tags'. *Nat. Biotechnol.* 17 (10), pp. 994–999.

Gygi, S. P., Y. Rochon, B. R. Franza, and R. Aebersold (1999). 'Correlation between protein and mRNA abundance in yeast'. *Mol. Cell. Biol.* 19 (3), pp. 1720–1730.

Haag, A. M. (2016). 'Mass Analyzers and Mass Spectrometers'. *Modern Proteomics – Sample Preparation, Analysis and Practical Applications.* Adv. Exp. Med. Biol. Cham, CH: Springer, Cham, pp. 157–169.

Hajnsdorf, E. and V. R. Kaberdin (2018). 'RNA polyadenylation and its consequences in prokaryotes'. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 373 (1762).

Hall, D. A., J. Ptacek, and M. Snyder (2007). 'Protein microarray technology'. *Mech. Ageing Dev.* 128 (1), pp. 161–167.

Halloran, J. T., H. Zhang, K. Kara, C. Renggli, M. The, C. Zhang, D. M. Rocke, L. Käll, and W. S. Noble (2019). 'Speeding Up Percolator'. *J. Proteome Res.* 18 (9), pp. 3353–3359.

Hampel, F. R. (1971). 'A General Qualitative Definition of Robustness'. *Ann. Math. Statist.* 42 (6), pp. 1887–1896.

Hampel, F. R. (1974). 'The Influence Curve and its Role in Robust Estimation'. *J. Am. Stat. Assoc.* 69 (346), pp. 383–393.

Han, H., Y. Xia, and S. A. McLuckey (2007). 'Ion trap collisional activation of c and z* ions formed via gas-phase ion/ion electron-transfer dissociation'. *J. Proteome Res.* 6 (8), pp. 3062–3069.

Hansen, K. D., S. E. Brenner, and S. Dudoit (2010). 'Biases in Illumina transcriptome sequencing caused by random hexamer priming'. *Nucleic Acids Res.* 38 (12), e131.

Hansen, K. D., R. A. Irizarry, and Z. Wu (2012). ‘Removing technical variability in RNA-seq data using conditional quantile normalization’. *Biostatistics* 13 (2), pp. 204–216.

Harbers, M. (2008). ‘The current status of cDNA cloning’. *Genomics* 91 (3), pp. 232–242.

Hardwick, S. A., I. W. Deveson, and T. R. Mercer (2017). ‘Reference standards for next-generation sequencing’. *Nat. Rev. Genet.* 18 (6).

Hawrylycz, M. J., E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones (2012). ‘An anatomically comprehensive atlas of the adult human brain transcriptome’. *Nature* 489 (7416), pp. 391–399.

He, Z., T. Huang, C. Zhao, and B. Teng (2016). ‘Protein Inference’. *Adv. Exp. Med. Biol.* 919, pp. 237–242.

Hebenstreit, D., M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann (2011). ‘RNA sequencing reveals two major classes of gene expression levels in metazoan cells’. *Mol. Syst. Biol.* 7 (1).

Higgs, R. E., J. P. Butler, B. Han, and M. D. Knierman (2013). ‘Quantitative Proteomics via High Resolution MS Quantification: Capabilities and Limitations’. *Int. J. Proteomics* 2013, p. 674282.

Hilbrig, F. and R. Freitag (2003). ‘Protein purification by affinity precipitation’. *J. Chromatogr. B* 790 (1-2), pp. 79–90.

Hillen, H. S., D. Temiakov, and P. Cramer (2018). ‘Structural basis of mitochondrial transcription’. *Nat. Struct. Mol. Biol.* 25 (9), pp. 754–765.

Hochbaum, D. S. (1997). ‘Approximation Algorithms for NP-hard Problems’. Ed. by D. S. Hochbaum. Boston, MA, US: PWS Publishing Co. Chap. Approximating Covering and Packing Problems: Set Cover, Vertex Cover, Independent Set, and Related Problems, pp. 94–143.

Hoheisel, J. D. (2006). ‘Microarray technology: beyond transcript profiling and genotype analysis’. *Nat. Rev. Genet.* 7 (3), pp. 200–210.

Holman, J. D., D. L. Tabb, and P. Mallick (2014). ‘Employing ProteoWizard to Convert Raw Mass Spectrometry Data’. *Curr. Protoc. Bioinformatics* 46 (13.24), pp. 1–9.

Hou, Z., P. Jiang, S. A. Swanson, A. L. Elwell, B. K. S. Nguyen, J. M. Bolin, R. Stewart, and J. A. Thomson (2015). ‘A cost-effective RNA sequencing protocol for large-scale gene expression studies’. *Sci. Rep.* 5, p. 9570.

Hsu, J.-L., S.-Y. Huang, N.-H. Chow, and S.-H. Chen (2003). ‘Stable-isotope dimethyl labeling for quantitative proteomics’. *Anal. Chem.* 75 (24), pp. 6843–6852.

Hu, A., W. S. Noble, and A. Wolf-Yadlin (2016). ‘Technical advances in proteomics: new developments in data-independent acquisition’. *F1000Research* 5.

Hu, Q., R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks (2005). ‘The Orbitrap: a new mass spectrometer’. *J. Mass Spectrom.* 40 (4), pp. 430–443.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). ‘Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists’. *Nucleic Acids Res.* 37 (1), pp. 1–13.

Huang, T., J. Wang, W. Yu, and Z. He (2012). ‘Protein inference: a review’. *Briefings Bioinf.* 13 (5), pp. 586–614.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan (2015). 'Orchestrating high-throughput genomic analysis with Bioconductor'. *Nat. Methods* 12 (2), pp. 115–121.

Ilicic, T., J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann (2016). 'Classification of low quality cells from single-cell RNA-seq data'. *Genome Biol.* 17, p. 29.

Illumina (2016). *Illumina Sequencing by Synthesis.* URL: https://youtu.be/fCd6B5HRaZ8.

International Human Genome Sequencing Consortium (2004). 'Finishing the euchromatic sequence of the human genome'. *Nature* 431 (7011), pp. 931–945.

Irizarry, R. A., C. Wang, Y. Zhou, and T. P. Speed (2009). 'Gene set enrichment analysis made simple'. *Stat. Methods Med. Res.* 18 (6), pp. 565–575.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martínez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu (2005). 'Multiple-laboratory comparison of microarray platforms'. *Nat. Methods* 2 (5), pp. 345–350.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann (2005). 'Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein'. *Mol. Cell. Proteom.* 4 (9), pp. 1265–1272.

Iwakiri, J., M. Hamada, and K. Asai (2016). 'Bioinformatics tools for lncRNA research'. *BBA* 1859 (1), pp. 23–30.

Jaccard, P. (1901). 'Etude de la distribution florale dans une portion des Alpes et du Jura'. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (142), pp. 547–579.

Jahn, N. (2018). *europepmc: R Interface to the Europe PubMed Central RESTful Web Service.* R package version 0.3.

Jänes, J., F. Hu, A. Lewin, and E. Turro (2015). 'A comparative study of RNA-seq analysis strategies'. *Briefings Bioinf.* 16 (6), pp. 932–940.

Jaskowiak, P. A., R. J. G. B. Campello, and I. G. Costa (2014). 'On the selection of appropriate distances for gene expression data clustering'. *BMC Bioinf.* 15 Suppl. 2, S2.

Jeong, K., S. Kim, and P. A. Pevzner (2013). 'UniNovo: a universal tool for de novo peptide sequencing'. *Bioinformatics* 29 (16), pp. 1953–1962.

Jiang, C., Y. Li, Z. Zhao, J. Lu, H. Chen, N. Ding, G. Wang, J. Xu, and X. Li (2016). 'Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs'. *Oncotarget* 7 (6), pp. 7120–7133.

Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver (2011). 'Synthetic spike-in standards for RNA-seq experiments'. *Genome Res.* 21 (9), pp. 1543–1551.

Jiménez-Lozano, N., J. Segura, J. R. Macías, J. Vega, and J. M. Carazo (2012). 'Integrating human and murine anatomical gene expression data for improved comparisons'. *Bioinformatics* 28 (3), pp. 397–402.

Johnson, N. L., A. W. Kemp, and S. Kotz (2005). *Univariate Discrete Distributions.* Hoboken, NJ, US: Wiley.

Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson (1987). 'Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine'. *Anal. Chem.* 59 (21), pp. 2621–2625.

Johnson, W. E., C. Li, and A. Rabinovic (2007). 'Adjusting batch effects in microarray expression data using empirical Bayes methods'. *Biostatistics* 8 (1), pp. 118–127.

Jovanovic, M., M. S. Rooney, P. Mertins, D. Przybylski, N. Chevrier, R. Satija, E. H. Rodriguez, A. P. Fields, S. Schwartz, R. Raychowdhury, M. R. Mumbach, T. Eisenhaure, M. Rabani, D. Gennert, D. Lu, T. Delorey, J. S. Weissman, S. A. Carr, N. Hacohen, and A. Regev (2015). 'Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens'. *Science* 347 (6226), p. 1259038.

Kadota, K., J. Ye, Y. Nakai, T. Terada, and K. Shimizu (2006). 'ROKU: a novel method for identification of tissue-specific genes'. *BMC Bioinf.* 7, p. 294.

Käll, L., J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss (2007). 'Semi-supervised learning for peptide identification from shotgun proteomics datasets'. *Nat. Methods* 4 (11), pp. 923–925.

Käll, L., J. D. Storey, M. J. MacCoss, and W. S. Noble (2008). 'Posterior error probabilities and false discovery rates: two sides of the same coin'. *J. Proteome Res.* 7 (1), pp. 40–44.

Karro, J. E., Y. Yan, D. Zheng, Z. Zhang, N. Carriero, P. Cayting, P. Harrrison, and M. Gerstein (2007). 'Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation'. *Nucleic Acids Res.* 35 (Database issue), pp. D55–60.

Karthik, D., G. Stelzer, S. Gershanov, D. Baranes, and M. Salmon-Divon (2016). 'Elucidating tissue specific genes using the Benford distribution'. *BMC Genom.* 17, p. 595.

Kechavarzi, B. and S. C. Janga (2014). 'Dissecting the expression landscape of RNA-binding proteins in human cancers'. *Genome Biol.* 15 (1), R14.

Keller, A., A. I. Nesvizhskii, E. Kolker, and R. Aebersold (2002). 'Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search'. *Anal. Chem.* 74 (20), pp. 5383–5392.

Kern, D. M., P. K. Nicholls, D. C. Page, and I. M. Cheeseman (2016). 'A mitotic SKAP isoform regulates spindle positioning at astral microtubule plus ends'. *J. Cell Biol.* 213 (3), pp. 315–328.

Khang, T. F. and C. Y. Lau (2015). 'Getting the most out of RNA-seq data analysis'. *PeerJ* 3, e1360.

Khatri, P. and S. Drăghici (2005). 'Ontological analysis of gene expression data: current tools, limitations, and open problems'. *Bioinformatics* 21 (18), pp. 3587–3595.

Khatri, P., M. Sirota, and A. J. Butte (2012). 'Ten years of pathway analysis: current approaches and outstanding challenges'. *PLOS Comput. Biol.* 8 (2), e1002375.

Kim, D., B. Langmead, and S. L. Salzberg (2015). 'HISAT: a fast spliced aligner with low memory requirements'. *Nat. Methods* 12 (4), pp. 357–360.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (2013). 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions'. *Genome Biol.* 14 (4), R36.

Kim, M.-S., S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. K. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey (2014). 'A draft map of the human proteome'. *Nature* 509 (7502), pp. 575–581.

Kim, M., A. Eetemadi, and I. Tagkopoulos (2017). 'DeepPep: Deep proteome inference from peptide profiles'. *PLOS Comput. Biol.* 13 (9), e1005661.

Kim, P., A. Park, G. Han, H. Sun, P. Jia, and Z. Zhao (2017). 'TissGDB: tissue-specific gene database in cancer'. *Nucleic Acids Res.* 46 (D1).

Kim, S. and P. A. Pevzner (2014). 'MS-GF+ makes progress towards a universal database search tool for proteomics'. *Nat. Commun.* 5, p. 5277.

Kohlbacher, O., K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm (2007). 'TOPP–the OpenMS proteomics pipeline'. *Bioinformatics* 23 (2), e191–7.

Kong, A. T., F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii (2017). 'MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics'. *Nat. Methods* 14 (5), pp. 513–520.

Kosti, I., N. Jain, D. Aran, A. J. Butte, and M. Sirota (2016). 'Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues'. *Sci. Rep.* 6, 24799ep.

Kotrys, A. V. and R. J. Szczesny (2019). 'Mitochondrial Gene Expression and Beyond-Novel Aspects of Cellular Physiology'. *Cells* 9 (1).

Koziol, J., N. Griffin, F. Long, Y. Li, M. Latterich, and J. Schnitzer (2013). 'On protein abundance distributions in complex mixtures'. *Proteome Sci.* 11 (1), p. 5.

Kratz, A. and P. Carninci (2014). 'The devil in the details of RNA-seq'. *Nat. Biotechnol.* 32 (9), pp. 882–884.

Kroll, K. W., N. E. Mokaram, A. R. Pelletier, D. E. Frankhouser, M. S. Westphal, P. A. Stump, C. L. Stump, R. Bundschuh, J. S. Blachly, and P. Yan (2014). 'Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline'. *Cancer Informat.* 13 (Suppl 3), pp. 7–14.

Krupp, M., J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel (2012). 'RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing'. *Bioinformatics* 28 (8), pp. 1184–1185.

Kryuchkova-Mostacci, N. and M. Robinson-Rechavi (2017). 'A benchmark of gene expression tissue-specificity metrics'. *Briefings Bioinf.* 18 (2), pp. 205–214.

Kumar, A., S. K. Ghosh, M. A. Faiq, V. R. Deshmukh, C. Kumari, and V. Pareek (2019). 'A brief review of recent discoveries in human anatomy'. *QJM* 112 (8), pp. 567–573.

Kurt, W. (2019). *Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks.* San Francisco, CA, US: No Starch Press, Incorporated.

Ladoukakis, E. D. and E. Zouros (2017). 'Evolution and inheritance of animal mitochondrial DNA: rules and exceptions'. *J. Biol. Res.* 24, p. 2.

Lam, H., E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold (2008). 'Building consensus spectral libraries for peptide identification in proteomics'. *Nat. Methods* 5 (10), pp. 873–875.

Lander, E. S. et al. (2001). 'Initial sequencing and analysis of the human genome'. *Nature* 409 (6822), pp. 860–921.

Langfelder, P. and S. Horvath (2008). 'WGCNA: an R package for weighted correlation network analysis'. *BMC Bioinf.* 9, p. 559.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome'. *Genome Biol.* 10 (3), R25.

Laskay, Ü. A., A. A. Lobas, K. Srzentić, M. V. Gorshkov, and Y. O. Tsybin (2013). 'Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments'. *J. Proteome Res.* 12 (12), pp. 5558–5569.

Laurent, J. M., C. Vogel, T. Kwon, S. A. Craig, D. R. Boutz, H. K. Huse, K. Nozue, H. Walia, M. Whiteley, P. C. Ronald, and E. M. Marcotte (2010). 'Protein abundances are more conserved than mRNA abundances across diverse taxa'. *Proteomics* 10 (23), pp. 4209–4212.

Lazar, C., L. Gatto, M. Ferro, C. Bruley, and T. Burger (2016). 'Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies'. *J. Proteome Res.* 15 (4), pp. 1116–1125.

Lazar, C., S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé (2013). 'Batch effect removal methods for microarray gene expression data integration: a survey'. *Briefings Bioinf.* 14 (4), pp. 469–490.

Lee, H. Y., E. G. Kim, H. R. Jung, J. W. Jung, H. B. Kim, J. W. Cho, K. M. Kim, and E. C. Yi (2019). 'Refinements of LC-MS/MS Spectral Counting Statistics Improve Quantification of Low Abundance Proteins'. *Sci. Rep.* 9, p. 13653.

Lee, L. H. (2015). 'Quantitative and functional analysis pipeline for label-free metaproteomics data and its applications'. PhD thesis. University of Tennessee.

Lee, M.-L. (2006). *Analysis of Microarray Gene Expression Data.* New York, NY, US: Springer.

Leek, J. T. (2014). 'svaseq: removing batch effects and other unwanted noise from sequencing data'. *Nucleic Acids Res.* 42 (21).

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010). 'Tackling the widespread and critical impact of batch effects in high-throughput data'. *Nat. Rev. Genet.* 11 (10), pp. 733–739.

Levin, J. Z., M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev (2010). 'Comprehensive comparative analysis of strand-specific RNA sequencing methods'. *Nat. Methods* 7 (9), pp. 709–715.

Lewczuk, P., G. Beck, O. Ganslandt, H. Esselmann, F. Deisenhammer, A. Regeniter, H.-F. Petereit, H. Tumani, A. Gerritzen, P. Oschmann, J. Schröder, P. Schönknecht, K. Zimmermann, H. Hampel, K. Bürger, M. Otto, S. Haustein, K. Herzog, R. Dannenberg, U. Wurster, M. Bibl, J. M. Maler, U. Reubach, J. Kornhuber, and J. Wiltfang (2006). 'International quality control survey of neurochemical dementia diagnostics'. *Neurosci. Lett.* 409 (1), pp. 1–4.

Li, B. and G. J. Babu (2019). 'Bayesian Inference'. *A Graduate Course on Statistical Inference.* Ed. by B. Li and G. J. Babu. New York, NY, US: Springer New York, pp. 173–201.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup (2009). 'The Sequence Alignment/Map format and SAMtools'. *Bioinformatics* 25 (16), pp. 2078–2079.

Li, J. J., P. J. Bickel, and M. D. Biggin (2014). 'System wide analyses have underestimated protein abundances and the importance of transcription in mammals'. *PeerJ* 2, e270.

Li, P., Y. Piao, H. S. Shon, and K. H. Ryu (2015). 'Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.' *BMC Bioinf.* 16, p. 347.

Li, S., P. P. Łabaj, P. Zumbo, P. Sykacek, W. Shi, L. Shi, J. Phan, P.-Y. Wu, M. Wang, C. Wang, D. Thierry-Mieg, J. Thierry-Mieg, D. P. Kreil, and C. E. Mason (2014). 'Detecting and correcting systematic variation in large-scale RNA sequencing data'. *Nat. Biotechnol.* 32 (9), pp. 888–895.

Liang, S., Y. Li, X. Be, S. Howes, and W. Liu (2006). 'Detecting and profiling tissue-selective genes'. *Physiol. Genomics* 26 (2), pp. 158–162.

Lide, D. R., ed. (2005). *Handbook of Chemistry and Physics, 85th edition.* Boca Raton, FL, US: CRC Press.

Liebal, U. W., A. N. T. Phan, M. Sudhakar, K. Raman, and L. M. Blank (2020). 'Machine Learning Applications for Mass Spectrometry-Based Metabolomics'. *Metabolites* 10 (6).

Lin, T. Y., Y. Xie, A. Wasilewska, and C.-J. Liau, eds. (2008). *Data Mining: Foundations and Practice.* Berlin, DE: Berlin Springer.

Lindemann, C., N. Thomanek, F. Hundt, T. Lerari, H. E. Meyer, D. Wolters, and K. Marcus (2017). 'Strategies in relative and absolute quantitative mass spectrometry based proteomics'. *Biol. Chem.* 398 (5-6), pp. 687–699.

Lindner, M. D., K. D. Torralba, and N. A. Khan (2018). 'Scientific productivity: An exploratory study of metrics and incentives'. *PLOS ONE* 13 (4), e0195321.

Linsinger, T. P. J., W. Kandler, R. Krska, and M. Grasserbauer (1998). 'The influence of different evaluation techniques on the results of interlaboratory comparisons'. *Accredit. Qual. Assur.* 3 (8), pp. 322–327.

Liu, H., S. Shah, and W. Jiang (2004). 'On-line outlier detection and data cleaning'. *Comput. Chem. Eng.* 28 (9), pp. 1635–1647.

Liu, H., R. G. Sadygov, and J. R. Yates 3rd (2004). 'A model for random sampling and estimation of relative protein abundance in shotgun proteomics'. *Anal. Chem.* 76 (14), pp. 4193–4201.

Liu, K., J. Zhang, J. Wang, L. Zhao, X. Peng, W. Jia, W. Ying, Y. Zhu, H. Xie, F. He, and X. Qian (2009). 'Relationship between sample loading amount and peptide identification and its effects on quantitative proteomics'. *Anal. Chem.* 81 (4), pp. 1307–1314.

Liu, W., J. Wang, T. Wang, and H. Xie (2014). 'Construction and analyses of human large-scale tissue specific networks'. *PLOS ONE* 9 (12), e115074.

Liu, X., X. Yu, D. J. Zack, H. Zhu, and J. Qian (2008). 'TiGER: a database for tissue-specific gene expression and regulation'. *BMC Bioinf.* 9, p. 271.

Liu, Y., A. Beyer, and R. Aebersold (2016). 'On the Dependency of Cellular Protein Levels on mRNA Abundance'. *Cell* 165 (3), pp. 535–550.

Love, M. I., W. Huber, and S. Anders (2014). 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2'. *Genome Biol.* 15 (12), p. 550.

Lowe, R., N. Shirley, M. Bleackley, S. Dolan, and T. Shafee (2017). 'Transcriptomics technologies'. *PLOS Comput. Biol.* 13 (5), e1005457.

Lukk, M., M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma (2010). 'A global map of human gene expression'. *Nat. Biotechnol.* 28 (4), pp. 322–324.

Lundberg, E., L. Fagerberg, D. Klevebring, I. Matic, T. Geiger, J. Cox, C. Algenäs, J. Lundeberg, M. Mann, and M. Uhlen (2010). 'Defining the transcriptome and proteome in three functionally different human cell lines'. *Mol. Syst. Biol.* 6 (1), p. 450.

Lyu, Z., K. Peng, and C.-P. Hu (2018). 'P-Value, Confidence Intervals, and Statistical Inference: A New Dataset of Misinterpretation'. *Front. Psychol.* 9, p. 868.

Ma, B. (2015). 'Novor: real-time peptide de novo sequencing software'. *J. Am. Soc. Mass Spectrom.* 26 (11), pp. 1885–1894.

Ma, B., K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie (2003). 'PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry'. *Rapid Commun. Mass Spectrom.* 17 (20), pp. 2337–2342.

Ma, K., O. Vitek, and A. I. Nesvizhskii (2012). 'A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet'. *BMC Bioinf.* 13 Suppl 16, S1.

Mabbott, N. A., J. K. Baillie, H. Brown, T. C. Freeman, and D. A. Hume (2013). 'An expression atlas of human primary cells: inference of gene function from coexpression networks'. *BMC Genom.* 14, p. 632.

Macias, L. A., I. C. Santos, and J. S. Brodbelt (2020). 'Ion Activation Methods for Peptides and Proteins'. *Anal. Chem.* 92 (1), pp. 227–251.

Mackay, K. I. (2015). 'A Comparative Study of Analysis Methods in Quantitative Label-free Proteomics'. PhD thesis. University of Liverpool.

MacLean, B., J. K. Eng, R. C. Beavis, and M. McIntosh (2006). 'General framework for developing and evaluating database scoring algorithms using the TANDEM search engine'. *Bioinformatics* 22 (22), pp. 2830–2832.

Madalinski, G., E. Godat, S. Alves, D. Lesage, E. Genin, P. Levi, J. Labarre, J.-C. Tabet, E. Ezan, and C. Junot (2008). 'Direct introduction of biological samples into a LTQ-Orbitrap hybrid mass spectrometer as a tool for fast metabolome analysis'. *Anal. Chem.* 80 (9), pp. 3291–3303.

Maes, E., P. Kelchtermans, W. Bittremieux, K. De Grave, S. Degroeve, J. Hooyberghs, I. Mertens, G. Baggerman, J. Ramon, K. Laukens, L. Martens, and D. Valkenborg (2016). 'Designing biomedical proteomics experiments: state-of-the-art and future perspectives'. *Expert Rev. Proteomics* 13 (5), pp. 495–511.

Maier, T., M. Güell, and L. Serrano (2009). 'Correlation of mRNA and protein in complex biological samples'. *FEBS Lett.* 583 (24), pp. 3966–3973.

Maier, T., A. Schmidt, M. Güell, S. Kühner, A.-C. Gavin, R. Aebersold, and L. Serrano (2011). 'Quantification of mRNA and protein and integration with protein turnover in a bacterium'. *Mol. Syst. Biol.* 7 (1), p. 511.

Makarov, A. (2000). 'Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis'. *Anal. Chem.* 72 (6), pp. 1156–1162.

Mallick, P., M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold (2007). 'Computational prediction of proteotypic peptides for quantitative proteomics'. *Nat. Biotechnol.* 25 (1), pp. 125–131.

Manza, L. L., S. L. Stamer, A.-J. L. Ham, S. G. Codreanu, and D. C. Liebler (2005). 'Sample preparation and digestion for proteomic analyses using spin filters'. *Proteomics* 5 (7), pp. 1742–1745.

Marguerat, S., A. Schmidt, S. Codlin, W. Chen, R. Aebersold, and J. Bähler (2012). 'Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells'. *Cell* 151 (3), pp. 671–683.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008). 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays'. *Genome Res.* 18 (9), pp. 1509–1517.

Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch (2011). 'mzML — A Community Standard for Mass Spectrometry Data'. *Mol. Cell. Proteom.* 10 (1).

Martin, J. A. and Z. Wang (2011). 'Next-generation transcriptome assembly'. *Nat. Rev. Genet.* 12 (10), pp. 671–682.

Martin, W. F., S. Garg, and V. Zimorski (2015). 'Endosymbiotic theories for eukaryote origin'. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 370 (1678), p. 20140330.

Martínez, O. and M. H. Reyes-Valdés (2008). 'Defining diversity, specialization, and gene specificity in transcriptomes through information theory'. *PNAS* 105 (28), pp. 9709–9714.

Marx, V. (2019). 'A dream of single-cell proteomics'. *Nat. Methods* 16 (9), pp. 809–812.

Mathur, R., D. Rotroff, J. Ma, A. Shojaie, and A. Motsinger-Reif (2018). 'Gene set analysis methods: a systematic comparison'. *BioData Min.* 11, p. 8.

McCaffrey, A. (1964). *Dragonflight.* New York, NY, US: Ballantine Books.

McIlwain, S., K. Tamura, A. Kertesz-Farkas, C. E. Grant, B. Diament, B. Frewen, J. J. Howbert, M. R. Hoopmann, L. Käll, J. K. Eng, M. J. MacCoss, and W. S. Noble (2014). 'Crux: rapid open source protein tandem mass spectrometry analysis'. *J. Proteome Res.* 13 (10), pp. 4488–4491.

McPherson, J. D. (2014). 'A defining decade in DNA sequencing'. *Nat. Methods* 11 (10), pp. 1003–1005.

Medzihradszky, K. F. and R. J. Chalkley (2015). 'Lessons in de novo peptide sequencing by tandem mass spectrometry'. *Mass Spectrom. Rev.* 34 (1), pp. 43–63.

Melé, M., P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, a. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Ségrè, S. Djebali, A. Niarchou, T. G. Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó (2015). 'The human transcriptome across tissues and individuals'. *Science* 348 (6235), pp. 660–665.

Menschaert, G. and D. Fenyö (2017). 'Proteogenomics from a bioinformatics angle: A growing field'. *Mass Spectrom. Rev.* 36 (5), pp. 584–599.

Miller, R. M., R. J. Millikin, C. V. Hoffmann, S. K. Solntsev, G. M. Sheynkman, M. R. Shortreed, and L. M. Smith (2019). 'Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data'. *J. Proteome Res.* 18 (9), pp. 3429–3438.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer (2011). 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems'. *Genome Biol.* 12 (11), R112.

Morlan, J. D., K. Qu, and D. V. Sinicropi (2012). 'Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue'. *PLOS ONE* 7 (8), e42882.

Morris, J., D. Hartl, A. Knoll, R. Lue, M. Michael, A. Berry, A. Biewener, B. Farrell, and N. M. Holbrook (2016). *Biology: How Life Works, 2nd edition.* New York, NY, US: W. H. Freeman.

Morrison, S. J. (2014). 'Time to do something about reproducibility'. *eLife* 3.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). 'Mapping and quantifying mammalian transcriptomes by RNA-Seq'. *Nat. Methods* 5 (7), pp. 621–628.

Murrell, P. (2014). *gridBase: Integration of base and grid graphics.* R package version 0.4-7.

Muth, T., F. Hartkopf, M. Vaudel, and B. Y. Renard (2018). 'A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics'. *Proteomics* 18 (18), e1700150.

Nagaraj, S. H., R. B. Gasser, and S. Ranganathan (2007). 'A hitchhiker's guide to expressed sequence tag (EST) analysis'. *Briefings Bioinf.* 8 (1), pp. 6–21.

Nahnsen, S., C. Bielow, K. Reinert, and O. Kohlbacher (2013). 'Tools for label-free peptide quantification'. *Mol. Cell. Proteom.* 12 (3), pp. 549–556.

Navarro, P., J. Kuharev, L. C. Gillet, O. M. Bernhardt, B. MacLean, H. L. Röst, S. A. Tate, C.-C. Tsou, L. Reiter, U. Distler, G. Rosenberger, Y. Perez-Riverol, A. I. Nesvizhskii, R. Aebersold, and S. Tenzer (2016). 'A multicenter study benchmarks software tools for label-free proteome quantification'. *Nat. Biotechnol.* 34 (11), pp. 1130–1136.

Neilson, K. A., N. A. Ali, S. Muralidharan, M. Mirzaei, M. Mariani, G. Assadourian, A. Lee, S. C. van Sluyter, and P. A. Haynes (2011). 'Less label, more free: approaches in label-free quantitative mass spectrometry'. *Proteomics* 11 (4), pp. 535–553.

Nesvizhskii, A. (2006). *Statistical Validation of Protein Identifications.*

Nesvizhskii, A. I. (2010). 'A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics'. *J. Proteom.* 73 (11), pp. 2092–2123.

Nesvizhskii, A. I. and R. Aebersold (2005). 'Interpretation of shotgun proteomic data: the protein inference problem'. *Mol. Cell. Proteom.* 4 (10), pp. 1419–1440.

Nesvizhskii, A. I., A. Keller, E. Kolker, and R. Aebersold (2003). 'A statistical model for identifying proteins by tandem mass spectrometry'. *Anal. Chem.* 75 (17), pp. 4646–4658.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes.* R package version 1.1-2.

Nilsson, T., M. Mann, R. Aebersold, J. R. Yates 3rd, A. Bairoch, and J. J. M. Bergeron (2010). 'Mass spectrometry in high-throughput proteomics: ready for the big time'. *Nat. Methods* 7 (9), pp. 681–685.

Noor, Z., S. B. Ahn, M. S. Baker, S. Ranganathan, and A. Mohamedali (2020). 'Mass spectrometry-based protein identification in proteomics-a review'. *Briefings Bioinf.*

O'Bryon, I., S. C. Jenson, and E. D. Merkley (2020). 'Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide identification'. *Protein Sci.* 29 (9), pp. 1864–1878.

O'Malley, C. D. (1964). *Andreas Vesalius of Brussels, 1514–1564.* Berkeley, CA, US: Berkeley: University of California Press.

Oytam, Y., F. Sobhanmanesh, K. Duesing, J. C. Bowden, M. Osmond-McLeod, and J. Ross (2016). 'Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets'. *BMC Bioinf.* 17 (1), p. 332.

Pachter, L. (2015). *What is a read mapping?* Accessed: 2017-6-27. URL: https://liorpachter.wordpress.com/2015/11/01/what-is-a-read-mapping/.

Palasca, O., A. Santos, C. Stolte, J. Gorodkin, and L. J. Jensen (2018). 'TISSUES 2.0: an integrative web resource on mammalian tissue expression'. *Database* 2018.

Palmblad, M., C. V. Henkel, R. P. Dirks, A. H. Meijer, A. M. Deelder, and H. P. Spaink (2013). 'Parallel deep transcriptome and proteome analysis of zebrafish larvae'. *BMC Res. Notes* 6, p. 428.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe (2008). 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing'. *Nat. Genet.* 40 (12), pp. 1413–1415.

Papachristodoulou, D., A. Snape, W. H. Elliott, and D. C. Elliott (2014). *Biochemistry and Molecular Biology.* Oxford, UK: OUP Oxford.

Pappireddi, N., L. Martin, and M. Wühr (2019). 'A Review on Quantitative Multiplexed Proteomics'. *Chembiochem* 20 (10), pp. 1210–1224.

Paradis, E. and K. Schliep (2019). 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R'. *Bioinformatics* 35 (3), pp. 526–528.

Park, C. Y., A. A. Klammer, L. Käll, M. J. MacCoss, and W. S. Noble (2008). 'Rapid and accurate peptide identification from tandem mass spectra'. *J. Proteome Res.* 7 (7), pp. 3022–3027.

Parkhomchuk, D., T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, and A. Soldatov (2009). 'Transcriptome analysis by strand-specific sequencing of complementary DNA'. *Nucleic Acids Res.* 37 (18), e123.

Parkinson, J. and M. Blaxter (2009). 'Expressed Sequence Tags: An overview'. *Expressed Sequence Tags (ESTs): Generation and Analysis.* Ed. by J. Parkinson. Vol. 533. Methods in Molecular Biology. Totowa, NJ, US: Humana Press, pp. 1–12.

Pascal, L. E., L. D. True, D. S. Campbell, E. W. Deutsch, M. Risk, I. M. Coleman, L. J. Eichner, P. S. Nelson, and A. Y. Liu (2008). 'Correlation of mRNA and protein levels: Cell type-specific gene expression of cluster designation antigens in the prostate'. *BMC Genom.* 9, p. 246.

Pasquali, L., K. J. Gaulton, S. A. Rodríguez-Seguí, L. Mularoni, I. Miguel-Escalada, İ. Akerman, J. J. Tena, I. Morán, C. Gómez-Marín, M. van de Bunt, J. Ponsa-Cobas, N. Castro, T. Nammo, I. Cebola, J. García-Hurtado, M. A. Maestro, F. Pattou, L. Piemonti, T. Berney, A. L. Gloyn, P. Ravassard, J. L. G. Skarmeta, F. Müller, M. I. McCarthy, and J. Ferrer (2014). 'Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants'. *Nat. Genet.* 46 (2), pp. 136–143.

Pazhamala, L. T., S. Purohit, R. K. Saxena, V. Garg, L. Krishnamurthy, J. Verdier, and R. K. Varshney (2017). 'Gene expression atlas of pigeonpea and its application to gain insights into genes associated with pollen fertility implicated in seed formation'. *J. Exp. Bot.* 68 (8), pp. 2037–2054.

Pearce, S., H. Vazquez-Gross, S. Y. Herin, D. Hane, Y. Wang, Y. Q. Gu, and J. Dubcovsky (2015). 'WheatExp: an RNA-seq expression database for polyploid wheat'. *BMC Plant Biol.* 15, p. 299.

Pearson, R. K. (2002). 'Outliers in process modeling and identification'. *IEEE Trans. Control Syst. Technol.* 10 (1), pp. 55–63.

Peixoto, L., D. Risso, S. G. Poplawski, M. E. Wimmer, T. P. Speed, M. A. Wood, and T. Abel (2015). 'How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets'. *Nucleic Acids Res.* 43 (16), pp. 7664–7674.

Penkert, M., A. Hauser, R. Harmel, D. Fiedler, C. P. R. Hackenberger, and E. Krause (2019). 'Electron Transfer/Higher Energy Collisional Dissociation of Doubly Charged Peptide Ions: Identification of Labile Protein Phosphorylations'. *J. Am. Soc. Mass Spectrom.* 30 (9), pp. 1578–1585.

Penkert, M., L. M. Yates, M. Schümann, D. Perlman, D. Fiedler, and E. Krause (2017). 'Unambiguous Identification of Serine and Threonine Pyrophosphorylation Using Neutral-Loss-Triggered Electron-Transfer/Higher-Energy Collision Dissociation'. *Anal. Chem.* 89.6, pp. 3672–3680.

Perkins, D. N., D. J. Pappin, D. M. Creasy, and J. S. Cottrell (1999). 'Probability-based protein identification by searching sequence databases using mass spectrometry data'. *Electrophoresis* 20 (18), pp. 3551–3567.

Petryszak, R., T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, J. McMurry, J. C. Marioni, J. Malone, K. Megy, G. Rustici, A. Y. Tang, J. Taubert, E. Williams, O. Mannion, H. E. Parkinson, and A. Brazma (2014). 'Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments'. *Nucleic Acids Res.* 42 (Database issue), pp. D926–32.

Petryszak, R., M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, A. Füllgrabe, A. M.-P. Fuentes, S. Jupp, S. Koskinen, O. Mannion, L. Huerta, K. Megy, C. Snow, E. Williams, M. Barzine, E. Hastings, H. Weisser, J. Wright, P. Jaiswal, W. Huber, J. Choudhary, H. E. Parkinson, and A. Brazma (2015). 'Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants'. *Nucleic Acids Res.* 44 (D1), pp. D746–52.

Pfeuffer, J., T. Sachsenberg, O. Alka, M. Walzer, A. Fillbrunn, L. Nilse, O. Schilling, K. Reinert, and O. Kohlbacher (2017). 'OpenMS - A platform for reproducible analysis of mass spectrometry data'. *J. Biotechnol.* 261, pp. 142–148.

Pfeuffer, J., T. Sachsenberg, T. M. H. Dijkstra, O. Serang, K. Reinert, and O. Kohlbacher (2020). 'EPIFANY: A Method for Efficient High-Confidence Protein Inference'. *J. Proteome Res.* 19 (3), pp. 1060–1072.

Picotti, P. and R. Aebersold (2012). 'Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions'. *Nat. Methods* 9 (6), pp. 555–566.

Pierce, B. A. (2005). *Genetics: A conceptual approach, 2nd edition.* New York, NY, US: Macmillan.

Piétu, G., R. Mariage-Samson, N. A. Fayein, C. Matingou, E. Eveno, R. Houlgatte, C. Decraene, Y. Vandenbrouck, F. Tahi, M. D. Devignes, U. Wirkner, W. Ansorge, D. Cox, T. Nagase, N. Nomura, and C. Auffray (1999). 'The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics'. *Genome Res.* 9 (2), pp. 195–209.

Pino, L. K., B. C. Searle, J. G. Bollinger, B. Nunn, B. MacLean, and M. J. MacCoss (2020). 'The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics'. *Mass Spectrom. Rev.* 39 (3), pp. 229–244.

Plotkin, J. B. (2010). 'Transcriptional regulation is only half the story'. *Mol. Syst. Biol.* 6 (1), p. 406.

Pontius, Joan U. and Wagner, Lukas and Schuler, Gregory D. (2002). 'UniGene: A Unified View of the Transcriptome'. *The NCBI Handbook [Internet].* Ed. by O. J. McEntyre J. Bethesda, MD, US: National Center for Biotechnology Information (US).

Pootakham, W., W. Mhuantong, T. Yoocha, L. Putchim, C. Sonthirod, C. Naktang, N. Thongtham, and S. Tangphatsornruang (2017). 'High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system'. *Sci. Rep.* 7, p. 2774.

Poverennaya, E. V., E. V. Ilgisonis, E. A. Ponomarenko, A. T. Kopylov, V. G. Zgoda, S. P. Radko, A. V. Lisitsa, and A. I. Archakov (2017). 'Why Are the Correlations between mRNA and Protein Levels so Low among the 275 Predicted Protein-Coding Genes on Human Chromosome 18?' *J. Proteome Res.* 16 (12), pp. 4311–4318.

Prieto, G., K. Aloria, N. Osinalde, A. Fullaondo, J. M. Arizmendi, and R. Matthiesen (2012). 'PAnalyzer: a software tool for protein inference in shotgun proteomics'. *BMC Bioinf.* 13, p. 288.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramakrishnan, S. R., C. Vogel, J. T. Prince, Z. Li, L. O. Penalva, M. Myers, E. M. Marcotte, D. P. Miranker, and R. Wang (2009). 'Integrating shotgun proteomics and mRNA expression data to improve protein identification'. *Bioinformatics* 25 (11), pp. 1397–1403.

Ramsköld, D., E. T. Wang, C. B. Burge, and R. Sandberg (2009). 'An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.' *PLOS Comput. Biol.* 5 (12), e1000598.

Rau, A., G. Marot, and F. Jaffrézic (2014). 'Differential meta-analysis of RNA-seq data from multiple studies'. *BMC Bioinf.* 15, p. 91.

Rechenberger, J., P. Samaras, A. Jarzab, J. Behr, M. Frejno, A. Djukovic, J. Sanz, E. M. González-Barberá, M. Salavert, J. L. López-Hontangas, K. B. Xavier, L. Debrauwer, J.-M. Rolain, M. Sanz, M. Garcia-Garcera, M. Wilhelm, C. Ubeda, and B. Kuster (2019). 'Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae'. *Proteomes* 7 (1).

Renard, B. Y., M. Kirchner, H. Steen, J. A. J. Steen, and F. A. Hamprecht (2008). 'NITPICK: peak identification for mass spectrometry data'. *BMC Bioinf.* 9, p. 355.

Révész, Á., M. G. Milley, K. Nagy, D. Szabó, G. Kalló, É. Csősz, K. Vékey, and L. Drahos (2021). 'Tailoring to Search Engines: Bottom-Up Proteomics with Collision Energies Optimized for Identification Confidence'. *J. Proteome Res.* 20, pp. 474–484.

Ringwald, M., C. Wu, and A. I. Su (2012). 'BioGPS and GXD: mouse gene expression data-the benefits and challenges of data integration'. *Mamm. Genome* 23 (9-10), pp. 550–558.

Risso, D., J. Ngai, T. P. Speed, and S. Dudoit (2014). 'Normalization of RNA-seq data using factor analysis of control genes or samples'. *Nat. Biotechnol.* 32 (9), pp. 896–902.

Robert, C. and M. Watson (2015). 'Errors in RNA-Seq quantification affect genes of relevance to human disease'. *Genome Biol.* 16, p. 177.

Roberts, A., H. Pimentel, C. Trapnell, and L. Pachter (2011). 'Identification of novel transcripts in annotated genomes using RNA-Seq'. *Bioinformatics* 27 (17), pp. 2325–2329.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol (2010). 'De novo assembly and analysis of RNA-seq data'. *Nat. Methods* 7 (11), pp. 909–912.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data'. *Bioinformatics* 26 (1), pp. 139–140.

Robotham, S. A., A. P. Horton, J. R. Cannon, V. C. Cotham, E. M. Marcotte, and J. S. Brodbelt (2016). 'UVnovo: A de Novo Sequencing Algorithm Using Single Series of Fragment Ions via Chromophore Tagging and 351 nm Ultraviolet Photodissociation Mass Spectrometry'. *Anal. Chem.* 88 (7), pp. 3990–3997.

Rocke, D. M. (1983). 'Robust statistical analysis of interlaboratory studies'. *Biometrika* 70 (2), pp. 421–431.

Rodriguez, J. M., J. Rodriguez-Rivas, T. Di Domenico, J. Vázquez, A. Valencia, and M. L. Tress (2018). 'APPRIS 2017: principal isoforms for multiple gene sets'. *Nucleic Acids Res.* 46 (D1), pp. D213–D217.

Roepstorff, P. and J. Fohlman (1984). 'Proposal for a common nomenclature for sequence ions in mass spectra of peptides'. *Biomed. Mass Spectrom.* 11 (11), p. 601.

Rorbach, J., A. Bobrowicz, S. Pearce, and M. Minczuk (2014). 'Polyadenylation in Bacteria and Organelles'. *Polyadenylation: Methods and Protocols.* Ed. by J. Rorbach and A. J. Bobrowicz. Totowa, NJ, US: Humana Press, pp. 211–227.

Rosenberger, G., Y. Liu, H. L. Röst, C. Ludwig, A. Buil, A. Bensimon, M. Soste, T. D. Spector, E. T. Dermitzakis, B. C. Collins, L. Malmström, and R. Aebersold (2017). 'Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS'. *Nat. biotechnol.* 35 (8), pp. 781–788.

Rosman, K. J. R. and P. D. P. Taylor (1998). ‘Isotopic compositions of the elements 1997’. *Pure Appl. Chem.* 70 (1), pp. 217–235.

Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin (2004). ‘Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents’. *Mol. Cell. Proteom.* 3 (12), pp. 1154–1169.

Röst, H. L., T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher (2016). ‘OpenMS: a flexible open-source software platform for mass spectrometry data analysis’. *Nat. Methods* 13 (9), pp. 741–748.

Rouillard, A. D., G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma’ayan (2016). ‘The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins’. *Database* 2016.

Rudnick, P. A., X. Wang, X. Yan, N. Sedransk, and S. E. Stein (2014). ‘Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data’. *Mol. Cell. Proteom.* 13 (5), pp. 1341–1351.

Rudolph, K. L. M. (2014). *ebits: An alternative module system for R.*

Rung, J. and A. Brazma (2013). ‘Reuse of public genome-wide gene expression data’. *Nat. Rev. Genet.* 14 (2), pp. 89–99.

Sadygov, R. G., D. Cociorva, and J. R. Yates 3rd (2004). ‘Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book’. *Nat. Methods* 1 (3), pp. 195–202.

Saltzman, A. B., M. Leng, B. Bhatt, P. Singh, D. W. Chan, L. Dobrolecki, H. Chandrasekaran, J. M. Choi, A. Jain, S. Y. Jung, M. T. Lewis, M. J. Ellis, and A. Malovannaya (2018). ‘gpGrouper: A Peptide Grouping Algorithm for Gene-Centric Inference and Quantitation of Bottom-Up Proteomics Data’. *Mol. Cell. Proteom.* 17 (11), pp. 2270–2283.

Sandin, M., J. Teleman, J. Malmström, and F. Levander (2014). ‘Data processing methods and quality control strategies for label-free LC-MS protein quantification’. *BBA* 1844 (1 Pt A), pp. 29–41.

Sanger, F. and A. R. Coulson (1975). ‘A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase’. *J. Mol. Biol.* 94 (3), pp. 441–448.

Santos, A., K. Tsafou, C. Stolte, S. Pletscher-Frankild, S. I. O’Donoghue, and L. J. Jensen (2015). ‘Comprehensive comparison of large-scale tissue expression datasets’. *PeerJ* 3, e1054.

Savitski, M. M., M. Wilhelm, H. Hahne, B. Kuster, and M. Bantscheff (2015). ‘A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets’. *Mol. Cell. Proteom.* 14 (9), pp. 2394–2404.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). ‘Quantitative monitoring of gene expression patterns with a complementary DNA microarray’. *Science* 270 (5235), pp. 467–470.

Schubert, M. (2019). ‘clustermq enables efficient parallelisation of genomic analyses’. *Bioinformatics* 35 (21).

Schubert, M. and K. L. M. Rudolph (2014). *modules: An alternative module system for R.*

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach (2011). ‘Global quantification of mammalian gene expression control.’ *Nature* 473 (7347), pp. 337–342.

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach (2013). ‘Corrigendum: Global quantification of mammalian gene expression control’. *Nature* 495 (7439), pp. 126–127.

Scigelova, M., M. Hornshaw, A. Giannakopulos, and A. Makarov (2011). ‘Fourier transform mass spectrometry’. *Mol. Cell. Proteom.* 10 (7), p. M111.009431.

Searle, B. C., M. Turner, and A. I. Nesvizhskii (2008). ‘Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies’. *J. Proteome Res.* 7 (1), pp. 245–253.

SEQC/MAQC-III Consortium (2014). ‘A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium’. *Nat. Biotechnol.* 32 (9), pp. 903–914.

Serang, O., M. J. MacCoss, and W. S. Noble (2010). ‘Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data’. *J. Proteome Res.* 9 (10), pp. 5346–5357.

Serang, O. and W. Noble (2012). ‘A review of statistical methods for protein identification using tandem mass spectrometry’. *Stat. Interface* 5 (1), pp. 3–20.

Shaffer, J. P. (1995). ‘Multiple Hypothesis Testing’. *Annu. Rev. Psychol.* 46 (1), pp. 561–584.

Shevchenko, A., H. Tomas, J. Havlis, J. V. Olsen, and M. Mann (2006). ‘In-gel digestion for mass spectrometric characterization of proteins and proteomes’. *Nat. Protoc.* 1 (6), pp. 2856–2860.

Shi Jing, L., L. S. Jing, F. F. M. Shah, M. S. Mohamad, K. Moorthy, S. Deris, Z. Zakaria, and S. Napis (2015). ‘A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data’. *Curr. Proteomics* 12 (1), pp. 14–27.

Shi, T., E. Song, S. Nie, K. D. Rodland, T. Liu, W.-J. Qian, and R. D. Smith (2016). ‘Advances in targeted proteomics and applications to biomedical research’. *Proteomics* 16 (15-16), pp. 2160–2182.

Shiferaw, G. A., E. Vandermarliere, N. Hulstaert, R. Gabriels, L. Martens, and P.-J. Volders (2020). ‘COSS: A Fast and User-Friendly Tool for Spectral Library Searching’. *J. Proteome Res.* 19 (7), pp. 2786–2793.

Shokolenko, I. N. and M. F. Alexeyev (2017). ‘Mitochondrial transcription in mammalian cells’. *Front. Biosci.* 22, pp. 835–853.

Shteynberg, D., E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold, and A. I. Nesvizhskii (2011). ‘iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates’. *Mol. Cell. Proteom.* 10 (12), p. M111.007690.

Shteynberg, D., A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch (2013). ‘Combining results of multiple search engines in proteomics’. *Mol. Cell. Proteom.* 12 (9), pp. 2383–2393.

Sikdar, S., R. Gill, and S. Datta (2016). ‘Improving protein identification from tandem mass spectrometry data by one-step methods and integrating data from other platforms’. *Brief. Bioinformatics* 17 (2), pp. 262–269.

Silva, J. C., M. V. Gorenstein, G.-Z. Li, J. P. C. Vissers, and S. J. Geromanos (2006). ‘Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition’. *Mol. Cell. Proteom.* 5 (1), pp. 144–156.

Sinitcyn, P., J. D. Rudolph, and J. Cox (2018). ‘Computational Methods for Understanding Mass Spectrometry-Based Shotgun Proteomics Data’. *Annu. Rev. Biomed. Data Sci.* Annual Review of Biomedical Data Science 1. Ed. by R. B. Altman and M. Levitt, pp. 207–234.

Slavin, R. E. (1986). ‘Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews’. *Educ. Res.* 15 (9), pp. 5–11.

Smith, R. N., J. Aleksic, D. Butano, A. Carr, S. Contrino, F. Hu, M. Lyne, R. Lyne, A. Kalderimis, K. Rutherford, R. Stepan, J. Sullivan, M. Wakeling, X. Watkins, and G. Micklem (2012). ‘InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data’. *Bioinformatics* 28 (23), pp. 3163–3165.

Spivak, M., J. Weston, L. Bottou, L. Käll, and W. S. Noble (2009). ‘Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets’. *J. Proteome Res.* 8 (7), pp. 3737–3745.

Spivak, M., J. Weston, D. Tomazela, M. J. MacCoss, and W. S. Noble (2012). 'Direct maximization of protein identifications from tandem mass spectra'. *Mol. Cell. Proteom.* 11 (2), p. M111.012161.

Stairs, C. W., M. M. Leger, and A. J. Roger (2015). 'Diversity and origins of anaerobic metabolism in mitochondria and related organelles'. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 370 (1678), p. 20140326.

Stanke, M., R. Steinkamp, S. Waack, and B. Morgenstern (2004). 'AUGUSTUS: a web server for gene finding in eukaryotes'. *Nucleic Acids Res.* 32 (Web Server issue), W309–12.

Starr, A. E., S. A. Deeke, L. Li, X. Zhang, R. Daoud, J. Ryan, Z. Ning, K. Cheng, L. V. H. Nguyen, E. Abou-Samra, M. Lavallée-Adam, and D. Figeys (2018). 'Proteomic and Metaproteomic Approaches to Understand Host-Microbe Interactions'. *Anal. Chem.* 90 (1), pp. 86–109.

Stead, D. A., N. W. Paton, P. Missier, S. M. Embury, C. Hedeler, B. Jin, A. J. P. Brown, and A. Preece (2008). 'Information quality in proteomics'. *Briefings Bioinf.* 9 (2), pp. 174–188.

Steen, H. and M. Mann (2004). 'The ABC's (and XYZ's) of peptide sequencing'. *Nat. Rev. Mol. Cell Biol.* 5 (9), pp. 699–711.

Stegle, O., L. Parts, M. Piipari, J. Winn, and R. Durbin (2012). 'Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses'. *Nat. Protoc.* 7 (3), pp. 500–507.

Steiner, C., A. Ducret, J.-C. Tille, M. Thomas, T. A. McKee, L. Rubbia-Brandt, A. Scherl, P. Lescuyer, and P. Cutler (2014). 'Applications of mass spectrometry for quantitative protein analysis in formalin-fixed paraffin-embedded tissues'. *Proteomics* 14 (4-5), pp. 441–451.

Stelpflug, S. C., R. S. Sekhon, B. Vaillancourt, C. N. Hirsch, C. R. Buell, N. de Leon, and S. M. Kaeppler (2016). 'An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development'. *Plant Genome* 9 (1).

Student (Gosset, W. S. (1908). 'The probable error of a mean'. *Biometrika* 6 (1), pp. 1–25.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch (2004). 'A gene atlas of the mouse and human protein-encoding transcriptomes'. *PNAS* 101 (16), pp. 6062–6067.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles'. *PNAS* 102 (43), pp. 15545–15550.

Sudmant, P. H., M. S. Alexis, and C. B. Burge (2015). 'Meta-analysis of RNA-seq expression data across species, tissues and studies'. *Genome Biol.* 16, p. 287.

Sultan, M., V. Amstislavskiy, T. Risch, M. Schuette, S. Dökel, M. Ralser, D. Balzereit, H. Lehrach, and M.-L. Yaspo (2014). 'Influence of RNA extraction methods and library selection schemes on RNA-seq data'. *BMC Genom.* 15 (1), p. 675.

Sun, Y., U. Braga-Neto, and E. R. Dougherty (2012). 'A systematic model of the LC-MS proteomics pipeline'. *BMC Genom.* 13 Suppl 6, S2.

Suntsova, M., N. Gaifullin, D. Allina, A. Reshetun, X. Li, L. Mendeleeva, V. Surin, A. Sergeeva, P. Spirin, V. Prassolov, A. Morgan, A. Garazha, M. Sorokin, and A. Buzdin (2019). 'Atlas of RNA sequencing profiles for normal human tissues'. *Sci. Data* 6, p. 36.

Sutandy, F. X. R., J. Qian, C.-S. Chen, and H. Zhu (2013). 'Overview of protein microarrays'. *Curr. Protoc. Protein Sci.* Chapter 27, Unit 27.1.

Syka, J. E. P., J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt (2004). 'Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry'. *PNAS* 101 (26), pp. 9528–9533.

Tabb, D. L. (2015). 'The SEQUEST family tree'. *J. Am. Soc. Mass Spectrom.* 26 (11), pp. 1814–1819.

Tabb, D. L., Z.-Q. Ma, D. B. Martin, A.-J. L. Ham, and M. C. Chambers (2008). 'DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring'. *J. Proteome Res.* 7 (9), pp. 3838–3846.

Tabb, D. L., W. H. McDonald, and J. R. Yates 3rd (2002). 'DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics'. *J. Proteome Res.* 1 (1), pp. 21–26.

Tabb, D. L., A. Saraf, and J. R. Yates 3rd (2003). 'GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model'. *Anal. Chem.* 75 (23), pp. 6415–6421.

Tabb, D. L., L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A.-J. L. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. A. Carr, K. R. Clauser, J. D. Jaffe, K. A. Kowalski, T. A. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman (2010). 'Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry'. *J. Proteome Res.* 9 (2), pp. 761–776.

Tamayo, P., G. Steinhardt, A. Liberzon, and J. P. Mesirov (2012). 'The limitations of simple gene set enrichment analysis assuming gene independence'. *Stat. Methods Med. Res.* 25 (1), pp. 472–487.

Taminau, J., C. Lazar, S. Meganck, and A. Nowé (2014). 'Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis'. *ISRN Bioinform.* 2014, p. 345106.

Tang, H., R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly, and P. Radivojac (2006). 'A computational approach toward label-free protein quantification using predicted peptide detectability'. *Bioinformatics* 22 (14), e481–8.

Tanner, S., H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna (2005). 'InsPecT: identification of posttranslationally modified peptides from tandem mass spectra'. *Anal. Chem.* 77 (14), pp. 4626–4639.

Taylor, J. A. and R. S. Johnson (1997). 'Sequence database searches via de novo peptide sequencing by tandem mass spectrometry'. *Rapid Commun. Mass Spectrom.* 11 (9), pp. 1067–1075.

The UniProt Consortium (2017). 'UniProt: the universal protein knowledgebase'. *Nucleic Acids Res.* 45 (D1), pp. D158–D169.

The, M., F. Edfors, Y. Perez-Riverol, S. H. Payne, M. R. Hoopmann, M. Palmblad, B. Forsström, and L. Käll (2018). 'A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms'. *J. Proteome Res.* 17 (5), pp. 1879–1886.

The, M., M. J. MacCoss, W. S. Noble, and L. Käll (2016). 'Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0'. *J. Am. Soc. Mass Spectrom.* 27 (11), pp. 1719–1727.

The, M., A. Tasnim, and L. Käll (2016). 'How to talk about protein-level false discovery rates in shotgun proteomics'. *Proteomics* 16 (18), pp. 2461–2469.

Thompson, A., J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. A. Mohammed, and C. Hamon (2003). 'Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS'. *Anal. Chem.* 75 (8), pp. 1895–1904.

Thomson, J. J. (1913). 'Rays of Positive Electricity'. *Proc. R. Soc. A* 89 (607), pp. 1–20.

Tian, Q., S. B. Stepaniants, M. Mao, L. Weng, M. C. Feetham, M. J. Doyle, E. C. Yi, H. Dai, V. Thorsson, J. Eng, D. Goodlett, J. P. Berger, B. Gunter, P. S. Linseley, R. B. Stoughton, R. Aebersold, S. J. Collins, W. A. Hanlon, and L. E. Hood (2004). 'Integrated genomic and proteomic analyses of gene expression in Mammalian cells'. *Mol. Cell. Proteom.* 3 (10), pp. 960–969.

Tipney, H. and L. Hunter (2010). 'An introduction to effective use of enrichment analysis software'. *Hum. Genomics* 4 (3), pp. 202–206.

Tran, N. H., X. Zhang, L. Xin, B. Shan, and M. Li (2017). ‘De novo peptide sequencing by deep learning’. *PNAS* 114 (31), pp. 8247–8252.

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

Tsiatsiani, L. and A. J. R. Heck (2015). ‘Proteomics beyond trypsin’. *FEBS J.* 282 (14), pp. 2612–2626.

Tu, C., J. Li, S. Shen, Q. Sheng, Y. Shyr, and J. Qu (2016). ‘Performance Investigation of Proteomic Identification by HCD/CID Fragmentations in Combination with High/Low-Resolution Detectors on a Tribrid, High-Field Orbitrap Instrument’. *PLOS ONE* 11 (7), e0160160.

Tu, C., J. Li, Q. Sheng, M. Zhang, and J. Qu (2014). ‘Systematic assessment of survey scan and MS2-based abundance strategies for label-free quantitative proteomics using high-resolution MS data’. *J. Proteome Res.* 13 (4), pp. 2069–2079.

Tu, C., Q. Sheng, J. Li, D. Ma, X. Shen, X. Wang, Y. Shyr, Z. Yi, and J. Qu (2015). ‘Optimization of Search Engines and Postprocessing Approaches to Maximize Peptide and Protein Identification for High-Resolution Mass Data’. *J. Proteome Res.* 14 (11), pp. 4662–4673.

Turner, S. D. (2015). *RNA-SEQ quality control and analysis of Differential gene expression using the TUXEDO software suite.*

Tyanova, S., T. Temu, and J. Cox (2016). ‘The MaxQuant computational platform for mass spectrometry-based shotgun proteomics’. *Nat. Protoc.* 11 (12), pp. 2301–2319.

Tyanova, S., T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox (2016). ‘The Perseus computational platform for comprehensive analysis of (prote)omics data’. *Nat. Methods* 13 (9), pp. 731–740.

Uhlén, M., 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. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén (2015). ‘Tissue-based map of the human proteome’. *Science* 347 (6220).

Uhlén, M., B. M. Hallström, C. Lindskog, A. Mardinoglu, F. Pontén, and J. Nielsen (2016). ‘Transcriptomics resources of human tissues and organs’. *Mol. Syst. Biol.* 12 (4).

Unlü, M., M. E. Morgan, and J. S. Minden (1997). ‘Difference gel electrophoresis: a single gel method for detecting changes in protein extracts’. *Electrophoresis* 18 (11), pp. 2071–2077.

Urbanek, S. and J. Horner (2019). *Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output.* R package version 1.5-10.

Ushey, K., J. McPherson, J. Cheng, A. Atkins, and J. Allaire (2018). *packrat: A Dependency Management System for Projects and their R Package Dependencies.* R package version 0.5.0.

Uszkoreit, J., A. Maerkens, Y. Perez-Riverol, H. E. Meyer, K. Marcus, C. Stephan, O. Kohlbacher, and M. Eisenacher (2015). ‘PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface’. *J. Proteome Res.* 14 (7), pp. 2988–2997.

Välikangas, T., T. Suomi, and L. L. Elo (2018a). ‘A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation’. *Briefings Bioinf.* 19 (6), pp. 1344–1355.

Välikangas, T., T. Suomi, and L. L. Elo (2018b). ‘A systematic evaluation of normalization methods in quantitative label-free proteomics’. *Briefings Bioinf.* 19 (1), pp. 1–11.

Van Dijk, E. L., H. Auger, Y. Jaszczyszyn, and C. Thermes (2014). ‘Ten years of next-generation sequencing technology’. *Trends Genet.* 30 (9), pp. 418–426.

Van Leeuwen, J. and J. Leeuwen (1990). *Handbook of Theoretical Computer Science*. Cambridge, MA, US: Elsevier.

VanGuilder, H. D., K. E. Vrana, and W. M. Freeman (2008). 'Twenty-five years of quantitative PCR for gene expression analysis'. *BioTechniques* 44 (5), pp. 619–626.

Väremo, L., C. Scheele, C. Broholm, A. Mardinoglu, C. Kampf, A. Asplund, I. Nookaew, M. Uhlén, B. K. Pedersen, and J. Nielsen (2015). 'Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes'. *Cell Rep.* 11 (6), pp. 921–933.

Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett Jr, P. Hieter, B. Vogelstein, and K. W. Kinzler (1997). 'Characterization of the yeast transcriptome'. *Cell* 88 (2), pp. 243–251.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. New York, NY, US: Springer.

Venter, J. C. et al. (2001). 'The sequence of the human genome'. *Science* 291 (5507), pp. 1304–1351.

Visser, N. F. C., H. Lingeman, and H. Irth (2005). 'Sample preparation for peptides and proteins in biological matrices prior to liquid chromatography and capillary zone electrophoresis'. *Anal. Bioanal. Chem.* 382 (3), pp. 535–558.

Vitek, O. (2009). 'Getting started in computational mass spectrometry-based proteomics'. *PLOS Comput. Biol.* 5 (5), e1000366.

Vogel, C., R. d. S. Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva (2010). 'Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line'. *Mol. Syst. Biol.* 6 (1), p. 400.

Vogel, C. and E. M. Marcotte (2012). 'Insights into the regulation of protein abundance from proteomic and transcriptomic analyses'. *Nat. Rev. Genet.* 13 (4), pp. 227–232.

Vyatkina, K., S. Wu, L. J. M. Dekker, M. M. VanDuijn, X. Liu, N. Tolić, M. Dvorkin, S. Alexandrova, T. M. Luider, L. Paša-Tolić, and P. A. Pevzner (2015). 'De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra'. *J. Proteome Res.* 14 (11), pp. 4450–4462.

Wagner, G. P., K. Kin, and V. J. Lynch (2012). 'Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples'. *Theory Biosci.* 131 (4), pp. 281–285.

Wajid, B. and E. Serpedin (2012). 'Review of general algorithmic features for genome assemblers for next generation sequencers'. *Genomics, Proteomics & Bioinformatics* 10 (2), pp. 58–73.

Walsh, C. J., P. Hu, J. Batt, and C. C. D. Santos (2015). 'Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery'. *Microarrays* 4 (3), pp. 389–406.

Walther, T. C. and M. Mann (2010). 'Mass spectrometry–based proteomics in cell biology'. *J. Cell Biol.* 190 (4), pp. 491–500.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander (1998). 'Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome'. *Science* 280 (5366), pp. 1077–1082.

Wang, D., B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, and B. Kuster (2019). 'A deep proteome and transcriptome abundance atlas of 29 healthy human tissues'. *Mol. Syst. Biol.* 15 (2), e8503.

Wang, E. T., R. Sandberg, R. Sandberg, R. Sandberg, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge (2008). 'Alternative isoform regulation in human tissue transcriptomes'. *Nature* 456 (7221), pp. 470–476.

Wang, G., W. W. Wu, Z. Zhang, S. Masilamani, and R.-F. Shen (2009). ‘Decoy methods for assessing false positives and false discovery rates in shotgun proteomics’. *Anal. Chem.* 81 (1), pp. 146–159.

Wang, J., J. Pérez-Santiago, J. E. Katz, P. Mallick, and N. Bandeira (2010). ‘Peptide identification from mixture tandem mass spectra’. *Mol. Cell. Proteom.* 9 (7), pp. 1476–1485.

Wang, Q., J. Armenia, C. Zhang, A. V. Penson, E. Reznik, L. Zhang, T. Minet, A. Ochoa, B. E. Gross, C. A. Iacobuzio-Donahue, D. Betel, B. S. Taylor, J. Gao, and N. Schultz (2017). ‘Enabling cross-study analysis of RNA-Sequencing data’. *bioRxiv* (110734).

Wang, S., I. Pandis, D. Johnson, I. Emam, F. Guitton, A. Oehmichen, and Y. Guo (2014). ‘Optimising parallel R correlation matrix calculations on gene expression data using MapReduce’. *BMC Bioinf.* 15, p. 351.

Wang, X., S. Shen, S. S. Rasam, and J. Qu (2019a). ‘MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts’. *Mass Spectrom. Rev.* 38 (6), pp. 461–482.

Wang, X., S. Shen, S. S. Rasam, and J. Qu (2019b). ‘MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts’. *Mass Spectrom. Rev.* 38 (6), pp. 461–482.

Wang, Z., M. Gerstein, and M. Snyder (2009). ‘RNA-Seq: a revolutionary tool for transcriptomics’. *Nat. Rev. Genet.* 10 (1), pp. 57–63.

Ward, J. H. (1963). ‘Hierarchical Grouping to Optimize an Objective Function’. *J. Am. Stat. Assoc.* 58 (301), pp. 236–244.

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables (2019). *gplots: Various R Programming Tools for Plotting Data.* R package version 3.0.1.1.

Webb-Robertson, B.-J. M., H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, and K. M. Waters (2015). ‘Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics’. *J. Proteome Res.* 14 (5), pp. 1993–2001.

Weir, A. (2014). *The Martian.* New York, NY, US: Penguin Random House.

Weisser, H., J. C. Wright, J. M. Mudge, P. Gutenbrunner, and J. S. Choudhary (2016). ‘Flexible Data Analysis Pipeline for High-Confidence Proteogenomics’. *J. Proteome Res.* 15 (12), pp. 4686–4695.

Welch, B. L. (1947). ‘The generalization of Student’s problem when several different population variances are involved’. *Biometrika* 34 (1-2), pp. 28–35.

Welch, B. L. (1951). ‘On the Comparison of Several Mean Values: An Alternative Approach’. *Biometrika* 38 (3/4), pp. 330–336.

Wenger, C. D. and J. J. Coon (2013). ‘A proteomics search algorithm specifically designed for high-resolution tandem mass spectra’. *J. Proteome Res.* 12 (3), pp. 1377–1386.

Westerhoff, H. V., C. Winder, H. Messiha, E. Simeonidis, M. Adamczyk, M. Verma, F. J. Bruggeman, and W. Dunn (2009). ‘Systems biology: the elements and principles of life’. *FEBS Lett.* 583 (24), pp. 3882–3890.

Wheeler, D. L., D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, and L. Wagner (2003). ‘Database resources of the National Center for Biotechnology’. *Nucleic Acids Res.* 31 (1), pp. 28–33.

Wickham, H. (2007). ‘Reshaping Data with the reshape Package’. *J. Stat. Softw.* 21 (12), pp. 1–20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY, US: Springer-Verlag New York.

Wickham, H. (2018). *scales: Scale Functions for Visualization.* R package version 1.0.0.

Wickham, H., J. Hester, and W. Chang (2019). *devtools: Tools to Make Developing R Packages Easier.* R package version 2.1.0.

Wilhelm, M., J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster (2014). 'Mass-spectrometry-based draft of the human proteome'. *Nature* 509 (7502), pp. 582–587.

Williams, C. R., A. Baccarella, J. Z. Parrish, and C. C. Kim (2016). 'Trimming of sequence reads alters RNA-Seq gene expression estimates'. *BMC Bioinf.* 17, p. 103.

Wiśniewski, J. R., M. Y. Hein, J. Cox, and M. Mann (2014). 'A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards'. *Mol. Cell. Proteom.* 13 (12), pp. 3497–3506.

Wiśniewski, J. R., A. Zougman, N. Nagaraj, and M. Mann (2009). 'Universal sample preparation method for proteome analysis'. *Nat. Methods* 6 (5), pp. 359–362.

Wood, S. N. (2004). 'Stable and efficient multiple smoothing parameter estimation for generalized additive models'. *J. Am. Stat. Assoc.* 99 (467), pp. 673–686.

Wright, J. C. and J. S. Choudhary (2016). 'DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics'. *J. Proteom. Bioinform.* 9 (6), pp. 176–180.

Wright, J. C., M. O. Collins, L. Yu, L. Käll, M. Brosch, and J. S. Choudhary (2012). 'Enhanced peptide identification by electron transfer dissociation using an improved Mascot Percolator'. *Mol. Cell. Proteom.* 11 (8), pp. 478–491.

Wright, J. C., J. Mudge, H. Weisser, M. P. Barzine, J. M. Gonzalez, A. Brazma, J. S. Choudhary, and J. Harrow (2016). 'Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow'. *Nat. Commun.* 7, p. 11778.

Wu, C., I. Macleod, and A. I. Su (2013). 'BioGPS and MyGene.info: organizing online, gene-centric information'. *Nucleic Acids Res.* 41 (Database issue), pp. D561–5.

Wu, C., C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss 3rd, and A. I. Su (2009). 'BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources'. *Genome Biol.* 10 (11), R130.

Wu, C., J. C. Tran, L. Zamdborg, K. R. Durbin, M. Li, D. R. Ahlf, B. P. Early, P. M. Thomas, J. V. Sweedler, and N. L. Kelleher (2012). 'A protease for 'middle-down' proteomics'. *Nat. Methods* 9 (8), pp. 822–824.

Wu, X., C.-W. Tseng, and N. Edwards (2007). 'HMMatch: peptide identification by spectral matching of tandem mass spectra using hidden Markov models'. *J. Comput. Biol.* 14 (8), pp. 1025–1043.

Wysocki, V. H., K. A. Resing, Q. Zhang, and G. Cheng (2005). 'Mass spectrometry of peptides and proteins'. *Methods* 35 (3), pp. 211–222.

Xiao, S.-J., C. Zhang, Q. Zou, and Z.-L. Ji (2010). 'TiSGeD: a database for tissue-specific genes'. *Bioinformatics* 26 (9), pp. 1273–1275.

Xie, Y., J. Allaire, and G. Grolemund (2018). *R Markdown: The Definitive Guide.* Boca Raton, FL, US: Chapman and Hall/CRC.

Xie, Y., J. Cheng, and X. Tan (2019). *DT: A Wrapper of the JavaScript Library 'DataTables'.* R package version 0.6.

Xu, M., Z. Li, and L. Li (2013). 'Combining percolator with X!Tandem for accurate and sensitive peptide identification'. *J. Proteome Res.* 12 (6), pp. 3026–3033.

Yagüe, J., A. Paradela, M. Ramos, S. Ogueta, A. Marina, F. Barahona, J. A. López de Castro, and J. Vázquez (2003). 'Peptide rearrangement during quadrupole ion trap fragmentation: added complexity to MS/MS spectra'. *Anal. Chem.* 75 (6), pp. 1524–1535.

Yang, H., H. Chi, W.-F. Zeng, W.-J. Zhou, and S.-M. He (2019). 'pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework'. *Bioinformatics* 35 (14), pp. i183–i190.

Yang, X., V. Dondeti, R. Dezube, D. M. Maynard, L. Y. Geer, J. Epstein, X. Chen, S. P. Markey, and J. A. Kowalak (2004). 'DBParser: web-based software for shotgun proteomic data analyses'. *J. Proteome Res.* 3 (5), pp. 1002–1008.

Yao, L., H. Wang, Y. Song, and G. Sui (2017). 'BioQueue: a novel pipeline framework to accelerate bioinformatics analysis'. *Bioinformatics* 33 (20), pp. 3286–3288.

Yao, S., C. Jiang, Z. Huang, I. Torres-Jerez, J. Chang, H. Zhang, M. Udvardi, R. Liu, and J. Verdier (2016). 'The Vigna unguiculata Gene Expression Atlas (VuGEA) from de novo assembly and quantification of RNA-seq data provides insights into seed maturation mechanisms'. *Plant J.* 88 (2), pp. 318–327.

Ye, D., Y. Fu, R.-X. Sun, H.-P. Wang, Z.-F. Yuan, H. Chi, and S.-M. He (2010). 'Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate'. *Bioinformatics* 26 (12), pp. i399–406.

Ye, X., B. Luke, T. Andresson, and J. Blonder (2009). '18O stable isotope labeling in MS-based proteomics'. *Brief. Funct. Genom. Proteomics* 8 (2), pp. 136–144.

Yeung, E. S. (2011). 'Genome-wide correlation between mRNA and protein in a single cell'. *Angew. Chem.* 50 (3), pp. 583–585.

Yi, H., A. T. Raman, H. Zhang, G. I. Allen, and Z. Liu (2018). 'Detecting hidden batch factors through data-adaptive adjustment for biological effects'. *Bioinformatics* 34 (7), pp. 1141–1147.

Yi, H., L. Xue, M.-X. Guo, J. Ma, Y. Zeng, W. Wang, J.-Y. Cai, H.-M. Hu, H.-B. Shu, Y.-B. Shi, and W.-X. Li (2010). 'Gene expression atlas for human embryogenesis'. *FASEB J.* 24 (9), pp. 3341–3350.

Yu, G., L.-G. Wang, Y. Han, and Q.-Y. He (2012). 'clusterProfiler: an R package for comparing biological themes among gene clusters'. *OMICS* 16 (5), pp. 284–287.

Yu, N. Y.-L., B. M. Hallström, L. Fagerberg, F. Ponten, H. Kawaji, P. Carninci, A. R. R. Forrest, Fantom Consortium, Y. Hayashizaki, M. Uhlén, and C. O. Daub (2015). 'Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium'. *Nucleic Acids Res.* 43 (14), pp. 6787–6798.

Yu, X., J. Lin, D. J. Zack, and J. Qian (2006). 'Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues'. *Nucleic Acids Res.* 34 (17), pp. 4925–4936.

Zhang, J., L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie, and B. Ma (2012). 'PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification'. *Mol. Cell. Proteom.* 11 (4), p. M111.010587.

Zhang, Y., Q. Li, F. Wu, R. Zhou, Y. Qi, N. Su, L. Chen, S. Xu, T. Jiang, C. Zhang, G. Cheng, X. Chen, D. Kong, Y. Wang, T. Zhang, J. Zi, W. Wei, Y. Gao, B. Zhen, Z. Xiong, S. Wu, P. Yang, Q. Wang, B. Wen, F. He, P. Xu, and S. Liu (2015). 'Tissue-Based Proteogenomics Reveals that Human Testis Endows Plentiful Missing Proteins'. *J. Proteome Res.* 14 (9), pp. 3583–3594.

Zhang, Y., B. R. Fonslow, B. Shan, M.-C. Baek, and J. R. Yates 3rd (2013). 'Protein analysis by shotgun/bottom-up proteomics'. *Chemical Rev.* 113 (4), pp. 2343–2394.

Zhang, Z., S. Wu, D. L. Stenoien, and L. Paša-Tolić (2014). 'High-throughput proteomics'. *Annu. Rev. Anal. Chem.* 7, pp. 427–454.

Zhao, S. (2014). 'Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads'. *PLOS ONE* 9 (7), e101374.

Zhao, S., Y. Zhang, R. Gamini, B. Zhang, and D. von Schack (2018). 'Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion'. *Sci. Rep.* 8, p. 4781.

Zhou, Y., Y. Shan, L. Zhang, and Y. Zhang (2014). 'Recent advances in stable isotope labeling based techniques for proteome relative quantification'. *J. Chromatogr. A* 1365, pp. 1–11.

Zhu, J., G. Chen, S. Zhu, S. Li, Z. Wen, Bin Li, Y. Zheng, and L. Shi (2016). ‘Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq’. *Sci. Rep.* 6, p. 28400.

Zhuang, F., R. T. Fuchs, and G. B. Robb (2012). ‘Small RNA Expression Profiling by High-Throughput Sequencing: Implications of Enzymatic Manipulation’. *J. Nucleic Acids* 2012.

Zhuo, B., S. Emerson, J. H. Chang, and Y. Di (2016). ‘Identifying stably expressed genes from multiple RNA-Seq data sets’. *PeerJ* 4, e2791.

Zwiener, I., B. Frisch, and H. Binder (2014). ‘Transforming RNA-Seq data to improve the performance of prognostic gene signatures’. *PLOS ONE* 9 (1), e85150.

Zyprych-Walczak, J., A. Szabelska, L. Handschuh, K. Górczak, K. Klamecka, M. Figlerowicz, and I. Siatkowski (2015). ‘The Impact of Normalization Methods on RNA-Seq Data Analysis’. *BioMed Res. Int.* 2015, p. 621690.